INSTITUT POLYTECHNIQUE DE PARIS

NNT : 2024IPPAT032

Thèse de doctorat

Statistical Learning for Spatial Data: Theory and Algorithms

Thèse de doctorat de l'Institut Polytechnique de Paris
préparée à Télécom Paris

École doctorale n°574 École doctorale de Mathématiques Hadamard (EDMH)
Spécialité de doctorat : Mathématiques appliquées

Thèse présentée et soutenue à Palaiseau, le 2 décembre 2024, par

EMILIA SIVIERO

Composition du Jury :

Viet-Chi Tran Professeur, Université Gustave Eiffel (LAMA)	Président/Examinateur
Céline Lévy-Leduc Professeure, Université Paris Cité (LPSM)	Rapportrice
Christophe Denis Professeur, Université Paris 1 Panthéon-Sorbonne (SAMM)	Rapporteur
Florence d'Alché-Buc Professeure, Télécom Paris (LTCI)	Examinatrice
Odalric-Ambrym Maillard Chargé de recherche, Inria Lille - Nord Europe	Examinateur
Stéphan Clémençon Professeur, Télécom Paris (LTCI)	Directeur de thèse

À ma mère, mon père, et mon frère.
À ma grand-mère Mayo, a nonno Sergio.

Contents

Remerciements

Je souhaite débuter cette thèse en remerciant celles et ceux qui m'ont accompagnée et soutenue durant ces années de doctorat.

Mes premiers remerciements vont à ma famille. Merci Maman pour tout, pour tes conseils, ton écoute et ta motivation. Sans toi, je ne serais pas ici à écrire les dernières lignes de cette thèse. Merci également d'être le meilleur exemple à suivre, aussi bien en tant que chercheuse qu'en tant que personne. Merci Papà pour m'avoir toujours aidée dans tout ce que j'entreprends, pour avoir toujours su être près de moi malgrè la distance, ainsi que pour ta patience et tes précieux conseils. Merci Julien, pour nos discussions pendant les pauses durant l'écriture de ce manuscrit, pour m'avoir remonté le moral de nombreuses fois et pour m'aider à voir le monde sous une autre perspective. Je suis très fière et heureuse de pouvoir t'appeler grand frère. Merci Clément pour ton hospitalité de nombreuses fois à Marseille, je n'oublie pas tous ces beaux moments passés à la plage ou, principalement, au cours Ju. J'espère qu'on pourra refaire cela à Nice ou à Venise, on a le choix !

Merci à Stéphan, mon directeur de thèse. Je te suis reconnaissante de m'avoir donné l'opportunité de réaliser cette thèse et de m'avoir initiée aux statistiques spatiales. J'ai beaucoup appris pendant ces années de thèse et je le dois en grande partie à toi. Merci également à Emilie, ton aide et tes conseils lors de la première année de thèse ont été précieux.

Je souhaite vivement remercier les membres de mon jury d'avoir accepté d'évaluer mon travail. Je remercie sincèrement Céline Lévy-Leduc et Christophe Denis d'avoir rapporté cette thèse. Un grand merci à l'ensemble de mon jury, Viet-Chi Tran, Odalric-Ambrym Maillard et Florence d'Alché-Buc, pour avoir accepté d'être présents lors de ma soutenance de thèse. Je vous suis très reconnaissante.

Merci Guillaume, la deuxième partie de cette thèse n'aurait pas été la même sans toi. Merci pour ta patience, tes conseils et ton expertise. Tu m'as redonné goût à la recherche à un moment difficile de la thèse, je t'en suis très reconnaissante.

Merci Tamim, Marc et Jr, même si ces années ont parfois pu paraître longues et fatiguantes, vous les avez rendues bien plus belles et souriantes que prévu. Merci Jr pour ta présence constante au *laBorAtoiRe*, toujours au top ! Bon courage pour la fin de ta thèse, j'ai très hâte qu'on puisse fêter tous les quatre ensemble. Merci Marc d'avoir supporté toutes mes discussions pendant les pauses, pour tes précieux conseils en anglais et ta bonne humeur. Merci Tamim pour avoir été là du début à la fin de cette thèse, pour ta gentillesse, ton écoute et ton aide ; je ne pouvais pas espérer de meilleur co-bureau (et ami) ! Merci également au reste de la famille El Ahmad, Nour, Zahira et Youssef, pour votre gentillesse et votre hospitalité chaleureuse. Je garderai précieusement en mémoire toutes les discussions et moments échangés lors des manifestations et autour d'excellents repas.

Merci Junjie pour ta joie débordante, ton sourire tous les jours au labo, et pour m'avoir réconfortée de nombreuses fois. Je suis de tout cœur avec toi pour la fin de ta thèse, c'est presque terminé pour toi aussi ! Merci beaucoup Mathilde, tu es le rayon de soleil

qui manquait à l'équipe, tu as changé beaucoup de choses au labo simplement avec ta bonne humeur et ton sourire, et tu as été la sœur de thèse dont je ne savais pas avoir besoin. Je te souhaite le meilleur pour la suite de ta thèse, et même si je suis contente d'en avoir fini avec la mienne, nos pauses ensemble me manquent déjà. Merci d'avance à mon futur coloc Nathan pour cuisiner tous les jours pour nous à Venise... J'ai très hâte de commencer cette nouvelle expérience, et je suis très heureuse qu'on puisse la vivre ensemble !

Vous six êtes les plus belles rencontres que j'ai faites au cours de cette thèse !

Merci à tous les gens du labo, ancien.ne.s comme nouvelles.aux, pour toutes les discussions autour d'un café ou d'un déjeuner: Dimitri, Lucien, Anass, Louise, Iyad, Ikhlas, Jérémy, Joël, Quentin, Lilian, Paul, Arturo... Un grand merci à Louise et Ikhlas pour avoir repris l'atelier de français, et bon courage à vous deux pour la suite de votre thèse.

Je souhaite remercier spécialement mes amies, ou plutôt mes soeurs, vous avez eu un impact non négligeable sur cette thèse. Merci à Ariane et Morgane, pour tous les weekends à Bordeaux, les soirées crochet/tarot, et toutes nos discussions. Merci d'avoir toujours été proches de moi, j'espère pouvoir en faire autant pour vous à l'avenir. Merci à Prune, Garance et Margot. Il s'en est passé des choses depuis toutes ces années, mais nous sommes restées les mêmes toutes les quatre. Même si on se retrouve aux quatre coins du monde, je sais que je peux compter sur chacune de vous. Merci pour ces longs appels et pour m'avoir fait faire le tour du monde, même à distance. Je vous aime très fort. Un grand merci à Garance, pour ta patience de nombreux matins, sans toi j'aurais raté beaucoup de réveils... Merci également à Mathilde, se retrouver après le lycée a été une magnifique surprise, et je garde précieusement en mémoire tous nos moments à Rome ou sur une plage avec un spritz ! J'espère qu'on continuera nos discussions, partout en Italie, sur le féminisme et tous les sujets de révolte.

Merci à tou.te.s mes ami.e.s rencontré.e.s au fil de ces longues années d'études : vous représentez la meilleure partie de chaque année. Le TD6 (Ariane, Valentin, Nacim, Lucie) pour la L1 ; Ezio, Morgane, Rémi (eh oui, ça m'arrive d'utiliser ton vrai prénom) et Jeanne pour la L2 et la L3 ; Victor et Olivier pour le M1 ; Valentin pour le M2. Lorsque je repense à mes études, c'est à vous que je pense en premier.

Merci à Madame Sylvie Coussot, vous m'avez aidée à un moment important et vous avez eu un impact décisif pour la suite de la thèse. Je vous en serai toujours reconnaissante. J'espère que vous pourrez continuer à aider d'autres doctorant.e.s, et que chacun ait la chance de pouvoir être soutenu par vous.

Merci à Laurence, Delphine et Janique pour tout ce que vous faites pour les doctorants du département ! Merci pour votre patience et votre aide, la vie doctorante au sein de l'IDS ne serait pas la même sans vous.

Merci Avner pour ton soutien tout au long de ces années d'études, depuis le stage en licence jusqu'à la fin de cette thèse. Nos discussions m'ont aidée à plusieurs reprises à choisir la voie à suivre. Merci également pour l'entraînement de soutenance, tes remarques m'ont permis d'améliorer la présentation et m'ont donné confiance pour cette dernière étape. Merci Stéphane pour tes précieux conseils sur la recherche, ainsi que pour ton aide à comprendre cette mystérieuse transformation de Fourier.

Abstract

In the Big Data era, thanks to the ubiquity of geolocation sensors in particular, massive datasets exhibiting a possibly complex spatial dependence structure are becoming increasingly available. In this thesis, we aim at developing approaches to efficiently exploit the dependence structure of spatial (and spatio-temporal) data.

We first analyze the *simple Kriging* task, the flagship problem in Geostatistics, from a statistical learning perspective, *i.e.* by carrying out a non-parametric finite-sample predictive analysis. In this context, the standard probabilistic theory of statistical learning does not apply directly and theoretical guarantees of the generalization capacity of the Kriging predictive rule learned from spatial data are left to be established. Given $d \geq 1$ values taken by a realization of a square integrable random field $\mathbf{X} = \{\mathbf{X}_s\}_{s \in \mathcal{S}}$, $\mathcal{S} \subset \mathbb{R}^2$, with unknown covariance structure, at sites $s_1, \ldots, s_d$ in $\mathcal{S}$, the goal is to predict the unknown values that $\mathbf{X}$ takes at any other location $s \in \mathcal{S}$ with minimum quadratic risk. The prediction rule is derived from a training spatial dataset: a single realization $\mathbf{X}'$ of $\mathbf{X}$, independent from those to be predicted, observed at $n \geq 1$ locations $\sigma_1, \ldots, \sigma_n$ in $\mathcal{S}$. Despite the connection of this minimization problem with kernel ridge regression, establishing the generalization capacity of empirical risk minimizer is far from straightforward, due to the non independent and identically distributed nature of the training data $\mathbf{X}'_{\sigma_1}, \ldots, \mathbf{X}'_{\sigma_n}$ involved in the learning procedure. In the first part of this thesis, non-asymptotic bounds of order $O_{\mathbb{P}}(1/\sqrt{n})$ are proved for the excess risk of a *plug-in* predictive rule mimicking the true minimizer in the case of isotropic stationary Gaussian processes, observed at locations forming a regular grid in the learning stage. These theoretical results, as well as the role played by the technical conditions required to establish them, are illustrated by various numerical experiments, on simulated data and on real-world datasets, and may hopefully pave the way for further developments in statistical learning based on spatial data.

In the second part of this thesis, we focus on space-time Hawkes processes. Many modern spatio-temporal data sets, in sociology, epidemiology or seismology, for example, exhibit self-exciting characteristics, with simultaneous triggering and clustering behaviors, that a suitable spatio-temporal Hawkes process can accurately capture. However, dealing efficiently with the high volumes of data now available is challenging. We aim at developing a fast and flexible parametric inference technique to recover the parameters of the kernel functions involved in the intensity function of a spatio-temporal Hawkes process based on such data. Our statistical approach combines three key ingredients: (1) kernels with finite support are considered, (2) the space-time domain is appropriately discretized, and (3) (approximate) precomputations are used. The inference technique we propose consists of a ℓ_2 gradient-based

solver that is fast and statistically accurate. In addition to describing the algorithmic aspects, numerical experiments have been carried out on synthetic and real spatio-temporal data, providing solid empirical evidence of the relevance of the proposed methodology.

Résumé

À l'époque des grandes données, et en particulier avec la prolifération des capteurs de géolocalisation, l'accès à des ensembles de données massives, présentant une structure de dépendance spatiale possiblement complexe, augmente de plus en plus. Dans cette thèse, notre objectif est de surmonter les enjeux liés à la structure de dépendance des données spatiales (et spatio-temporelles).

En un premier temps, nous analysons le *Krigeage simple*, problème clé en Géostatistique, en adoptant le point de vue de l'apprentissage statistique, *i.e.* en effectuant une analyse prédictive non paramétrique à partir d'un échantillon fini. Dans ce contexte, la théorie probabiliste standard de l'apprentissage statistique ne s'applique pas directement. De nouvelles garanties sur la capacité de généralisation du prédicteur par Krigeage doivent être établies. Étant donné une réalisation d'un champ aléatoire de carré intégrable $\mathbf{X} = \{\mathbf{X}_s\}_{s\in\mathcal{S}}$, $\mathcal{S} \subset \mathbb{R}^2$ de covariance inconnue, observé en $d \geq 1$ sites $s_1, \cdots, s_d$ du domaine spatial $\mathcal{S}$, l'objectif est de prédire les valeurs inconnues de $\mathbf{X}_s$ à n'importe quel point $s \in \mathcal{S}$, tout en minimisant le risque quadratique. La règle de prédiction est dérivée d'un ensemble de données spatiales d'apprentissage : une unique réalisation $\mathbf{X}'$ de $\mathbf{X}$, indépendante de celles à prédire, observée en $n \geq 1$ points $\sigma_1, \cdots, \sigma_n$ dans $\mathcal{S}$. Malgré le lien avec la régression ridge à noyau, déterminer la capacité de généralisation des minimiseurs de risque empiriques reste un défi complexe, en raison du caractère non indépendant et non identiquement distribué des données d'apprentissage $\mathbf{X}'_{\sigma_1}, \cdots, \mathbf{X}'_{\sigma_n}$ impliquées dans la procédure. Dans la première partie de cette thèse, nous présentons des bornes non asymptotiques d'ordre $O_{\mathbb{P}}(1/\sqrt{n})$ pour l'excès de risque d'une règle prédictive *plug-in* imitant le vrai minimiseur. Ces bornes sont établies pour des processus gaussiens stationnaires avec une fonction de covariance isotrope, observés lors de la phase d'apprentissage à des emplacements formant une grille régulière. Nos résultats théoriques, ainsi que le rôle joué par les conditions techniques requises pour les définir, sont illustrés par diverses expériences numériques, sur des données simulées ainsi que sur des données réelles, et ouvrent, nous l'espérons, la voie à de nouveaux développements dans l'apprentissage statistique basé sur des données spatiales.

En un second temps, nous nous concentrons sur les processus de Hawkes spatio-temporels. De nombreux ensembles de données spatio-temporelles, en sociologie, épidémiologie ou sismologie, par exemple, présentent des caractéristiques d'auto-excitation: les événements ont tendance à se regrouper ou à déclencher une série d'événements successifs, ou encore les deux à la fois. Dans ce contexte, les processus de Hawkes spatio-temporels se révèlent être un outil puissant grâce à leur capacité à capturer ces comportements avec précision. Cependant, traiter efficacement le grand

volume de données actuellement disponible s'avère difficile. La deuxième partie de cette thèse vise à développer une technique d'inférence paramétrique rapide et flexible pour obtenir les paramètres des fonctions noyaux impliquées dans la fonction d'intensité d'un processus de Hawkes spatio-temporel. Notre approche statistique combine trois ingrédients clés : (1) nous considérons des fonctions noyaux à support, (2) le domaine spatio-temporel est discrétisé de manière appropriée, et (3) des calculs préalables (approximatifs) sont utilisés. La technique d'inférence que nous proposons consiste en un solveur basé sur le gradient ℓ_2 qui est rapide et statistiquement précis. En complément de la description des aspects algorithmiques, des expériences numériques ont été menées sur des données spatio-temporelles, tant synthétiques que réelles, apportant des preuves empiriques solides de la pertinence de la méthodologie proposée.

Notation

$:=$	Equal by definition
$\mathbb{N}^*$	Set of strictly positive integers
$\mathbb{R}$	Set of real numbers
$\mathbb{R}^d$	Set of d-dimensional real-valued vectors
$\|\cdot\|$	Euclidean norm in $\mathbb{R}^d$, for any $d \in \mathbb{N}^*$
$\langle \cdot, \cdot \rangle$	Corresponding inner product
$\|\cdot\|_\infty$	Maximum norm
$\|\|\|\cdot\|\|\|$	Operator norm of any $d \times d$ matrix M, such that $\|\|\|M\|\|\| = \sup\{\|Mv\| : v \in \mathbb{R}^d, \|v\| = 1\}$
$\mathbb{I}\{\mathcal{E}\}$	Indicator function of any event $\mathcal{E}$
$\|E\|$	Cardinality of any finite set E
δ_x	Dirac mass at any point x
$\mathbf{I}_d$	Identity matrix of size $d \times d$
$M^\top$	Transpose of any matrix M
$Rank(M)$	Rank of any matrix M
$Tr(M)$	Trace of any matrix M
$(\Omega, \mathcal{F}, \mathbb{P})$	Probability space
$\mathbb{P}(\cdot)$	Probability of an event
$\mathbb{E}[Z]$	Expectation of any square-integrable real-valued random variable Z
$Cov(Z_1, Z_2)$	Covariance of any pair (Z_1, Z_2) of square-integrable real-valued random variables defined on a same probability space
$Var(Z)$	Covariance matrix of any square-integrable random vector Z
$\mathcal{N}(\mu, \Sigma)$	Gaussian distribution with mean μ and covariance matrix Σ

$\mathcal{S}$	Spatial set $\mathcal{S} \subset \mathbb{R}^d$
s	Location in $\mathcal{S}$
$\mathbf{s}_d = (s_1, \ldots, s_d)$	Set of d observed locations in $\mathcal{S}$
$\mathbf{X}$	Random field on $\mathcal{S}$ with $\mathbb{R}$
$\mathbf{X}(\mathbf{s}_d) = (\mathbf{X}_{s_1}, \ldots, \mathbf{X}_{s_d})$	Set of d observations of $\mathbf{X}$ on $\mathcal{S}$
$\Sigma(\mathbf{s}_d) = Var(\mathbf{X}(\mathbf{s}_d))$	Covariance matrix of the observations
$\mathbf{c}_d(s) = (Cov(\mathbf{X}_s, \mathbf{X}_{s_i}))_{1 \leq i \leq d}$	Covariance vector of the observations
$\gamma(\cdot)$	Semi-variogram function of any (second-order or intrinsic) stationary random field

Abbreviations

ERM	Empirical Risk Minimization
MSE	Mean Squared Error
IMSE	Integrated Mean Squared Error
AMSE	Averaged Mean Squared Error
MoM	Median-of-Means
LRR	Linear Ridge Regression
KRR	Kernel Ridge Regression
RKHS	Reproducing Kernel Hilbert Space
EM	Expectation Maximization
PP	Point Process
TPP	Temporal Point Process
HP	Hawkes Process
STHP	Spatio-Temporal Hawkes Process
MSTHP	Multivariate Spatio-Temporal Hawkes Process
ETAS	Epidemic-Type Aftershock Sequence
NLL	Negative Log-Likelihood
i.i.d.	independent and identically distributed
r.v.	random variable
w.r.t.	with respect to

1 Introduction

Contents

1.1 Spatial Statistics

In Machine Learning, the theory generally relies on independent and identically distributed (i.i.d.) characteristics of data. In other words, observations of a phenomenon are assumed to be collected under uniform conditions, with each observation being independent of the others. This allows employing standard statistical methods to construct an accurate and robust model and predicting new phenomena from it. Machine Learning techniques are supported by a very sound probabilistic theory (Devroye et al., 1996; Boucheron et al., 2013) guaranteeing the generalization capacity of empirically learned predictive rules under mild assumptions.

The independent assumption is very convenient. It makes Machine Learning flexible, easily implementable, and hence a successful tool with efficient algorithms. In recent years, a variety of statistical learning techniques – including boosting methods, support vector machines, neural networks among others – have been successfully developed for performing various tasks such as classification, regression or clustering. It can be applied to a wide variety of applications, such as image recognition, healthcare diagnostics, and natural language processing.

Example 1.1. *(Image recognition) Image recognition, a branch of computer vision, focuses on developing algorithms and models that interpret and categorize visual data from images. It involves extracting meaningful features, patterns, and relationships within images to recognize objects, scenes, or actions. Image recognition may have various scopes, such as:*

- *Image classification: to classify images into categories, for example into images of dogs or cats. Each image can be considered as an independent observation, assuming images are from different instances and the features extracted from these images are identically distributed across the dataset.*

- *Object detection: the focus here is on detecting and localizing objects within an image. If the desired goal is to localize multiple instances of the same object (for example cars), each task for a specific object can be treated as i.i.d., assuming the features of different instances of the same object as identically distributed.*

Still, in image recognition tasks such as object detection, understanding the context and relationships between objects or regions within an image is crucial. Objects may have complex interactions that cannot be modeled effectively under the i.i.d. assumption.

The validity of statistical learning remains mainly confined to the case of i.i.d. training data. At the same time, spectacular progress has been made in the collection, management and warehousing of massive datasets for scientific, engineering, medical or commercial purposes, relying on modern technologies, such as satellite imagery or geophysical tomography. These data tend to exhibit complex dependence structures, resulting in the violation of the i.i.d. assumption.

Moreover, we are facing more and more situations where data are of spatial nature and exhibit a strong dependence structure. In the context of spatial data, dependencies exist in all directions. Specifically, data points that are spatially close to each other are likely to exhibit correlation and the dependence becomes weaker as data locations become more distant.

Example 1.2. *(Meteorology) In meteorology, one's goal is typically to understand and predict weather patterns, climate trends, and atmospheric phenomena. Meteorology plays a crucial role in various sectors, such as agriculture, transportation, energy, and disaster management. One of the defining characteristics of meteorological data is of course its spatial nature. Spatial variability in meteorological data is influenced by factors such as geography, topography, or proximity to water bodies. Therefore, meteorological data present a strong dependence structure, which is essential for both observation and modeling tasks. The spatial dependencies are quite obvious in certain cases: nearby locations tend to have similar meteorological conditions (e.g. rainfall, Goovaerts, 2000), and show gradual changes over geographical regions.*

The first models adapted to dependent data appeared in temporal studies (or time series) (Box and Jenkins, 1970; Steinwart and Christmann, 2008; Steinwart et al., 2009; Kuznetsov and Mohri, 2014; Hanneke, 2017; Clémençon et al., 2019). These models assumed that the observations, identically distributed and occurring at regularly spaced time intervals, exhibit dependence and that this dependence is based on the natural unidirectional flow of time. This implies that the modeling in temporal studies is causal.

Whereas the case of time series, which can rely on concentration results for ergodic processes, is receiving increasing attention (Steinwart and Christmann, 2009; Kuznetsov and Mohri, 2014), that of spatial data is in contrast less intensively studied in the statistical learning literature. As in the case of temporal data, spatial statistics differ from classical statistics by the dependence characteristics of the observations. However, spatial models differ from temporal models in two main respects: they must possess greater flexibility, as there are no equivalences to the concepts of past, present, and future in spatial contexts; and, they must take into account the spatial position of the collected data, as it is crucial information.

Let us now introduce the general spatial model. Let $s \in \mathbb{R}^p$ be a data location, where $p \in \mathbb{N}^*$. Suppose that $\mathbf{X}_s$ is a random quantity. Let s varies over an index set $\mathcal{S} \subset \mathbb{R}^p$, so as to generate the random process:

$$\mathbf{X} = \left\{\mathbf{X}_s, s \in \mathcal{S}\right\}. \tag{1.1}$$

Note that we distinguish X the variable under study, also called the regionalized variable by Matheron (1965), and $\mathbf{X}$, the modeling of X by a random field. With this notation, X is a realization of $\mathbf{X}$.

Assumptions on $\mathcal{S}$ can vary from it being a fixed (non-random) subset of $\mathbb{R}^p$ to it being a random set (which implies that $\mathcal{S}$ may vary from one realization to another). Spatial statistics studies can be divided into three categories based on the nature of the index set $\mathcal{S}$ (Cressie, 1993, Chapter 1):

1. **Geostatistical data:** $\mathcal{S}$ is a fixed subset of $\mathbb{R}^p$ ($\mathcal{S}$ is continuous) and $\{\mathbf{X}_s, s \in \mathcal{S}\}$ is a random vector at location $s \in \mathcal{S}$. We assume that the random field is observed at n fixed points $\{s_1, s_2, \cdots, s_n\}$. The observations can be either randomly sampled over $\mathcal{S}$ or selected over a regular grid. Geostatistics deals with tasks like modeling, prediction (called Kriging) on an unobserved site s, and constructing a complete map of the random field over the entire domain $\mathcal{S}$.

   Example 1.3. *(Mining) Geostatistics emerged as an interdisciplinary study that involves both mining engineering and statistics. Previous methods adopted in mining usually employed histograms of ore grade observations, focusing only on the rate of these samples and thus neglecting the spatial position of the observations. Still, the spatial location, as well as the possible patterns (like clustering) over the ore body, are valuable information in mining operations. Matheron (1963) proposed geostatistics as a new approach to estimate the ore grades and ore reserves for mining operations. Based on a set of observations over the mineral deposit, he developed a prediction method that takes into account the spatial position of the samples, as well as the dependence structure of the ore grades. See Example 2.1 in Chapter 2 for further details on geostatistics methodology and scopes for mining issues.*

2. **Lattice data:** $\mathcal{S}$ is a fixed collection of countably many points of $\mathbb{R}^p$ ($\mathcal{S}$ is discrete) and $\mathbf{X}_s$ is a random vector at location $s \in \mathcal{S}$. The data is linked to spatial units or regions, thus forming a network. In lattice data, one may be interested in the study of spatial correlation, prediction, or for example in image analysis and image restoration.

   Example 1.4. *(Image analysis) In the context of image analysis and image restoration, lattice data methodology applies since images can be viewed as a grid of pixels, each representing a spatial unit. Lattice data methods can help understand the spatial structure and dependencies within the image, and measure the spatial correlation (which is informative as to how pixel value in an image is correlated with its neighbors). See for example Cressie, 1993, Section 7.4 and Besag (1974); Ripley (2005) for more details.*

3. **Point patterns:** $\mathcal{S}$ is a point process in $\mathbb{R}^p$ ($\mathcal{S}$ is random) and $\mathbf{X}_s$ is a random vector at location $s \in \mathcal{S}$. Here, the data locations of observations $\{s_1, s_2, \cdots, s_n\}$ and the number of observations n are random. In this case, the data locations carry

the desired information, and the observation points are considered as events of a point process. The statistician will be interested in capturing a pattern in the data, such as clustering, triggering, regularity, or complete randomness.

Example 1.5. *(Seismology) In seismology, clustering patterns typically appear: some regions are particularly affected by earthquake occurrences, whereas other regions may never suffer from it. Ogata (1988) introduced point processes to study the occurrences of an earthquake in a given region. The events have also a triggering pattern: an earthquake may trigger following earthquakes, called aftershocks. Therefore, for safety reasons, it is essential to identify clustering patterns to determine the most affected regions and to understand triggering relationships to prevent further damage. It is also possible to use a marked process, where the mark may represent the magnitude of the earthquake. See subsections 5.1 and 5.2 for further details on seismological studies in spatial statistics.*

In this thesis, we choose to focus on two categories of spatial data: geostatistical data and point patterns data.

1.2 Motivations

In this section, we present the research questions that motivated this thesis and the challenges stemming from it.

The main goal of this thesis is to develop methods to take into account the strong dependence structure of spatial data, based on an observation of the phenomenon at a finite number of spatial locations, in order to model, predict and learn from spatial data. The thesis is divided into two main parts: the first one concerns a Geostatistics method and aims at providing theoretical guarantees for this prediction method, while the second part focuses on designing a new approach for a specific category of spatio-temporal processes. Even if Geostatistics and point processes are two different categories of spatial statistics, differing by their assumption on the spatial domain, they partially share motivations and difficulties.

Domains of Application. Most elements around us possess a spatial dimension. This includes natural phenomena, like weather and natural disasters, as well as human-made infrastructures, such as water wells and urban planning. All of these are heavily influenced by spatial factors. For instance, on a small scale, geographically close cities tend to experience similar weather conditions. Similarly, the proper organization of streets and traffic, which is crucial for the effective functioning of a city, requires that interdependencies in traffic among different city areas be properly taken into account.

Consider the example in Figure 1.1 which shows the average daily temperature in France (expressed in Kelvin). On the right, we see a color map of the temperatures observed at each point within the sampled square on the left. There is a yellow area indicating slightly lower temperature values and some red areas indicating higher temperatures (the difference between these temperatures is relatively small). What stands out is the smooth gradient of temperatures (*i.e.* gradual changes) across the entire spatial domain and the presence of zones with similar values. Properly modeling these temperature interdependencies is crucial for improving weather forecasts, which are essential as they may help to anticipate extreme situations such as storms, cyclones, or periods of severe drought. Let's now take the example in Figure 1.5. This

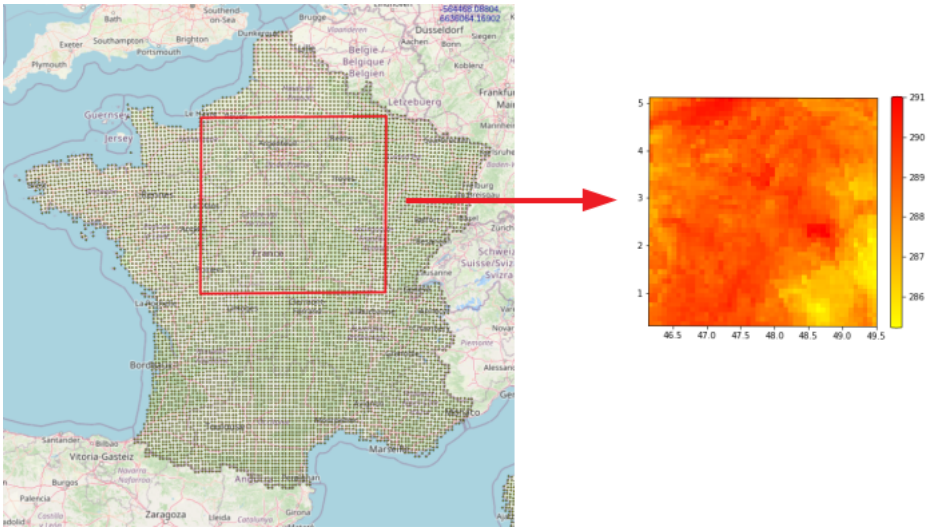


Figure 1.1: Map of France with a sampled square grid (left); color map of the temperatures (in Kelvin) of the sampled square grid, 2nd June 2005 (right).
Source: DRIAS (https://drias-prod.meteo.fr/okapi/accueil/okapiWebDrias/).

map depicts all earthquakes recorded in the volcanic region of the Phlegraean Fields (west of Naples, Italy) from January to July 2024. It is evident that there is a significant concentration of seismic activity in this area, mainly clustered along the coastline near the dormant volcano of Agnano. The occurrence of potential tremors is associated with a cyclic volcanic phenomenon that causes periodic lifting and subsidence of the ground. It is precisely during the lifting phases that an increase in seismic activity can be observed. While the majority of these earthquakes are of low magnitude (in comparison to the earthquake that occurred in Turkey and Syria in February 2023, see Figure 1.4), the high concentration in this region might indicate the possibility of an upcoming eruption or stronger earthquakes. Therefore, it is crucial to accurately predict future seismic events by considering both historical data and spatial distribution.

We have mentioned only two examples of spatial data (weather forecasting and earthquakes prediction). There are undoubtedly numerous other application areas, several of which are briefly covered in this thesis, including mining, hydrology, ecology, epidemiology, finance, and criminology studies.

Violation of the i.i.d. Assumption. Classical statistical techniques generally assume that observations of a phenomenon are independent and identically distributed. However, due to the presence of a dependency structure in the spatial data framework, the assumption of i.i.d. observations of a phenomenon is not satisfied. Thus, new methods and theoretical results must be established in this context. The main difficulty in learning from spatial data is to obtain information about the underlying dependency structure, so that it can be taken into account when modeling and predicting from these data.

This leads to our first research questions and corresponding challenge, which are the overarching goals of this thesis.

Research Questions 1: How to learn from spatial data that presents a **strong dependence structure**? How does the dependence structure of the observed phenomenon affect the **performance** of the algorithms?

Challenge 1: Provide statistical guarantees for methods used to predict spatial data. Develop new, efficient, and accurate methods to predict from spatial data.

As previously mentioned, the first part of this thesis concerns geostatistical data, and aims at contributing to address the first challenge for spatial data by providing statistical guarantees for spatial prediction methods.

Geostatistical Data with a Singular Phenomenon Realization. Learning from geostatistical data implies two main challenges. The first one, already mentioned, is the presence of a strong dependence structure within the data. The second one is the fact that typically only a single realization of the phenomenon is available. For example, a specific natural event, such as a storm, happens only once and no other independent realization of it can be observed. Other instances are the high economic cost of the data collection and the possible deterioration of the environment. This is the case in the hydrogeological dataset presented in Figure 1.2. Hydrogeology aims at assessing groundwater quality (see Example 2.2 in Chapter 2 for further details) based on characteristics of the water, such as pH level, water conductivity and temperature. To do so, observations are collected over a spatial region, here in the department of La Guajira in Colombia, and special measurements are made. However, this procedure involves a significant economic cost.

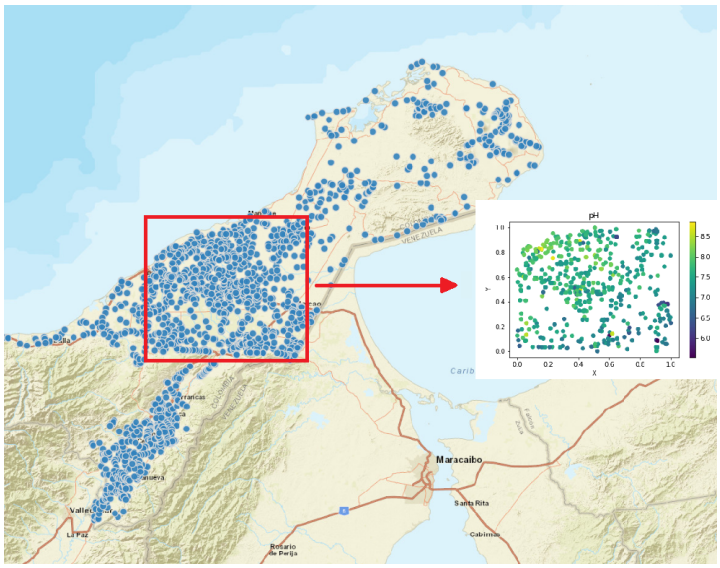


Figure 1.2: Hydrogeological map of the department of La Guajira (Colombia) in 2016. Each point on this map represents a water body. On the right, the color map of a sampled square grid representing the pH values at each spatial point.
Source: Servicio Geológico Colombiano (https://datos.sgc.gov.co/). Map created in 2016 by the Groundwater Group of the Colombian Geological Service.

In Geostatistics, a phenomenon is modeled by a random field, assumed to be observed at a finite number of locations on the spatial domain $\mathcal{S} \subset \mathbb{R}^p$. The dependence characteristics of the data are modeled by the covariance function of the random field. Our setting is the following: we are interested in predicting the values at all spatial locations $s \in \mathcal{S}$ of a random field $\mathbf{X}$, observed at a fixed set of locations $\{s_1, \dots, s_d\}$. The random field is assumed to be a second-order stationary Gaussian random process with isotropic covariance function. These assumptions are often formulated in Geostatistics as they ensure a successful frequentist approach. The resulting interpolation method is called *Kriging* (Matheron, 1962) and aims at building a predictor $\widehat{\mathbf{X}}_s$ of $\mathbf{X}_s$, defined as a linear combination of the observations. The Kriging weights depend on the covariance function of the random field.

Non-Parametric Covariance Estimation. When dealing with real data, the covariance function is unknown. Based on a training dataset, defined as a single realization $\mathbf{X}'$ of $\mathbf{X}$ observed at n locations $\{\sigma_1, \dots, \sigma_n\}$, the covariance function can be estimated. Previous results for the estimation of the covariance function have been investigated, either in an asymptotic perspective (Stein, 1999), or by means of a parametric approach (Zimmerman, 1989; Zimmerman and Cressie, 1992). In contrast, we are interested by the finite sample behavior of the non-parametric covariance estimator, under an *infill* asymptotic setting (*i.e.* assuming that new observations occur within the same fixed spatial domain, which becomes denser and denser).

This brings us to our second research question and challenge, aiming at defining the accuracy of the non-parametric covariance estimation.

Research Question 2: How **accurate** is the empirical covariance estimator, based on a **finite number of observations** on a regular grid and with **one unique realization**?

Challenge 2: Derive non-asymptotic bounds for the **non-parametric covariance estimator**, under the in-fill asymptotic.

Challenge 2 can be seen as an intermediary challenge to answer our Research Question 3 (see below), namely deriving non-asymptotic guarantees for the Kriging method. Indeed, as explained below, the prediction Kriging method depends on the accuracy of the covariance estimator. Thus, computing an accurate estimation of the covariance function and identifying the possible uncertainty effects of this estimation on the Kriging predictor is of prime importance.

Non-asymptotic Guarantees for the Empirical Kriging Method. When the dependence structure of the random field is known, the Kriging method is optimal (we refer to it as *theoretical Kriging*). Yet, in the case of unknown covariance function, the generalization capacity of the resulting method (we refer to it as *empirical Kriging*) is left to be established. The objective is to develop a novel theoretical framework offering non-asymptotic guarantees for empirical simple Kriging predictions. The generalization guarantees of the empirical predictor are given by means of a bound on the global excess risk. This risk is defined as the global gap between the prediction errors of the theoretical and the empirical Kriging predictors.

Research Questions 3: What is the **non-asymptotic** behavior of the Kriging predictor when the dependence structure is **unknown** and with a finite number of observations? To what extent the Kriging weights depend on the accuracy of the covariance function estimation and on the location of samples?

Challenge 3: Derive **non-asymptotic** bounds for the **global excess risk** of the Kriging method. These theoretical results must depend on the covariance estimation and on the choice of the sampling setting.

The second part of this thesis concerns point patterns data, and more precisely spatio-temporal Hawkes processes. In this context, the observations are considered as events of a process. Hawkes processes find applications in various domains, such as the study of natural disasters, as explained below.

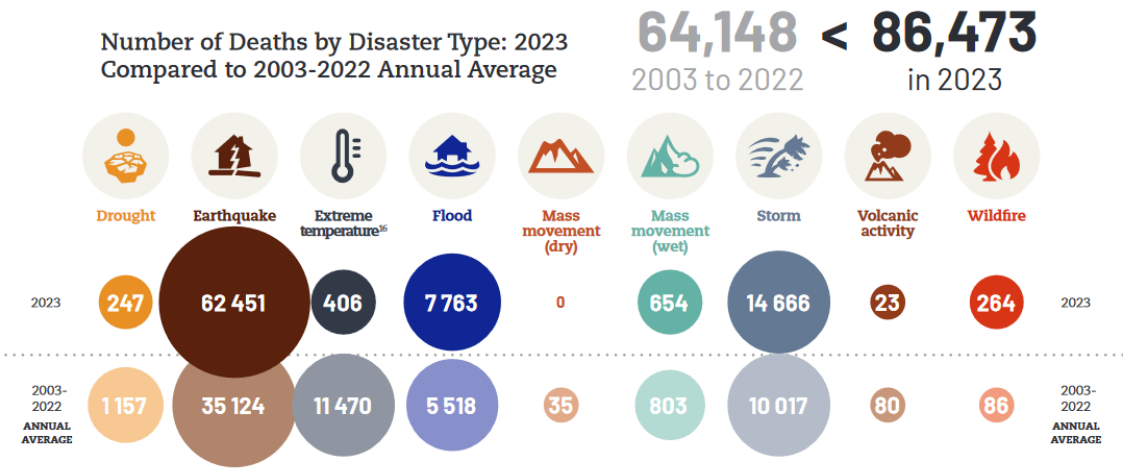


Figure 1.3: Number of deaths by disaster type: a comparison between the number in 2023 and the annual average number from 2003 to 2022.
Source: EM-DAT (https://www.emdat.be/), CRED annual report, *2023 Disasters in numbers*.

Earthquake Prediction and Risk Assessment. In 2023, the total number of deaths due to earthquakes rose to 62 451, according to the CRED (Center for Research on the Epidemiology of Disasters) annual report, almost double the average over the past 20 years (see Figure 1.3). This number was particularly high that year because of the earthquake in Turkey and Syria in February 2023 which was, as written by CRED, '*the most catastrophic event of the year regarding mortality and economic damage, counting for two-thirds of the total deaths*'. On 6 February 2023, a sequence of earthquakes occurred in south-eastern Turkey, at the border with Syria. A first major earthquake of magnitude 7.8 struck near the city of Gaziantep, followed by aftershocks of lower magnitude, spreading in all the surrounding areas (see Figure 1.4). This region frequently experiences seismic events. Therefore, there is an urgent need of robust and accurate models for predicting seismic activity and improving risk assessment in regions highly impacted by earthquakes. Another region of this kind is the Phlegraean Fields, whose seismic activity in 2024 is depicted in Figure 1.5 (where the dimension of a point gives the magnitude of the event, and the color indicates the period of time at which it occurred).

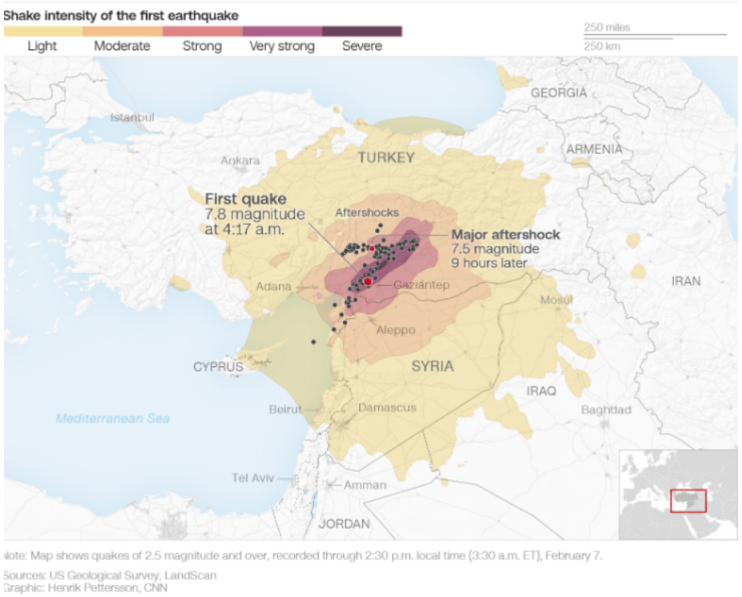


Figure 1.4: The earthquake in Turkey and Syria, 6 February 2023. The epicenters are in red, and the scale of colours shows the shake intensity of the first earthquake in the area.
Source: USGS (https://www.usgs.gov/), United States Geological Survey, LandScan.

Observing Figures 1.4 and 1.5, one can clearly see a clustering behavior of the events. Indeed, in Figure 1.5, the events are clustered both in the temporal and spatial dimensions: points of the same color are gathered together in the spatial region. This observation reveals the triggering characteristics of earthquakes. Indeed, a first major earthquake of high magnitude (called a mainshock) can trigger a new occurrence, generally of lower magnitude (called an aftershock). This triggering effect mainly happens with a clustering pattern, *i.e.* the appearance of new occurrences takes place in a specific time window and within the neighborhood of the first epicenter. The triggering and clustering behaviors of a seismic phenomenon are thus crucial to understand the underlying seismic activity and improving the prediction of future events.

Space-Time Hawkes Processes. Among point processes, Hawkes models (Hawkes, 1971) have recently received a great deal of attention, as they take into account the self-exciting nature of observed events, space-time interaction, and spatial anisotropy in a very flexible way. For well-chosen intensity functions, the probability of future events occurring over a given period increases with these point processes (Reinhart, 2018). Vere-Jones (1970) and Ogata (1988) introduced these processes to seismology because of the triggering behavior of earthquakes. Indeed, *Epidemic-Type Aftershock Sequence* (ETAS) models are well-suited for modeling seismic activities since they imply that each earthquake can initiate aftershocks, which in turn can cause additional aftershocks, resulting in a chain reaction of seismic activity. Thus, Hawkes processes turn out to be powerful tools with data presenting a self-exciting nature. However, the first Hawkes models were purely temporal (Ogata, 1988), thus neglecting the spatial dimension of the phenomena. Indeed, as observed in Figure 1.5, the complex dynamics of earthquakes show both spatial and temporal clustering. In this thesis, we investigate Space-Time Hawkes Processes (STHPs) to take into account the spatio-temporal dependencies between the events.

The self-exciting properties of space-time Hawkes processes are the reason why they are being increasingly exploited in many fields requiring spatio-temporal analysis, such as epidemiology (Holbrook et al., 2022; Kresin et al., 2022; Rambhatla et al., 2022; Dong et al., 2023), criminology (Mohler et al., 2011; Mohler, 2014; D'Angelo et al., 2022; Zhu and Xie, 2022), and seismology (Ogata, 1998; Musmeci and Vere-Jones, 1992; Kwon et al., 2023) for instance. The main methodological challenge is then to design efficient inference techniques to fit Hawkes models to spatio-temporal datasets.

Research Question 4: How to **learn from a multivariate spatio-temporal Hawkes process**, despite the modeling and numerical **challenges** posed by parametric STHP's complexity?

Challenge 4: Develop a **new efficient and flexible method** for parametric inference for spatio-temporal Hawkes processes, consisting of a **fast** ℓ_2 gradient-based solver.

For computational and simplicity reasons, most of previous methods are restricted to space-time separable kernels, where the temporal kernel is often chosen as exponential and the spatial influence is modeled by a Gaussian kernel (Mohler, 2014; Yuan et al., 2019; Ilhan and Kozat, 2020).

General Parametric Kernels. The generally assumed exponential temporal kernel, even if it brings computational efficiency, implies major limitations in real-world situations, as it assumes that an event immediately trigger a future event. However, in the case of earthquakes, this assumption is generally not valid. For instance, in the case of the seismic activity that affected Turkey and Syria in 2023, a first mainshock (of magnitude 7.8) occurred around 4 AM. A second mainshock (of magnitude 7.5) arrived 9 hours later, around 1 PM. The epicenters of these two earthquakes are spatially close, as shown in Figure 1.4. In this case, the exponential kernel is not suitable, since a latency is observed between the two major earthquakes. On the contrary, on 20 May 2024, several earthquakes hit the region of the Phlegraean Fields, between 7:51 PM and 9:55 PM (of magnitudes between 3.1 and 4.4), implying a more immediate influence. Thus, depending on several factors (such as the underlying tectonic plates of the region, the presence of a volcano, etc), the triggering and clustering temporal behaviors of earthquakes may vary from a region to another. Furthermore, for the spatial dimension, the spread of the aftershocks in Figure 1.5 does not seem to follow a Gaussian distribution.

Space-Time Interactions. Space-time separability for the kernel of a Hawkes process is a common assumption (see *e.g.* Mohler, 2014; Yuan et al., 2019; Ilhan and Kozat, 2020). Indeed, it brings simplicity, since it implies that the kernel is a product of spatial and temporal influences that can be modeled separately. However, when dealing with natural phenomena such as earthquakes, a space-time interaction can generally be observed.

These two limits of previous approaches motivate the need of a new efficient and flexible method for modeling spatio-temporal Hawkes processes. This new method must be suitable for both general parametric kernels and space-time non-separable

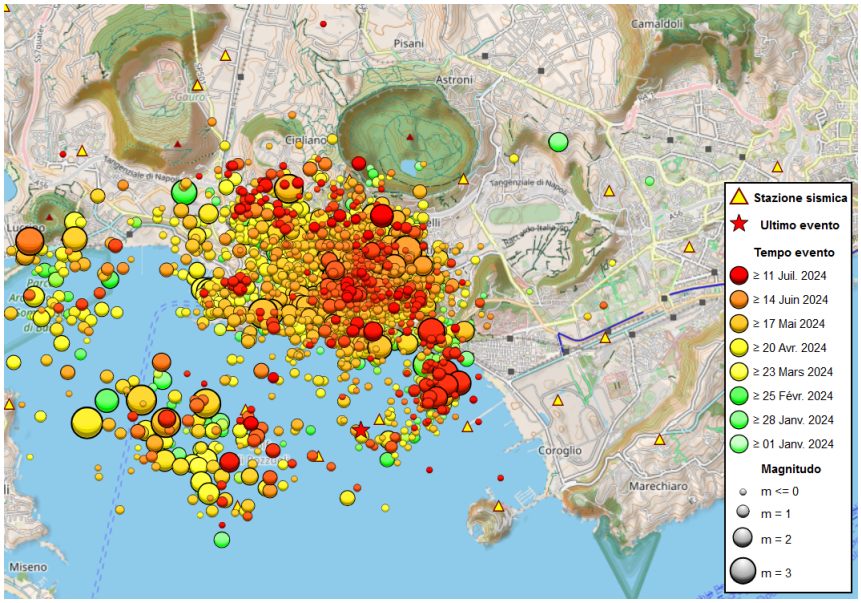


Figure 1.5: Earthquakes occurrences in 2024, in Phlegraean Fields, west of Naples (Italy). The scale of colors represents the time of each event, and the size of the point, its magnitude.
Source: INGV (https://terremoti.ov.ingv.it/gossip/flegrei/2024/), National Institute of Geophysics and Volcanology.

kernels, allowing a better prediction based on the characteristics of the spatial domain or of the phenomenon under study.

Research Question 5: How to **accurately model real-world situations**, where **space-time interactions** occur and where a latency between aftershocks may be observed, by means of Hawkes processes?

Challenge 5: Adapt the parametric method in such a way that it allows for **any kind of kernels** and enables the estimation of parameters of a **space-time non-separable** Hawkes process, thus providing **flexibility and accuracy** in modeling **complex dependencies in real-world datasets**.

Flexible and Efficient Parametric Method. We develop a fast parametric method that allows for any kind of kernels and for space-time non-separable kernels. The method derived is inspired by the work of Staerman et al. (2023) for temporal Hawkes processes, extending the method to capture space-time interactions.

1.3 Contributions

To overcome the **challenges** just described and answer our **research questions**, our contributions are the following (see Table 1.1 for a summary of these contributions).

Part I. The first part of this thesis aims at contributing to the design and the study of statistical learning methods applied to spatial data, by investigating the Kriging problem. The objective of Kriging is to predict the values of a random field $\mathbf{X} = \{\mathbf{X}_s\}_{s \in \mathcal{S}}$,

$\mathcal{S} \subset \mathbb{R}^2$, at all unobserved locations in $\mathcal{S}$, based on a finite number $d \geq 1$ of observations $\mathbf{X}(\mathbf{s}_d) := (\mathbf{X}_{s_1}, \ldots, \mathbf{X}_{s_d})$, with $\mathbf{s}_d = (s_1, \ldots, s_d)$. For this set of observations, let $\Sigma(\mathbf{s}_d) = Var(\mathbf{X}(\mathbf{s}_d))$ be the covariance matrix and $\mathbf{c}_d(s) = (Cov(\mathbf{X}_s, \mathbf{X}_{s_1}), \ldots, Cov(\mathbf{X}_s, \mathbf{X}_{s_d}))$ the covariance vector. The goal is to build a predictive map $f(s) = f_{\Lambda_d}(s, \mathbf{X}(\mathbf{s}_d)) = \lambda_1(s)\mathbf{X}_{s_1} + \ldots + \lambda_d(s)\mathbf{X}_{s_d}$ that is linear in $\mathbf{X}(\mathbf{s}_d)$ and minimizes the integrated mean squared error

$$L_{\mathcal{S}}(f_{\Lambda_d}) = \mathbb{E}_{\mathbf{X}}\left[\int_{s\in\mathcal{S}} \left(f_{\Lambda_d}(s, \mathbf{X}(\mathbf{s}_d)) - \mathbf{X}_s\right)^2 ds\right],$$

where $\Lambda_d : s \in \mathcal{S} \mapsto (\lambda_1(s), \ldots, \lambda_d(s))$ is a measurable function valued in $\mathbb{R}^d$. When the true covariance function $c(\cdot)$ of $\mathbf{X}$ is known and when the matrix $\Sigma(\mathbf{s}_d)$ is positive definite, the Kriging predictor $f_{\Lambda_d^*}(s, \mathbf{X}(\mathbf{s}_d)) = \mathbf{X}(\mathbf{s}_d)^\top \Sigma(\mathbf{s}_d)^{-1}\mathbf{c}_d(s)$ achieves optimal performance. Let's denote the minimum global error as $L_{\mathcal{S}}^* := L_{\mathcal{S}}(f_{\Lambda_d^*})$ and the optimal Kriging weights $\Lambda_d^*(s) = \Sigma(\mathbf{s}_d)^{-1}\mathbf{c}_d(s)$. However, this optimality does not always hold in practice because the true covariance structure of real data remains unknown. Thus, based on a training dataset $\mathbf{X}'$, defined as a single realization of $\mathbf{X}$, observed at $n \geq 1$ spatial locations $\sigma_1, \ldots, \sigma_n$ forming a regular dyadic grid, an empirical estimation $\widehat{c}(\cdot)$ of the covariance function can be obtained. From $\widehat{c}(\cdot)$, the empirical estimators $\widehat{\Sigma}(\mathbf{s}_d)$ and $\widehat{\mathbf{c}}_d(s)$ of $\Sigma(\mathbf{s}_d)$ and $\mathbf{c}_d(s)$ respectively can be computed. Now, replacing $\Sigma(\mathbf{s}_d)^{-1}$ and $\mathbf{c}_d(s)$ by their estimators, a natural empirical counterpart of Λ_d^* is built by means of the *plug-in* method and an empirical version of the Kriging predictor is

$$f_{\widehat{\Lambda}_d}(s, \mathbf{X}(\mathbf{s}_d)) = \mathbf{X}(\mathbf{s}_d)^\top \widehat{\Sigma}(\mathbf{s}_d)^{-1}\widehat{\mathbf{c}}_d(s).$$

Viewing Dual Kriging as a Kernel Ridge Regression Problem. We show that the optimal predictor f_{Λ_d} has the same form as a kernel ridge regressor, once the Gram matrix of the selected kernel for the regression is replaced with the true covariance matrix of $\mathbf{X}(\mathbf{s}_d)$ (*cf* Chapter 4).

The goal now is to provide theoretical guarantees for the empirical Kriging predictor by means of non-asymptotic bounds on the global excess risk $L_{\mathcal{S}}(f_{\widehat{\Lambda}_d}) - L_{\mathcal{S}}^*$. Since the empirical predictor $f_{\widehat{\Lambda}_d}$ depends on the the covariance function estimation $\widehat{c}(\cdot)$, our first goal is to assess the accuracy of this estimation.

Non-asymptotic Bounds for the Covariance Function Estimation. In Geostatistics, when the random field is (second-order) stationary, one uses the semi-variogram $\gamma(\cdot)$ to characterize the spatial dependence structure of the observations. The relationship between the isotropic covariance and semi-variogram functions is given by the following equation $\gamma(h) = c(0) - c(h)$. We extend this relationship to their estimators based on the observations $\mathbf{X}'_{\sigma_1}, \cdots, \mathbf{X}'_{\sigma_n}$: $\widehat{\gamma}(h) = \widehat{c}_h(0) - \widehat{c}(h)$. In Chapter 3, under the assumption that $\mathbf{X}$ is a second-order stationary Gaussian random field with isotropic covariance function, we first identify the distribution of the non-parametric estimators $\widehat{\gamma}(h)$ and $\widehat{c}_h(0)$, which is given by a weighted sum of χ^2 random variables. Under appropriate conditions, we derive Poisson tail bounds for these estimators, based on new concentration results for Gamma and χ^2 variables (Bercu et al., 2015; Wang and Ma, 2020). These bounds are derived only for the observed lags h of the sampled regular grid $\sigma_1, \cdots, \sigma_n$. Thanks to the relationship between the estimators, corresponding bounds can be derived also for the covariance function estimation. Finally, assuming that $c(\cdot)$ is of class $\mathcal{C}^1$ with gradient bounded by a constant $Q < +\infty$, we extend the previous bounds for all lags of the supposedly bounded spatial domain. These contributions allow us to answer to our **Research Question 2**.

Then, thanks to the above contributions, we analyze the impact of the covariance function estimation accuracy on the performance of the empirical Kriging predictor.

Statistical Guarantees for the Kriging Method. First, we provide non-asymptotic bounds for the accuracy of the covariance matrix and precision matrix estimations ($\widehat{\Sigma}(\mathbf{s}_d)$ and $\widehat{\Sigma}(\mathbf{s}_d)^{-1}$ respectively) in Chapter 4. These bounds stem from the previous results for the covariance function estimation $\widehat{c}(\cdot)$, under an additional assumption on the eigenvalues of $\Sigma(\mathbf{s}_d)$. Then, in Chapter 4, we assess the generalization capacity of the empirical Kriging predictor at all unobserved sites of the spatial domain, by deriving non-asymptotic tail bounds for the global excess risk of the Kriging method. The final result is provided by Theorem 4.8, where learning rate bounds of order $O_{\mathbb{P}}(1/\sqrt{n})$ are established for the empirical predictor, under appropriate conditions. Our main result is the following:

For any $\delta \in (0,1)$, we have with probability at least $1-\delta$:

$$L_{\mathcal{S}}(f_{\widehat{\Lambda}_d}) - L^*_{\mathcal{S}} \leq C_6\, d^2 \sqrt{\log(4n/\delta)/n} + C'_6\, d^2\, Q/(\sqrt{n}-1),$$

as soon as $n \geq C''_6 \log(4n/\delta)$, where C_6, C'_6 and C''_6 are positive constants.

This result allow us to answer to our **Research Question 3**.

Numerical Experiments. The theoretical results, as well as the role played by the technical conditions required to establish them, are illustrated in Chapter 4 by various numerical experiments on simulated data. We compute the experiments for different covariance models, some of which satisfy all the required conditions for our results while others do not. We repeat the experiments for different sizes of the observations grids. Our numerical experiments fully corroborate our theoretical results for all covariance models that satisfy the assumptions. Furthermore, we investigate other possible future extensions of our work by providing additional experiments that tackle the following cases: (1) the d observation points are taken from different configurations than the random uniform procedure, (2) the case of anisotropic covariance models, (3) the setting of irregular grids for the training sample. The results of our experiments show that the prediction method is robust in case of slight violations of the above assumptions.

Applications to Real Data. We illustrate the strength and advantages of the empirical Kriging predictor by numerical experiments on real meteorological data in Chapter 4. The DRIAS dataset gives the daily averaged temperature in France, observed over a regular grid (see Figure 1.1). A parametric Kriging predictor, constructed by means of a parametric covariance function, as well as the non-parametric Kriging predictor of interest, are applied to these data. Our results corroborate the established theoretical guarantees and demonstrate that a straightforward application of the empirical Kriging prediction method can yield strong performance and better flexibility compared to a parametric method.

Code. Our experiments are fully reproducible and can be replicated with the codes available on GitHub[1].

Part II. The second part of this thesis aims at designing a new inference method for multivariate spatio-temporal Hawkes processes. The main characteristic of a Hawkes process is that it takes into account the self-exciting nature of the underlying phe-

[1] https://github.com/EmiliaSiv/Simple-Kriging-Code

nomenon. Let $T \in \mathbb{R}_+$ be a stopping time, and consider $[0, T]$ as the resulting observation period. Additionally, let $\mathcal{S} = \mathcal{X} \times \mathcal{Y} \subset \mathbb{R}^2$ be a compact set within the spatial domain that contains the locations of the observed events up to time T. Let $D \in \mathbb{N}^*$ be the dimension of the multivariate spatio-temporal Hawkes process. Then, a realization consists of D sets of distinct events: $\mathcal{H}_T^i = \left\{ u_n^i = (x_n^i, y_n^i, t_n^i), (x_n^i, y_n^i) \in \mathcal{S}, t_n^i \in [0, T] \right\}, \forall i \in \{1, \dots, D\}$ occurring in continuous space-time, with an associated time t_n^i and a location (x_n^i, y_n^i). The behavior of the process is entirely described by its D intensity functions, which depend on the times and locations of past events. The conditional intensity function for the i-th process is:

$$\lambda_i(x, y, t | \mathcal{H}_t) = \boldsymbol{\mu}_i + \sum_{j=1}^{D} \sum_{u_n^j \in \mathcal{H}_t^j} \boldsymbol{\alpha}_{ij}\, g_{ij}(x - x_n^j, y - y_n^j, t - t_n^j),$$

where $\boldsymbol{\mu}_i > 0$ is the baseline parameter, $\boldsymbol{\alpha}_{ij} > 0$ is the excitation scaling parameter, and $g_{ij} : \mathcal{S} \times [0, T] \mapsto \mathbb{R}_+$ is the spatio-temporal kernel with parameters $\boldsymbol{\eta}_{ij}$. Note that we use the same notation as for the weights of the Kriging predictor in Part I to respect the usual notations of these domains.

The goal of our method is to be able to infer the parameters of (1) any parametric kernels, including (2) space-time non-separable kernels. Our work is inspired from the approach proposed in Staerman et al. (2023) for temporal Hawkes processes, which procedure relies on three key ideas that we extend to spatio-temporal data. The first concept is that the spatio-temporal domain of observations is discretized into a three-dimensional regular grid and the observations are projected on it. Next, we assume that the kernel functions are of finite length. Combining these first two ideas, the triggering function in the conditional intensity λ_i can be replaced by a discretized version, thus replacing the sum over past events by a sum over a finite number of grid elements. Then, we focus on the least squares loss and derive a discretized version

$$\mathcal{L}_{\mathcal{G}}(\boldsymbol{\theta}, \widetilde{\mathcal{H}}_T) = \sum_{i=1}^{D} \left(\Delta_{\mathcal{X}} \Delta_{\mathcal{Y}} \Delta_T \sum_{v_x=0}^{G_{\mathcal{X}}} \sum_{v_y=0}^{G_{\mathcal{Y}}} \sum_{v_t=0}^{G_T} \left(\tilde{\lambda}_i[v_x, v_y, v_t] \right)^2 - 2 \sum_{\tilde{u}_n^i \in \widetilde{\mathcal{H}}_T^i} \tilde{\lambda}_i \left[\frac{\tilde{x}_n^i}{\Delta_{\mathcal{X}}}, \frac{\tilde{y}_n^i}{\Delta_{\mathcal{Y}}}, \frac{\tilde{t}_n^i}{\Delta_T} \right] \right),$$

where $(\Delta_{\mathcal{X}}, \Delta_{\mathcal{Y}}, \Delta_T)$ are the step sizes of the discretization, $(G_{\mathcal{X}}, G_{\mathcal{Y}}, G_T)$ are the size of the discretized grids, and $\widetilde{\mathcal{H}}_T^i$ are the projected space-time stamps of $\mathcal{H}_T^i$. This takes us to the third key component of our approach, which is the identification of precomputation terms (that do not depend on the parameter $\boldsymbol{\theta} = \{\boldsymbol{\mu}_i, \boldsymbol{\alpha}_{ij}, \boldsymbol{\eta}_{ij}\}_{i,j}$) in the discretized loss function. Thanks to these precomputation terms, our approach is efficient and allows fast inference.

Combining all three key ideas, we design a method answering our **Research Questions 4 and 5**, defined in Section 1.2. We also provide theoretical guarantees on the bias induced by the discretization, which show its low impact on parameter estimation accuracy.

Efficient and Flexible Method for Parametric Inference in Space-Time Hawkes Models. In Chapter 6, we develop a fast method for inferring kernel parameters in spatio-temporal Hawkes models. The method we design enables the incorporation of any parametric kernels for the triggering function, extending beyond the traditional Gaus-

sian and exponential forms. Furthermore, to better fit real data, the approach accounts also for space-time interactions, as it extends to the case of space-time non-separable kernels. These two innovations improve the accuracy and flexibility in modeling complex dependencies in real-world datasets.

Numerical Experiments. We show the advantages of our approach with various experiments on simulated data in Chapter 6. First, we study the impact of the discretization step on the accuracy of the method by repeating the experiments for different values of $(\Delta_{\mathcal{X}}, \Delta_{\mathcal{Y}}, \Delta_T)$. Our results show that the estimation error goes towards zero as the steps simultaneously decrease, supporting our theoretical results on the discretization. Next, our experiments computed for various ending time T and spatial bounds $\mathcal{S}$ prove the accuracy of the method. The computational time with respect to the discretization step and with respect to $(\mathcal{S}, T)$ is also investigated, proving the efficiency of our method. Finally, all the experiments are computed with varying spatial and temporal kernels, thus showing the flexibility of our approach.

Applications to Real Data. The advantages of our inference method are also proved by applying it to two real-world datasets in Chapter 6: (1) real earthquake data, based on the *Northern California Earthquake Data Center* [2] dataset (NCEDC; nce, 2014) and (2) burglary data from the *Chicago Crime Dataset* [3]. Indeed, both real-world datasets violate the two conditions assumed by the majority of previous approaches. Generally, an earthquake does not immediately trigger aftershocks (so the exponential temporal kernel is not well suited) and the triggering effects may vary for different spatial directions (which implies that the Gaussian spatial kernel does not reflect the underlying process). Furthermore, burglary events present space-time dependencies, due to the 'near-repeat victimization' pattern (Johnson, 2008): burglars often target the same neighborhood repeatedly within a short time frame.

Code. The implementation of our approach is available on GitHub[4].

1.4 Outline of the Thesis

Part I focuses on a statistical learning view of simple Kriging and develops a novel theoretical framework offering non-asymptotic guarantees for empirical simple Kriging rules. The main goal in this first part is to overcome the challenges posed by the characteristics of spatial data, mainly the presence of a strong dependence structure and the observation of one single realization of the phenomenon under study.

- Chapter 2 provides the necessary background to study spatial data with geostatistical tools, and presents the basics of statistical learning, focusing on the empirical risk minimization principle.
- In Chapter 3 we propose tail bounds for the non-parametric covariance estimation of a random field based on a finite and unique sample of observations.
- The final results are provided in Chapter 4 where learning rate bounds are retrieved for the empirical simple Kriging predictor.

[2] https://ncedc.org/
[3] https://data.cityofchicago.org/Public-Safety/Crimes-2001-to-Present/ijzp-q8t2
[4] https://github.com/EmiliaSiv/Flexible-Parametric-Inference-for-Space-Time-Hawkes-Processes

Table 1.1: Summary of the contributions.

Chapters	Contributions
Chapter 3	• Theoretical guarantees for the non-parametric covariance function estimation, under appropriate conditions. • Numerical experiments on simulated data that validates the use of the technical assumptions.
Chapter 4	• Statistical Garantees for the Kriging method. • Numerical experiments on simulated data that verify our theoretical results. • Application to real meteorological data.
Chapter 6	• New, efficient, and flexible method for inference of spatio-temporal Hawkes models. • Numerical experiments on simulated data proving the accuracy and flebility of our method. • Application to real earthquake and burglary data.

Part II is devoted to Hawkes processes and to the study of space-time data from real-world datasets that present heterogeneity. The aim of this part is to develop an efficient and flexible method for parametric inference in space-time Hawkes models.

- Chapter 5 gives the background on point processes useful for our approach and highlights the advantages of developing such a method by investigating real-world examples where heterogeneity and space-time interactions are observed.
- Chapter 6 introduces a novel, flexible, and efficient approach to infer any parametric kernels in the context of space-time Hawkes processes.

Chapter 7 provides a global conclusion and offers an overview of future lines of research and perspectives stemming from the work developed in this thesis.

1.5 Publications

The contributions presented here have resulted in the following publication and preprint:

- (Siviero et al., 2024a) Emilia Siviero, Emilie Chautru, Stephan Clémençon. A Statistical Learning View of Simple Kriging. In *TEST, vol. 33, no 1*, pages 271-296, 2024. Reproduced in Chapters 3 and 4.
- (Siviero et al., 2024b) Emilia Siviero, Guillaume Staerman, Stephan Clémençon, Thomas Moreau. Flexible Parametric Inference for Space-time Hawkes Processes. *arXiv preprint arXiv:2406.06849*, 2024. Reproduced in Chapter 6.

The publication was presented in the following conferences and seminars:

- July 2022: Poster presentation of the paper 'A Statistical Learning View of Simple Kriging', at the French conference on Machine Learning (CAp 2022), Vannes (France).
- August 2022: Oral presentation of the paper 'A Statistical Learning View of Simple Kriging', at the International Conference on Computational Statistics (COMPSTAT 2022), Bologna (Italy).
- March 2023: Oral presentation of the paper 'A Statistical Learning View of Simple Kriging', at the MIND team Seminar, Inria, Palaiseau (France).
- August 2024: Oral presentation of the paper 'Flexible Parametric Inference for Space-time Hawkes Processes', at the International Conference on Computational Statistics (COMPSTAT 2024), Giessen (Germany).

Part I

Statistical Learning for Spatial Data

2
Background

Contents

In this chapter, we provide the necessary background to study spatial data. We first adopt a geostatistical point of view, and then a statistical learning one. In Section 2.1 we introduce the motivations and some examples of real-world spatial data, the key definitions and properties, the main challenges encountered when studying such data, and finally the Kriging method, which is studied in details in the next chapters. In Section 2.2, we present the empirical risk minimization principle, for both the classification and regression problems, and give some results of concentration inequalities. We discuss in detail the strengths and advantages of the definitions, assumptions, and concepts useful for our study, which is the subject of Chapters 3 and 4, focusing on their role in the following chapters.

2.1 Geostatistics

Geostatistics includes powerful tools to study, define, predict, and simulate spatial data. In this section, we first motivate the importance of developing tools to study spatial data and illustrate the discussion by means of real-world examples (subsection 2.1.1). In subsection 2.1.2, we give the key definitions useful for the study of spatial data, with particular attention to the properties of the associated covariance (or semi-variogram) function, such as stationarity and isotropy. Next, in subsection 2.1.3, we present and discuss the main challenges encountered with spatial data, which are the dependence structure that typically characterizes such data and the uniqueness of the observed phenomenon. Then, in subsection 2.1.4, we give a short review of the

estimation of the dependence structure, both in a parametric and non-parametric way. Afterward, we present the main methodologies to simulate spatial data in subsection 2.1.5. Finally, in subsection 2.1.6 we present the key method in geostatistics: the Kriging method, for which we provide the necessary information used in the following chapters.

2.1.1 Motivations – Spatial Data

In this first part of the thesis, we confine our study to geostatistical data. In this case, data are sampled (regularly or irregularly) over the spatial set $\mathcal{S}$ and can be measured on each location of a continuous domain: usually, the data is a (partial) realization of a random process.

Originally, Geostatistics emerged to face the problem of ore grade prediction in a mining block from observed samples. Georges Matheron laid down the theoretical bases for Geostatistics in Matheron (1962) while working on practical problems in mining. About a decade earlier, Danie G. Krige, a mining engineer, had developed practical methods for estimating ore reserves in mining operations (Krige, 1951). His work focused on improving the accuracy of mineral resource estimates, which led to the development of Kriging, a statistical interpolation method named in his honor.

Matheron aimed to give a theoretical and methodological framework to address two issues:

1. Defining a statistical framework for the study of **one unique** realization of a phenomenon, and
2. Taking into account the **spatial correlations** of the data.

The study of spatial data thus differs from the classic theory of statistical learning, where the observed data are assumed to be independent and identically distributed, and for which a large number of repetitions of the phenomenon are available.

Example 2.1. *(Mining) As previously seen in Example 1.3 in Chapter 1, an ore grade within a deposit violates the i.i.d. assumptions of classic statistical learning. Firstly, each ore body is unique, and thus a repetition of the observation is practically impossible. Secondly, mining data has a strong dependence structure, due to natural geological processes that form mineral deposits. A consequence of these natural processes is that ore samples that are spatially close together are more similar to one another than those that are further apart, so that clustering patterns (like high-grade mineral zones) are generally observed.*

In mining, geostatistical methods help to determine the quantity and quality of minerals in the ground, optimizing the extraction process, and reducing financial risk (Matheron, 1963). For example, in the estimation step, Kriging is used to create detailed maps of mineral concentration within the deposit. Using such efficient and accurate estimation methods helps estimate the total ore reserves with greater accuracy and minimize waste extraction.

Example 2.2. *(Hydrology) Hydrology plays a crucial role in managing water resources and assessing groundwater quality. A key characteristic of hydrological data is its spatial dependence. For example, the groundwater levels within an aquifer present spatial dependence induced by the geological characteristics of the area. Understanding and modeling this dependence structure is essential for accurate predictions, effective management of water*

resources, and contamination risk assessment. Geostatistical methods like Kriging, can interpolate the concentration of contaminants to predict the extent of pollution across a given area.

The central goal of Kriging is to compute predicted complete maps of the random process over all the spatial domain $\mathcal{S}$, using a finite number of observations. Before applying this prediction method, some essential steps when dealing with spatial data must be carried out.

Selecting the Optimal Sampling and the Asymptotic Setting. Firstly, one needs to design an optimal sampling plan and decide where to collect the observation data. Two questions arise in this context: How many samples are needed to ensure an efficient statistical study? And how do they need to be distributed over the spatial domain? In order to have a robust estimation, it is preferable that the samples are collected over a regular grid. But, in some real-world applications, this assumption is not satisfied and the samples are randomly distributed over the spatial domain. Furthermore, when determining a sampling plan, one needs to choose the asymptotic setting, based on the nature of the data. Two main asymptotic designs arise, with two different behaviors when the number of observations tends to infinity. The *in-fill* asymptotic stipulates that new observations are made in the same fixed spatial domain. Thus, the spatial domain becomes denser and denser as the number of observations grows. By increasing the density of sampling points, one can capture small-scale spatial variability and improve the precision of spatial predictions. The in-fill setting is appropriate for real-world situations where the domain of interest is fixed and where high-resolution spatial data are needed. For example, this setting is used in meteorology to study localized weather phenomena or weather forecasts in a country or in a specific region, thanks to its ability to capture small-scale variations in temperature. The in-fill setting is also preferred in other domains, such as mining and hydrology, since the area of interest is often thought of as bounded. See for example Stein (1988); Yakowitz and Szidarovszky (1985); Stein (1995) for a theoretical study of geostatistical methods under the in-fill asymptotic. On the contrary, the *increasing domain* (or *out-fill*) asymptotic assumes that new samples are taken outside the current spatial domain of observations. In this case, the spatial domain becomes wider and wider as new observations are collected while the spatial density remains fixed. A minimum distance between neighboring sampling locations is assumed (Mardia and Marshall, 1984; Sherman and Carlstein, 1994). This is particularly useful in situations where large-scale spatial patterns and trends are of interest. By expanding the spatial domain while maintaining a consistent sampling density, one can gain insights into the wider spatial dependencies and variations in the data. In epidemiology, increasing domain asymptotics are useful to study the spread of a disease across large geographical regions. Finally, some researchers considered a hybrid approach (see *e.g.* Hall and Patil, 1994; Lahiri, 1999; Lahiri et al., 1999; Putter and Young, 2001), where a combination of in-fill and increasing domain asymptotics is assumed. Often one assumes that both the size of the spatial domain of the observations and the number of observations in each of its subsets grow with the number of observations. This approach makes it possible to capture both large-scale patterns and small-scale variability, and thus gives a comprehensive spatial understanding at multiple scales. For example, in the case of urban planning, one needs to understand both the overall growth patterns of metropolitan areas and the detailed dynamics within neighborhoods. The statistical properties of the estimators may largely vary depending on the chosen asymptotic

(Cheng et al., 2020; Cressie, 1993). Still, if one is interested in obtaining an efficient and accurate interpolation, the in-fill asymptotic is preferable (Stein, 1999; Chang and Stein, 2013).

In our study

Asymptotic Setting: Since our focus in this thesis is on the accuracy of the Kriging interpolation method, we opt for the in-fill asymptotic setting (see Section 3.2): the number of observations inside the fixed and bounded spatial domain $\mathcal{S}$ increases. This setting is adapted for empirical semi-variogram estimation, since the maximum distance h_{max} is fixed. More observations will induce more pairs of points that are at a distance h from one another (for all $h < h_{max}$), thus providing more elements for the computation of the empirical semi-variogram $\widehat{\gamma}(h)$ defined in Equation (2.3).

Assessing the Stationarity of the Data. In order to address the difficulties posed by the fact that spatial phenomena typically consist of a single and irreproducible observation, one may formulate the assumption of stationarity (Hall and Patil, 1994; Hall et al., 1994). This assumption is an alternative to the hypothesis on i.i.d. repetitions of the phenomenon in the classic statistical learning setting, and makes it possible to rely on a successful frequentist approach (Cressie, 1993; Gaetan and Guyon, 2009; Kanevski et al., 2009). Still, in cases where the data does not fit the stationarity assumption, one may prefer to assume spatial non-stationarity (or only local stationarity) and try to detect it (see *e.g.* Bel (2004) for non-stationary random fields and Donoho et al. (1996) for locally stationary processes). See subsection 2.1.2 for further details on the stationarity assumption, together with other key assumptions in Geostatistics (such as isotropy and ergodicity).

Estimating the Spatial Correlation within the Data. Since the main characteristic of spatial data is the presence of correlation within the data, an important step in Geostatistics is to understand the dependence structure of the data. This process involves quantifying how data points are related to each other based on their spatial proximity. The key tools for this purpose are the covariance and semi-variogram functions, which describe how the spatial correlation between data points changes with distance. However, in practice, the semi-variogram (and the covariance) of a random field is unknown and thus needs to be estimated. The experimental semi-variogram modeling (see subsection 2.1.3 for further details) is the first stage to understand the dependence structure. It allows one to identify the main characteristics of the dependencies within the data and construct a first experimental version of the semi-variogram. A detailed discussion on the key tools to define the dependence structure of the data can be found in subsection 2.1.3. Based on the information collected during the experimental modeling, the next stage is to estimate the semi-variogram function accurately. A review of both parametric and non-parametric dependence structure estimations is given in subsection 2.1.4.

Interpolation and Complete Maps. The final goal of Kriging is the estimation of the random process at the unobserved locations of the spatial domain, based on a finite number of observations. The predictor is constructed as a linear combination of the observations. By applying the Kriging method to all unobserved points of the

domain, one may draw a complete map of the random field. We refer the reader to subsection 2.1.6 for further details on the Kriging prediction method. When the semi-variogram function is known, the Kriging predictor is optimal. In practice the dependence structure of the data is unknown and the true semi-variogram values are replaced in the Kriging equations by their estimated counterparts. In subsection 2.1.6 we provide a quick review of previous results concerning the robustness of the Kriging method to the uncertainty effects of the semi-variogram estimation. In the case of an unknown semi-variogram function, there are no theoretical guarantees of optimality. It is then necessary to establish rate bounds that assess the generalization capacity of the resulting empirical predictive map. The main motivation of the first part of this thesis is to develop a framework for Kriging based on a non-asymptotic study of the performances of a non-parametric semi-variogram estimator. The methodology and results are given in Chapters 3 and 4.

2.1.2 Definitions – Geostatistical Tools

In this subsection, we recall the main definitions and properties useful for the study of geostatistical data. The results presented here are drawn from Chapter 2 in Cressie (1993) and Chapter 1 in Gaetan and Guyon (2009).

Let $\mathcal{S} \subset \mathbb{R}^p$ be a spatial set and $\mathbf{X}$ be a random field on $\mathcal{S}$ with $\mathbb{R}$ as state space, *i.e.* a collection $\mathbf{X} = \{\mathbf{X}_s : s \in \mathcal{S}\}$ of real-valued random variables (r.v.'s) defined on the same probability space, $(\Omega, \mathcal{F}, \mathbb{P})$ say, indexed by $s \in \mathcal{S}$. Suppose that we observe $\mathbf{X}$ at n spatial locations $(s_i)_{1\leq i\leq n} \in \mathcal{S} \subset \mathbb{R}^p$. The random field $\mathbf{X}$ is fully characterized by its distribution function:

$$F\left(s_1, \cdots, s_n; x_1, \cdots, x_n\right) = \mathbb{P}\left(\mathbf{X}_{s_1} \leq x_1, \cdots, \mathbf{X}_{s_n} \leq x_n\right),$$

$\forall n \in \mathbb{N}^*, \forall (s_i)_{1\leq i\leq n} \in \mathcal{S}^n, \forall (x_i)_{1\leq i\leq n}$.

Strict Stationarity. In spatial statistics, one unique realization of the phenomenon is generally available. In order to realize statistical inference for one unique event, one must somehow replace the hypothesis on independent repetitions of the phenomenon with a hypothesis on the random field: it is the role of the stationarity assumption. Intuitively, a random process can be considered stationary if its characteristics do not vary along the spatial domain: thus, thanks to multiple observations of the process in distinct locations, we effectively have access to multiple instances of a single random process, enabling statistical analysis.

Definition 2.3. *(Strict Stationarity) A random field is strictly stationary if its spatial distribution is invariant by translation:*

$$\mathbb{P}\left(\mathbf{X}_{s_1} \leq x_1, \cdots, \mathbf{X}_{s_n} \leq x_n\right) = \mathbb{P}\left(\mathbf{X}_{s_1+\boldsymbol{h}} \leq x_1, \cdots, \mathbf{X}_{s_n+\boldsymbol{h}} \leq x_n\right),$$

$\forall (s_i)_{1\leq i\leq n} \in \mathcal{S}^n, \forall (x_i)_{1\leq i\leq n}$ *and* $\forall \boldsymbol{h} \in \mathbb{R}^p$.

The notion of stationarity implies that the phenomenon is sufficiently homogeneous (its characteristics are identical from one point to another) within the spatial domain so that we can replace a repetition of the random field with a repetition in the space.

However, in practice, this assumption is often not feasible since the finite number of observations provides insufficient information about the mean and variance.

We present a practicable relaxation of the strict stationarity. To do so, let's first define the class of second-order random fields and their associated covariance functions.

Definition 2.4. *(Second-order Random Field) A random field is of second-order if* $\forall s \in \mathcal{S}, \mathbb{E}\left[\mathbf{X}_s^2\right] < \infty$.

Thus, the mean of the process $\mathbf{X}$ exists and is denoted $\mu(s) = \mathbb{E}[\mathbf{X}_s], \forall s \in \mathcal{S}$.

Covariance Function. The covariance function for a second-order random process can be defined as follows.

Definition 2.5. *(Covariance Function) The covariance function of a second-order random field is defined as*

$$\forall (s, s') \in \mathcal{S}^2, Cov\left(\mathbf{X}_s, \mathbf{X}_{s'}\right) = \mathbb{E}\left[\left(\mathbf{X}_s - \mu(s)\right)\left(\mathbf{X}_{s'} - \mu(s')\right)\right].$$

Let C be the covariance function:

$$\forall (s, s') \in \mathcal{S}^2, C(s, s') = Cov\left(\mathbf{X}_s, \mathbf{X}_{s'}\right).$$

Proposition 2.6. *A covariance function C of a second-order random field has the following properties:*

1. *the covariance function is symmetric:* $\forall (s, s') \in \mathcal{S}^2, C(s, s') = C(s', s)$.
2. $\forall s \in \mathcal{S}, C(s, s) = Var(\mathbf{X}_s)$.
3. *the covariance function is positive semi-definite:* $\forall \lambda \in \mathbb{R}^n, \sum_{i=1}^{n} \sum_{j=1}^{n} \lambda_i \lambda_j C(s_i, s_j) \geq 0$.

Notice that the third property is a necessary condition for covariance functions since it results from $Var\left(\sum_{i=1}^{n} \lambda_i \mathbf{X}_{s_i}\right) = \sum_{i=1}^{n} \sum_{j=1}^{n} \lambda_i \lambda_j C(s_i, s_j)$ and that the variance is always non-negative.

Second-order Stationarity. Finally, we are able to define second-order (or weak) stationarity, a relaxed and more practicable version of strict stationarity.

Definition 2.7. *(Second-order Stationarity) A random field is second-order (or weak) stationary if its two first moments exist and are invariant:*

1. *the mean is finite and invariant by translation (constant):* $\forall s \in \mathcal{S}, \mathbb{E}[\mathbf{X}_s] = \mu \in \mathbb{R}$, *and*
2. *the covariance is invariant by translation:* $\forall s, s' \in \mathcal{S}, C(s, s') = C\left(s + \boldsymbol{h}, s' + \boldsymbol{h}\right)$, *for all translation* $\boldsymbol{h} \in \mathbb{R}^p$.

Note that the covariance depends only on the vector of distance $\boldsymbol{h}$. With a slight abuse of notation, we let C the function such that: $C(\boldsymbol{h}) = Cov\left(\mathbf{X}_s, \mathbf{X}_{s+\boldsymbol{h}}\right)$. The function C satisfies the following properties, similar to those of Proposition 2.6:

Proposition 2.8. *The covariance function C of a second-order stationary random field has the following properties:*

1. *the covariance is an even function:* $C(-\boldsymbol{h}) = C(\boldsymbol{h})$.
2. *the covariance in* $\boldsymbol{h} = \mathbf{0}$ *is equal to the variance of the random field:* $C(\mathbf{0}) = Var(\mathbf{X}_s) > 0$.
3. *the covariance function is positive semi-definite:* $\forall \lambda \in \mathbb{R}^n$, $\sum_{i=1}^{n} \sum_{j=1}^{n} \lambda_i \lambda_j C(s_i - s_j) \geq 0$.

Thanks to the second-order stationarity assumption, we can define a covariance function depending only on the difference $\boldsymbol{h} = s - s'$ between two data locations.

Isotropy. In the next definition, we add another definition for the case when the covariance depends only on the (Euclidean) distance $h = \|s - s'\|$:

Definition 2.9. *(Isotropy) The covariance is said to be isotropic if it depends only on the (Euclidean) distance between the spatial points: it exists a function* $c : \mathbb{R}_+ \to \mathbb{R}$ *such that* $Cov\left(\mathbf{X}_s, \mathbf{X}_{s'}\right) = C(s - s') = c(\|s - s'\|)$.

On the contrary, when the direction of the difference $\boldsymbol{h} = s - s'$ between two data locations matters, the covariance is said to be anisotropic.

In our study

Stationarity and Isotropy: In Chapters 3 and 4, we assume that we have access to one unique realization of the phenomenon under study, that presents a strong dependence structure. To ensure an accurate estimation of the dependence structure, and thus an efficient prediction through Kriging, we must assume the following: the random field is second-order stationary and its covariance function is isotropic (see Section 3.2 and Assumption 3.2).

Intrinsic Stationarity. In order to relax even more the stationary assumption, we consider the increment process $\left\{\mathbf{X}_{s+\boldsymbol{h}} - \mathbf{X}_s, s \in \mathcal{S}\right\}$.

Definition 2.10. *(Intrinsic Stationarity) A random process* $\mathbf{X}$ *is intrinsically stationary if its increments* $\mathbf{X}_{s+\boldsymbol{h}} - \mathbf{X}_s$ *are second-order stationary:*

1. $\forall s \in \mathcal{S}, \forall \boldsymbol{h} \in \mathbb{R}^p, \mathbb{E}\left[\mathbf{X}_{s+\boldsymbol{h}} - \mathbf{X}_s\right] = 0$, *and*
2. $\forall s \in \mathcal{S}, \forall \boldsymbol{h} \in \mathbb{R}^p, Var\left(\mathbf{X}_{s+\boldsymbol{h}} - \mathbf{X}_s\right) = 2\gamma(\boldsymbol{h})$, *where* $\gamma(\boldsymbol{h})$ *is the semi-variogram of the random field.*

Semi-variogram Function. The intrinsic stationarity assumption introduces the semi-variogram function of a random field. Notice that the semi-variogram function exists for all second-order and intrinsic stationary functions. We present its definition and give its main properties below.

Definition 2.11. *(Semi-variogram) The semi-variogram of a second-order (or intrinsic) stationary r.v. is defined as:*

$$\begin{aligned}\gamma(\boldsymbol{h}) &= \frac{1}{2}\mathbb{E}\left[\left(\mathbf{X}_{s+\boldsymbol{h}} - \mathbf{X}_s\right)^2\right] \\ &= \frac{1}{2}Var\left(\mathbf{X}_{s+\boldsymbol{h}} - \mathbf{X}_s\right).\end{aligned}$$

The term variogram refers to the quantity $2\gamma(\boldsymbol{h})$.

In Geostatistics, the use of the semi-variogram is often preferred to the covariance function for several reasons. More information about the semi-variogram function and a discussion about its advantages over the covariance function can be found in subsection 2.1.3, dedicated to the dependence structure of the random field. Still, one can easily go from the semi-variogram to the covariance function thanks to the following equation.

Proposition 2.12. *(Relationship between covariance and semi-variogram) Let* $\mathbf{X}$ *be a second-order or intrinsic stationary random field, and define by* C *(respectively* γ*) its covariance function (resp. its semi-variogram function). Then, if the semi-variogram* γ *is bounded, for all* $\boldsymbol{h} \in \mathbb{R}^p$,

$$\gamma(\boldsymbol{h}) = C(0) - C(\boldsymbol{h}). \tag{2.1}$$

Proposition 2.13. *The semi-variogram function* γ *of a second-order (or intrinsic) stationary random field has the following properties:*

1. *the semi-variogram is an even function:* $\gamma(-\boldsymbol{h}) = \gamma(\boldsymbol{h})$.
2. *the semi-variogram in* $\boldsymbol{h} = \mathbf{0}$ *is null:* $\gamma(\mathbf{0}) = 0$.

The proofs are easily derived from Definition 2.11 and Proposition 2.8.

Note that every stationary process is intrinsic. The reciprocal is generally false. The following property states one case where an intrinsic stationary process $\mathbf{X}$ is second-order stationary, if its semi-variogram function satisfies a condition.

Proposition 2.14. *If the semi-variogram* $\gamma(\boldsymbol{h})$ *of an intrinsic random field* $\mathbf{X}$ *satisfies* $\lim_{\|\boldsymbol{h}\|\to\infty} \gamma(\boldsymbol{h}) < \infty$, *then* $\mathbf{X}$ *is a second-order stationary random field.*

Proposition 2.15. *The semi-variogram* γ *of an intrinsic random field* $\mathbf{X}$ *is conditionally negative definite, i.e. it satisfies:* $\forall(\lambda_i)_{i\leq n}$ *such that* $\sum_{i=1}^{n} \lambda_i = 0$ *(called authorized linear combinations),* $\mathbb{E}\left[\sum_{i=1}^{n} \lambda_i \mathbf{X}_{s_i}\right] = 0$ *and* $Var\left(\sum_{i=1}^{n} \lambda_i \mathbf{X}_{s_i}\right) = -\sum_{i=1}^{n}\sum_{j=1}^{n} \lambda_i\lambda_j\gamma(s_i - s_j) \geq 0.$

Ergodicity. A complementary assumption to enable statistical inference for spatial data is the ergodic condition, stated below. Ergodicity allows for an accurate estimation of the statistical properties of a random field, from a single realization of it. Let $\mathbf{X}(\mathcal{S}) = \frac{1}{|\mathcal{S}|}\int_{\mathcal{S}} \mathbf{X}_s ds$.

Definition 2.16. *(Ergodicity) A second-order (or intrinsic) stationary random field* $\mathbf{X}$*, with constant mean* μ*, is said to be ergodic if* $\lim_{|\mathcal{S}|\to\infty} \mathbb{E}\left[\left(\mathbf{X}(\mathcal{S}) - \mu\right)^2\right] = 0.$

This means that the spatial averages taken over a large enough spatial domain will almost surely converge to the expected values over all possible realizations. Still, this unique realization must be large enough for the estimation of the parameters to be accurate (Lantuéjoul, 1991).

In our study

Ergodicity: In Chapters 3 and 4, we make the widely used assumption of ergodicity to ensure an efficient prediction of the random field $\mathbf{X}$.

Non- and Local Stationarity. Notice that other categories of random processes have been getting more and more attention in the last few years. It is the case of non-stationary and locally stationary random processes. Non-stationarity refers to a random process in which distribution varies in the spatial domain (the mean and the variance may change from one point to another). Indeed, in some real-world situations, non-stationary assumptions are well-adapted and allow more flexible models, for example in the case of the spread of diseases (Dong et al., 2023) and in the study of air pollution (Sampson and Guttorp, 1992; Bel, 2004). Locally stationary variables exhibit the stationary property within small regions of the spatial domain. Thus, the statistical properties, such as mean and variance, gradually change over space. This means that within these local regions, traditional statistical techniques for stationary processes may be applied locally. In Kurisu (2022), locally stationary random fields are expressed as non-stationary random fields that can be locally approximated by stationary ones.

Gaussian Process. We now briefly present Gaussian processes, a powerful tool for modeling complex data in Geostatistics (Cressie, 1993).

Definition 2.17. *A random field* $\mathbf{X}$ *is Gaussian if each finite linear combination follows a Gaussian distribution:* $\forall A \subset \mathcal{S}, \forall a = \{a_s, s \in A\}$, $\sum_{s\in A} a_s \mathbf{X}_s$ *is a Gaussian distribution.*

Let $\mathbf{X}_A = \{X_s, s \in A\}$ a Gaussian process of mean $\mu_A = \mathbb{E}[X_A]$ and covariance Σ_A. Then, the density is: $f_A(x_A) = \frac{1}{(2\pi)^{-|A|/2}\sqrt{Det(\Sigma_A)}} \exp\left(-\frac{1}{2}(x_A - \mu_A)^\top \Sigma_A^{-1}(x_A - \mu_A)\right)$.

Since a Gaussian process is fully characterized by its mean and its covariance function, second-order stationarity is equivalent to strict stationarity (Cressie, 1993, Chapter 2). Furthermore, a sufficient condition for ergodicity is $\lim_{\|\boldsymbol{h}\|\to\infty} C(\boldsymbol{h}) = 0$ (Cressie, 1993, Chapter 2).

In our study

Gaussian Process: In Chapters 3 and 4, we assume that the random field $\mathbf{X}$ is Gaussian (see Assumption 3.5). This assumption allows us to obtain strict stationarity of the random field and is a key argument to develop non-asymptotic bounds for the accuracy of the Kriging predictor. Indeed, under the Gaussian assumption, the empirical semi-variogram can be seen as a sum of independent χ^2 variables (Proposition 3.7). We refer the reader to Section 3.2 for further details.

Spectral Representation. The spectral representation is a key concept in Geostatistics. The main result comes from Bochner's theorem (see *e.g.* Stein, 1999, Chapter 2). It states a sufficient and necessary condition for a function C to be the covariance function of a second-order stationary random field.

Theorem 2.18. *(Bochner's Theorem, (Stein, 1999, Section 2.5)) A function C is the covariance function of a second-order stationary random field if and only if it satisfies*

$$C(\boldsymbol{h}) = \int_{\mathbb{R}^p} \exp\left(iu^\top \boldsymbol{h}\right) F(du),$$

where F is a positive measure on $\mathbb{R}^p$ such that $\int_{\mathbb{R}^p} F(du) = C(0) < \infty$.

When F has a density with respect to the Lebesgue measure on $\mathbb{R}^p$, we denote it Φ. The function Φ is called the *spectral density* of $\mathbf{X}$ and can be expressed in terms of the covariance function C thanks to the inverse Fourier transform (see *e.g.* Stein, 1999, Chapter 2; Yaglom, 1987)

$$\Phi(u) = (2\pi)^{-p} \int_{\mathbb{R}^p} \exp\left(-iu^\top \boldsymbol{h}\right) C(\boldsymbol{h}) d\boldsymbol{h}.$$

Furthermore, if $\mathbf{X}$ is a second-order random field with isotropic covariance function c, then the spectral density Φ is also isotropic and defined by

$$\Phi(u) = (2\pi)^{-p} \int_{\mathbb{R}^p} \exp\left(-iu^\top \boldsymbol{h}\right) c(\|\boldsymbol{h}\|) d\boldsymbol{h} = \phi(\|u\|). \tag{2.2}$$

As stated in subsection 2.1.5, the notion of the spectral representation is also used for the simulation of second-order stationary random fields.

In our study

Spectral Representation: The spectral representation and Bochner's theorem come into play in Section 3.2 to define an additional assumption necessary for our theoretical results (see Assumption 3.8 in Chapter 3).

To summarize, the key concepts and assumptions on the statistical properties of a random function are the following:

- The covariance and semi-variogram functions of a second-order or intrinsic stationary random process depend only on the distance between the data locations: the spatial coordinates of the data points do not have an influence on the statistical properties of the random process, and there are no spatial trends.
- The stationarity and ergodicity assumptions allow sound statistical inference: the observations at different regions of the spatial domain can be seen as independent repetitions of the process, and statistical averaging over the repetitions of the random field can be replaced by averaging over the repetitions of the spatial domain.

From now on, unless the contrary is mentioned, we assume that the random field $\mathbf{X}$ is ergodic, second-order stationary, and isotropic.

2.1.3 Dependence Structure within Spatial Data

The main characteristic of spatial data is that it presents a strong dependence structure, which reflects the fact that close data (*i.e.* their locations are close in the spatial domain) are correlated, while, as data locations get more distant, the correlation decreases. This spatial dependence is central in many fields, where understanding how a process varies across a spatial domain informs predictions.

Example 2.19. *(Weather Forecasting) For example, in temperature forecasting, weather stations located close to each other often record similar temperatures because they are both affected by local weather conditions, like mountains or bodies of water. This spatial coherence means that knowing the temperature at one location can provide valuable information about nearby areas. Understanding the dependence structure of spatial data allows better anticipating local variations in weather phenomena, aiding in more precise and reliable predictions for various applications. Meteorologists use techniques like spatial interpolation and Geostatistics to model and predict these spatial dependencies, helping to improve the accuracy of weather forecasts over larger regions.*

The dependence structure must be taken into account when predicting spatial data. The spatial dependence structure can be quantified and modeled using statistical tools, primarily the covariance function and the semi-variogram function of the process (both described in the previous subsection).

While the covariance function is useful, the semi-variogram function is often preferred in practice for several reasons (Robinson, 1990; Cressie and Zimmerman, 1992); (Cressie, 1993, Chapter 2). First, unlike the covariance function, the computation of the semi-variogram does not require the knowledge of the (supposedly constant) mean. The semi-variogram empirical estimation is unbiased (see subsection 2.1.4), while the mean of the random field in the covariance empirical estimation introduces a bias. Second, the class of variogram functions is broader (implying less restrictive assumptions). Third, it is well-known that the sum of the square of dependent random variables of normal laws gives a χ^2 random variable. Thus, all the distributions in the empirical semi-variogram function are known (see Proposition 3.7 in Chapter 3). However, the product of dependent random variables (for the covariance estimation, see Equation (3.2) in Chapter 3) gives a much more complicated distribution (see the result in Nadarajah and Pogány (2015) on the distribution of the product of correlated normal r.v.'s). Furthermore, other advantages of the semi-variogram function

are: a more intuitive understanding of the spatial variability since it is equal to the average squared difference based on the distance; and, the semi-variogram explicitly accounts for the nugget effect (see below for a definition), observed as a discontinuity at the origin of the semi-variogram.

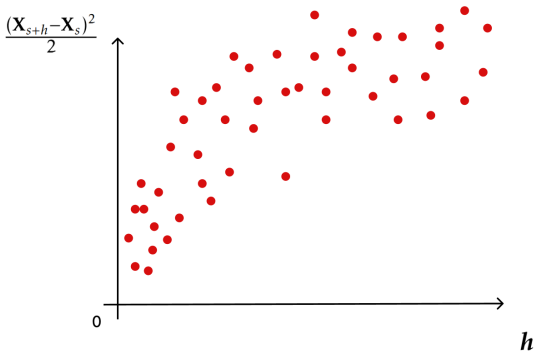
In our study

Semi-variogram Function: In Chapter 3, we move from the semi-variogram to the covariance function thanks to the relationship (2.1) between them, for the following reasons: under the Gaussian assumption, the empirical semi-variogram can be written as a sum of independent χ^2 variables (see Proposition 3.7), thus tail bounds can be derived for the semi-variogram; then, tail bounds stem from for the covariance matrix.

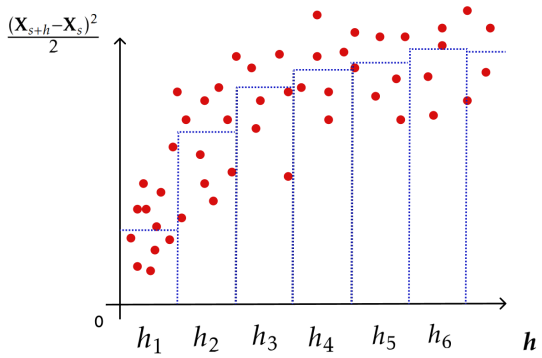
Before estimating the dependence structure of the process, it is essential to identify its main characteristics. Indeed, each phenomenon has its spatial variability characteristics, that are valuable information. In the following, we consider an isotropic semi-variogram function (refer to subsection 2.1.2 for a definition of isotropy). To do so, one can use experimental semi-variogram modeling (Gaetan and Guyon, 2009, Chapter 5) by plotting the differences between values at pairs of locations given by $\frac{(\mathbf{X}_{s+h}-\mathbf{X}_s)^2}{2}$ against the distance h separating them. The resulting graphic is called the variogram cloud (see Figure 2.1a for an example). Then, one can add the class of distances, and obtain a plot similar to a histogram (see Figure 2.1b): the distances are divided into intervals (or lags), and the average value of the differences between pairs contained in the intervals is plotted. Finally, this allows one to obtain a theoretical model (see Figure 2.1c) by plotting a function adapted to the obtained values for each class of distances. One can observe the correlation principle: pairs of locations that are spatially closed together should have similar values. Indeed, points on the left side of the variogram cloud (that represent points separated by a small distance h), have small values, while as the distance h becomes larger, the values become higher.

Studying precisely the form of the variogram allows a better understanding of the spatial correlation between the data and of the underlying process. Three components become visible on the plot of the empirical semi-variogram (see Figure 2.2): the nugget effect, the range, and the sill. These parameters are critical for interpreting the spatial structure of the data and building models that accurately reflect this structure.

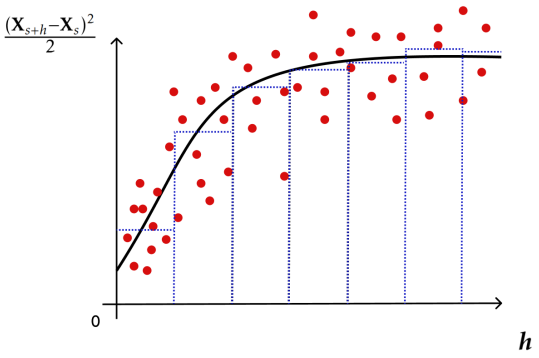
Nugget effect. As seen in the previous subsection, the value of the semi-variogram at the origin $h = 0$ is equal to 0. However, for very small distances, the empirical semi-variogram generally exhibits an abrupt change, which is called the *nugget effect* (Matheron, 1962; Cressie, 1993). In Figure 2.2, the nugget effect on the semi-variogram can be seen at the origin (in orange): the value of the empirical semi-variogram is not equal to 0. In Geostatistics, the term 'nugget effect' is used to describe the behavior of the semi-variogram near the origin. Indeed, it refers to the variability observed at very short distances within the data. Since no information from the data is generally available for infinitely small distances, the nugget effect is attributed to measurement errors or sources of spatial variation at distances smaller than the minimum spacing between sampling points. The nugget effect represents the variability in data that is not explained by spatial correlation. For example, in environmental studies, measure-



(a) The variogram cloud.



(b) The class of distances h.



(c) The experimental semi-variogram.

Figure 2.1: The three steps of the empirical semi-variogram modeling. The red dots represent the values of $\frac{(\mathbf{X}_{s+h}-\mathbf{X}_s)^2}{2}$ for each pair of data samples $(\mathbf{X}_{s+h}, \mathbf{X}_s)$, the dotted blue lines define classes of distances, and the fitted experimental semi-variogram is in black.

ments of soil contamination at spatially closed locations may exhibit a nugget effect due to small-scale contamination levels that contribute to local variations.

Range. In Figure 2.2, it is possible to observe that the semi-variogram function stabilizes after a certain distance h. This distance is the *range* of the semi-variogram (represented in blue in the graphic). It represents the distance at which the correlation between spatial locations becomes negligible. The knowledge of the range gives valuable information on the dependence structure of the data: locations separated by a distance less than the range are spatially correlated, whereas locations separated by a distance greater than the range become statistically independent. For example, in geological surveys, the range defines the distance scale at which rock properties such as mineral grades are correlated, influencing resource exploration and extraction strategies. As it is pointed out in Chapter 3, the range is an important tool in Kriging prediction since it determines the spatial distance range within which values at unobserved locations can be accurately estimated using available data.

Sill. The value at which the semi-variogram reaches the range is the *sill* (represented in green in Figure 2.2). The sill appears as the asymptotic value of the variogram at large distances. It is essential in spatial modeling since it provides insights into how much spatial variability is present.

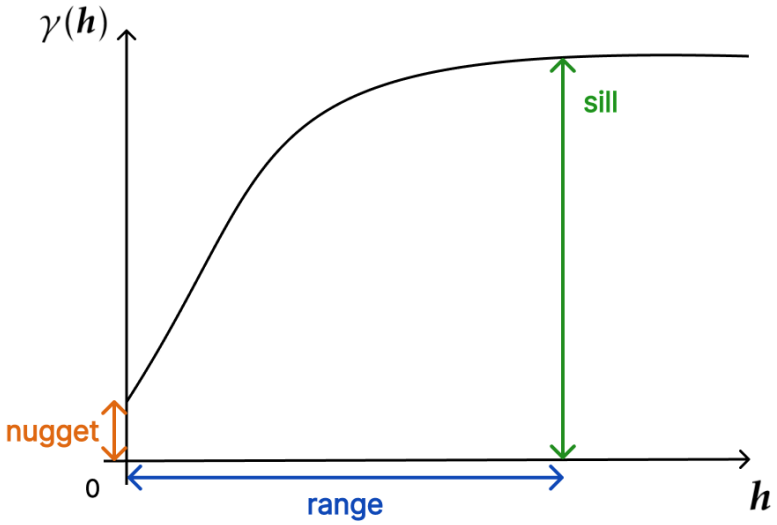


Figure 2.2: Parameters of the semi-variogram: the nugget effect (in orange), the range (in blue), and the sill (in green).

2.1.4 Covariance and Semi-variogram Estimation

Now that the basis for the dependence structure has been laid, we discuss the different procedures for the dependence structure estimation, separating them into two main categories: the non-parametric approach and the parametric models.

The estimation of the dependence structure of a random field is a delicate phase, for several reasons. In spatial prediction such as Kriging, its accuracy heavily depends on the accurate estimation of the covariance (or semi-variogram) function. Indeed, in Kriging, the covariance (or semi-variogram) function defines the weights given to each sample value in the prediction of unobserved locations.

The first step of the semi-variogram estimation of a random field is, as stated in subsection 2.1.3, the experimental semi-variogram modeling. This allows one to obtain a variogram cloud, calculated directly from the data, and to analyze graphically the dependence structure of the process. Then, one can either opt for a non-parametric approach or for a parametric point of view. The non-parametric approach consists of using the derived experimental semi-variogram on the observations without assuming any specific model. In parametric modeling, one should choose a semi-variogram model to fit the observed variogram cloud: the experimental semi-variogram is used as a preliminary step to understand the spatial structure before fitting a more complex model.

In the following, we assume that the random field $\mathbf{X}$ is observed at n spatial locations $\{s_1, \cdots, s_n\}$ of the domain $\mathcal{S}$.

Non-Parametric Estimation. The most common non-parametric approach is the Matheron (1962) empirical semi-variogram estimator, based on the method of moments, under the assumption that the mean of the process is constant. It is defined by the

following equation:

$$\widehat{\gamma}(h) = \frac{1}{2n_h} \sum_{(s_i,s_j)\in N(h)} \left(\mathbf{X}_{s_i} - \mathbf{X}_{s_j}\right)^2, \tag{2.3}$$

where $N(h) = \left\{(s_i,s_j) \in \mathcal{S}^2, \|s_i - s_j\| = h, (i,j) \in [\![1,n]\!]^2\right\}$ is the set of pairs of sites that are at a distance h from one another (also called the set of neighbors) and $n_h = |N(h)|$ denotes its cardinality. This estimator is unbiased when the random field is intrinsic or second-order stationary (Cressie, 1993, Section2).

The first property for the theoretical semi-variogram function in Proposition 2.13 is satisfied by the empirical semi-variogram estimator: $\widehat{\gamma}$ is an even function such that $\widehat{\gamma}(h) = \widehat{\gamma}(-h)$. Note that this estimator can be defined for second-order stationary fields that do not necessarily have an isotropic covariance function. Still, since in the following we make the isotropic assumption, the focus here is on the isotropic definition only.

This estimator is well-suited when one assumes that the observations are on a regular grid. Most importantly, one must have access to a large number of observations that are at a distance h from one another to ensure an accurate and stable estimator (see Section 3.2 and Lemma 3.6 for further details). For the irregular case, one may slightly modify the empirical estimator by introducing a tolerance term in the definition of the neighboring sets. For example, define the set of pairs that are at a distance more or less h with error $\varepsilon > 0$ as $N_\varepsilon(h) = \left\{(s_i,s_j), h-\varepsilon \leq \|s_i - s_j\| \leq h+\varepsilon, (i,j) \in [\![1,n]\!]^2\right\}$. Still, the introduction of a tolerance term ε may induce a bias in the semi-variogram estimation.

However, this empirical estimator may be very affected by the presence of outliers.

Cressie and Hawkins (1980) proposed an alternative estimator that is more robust against outliers in the case of normal-like distributions:

$$\widehat{\gamma}(h) = \frac{1}{C_h}\left(\frac{1}{2n_h} \sum_{(s_i,s_j)\in N(h)} \left|\mathbf{X}_{s_i} - \mathbf{X}_{s_j}\right|^{1/2}\right)^4,$$

where $C_h = \left(0.457 + \frac{0.494}{n_h} + \frac{0.045}{n_h^2}\right)$ is a correction factor for bias. This estimator is robust against non-normality for distributions that are normal in the central region and heavier in the tails.

Another robust alternative is the estimator proposed in Rousseeuw and Croux (1992), based on a highly robust estimator of scale $V(h) = \mathbf{X}_{s+h} - \mathbf{X}_s$ (see also Genton, 1998).

One can also define the Median estimator (see *e.g.* Cressie, 1993) by:

$$\widehat{\gamma}(h) = \frac{1}{2B_h}\left(med\left\{\left|\mathbf{X}_{s_i} - \mathbf{X}_{s_j}\right|^{1/2}, (s_i,s_j) \in N(h)\right\}\right)^4,$$

where B_h is a correction factor for bias.

A comparison of these estimators is given in Mingoti and Rosa (2008). The results are that the best estimators for contaminated data are the Genton (1998) and Median estimators, while in the absence of outliers, the Matheron (1962) estimator gives the best results.

Further non-parametric estimators have been developed in Hall et al. (1994); Elogne et al. (2008) for the covariance function of a stationary stochastic process (temporal process) and in Hall and Patil (1994) in the case of a stationary random field (spatial process). The common motivation behind these works is the fact that non-parametric approaches usually have difficulties in imposing the positive definition property of the covariance estimator. Indeed, this property is a fundamental assumption since it ensures that the prediction variances are non-negative for the optimal Kriging predictor. Furthermore, if the resulting estimator is positive definite, then it is itself a covariance function. Hall and Patil (1994) use Fourier characterization and Bochner's theorem to obtain a positive definite kernel type estimator. It is proven that the derived estimator is consistent, under some assumptions.

Empirical estimators are data-driven methods (since they rely solely on the observations of the phenomenon) that present several advantages over parametric estimators. First of all, the non-parametric estimation provides flexibility, since it does not assume any specific model for the semi-variogram. This flexibility allows non-parametric methods to capture a wide range of spatial dependencies directly from the data. Secondly, non-parametric approaches provide adaptability, which is a valuable advantage when dealing with complex and heterogeneous spatial data. The complexity of the data may be encountered in several forms, like anisotropy or non-stationarity of the random field. Indeed, since non-parametric methods do not assume a structure of dependence, they may adapt more easily to complex spatial data, whereas this complexity may not be well-represented by classical parametric models.

In our study

Non-parametric Estimation: The massive character of spatial datasets now available suggests resorting to more flexible, non-parametric, approaches to analyze spatial observations. In Chapter 3, the empirical Matheron (1962) semi-variogram estimator is chosen and its accuracy is analyzed by means of tail bounds.

However, one must be careful when applying non-parametric methods for the following two reasons. First, the empirical estimation should be done on a finite number of distances h, for which a large enough number of observations is available in the sampled data. In Section 3.2, we define $\mathcal{H}_n = \left\{ \|s_i - s_j\| : (i,j) \in [\![1,n]\!]^2 \right\}$ the set of observed lags and we state a lemma ensuring that the number of terms averaged in the empirical semi-variogram is large enough to obtain accurate estimations for these lags (see Lemma 3.6). Second, for unobserved distances, the empirical estimation can be extrapolated by means of various non-parametric procedures. In Section 3.2, we choose the 1-NN piecewise constant estimator. As it is shown in the subsequent Chapter 3, we are able to give a bound on the estimation error at all distances (both observed lags from $\mathcal{H}_n$ and unobserved ones), assuming a smoothness condition for the covariance function.

In our study

Accurate Empirical Estimation: In Chapter 3, we assume that the observations forms a regular grid. To obtain an accurate empirical estimation, we derive a result ensuring that the number of terms averaged in the empirical semi-variogram estimation is large enough (see Lemma 3.6).

In our study

Estimation at Unobserved Lags: In Chapter 3, we introduce a smoothness condition on the covariance function of the random field (see Assumption 3.10) and define a piecewise constant estimator to compute the empirical estimation at unobserved lags. This assumption allows us to obtain a bound on the estimation error at all lags.

Parametric Estimation, Model Fitting. Once the experimental variogram is computed as in subsection 2.1.3, the next step is to fit a valid model to the experimental semi-variogram values for all lags. To achieve accurate interpolation, the values of the semi-variogram must be known for all distances, including unobserved lags.

Furthermore, we desire the semi-variogram estimation to verify the conditionally negative definite property (as in Proposition 2.15). We must define a parametric family of functions that verify this condition. The set of valid semi-variogram models is defined as follows:

$$\Big\{\gamma(\cdot) = \gamma(\cdot;\theta),\, \theta \in \Theta\Big\}. \tag{2.4}$$

This family of models depends on a parameter $\theta \in \Theta \subset \mathbb{R}^q$ generally unknown. The focus of model fitting is on estimating the values of the parameter θ.

Some examples of isotropic semi-variogram models are given below. Figure 2.3 shows the graphical representation of these models. We refer the reader to Chiles and Delfiner (1999); Yaglom (1987); Armstrong (1998) for additional models.

Exponential model:

$$\gamma(h;\theta) = \begin{cases} 0, & h = 0 \\ c_0 + c_s\left(1 - \exp\left(-\frac{h}{\alpha}\right)\right), & h > 0, \end{cases}$$

where $\theta = (c_0, c_s, \alpha)$, with $c_0 \geq 0$, $c_s \geq 0$, and $\alpha > 0$.

Gaussian model:

$$\gamma(h;\theta) = \begin{cases} 0, & h = 0 \\ c_0 + c_s\left(1 - \exp\left(-\frac{h^2}{\alpha^2}\right)\right), & h > 0, \end{cases}$$

where $\theta = (c_0, c_s, \alpha)$, with $c_0 \geq 0$, $c_s \geq 0$, and $\alpha > 0$.

Spherical model: This model is valid in $\mathbb{R}^p$ only for $p \leq 3$.

$$\gamma(h;\theta) = \begin{cases} 0, & h = 0 \\ c_0 + c_s\left(\frac{3}{2}\frac{h}{\alpha} - \frac{1}{2}\frac{h^3}{\alpha^3}\right), & 0 < h \leq \alpha \\ c_0 + c_s, & h > \alpha, \end{cases}$$

where $\theta = (c_0, c_s, \alpha)$, with $c_0 \geq 0$, $c_s \geq 0$, and $\alpha > 0$.

Cubic model:

$$\gamma(h;\theta) = \begin{cases} 0, & h = 0 \\ c_0 + c_s\left(7\frac{h^2}{\alpha^2} - \frac{35}{4}\frac{h^3}{\alpha^3} + \frac{7}{2}\frac{h^5}{\alpha^5} - \frac{3}{4}\frac{h^7}{\alpha^7}\right), & 0 < h \leq \alpha \\ c_0 + c_s, & h > \alpha, \end{cases}$$

where $\theta = (c_0, c_s, \alpha)$, with $c_0 \geq 0$, $c_s \geq 0$ and $\alpha > 0$.

As observed in Figure 2.3, the exponential and Gaussian semi-variograms reach their sill only asymptotically (when $h \to \infty$), while the spherical and cubic models reach their sill at $h = \alpha$. Simulated Gaussian random fields using these different semi-variogram models are depicted in Figure 2.4.

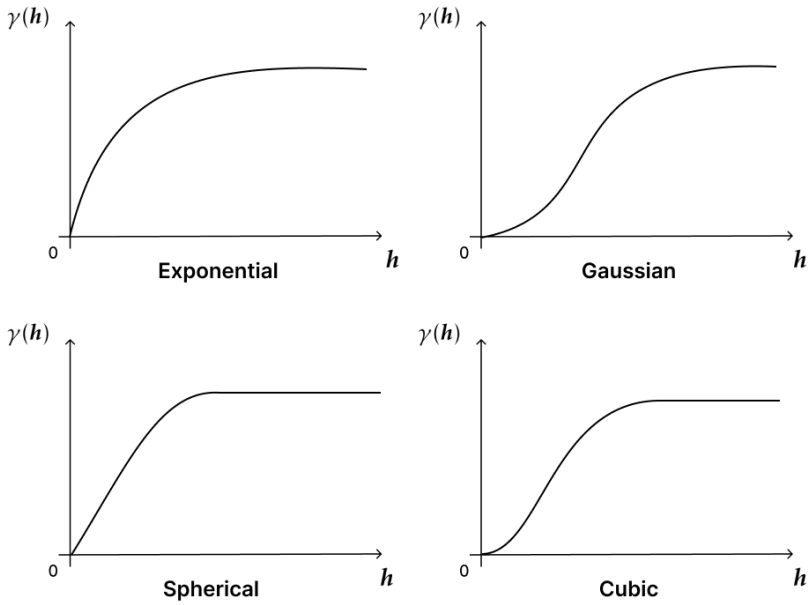


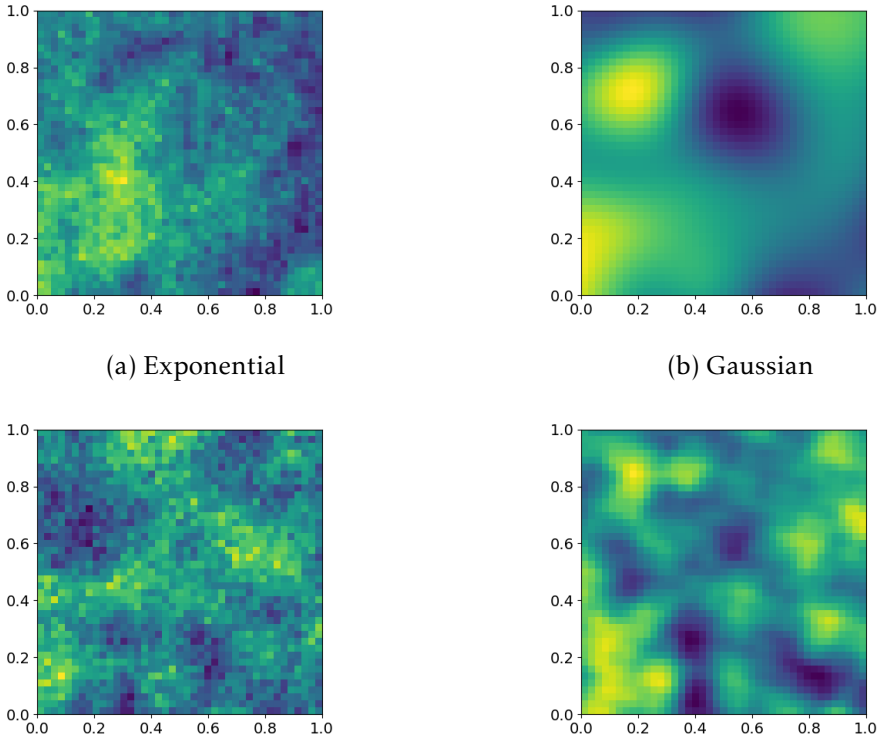
Figure 2.3: Semi-variogram models: exponential (top left), Gaussian (top right), spherical (bottom left), and cubic (bottom right), where $c_0 = 0$ (no nugget effect).

Matheron (1962) stressed the fact that modeling is a matter of choice. The parametric estimation consists of choosing a valid model. This choice must be based on the empirical knowledge given by the observations and on the shape of the experimental semi-variogram. We present two estimation methods for the parameter θ of a chosen semi-variogram model: the least squares and the maximum likelihood.

Least Squares. The ordinary least squares estimator is given by

$$\widehat{\theta} = \arg\min_{\theta \in \Theta} \sum_{k=1}^{K} \left(\widehat{\gamma}(h_k) - \gamma(h_k;\theta)\right)^2,$$

where K is the number of classes of distances defined during the experimental semi-variogram modeling (see subsection 2.1.3 and Figure 2.1b).



(a) Exponential

(b) Gaussian

(c) Spherical

(d) Cubic

Figure 2.4: Simulated Gaussian random fields with different semi-variogram models: exponential (top left), Gaussian (top right), spherical (bottom left), and cubic (bottom right).

However, ordinary least squares estimation is generally not accurate since it does not take into account the number of terms $N(h)$ for each distance. Notice that this value may largely vary from one class of distances to another. The method of weighted least squares overcomes this issue by incorporating the value $N(h)$ for each class:

$$\widehat{\theta} = \arg\min_{\theta\in\Theta} \sum_{k=1}^{K} \frac{N(h_k)}{\gamma^2(h_k;\theta)} \left(\widehat{\gamma}(h_k) - \gamma(h_k;\theta)\right)^2.$$

We refer the reader to Gaetan and Guyon, 2009, Chapter 5 for consistent and asymptotic results of the least squares estimators.

Maximum Likelihood. We assume that the vector of observations $\mathbf{X}(s_n) = \left\{\mathbf{X}_{s_1}, \cdots, \mathbf{X}_{s_n}\right\}$ is a Gaussian vector with mean μ and covariance matrix $\Sigma(\theta) = Var(\mathbf{X})$. Let $\gamma(h;\theta)$ the semi-variogram function with parameter θ. Then, the negative loglikelihood is given by

$$l(\mu,\theta) = -\frac{1}{2}\left(\log\left(Det\left(\Sigma(\theta)\right)\right) + (\mathbf{X}-\mu)^\top \Sigma(\theta)^{-1}(\mathbf{X}-\mu)\right).$$

We denote $\widehat{\mu}$ and $\widehat{\theta}$ (called the maximum likelihood estimators), the quantities that satisfy: $l(\widehat{\mu},\widehat{\theta}) = \arg\min\left\{l(\mu,\theta) : \mu \in \mathbb{R}^n, \theta \in \Theta\right\}$. Thus, the semi-variogram estimator by maximum likelihood is the resulting semi-variogram function $\gamma(h;\widehat{\theta})$.

For asymptotic properties of the maximum likelihood estimator, we refer to Gaetan and Guyon, 2009, Chapter 5 and Cressie, 1993, Chapter 7. See also Chiles and Delfiner, 1999, Section 2.6 for further details.

Once the parameter θ is estimated, one needs to validate this estimation. The two most widely used methods are the (leave-one-out) cross-validation and the bootstrap. The idea of the cross-validation method is to iteratively remove each observation and predict its value using the remaining observations via Kriging. Then, one may choose to validate the model by analyzing the prediction error distributions (Cressie, 1993). Bootstrap methods, by resampling the data and repeatedly estimating the semi-variogram parameters, provide a robust way to validate the parametric estimation (Cressie, 1993; Pardo-Igúzquiza and Olea, 2012).

Parametric estimation offers several advantages. By fitting a continuous function to the data, it provides a smooth representation of spatial dependence. Furthermore, parametric estimation provides interpretable parameters that allow valuable insights into the spatial behavior of the studied phenomenon. For instance, in a spherical covariance model, the range parameter α represents the distance beyond which spatial correlation becomes negligible (see Figure 2.3). Finally, as already mentioned, model fitting gives semi-variogram estimators that are valid models (*i.e.* that satisfy the conditionally negative definite property).

However, parametric models risk misspecifying the true dependence structure if the selected model does not accurately reflect the underlying process. On the contrary, non-parametric methods avoid this risk by not assuming a predefined model.

Semi-parametric. Semi-parametric methods may offer great flexibility by combining parametric and non-parametric components. These methods do not fully specify the semi-variogram model, allowing for a more accurate representation of the spatial dependence structure. In Im et al. (2007) a semi-parametric estimation of the spectral density of a Gaussian random field is given. Desassis et al. (2015) proposed a semi-parametric method based on pairwise likelihood estimation for the empirical semi-variogram function, under the Gaussianity assumption of the random field. The approach is the following: the variogram at a given lag is viewed as a parameter of the model; then, the pairs of data locations separated (approximately) by this lag are grouped, following the same idea as for the non-parametric Matheron (1962) estimator; and thus, the pairwise likelihood is maximized on these groups.

2.1.5 Simulation of Gaussian Processes

Simulation methods are used to generate multiple realizations of spatial variables. Undoubtedly, it is impossible to recover the true underlying spatial process. Therefore, the goal of a simulation is to construct a representation of the phenomenon while maintaining the spatial structure of the observed data.

One may be interested in simulations for several reasons: quantifying prediction uncertainty, evaluating risks associated with spatial decisions, and finding the optimal sampling design.

The classical simulation methods are the following. See Chiles and Delfiner, 1999, Chapter 7, for further details about (conditional) simulation methods.

Sequential Gaussian Simulation. Sequential Gaussian is one of the most widely used methods for simulating continuous Gaussian random fields. It generates realizations by sequentially drawing values from a multivariate normal distribution conditioned on previously simulated values and observed data. This sequential approach based on previous simulations ensures spatial coherence. However, the Gaussian assumption of the data may be restrictive.

Turning Bands Method. The turning bands method simulates an isotropic random field on $\mathbb{R}^p$, based on a stationary process in $\mathbb{R}$. The general idea is to first generate one-dimensional simulations independently and then combine them along multiple directions (called bands). This method is efficient for large-scale simulations.

Spectral Methods. As previously announced in subsection 2.1.2, the notion of the spectral representation can be used to develop simulation methods. It's the case of spectral methods, that use the Fourier transform to generate random fields in the frequency domain, and then transform them back to the spatial domain by means of inverse Fourier transform. Spectral methods are efficient for large grids.

In our study

Spectral Methods: The experiments in Sections 3.4 and 4.4 are implemented using the `gstools` library. This simulation method belongs to the family of the spectral methods (Müller and Schüler, 2020).

2.1.6 Kriging Interpolation Method

In this subsection, we present the Kriging method, a principal problem in Geostatistics. The goal here is interpolation, which aims at predicting the value of a random process at an unobserved spatial location, based on one single realization of the phenomenon observed at a finite number of sites. We discuss the optimality of the Kriging method when the dependence structure of the random field is known. Finally, we discuss previous results on the robustness of the Kriging method and the importance of selecting an efficient procedure for estimating the dependence structure, highlighting the effects of the uncertainty arising from this estimation on the Kriging prediction weights.

Let $\mathcal{S} \subset \mathbb{R}^p$ be a spatial set and $\mathbf{X}$ be a second-order random field on $\mathcal{S}$ with $\mathbb{R}$ as state space, *i.e.* a collection $\mathbf{X} = \{\mathbf{X}_s : s \in \mathcal{S}\}$ of real-valued square-integrable r.v.'s defined on the same probability space, $(\Omega, \mathcal{F}, \mathbb{P})$ say, indexed by $s \in \mathcal{S}$. We denote by $\mu : s \in S \mapsto \mathbb{E}[\mathbf{X}_s]$ its mean and by $C : (s,t) \in \mathcal{S}^2 \mapsto C(s,t) = Cov(\mathbf{X}_s, \mathbf{X}_t)$ its covariance functions. Assume that the spatial process is observed at a finite number of points $s_1, \ldots, s_d$ in the spatial set $\mathcal{S}$.

As formulated in the seminal contribution of Matheron (1962), the *Kriging* method can be described as follows. Given a set of observations, the goal pursued is to build a predictor $\widehat{\mathbf{X}}_s$ of $\mathbf{X}_s$ at a given unobserved site $s \in \mathcal{S}$. The accuracy of the prediction can be measured by the Mean Squared Error (MSE)

$$L\left(s, \widehat{\mathbf{X}}_s\right) = \mathbb{E}_{\mathbf{X}}\left[\left(\widehat{\mathbf{X}}_s - \mathbf{X}_s\right)^2\right]. \tag{2.5}$$

The Kriging method assumes that the predictor $\widehat{\mathbf{X}}_s$ is of the form of a *linear* combination of the $\mathbf{X}_{s_i}$'s

$$\widehat{\mathbf{X}}_{s,\Lambda_d(s)} = \lambda_1(s)\mathbf{X}_{s_1} + \ldots + \lambda_d(s)\mathbf{X}_{s_d}, \tag{2.6}$$

where $\Lambda_d(s) = (\lambda_1(s), \ldots, \lambda_d(s)) \in \mathbb{R}^d$ is the vector of weights to be determined.

The goal is to find the vector of weights $\Lambda_d^*(s)$ that minimizes the MSE, such that the predictor is unbiased.

We present two interpolation methods: Simple Kriging and Ordinary Kriging. Notice that other methods have been proposed for random fields: we refer the reader to Chiles and Delfiner, 1999, Section 3, and Gaetan and Guyon, 2009, Section 1, for a presentation of the Universal Kriging approach, which is useful in the presence of drift, and to Cressie, 1993, Section 3, for the Cokriging method, which enables multivariate interpolation.

Simple Kriging. In Kriging in its simplest form, the mean $\mu(\cdot)$ is supposed to be known. Rather than recentering it, it is assumed that the random field $\mathbf{X}$ is centered: $\mu := 0$.

The optimal predictor of this form regarding the expected prediction error can be deduced from a basic variance computation, it is described below.

Lemma 2.20. *For $d \geq 1$, let $\mathbf{s}_d = (s_1, \ldots, s_d)$, $\mathbf{X}(\mathbf{s}_d) = (\mathbf{X}_{s_1}, \ldots, \mathbf{X}_{s_d})$, $\Sigma(\mathbf{s}_d) = Var(\mathbf{X}(\mathbf{s}_d))$ and define $\mathbf{c}_d(s) = (Cov(\mathbf{X}_s, \mathbf{X}_{s_1}), \ldots, Cov(\mathbf{X}_s, \mathbf{X}_{s_d}))$. Suppose that the matrix $\Sigma(\mathbf{s}_d)$ is positive definite. Then, the solution of the minimization problem*

$$\min_{\Lambda_d(s)\in\mathbb{R}^d} L\left(s, \widehat{\mathbf{X}}_{s,\Lambda_d(s)}\right)$$

is unique and given by

$$\Lambda_d^*(s) = \Sigma(\mathbf{s}_d)^{-1}\mathbf{c}_d(s). \tag{2.7}$$

In addition, the minimum is equal to

$$L\left(s, \widehat{\mathbf{X}}_{s,\Lambda_d^*(s)}\right) = Var(\mathbf{X}_s) - \mathbf{c}_d(s)^\top \Sigma(\mathbf{s}_d)^{-1}\mathbf{c}_d(s).$$

Proof. The optimality derives from the basic properties of orthogonal projection in the L_2 space and the closed analytical form for the minimizer of

$$\begin{aligned} L\left(s, \widehat{\mathbf{X}}_{s,\Lambda_d(s)}\right) &= Var\left(\Lambda_d(s)^\top \mathbf{X}(\mathbf{s}_d) - \mathbf{X}_s\right) \\ &= Var(\mathbf{X}_s) + \Lambda_d(s)^\top \Sigma(\mathbf{s}_d)\Lambda_d(s) - 2\mathbf{c}_d(s)^\top \Lambda_d(s), \end{aligned} \tag{2.8}$$

is obtained by solving a linear system: $\Lambda_d^*(s) = \Sigma(\mathbf{s}_d)^{-1}\mathbf{c}_d(s)$. The minimal mean squared error is then:

$$\begin{aligned} L\left(s, \widehat{\mathbf{X}}_{s,\Lambda_d^*(s)}\right) = Var(\mathbf{X}_s) &+ \mathbf{c}_d(s)^\top \Sigma(\mathbf{s}_d)^{-1}\Sigma(\mathbf{s}_d)\left(\Sigma(\mathbf{s}_d)^{-1}\right)^\top \mathbf{c}_d(s) \\ &- 2\mathbf{c}_d(s)^\top \Sigma(\mathbf{s}_d)^{-1}\mathbf{c}_d(s) = Var(\mathbf{X}_s) - \mathbf{c}_d(s)^\top \Sigma(\mathbf{s}_d)^{-1}\mathbf{c}_d(s). \end{aligned}$$

■

Remark 2.21. (Gaussian random fields) *Notice that, when the random field* $\mathbf{X}$ *is Gaussian, we have:* $\widehat{\mathbf{X}}_{s,\Lambda_d^*(s)} = \mathbb{E}[\mathbf{X}_s \mid \mathbf{X}(\mathbf{s}_d)]$. *Hence* $\widehat{\mathbf{X}}_{s,\Lambda_d^*(s)}$ *is the minimizer of the quadratic error* $L\left(s, \widehat{\mathbf{X}}_s\right)$ *over the set of all predictors* $\widehat{\mathbf{X}}_s$.

Remark 2.22. (Exact interpolation) *The solution of the simple Kriging problem is an exact interpolator, in the sense that* $\widehat{\mathbf{X}}_{s_i,\Lambda_d^*(s_i)} = \mathbf{X}_{s_i}$ *for all* $i \in \{1, \ldots, d\}$ *(Cressie, 1993, p. 129).*

Ordinary Kriging. In most practical situations, the mean is unknown. Thus, simple Kriging cannot be applied. An obvious approach is to estimate the mean and subtract it from the data (in order to go back to the zero mean case). In this case, however, the estimated residuals are not the same as the true residuals and the statistical procedure properties are difficult to analyze. An optimal solution is Ordinary Kriging, which does not require any knowledge of the mean.

Chiles and Delfiner (1999) proposed a version of Ordinary Kriging using the semi-variogram. The Kriging system to estimate the random process at a location s in a matrix version (Gratton, 2002) is:

$$\Lambda_{OK}^*(s) = \Gamma(\mathbf{s}_d)^{-1}\boldsymbol{b}_d(s), \tag{2.9}$$

where

$$\Gamma(\mathbf{s}_d) = \begin{pmatrix} \gamma(h_{1,1}) & \cdots & \gamma(h_{1,d}) & 1 \\ \vdots & & \vdots & \vdots \\ \gamma(h_{d,1}) & \cdots & \gamma(h_{d,d}) & 1 \\ 1 & \cdots & 1 & 0 \end{pmatrix}, \qquad \boldsymbol{b}_d(s) = \begin{pmatrix} \gamma(h_{1,0}) \\ \vdots \\ \gamma(h_{d,0}) \\ 1 \end{pmatrix}, \qquad \lambda = \begin{pmatrix} \lambda_1(s) \\ \vdots \\ \lambda_d(s) \\ \beta \end{pmatrix},$$

with $h_{i,j} = \|s_i - s_j\|$ being the distance between two observed locations $(s_i, s_j) \in \mathcal{S}^2$ and $h_{i,0} = \|s_i - s\|$ is the distance between the point of interest s and each observed location $(s_i)_{1 \le i \le d}$.

The parameter β is introduced in order to obtain an unbiased estimator: it induces that the weights $\lambda_1(s), \cdots, \lambda_d(s)$ sum up to 1.

Notice that Remarks 2.21 and 2.22 also apply to the Ordinary Kriging approach.

Optimality / BLUP. When the covariance function (respectively the semi-variogram function) is known, the simple Kriging predictor (resp. Ordinary Kriging predictor) is optimal. In the literature, the Kriging predictor is said to be BLUP, for *Best Linear Unbiased Predictor*. In the following chapters, we refer to this predictor as the *theoretical Kriging predictor*. For further details, we refer the reader to (Stein, 1999, Section 1) and (Cressie, 1993, Section 3).

In practice, the covariance (or semi-variogram) function is unknown and needs to be estimated. See subsection 2.1.4 for a discussion on the estimation of the dependence structure. Once the covariance (respectively the semi-variogram) estimation is obtained one can construct the empirical counterpart of $\Sigma(\mathbf{s}_d)$ and $\mathbf{c}_d(s)$ (resp. $\Gamma(\mathbf{s}_d)$ and $\boldsymbol{b}_d(s)$). Finally, replacing $\Sigma(\mathbf{s}_d)$ and $\mathbf{c}_d(s)$ in Equation (2.7) (resp. $\Gamma(\mathbf{s}_d)$ and $\boldsymbol{b}_d(s)$ in Equation (2.9)) by their estimators, an empirical version of the Kriging predictor can

be obtained employing the *plug-in* rule. In the following chapters, we refer to this predictor as the *empirical Kriging predictor*.

Both Kriging methods depend on the dependence structure of the data: the simple Kriging's optimal vector of weights depends on the covariance function, while the Ordinary Kriging's weights depend on the semi-variogram function. Furthermore, both linear predictors depend also on the point location of interest s. This implies that the Kriging method takes advantage of the spatial dependence structure of the data: point locations that are close to some observations will get more weight in the prediction function, and the weights are determined by the degree of similarity between the data points.

Uncertainty effects of the dependence structure estimation on Kriging prediction. The semi-variogram plays a crucial role in Kriging, as the prediction weights are determined based on the semi-variogram values. Thus, it is important to be able to quantify the variogram estimation uncertainty to ensure that the estimator is sufficiently accurate for Kriging prediction. As in Cressie and Zimmerman (1992), we say that a method is stable when model misspecification or parameter estimation has little effect on its accuracy. Cressie and Zimmerman (1992) distinguish two broad categories of approaches: the mathematical stability approach and the statistical stability method. For the first approach, one may consider the effects on the Kriging prediction weights due to the use of a perturbed semi-variogram instead of the true function. The second approach investigates the sensitivity of the results to estimates of the semi-variogram's parameter θ.

Matheron (1965) warned about the variability of the semi-variogram estimator when computed for large distances h. Stein (1999) and Putter and Young (2001) supported this idea and stressed the necessity to take into account the effects of the uncertainty about the dependence structure estimation for classical Geostatistics methods. However, ignoring these effects may lead to an inappropriate model choice, as it may induce a bias for the Kriging predictor, and thus result in larger prediction errors (Todini, 2001).

Various approaches have been proposed to quantify the effects stemming from uncertainty in the estimation. For example, Diamond and Armstrong (1984) and Armstrong (1984) introduced the definition of neighborhoods of semi-variograms: a semi-variogram g lies in the σ-neighborhood (with $\sigma \in]0,1[$) of a semi-variogram γ, if for all distances $h > 0$, $1 - \sigma < \frac{g(h)}{\gamma(h)} < 1 + \sigma$. The σ-neighborhood is then noted $N_\sigma(\gamma)$. The idea is that similar-looking variograms may lead to similar results for Kriging interpolation. They identified two categories of causes that imply a change in the Kriging weights between two semi-variogram models that belong to the same σ-neighborhood: changes due to a slight perturbation in the semi-variogram estimation, and changes that result from a different sampling of the observations locations. Bardossy (1988) discussed two drawbacks of this approach. First, these bounds are limited to semi-variogram models belonging to the same σ-neighborhood. Second, the bounds depend mostly on the observations locations and on the semi-variogram, neglecting the information carried by the spatial position of the point of interest on which Kriging is applied. Other studies focused on the estimation variance. Brooker (1986) studied the effect of parameter changes on estimation variance for spherical semi-variogram, later extended by Bardossy (1988) for a modified nugget effect.

Bardossy (1988) warned about the lack of results for the implication of the sampling configuration in these effects. To ensure robust Kriging predictions, one must select an optimal sampling setting. Indeed, the empirical estimation of the semi-variogram is unavoidably affected by the spatial configuration of the observations. The issue of selecting an optimal sampling recovers: *(i)* the number of observations needed, and *(ii)* their distribution over the spatial domain (see subsection 2.1.1). Russo (1984) developed a method to define an optimal design, based on a criterion that tries to minimize the repartition of lags within each class of the empirical semi-variogram. McBratney and Webster (1981) and Yfantis et al. (1987) compared three sampling schemes: the triangular, the rectangular, and the hexagonal grids. They found that the equilateral triangular design gives more reliable estimate of the semi-variogram. However, the rectangular design usually gives similar results, and is often preferred in real-data applications since it fits real-world situations well. Wang et al. (2020) showed that an optimal sampling design must minimize the fill distance, which is the maximum distance from any site in the domain to the nearest sample point location. It measures how well the set of observations covers the spatial domain: the more the fill distance is small, the more the sample points are well-distributed and provide better coverage. Wang et al. (2020) also discussed the link between prediction robustness and smoothness of the covariance function: a less smooth function is more robust.

In some studies, the accuracy of the dependence structure estimation was discussed when independent copies of the random process were available. The focus is on defining the optimal number of independent copies, referred to as the sample size. Adamczak et al. (2010) and Rudelson (1999) proposed an optimal sample size for specific distributions, such as sub-exponential or sub-Gaussian distributions. Vershynin (2012) extended the study for all distributions with finite fourth moment. Loukas (2017) gave non-asymptotic bounds for the concentration of inner products involving eigenvectors of the estimated and the true covariance matrices. The result is that few independent copies can be sufficient. Marchant and Lark (2004) gave a comparison between uncertainty effects when the semi-variogram is computed over one single simulation and when it is averaged over several simulations. The uncertainty effects have different origins depending on the setting. For the one single simulation case, the errors may be induced by the sampling design only. When several simulations are available, errors may appear due to the variation of the random field over each realization.

These last approaches violate our main assumption, since they assume independent copies of the random process. In Chapters 3 and 4, the focus is on the accuracy of the semi-variogram estimation and of the Kriging predictor when the phenomenon is assumed to be observed from one single realization. This setting is closer to real-world situations but involves deeper technical difficulties, which are discussed in the following chapters.

2.2 Statistical Learning Theory

Statistical Learning Theory provides the mathematical and theoretical foundations for understanding and developing Machine Learning algorithms. The main goal is to build models that can make accurate predictions based on data, and identify underlying patterns and structures within the data.

The different kinds of learning are:

- Supervised Learning: The goal is to train a function to associate inputs with outputs by leveraging observations of pairs of inputs and their corresponding labels.
- Unsupervised Learning: In contrast to Supervised Learning, the data lack labels. The goal in this context is to deduce patterns or structures in the data based on the observations (only inputs).
- Semi-supervised Learning: Combining both previous learning methods, the idea is to improve the learning accuracy using both labeled and unlabeled data.
- Reinforcement Learning: The learning evolves through feedbacks (in the form of rewards or penalties) from an environment.

We restrict ourselves to the Supervised Learning setting. In Supervised Learning, the output nature can vary: an output can be either a quantitative value or a qualitative value (also called a categorical variable). For each type of output, there is the corresponding prediction method: for a quantitative output, we apply a regression method, while for a categorical variable, the prediction method is called classification. Both methods share a similar setup, which is the following. Let $\mathbf{Z} = (Z_1, \cdots, Z_d)$ a random $d-$dimensional vector (with $\mathbb{R}^d$ as input space) and Y a square integrable random output (the label) that takes its values in an output space $\mathcal{Y}$. Let $D_N = (\mathbf{Z}_i, Y_i)_{1 \leq i \leq N}$ be a training sample, consisting of independent copies of the random pair $(\mathbf{Z}, Y)$, distributed according to a distribution P. The goal here is to estimate a function $g : \mathbb{R}^d \to \mathcal{Y}$ (a predictor) which predicts Y from $\mathbf{Z}$. The two settings will differ from the definition of the output space $\mathcal{Y}$.

This section is structured as follows. In subsection 2.2.1, the principle of Empirical Risk Minimization (ERM) is discussed. In the next two subsections, we present the two prediction methods of classification (subsection 2.2.2) and regression (subsection 2.2.3). Finally, subsection 2.2.4 provides a short summary about the main concentration inequality results in Statistical Learning Theory.

The results presented in this section are drawn from Lugosi (2002) and (Boucheron et al., 2005). For further details on Statistical Learning Theory, we refer the reader to Bousquet et al. (2003), and to Kanevski et al. (2009) for an overview of Machine Learning for spatial data, with a comparison between the Geostatistical and the Machine Learning point of views.

2.2.1 Empirical Risk Minimization

Under the setting stated above, the objective of this subsection is to evaluate the accuracy of the predictor g. To this purpose, we first define a loss function $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$, whose definition depends on the nature of the problem under consideration. The loss function $\ell(y, y')$, given two variables y and y', gives information about their similarity: the more the loss value is small, the more the two variables have close values. Since our focus is on the accuracy of the prediction of a label Y_{N+1} for an unobserved input $\mathbf{Z}_{N+1}$ by the classifier F, the quantity of interest is $\ell\Big(g(\mathbf{Z}_{N+1}), Y_{N+1}\Big)$: the more this quantity is small, the more the prediction is good. We define the *risk* of a predictor g as:

$$\mathcal{R}(F) = \mathbb{E}\Big[\ell\Big(g(\mathbf{Z}), Y\Big)\Big]. \tag{2.10}$$

Thus, a good predictor is such that it minimizes the risk over all measurable functions. The problem becomes

$$\inf_g \mathcal{R}(g). \tag{2.11}$$

When the distibution P is known, we define the *Bayes risk* $\mathcal{R}^* = \inf_g \mathcal{R}(g)$ as the smallest possible risk value, and the *Bayes rule* $g^* = \arg\min_g \mathcal{R}(g)$ as the optimal predictive function. However, the distribution P is unknown in practice. The Empirical Risk Minimization principle consists in replacing the unknown risk (2.10) in the optimization problem (2.11) by a statistical version computed from the training sample available:

$$\widehat{\mathcal{R}}_N(g) = \frac{1}{N}\sum_{i=1}^{N} \ell\Big(g(\mathbf{Z}_i), Y_i\Big), \tag{2.12}$$

and restricting minimization to a class $\mathcal{G}$ of predictors of controlled complexity, see *e.g.* Györfi et al. (2002). The quantity $\widehat{\mathcal{R}}_N(g)$ is called the *empirical error* of the classifier g, and depicts the mean prediction error of g over all the training sample D_N.

In the two following subsections, we present the particular cases of classification and regression for ERM.

2.2.2 Classification Problem

Classification aims at defining a rule that assigns a class for an observation. For simplicity, we restrict ourselves to binary classification, *i.e.* $\mathcal{Y} = \{0, 1\}$. A mapping $g : \mathbb{R}^d \to \{0, 1\}$ is called a classifier.

We introduce the $0-1$ loss function, given by $\ell(y, y') = \mathbb{I}\{y \neq y'\}$. Thus, the risk of a classifier g is defined as the probability of error of g

$$\mathcal{R}(g) = \mathbb{P}(g(\mathbf{Z}) \neq Y).$$

Define $\eta(\mathbf{Z}) = \mathbb{P}(Y = 1|\mathbf{Z}) = \mathbb{E}[Y|\mathbf{Z}]$. Then, the Bayes classifier (*i.e.* the Bayes rule function for classification) is

$$g^*(\mathbf{Z}) = \begin{cases} 1 & \text{if } \eta(\mathbf{Z}) > 1/2, \\ 0 & \text{otherwise.} \end{cases}$$

Indeed, it minimizes the risk and we have $\mathbb{P}\Big(g^*(\mathbf{Z}) \neq Y\Big) \leq \mathbb{P}\Big(g(\mathbf{Z}) \neq Y\Big)$ (Lugosi, 2002, Theorem 1.1).

The empirical error for the classification problem is

$$\widehat{\mathcal{R}}_N(F) = \frac{1}{N}\sum_{i=1}^{N} \mathbb{I}\{g(\mathbf{Z}_i) \neq Y_i\}.$$

For further details on classification, we refer the reader to Devroye et al. (1996); Boucheron et al. (2005).

2.2.3 Regression Problem

In the basic and usual regression setup, the label is assumed to be a continuous output, *i.e.* $\mathcal{Y} = \mathbb{R}$. One of the most widely used loss functions in the simple regression setting is the quadratic loss $\ell(y, y') = (y - y')^2$.

Based on a training sample $\{(\mathbf{Z}_1, Y_1), \ldots, (\mathbf{Z}_N, Y_N)\}$ composed of independent copies of the random pair $(\mathbf{Z}, Y)$, the goal pursued is to build a predictor g that minimizes the MSE

$$\mathcal{R}(g) = \mathbb{E}\left[\left(Y - g(\mathbf{Z})\right)^2\right] \tag{2.13}$$

over the ensemble of all possible measurable mappings $g : \mathbb{R}^d \to \mathbb{R}$ that are square integrable with respect to the distribution of $\mathbf{Z}$. The Bayes rule is given by $g^*(\mathbf{Z}) = \mathbb{E}[Y \mid \mathbf{Z}]$. Based on the Empirical Risk Minimization principle, the empirical error is

$$\widehat{\mathcal{R}}_N(g) = \frac{1}{N}\sum_{i=1}^{N}\left(Y_i - g(\mathbf{Z}_i)\right)^2.$$

Under the assumption that the random variables Y and $\{g(\mathbf{Z}) : \ g \in \mathcal{G}\}$ have sub-Gaussian tails, the order of magnitude of the fluctuations of the maximal deviations $\sup_{g\in\mathcal{G}} |\widehat{L}_N(g) - L(g)|$ can be estimated and generalization bounds for the MSE of empirical risk minimizers can be established (Lecué and Mendelson, 2013).

Linear Ridge Regression. In linear regression, the class considered is that composed of all linear functionals on $\mathbb{R}^d$, namely $\mathcal{G} = \{\langle \zeta, \cdot \rangle : \ \zeta \in \mathbb{R}^d\}$, where $\langle \cdot, \cdot \rangle$ denotes the usual inner product on $\mathbb{R}^d$. In Linear Ridge Regression (LRR), one thus minimizes the empirical error plus a quadratic penalty term to avoid overfitting

$$\sum_{i=1}^{N}\left(Y_i - \ \zeta^\top \mathbf{Z}_i\right)^2 + \alpha\|\zeta\|^2,$$

where $\alpha \geq 0$ is a tuning parameter that rules the trade-off between complexity penalization and goodness-of-fit (generally selected via cross-validation in practice). It yields

$$\widehat{\zeta}_\alpha = \ \mathbf{Y}_N^\top \mathcal{Z}_N \left(\alpha \mathbf{I}_d + \mathcal{Z}_N \ \mathcal{Z}_N^\top\right)^{-1}, \tag{2.14}$$

where $\mathbf{I}_d$ is the $d \times d$ identity matrix, $\mathbf{Y}_N = (Y_1, \ldots, \ Y_N)$ and $\mathcal{Z}_N$ is the $d \times N$ matrix with $\mathbf{Z}_1, \ldots, \mathbf{Z}_N$ as column vectors, as well as the predictive mapping $\widehat{g}_N(z) = \ \widehat{\zeta}_\alpha^\top z$, $z \in \mathbb{R}^d$. The regularization term ensures that the matrix inversion involved in (2.14) is always well-defined.

Kernel Ridge Regression. However, LRR is limited to linear problems. To overcome this boundary, we introduce the kernel trick. The idea is to transform the data space into a space of larger dimension, namely the Reproducing Kernel Hilbert Space (RKHS), thanks to the use of a kernel function.

In Kernel Ridge Regression (KRR), one applies LRR in the feature space, *i.e.* to the data $(Y_1, \Phi(\mathbf{Z}_1)), \ldots, (Y_N, \Phi(\mathbf{Z}_N))$, where $\Phi : \mathbb{R}^d \to \mathcal{H}$ is a *feature map* taking its values in a RKHS $(\mathcal{H}, \langle \cdot, \cdot \rangle_{\mathcal{H}})$ associated to a (positive definite) kernel K such that $\forall (z, z') \in \mathbb{R}^d \times \mathbb{R}^d$, $K(z, z') = \langle \Phi(z), \Phi(z') \rangle_{\mathcal{H}}$. By means of the kernel trick, the predictive mapping, linear in the feature space, can be written as

$$\widetilde{g}_N(z) = \ \mathbf{Y}_N^\top (\alpha \mathbf{I}_N + \mathcal{K}_N)^{-1} \kappa_N(z), \tag{2.15}$$

where $\mathbf{I}_N$ is the $N \times N$ identity matrix, $\mathcal{K}_N$ is the Gram matrix with entries $K(\mathbf{Z}_i, \mathbf{Z}_j) = \langle \Phi(\mathbf{Z}_i), \Phi(\mathbf{Z}_j) \rangle_{\mathcal{H}}$ for $1 \leq i,\ j \leq N$ and $\kappa_N(z) = (K(\mathbf{Z}_1, z),\ \ldots,\ K(\mathbf{Z}_N, z)) \in \mathbb{R}^N$. One may refer to Steinwart and Christmann, 2008, Chapter 9, for more details on support vector machines for regression, and to Györfi et al. (2002) for additional details on regression.

2.2.4 Concentration Inequalities – Theoretical Guarantees

Excess of Risk. Recall the general empirical error $\widehat{\mathcal{R}}_N(g)$ definition in Equation (2.12) (for both classification and regression problems). Denote g_N^* the classifier such that $g_N^* \in \arg\min_{g \in \mathcal{G}} \widehat{\mathcal{R}}_N(g)$. The *excess of risk* of g_N^* is given by $\mathcal{R}(g_N^*) - \mathcal{R}^*$. We can decompose the excess of risk into two errors:

$$\mathcal{R}(g_N^*) - \mathcal{R}^* = \underbrace{\inf_{g \in \mathcal{G}} \mathcal{R}(g) - \mathcal{R}^*}_{\text{approximation error}} + \underbrace{\mathcal{R}(g_N^*) - \inf_{g \in \mathcal{G}} \mathcal{R}(g)}_{\text{estimation error}}.$$

The estimation error can be bounded by the following inequality (Lugosi, 2002, Lemma 1.1)

$$\mathcal{R}(g_N^*) - \inf_{g \in \mathcal{G}} \mathcal{R}(g) \leq 2 \sup_{g \in \mathcal{G}} \left| \widehat{\mathcal{R}}_N(g) - \mathcal{R}(g) \right|. \tag{2.16}$$

Basic Concentration Inequalities. Notice that $\widehat{\mathcal{R}}_N(g) - \mathcal{R}(g) = \frac{1}{N} \sum_{i=1}^{N} X_i - \mathbb{E}[X]$, with $X = \ell(g(\mathbf{Z}), Y)$. Concentration inequalities provide bounds on the probability of deviation of a random variable from some value (generally, its expected value). We give hereafter some of the most widely used results. We refer the reader to Devroye et al. (1996); Boucheron et al. (2013) for a complete review of concentration inequalities, together with the proofs of the following results.

Markov's Inequality. Let X be a non-negative random variable. Then, for any $t > 0$,

$$\mathbb{P}\Big(X \geq t\Big) \leq \frac{\mathbb{E}[X]}{t}.$$

Chebyshev's Inequality. Let X be an arbitrary random variable. Then, for any $t > 0$,

$$\mathbb{P}\Big(|X - \mathbb{E}[X]| \geq t\Big) \leq \frac{Var(X)}{t^2}.$$

Hoeffding's Inequality. Let $X_1, \cdots, X_N$ be independent bounded random variables, where X_i is in $[a_i, b_i]$ with probability 1 and let $S = \sum_{i=1}^{N} \Big(X_i - \mathbb{E}[X_i]\Big)$. Then, for any $t > 0$,

$$\mathbb{P}\Big(S \geq t\Big) \leq \exp\left(- \frac{2t^2}{\sum_{i=1}^{N} (b_i - a_i)^2} \right),$$

and

$$\mathbb{P}\Big(S \leq -t\Big) \leq \exp\left(- \frac{2t^2}{\sum_{i=1}^{N} (b_i - a_i)^2} \right).$$

Combining Hoeffding's inequality and the bound in Equation (2.16), the following bound on the estimation error of the ERM is obtained.

Theorem 2.23. *(Lugosi, 2002, Theorem 1.3) Let $\mathcal{F}$ be a class of predictors with cardinality bounded by M. Then, for any $t > 0$,*

$$\mathbb{P}\left(\sup_{g\in\mathcal{G}}\left|\widehat{\mathcal{R}}_N(g) - \mathcal{R}(g)\right| > t\right) \leq 2M\exp(-2Nt^2).$$

Concentration Inequalities for Sum of Independent Gamma Variables. We now present two concentration inequalities on the sum of independent Gamma variables, from Bercu et al. (2015) and Wang and Ma (2020), that we use in Chapter 3 to derive tail bounds for the empirical semi-variogram estimator. We refer the reader to the corresponding references for a proof of these results.

Let $X \sim \Gamma(\alpha, \beta)$ be a Gamma random variable with shape parameter $\alpha \in \mathbb{R}_+$ and rate parameter $\beta \in \mathbb{R}_+$, such that $\mathbb{E}[X] = \frac{\alpha}{\beta}$ (see subsection A.1 on Gamma r.v.'s).

Theorem 2.24. *(Bercu et al., 2015, Theorem 2.57) Let $X_1, \cdots, X_N$ be a finite sequence of independent r.v.'s such that, for all $1 \leq i \leq N$, $X_i \sim \Gamma(\alpha_i, \beta_i)$, with $\alpha_i, \beta_i \in \mathbb{R}_+$ and let $S_N = X_1 + \cdots + X_N$. Then, for any $t \in]0, 1[$,*

$$\mathbb{P}\Big(S_N \leq \mu - t\mu\Big) \leq \exp\left(-\frac{\mu^2}{\sigma^2}\frac{t^2}{2}\right),$$

where $\mu = \mathbb{E}[S_N]$ and $\sigma^2 = Var(S_N)$.

Theorem 2.25. *(Wang and Ma, 2020, Theorem 4.1) Let $X_1, \cdots, X_N$ be a finite sequence of independent r.v.'s such that, for all $1 \leq i \leq N$, $X_i \sim \Gamma(\alpha_i, \beta_i)$, with $\alpha_i, \beta_i \in \mathbb{R}_+$ and let $S_N = X_1 + \cdots + X_N$. Then, for any $t \geq 1$,*

$$\mathbb{P}\Big(S_N \geq t\mu\Big) \leq \exp\Big(-\beta_*\mu(t - 1 - \log(t))\Big),$$

where $\mu = \mathbb{E}[S_N]$ and $\beta_* = \min_i \beta_i$.

We propose the following result obtained thanks to a slight modification of Theorem 2.25:

Corollary 2.26. *Let $X_1, \cdots, X_N$ be a finite sequence of independent r.v.'s such that, for all $1 \leq i \leq N$, $X_i \sim \Gamma(\alpha_i, \beta_i)$, with $\alpha_i, \beta_i \in \mathbb{R}_+$ and let $S_N = \sum_{i=1}^{N}\Big(X_i - \mathbb{E}[X_i]\Big)$. Then, for any $t \geq 1$,*

$$\mathbb{P}\left(\frac{1}{N}S_N \geq t\mu\right) \leq \exp\Big(-\beta_*\mu(tN - \log(1 + tN))\Big),$$

where $\mu = \sum_{i=1}^{N}\mathbb{E}[X_i]$ and $\beta_* = \min_i \beta_i$.

In our study

Concentration Inequalities for the sum of Independent Gamma Variables: Assuming a Gaussian distribution, the empirical semi-variogram function can be regarded as the sum of independent χ^2 random variables (see Proposition 3.7 in Chapter 3), and thanks to the relationship between Gamma and χ^2 random variables (see Proposition .14 in Appendix A), as the sum of independent Gamma random variables. Thus, combining this with the above results, we obtain concentration inequalities for the empirical semi-variogram (see Section 3.2 for all the details).

2.3 Conclusion

The purpose of Kriging can be more ambitious than the construction of a pointwise prediction for the random field $\mathbf{X}$ at an unobserved site $s \in \mathcal{S}$. The goal pursued may consist of building a decision function $f : \mathcal{S} \times \mathbb{R}^d \to \mathbb{R}$ in order to predict $\mathbf{X}$ over all $\mathcal{S}$ based on the observation of the spatial process at a finite number of points $s_1, \ldots, s_d$ in the spatial set $\mathcal{S}$. Thus, one can obtain a complete map, which gives the predicted values of the random field at each site of the spatial domain. This point of view is the one adopted in Chapter 4.

When the dependence structure of the phenomenon is known, the Kriging predictor is optimal. But, in practice, there are no knowledge about this structure, which thus needs to be estimated. In the case of an unknown dependence structure, there are no theoretical guarantees of optimality.

The nature of currently available spatial datasets calls for flexible, non-parametric approaches to analyze spatial phenomenon. Nevertheless, the inherent dependence structure within the data prevents significant theoretical advancements in spatial interpolation using finite samples. Thus, the theory available in spatial statistics is mainly asymptotic (see *e.g.* Stein, 1999).

New (non-asymptotic) results must be developed, in order to establish generalization guarantees. The first part of this thesis aims at establishing non-asymptotic bounds that assess the generalization capacity of the empirical Kriging predictive map. This research could provide valuable insights and tools for geostatistical modeling. Deriving robust non-asymptotic bounds for the Kriging predictor when the covariance function is unknown could lead to more reliable and interpretable spatial predictions.

Key points of this Chapter.

Definitions and Assumptions used in this thesis

- **In-fill asymptotic:** the spatial domain becomes denser and denser as the number of observations grows ⟶ allows efficient semi-variogram estimation for all distances (subsection 2.1.1).
- **Second-order stationarity and isotropy:** ensure that the process is sufficiently homogeneous ⟶ allows accurate estimation of the dependence structure (definitions 2.7 and 2.9).
- **Ergodicity:** ensures efficient prediction of the random field from a single realization (definition 2.16).
- **Gaussian process:** the empirical semi-variogram estimator can be seen as the sum of independent χ^2 variables (definition 2.17 and Proposition 3.7 in Chapter 3).

Concepts and Methodologies used in this thesis

- **Non-parametric estimation:** Matheron (1962) empirical semi-variogram estimator ⟶ more flexible approach (subsection 2.1.4).
- **Simple Kriging:** interpolation method to predict the value of a random process at unobserved spatial location, taking into account the dependence structure of the data ⟶ predictor of the form of a linear combination of the (finite number of) observations, based on one unique realization of the phenomenon (subsection 2.1.6).
- **Non-asymptotic results:** tail bounds for the global excess risk of the Kriging method ⟶ lead to more reliable and interpretable spatial predictions.
- **Kernel ridge regression:** similarities between the Kriging and the KRR methods ⟶ viewing Kriging as a KRR problem (subsection 2.2.3).
- **Statistical theory:** concentration inequalities for the sum of independent Gamma variables ⟶ non-asymptotic bounds on the accuracy of the empirical semi-variogram estimation for Gaussian processes (subsection 2.2.4).

3

Estimation of the Spatial Dependence Structure

Contents

3.1 Introduction

In this chapter, we focus on the non-parametric covariance function estimation of an isotropic stationary Gaussian process $\mathbf{X}$, based on a single realization $\mathbf{X}'$ observed at spatial locations $\sigma_1, \ldots, \sigma_n$ forming a regular grid. We place ourselves in the *in-fill* asymptotic setting, assuming that the observed sites $\sigma_1, \ldots, \sigma_n$ form a denser and denser grid of $\mathcal{S}$ as n grows. We provide non-asymptotic bounds for the covariance estimation, thanks to recent concentration results for sums of Gamma random variables (Bercu et al., 2015; Wang and Ma, 2020). To the best of our knowledge, our non-asymptotic bounds for a non-parametric covariance estimation in the in-fill setting are the first results of this nature. These theoretical results, as well as the role played by the technical conditions required to establish them, are illustrated by various numerical experiments on simulated data.

This contribution answers to our **Research Question 2:**

How accurate is the empirical covariance estimator, based on a finite number of observations on a regular grid and with one unique realization? (see Section 1.2 in Chapter 1).

The theoretical results obtained in this chapter are fundamental for deriving the non-asymptotic bounds to establish generalization guarantees of the simple Kriging predictor, that are presented and discussed in Chapter 4, thus answering to our **Research Question 3**.

This chapter is organized as follows. The general setup, together with the main assumptions and concepts to ensure an efficient and flexible estimation of the spatial dependence structure are stated in Section 3.2. The main results of the chapter are

given in Section 3.3, where a non-asymptotic bound for a non-parametric estimator of the covariance function is established. Illustrative experiments are presented in Section 3.4. Finally, Section 3.5 gathers concluding remarks. The proofs of the main results are provided in Appendix A.

Contributions. Our goal is to overcome the challenges posed by the spatial dependency within the data. For that purpose, under appropriate conditions, we assess the accuracy of the non-parametric covariance estimator for second-order stationary Gaussian processes whose covariance function, unknown in practice, is assumed to be isotropic. The code for the numerical experiments on simulated data (which allow assessing the role of each assumption) are available on GitHub.

Context and Motivation. The goal of Kriging is interpolating, *i.e.* predicting the values of a square integrable random process $\mathbf{X} = \{\mathbf{X}_s\}_{s\in\mathcal{S}}$, $\mathcal{S} \subset \mathbb{R}^2$, at all unobserved locations in $\mathcal{S}$, based on one unique realization observed at a finite number $d \geq 1$ of sites $s_1, \ldots, s_d$. When the covariance function is known, the Kriging predictor is optimal (or BLUP, for *Best Linear Unbiased Predictor*, see *e.g.* Stein, 1999, Section 1, and Cressie, 1993, Section 3). However, in practice, the covariance structure of $\mathbf{X}$ is unknown. Thus, the prediction rule is derived from a (generally non i.i.d.) training spatial dataset: a single realization $\mathbf{X}'$ of $\mathbf{X}$, independent from those to be predicted, observed at $n \geq 1$ locations $\sigma_1, \ldots, \sigma_n$ in $\mathcal{S}$. From a non-parametric statistical perspective, an empirical version of the simple Kriging predictor can be built, based on the spatial observations $\mathbf{X}'_{\sigma_1}, \ldots, \mathbf{X}'_{\sigma_n}$ involved in the learning procedure, by means of a plug-in strategy, where the (unknown) covariance is replaced with its estimator. Thus, the covariance function estimation is of crucial importance, since the simple Kriging problem heavily depends on it.

Notation. Let $\mathcal{S} \subset \mathbb{R}^2$ be a (Borel measurable) spatial set and $\mathbf{X}$ be a second-order random field on $\mathcal{S}$ with $\mathbb{R}$ as state space. We denote by $\mu : s \in \mathcal{S} \mapsto \mathbb{E}[\mathbf{X}_s]$ its mean and by $C : (s,t) \in \mathcal{S}^2 \mapsto C(s,t) = Cov(\mathbf{X}_s, \mathbf{X}_t)$ its covariance functions. The main definitions and methodologies are presented in Chapter 2.

3.2 Non-parametric Estimation

The estimation of the covariance function, which the plug-in predictive approach considered in Chapter 4 fully relies on, is based on a 'large' number $n \geq 1$ of observations $\mathbf{X}'(\sigma_n) := (\mathbf{X}'_{\sigma_1}, \ldots, \mathbf{X}'_{\sigma_n})$ exhibiting a certain dependence structure and cannot rely on independent realizations of the random field $\mathbf{X}$, in contrast to the usual statistical learning setup. In this section, we first give the main setting about the sampling scheme and the asymptotic regime (refer to subsection 2.1.1 and the corresponding paragraph in subsection 2.1.6 for a discussion about optimal sampling). Then, we present the covariance estimation and the main assumptions made to ensure a successful approach. Finally, we consider the semi-variogram estimation, which is generally used in Geostatistics since it has several advantages compared to the covariance estimation.

3.2.1 Sampling Setting – Regular Grids and In-fill Asymptotic

Before computing the spatial dependence structure estimation, two crucial questions are: What is the necessary number of samples for an efficient statistical study, and how should they be distributed across the spatial domain? Indeed, the empirical estimation of the dependence structure is inevitably influenced by the spatial arrangement of the observations (see subsection 2.1.6). The question of an optimal sampling scheme has been the center of interest in various studies, see *e.g.* Russo, 1984; McBratney and Webster, 1981; Wang et al., 2020.

Regular grids. In this thesis, the observed sites are supposed to be equispaced, forming a regular grid. Since the goal of this chapter is to explain the main ideas rather than dealing with the problem in full generality, we assume for simplicity that the spatial set $\mathcal{S}$ is equal to the unit square $[0,1]^2$ and the observed sites are the points of the dyadic grid at scale $J \geq 1$ (see Figure 3.1):

$$\mathcal{G}_J = \left\{ (k2^{-J}, l2^{-J}) : \ 0 \leq k, \, l \leq 2^J \right\}. \tag{3.1}$$

In this case, we have $n = (1+2^J)^2$. The gridpoints are indexed using the lexicographic order on $\mathbb{R}^2$ by assigning index $i = k(1+2^J) + (l+1)$ to point $(k2^{-J}, l2^{-J})$, which is denoted by σ_i. We point out that the regularity of the grid formed by the observed sites is key to the present analysis, to control the spectrum of the covariance matrix of the sampled points in a non-asymptotic fashion namely (see also Assumption 3.8 below), so as to define an unbiased semi-variogram estimator at the observed lags with provable accuracy (*cf* Proposition 3.9). Proving that such a control still holds true for irregular grids (under specific assumptions unavoidably, remaining to be formulated precisely) is a great mathematical challenge (even in the 1-d time series case, see e.g. Brockwell and Davis, 1987). Extending the present theoretical study to observations on irregular grids is of importance undeniably, insofar that it may cover various situations in practice, but will be the subject of further research. However, beyond technical barriers, one should pay attention to the fact that measurements at equispaced spatial sites are extremely common and of great interest in practice, since they are precisely those produced by numerous image-producing systems, in Geophysics for instance. For this reason, the geostatistical literature (see *e.g.* Cressie, 1993, Section 2.4) has mainly focused on the regular case, while the irregular case is in contrast poorly documented and no dedicated statistical study of the (even non-asymptotic) behavior of the variogram/covariance estimator has been carried out yet.

Asymptotic Setting. Here, we place ourselves in the *in-fill* asymptotic: the number of observations inside the fixed and bounded domain $\mathcal{S}$ increases, the latter forming a denser and denser grid as $n \to +\infty$. Indeed, the in-fill asymptotic is well-suited for interpolation problems (Stein, 1999; Chang and Stein, 2013). Furthermore, the in-fill asymptotic is obviously preferred in real-world applications where the domain of interest is fixed, for example in the case of temperature prediction in a given country (*cf* Section 4.5 in Chapter 4).

3.2.2 Covariance Estimator

The strong dependence structure in spatial data explains the relationships between data points influenced by their spatial proximity. Understanding this structure is of prime importance before incorporating it into modeling. In Geostatistics, the spatial

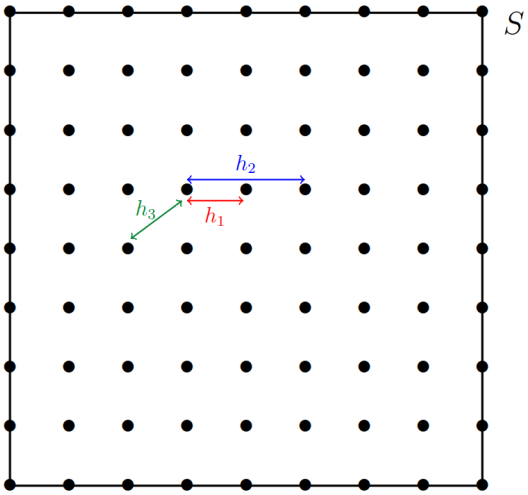


Figure 3.1: Dyadic grid at scale $J = 3$ ($n = 81$). Depicted lags: $h_1 = 2^{-J}$ (in red), $h_2 = 2/2^J$ (in blue) and $h_3 = \sqrt{2}/2^J$ (in green).

dependence is generally modeled using either the covariance function or the semi-variogram function of the process. However, in practice, the dependence structure of an observed phenomenon is unknown and must be estimated.

Recall that in our study, we have access to a single realization of the random field $\mathbf{X}$ to compute the estimation of the covariance function. Therefore, to achieve an accurate estimation, we need to make several assumptions about the random field, as listed below.

Gaussianity, stationarity and isotropy. The assumption of (second-order) stationarity is required so that a frequentist approach can be successful in such a non-parametric framework (see subsection 2.1.2 for further details and the definitions of stationarity, isotropy, and Gaussian processes). Furthermore, since the final objective is the study of the simple Kriging method, the mean of the random field is supposed to be known and equal to zero. In the subsequent analysis, we suppose that the following hypotheses are fulfilled.

Assumption 3.1. *The random field* $\mathbf{X}$ *is centered:* $\mu := 0$.

Assumption 3.2. *The centered random field* $\mathbf{X}$ *is stationary in the second-order sense and its covariance function* C *is isotropic w.r.t. the Euclidean norm, i.e. there exists* $c : \mathbb{R}_+ \to \mathbb{R}$ *s.t.* $c(\|t-s\|) = C(s,t)$ *for all* $(s,t) \in \mathcal{S}^2$.

Assumption 3.3. *There exists a known integer* $j_1 \geq 1$ *s.t.* $c(h) = 0$ *as soon as* $h \geq \sqrt{2} - 2^{-j_1}$.

Remark 3.4. (On Assumption 3.3) *We point out that many popular covariance models fulfill Assumption 3.3. This is the case of the truncated power law, the cubic and the spherical covariance models, used to generate the datasets analyzed in the experiments presented in Section 3.4. On the contrary, the Gaussian, the exponential and the Matern covariance models do not satisfy this hypothesis. However, as shown in Section 4.4 in Chapter 4, the Kriging prediction methodology still performs satisfactorily in these situations, provided that the decorrelation rate is fast enough.*

Assumption 3.5. *The random field* $\mathbf{X}$ *is Gaussian with positive definite covariance function (see Definition 2.17).*

Under Assumption 3.5, the weak stationarity guaranteed by Assumption 3.2 is of course equivalent to strong stationarity (*i.e.* shift-invariance of the finite dimensional marginals) insofar as Gaussian laws are fully characterized by the mean and covariance functions. Under Assumptions 3.1-3.2, we have $C(s,t) = \mathbb{E}[\mathbf{X}_t\mathbf{X}_s] = c(h)$ for all $(s,t) \in \mathcal{S}^2$ s.t. $h = \|s-t\|$. Supposing in addition that Assumption 3.3 is fulfilled (the function c can be then extended to $\mathbb{R}^2$ by setting $c(h) = 0$ for all $h > \sqrt{2}$), a natural estimator $\widehat{c}$ of the covariance function is then defined by $\widehat{c}(h) = 0$ if $h \geq \sqrt{2} - 2^{-j_1}$ and otherwise by

$$\widehat{c}(h) = \frac{1}{n_h} \sum_{(\sigma_i,\sigma_j)\in N(h)} \mathbf{X}'_{\sigma_i}\mathbf{X}'_{\sigma_j}, \tag{3.2}$$

where $N(h) = \left\{(\sigma_i,\sigma_j), \|\sigma_i - \sigma_j\| = h, (i,j) \in [\![1,n]\!]^2\right\}$ is the set of pairs of sites that are at distance h from one another and $n_h = |N(h)|$ denotes its cardinality (notice that $n_0 = n$). Equipped with these notations, notice that a pair of sites $(\sigma_i,\sigma_j) \in N(h)$ at distance h from one another is taken into account twice in the set $N(h)$, since $\|\sigma_i - \sigma_j\| = \|\sigma_j - \sigma_i\| = h$ and so $(\sigma_j,\sigma_i) \in N(h)$. Define $\mathcal{H}_n = \{\|\sigma_i - \sigma_j\| : (i,j) \in [\![1,n]\!]^2\}$ the set of observed lags, which are all less than $\sqrt{2}$. The lemma stated below shows that n_h is of order n for observed lags $h < \sqrt{2} - 2^{-j_1}$ as soon as $n > (\sqrt{2} - 2^{-j_1})^2$, thus ensuring that the number of terms averaged in (3.2) is large enough, of order n namely. Refer to Appendix A for the technical proof.

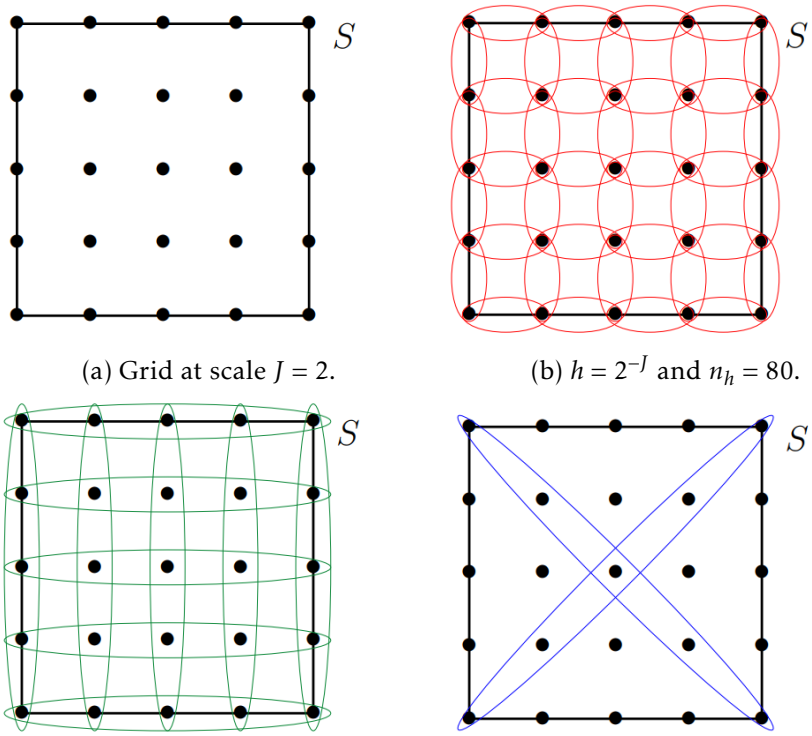
Lemma 3.6. *Suppose that Assumption 3.3 is fulfilled and let $n > (\sqrt{2} - 2^{-j_1})^2$. Then, there exists a constant $0 < \nu \leq 1$ depending on j_1 only such that: $\forall h \in \mathcal{H}_n$, s.t. $h < \sqrt{2} - 2^{-j_1}$, $n_h > \nu n$.*

Figure 3.2a depicts the case of a dyadic grid at scale $J = 2$ for illustration purpose. For small values of h (see Figure 3.2b), the number n_h is large, so that a sufficient number of pairs of sites can be involved in the computation of $\widehat{c}(h)$ in Equation (3.2). As h grows, the number n_h decreases (Figure 3.2c). In the extreme situation, *i.e.* when $h = \sqrt{2}$, the number of pairs is reduced to 4, see Figure 3.2d. Hence, Assumption 3.3, stipulating that the covariance function vanishes for lags exceeding the threshold $\sqrt{2} - 2^{-j_1}$, guarantees that a sufficient number of pairs of observations can be used to estimate the non zero values taken by $c(\cdot)$ on the set of lags formed by the grid-points. Of course, this hypothesis can be relaxed, by assuming a specific decay rate for $c(h)$ as h tends to $\sqrt{2}$ (or equivalently for $c(\sqrt{2} - 2^{-j_1})$ as j_1 tends to ∞). For the sake of simplicity, in order to avoid an excessive number of parameters involved in the problem statement, the inference and statistical learning results are established under Assumption 3.3. As shown below, beyond the number of pairs over which one averages to compute the statistic (3.2), the Gaussian hypothesis, Assumption 3.5, plays a crucial role in describing (the concentration properties of) its distribution.

3.2.3 Semi-variogram Estimator

In Geostatistics, rather than the covariance, one uses the semi-variogram to characterize the second-order dependence structure of the observations (Cressie and Zimmerman, 1992; Cressie, 1993), namely

$$\gamma(h) = \frac{1}{2} Var(\mathbf{X}_{s+h} - \mathbf{X}_s) = \frac{1}{2}\mathbb{E}[(\mathbf{X}_{s+h} - \mathbf{X}_s)^2]. \tag{3.3}$$



(a) Grid at scale $J = 2$.

(b) $h = 2^{-J}$ and $n_h = 80$.

(c) $h = 1$ and $n_h = 20$.

(d) $h = \sqrt{2}$ and $n_h = 4$.

Figure 3.2: Changes in the value of the number n_h of pairs of sites that are at distance h from one another, for different values of the lag h, on a dyadic grid at scale $J = 2$ ($n = 25$).

Its main advantage lies in the fact that its computation does not require the knowledge of the (supposedly constant) mean. We refer the reader to subsection 2.1.3 for a discussion about the advantages of the semi-variogram function. We recall that it is linked to the covariance by the equation $\gamma(h) = c(0) - c(h)$. Observe that, for any lag $h \in \mathcal{H}_n$, unbiased estimators of $\gamma(h)$ based on the n observations are:

$$\forall h < \sqrt{2} - 2^{-j_1}, \quad \widehat{\gamma}(h) = \frac{1}{2n_h} \sum_{(\sigma_i,\sigma_j)\in N(h)} (\mathbf{X}'_{\sigma_i} - \mathbf{X}'_{\sigma_j})^2, \tag{3.4}$$

and $\widehat{\gamma}(h) = 0$ otherwise. This estimator is referred to as the Matheron (1962) semi-variogram estimator. Observe also that

$$\forall h < \sqrt{2} - 2^{-j_1}, \quad \widehat{c}_h(0) := \widehat{\gamma}(h) + \widehat{c}(h) = \frac{1}{n_h} \sum_{i=1}^{n} \mathbf{X}'^2_{\sigma_i} n_h(i), \tag{3.5}$$

and $\widehat{c}_h(0) = 0$ otherwise, where $n_h(i) = |\{j \in \{1, \ldots, n\} : (\sigma_i, \sigma_j) \in N(h)\}|$ for $i \in \{1, \ldots, n\}$. Indeed, one may write: $\forall h \geq 0$,

$$\widehat{\gamma}(h) = \frac{1}{2n_h} \sum_{(\sigma_i,\sigma_j)\in N(h)} (\mathbf{X}'^2_{\sigma_i} + \mathbf{X}'^2_{\sigma_j}) - \frac{1}{n_h} \sum_{(\sigma_i,\sigma_j)\in N(h)} \mathbf{X}'_{\sigma_i}\mathbf{X}'_{\sigma_j}$$

$$= \frac{1}{2n_h} \sum_{(\sigma_i,\sigma_j)\in N(h)} (\mathbf{X}'^2_{\sigma_i} + \mathbf{X}'^2_{\sigma_j}) - \widehat{c}(h),$$

and

$$\frac{1}{2n_h}\sum_{(\sigma_i,\sigma_j)\in N(h)}(\mathbf{X}_{\sigma_i}^{'2}+\mathbf{X}_{\sigma_j}^{'2})=\frac{1}{2n_h}\sum_{i=1}^{n}\mathbf{X}_{\sigma_i}^{'2}\sum_{j=1}^{n}\mathbb{I}\{(\sigma_i,\sigma_j)\in N(h)\}$$
$$+\frac{1}{2n_h}\sum_{j=1}^{n}\mathbf{X}_{\sigma_j}^{'2}\sum_{i=1}^{n}\mathbb{I}\{(\sigma_i,\sigma_j)\in N(h)\}=\frac{1}{n_h}\sum_{i=1}^{n}\mathbf{X}_{\sigma_i}^{'2}n_h(i).$$

Under Assumption 3.5, the distributions of the estimators (3.4)-(3.5) can be classically made explicit, as revealed by the result stated below, see *e.g.* Gaetan and Guyon, 2009; Cressie, 1993.

Proposition 3.7. *Suppose that Assumption 3.5 is fulfilled. Let $h \in \mathcal{H}_n$. Denote by $L(n,h)$ the symmetric positive semi-definite (Laplacian) matrix with entries $L_{i,j}(n,h) = -\mathbb{I}\{(\sigma_i,\sigma_j)\in N(h)\}$ if $i \neq j$ and $L_{i,i}(n,h) = n_h(i)$ and by $\ell_i(h)$'s the n_h eigenvalues of the symmetric positive semi-definite matrix $L(n,h)\Sigma_n$, where $\Sigma_n = \Sigma(\sigma_1, \ldots, \sigma_n)$. Denote also by $D(n,h)$ the diagonal matrix with entries $D_{i,i}(n,h) = n_h(i)$ and by $\rho_i(h)$'s the n_h eigenvalues of the symmetric positive semi-definite matrix $D(n,h)\Sigma_n$. The following assertions hold true.*

(i) The estimators (3.4)-(3.5) are distributed as follows:

$$\widehat{\gamma}(h)\sim\frac{1}{n_h}\sum_{i=1}^{n_h}\ell_i(h)\chi_i^2 \text{ and } \widehat{c_h}(0)\sim\frac{1}{n_h}\sum_{i=1}^{n_h}\rho_i(h)\chi_i^2, \tag{3.6}$$

where the χ_i^2's are independent χ^2 random variables with one degree of freedom.

(ii) The $\ell_i(h)$'s and the $\rho_i(h)$'s are strictly positive, i.e. the matrices $L(n,h)\Sigma_n$ and $D(n,h)\Sigma_n$ are positive definite.

Refer to Appendix A for the technical argument. This result shows one of the great advantages of the semi-variogram estimator: since the average squared difference of dependent normal random variables gives a χ^2 r.v., all the distributions are known. On the contrary, the product of dependent random variables appearing in the covariance function estimation in Equation (3.2) are much more complicated to control (see *e.g.* Nadarajah and Pogány, 2015).

3.3 Tail Bounds Inequalities on Estimation – Main Results

Thanks to the knowledge of the distributions of the estimators (3.4)-(3.5) obtained in Proposition 3.7, we are able to define tail bounds for these estimators, based on recent results in Bercu et al. (2015); Wang and Ma (2020). Recall that, under Assumptions 3.2-3.3, the spatial process $\mathbf{X}$ has an isotropic spectral density (see Equation (2.2) and subsection 2.1.2 for further details) $\Phi(u) = (2\pi)^{-2}\int_{s\in\mathbb{R}^2}\exp(-i\ s^\top u)c(\|s\|)ds = \phi(\|u\|)$. The additional hypothesis below is required in the subsequent analysis. It classically guarantees that the eigenvalues of the covariance matrix of the spatial process sampled on the regular grid $\mathcal{G}_J$ are bounded and bounded away from 0, see *e.g.* Brockwell and Davis, 1987.

Assumption 3.8. *There exist $0 < m \leq M < +\infty$ such that: $\forall u \in \mathbb{R}^2$,*

$$m \leq \sum_{k\in\mathbb{Z}^2}\Phi(u+2\pi k) \leq M.$$

Combining then Lemma 3.6 with Proposition 3.7 and classic bounds for the largest eigenvalues of Laplacian matrices (see the auxiliary results in Appendix A), one may deduce Poisson tail bounds for the deviations between the unbiased estimators and their expectations, from the exponential inequalities established for Gamma r.v.'s in Bercu et al. (2015); Wang and Ma (2020) (see subsection 2.2.4 in Chapter 2 for a presentation of these results). The proof is detailed in Appendix A, together with intermediary results involved in its argument.

Proposition 3.9. *Suppose that Assumptions 3.1–3.8 are fulfilled. Let $h \in \mathcal{H}_n$. For all $t > 0$, we have:*

$$\begin{aligned} \mathbb{P}\left(\left|\widehat{\gamma}(h) - \gamma(h)\right| \geq t\right) &\leq e^{-C_1 nt} + e^{-C_1' nt^2}, \\ \mathbb{P}\left(\left|\widehat{c}_h(0) - c(0)\right| \geq t\right) &\leq e^{-C_2 nt} + e^{-C_2' nt^2}, \end{aligned}$$

where C_i and C_i', $i \in \{1, 2\}$, are positive constants depending on j_1, m and M solely.

Estimation of the covariance function for all lags. The empirical covariance function $\widehat{c}(h)$ can be extrapolated at unobserved lags $h \in [0, \sqrt{2} - 2^{-j_1}] \setminus \mathcal{H}_n$ by means of various non-parametric procedures, such as local averaging methods. For simplicity, one may consider a piecewise constant estimator, for instance the 1-NN estimator $\widehat{c}(h) = \widehat{c}(l_h)$, where $l_h = \arg\min_{l \in \mathcal{H}_n} \|h - l\|$ (breaking ties in an arbitrary fashion). As $\|h - l_h\| \leq 2^{-J} = 1/(\sqrt{n} - 1)$, the (weak) smoothness hypothesis below then permits to control the covariance estimation error at unobserved lags.

Assumption 3.10. *The mapping $h \in [0,\ \sqrt{2} - 2^{-j_1}] \mapsto c(h)$ is of class $\mathcal{C}^1$ with gradient bounded by $Q < +\infty$ and there exists $0 < B < +\infty$, such that $\sup_{h \geq 0} |c(h)| \leq B$.*

Of course, under more restrictive regularity assumptions, the accuracy of other classic non-parametric estimation techniques (*e.g.* splines of degree larger than 2) can be established, and can be the subject of further research. The result below is proved at length in Appendix A. Under Assumption 3.10, it simply follows from Proposition 3.9 combined with the union bound and the finite increment inequality.

Corollary 3.11. *Suppose that Assumptions 3.1–3.10 are satisfied. Then, for any $\delta \in (0, 1)$, we have with probability at least $1 - \delta$:*

$$\sup_{h \geq 0} \left|\widehat{c}(h) - c(h)\right| \leq C_3 \sqrt{\log(4n/\delta)/n} + Q/(\sqrt{n} - 1),$$

as soon as $n \geq C_3' \log(4n/\delta)$, where C_3 and C_3' are positive constants depending on j_1, m and M solely.

To the best of our knowledge, this non-asymptotic bound for a non-parametric estimator of the covariance function in the in-fill setup is the first result of this kind, the vast majority of the results documented in the literature being either of asymptotic nature and/or related to parametric inference. One may refer to Hall et al. (1994) for limit results related to kernel smoothing methods applied to covariance estimation.

Simple Kriging. As the main objective of the first part of this thesis is to provide statistical guarantees to the simple Kriging method, Chapter 4 aims at establishing generalization bounds for the performance of the predicting function constructed by

means of the plug-in approach. Thus, based on the estimator $\widehat{c}$ of the covariance function, one may derive empirical counterparts of the quantities involved in the simple Kriging predictor (see Equation (4.4) for its definition), namely $\Sigma(\mathbf{s}_d)$ and $\mathbf{c}_d(s)$ and form an empirical version of the optimal Kriging rule. Therefore, these bounds can be established from the non-asymptotic guarantees stated in the corollary above.

3.4 Illustrative Experiments

We now illustrate the theoretical analysis carried out above by numerical results, illuminating the impact of the assumptions made to get finite sample guarantees.

The (fully reproducible) experiments, available on GitHub, were implemented in Python 3.6, using the library `gstools` (Müller and Schüler, 2020). Below, two covariance models of a Gaussian random field $\mathbf{X}$ are considered. First, an isotropic *truncated power law* (TPL) covariance function is considered:

$$c : h \in [0, +\infty) \mapsto \left(1 - h/\theta\right)^{\frac{3}{2}} \mathbb{I}\Big\{h \leq \theta\Big\}, \tag{3.7}$$

where $\theta \in \mathbb{R}_+^*$ is the correlation length. It satisfies all the stated assumptions (Golubov, 1981). Next, covariance estimation is applied to a Gaussian field $\mathbf{X}$ with *Gaussian* covariance function:

$$c : h \in [0, +\infty) \mapsto \exp\left(-h^2/\theta^2\right), \tag{3.8}$$

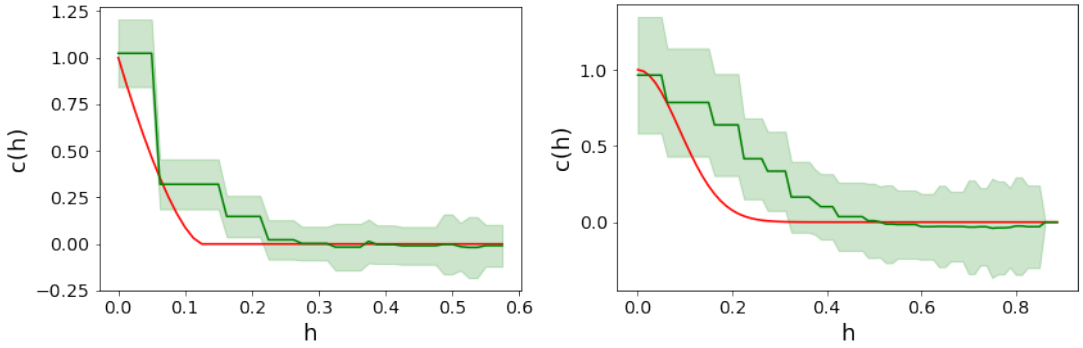
which does not satisfy Assumption 3.3 but vanishes very quickly.

Given the spatial domain $\mathcal{S} = [0,1]^2$, an independent realization of $\mathbf{X}$ was simulated, serving for the non-parametric covariance estimation and referred to as the training spatial dataset $\mathbf{X}'$. In accordance with the statistical framework considered in Section 3.2, this realization is observed at n sites $(\sigma_1, \cdots, \sigma_n)$ supposed to form a dyadic grid at scale $J \geq 1$, with $n = (1 + 2^J)^2$. The estimation of the covariance function of the random field is then computed using Equation (3.2) from the n observations. In order to illustrate its accuracy, the estimator is computed for 100 independent simulations of X observed at the same fixed sites.

The corresponding results for the two covariance functions are depicted in Figure 3.3 (for a dyadic grid at scale $J = 3$ and with correlation length fixed at $\theta = 5$), where, for each model, the true covariance function appears in red and the mean of the estimated one in green, together with the corresponding mean standard deviation over the 100 replications. Observe in Figure 3.3a, for the truncated power law function, that the estimation is close to the true value and equal to zero beyond a certain threshold. For the Gaussian model, Figure 3.3b shows that the estimation method is less accurate than for the previous covariance model. In particular, it unsuccessfully detects the true correlation parameter θ.

Additional Covariance Models. Based on several covariance models, fulfilling or not our assumptions, extra numerical experiments were performed.

Besides the covariance functions depicted above, the following covariance functions are considered (where $\theta \in \mathbb{R}_+^*$ is the correlation length):



(a) TPL (b) Gaussian

Figure 3.3: Estimation of the truncated power law (left) and the Gaussian (right) covariance functions, on a dyadic grid at scale $J = 3$ ($n = 81$), with $\theta = 5$. For each model, the red line corresponds to the true covariance function and the green line to the mean of the estimated one, together with the corresponding mean standard deviation (in green shaded bands), over 100 replications.

- the *cubic* covariance function:

$$c : h \in [0, +\infty) \mapsto \left(1 - \left(7\frac{h^2}{\theta^2} - \frac{35}{4}\frac{h^3}{\theta^3} + \frac{7}{2}\frac{h^5}{\theta^5} - \frac{3}{4}\frac{h^7}{\theta^7}\right)\right)\mathbb{I}\{h \leq \theta\}. \tag{3.9}$$

- the *spherical* covariance function:

$$c : h \in [0, +\infty) \mapsto \left(1 - \left(\frac{3}{2}\frac{h}{\theta} - \frac{1}{2}\frac{h^3}{\theta^3}\right)\right)\mathbb{I}\{h \leq \theta\}. \tag{3.10}$$

- the *exponential* covariance function:

$$c : h \in [0, +\infty) \mapsto \exp\left(-h/\theta\right). \tag{3.11}$$

- the *Matern* covariance function with smoothness parameter ν_m:

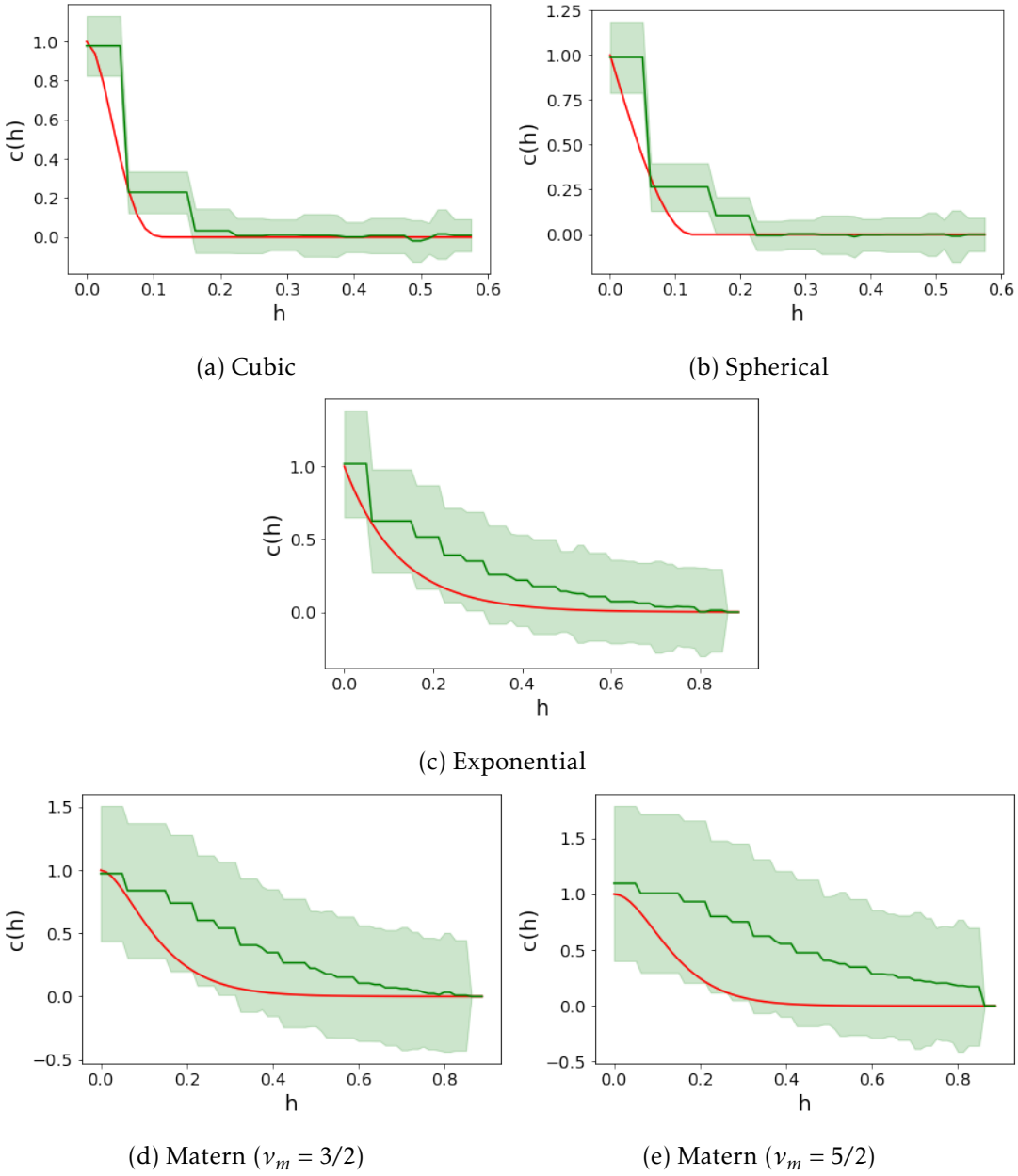
$$c : h \in [0, +\infty) \mapsto \frac{2^{1-\nu_m}}{\Gamma(\nu_m)}\left(\sqrt{2\nu_m}\frac{h}{\theta}\right)^{\nu_m} K_{\nu_m}\left(\sqrt{2\nu_m}\frac{h}{\theta}\right), \tag{3.12}$$

  where Γ is the gamma function and K_{ν_m} is the modified Bessel function of the second kind.

Remark 3.12. (Matern model) *When $\nu_m = p + \frac{1}{2}$ where p is an integer, the Matern covariance function is a product of an exponential function and a polynomial function of order p.*

Following this remark, we select the two following smoothness parameter values for the Matern model:

- when $\nu_m = \frac{3}{2}$: $c(h) = \left(1 + \sqrt{3}\frac{h}{\theta}\right)\exp\left(-\sqrt{3}\frac{h}{\theta}\right)$, $\forall h \in [0, +\infty)$, and



(a) Cubic

(b) Spherical

(c) Exponential

(d) Matern ($\nu_m = 3/2$)

(e) Matern ($\nu_m = 5/2$)

Figure 3.4: Estimation of the cubic (top left), the spherical (top right), the exponential (center), the Matern with $\nu_m = 3/2$ (bottom left) and $\nu_m = 5/2$ (bottom right) covariance functions, on a dyadic grid at scale $J = 3$ ($n = 81$), with $\theta = 5$.

- when $\nu_m = \frac{5}{2}$: $\forall h \in [0, +\infty)$, $c(h) = \left(1 + \sqrt{5}\frac{h}{\theta} + \frac{5}{3}\left(\frac{h}{\theta}\right)^2\right)\exp\left(-\sqrt{5}\frac{h}{\theta}\right)$.

Note that the cubic and spherical covariance models satisfy all the assumptions stated in this chapter, whereas the exponential and Matern covariance functions do not verify Assumption 3.3. We apply the same procedure as before for the five additional covariance models, with the same setting: the training dataset is composed of observations sampled on a dyadic grid at scale $J = 3$ ($n = 81$). As an illustration of the covariance estimation, Figure 3.4 shows, for each model, the true covariance function in red and the mean of the estimated covariance function in green (with the

corresponding mean standard deviation), on 100 independent simulations. Observe that, for both the cubic (top left) and the spherical (top right) covariance functions, that satisfy Assumption 3.3, the estimation is accurate and it successfully estimates the threshold after which the covariance is equal to zero. For the covariance functions that do not satisfy Assumption 3.3, the exponential function is quite accurate but does not detect the correlation length, whereas for the Matern function, when $\nu_m = 5/2$, the estimation is less accurate than for the smaller value of the smoothness parameter. In particular, for these three covariance models, when the true covariance function tends to zero, the estimation is still considerably different from zero for an important number of lags and the mean standard deviation is large.

3.5 Conclusion

In this chapter, we have derived tail bounds for the non-parametric covariance estimator, overcoming our **Challenge 2** (Section 1.2 in Chapter 1). In the first part of this thesis, the main objective is to develop a novel theoretical framework offering non-asymptotic guarantees for the empirical simple Kriging predictor. The simple Kriging problem heavily depends on the spatial dependence structure of the observed phenomenon, unknown in practice. Thus, computing an accurate estimation of the covariance function is of prime importance. Furthermore, the massive character of spatial datasets now available suggests resorting to more flexible, non-parametric, approaches to analyze spatial observations.

Key points of this Chapter.

Technical Assumptions involved in this Chapter and their role

- **Regular grid** of observations and **in-fill asymptotic** $\longrightarrow$ the number of terms averaged in (3.2) and (3.4) is large enough.
- **Assumption 3.1:** in simple Kriging, the mean is supposed to be known, and for simplicity $\mathbf{X}$ is assumed to be centered.
- **Assumption 3.2:** the phenomenon is sufficiently homogeneous within $\mathcal{S}$ $\longrightarrow$ ensures a successful frequentist approach (see subsec. 2.1.2, Chap. 2).
- **Assumption 3.3:** guarantees that a sufficient number of pairs of observations can be used to estimate (3.3) for the observed lags.
- **Assumption 3.5:** the second-order stationarity is equivalent to strong stationarity (see subsec. 2.1.2, Chap. 2), and the distribution of the semi-variogram estimator can be made explicit in Proposition 3.7.
- **Assumption 3.8:** combined with the previous assumptions, allows to deduce non-asymptotic bounds for the estimators (3.4) and (3.5) $\longrightarrow$ Proposition 3.9 for observed lags on the regular grid.
- **Assumption 3.10:** allows to extend the results in Proposition 3.9 to all lags.

Main results of this Chapter

- **Lemma 3.6:** ensures that the number of terms averaged in (3.2) and (3.4) is large enough $\longrightarrow$ accurate and stable estimator.
- **Proposition 3.7:** derives the distributions of the estimators (3.4) and (3.5) $\longrightarrow$ the estimators can be seen as a weighted sum of χ^2 random variables.
- **Proposition 3.9:** gives Poisson tail bounds for the deviations between the unbiased estimators (3.4) and (3.5) and their expectations, for all observed lags on the regular dyadic grid.
- **Corollary 3.11:** extends the previous result to unobserved lags $\longrightarrow$ provides a non-asymptotic bound for a non-parametric estimation of the covariance function.

  $\Longrightarrow$ **OBJECTIVE:** These results allow us to establish **statistical guarantees for the simple Kriging method** (**Research Question 3**, Section 1.2): see Chapter 4 for generalization bounds for the performance of the empirical Kriging predictor.

4

Statistical Learning Guarantees

Contents

4.1 Introduction

This chapter, which is in continuity of Chapter 3, aims at contributing to our **Research Questions 3** (see Section 1.2 in Chapter 1):

What is the non-asymptotic behavior of the Kriging predictor when the dependence structure is unknown and with a finite number of observations? To what extent the Kriging weights depend on the accuracy of the covariance function estimation and on the location of samples?

In this chapter, we investigate *Kriging*, the flagship problem in Geostatistics, introduced by Krige (1951) and later in the work of Matheron (1962). In the standard Kriging setup, the spatial process $\mathbf{X}$ is observed at $d \geq 1$ sites $s_1, \ldots, s_d$ in the domain $\mathcal{S}$. Based on a (generally non i.i.d.) training dataset, a single realization $\mathbf{X}'$ of $\mathbf{X}$ observed at $n \geq 1$ locations $\sigma_1, \ldots, \sigma_n$ in $\mathcal{S}$, the goal is to build a map $f : \mathcal{S} \times \mathbb{R}^d \to \mathbb{R}$ in order to predict $\mathbf{X}$ at all unobserved sites $s \in \mathcal{S}$ with minimum Mean Squared Error (MSE). In *simple Kriging*, the mean of $\mathbf{X}$ is supposed to be known and the goal is to search for a predictive map $f(s) = f(s, (\mathbf{X}_{s_1}, \ldots, \mathbf{X}_{s_d}))$ that is linear in $\mathbf{X}(\mathbf{s}_d) := (\mathbf{X}_{s_1}, \ldots, \mathbf{X}_{s_d})$. In this chapter, we start with recalling that the optimal predictor of this type (which can be derived by ordinary least squares) has the same form as a Kernel Ridge Regressor (KRR), once the Gram matrix is replaced with the true covariance matrix of the random vector $\mathbf{X}(\mathbf{s}_d)$. Then, using a non-parametric estimation of the covariance function,

an empirical version of the optimal linear predictor is obtained by means of a plug-in approach, *i.e.* by replacing the (unknown) covariance with the estimator in the KRR type formula. This strategy can also be viewed as Empirical Risk Minimization (ERM), under the hypothesis of stationarity (combined with appropriate conditions on the decorrelation rate). Based on the non-asymptotic results for the non-parametric covariance function of a Gaussian process derived in Chapter 3, under in-fill and regular grid assumptions for the observations, we establish non-asymptotic bounds for the excess risk of the predictive function thus constructed. To the best of our knowledge, this is the only theoretical analysis of this nature documented in the statistical and machine learning literature. In order to illustrate the generalization capacity of the non-parametric predictive approach analyzed in this chapter and the impact of the conditions stipulated to guarantee it, a number of numerical experiments, on both simulated and real data, are carried out.

This chapter is structured as follows. As a first go, in Section 4.2, we highlight the similarity between KRR and the dual Kriging problem. In Section 4.3, excess risk bounds for the empirical version of the simple Kriging rule are established under appropriate assumptions, which are discussed at length. Numerical results on simulated data are presented in Section 4.4 for illustration purposes, while experimental results on meteorological real data are presented in Section 4.5. Section 4.6 shows additional numerical experiments concerning the possible extension of our study. Finally, concluding remarks are collected in Section 4.7. The proofs of the main results are provided in Appendix B.

Contributions. Our aim is to develop a new theoretical framework for the simple Kriging problem, adopting a statistical learning view of the well-known geostatistical method. A major question here is the following: To what extent the accuracy of the covariance function estimation of a random field and the locations of the samples can influence the Kriging predictor? Based on the theoretical and experimental results obtained in Chapter 3, we aim at defining the accuracy of the empirical Kriging predictor. To do so, our contributions are:

- We provide non-asymptotic bounds for the accuracy of the covariance matrix and precision matrix estimations.
- We assess the generalization capacity of the empirical simple Kriging predictor, at all unobserved sites of the spatial domain. This accuracy is determined by the values of the global excess risk, defined as the global gap between the prediction errors obtained when the covariance function is known and when the true covariance function is replaced by its empirical counterpart.
- The implementation of the numerical experiments on both simulated and real data presented here are available on GitHub.

Setting and Notations. We place ourselves in the same setting as the one adopted in Chapter 3: the observations for the training spatial dataset are assumed to be taken on a regular grid, under the in-fill asymptotic setting (see subsection 3.2.1). When mentioned, we also make the same assumptions stated before: Assumptions 3.1, 3.2, 3.3 and 3.5, introduced in subsection 3.2.2, and Assumptions 3.8 and 3.10, formulated in Section 3.3. We refer the reader to Chapter 2, where the main notations are

introduced and the basic concepts pertaining to the theory of Kriging in Geostatistics – involved in the subsequent analysis – are briefly recalled, together with some key results related to kernel ridge regression.

4.2 Viewing Dual Kriging as a KRR Problem

In spite of the major difference regarding the statistical setup, the Kriging and regression problems (introduced in Chapter 2) share similarities in the optimization approach considered to solve them. See also Kanagawa et al., 2018; Kanevski et al., 2009 for a detailed discussion about the connections between Gaussian processes and kernel methods.

As discussed in Section 2.3 in Chapter 2, the aim of simple Kriging may be to develop a decision function $f : \mathcal{S} \times \mathbb{R}^d \to \mathbb{R}$ to predict $\mathbf{X}$ across the entire spatial set $\mathcal{S}$ based on observations of the spatial process at a finite number of points $s_1, \ldots, s_d$ within $\mathcal{S}$. The global accuracy of the predictive map f over the entire set $\mathcal{S}$ can be measured by the Integrated Mean Squared Error (IMSE)

$$L_{\mathcal{S}}(f) = \int_{s\in\mathcal{S}} L(s, f(s))\, ds = \mathbb{E}_{\mathbf{X}}\left[\int_{s\in\mathcal{S}} \Big(f(s,\, \mathbf{X}(\mathbf{s}_d)) - \mathbf{X}_s\Big)^2\, ds\right]. \tag{4.1}$$

Remark 4.1. (Alternative to ERM) *We point out that, in order to avoid the restrictions stemming from the sub-Gaussianity assumptions, an alternative to ERM in regression, relying on a tournament method combined with the Median-of-Means (MoM) estimation procedure, has recently received much attention in the statistical learning literature, see Lugosi and Mendelson (2016) for further details. Extension of the MoM approach to Kriging will be the subject of future work.*

Hence, the objective is to predict an infinite number of output variables $\mathbf{X}_s$, with $s \in \mathcal{S}$, based on the input variables $\mathbf{X}_{s_1}, \ldots, \mathbf{X}_{s_d}$. It can be viewed as a multitask predictive problem with an infinite number of tasks. In the ordinary formulation, the prediction $f(s)$ at any point $s \in \mathcal{S}$ is assumed to be a linear combination of the $\mathbf{X}_{s_i}$'s

$$f_{\Lambda_d}(\cdot,\, \mathbf{X}(\mathbf{s}_d)) = \lambda_1(\cdot)\mathbf{X}_{s_1} + \ldots + \lambda_d(\cdot)\mathbf{X}_{s_d}, \tag{4.2}$$

where $\Lambda_d : s \in \mathcal{S} \mapsto (\lambda_1(s),\, \ldots,\, \lambda_d(s))$ is a measurable function valued in $\mathbb{R}^d$. In this case, we have:

$$L_{\mathcal{S}}(f_{\Lambda_d}) = \int_{s\in S} \Big(Var(\mathbf{X}_s) + \Lambda_d(s)^\top \Sigma(\mathbf{s}_d)\Lambda_d(s) - 2\, \mathbf{c}_d(s)^\top \Lambda_d(s)\Big) ds. \tag{4.3}$$

The optimal predictive rule of this form regarding the expected prediction error can be straightforwardly deduced from Lemma 2.20 in Chapter 2. It is described in the result stated below, the proof of which is omitted.

Lemma 4.2. *Suppose that the hypotheses of Lemma 2.20 are fulfilled. Define the predictive mapping:* $\forall s \in \mathcal{S}$,

$$f_{\Lambda_d^*}(s, \mathbf{X}(\mathbf{s}_d)) = \langle \Lambda_d^*(s), \mathbf{X}(\mathbf{s}_d)\rangle = \mathbf{X}(\mathbf{s}_d)^\top \Sigma(\mathbf{s}_d)^{-1} \mathbf{c}_d(s). \tag{4.4}$$

We have

$$f_{\Lambda_d^*} = \arg\min_f L_{\mathcal{S}}(f). \tag{4.5}$$

In addition, the minimum global error is

$$L_{\mathcal{S}}(f_{\Lambda_d^*}) = \int_{s\in\mathcal{S}} \Big(Var(\mathbf{X}_s) - \mathbf{c}_d(s)^\top \Sigma(\mathbf{s}_d)^{-1} \mathbf{c}_d(s) \Big)\, ds.$$

Observe that the mapping (4.4) at $s \in \mathcal{S}$ has the same form as that of the kernel ridge regressor at $z \in \mathbb{R}^d$ in Equation (2.15) in Chapter 2, except that the regularized Gram matrix $\alpha \mathbf{I}_N + \mathcal{K}_N$ is replaced by $\Sigma(\mathbf{s}_d)$, the vector $\mathbf{Y}_N$ by $\mathbf{X}(\mathbf{s}_d)$ and $\kappa_N(z)$ by $\mathbf{c}_d(s)$.

The major difference between the two frameworks is of statistical nature. If one tries to predict $\mathbf{X}_s$ by a linear combination of the components of the observed random vector $\mathbf{X}(\mathbf{s}_d)$ in simple Kriging, statistical fitting does not rely on the observation of $n \geq 1$ independent copies of the pair output-input $(\mathbf{X}_s, \mathbf{X}(\mathbf{s}_d))$ but on the observation of a single realization $\mathbf{X}'$ of the random field $\mathbf{X}$ at certain sites $\sigma_1, \ldots, \sigma_n$ solely and some appropriate structural (possibly parametric) assumptions regarding the second-order structure of the random field $\mathbf{X}$, see *e.g.* Chiles and Delfiner (1999). Before analyzing simple Kriging from a statistical learning perspective, a few remarks are in order.

Remark 4.3. (Gaussian random fields, bis) *We point out that, in the case where the random field $\mathbf{X}$ is Gaussian, the mapping $f_{\Lambda_d^*}$ is a minimizer of the global error $L_{\mathcal{S}}$ over the set of all predictive rules $f(s, \mathbf{X}(\mathbf{s}_d))$ such that the IMSE (4.3) is well-defined.*

Remark 4.4. (Worst case error vs integrated error) *Rather than integrating the pointwise MSE over the spatial domain $\mathcal{S}$ to define the global accuracy of a predictive map f (see (4.1)), one may naturally consider the supremum of the MSE over $\mathcal{S}$, namely $\sup_{s\in\mathcal{S}} L(s, f(s))$. Notice that, under the assumptions stipulated, the rate-bound results obtained in the subsequent analysis obviously remain valid when substituting the IMSE with it.*

Remark 4.5. (Alternative framework) *The spatial prediction problem has been investigated in a different statistical framework, much more restrictive regarding practical applications, assuming the observation of $N \geq 1$ independent copies of $\mathbf{X}(\mathbf{s}_d)$, in Qiao et al. (2018), where a non-asymptotic analysis is carried out as N increases. We also point out that, instead of the classic in-fill setting considered here (stipulating that the grid $\sigma_1, \ldots, \sigma_n$ formed by the observed sites in $\mathcal{S}$ in the learning/estimation stage is denser and denser, while $\mathcal{S}$ is fixed), the out-fill framework can be considered alternatively (the prediction accuracy is then analyzed as the spatial domain $\mathcal{S}$ becomes wider and wider) or combined with the in-fill model in a hybrid fashion, see e.g. Hall and Patil (1994) and the references mentioned in subsection 2.1.1 in Chapter 2.*

Plug-in predictive rules. The quantities $\Sigma(\mathbf{s}_d)$ and $\mathbf{c}_d(s)$ are unknown in practice just like the risk (4.3) and must be replaced by estimators in order to form an estimator $\widehat{\Lambda}_d$ of Λ_d^* (or an empirical version of (4.3)). For this reason, establishing rate bounds that assess the generalization capacity of the resulting predictive map $f_{\widehat{\Lambda}_d}$ is far from straightforward. It is the aim of the subsequent analysis to develop a non-asymptotic and non-parametric framework for simple Kriging with statistical guarantees, based on the preliminary finite-sample study of the performance of a covariance estimator $\widehat{c}(\cdot)$ developed in Chapter 3. The angle embraced here is thus different from that usually adopted in the traditional Kriging literature, often calling forth the use of MLE methods (see *e.g.* Section 5.3.3 in Gaetan and Guyon, 2009). In contrast, it is akin to that of statistical learning, particularly relevant when the availability of large training datasets permits to consider flexible techniques, avoiding the specification of a parametric class of probability laws.

4.3 Excess Risk Bounds in Simple Kriging – Main Results

In this section, we explain how a solution to Kriging with statistical guarantees in the form of non-asymptotic learning rate bounds can be derived from an accurate estimator of the covariance under appropriate conditions.

Equipped with the non-asymptotic results established in Section 3.3 in Chapter 3, we now address the simple Kriging problem from a predictive learning perspective. Let $d \geq 1$ and consider arbitrary pairwise distinct sites $s_1, \ldots, s_d$ in $\mathcal{S} = [0,1]^2$. The goal is to predict the value $\mathbf{X}_s$ taken by $\mathbf{X}$ at any site $s \in \mathcal{S}$ based on $(\mathbf{X}_{s_1}, \ldots, \mathbf{X}_{s_d})$, nearly as accurately as the optimal Kriging rule (4.4) would do it. For this purpose, one uses a training dataset, composed of observations $\mathbf{X}'_{\sigma_1}, \ldots, \mathbf{X}'_{\sigma_n}$ of $\mathbf{X}'$, an independent copy of $\mathbf{X}$, at sites $\sigma_1, \ldots, \sigma_n$ forming a regular dyadic grid (see subsection 3.2.1 in Chapter 3). Consider $\widehat{c}(\cdot)$, the estimator of the covariance function studied in Sections 3.2 and 3.3 in Chapter 3, based on the $\mathbf{X}'_{\sigma_i}$'s. From $\widehat{c}(\cdot)$, the covariance matrix $\Sigma(\mathbf{s}_d)$ and the covariance vector $\mathbf{c}_d(s)$ can be naturally estimated as follows:

$$\widehat{\mathbf{c}}_d(s) = \left(\widehat{c}(\|s - s_1\|, \ldots, \widehat{c}(\|s - s_d\|\right) \text{ for } s \in \mathcal{S}, \tag{4.6}$$

$$\widehat{\Sigma}(\mathbf{s}_d) = \left(\widehat{c}(\|s_i - s_j\|)\right)_{1 \leq i,\, j \leq d}. \tag{4.7}$$

Assumption 4.6. *Let $0 < \underline{m} \leq \overline{M} < +\infty$ and assume that the eigenvalues of the covariance matrix $\Sigma(\mathbf{s}_d)$ are upper bounded by $\overline{M}$, lower bounded by $\underline{m}$.*

4.3.1 Estimation of the Precision Matrix

Under Assumption 3.5, the matrix $\Sigma(\mathbf{s}_d)$ is always invertible for any pairwise distinct sites $s_1, \ldots, s_d$, which permits to define the Kriging rule (4.4), involving the precision matrix $\Sigma(\mathbf{s}_d)^{-1}$. The simplest way of building an estimate of the precision matrix is to invert (4.7), when it is positive definite. This theoretically happens with overwhelming probability, as shown by the result stated below, and turned out to be true in all the numerical experiments presented in Section 4.4. Hence, the estimator of the precision matrix we consider here in order to build an empirical version of the predicting function (4.4) is the inverse of (4.7), when the latter is definite positive, and that of any definite positive regularized version (e.g. Tikhonov, 1943) of the latter otherwise. It is (possibly abusively) denoted by $\widehat{\Sigma}(\mathbf{s}_d)^{-1}$ in both situations. Its accuracy is described in a non-asymptotic fashion by the bound stated in the result below.

Proposition 4.7. *Suppose that Assumptions 3.1–4.6 are satisfied. The following assertions hold true.*

(i) For any $\delta \in (0,1)$, we have with probability at least $1 - \delta$:

$$|||\widehat{\Sigma}(\mathbf{s}_d) - \Sigma(\mathbf{s}_d)||| \leq C_3\, d\, \sqrt{\log(4n/\delta)/n} + d\, Q/(\sqrt{n} - 1),$$

as soon as $n \geq C'_3 \log(4n/\delta)$, where C_3 and C'_3 are positive constants depending on j_1, m and M solely (see Corollary 3.11).

(ii) For any $\delta \in (0,1)$, we have with probability at least $1 - \delta$:

$$|||\widehat{\Sigma}(\mathbf{s}_d)^{-1} - \Sigma(\mathbf{s}_d)^{-1}||| \leq C_4\, d\, \sqrt{\log(4n/\delta)/n} + C'_4\, d\, Q/(\sqrt{n} - 1),$$

as soon as $n \geq C''_4 \log(4n/\delta)$, where C_4, C'_4 and C''_4 are positive constants depending on j_1, m, M, $\underline{m}$ and $\overline{M}$ solely.

Assertion *(i)* simply follows from Corollary 3.11 in Chapter 3, the operator norm and the max norm being equivalent in finite dimension. The proof of the second assertion uses a classic inequality for inverse matrices as in Wedin (1973) combined with Assertion *(i)*. Refer to Appendix B for further technical details.

4.3.2 Empirical Risk Minimization and Generalization Capacity

Now, replacing $\Sigma(\mathbf{s}_d)^{-1}$ and $\mathbf{c}_d(s)$ by the estimators introduced above, a natural empirical counterpart of Λ_d^* is built by means of the *plug-in* method:

$$\widehat{\Lambda}_d(s) = \widehat{\Sigma}(\mathbf{s}_d)^{-1}\widehat{\mathbf{c}}_d(s). \tag{4.8}$$

We point out that it actually corresponds to an empirical risk minimizer. Indeed, (4.8) is the minimizer of

$$\Lambda_d(s)^\top \widehat{\Sigma}(\mathbf{s}_d)\Lambda_d(s) - 2\widehat{\mathbf{c}}_d(s)^\top \Lambda_d(s)$$

over $\Lambda_d(s)$ in $\mathbb{R}^d$, which functional can be viewed as an empirical version of $L(s,\ f_{\Lambda_d}(s)) - c(0) = \Lambda_d(s)^\top \Sigma(\mathbf{s}_d)\Lambda_d(s) - 2\mathbf{c}_d(s)^\top \Lambda_d(s)$, the pointwise risk at s up to an additive term independent from $\Lambda_d(s)$ under the assumptions introduced in Section 3.2. Define thus the empirical predictive mapping $f_{\widehat{\Lambda}_d}$ by:

$$f_{\widehat{\Lambda}_d}(s, \mathbf{x}(\mathbf{s}_d)) = \langle \widehat{\Lambda}_d(s), \mathbf{x}(\mathbf{s}_d)\rangle = \mathbf{x}(\mathbf{s}_d)^\top \widehat{\Sigma}(\mathbf{s}_d)^{-1}\widehat{\mathbf{c}}_d(s), \tag{4.9}$$

for all $s \in \mathcal{S}$ and any $\mathbf{x}(\mathbf{s}_d) = (x_{s_1}, \ldots, x_{s_d}) \in \mathbb{R}^d$. The (random) predictive function (4.9) can be used to predict the values taken by $\mathbf{X}$, any independent copy of the random field $\mathbf{X}'$ partially observed in the learning/estimation phase, over the whole spatial domain $\mathcal{S}$ based on the input observations $\mathbf{X}(\mathbf{s}_d) = (\mathbf{X}_{s_1}, \ldots, \mathbf{X}_{\mathbf{s}_d})$. Conditioned upon the $\mathbf{X}'_{\sigma_i}$'s, it is of course a linear prediction rule which minimizes the empirical counterpart of $L_{\mathcal{S}}(f_{\Lambda_d})$ based on the $\mathbf{X}'_{\sigma_i}$'s, namely

$$\widehat{L}_{\mathcal{S}}(f_{\Lambda_d}) := \int_{s\in\mathcal{S}} \left(\widehat{c}(0) + \Lambda_d(s)^\top \widehat{\Sigma}(\mathbf{s}_d)\Lambda_d(s) - 2\widehat{\mathbf{c}}_d(s)^\top \Lambda_d(s)\right) ds, \tag{4.10}$$

over all Borel measurable functions $\Lambda_d : \mathcal{S} \to \mathbb{R}^d$ such that (4.10) is well-defined (as previously noticed, the quantity integrated over $\mathcal{S}$ then reaches its minimum at all s in $\mathcal{S}$). Hence, the plug-in predictive rule (4.9) can also be derived from empirical risk minimization, the main paradigm of statistical learning, see *e.g.* Devroye et al. (1996).

The predictive performance of the function $f_{\widehat{\Lambda}_d}$ constructed on the basis of the $\mathbf{X}'_{\sigma_i}$'s is then measured by the conditional expectation, obtained by replacing $\Lambda_d(s)$ by its empirical counterpart $\widehat{\Lambda}_d(s)$ in (4.1) :

$$\begin{aligned} L_{\mathcal{S}}(f_{\widehat{\Lambda}_d}) &= \mathbb{E}_X\left[\int_{s\in\mathcal{S}} \left(f_{\widehat{\Lambda}_d}(s, \mathbf{X}(\mathbf{s}_d)) - \mathbf{X}_s\right)^2 ds \mid \mathbf{X}'_{\sigma_1}, \ldots, \mathbf{X}'_{\sigma_n}\right] \\ &= \int_{s\in\mathcal{S}} \left(c(0) + \widehat{\Lambda}_d(s)^\top \Sigma(\mathbf{s}_d)\widehat{\Lambda}_d(s) - 2\,\mathbf{c}_d(s)^\top \widehat{\Lambda}_d(s)\right) ds. \end{aligned}$$

It is a random quantity since it depends upon the training data, that is larger than $L_{\mathcal{S}}^* := L_{\mathcal{S}}(f_{\Lambda_d^*})$ with probability one, see Lemma 4.2. The theorem below shows that, with large probability, the prediction error of the empirical simple Kriging rule $f_{\widehat{\Lambda}_d}$ is close to the minimal prediction error $L_{\mathcal{S}}^*$, assessing its generalization capacity at unobserved sites.

Theorem 4.8. *Suppose that Assumptions 3.1–4.6 are satisfied. The following assertions hold true.*

(i) For any $\delta \in (0,1)$, we have with probability at least $1-\delta$:

$$\sup_{s\in\mathcal{S}} \|\widehat{\Lambda}_d(s) - \Lambda_d^*(s)\| \leq C_5\, d\, \sqrt{d\, \log(4n/\delta)/n} + C_5'\, d\, \sqrt{d}\, Q/(\sqrt{n}-1),$$

as soon as $n \geq C_5'' \log(4n/\delta)$, where C_5, C_5' and C_5'' are positive constants depending on j_1, m, M, $\underline{m}$, $\overline{M}$, and B solely.

(ii) For any $\delta \in (0,1)$, we have with probability at least $1-\delta$:

$$L_{\mathcal{S}}(f_{\widehat{\Lambda}_d}) - L_{\mathcal{S}}^* \leq C_6\, d^2\, \sqrt{\log(4n/\delta)/n} + C_6'\, d^2\, Q/(\sqrt{n}-1),$$

as soon as $n \geq C_6'' \log(4n/\delta)$, where C_6, C_6' and C_6'' are positive constants depending on j_1, m, M, $\underline{m}$, $\overline{M}$, and B solely.

Assertion (i) can be proved by exploiting the bounds obtained in Section 3.3 in Chapter 3 combined with Proposition 4.7, while the upper confidence bound for the excess of risk stated in Assertion (ii) can be deduced from the latter by noticing that, with probability one, the excess of integrated quadratic risk can be written as follows:

$$L_{\mathcal{S}}(f_{\widehat{\Lambda}_d}) - L_{\mathcal{S}}^* = \int_{s\in\mathcal{S}} \left(\widehat{\Lambda}_d(s)^\top \Sigma(\mathbf{s}_d) \widehat{\Lambda}_d(s) - \Lambda_d^*(s)^\top \Sigma(\mathbf{s}_d) \Lambda_d^*(s) - 2\, \mathbf{c}_d(s)^\top \left(\widehat{\Lambda}_d(s) - \Lambda_d^*(s) \right) \right) ds. \quad (4.11)$$

Shedding light onto the role of the technical assumptions made, we give a brief idea of the proof's approach.

Sketch of Proof. The proof of Assertion *(i)* essentially relies on the following bound

$$\sup_{s\in\mathcal{S}} \|\widehat{\Lambda}_d(s) - \Lambda_d^*(s)\| \leq \underbrace{|||\Sigma(\boldsymbol{s}_d)^{-1}|||}_{N_1} \underbrace{\sup_{s\in\mathcal{S}} \|\widehat{\boldsymbol{c}}_d(s) - \boldsymbol{c}_d(s)\|}_{N_2} + \underbrace{|||\widehat{\Sigma}(\boldsymbol{s}_d)^{-1} - \Sigma(\boldsymbol{s}_d)^{-1}|||}_{N_3} \underbrace{\sup_{s\in\mathcal{S}} \|\widehat{\boldsymbol{c}}_d(s)\|}_{N_4},$$

where

- For *term N_1:* Assumption 3.5 (all eigenvalues of $\Sigma(\mathbf{s}_d)$ are strictly positives) in Chapter 3 and Assumption 4.6 (spectrum of $\Sigma(\mathbf{s}_d)$ is lower bounded by $\underline{m}$), imply that one has $|||\Sigma(\mathbf{s}_d)^{-1}||| \leq \underline{m}^{-1}$.
- A bound for *term N_2* can be deduced from the link between the max norm and the Euclidean norm, and the upper bound in Corollary 3.11 (Chapter 3).

- Refer to Proposition 4.7 Assertion *(ii)* for a bound with high probability of *term* N_3.
- *Term* N_4*:* From Corollary 3.11 and Assumption 3.10, we deduce that $\sup_{h \geq 0} |\widehat{c}(h)| < B$, with high probability.

Using Equation (4.11) and since the domain $\mathcal{S}$ is bounded, the remaining terms to study are

- As a consequence of Proposition 4.7 Assertion *(i)*, with probability at least $1 - \delta$, $\forall \delta \in (0, 1)$, the eigenvalues of $\widehat{\Sigma}(\mathbf{s}_d)$ are close to the eigenvalues of $\Sigma(\mathbf{s}_d)$. Thus, one can deduce the upper bound $|||\widehat{\Sigma}(\mathbf{s}_d)^{-1}||| \leq \underline{m}^{-1}$, with high probability.
- A bound for $\sup_{s \in \mathcal{S}} \|\mathbf{c}_d(s)\|$ is deduced from the link between the max norm and the Euclidean norm, together with Assumption 3.10: $\sup_{s \in \mathcal{S}} \|\mathbf{c}_d(s)\| \leq \sqrt{d} B$.

■

The detailed proof is given in Appendix B. These theoretical guarantees are illustrated by numerical results based on simulated/real spatial data in Sections 4.4 and 4.5. They clearly show that the prediction errors of the non-parametric empirical kriging method analyzed above get closer and closer to those of the theoretical one (based on the true covariance function) for a variety of spatial models, as the training size n increases. In Section 4.6, we investigate possible ways of extending these results to a more general framework by computing additional illustrative experiments.

Summary

Figure 4.1 shows a graphic summary of the proof. It points out how the theorem needs all the previous results presented in Chapter 3 (depicted in blue) to be established, as well as the role of each technical assumption (in green).

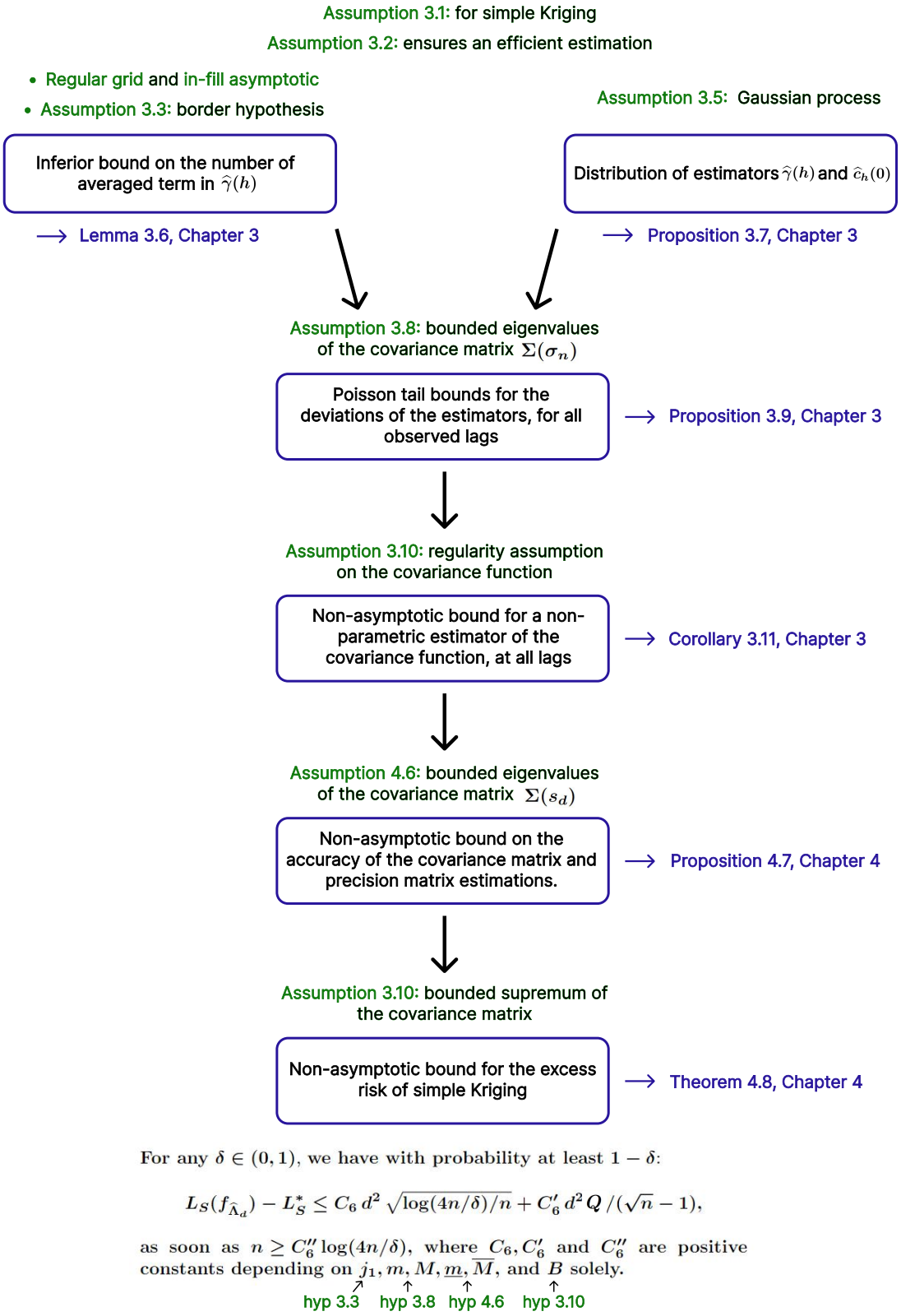


Figure 4.1: Summary of the contributions and technical assumptions.

4.4 Illustrative Experiments

The illustrative experiments of the theoretical analysis carried out in the previous section fits into the continuity of Section 3.4 in Chapter 3. Again, the experiments are available on GitHub.

We consider the two covariance models as in Section 3.4: an isotropic truncated power law covariance function defined as in (3.7), that satisfies all the assumptions involved in Theorem 4.8, and a Gaussian covariance function (3.8), which does not satisfy Assumption 3.3 but still vanishes quickly.

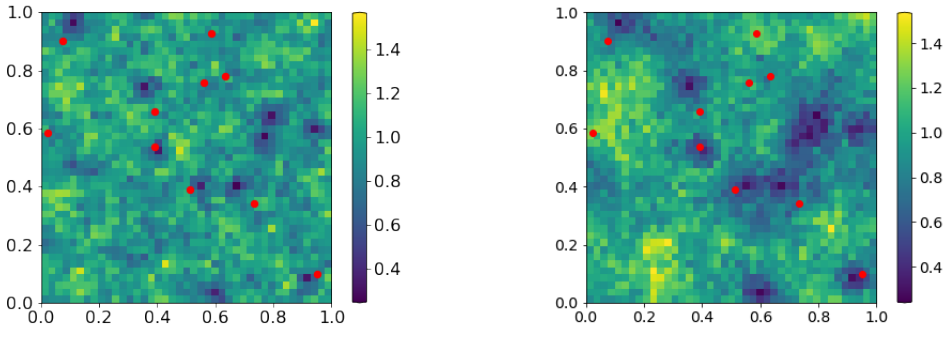
Table 4.1: Mean and standard deviation of the AMSE over 100 independent simulations of a Gaussian process with truncated power law (left) and Gaussian (right) covariance functions for theoretical and empirical Kriging with different values of θ (with $J = 3$, $N = 1681$ and $d = 10$).

TPL	Theoretical		Empirical	
θ	mean	std	mean	std
2.5	0.961	0.086	0.971	0.088
5	0.911	0.145	0.930	0.159
7.5	0.850	0.218	0.864	0.215
10	0.800	0.249	0.839	0.257

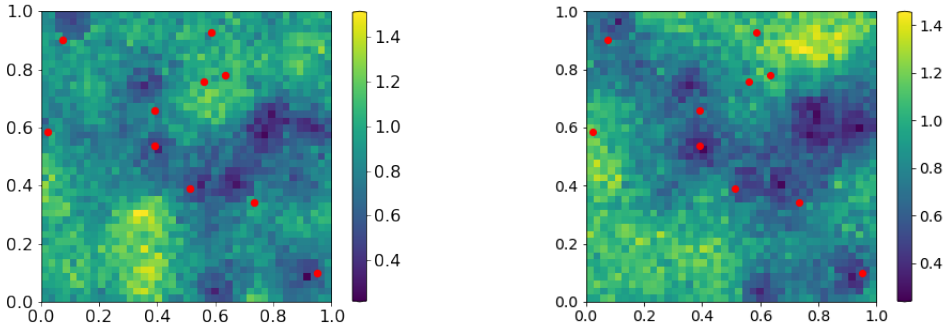
GAUSS	Theoretical		Empirical	
θ	mean	std	mean	std
2.5	0.891	0.196	0.899	0.208
5	0.635	0.269	0.686	0.340
7.5	0.421	0.257	0.703	1.498
10	0.247	0.202	0.536	1.048

Based on the covariance estimates obtained in Section 3.4 in Chapter 3 from observations of a regular grid of size $n = (1+2^J)^2$, we now consider the simple Kriging problem from a predictive point of view, and simulate a new independent realization of $\mathbf{X}$. The latter is observed at d sites $(s_1, \cdots, s_d)$, randomly selected over the spatial domain $\mathcal{S}$. As formulated in Section 4.2, the goal is to predict the value $\mathbf{X}_s$ taken by $\mathbf{X}$ at any site $s \in \mathcal{S}$ based on the d observations. Regarding the empirical evaluation of the predictive accuracy, the spatial domain $\mathcal{S}$ is (regularly) discretized: the goal is to predict the value taken by $\mathbf{X}$ at the corresponding $N \geq 1$ sites $s'_1, \ldots, s'_N$ in $\mathcal{S}$. In compliance with the methodology analyzed in Section 4.3, the predictive mapping is constructed by means of the plug-in technique from the covariance vector and the covariance matrix estimators defined in (4.6) and (4.7), see (4.8). The prediction error being the expected squared difference between the predicted random field and the true random field integrated (respectively, averaged) over (respectively, the discretized version of) the spatial domain $\mathcal{S}$, we performed 100 replications of the experiment, each one involving one simulation $\mathbf{X}'$ for the training step and one simulation $\mathbf{X}$ for the prediction test, the locations $(\sigma_1, \cdots, \sigma_n)$, $(s_1, \cdots, s_d)$ and $(s'_1, \ldots, s'_N)$ remaining fixed. In order to compare empirically the empirical Kriging method analyzed in the previous section to the 'Oracle' method based on the true covariance function (theoretical Kriging), the prediction techniques have been thus applied 100 times, so that 2×100 prediction maps have been obtained. For each replication $(\mathbf{X}', \mathbf{X})$ of the experiment, the (spatial) average over the discretized version of $\mathcal{S}$ of the MSE (2.5) (see Chapter 2) has been evaluated,

$$AMSE = \frac{1}{N} \sum_{t=1}^{N} (f(s'_t) - \mathbf{X}_{s'_t})^2, \tag{4.12}$$



(a) $\theta = 2.5$ (b) $\theta = 5$

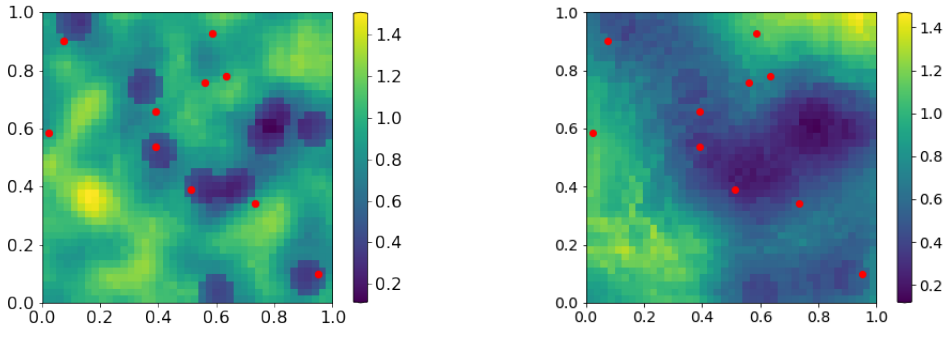


(c) $\theta = 7.5$ (d) $\theta = 10$

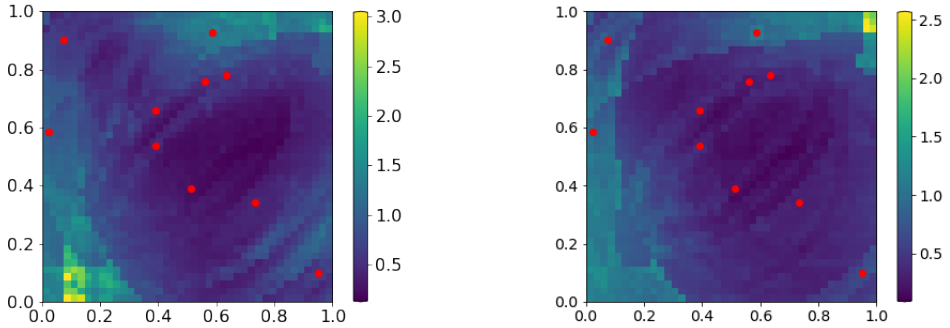
Figure 4.2: MSE maps over 100 realizations of a Gaussian process with truncated power law covariance function for the empirical Kriging predictor with different values of θ ($J = 3$, $N = 1681$, and $d = 10$).

for $f = f_{\Lambda_d^*}$ (theoretical Kriging) and $f = f_{\widehat{\Lambda}_d}$ (empirical Kriging). The mean and standard deviation of (4.12) have been computed over the 100 replications. To observe the effects of several parameters on the performance of the Kriging method, the experience was carried out for different sizes of the dyadic grid and different values of the correlation length θ ($\{2.5, 5, 7.5, 10\}$), the parameter used in the definition of the instrumental covariance functions. Note that, for the truncated power law covariance function, the parameter θ is linked to the parameter j_1 in Assumption 3.3: $j_1 = -\log(\sqrt{2}-\theta/\sqrt{n})/\log(2)$. The corresponding bound $h \geq \sqrt{2}-2^{-j_1}$ of Assumption 3.3 for the different values of θ are $\{0.061, 0.122, 0.183, 0.244\}$ respectively. The training dataset was drawn on a dyadic grid at scale $J = 3$ (with $n = 81$ observations), whereas the number of input observations for the prediction is equal to $d = 10$. The results for the two spatial models are displayed in Table 4.1: the mean and the standard deviation (std) of the AMSE, over the 100 replications, are given for different values of θ.

For the truncated power law covariance function, observe that the mean of the AMSE decreases slowly with the correlation length θ (and, by definition, with j_1), whereas the standard deviation increases slowly, for both the theoretical and empirical Kriging. Furthermore, in keeping with our theoretical results, the difference between the two AMSE (*i.e.* the excess of pointwise risk (4.11)) is small for all values of θ. The same observations can be made for the Gaussian covariance function, where the standard



(a) $\theta = 2.5$ (b) $\theta = 5$



(c) $\theta = 7.5$ (d) $\theta = 10$

Figure 4.3: MSE maps of over 100 realizations of a Gaussian process with Gaussian covariance function for the empirical Kriging predictor with different values of θ ($J = 3$, $N = 1681$, and $d = 10$).

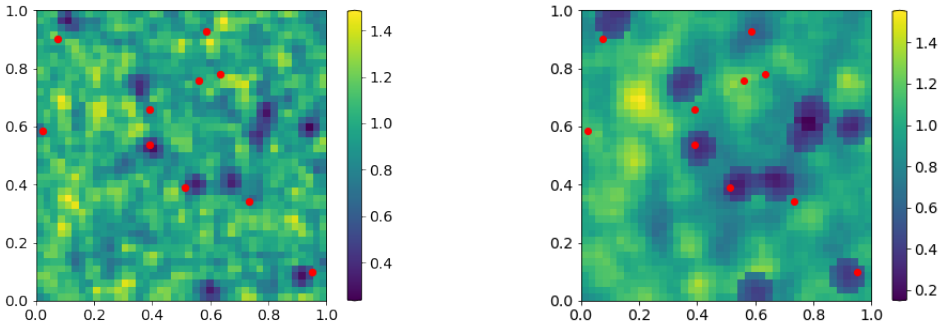
Table 4.2: Mean and standard deviation of the AMSE over 100 independent simulations of a Gaussian process with truncated power law (left) and Gaussian (right) covariance functions for theoretical and empirical Kriging with different values of θ (with $J = 4$, $N = 2401$ and $d = 10$).

TPL	Theoretical		Empirical	
θ	mean	std	mean	std
2.5	0.975	0.079	0.976	0.079
5	0.927	0.131	0.928	0.131
7.5	0.861	0.156	0.874	0.150
10	0.815	0.210	0.841	0.220

GAUSS	Theoretical		Empirical	
θ	mean	std	mean	std
2.5	0.913	0.167	0.890	0.167
5	0.708	0.272	0.745	0.306
7.5	0.529	0.263	0.582	0.288
10	0.312	0.188	1.675	12.085

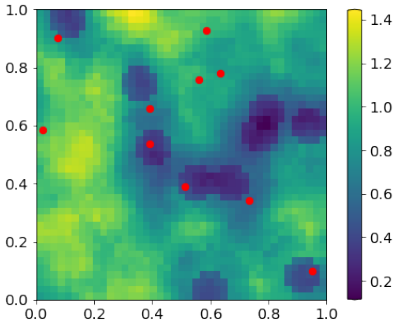
deviation is larger than for the other covariance model, the mean is slightly smaller however.

To better understand the spatial structure of these errors, the maps of the mean squared errors for the empirical Kriging predictors are depicted in Figures 4.2 and 4.3. The observed sites $(s_1, \cdots, s_d)$ are represented in red. Observe first that, as underlined in Remark 2.22, Kriging is an exact interpolator (the error is null at the ob-

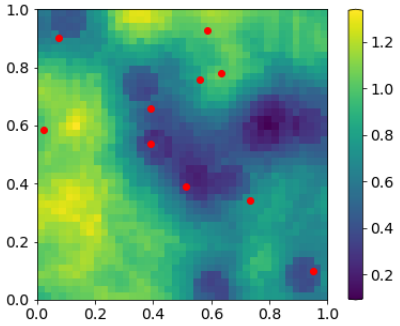


(a) $\theta = 2.5$

(b) $\theta = 5$



(c) $\theta = 7.5$

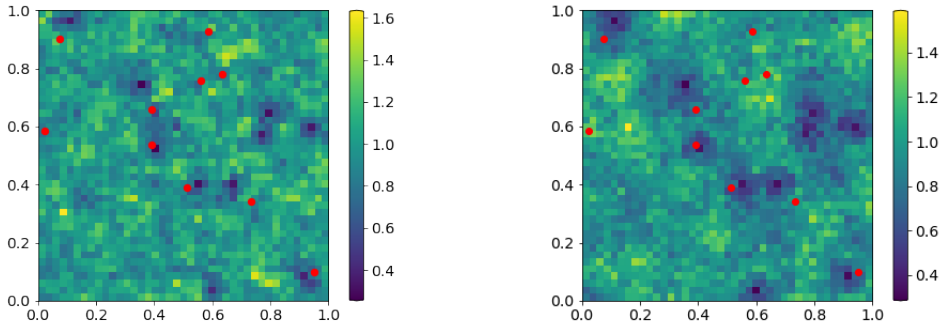


(d) $\theta = 10$

Figure 4.4: MSE maps over 100 realizations of a Gaussian process with cubic covariance function for the empirical Kriging predictor with different values of θ ($J = 3$ and $d = 10$).

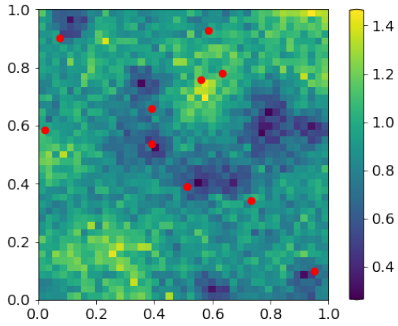
servation sites $(s_1, \cdots, s_d)$). For the truncated power law model, the complete maps of the MSE for different values of θ are comparable, with the same order of magnitude for the errors and similar location of the smallest errors. In the case of the Gaussian model, the results in Fig. 4.3c and 4.3d exhibit some border effects, where a higher AMSE can be observed near the boundaries of the spatial domain. This difference can be easily explained. For the truncated power law model, there is an absence of correlation between locations that are distant enough and, as can be seen in Figure 3.3a, the threshold is reached quickly. In contrast, for the Gaussian model, the covariance function vanishes only for large distances, especially for the empirical version, which fails to capture the correlation length value as noticed in the previous subsection. The predictive performance of the empirical Kriging method has been also evaluated for a larger number n of training observations, *i.e.* for a denser dyadic grid of scale $J = 4$ ($n = 289$). Table 4.2 (left) shows that the results for the truncated power law model are comparable to those in the case $J = 3$. This is also the case for the Gaussian model except that, when the covariance function is unknown, the mean and standard deviation of the AMSE become larger for $\theta = 10$ (see Table 4.2, right).

The numerical results for the Gaussian covariance function, which does not satisfy Assumption 3.3 but quickly vanishes, suggest that the validity framework of the empirical Kriging method can be extended.

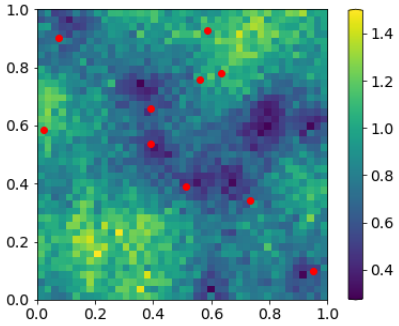


(a) $\theta = 2.5$

(b) $\theta = 5$



(c) $\theta = 7.5$



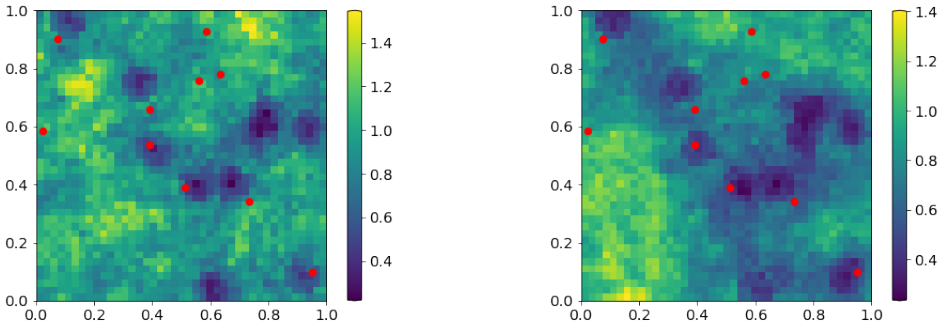
(d) $\theta = 10$

Figure 4.5: MSE maps over 100 realizations of a Gaussian process with spherical covariance function for the empirical Kriging predictor with different values of θ ($J = 3$ and $d = 10$).

Additional Covariance Models. We recall the additional covariance models considered in Section 3.4 in Chapter 3: the cubic (3.9) and spherical (3.10) covariance functions that satisfy all the assumptions; and the exponential (3.11) and Matern (3.12) (with varying smoothness parameter ν_m) covariance functions, that do not satisfy Assumption 3.3.

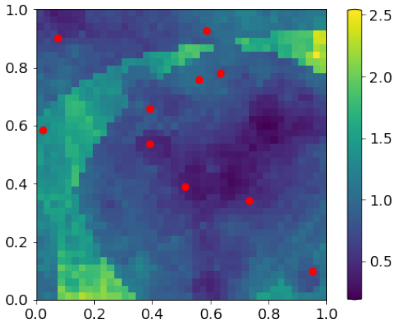
As for the previous covariance models, the maps of all mean squared errors for the empirical Kriging predictor were computed over 100 realizations of a Gaussian process, and the results are depicted in Figures 4.4 to 4.8, with varying values for the correlation length $\theta \in \{2.5, 5, 7.5, 10\}$. For the cubic covariance model, it can be noticed in Figure 4.4 that the MSE map seems to become smoother when θ increases. Still, as in the case of the spherical covariance model (see Figure 4.5), the complete maps are similar for all values of θ regarding the error scale and the allocation of the local area with small errors.

For the three covariance functions that do not satisfy Assumption 3.3, the same observations can be made: as θ grows, some border effects can be seen on the boundaries of the window of observation, and the error scale becomes larger. This lines up with the results obtained for the Gaussian covariance model.

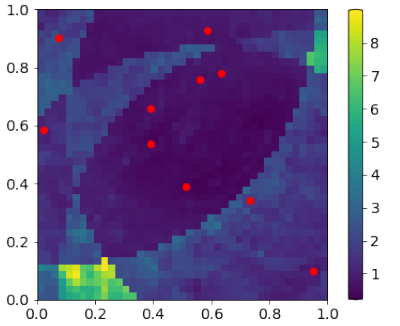


(a) $\theta = 2.5$

(b) $\theta = 5$



(c) $\theta = 7.5$

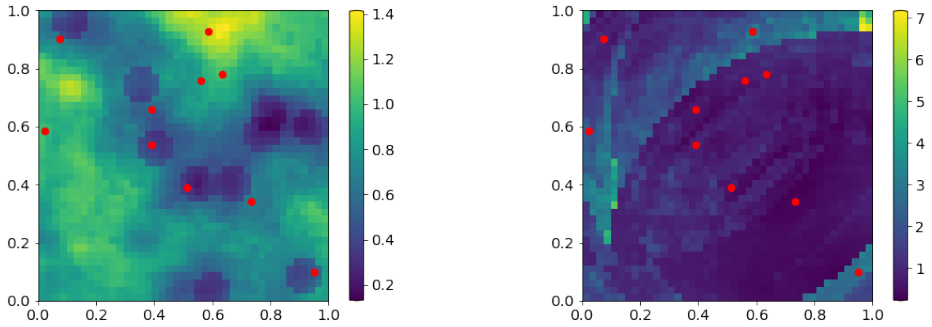


(d) $\theta = 10$

Figure 4.6: MSE maps over 100 realizations of a Gaussian process with exponential covariance function for the empirical Kriging predictor with different values of θ ($J = 3$ and $d = 10$).

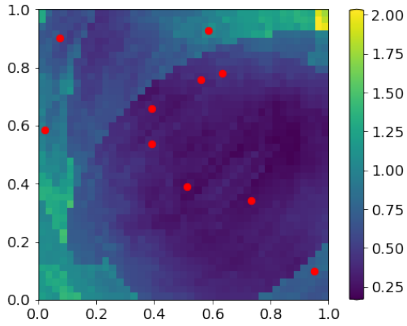
Finally, the AMSE for both theoretical and empirical predictive mappings for all the covariance functions was calculated. The results (the mean and the standard deviation for the 100 independent simulations for different values of θ) are presented in Table 4.3. For the cubic and spherical models, we observe the same trend as for the truncated power law model: as the correlation length θ increases, for both the theoretical and empirical Kriging, the mean of the AMSE decreases while the standard deviation increases slowly, and the excess risk is small for all values of θ. For the exponential covariance model, the mean and the standard deviation do not have a constant trend: first, the mean decreases while the standard deviation increases slightly, then they both increase more significantly, with a large standard deviation when $\theta = 10$. A significant increase of both the mean and standard deviation of the AMSE can also be observed for the Matern model, especially when $\nu_m = 5/2$: the mean and the standard deviation become large when $\theta = 7.5$ (see Table 4.3).

The results on additional covariance functions lead to the same conclusions about the truncated power law and the Gaussian covariance functions: the predictive method may perform well, even if Assumption 3.3 is slightly violated.

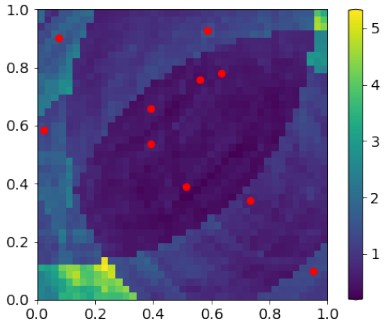


(a) $\theta = 2.5$

(b) $\theta = 5$



(c) $\theta = 7.5$



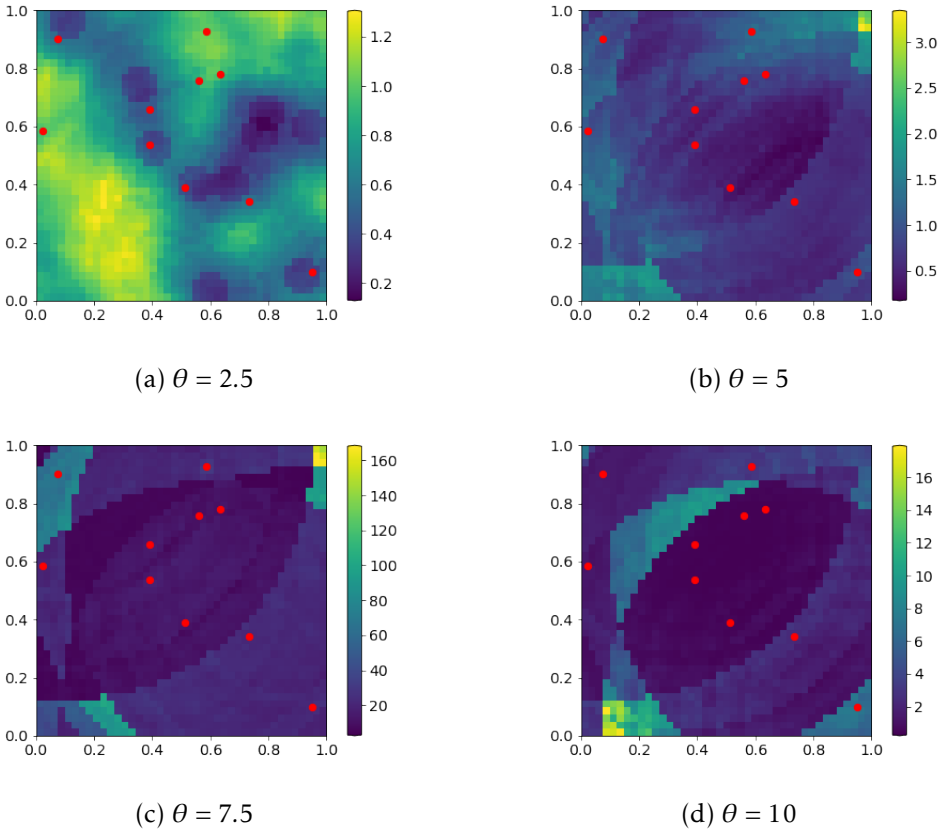
(d) $\theta = 10$

Figure 4.7: MSE maps over 100 realizations of a Gaussian process with Matern covariance ($\nu_m = 3/2$) for the empirical Kriging predictor with different values of θ ($J = 3$ and $d = 10$).

4.5 Application to Real Data – Mean Daily Temperature in France

For the sake of completeness, prediction via simple empirical Kriging has also been examined on real spatial observations, on datasets available on the web portal DRIAS, which provides the mean daily temperature in France (in Kelvin), observed on a regular grid, from 1951 to 2005.

The position (latitude and longitude) of the grid points are in decimal degrees (WGS84). The datasets that are used in this study are square grids of a total of 2401 point locations (referred to as the spatial domain $\mathcal{S}$), during the three months of summer (June, July, and August, for a total of 92 days) of the years 2004 and 2005 (refer to Figure 1.1 in the Introduction (Chapter 1) for the sampled square grid). The square grid is obtained directly on the web portal by selecting the desired points on the grid, in such a way that the complete grid is of the wanted dimension 49×49. Notice that the temporal dimension of the data is ignored here, it is assumed that the daily observations are independent from one year to the next. Under this simplifying hypothesis, we consider that a number of realizations of the phenomenon are available, large enough for computing significant AMSE's. The dataset of the year 2004 is used as training samples: for each day, $n = 289$ sites are observed, forming a dyadic grid at scale $J = 4$. These observations are used in order to estimate the (supposedly isotropic) covariance



(a) $\theta = 2.5$ (b) $\theta = 5$

(c) $\theta = 7.5$ (d) $\theta = 10$

Figure 4.8: MSE maps over 100 realizations of a Gaussian process with Matern covariance ($\nu_m = 5/2$) for the empirical Kriging predictor with different values of θ ($J = 3$ and $d = 10$).

function of the random field by means of the non-parametric statistics (3.2) studied in Chapter 3. Then, respectively on the same days of the year 2005, $d = 10$ sites are randomly selected over the spatial domain. The goal is to predict the value $\mathbf{X}_s$ taken by $\mathbf{X}$ at any unobserved site $s \in \mathcal{S}$ (*i.e.* predict the mean temperature on the same day for all unobserved locations) based on the d input observations and the estimated covariance function. The experiment has been performed 92 times (training samples from 2004 and data from 2005 for the prediction step).

We point out that, since the mean is unknown, we opted for the Ordinary Kriging variant: using the semi-variogram function rather than the covariance (see Chiles and Delfiner, 1999, Section 3, and Cressie, 1993, subsection 3.2, for a presentation of the estimator and of the method), Ordinary Kriging allows us to perform the prediction without any information about the mean of the random process (see subsection 2.1.6 in Chapter 2). The theoretical results of Section 4.3 easily extends to the case of Ordinary Kriging: indeed, Propositions 3.7 and 3.9 in Chapter 3 are based on the semi-variogram estimation and the following results up to Theorem 4.8 can be straightforwardly extended to the Ordinary Kriging predictor.

For the parametric Kriging method, the truncated power law model (see Equation (3.7)) has been selected, among several covariance models. Here we set $\theta = j_1$ (the parameter from Assumption 3.3): we fixed the value of θ high enough, so that a large number of correlations are taken into account. This means that the covariances for

Table 4.3: Mean and standard deviation of the AMSE over 100 independent simulations of a Gaussian process with cubic (top left), spherical (top right), exponential (center), Matern with $\nu_m = 3/2$ (bottom left) and $\nu_m = 5/2$ (bottom right) covariance functions for theoretical and empirical Kriging with different values of θ (with $J = 3$, $N = 1681$ and $d = 10$).

CUB	Theoretical		Empirical	
θ	mean	std	mean	std
2.5	0.979	0.091	0.983	0.088
5	0.920	0.147	0.934	0.150
7.5	0.847	0.232	0.869	0.241
10	0.748	0.294	0.778	0.294

SPH	Theoretical		Empirical	
θ	mean	std	mean	std
2.5	0.984	0.079	0.997	0.081
5	0.927	0.121	0.941	0.122
7.5	0.870	0.188	0.890	0.195
10	0.809	0.183	0.846	0.199

EXP	Theoretical		Empirical	
θ	mean	std	mean	std
2.5	0.890	0.178	0.923	0.184
5	0.679	0.180	0.781	0.230
7.5	0.548	0.158	0.922	1.976
10	0.439	0.150	1.481	5.393

3/2	Theoretical		Empirical	
θ	mean	std	mean	std
2.5	0.791	0.308	0.769	0.281
5	0.464	0.266	1.167	5.841
7.5	0.260	0.174	0.548	0.793
10	0.163	0.130	1.017	3.105

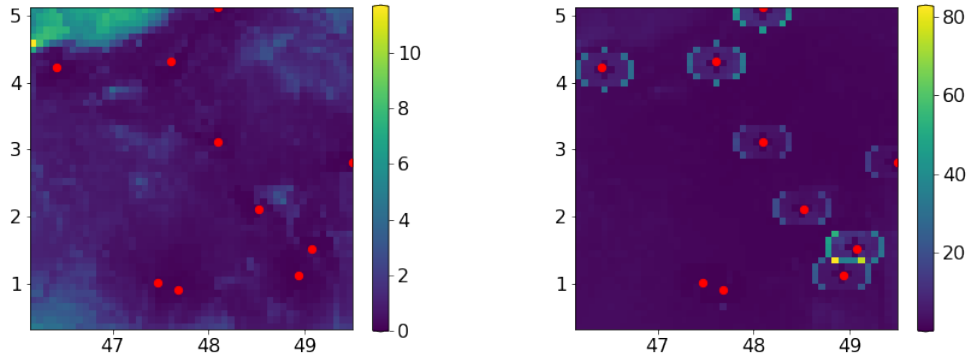
5/2	Theoretical		Empirical	
θ	mean	std	mean	std
2.5	0.791	0.269	0.756	0.249
5	0.471	0.277	0.800	1.842
7.5	0.236	0.191	20.945	179.206
10	0.129	0.106	2.455	18.476

almost all lags $h \in \mathcal{H}_n$ are involved in the computation of the parametric Kriging predictor. Yet, it is not surprising, in the case of temperature data, to obtain a better accuracy using almost all covariances, since the correlation is strong between all pairs of locations. The results are displayed in Table 4.4. For the non-parametric Kriging method, the parameter ν introduced in Lemma 3.6 (see Chapter 3) is set to 0.35 in order to use most of the observed distances $h \in \mathcal{H}_n$ for the covariance function estimation. Notice that the mean error, and the standard deviation as well, are low (see Table 4.4). The results are encouraging and corroborate the theoretical guarantees established.

Table 4.4: Mean and standard deviation (std) of all AMSE for parametric and non-parametric Kriging on Real Data (with $J = 4$, $N = 2401$, and $d = 10$).

	Parametric		Non-parametric	
	mean	std	mean	std
$J = 4$	2.581	0.564	2.944	1.931

The maps of all mean squared errors for parametric and non-parametric Kriging predictors are depicted in Figure 4.9.



(a) Parametric (b) Non-parametric

Figure 4.9: Complete maps of all MSE on Real Data on a dyadic grid of observations at scale $J = 4$ (with $N = 2401$ and $d = 10$).

As for Kriging applied to simulated data, the exact interpolator property (Remark 2.22 in Chapter 2) is verified. Notice that, for parametric Kriging, some border effects can be observed ; while, for non-parametric Kriging, there is a presence of local area with higher error in form of circles at a certain distance of the observed locations, where the mean error is higher and seems null everywhere else on the spatial domain. Still, the results on real data are encouraging to extend our theoretical results to a more general framework. Indeed, recall that these real data are irregular and violate some of the made assumptions in Chapter 3: Assumption 3.1 is not verified since the mean of the temperatures over several locations in France is not null; Assumption 3.3 is not satisfied, as discussed in the choice of the value of the parameter θ.

Though it is beyond the scope of this work, the statistical modeling and predictive analysis of such real data could be naturally refined in many ways, taking into account anisotropy and/or the temporal structure in particular. However, the only goal pursued here is to show that a simplistic application of the non-parametric empirical Kriging prediction method may perform well and can be competitive compared to a more rigid method based on a preliminary parametric modeling of the covariance structure.

4.6 Illustrative Experiments of Possible Extensions

In this section, we present additional illustrative experiments. The role of these experiments is to investigate possible future extensions of the present work. First, we present the case where the d observation points are taken from different configurations than the random uniform procedure used so far. Next, we discuss the use of anisotropic covariance models (thus relaxing Assumption 3.2). And finally, we investigate the setting of irregular grids for the training sample.

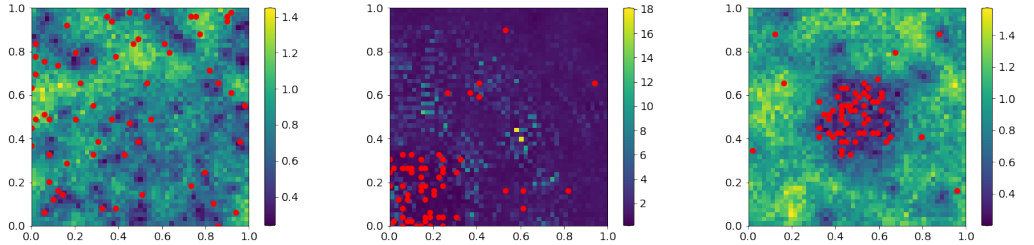
4.6.1 Extension to Different Configurations of the Observations' Locations

One may be interested in the influence of the configuration of the observation points $s_1, \cdots, s_d$ on the performance of the Kriging predictor. The results are presented here for two extreme situations: the first, called *Corner* (C), happens when the major number of observations are taken randomly in a small sub-region of the spatial domain $\mathcal{S}$, defined as one of the corners of the domain; for the second one, called *Ring* (R), the major number of observations are sampled randomly in a circle of center equal to the middle of the spatial domain $\mathcal{S}$.

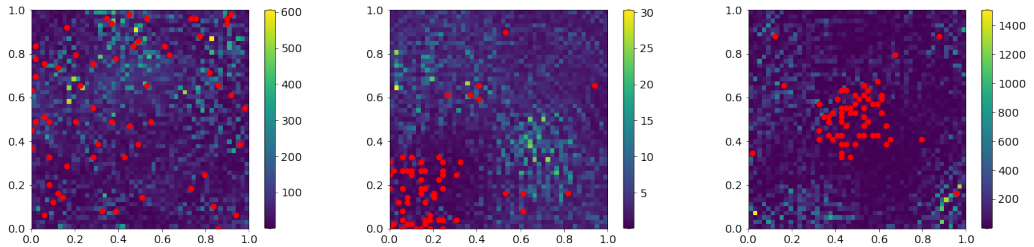
Table 4.5: Mean and standard deviation (std) of the AMSE on 100 independent simulations of a Gaussian process with truncated power law (left) and Gaussian (right) covariance functions for theoretical and empirical Kriging with different configurations of the observations' locations (where U: *Uniform* ; C: *Corner* ; R: *Ring*) (with $J = 4$, $N = 1681$, $d = 60$ and $\theta = 5$).

TPL	Theoretical		Empirical	
	mean	std	mean	std
U	0.747	0.113	0.797	0.120
C	0.850	0.126	1.596	6.899
R	0.866	0.126	0.917	0.140

GAU	Theoretical		Empirical	
	mean	std	mean	std
U	0.186	0.107	67.814	476.812
C	0.598	0.270	3.676	18.225
R	0.517	0.182	84.840	815.206



(a) Uniform (b) Corner (c) Ring



(d) Uniform (e) Corner (f) Ring

Figure 4.10: Complete maps of all MSE on 100 realizations of a Gaussian process with truncated power law covariance (top) and Gaussian (bottom) covariance functions for the empirical Kriging predictor with different configurations of the observations' locations $s_1, \cdots, s_d$ (with $J = 4$, $N = 1681$, $d = 60$ and $\theta = 5$).

The same procedure as in Sections 3.4 and 4.4 is then applied for the truncated power law (3.7) and the Gaussian (3.8) covariance models, with the following setting: the number of training observations is equal to $n = 289$ (dyadic scale $J = 4$); the total number of sampled locations for the prediction is fixed to $d = 60$, where 50 are taken in the sub-region of interest and the others 10 in the remaining area of the domain; and the correlation length is fixed at $\theta = 5$. The mean and the standard deviation of all AMSE are presented in Table 4.5 and the complete maps of all MSE are displayed in Figure 4.10, where *Uniform* (U) stands for the selection of the observations' locations as before, using a random uniform procedure over the spatial domain S. For the truncated power law model, the results in Table 4.5 (left) are similar when using the *Uniform* procedure or the *Ring* procedure and the errors are close for the theoretical and the empirical methods. In the *Corner* case, the mean AMSE for the empirical Kriging is larger, and especially its standard deviation increases. Looking at the corresponding complete map in Figure 4.10b, it can be seen that the predictor seems to succeed for the point locations near the observations but fails at some locations far from any observed sample (with a large error, going up to 18, as shown by the large error scale). This is a direct consequence of the fact that couples of point locations that are at a large distance from one another are under-represented in this setting. When looking at the Gaussian model results in Table 4.5 (right), it's obvious that this model is strongly linked to the configuration of the observations: the mean and the standard deviation for empirical Kriging are significantly larger than when the true covariance function is known. Indeed, let us observe that for the theoretical Kriging method, the mean and the standard deviation are more or less the same for the three configurations, whereas, for the empirical Kriging method, these two values rise abruptly when the observations are taken mainly in a circle. Thanks to Figure 4.10f, one can notice that the point locations that make the mean error explodes are located in the boundaries of the spatial domain, far from any observed sample (the maximum error value is more than 1400). Other observations that can be made with these results are on the influence of the number d of observed samples for the prediction step. When using the truncated power law model, the results for the *Uniform* configuration in Table 4.5 (left) are similar to the mean and standard deviation in Table 4.2 (left) when $\theta = 5$. So, the size d of observations does not seem to have an impact on the performance of the Kriging estimator. In contrast, for the Gaussian model, when the empirical version is used, the results significantly change between Table 4.5 (right) when the configuration is *Uniform* and Table 4.2 (right) when $\theta = 5$, with a strong increase for both the mean and the standard deviation in the case where $d = 60$.

Therefore, it can be of interest to explore how the performance of Kriging can be affected by a variation in the observations' locations.

4.6.2 Anisotropic Covariance Function

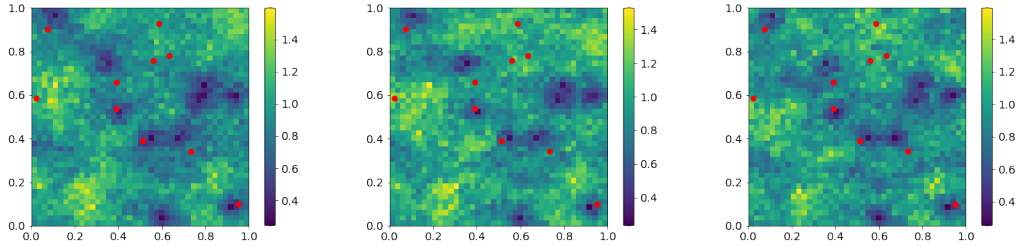
In this section, we study the role of the isotropy assumption for the covariance function (Assumption 3.2). Based on the truncated power law (3.7) and the Gaussian (3.8) covariance models, we apply the same procedure as in Sections 3.4 and 4.4 to the case where anisotropic covariance functions are selected. The `gstools` library allows one to simulate an anisotropic covariance function with varying anisotropic ratios α, where $\alpha = 1$ corresponds to the isotropy situation.

Table 4.6: Mean and standard deviation of the AMSE over 100 independent simulations of a Gaussian process with truncated power law (left) and Gaussian (right) covariance functions for theoretical and empirical Kriging with different values of the anisotropic ratio α (with $J = 3$, $N = 1681$, $d = 10$ and $\theta = 5$).

TPL	Theoretical		Empirical	
α	mean	std	mean	std
0.9	0.901	0.141	0.934	0.142
0.8	0.925	0.139	0.951	0.135
0.7	0.919	0.119	0.949	0.121

GAUSS	Theoretical		Empirical	
α	mean	std	mean	std
0.9	0.428	0.173	0.722	0.354
0.8	0.486	0.186	0.750	0.315
0.7	0.579	0.216	0.784	0.294

The same setting is used in order to compare the results with those in Section 4.4: the covariance estimation is done thanks to a training dataset, observed at $n = 81$ sites ($J = 3$) and the prediction over the whole spatial domain is computed based on $d = 10$ observations. We fix the value of the correlation length at $\theta = 5$, and repeat the experiments for different degrees of anisotropy $\alpha \in \{0.9, 0.8, 0.7\}$.



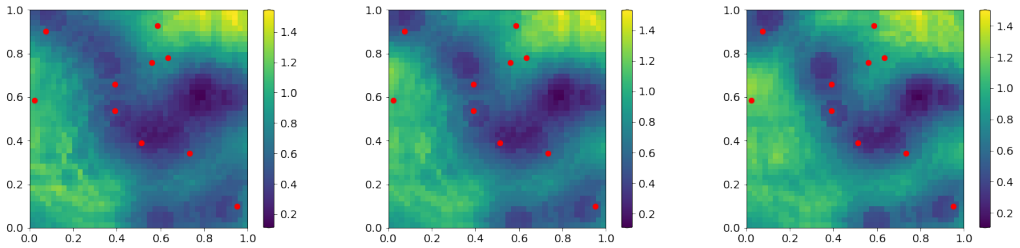
(a) $\alpha = 0.9$ (b) $\alpha = 0.8$ (c) $\alpha = 0.7$

Figure 4.11: MSE maps of over 100 realizations of a Gaussian process with truncated power law covariance function for the empirical Kriging predictor with different values of the anisotropic ratio α ($J = 3$, $N = 1681$, $d = 10$ and $\theta = 5$).

The mean and standard deviation of the AMSE computed over the 100 independent simulations of a Gaussian process, for both covariance models, are shown in Table 4.6. We observe that for both models, for the empirical Kriging prediction, the mean increases slightly when α decreases (so when the covariance function becomes more anisotropic), while the standard deviation decreases (see Table 4.1, when $\theta = 5$ for a comparison with the isotropic case).

Figures 4.11 and 4.12 (for the truncated power law and the Gaussian model, respectively), show the complete MSE maps over 100 realizations. It can be observed that the structure of the errors for both covariance models is similar to the maps obtained using an isotropic covariance function, with the same scale error and the same local area with small errors (see Figures 4.2b and 4.3b, respectively).

These results, which show that the prediction methodology is robust with respect to slight departures from isotropy, encourage us to relax also the isotropic assumption, in some future work.



(a) $\alpha = 0.9$ (b) $\alpha = 0.8$ (c) $\alpha = 0.7$

Figure 4.12: MSE maps of over 100 realizations of a Gaussian process with Gaussian covariance function for the empirical Kriging predictor with different values of the anisotropic ratio α ($J = 3$, $N = 1681$, $d = 10$ and $\theta = 5$).

4.6.3 Irregular Grids

With the purpose of extending the theoretical results to a more general framework, we present the numerical results within a different setting: we consider the case where the simulations are done over an irregular grid. In this new setting, the realization of the random field used for the non-parametric covariance estimation (the training spatial dataset) is no longer observed at n sites forming a dyadic grid. Instead of assuming that the training observations are made on a regular dyadic grid, we make the hypothesis that we have access only to a restricted number of these observations. The irregular grids are generated from regular grids using Bernoulli sampling, with varying probability p ($\{0.8, 0.6, 0.4\}$) of observing a spatial site. The number of observed locations is $\{65, 50, 35\}$ (respectively).

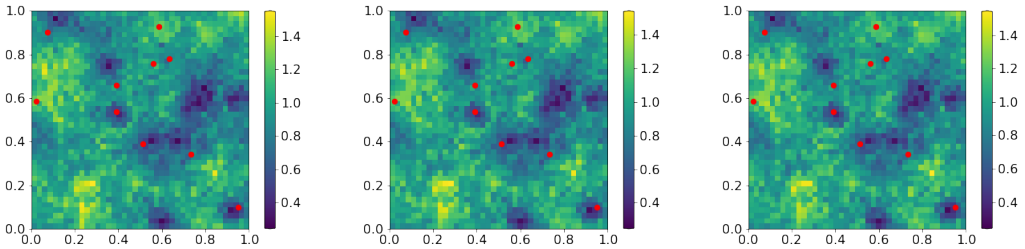
Table 4.7: Mean and standard deviation of the AMSE over 100 independent simulations of a Gaussian process with truncated power law (left) and Gaussian (right) covariance functions for empirical Kriging with different probabilities p for the Bernoulli sampling (with $N = 1681$, $d = 10$ and $\theta = 5$).

TPL	Empirical	
p	mean	std
0.8	0.932	0.159
0.6	0.934	0.156
0.4	0.936	0.154

GAUSS	Empirical	
p	mean	std
0.8	0.686	0.320
0.6	0.663	0.267
0.4	0.687	0.370

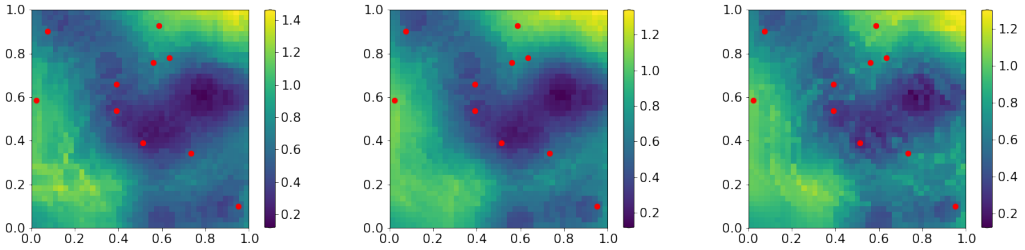
The estimation of the covariance function is computed as before, using Equation (3.2). Since this framework could imply (likely but not surely) situations where for some h previously observed on the complete dyadic grid, are not present anymore, we handle these cases by skipping the estimation of the covariance function for these values and simply applying the 1-NN estimator in the prediction step (as stated in Section 3.2). For the simple Kriging prediction, we use the same independent realization of $\mathbf{X}$ observed at $d = 10$ sites. We present the results only for the truncated power law and the Gaussian covariance models, for $\theta = 5$: the mean and the standard deviation of the AMSE are displayed in Table 4.7. Note that, the sampled locations over the irregular grid are fixed for the 100 replications of the experiment, as well as the d locations

for the prediction test. For both covariance models, we observe that, as fewer and fewer observations are selected, the AMSE and the standard deviation do not vary much from the regular grid situation. This shows that both models are robust against irregular sampling for the covariance estimation.



(a) $p = 0.8$ (b) $p = 0.6$ (c) $p = 0.4$

Figure 4.13: MSE maps of over 100 realizations of a Gaussian process with truncated power law covariance function for the empirical Kriging predictor with different probabilities p for the Bernoulli sampling ($N = 1681$, $d = 10$ and $\theta = 5$).



(a) $p = 0.8$ (b) $p = 0.6$ (c) $p = 0.4$

Figure 4.14: MSE maps of over 100 realizations of a Gaussian process with Gaussian covariance function for the empirical Kriging predictor with different probabilities p for the Bernoulli sampling ($N = 1681$, $d = 10$ and $\theta = 5$).

Regarding the structure of these errors for the truncated power law covariance model, as it can be seen in the maps of the mean squared errors in Figure 4.13, for all the values of p, the results are quite similar to the case where the observations are taken on a regular dyadic grid (see Figure 4.2b). For the Gaussian model, for $p = 0.8$ and $p = 0.6$, we recognize the same errors' structure as in Figure 4.3b with some border effects. Still, when p decreases, the errors seem to be expanded over the spatial domain.

As all previous results, the ones on the irregular sampling setup may be a motivation to extend the theoretical study to a more general framework, including new grid of observations.

4.7 Conclusion

In this chapter, we have proposed a statistical learning view of simple Kriging, which in the literature is usually addressed using a parametric and asymptotic approach. The major difficulty in analyzing this predictive problem lies in the complex dependence structure generally exhibited by spatial data. As explained at length, an empir-

ical version of the optimal simple Kriging rule (minimizing the MSE integrated over the spatial domain) can be constructed by means of a nonparametric estimator of the covariance function in a plug-in fashion. It is also shown that the predictive rule thus built can be viewed as a minimizer of the empirical counterpart of the risk, based on the covariance function estimator. We have developed a novel theoretical framework offering non-asymptotic guarantees for empirical simple Kriging rules in the form of non-asymptotic bounds for the integrated MSE in a classic in-fill setup, stipulating that the $n \geq 1$ sites at which the stationary isotropic Gaussian field under study is observed form a denser and denser regular grid. The learning rate bounds are of order $O_{\mathbb{P}}(1/\sqrt{n})$. To the best of our knowledge, these are the first results of this nature. These results allow us to overcome our **Challenge 3** (stated in Section 1.2 in Chapter 1).

Since the aim of this chapter is to elucidate the key concepts for achieving generalization guarantees in the spatial context, we employ several simplifying technical assumptions. However, some of the assumptions made to obtain our results may seem restrictive. Indeed, the regular grid assumption for the training data prevents the application of our theoretical results in some real situations. Assumption 3.3, stipulating that the covariance function is equal to zero after a given threshold, also limits the use of our results to a specific category of covariance models. Still, our additional numerical experiments on alternative settings motivate the extension of our results to more general frameworks, which will be the subject of future work.

Key points of this Chapter.

Technical Assumptions involved in this Chapter and their role

- **Assumptions 3.1–3.10:** see Chapter 3 for a discussion.
- **Assumption 4.6:** allows us to obtain non-asymptotic results for the precision matrix $\Sigma(\mathbf{s}_d)^{-1}$ (see Proposition 4.7, Assertion *(ii)*).
- See Figure 4.1 for a summary of the role of each assumption.

Main results of this Chapter

- Similarities between Kriging and KRR (Section 4.2).
- **Proposition 4.7:** we compute non-asymptotic bounds for the accuracy of the covariance matrix and precision matrix estimations.
- **Theorem 4.8:** we assess the generalization capacity at unobserved sites of the empirical simple Kriging predictor.

  $\Longrightarrow$ **MAIN RESULT:** learning rate bounds of order $O_{\mathbb{P}}(1/\sqrt{n})$

  See Figure 4.1 for a graphic summary of all the contributions of Chapters 3 and 4 leading to this main result.

Publication

- Emilia Siviero, Emilie Chautru, Stephan Clémençon. A Statistical Learning View of Simple Kriging. In *TEST, vol. 33, no 1*, pages 271-296, 2024.

Part II

Heterogeneity in Space-Time Data – Hawkes Models

5
Background

Contents

In this chapter, we present the third type of spatial data discussed in Chapter 1, Section 1.1 – point process data – and give the necessary tools to study space-time data from real-world datasets that present heterogeneity. In Section 5.1, we present the background of point processes, together with some motivations and real-world examples. In Section 5.2, we introduce the Hawkes process, a particular type of point process that has both triggering and clustering behaviors, together with the main challenges encountered when studying such data. Then, we give the necessary tools for studying Hawkes processes: the simulation methods in Section 5.3 and the estimation procedures in Section 5.4. In Section 5.5, we give more details about some real-world examples and a list of some available datasets for Hawkes process applications. Finally, in Section 5.6 we discuss the previous work presented in Staerman et al. (2023) that aims at inferring temporal Hawkes processes in an efficient and flexible way. In the next chapter, Chapter 6, we aim at extending the previous approach in Staerman et al. (2023) to spatio-temporal data.

We refer the reader to Daley and Vere-Jones (2003) for more details and insights about point processes and to Vere-Jones (2009); Diggle (2013); Reinhart (2018); Worrall et al. (2022) about Hawkes processes.

5.1 Point Processes

A Point Process (PP) is a stochastic process that describes the occurrence of events randomly distributed over time or space. It is defined as a collection of points in a one-dimensional (such as time) or multi-dimensional space (such as a geographical area). Point processes are binary-valued processes, indicating whether an event occurs at a specific time or location. Figure 5.1 depicts some examples of PPs. Figure 5.1a shows a multivariate temporal point process, formed by two processes: for each of them, the value at a time location in $[0, T]$ is equal to 1 if there is an event at this time and 0 otherwise. In Figure 5.1b, one can observe a univariate spatial point process, where each point is the location of an observed event. Finally, Figure 5.1c illustrates a univariate spatio-temporal point process where each point in the spatial domain indicates an event, and the color of this point gives the time stamp of the event. Note that, in these examples, the phenomenon occurring can be modeled to represent the time of the events, their location, or both.

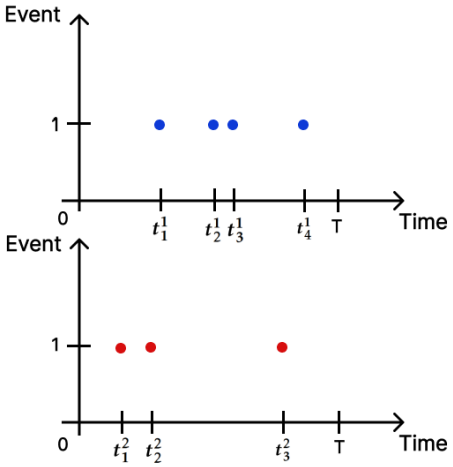
One can think of many real-world situations where a phenomenon occurring can be described by a point process. In the following, we give three examples of application domains where point processes are used to describe the observed phenomenon.

Example 5.1. *(Seismology) In seismology, point processes have been introduced by Vere-Jones (1970), driven by the need to better understand and predict earthquake occurrences. PPs are a powerful tool to model, analyze, and describe the random occurrence of earthquakes over time and space. PPs can also be used to provide insight into a possible clustering and triggering pattern. Indeed, an earthquake can trigger a new occurrence, called an aftershock, in a specific time window and spatial neighborhood. Thus, it is necessary to use models that can capture both clustering and triggering behaviors. The study on clustered and self-exciting point processes is still focused on seismological data, with the need to improve the accuracy and reliability of earthquake prediction and risk assessment. We refer the reader to Ogata (1988, 1999); Daley and Vere-Jones (2003) for more details on point processes for seismology.*

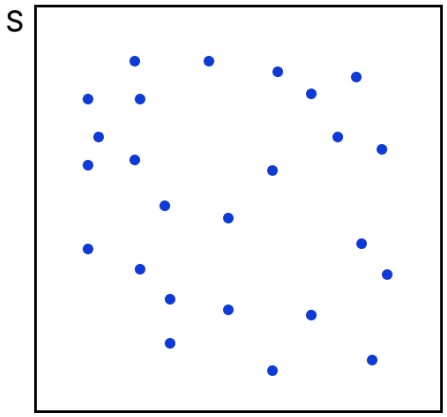
Example 5.2. *(Epidemiology) In epidemiology, point processes are essential for understanding and predicting the spread of infectious diseases, identifying patterns in disease occurrence, and implementing effective control measures. See for example Meyer and Held (2014) for more details on the modeling of an infectious disease spread, or Kresin et al. (2022) for a study on the spread of COVID-19.*

Example 5.3. *(Finance) In the context of finance, point processes help in modeling the timing and frequency of discrete events like trades, and transactions. These models are essential for understanding the dynamics of financial markets and for developing strategies to manage risk and optimize trading. We refer the reader to Bacry et al. (2015) for a review on the applications of point processes in finance.*

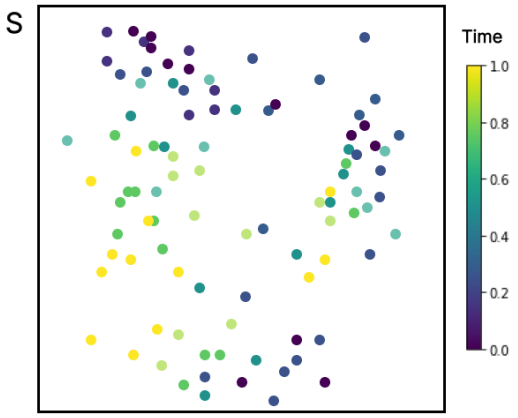
Point processes often exhibit clustering behavior, where points tend to cluster together rather than being randomly distributed. This clustering behavior can arise due to various factors such as spatial interactions, temporal dependencies, or underlying environmental conditions. Understanding and analyzing the clustering behavior of point processes is crucial in various fields, like ecology and epidemiology, as it provides insights into the underlying mechanisms driving the observed point patterns.



(a) Multivariate temporal point process



(b) Univariate spatial point process



(c) Univariate spatio-temporal point process

Figure 5.1: Several examples of point processes: realization of a multivariate temporal PP with two processes (top left); realization of a univariate spatial PP (top right); realization of a univariate spatio-temporal PP where the colors of the point locations give the time of the events (bottom).

In point process theory, the behavior of the phenomenon under study is characterized by its intensity function, which describes the expected rate of events occurrences at different points. The intensity function may rely on time, on past events, or be constant. Various forms of point processes exist, distinguished primarily by the structure of their intensity functions and the factors upon which these functions depend. Here are the main categories of point processes:

- *Homogeneous Poisson process:* The simplest and most widely used point process. It assumes events occur independently and uniformly over time or space: the occurrences of events are unrelated and do not influence others. As a consequence, the Poisson process does not exhibit clustering behavior. It exhibits stationarity, meaning that the phenomenon is sufficiently homogeneous (its characteristics remain consistent) within the domain. It is characterized by a constant intensity, indicating the average number of events per unit of time or space. The time

between events (interarrival time) follows an exponential distribution.

- *Renewal process:* Generalizes the Poisson process by allowing the interarrival time to follow an i.i.d. distribution other than the exponential distribution. Still, the renewal process lacks triggering behavior, since the occurrence of the next event is independent of past events.

- *Inhomogeneous Poisson process:* It is more flexible than the homogeneous Poisson process since it allows the event rate to vary over time or space. The number of events in any interval is assumed to follow a Poisson distribution, and the intensity function may depend on time or space.

- *Cox process:* Also known as a doubly stochastic Poisson process, it introduces randomness into the intensity function, making it a useful model for scenarios where the event rate varies over time or space.

- *Gibbs process:* Gibbs point processes are stochastic models used to describe the spatial distribution of points in a region of interest. These processes incorporate both random and deterministic components to model the spatial arrangement of points. The intensity function depends only on the location and configuration of neighboring points. Gibbs point processes are defined in terms of an interaction potential or energy function, which quantifies the strength of interactions between points. This potential function often depends on the distances between points and may incorporate other spatial attributes.

- *Hawkes process:* A Hawkes (or self-exciting) process (HP; Hawkes, 1971) presents both clustering and triggering behavior. In such a process, each event increases the likelihood of future events in its neighborhood. In a Hawkes process, events arrive randomly over time, and the rate of event arrivals is influenced by the history of past events.

Referring to the Examples 5.1, 5.2 and 5.3, we show the motivation behind the different types of point processes, their advantages and limitations.

Example 5.4. *(Seismology) Following a significant earthquake (called a mainshock), a series of smaller aftershocks typically occurs over the subsequent days. In the case of earthquake prediction, since the seismic data presents a strong cluster structure, the Poisson process often fails to capture the clustering of earthquakes and is not suitable in these situations. Clustered processes are already a better option since these models account for the tendency of earthquakes to cluster in time and space. The most notable among these is the Epidemic-Type Aftershock Sequence (ETAS) model. The ETAS model, introduced first by Ogata (1988) for purely temporal processes, suggests that each earthquake can trigger aftershocks, which themselves can trigger further aftershocks, creating a cascade of seismic activity. More precisely, self-exciting (or Hawkes) processes are particularly well suited for seismic data, since they stipulate that the occurrence of an event increases the likelihood of future events. Hawkes processes allow capturing both the triggering (immediate aftershocks) and clustering (subsequent aftershocks over time) behaviors observed in seismic data. Although predicting the exact time and location of an earthquake remains challenging, Hawkes processes improve the understanding of earthquake patterns and of potential future seismic activity.*

Example 5.5. *(Epidemiology) In the context of diseases' spread, Hawkes processes are used to represent self-exciting events where each case can potentially trigger further cases. This is particularly useful in diseases that exhibit secondary transmission, such as influenza or COVID-19. It captures the clustering behavior typical in infectious diseases. Following ETAS literature, Meyer et al. (2012) introduce self-exciting spatio-temporal point process adapted for predicting the incidence of invasive meningococcal disease, in which cases of infections can be divided into two categories (background and triggered events). On the contrary, in the case of a disease where all transmissions are from infected individuals to susceptible individuals, there are not background events producing new cases without having been exposed to an infected individual. Schoenberg et al. (2019) propose a recursive epidemic model, where the expected number of offspring of an event varies as a function of the conditional intensity. It accounts for the natural behavior of epidemics: as a disease becomes more prevalent, more people have already been exposed and active prevention measures slow the spread of the infection.*

Example 5.6. *(Finance) The Cox process allows for modeling time-varying and state-dependent event rates. The Hawkes process captures the clustering of trades and the contagion of financial shocks.*

We now give the main definitions and properties of Temporal Point Processes (TPPs).

Let $T \in \mathbb{R}_+$ be a stopping time and $[0, T]$ the resulting observation period. A temporal point process is a stochastic process whose realization consists of a set of distinct timestamps: $\mathcal{H}_T = \{t_n, t_n \in [0, T]\}$ occurring in continuous time. The behavior of a TPP is fully characterized by its intensity function, which represents the infinitesimal probability of an event occurring at time t. The *conditional intensity* function $\forall t \in [0, T]$ is

$$\lambda(t|\mathcal{H}_t) = \lim_{dt \to 0} \frac{\mathbb{P}(N_{t+dt} - N_t = 1|\mathcal{H}_t)}{dt}, \tag{5.1}$$

where $N_t = \sum_{n \geq 1} \mathbf{1}_{t_n \leq t}$ is the counting process associated with the TPP. The counting process is a non-decreasing and right-continuous process that counts the number of events up to time t.

The factors upon which the conditional intensity depends can vary for different types of processes. As previously said, in the case of homogeneous Poisson processes, the conditional intensity λ is constant and does not depend either on the past history or on the current time t. In the case of inhomogeneous Poisson processes, the conditional intensity function depends on the current time t and its value may vary over time. For Hawkes processes, the conditional intensity function depends on past history $\mathcal{H}_t$ (incorporating information up to but not including time t).

An important definition often used for statistical analysis of point processes is the compensator (or integrated intensity) function defined as follows.

Definition 5.7. *(Compensator function) The compensator or integrated intensity function $\Lambda(t)$ is the cumulative intensity over time and is defined as:*

$$\Lambda(t) = \int_0^t \lambda(\tau) d_\tau.$$

The main properties of point processes are given below.

Proposition 5.8. *(Stationarity) A point process is stationary if its statistical characteristics do not change over time.*

Proposition 5.9. *(Ergodicity) A point process is said to be ergodic if the time averages converge to ensemble averages. This allows long-term observations to be representative of the entire process.*

5.2 Hawkes Processes

In this subsection, we focus on self-exciting processes, motivating the choice of these models by seismology study. We first give the main definition of temporal Hawkes processes, together with a central result on the existence and unicity of self-exciting processes. Then, we present the particular case of spatio-temporal Hawkes processes.

5.2.1 Temporal Hawkes processes

Historically, the Poisson process was among the first statistical models used in seismology. It assumes that earthquakes occur randomly over time with a constant average rate. Because of its simplicity, the Poisson process fails to account for the observed clustering of earthquakes, particularly the aftershock sequences that follow major seismic events. Introduced in Hawkes (1971), the temporal Hawkes process was a significant advancement in modeling self-exciting phenomena. In seismology, Ogata (1988) applied this model to earthquakes, demonstrating its ability to capture the temporal clustering of seismic events. The temporal Hawkes process models the conditional intensity function, which represents the rate at which earthquakes occur, incorporating the effect of past events. See for example Lewis and Mohler (2011); Laub et al. (2015); Bompaire (2019) for an exhaustive overview of temporal Hawkes processes.

Univariate temporal Hawkes process. The conditional intensity function of a univariate Hawkes process can be expressed by the cluster representation:

$$\forall t \in [0,T], \quad \lambda(t|\mathcal{H}_t) = \mu + \int_0^t \alpha\, g(t-\tau)\mathrm{d}N_\tau \tag{5.2}$$

$$= \mu + \sum_{t_n \in \mathcal{H}_t} \alpha\, g(t-t_n), \tag{5.3}$$

where $\mu > 0$ is the constant baseline parameter (also referred to as the *background* of the process), α is the excitation scaling parameter, and $g : [0,T] \mapsto \mathbb{R}_+$ is the non-negative excitation function (called *kernel*) that represents the influence of past events onto future events.

Another point of view to define Hawkes processes is the Immigration-Birth representation (Hawkes and Oakes, 1974). This representation supposes that the events can be divided into two categories: first, the immigrant events that follow a Poisson process with rate μ and define the cluster centers; then, the offspring events that are induced by the clusters. The clusters are independent and present different branching structures.

In the following proposition, we recall the result from Hawkes and Oakes (1974) about the existence and unicity of a Hawkes process.

Proposition 5.10. *(Existence and unicity) Let N be a Hawkes process characterized by the intensity function in Equation (5.3), with a positive constant background $\mu > 0$. N is stable if $\int_0^\infty g(\tau)d\tau < 1$. If N is stable, then it exists a unique stationary process that satisfies Equation (5.3).*

Multivariate temporal Hawkes process. Multivariate Hawkes processes model the interactions of $D \in \mathbb{N}^*$ self-exciting TPPs. Given D sets of timestamps $\mathcal{H}_T^i = \{t_n^i, t_n^i \in [0,T]\}_{n=1}^{N_T^i}, i = 1,\dots,D$, where $N_T^i = \sum_{n\geq 1} \mathbf{1}_{t_n^i \leq T}$, each process i is described by the following intensity function:

$$\begin{aligned}\forall t \in [0,T], \quad \lambda_i(t|\mathcal{H}_t^i) &= \mu_i + \sum_{j=1}^{D} \int_0^t \alpha_{ij}\, g_{ij}(t-\tau)\mathrm{d}N_\tau^j \\ &= \mu_i + \sum_{j=1}^{D} \sum_{t_n^j \in \mathcal{H}_t^j} \alpha_{ij}\, g_{ij}(t-t_n^j),\end{aligned}$$

where the kernel function g_{ij} represents the influence of j^{th} process' past events onto the i^{th} process future events.

A multivariate Hawkes model composed of two processes is presented in Figure 5.2. On the left, the counting process values N_t^i with respect to time t are depicted for both processes $i \in \{1,2\}$. It's possible to observe the non-decreasing and the right-continuous properties of the counting processes. On the right, the respective conditional intensity values $\lambda^i(t|\mathcal{H}_t^i)$ are shown for each process $i \in \{1,2\}$, with respect to the time values t. The dotted green lines give the constant background values μ_i. We observe that the probability increases after an event occurrence, and then slowly decreases towards μ_i. The way the probability decreases depends on the chosen kernel for each couple of processes. The dotted orange arrow shows how an event from process 1 can influence either a future event of the same process or a future event of the other process 2. Note that the influence is ordered in time: only earlier events can influence the occurrence of subsequent events.

5.2.2 Spatio-temporal Hawkes processes

As previously seen, temporal Hawkes processes were introduced in seismology in the 80's by Ogata (1988), called the ETAS models. Compared to simpler point process models like Poisson processes, using temporal Hawkes processes offered a way to model the self-exciting nature of earthquakes in a purely temporal framework. However, the complex dynamics of seismic activity are not just confined to the temporal dimension. Earthquakes exhibit both spatial and temporal clustering, meaning that aftershocks are likely to occur near the location of the mainshock as well as shortly after it. To capture this spatio-temporal dependency, it is necessary to develop Spatio-Temporal Hawkes Processes (STHPs). These models extend the temporal Hawkes process by incorporating spatial dimensions, thereby providing a more comprehensive and realistic framework for modeling seismic activity. For further details on STHPs, please refer to Daley and Vere-Jones (2003); Reinhart (2018).

Univariate spatio-temporal Hawkes process. In the case where the spatial location of an event also gives information on the process, and in the case where the occurrence of an event is influenced by the location of previous events, one can model the phe-

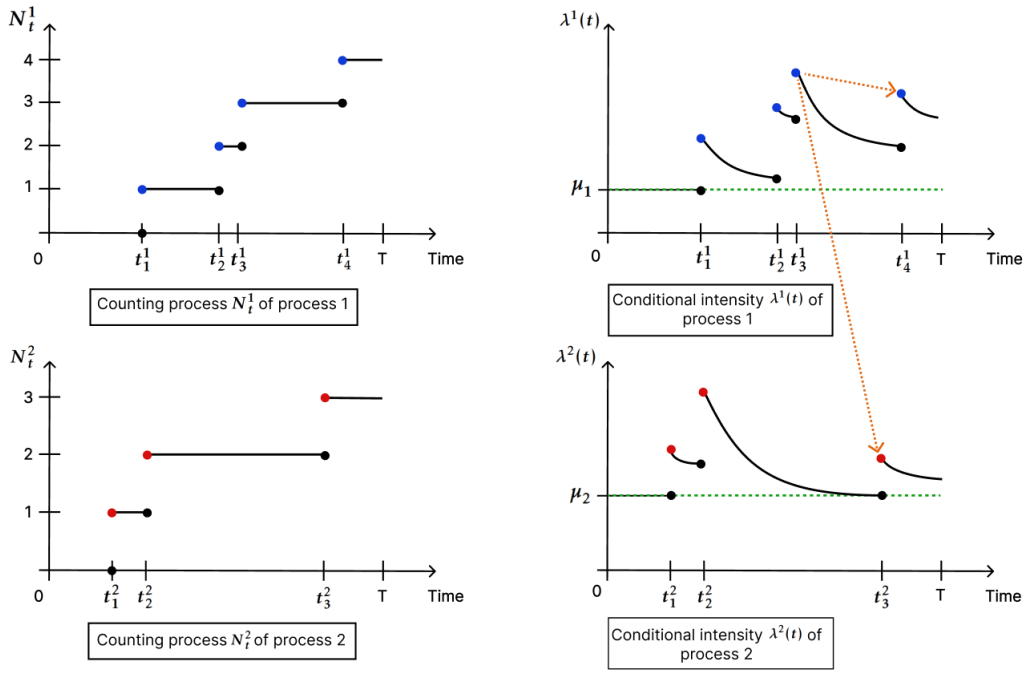


Figure 5.2: A multivariate temporal Hawkes process with two processes: the counting processes N_t^1 and N_t^2 with respect to time t (left), and the conditional intensity values with respect to time t (right). In blue are the events of process 1 (top) and in red are the events of process 2 (bottom). In the conditional intensity figure (right), the green dotted line represents the value of the constant backgrounds μ_1 and μ_2.

nomenon by a spatio-temporal point process. Let $\mathcal{S} \subset \mathbb{R}^2$ be a compact set of the space domain containing the locations of the observed events until time T. A STHP realization consists of a set of distinct events: $\mathcal{H}_T = \left\{u_n = (x_n, y_n, t_n), (x_n, y_n) \in \mathcal{S}, t_n \in [0, T]\right\}$ occurring in continuous space-time, with an associated time t_n and a location (x_n, y_n). The process's behavior is fully characterized by its intensity function, which relies on the time and location of past events. Denote by $\mathbf{N}$ the random counting measure defined on $\mathcal{S} \times [0, T]$, such that $\mathbf{N}(\mathrm{d}x, \mathrm{d}y, \mathrm{d}t) = \sum_{n=1}^{\infty} \delta_{(x_n, y_n, t_n)}(\mathrm{d}x, \mathrm{d}y, \mathrm{d}t)$, where (x, y) and t respectively denote the location and time of the events. The conditional intensity function of such a process is defined as the map from $\mathbb{R}^3$ to $\mathbb{R}_+$ such that: $\forall t \in [0, T], (x, y) \in \mathcal{S}$,

$$\lambda(x, y, t|\mathcal{H}_t) = \lim_{\mathrm{d}x, \mathrm{d}y, \mathrm{d}t \to 0} \frac{\mathbb{E}\Big[\mathbf{N}\Big([x, x+\mathrm{d}x] \times [y, y+\mathrm{d}y] \times [t, t+\mathrm{d}t]\Big)|\mathcal{H}_t\Big]}{\mathrm{d}x\mathrm{d}y\mathrm{d}t}.$$

The conditional intensity function of an STHP can also be expressed by the cluster representation as:

$$\begin{aligned}\lambda(x, y, t|\mathcal{H}_t) &= \mu + \int_0^t \int_{\mathcal{S}} \alpha\, g(x-u, v-y, t-\tau)\mathrm{d}N_u \mathrm{d}N_v \mathrm{d}N_\tau \\ &= \mu + \sum_{u_n \in \mathcal{H}_t} \alpha\, g(x-x_n, y-y_n, t-t_n),\end{aligned}$$

where g is a spatio-temporal kernel function.

Marked Hawkes process. An important feature in earthquake data is the magnitude of an earthquake. Indeed, a mainshock with high magnitude may have more probability to trigger a subsequent earthquake, compared to a mainshock with low magnitude. Thus, the magnitude of an earthquake can be a valuable information to include in the Hawkes models. A marked Hawkes process is a type of Hawkes process where each event is associated with an additional mark or attribute (Reinhart, 2018, Section 2). Marked Hawkes processes have been developed in various application domains, such as seismology (Ogata, 1988), ecology (Schoenberg, 2004), criminology (Mohler, 2014; Zhu and Xie, 2022), and social science (Yuan et al., 2019, 2021).

Finally, we present Multivariate Spatio-Temporal Hawkes Processes (MSTHPs), that are used in Chapter 6. See also Bompaire (2019) and Section 8 in Daley and Vere-Jones (2003) for further details on MSTHPs.

Multivariate spatio-temporal Hakwes process.

Given $D \geq 1$ type of events, for each $i \in [\![1, D]\!]$, the conditional intensity function of the i-th process has the following form:

$$\lambda_i(x, y, t|\mathcal{H}_t) = \boldsymbol{\mu}_i + \sum_{j=1}^{D} \sum_{u_n^j \in \mathcal{H}_t^j} \boldsymbol{\alpha}_{ij}\, g_{ij}(x - x_n^j, y - y_n^j, t - t_n^j), \tag{5.4}$$

where the strictly positive $\boldsymbol{\mu}_i$'s are the baseline parameters, controlling the spontaneous event apparition rate, the positive $\boldsymbol{\alpha}_{ij}$ are the *excitation scaling* parameters, and the $g_{ij} : \mathcal{S} \times [0, T] \mapsto \mathbb{R}_+$ are the spatio-temporal *kernel*, also referred as excitation functions. The parameters $\boldsymbol{\alpha}_{ij}$ and g_{ij} describe the excitation behavior between events. Here we assume that $0 \leq \boldsymbol{\alpha}_{ij} < 1$ and $\int_0^T \int_{\mathcal{S}} g_{ij}(x, y, t)\mathrm{d}x\mathrm{d}y\mathrm{d}t = 1$, ensuring the stability of the generated process.

From now on, we may simplify the notation of the conditional intensity function by $\lambda_i(x, y, t)$ for the sake of clarity.

Baseline. Notice that the baseline μ can either be defined as a function of the time t or of the spatial location (x, y), or either chosen as constant. The definition of the form of the baseline depends on the physical properties of the phenomenon under study. For example, for seismological data, the background sometimes depends on the spatial dimension (for instance, it may represents the tectonic plates), see *e.g.* Musmeci and Vere-Jones (1992); Ilhan and Kozat (2020); Kwon et al. (2023). However, in the first ETAS models, the baseline μ was assumed to be constant, and it represented the background seismic activity rate in a region of interest (see *e.g.* Marsan and Lengline, 2008). In epidemiology, the background is often chosen as constant, in particular, if the phenomenon of interest is an infectious disease (see *e.g.* Embrechts et al., 2011; Dong et al., 2023). In Mohler et al. (2011), where the variable under study is the occurrences of burglaries in a city, the background rate is defined as a function of the spatial location of the events, multiplied by a function depending on the temporal dimension. Indeed, the crime risk may depend on temporal and spatial fluctuations such as holiday seasons, population densities in some regions of the city, etc (see also Zhuang and Mateu, 2019; D'Angelo et al., 2022).

In our study

Constant Baseline: For simplicity reasons, in our study in Chapter 6, the background rate is supposed to be constant.

5.3 Simulation of Hawkes Processes

The simulation of Hawkes processes requires generating event sequences, based on the conditional intensity function of such processes. The main challenge is to preserve the temporal dependencies in the data, and the spatial dependence structure in the case of STHP. The simulation of Hawkes processes is a crucial step in the study of self-exciting processes, since it allows better understanding their behavior, validating theoretical models, and developing accurate predictions.

Several methods have been developed for this purpose, each with its advantages and challenges. We present two main approaches: the Ogata's thinning algorithm and the clustering algorithm.

5.3.1 Ogata's Thinning Simulation Algorithm

Ogata's thinning algorithm, introduced in Ogata (1998), is a widely used method for simulating inhomogeneous Poisson processes and self-exciting point processes. The key idea of this approach is the 'thinning' procedure: several events (called *candidate events*) are first generated, and then selectively accepted or rejected (*i.e.* thinned), based on the conditional intensity function. From a temporal point of view only, the thinning step is applied to the first generated event and then on the subsequent events (following the unidirectional flow of time, from present to future), such that the historical dependence of the process is preserved. Lewis and Shedler (1979) introduced this approach for inhomogeneous Poisson processes. Then, Ogata (1998) extended this procedure to allow the simulation of self-exciting point processes. Ogata's algorithm relies on the definition of an upper bound $\lambda^*(t|\mathcal{H}_t)$ for the intensity function $\lambda(t'|\mathcal{H}_t)$, for all $t' \geq t$. One can easily adapt this method to spatio-temporal Hawkes process simulations by introducing the spatial dimension: the spatial location of each events are simulated during the first phase of the algorithm, just after having simulated the time of the candidate events. The two main steps of the algorithm are the following (for a univariate spatio-temporal Hawkes process):

1. First, it draws the time t' of a new candidate event, sampled from a homogeneous Poisson process, based on λ^* (defined as the upper bound of $\lambda(x, y, t|\mathcal{H}_t)$, where t is the current time). After defining the time of the new event, it generates a candidate spatial location (x', y') uniformly over the spatial domain $\mathcal{S}$.
2. Then, it accepts or rejects the candidate event, based on the actual intensity function. For this purpose, it evaluates the actual intensity function at the candidate event (x', y', t') by $\lambda(x', y', t'|\mathcal{H}_t) = \mu + \sum_{u_n \in \mathcal{H}_t} g(x' - x_n, y' - y_n, t' - t_n)$. For the acceptance-rejection step, the candidate event (x', y', t') is accepted with probability $\lambda(x', y', t'|\mathcal{H}_t)/\lambda^*$.

We present the algorithm for univariate STHPs in Algorithm 5.1. Notice that for simplicity $\lambda(x', y', t'|\mathcal{H}_t)$ is written $\lambda(x', y', t')$.

Algorithm 5.1 Ogata's thinning algorithm for a univariate spatio-temporal Hawkes process

Require: ending time T, spatial set $\mathcal{S} \subset \mathbb{R}^2$

Set the initial time $t = 0$
while t < T **do**
  Define an upper bound λ^*, such that $\lambda^* \geq \lambda(x', y', t')$, $\forall (x', y') \in \mathcal{S}$, $\forall t' \geq t$
  Draw the candidate event time t' from an exponential distribution with rate λ^*
  Draw the candidate event location (x', y') uniformly on $\mathcal{S}$
  Sample $u \sim \mathcal{U}_{[0,1]}$
  if $u < \lambda(x', y', t')/\lambda^*$ **then**
    Accept the candidate event (x', y', t') and add it to the list of events
  end if
  $t = t'$
end while

Refer to Section 3 in Reinhart (2018) for a presentation of the Ogata's thinning procedure applied to spatio-temporal Hawkes processes. See also Ilhan and Kozat (2020) for a thinning procedure for marked STHPs, with a slight modification in order to generate data with multiple event types (the marks of the process).

However, this procedure is computationally expensive, since the intensity function must be evaluated for each new event.

5.3.2 Clustering Simulation Algorithm

The clustering method (also called the Immigration-Birth algorithm, see subsection 5.2 on the Immigration-Birth representation from Hawkes and Oakes, 1974) was initially proposed for earthquake simulation by Zhuang et al. (2004) and then improved by Møller and Rasmussen (2005). It alleviates the computational cost of Ogata's algorithm since it does not require the thinning step and the evaluation of the intensity function for each new event. The idea is that it directly exploits the cluster structure of the data: it first generates a sequence of immigrant events from the baseline and then simulates offspring events. Recall that the immigrant events are drawn from a Poisson process with rate μ and are used to define the cluster centers, while the offspring events are events induced by the clusters. The algorithm for univariate STHPs with space-time separable kernels is summarized in Algorithm 5.2 (see also Yuan et al., 2019, Algorithm 3, and Kwon et al., 2023 for similar cluster algorithms). Notice that it easily extends to the space-time non-separable case (see subsection 5.4.3) by generating offspring inter arrivals according to the space-time non-separable kernel directly.

For further details and other simulation algorithms, we refer the reader to Daley and Vere-Jones (2003) for TPPs and to Reinhart (2018) for a review of STHP's simulation methods.

Algorithm 5.2 Clustering algorithm for a spatio-temporal Hawkes process

Require: ending time T, spatial set $\mathcal{S} \subset \mathbb{R}^2$ of area $\mathcal{A}_{\mathcal{S}}$

Initialize an empty stack **L**

Define the number of immigrant events N_I from a Poisson distribution with parameter $\lambda_I = \mu \times T \times \mathcal{A}_{\mathcal{S}}$

Draw N_I immigrant samples defined by $(x_i, y_i, t_i), \forall i \in [1, N_I]$, where (x_i, y_i) is sampled uniformly on $\mathcal{S}$ and t_i uniformly on $[0, T]$

Add each immigrant event to the stack **L**

while **L** is not empty **do**

Remove the most recently added element to **L** denoted (x_i, y_i, t_i)

Generate the number n_i of offspring events from a Poisson distribution with rate α (the excitation scaling parameter)

Generate the offspring inter arrivals $(u_k, v_k, s_k), \forall k \in \{1, \dots, n_i\}$ according to the spatial and temporal kernels

Define the offspring events as $\left\{(x_i + u_1, y_i + v_1, t_i + t_1), \dots, (x_i + u_{n_i}, y_i + v_{n_i}, t_i + t_{n_i})\right\}$

Remove the offspring events that are outside the spatial domain $\mathcal{S}$ or outside the temporal window $[0, T]$

Add the immigrant (x_i, y_i, t_i) to the set of valid offspring events

Add the offspring events to the set **L**

end while

In our study

Clustering Simulation: We simulate events according to the clustering algorithm presented above, with a slight modification to account for finite support kernels (see Chapter 6 and the corresponding GitHub).

5.4 Estimation and Inference

Suppose that we have observed a realization of an MSTHP at n observations composed of event locations $\left\{(x_1, y_1), \dots, (x_n, y_n)\right\} \in \mathcal{S}$ and times $\{t_1, \dots, t_n\} \in [0, T]$. The goal of this subsection is to give a short review of available methods to estimate the parameters of a Hawkes process, based on a finite set of observations.

The statistical inference of a MSTHP defined as in (5.4) concerns, for all $i \in \{1, \dots, D\}$, the parameters μ_i, α_{ij} and the triggering kernels η_{ij} that can be parametric (Yuan et al., 2019; Reinhart, 2018) or non-parametric (Lewis and Mohler, 2011; Choi and Hall, 1999; Diggle et al., 1995; Kwon et al., 2023). Define the set of parameters to estimate as $\theta = \{\mu_i, \alpha_{ij}, \eta_{ij}\}_{i,j}, \forall i, j$.

We present here two inference approaches: the maximum log-likelihood, and the minimum least squares.

5.4.1 Maximum Log-likelihood

The log-likelihood of a MSTHP is given in Daley and Vere-Jones, 2003, Section 7.3:

$$L(\boldsymbol{\theta}, \mathcal{H}_T) = \sum_{i=1}^{D} \left(\sum_{u_n^i \in \mathcal{H}_T^i} \log\left(\lambda_i(x_n^i, y_n^i, t_n^i)\right) - \int_0^T \int_{\mathcal{S}} \lambda_i(x, y, t) \mathrm{d}x \mathrm{d}y \mathrm{d}t \right). \tag{5.5}$$

Notice that the log-likelihood function is composed of two parts: the first reflects the triggering behavior, while the second one is the compensator function (defined as the cumulative intensity over time and space, extending Definition 5.7 to the spatio-temporal case). However, performing maximum log-likelihood is generally impossible in practice. Indeed, it requires the computation of a sum over all the events of the intensity function. A solution is to consider Expectation Maximization (EM) algorithms (Veen and Schoenberg, 2008). The idea is to introduce a cluster representation of the events, labeling each event if it either comes from the background rate or if a previous event triggers it. Then, the two steps of the EM algorithm are the following: *(E-step)* the triggering probabilities (defined as the probability that an event does not arise directly from the background rate) are estimated based on the current parameters and the expected log-likelihood function can be computed, *(M-step)* the expected log-likelihood is maximized with respect to each parameter, and the parameters can be updated (then, return to the *E-step*). See *e.g.* subsection 3.1 in Reinhart (2018) for further details. However, this approach may introduce an estimation bias induced by boundary effects. A solution is the stochastic declustering approach in Zhuang et al. (2002), where the same cluster (or branching) structure of the events is computed. The idea is to fit in a non-parametric manner the background rate, based only on the background-labeled events. The iterative procedure is composed of the following main steps: starting from a constant background assumption, (1) it fits the parameters using the maximum log-likelihood principle; (2) then, it computes the triggering probabilities; (3) finally, it updates the estimated of the background rate (and returns to step 1). When the algorithm stops, a thinning (or stochastic declustering) procedure is applied to retrieve the background events. These two EM-based methods from Zhuang et al. (2002); Veen and Schoenberg (2008) assume a parametric form for the kernel. Marsan and Lengline (2008) proposed a non-parametric approach for piecewise constant kernel functions (called *model-independent stochastic declustering*) by estimating the shape of the kernel function directly from the observations, while Lewis and Mohler (2011) presented another non-parametric algorithm (called *maximum penalized likelihood estimation*) assuming that the kernel function exhibits at least a needed degree of regularity. Still, Marsan and Lengline (2008) assumed that the triggering function is space-time separable (see below for a discussion about space-time separability). In Kwon et al. (2023), a more flexible non-parametric approach is proposed, where the space-time interactions are taken into account.

5.4.2 Least Squares Minimization

The ERM-inspired least squares loss of a MSTHP (Reynaud-Bouret and Rivoirard, 2010; Reynaud-Bouret et al., 2014; Bacry et al., 2020) is:

$$\mathcal{L}(\boldsymbol{\theta}, \mathcal{H}_T) = \sum_{i=1}^{D} \left(\int_0^T \int_{\mathcal{S}} \lambda_i(x, y, t)^2 \mathrm{d}x \mathrm{d}y \mathrm{d}t - 2 \sum_{u_n^i \in \mathcal{H}_T^i} \lambda_i(x_n^i, y_n^i, t_n^i) \right). \tag{5.6}$$

Notice that the triggering part does not require the logarithmic function, compared to the above log-likelihood in Equation 5.5. In Staerman et al. (2023), an efficient and flexible inference method is proposed, based on a discretized version of the least squares loss (see Section 5.6 for a complete presentation of this method). Furthermore, this method takes advantage of the absence of the logarithm in the triggering part of the loss, thus allowing for precomputations and leveraging the computational burden. However, this method is suited only for temporal Hawkes processes, and its extension to spatio-temporal processes has yet to be developed.

In our study

ERM-inspired least squares loss: In Section 6.3, the least squares minimization approach is preferred and the objective is to extend the method in Staerman et al. (2023) for MSTHPs.

5.4.3 Space-time Separability

A key aspect of modeling with Hawkes processes is the concept of separability. For example, for marked Hawkes processes, it is generally assumed that the mark of an event does not depend either on the temporal dimension nor on the past events (see *e.g.* the ETAS models Ogata (1988, 1998) in seismology). Schoenberg (2004) and Díaz-Avalos et al. (2014) investigated non-parametric Monte-Carlo tests for the separability of a spatio-temporal marked point process and show how the separability assumption can be prohibitive. Another separability assumption can be made between the spatial and the temporal dimensions. In the literature (see *e.g.* Mohler, 2014; Yuan et al., 2019; Ilhan and Kozat, 2020), the kernel $g_{ij}(\cdot)$ is generally supposed to be first-order space-time separable, which means that the kernel is a product of spatial and temporal influences (see González et al., 2016, Section 4.2, and Reinhart, 2018, Section 2.2). This assumption is made for simplicity reasons since it implies that the temporal and spatial components can be modeled and estimated separately. Under the space-time separability assumption, the locations of the events are assumed to be independent of the time of the events, preventing any kind of space-time interaction. However, in practice, this is not always the case and the separability assumption can be restrictive when dealing with real-world situations. For instance, in criminology, burglars often revisit the same area to commit several burglaries consecutively within a short period of time (Johnson, 2008). Cressie and Huang (1999) proposed a class of kernel functions that do not satisfy the space-time separable assumption and Gneiting (2002) extended it to a more general class of space-time non-separable functions. In both works, an implementation of these approaches for various non-separable functions on real-world applications shows the accuracy of their proposed methods. In the context of earthquakes prediction, Kwon et al. (2023) used a space-time non-separable model and proves the flexibility induced by non-separable kernels.

5.4.4 Constrained Kernel Models

For computational purposes, the temporal kernel is often chosen as exponential, while the spatial kernel is chosen as Gaussian (see *e.g.* Mohler, 2014; Yuan et al., 2019; Ilhan and Kozat, 2020). However, there are various reasons to consider alternative kernels

for some datasets. An exponential temporal kernel assumes that the influence of past events decays exponentially, which may not accurately represent real-world situations with long-term memory or varying rates of decay. A spatial Gaussian kernel assumes that spatial correlations are symmetric around a point, which may not be realistic in all contexts, especially when directional dependencies are observed. Thus, restricting the modeling to these two kernels may be prohibitive. Indeed, selecting kernels that better reflect the underlying data characteristics and dependencies can significantly enhance the flexibility, adaptability, and accuracy of models. The choice of kernel should be guided by the specific nature of the data and by the goals of the analysis.

These specific choices – separability and constrained kernels– both limit the modeling flexibility between the temporality and spatiality of events, reducing the applicability of STHP models to real-world data.

In our study

Space-time Non-separable Kernels and General Parametric Form of Kernels: The objective of Chapter 6 is to extend the method developed in Staerman et al. (2023) to MSTHPs. Thus, the first goal is to develop an efficient inference method that allows the estimation of any parametric kernel. Furthermore, in order to take into account space-time interactions, the second goal is to allow estimating from space-time non-separable kernels.

5.5 Real-world Examples and Datasets

In this section, we present some of the main domains of application of Hawkes processes and we give some links to real-world datasets presenting self-exciting features. The aim of this section is to show the importance to develop flexible and accurate STHP models, demanded by various real-world situations.

Seismology. It is widely known that after a large earthquake (for example an earthquake of large magnitude, or an earthquake with a consequent spatial spread), a sequence of smaller earthquakes is typically observed. The first earthquake is referred to as the *mainshock*, whereas the subsequent ones are called the *aftershocks*. The aftershocks occur in the following days and in a spatial neighborhood of the mainshock. Thus, earthquakes exhibit a strong spatio-temporal clustering behavior (see Examples 5.1 and 5.4, and the discussion on seismology in Section 5.2). This clustering structure suggests to resort to spatio-temporal self-exciting models, such as Hawkes processes (see *e.g.* Musmeci and Vere-Jones, 1992; Ogata, 1998).

Dataset. The *National Earthquake Information Center* [1] provides earthquakes with a magnitude of 2.5 or higher since 1966. An earthquake record includes the date, the time, the location and the magnitude of the earthquake. The *Northern California Earthquake Data Center* [2] (nce, 2014) also provides an earthquake datasets, only for seismic events in California.

[1] https://earthquake.usgs.gov/earthquakes/search/
[2] https://ncedc.org/

Criminology. For various crimes (for example burglary, shootings, etc), a spatio-temporal clustering behavior can be observed. Indeed, a 'near-repeat victimization' (as described in Mohler et al., 2011) is commonly observed. For example, in the case of burglaries, Johnson (2008) noticed that burglars often steal repeatedly from a same neighborhood (or even a same street) in a short period of time. Another example of near-repeat victimization is the case of gang violence: a shooting may causes further vengeance shootings in the days after and in the same area (see *e.g.* Ratcliffe and Rengert, 2008). Therefore, a same pattern of mainshocks and aftershocks together with a spatial correlation can be observed in crime data, and thus crimes can be modeled by STHPs (Mohler et al., 2011; Mohler, 2014; D'Angelo et al., 2022; Zhu and Xie, 2022). The application of Hawkes processes to criminology began in the early 2000s as researchers recognized the parallels between the clustering of earthquakes and crime incidents. Mohler et al. (2011) were among the pioneers to apply Hawkes processes to crime data. They assume that the model is divided into the background (usually defined as spatial dependent, for factors such as the socioeconomical background of the neighborhood, the police activity in the sector, etc) and a self-exciting function that accounts for near-repeats events. Mohler (2014) introduced the use of marked Hawkes processes to identify hotspots of homicide (described as areas within a city where crime rates are significantly elevated). Furthermore, Mohler (2014) purpose was to see if other less violent crimes (such as burglary, criminal damage, etc) have an impact on more serious gun crimes. Thus, these other types of crimes, divided into marks, are incorporated into the conditional intensity by adding weights for each class. Later, D'Angelo et al. (2022) introduced a STHP model based on the underlying network structure of a region. This allows for a better understanding of the process behavior and leads to better predictions. In Zhu and Xie (2022), another marked STHP model was proposed, where the marks are defined as the textual descriptions of crimes.

Datasets. The *Chicago Crime Dataset* [3] gives the reported incidents in the City of Chicago from 2001 until now. An incident is defined by its location, time, and type (theft, burglary, assault, etc).

Epidemiology. The spread of a disease can be divided in two parts: the outbreak of a disease (that can be viewed as a mainshock) and the transmission of this disease in the neighboring areas (where the new infected cases are viewed as aftershocks). Spatio-temporal Hawkes processes offer a robust and flexible approach for the modeling of the spread of infectious diseases. For example, Holbrook et al. (2022) proposed a Hawkes process for the modeling of the Ebola outbreak in West Africa in the 2010s. With the COVID-19 pandemic, several studies (*e.g.* Rambhatla et al., 2022; Kresin et al., 2022; Dong et al., 2023) applied Hawkes processes to understand the spatio-temporal dynamics of the spread of this disease. By capturing the complex dynamics of disease transmission in both time and space, these models provide valuable insights for epidemiological research and public health interventions. As data availability continues to increase, spatio-temporal Hawkes processes play an important role in epidemiology.

[3] https://data.cityofchicago.org/Public-Safety/Crimes-2001-to-Present/ijzp-q8t2

Datasets. See for example subsection 3.2 in Holbrook et al. (2022), for a presentation and the corresponding links to real-data concerning the Ebola outbreak in West Africa (from 2014 to 2016). The *New York Times* website [4] provides the cumulative numbers of COVID-19 cases in the United States (it is used for example in Chiang et al., 2022 for the modeling of COVID-19 with Hawkes processes).

Climatology. In climatology, spatial variability is largely observed: meteorological data is influenced by geographical factors such as the presence of mountains, the latitude of the spatial observation domain, and the climate zone. Furthermore, the mainshock-aftershock pattern can also be observed in some situations: for example, a major hurricane may have a significant and immediate impact on the affected region, characterized by strong winds and heavy rainfall, thus conducting to subsequent climate events, such as tornadoes and landslides.

Datasets. Gneiting (2002) used the wind data from Ireland available on the StatLib website [5] in order to propose stationary covariance functions for spatio-temporal processes.

5.6 Fast and Flexible Inference for Temporal Hawkes Processes

In this section, we present the previous work in Staerman et al. (2023) aiming at inferring efficiently and flexibly temporal Hawkes processes with general parametric kernels.

A classical flexible approach for the inference of temporal Hawkes processes is the use of non-parametric kernels. Still, even if this approach brings flexibility to the modeling of Hawkes processes, it often comes with poor estimation when the data is limited. The increasing availability of large temporal datasets presenting a clustering and triggering behavior thus suggests resorting to more efficient approaches. On the other hand, parametric kernels reduce the computational burden and can be more efficient in estimation. However, they may introduce a bias if the assumed kernel shape does not fit the data well. The main motivation behind Staerman et al. (2023)'s work is to step outside of the classical setting where an exponential form is assumed for the temporal kernel. Indeed, the exponential kernel is often preferred since it is data-efficient and allows a simpler estimation than with other kernels. Yet, by its definition, the exponential kernel is adapted only for datasets where the events immediately trigger successive events. In practice, real-world situations do not fit this specific assumption: often, a latency between events is observed (for example, in the case of seismic data, an earthquake may not trigger aftershocks shortly after, but in the following days or weeks instead).

Thus, a more flexible and efficient parametric modeling of Hawkes processes is needed, allowing the estimation of any parametric kernel. Staerman et al. (2023) propose a new method, called *FaDIn* consisting of a fast ℓ_2 gradient based-solver. The method's goal is to find the parameters $\boldsymbol{\mu}, \boldsymbol{\alpha}$ and the kernels' parameters $\boldsymbol{\eta}$ that minimize the ERM-inspired least squares loss. Let $\boldsymbol{\theta} = (\boldsymbol{\mu}, \boldsymbol{\alpha}, \boldsymbol{\eta})$ be the set of parameters to estimate.

[4] https://www.nytimes.com/interactive/2021/us/covid-cases.html
[5] https://lib.stat.cmu.edu/

Furthermore, the method's efficiency and flexibility result from the three following key points.

Finite Support Kernels. Generally, it is assumed that the impact of a past event decreases over time. Restricting the class of parametric kernels to those with finite support implies that the influence of a past event is limited in time. Thus, not only does the influence diminish but it becomes null after a fixed window of time. Indeed, this assumption allows computing the conditional intensity function for all events, even for long-time-length processes, where previous parametric estimations generally failed because of the computational burden.

In our study

Finite Support Kernels: In subsection 6.2.2, we extend the notion of finite support temporal kernels to spatio-temporal ones: we consider spatio-temporal kernels to be of finite length and we assume that the influence of a past event is limited both in time and in space (for all directions).

Discretization of the Temporal Window. Inspired by the discretization method for non-parametric kernels, Staerman et al. (2023) propose to use a discretized version of the temporal Hawkes process: the events are projected on a regular temporal grid with finite elements. Combining the discretization step with the finite support kernels assumption, the triggering part in the conditional intensity function can be rewritten in a discretized version: the sum over the past events is replaced by a lighter sum over only a finite number L of grid elements, where L denotes the number of points on the discretized support.

Furthermore, the discretization has a low impact on the statistical performance of the estimator (see subsection 3.2 in Staerman et al., 2023).

In our study

Spatio-temporal Discretization: In subsection 6.2.3, we apply the same discretization procedure to the spatio-temporal dimension: we define a three-dimensional regular grid and project the observed events on these grids. Under the finite support kernels assumption, the triggering part in the intensity can be rewritten as a sum of kernels over a finite number $L_S \times L_S \times L_T$ of grid elements, where L_S (respectively L_T) denotes the number of points on the discretized spatial (resp. temporal) support. As for the temporal case, we show that the bias induced by the discretization is negligible compared to the statistical error (see Section 6.4 in Chapter 6).

Precomputations. Combining the finite support kernels and the discretization grid, Staerman et al. (2023) obtain a discretized version of ℓ_2 loss, by approximating the integral appearing in the ℓ_2 loss with a sum on the elements of the discretized grid. Furthermore, the sum over all the events in the ℓ_2 loss is replaced by a sum over the projected events on the grid. Developing and rearranging the terms in the discretized loss, some constants that do not depend on θ appear. These constants can be precomputed, thus reducing the complexity of the optimization step.

In our study

Precomputations: By developing the resulting discretized least squares loss for spatio-temporal processes, precomputation terms also appear (see subsection 6.3.2), thus significantly reducing the computational cost.

The flexibility and efficiency of this method are shown by numerical experiments on both simulated and neuroscience real-data in Sections 3 and 4 in Staerman et al. (2023).

5.7 Conclusion

Spatio-temporal Hawkes processes are well adapted to model natural phenomena presenting a self-exciting behavior both in the space and the time dimensions. Such data can be observed in a wide variety of domains, such as seismology, epidemiology, and climatology, and thus gaining more and more attention. Hawkes processes handle the self-exciting nature of a phenomenon by capturing its clustering and triggering behaviors. A temporal Hawkes process describes an occurrence of events where future events are influenced by the history of the process, composed of its past events. Generally, the impact of an event on future possible events decreases over time. In the case of spatio-temporal Hawkes processes, the occurrence of an event is affected by both the time and the location of previous events. However, in the spatial dimension, there is no notion of past, present, and future, since the flow is multi-directional. Thus, usually, an event can be triggered (or influenced) only by its spatial neighboring events that happened previously. An STHP is characterized by its intensity function (5.4), generally composed of two elements: a baseline, which controls the occurrence rate of spontaneous events, and a triggering kernel sum over the past history, which describes the excitation behavior between events. The events can be divided into two main categories: the background events (also called immigrants), which define the cluster centers; and the triggered events (also called the offsprings), whose occurrence was influenced by a previous event.

Key points of this Chapter.

Definitions used and Assumptions overcome in this thesis

- **Point Processes:** describe the occurrence of events, which behavior is characterized by the conditional intensity function ⟶ application to a wide variety of domains (Section 5.1 and examples within).
- **Self-exciting Processes:** clustering and triggering behavior ⟶ cluster representation (Section 5.2):

$$\lambda(x,y,t|\mathcal{H}_t) = \underbrace{\mu}_{\text{background}} + \overbrace{\sum_{u_n \in \mathcal{H}_t}}^{\text{based on history}} \underbrace{g(x-x_n, y-y_n, t-t_n)}_{\substack{\text{triggering kernel: influence of} \\ \text{past events onto future events}}} .$$

- **Spatio-Temporal Hawkes Processes:** the occurrence of an event is influenced by the time and location of previous events ⟶ better understand the spatial dependence structure of the data (subsection 5.2.2 and eq. (5.4)).
- **Space-time Separability:** for simplicity reasons, the kernel is often assumed to be a product of spatial and temporal kernels ⟶ need to develop a method to learn from space-time interactions (subsection 5.4.3).
- **General Kernels:** the temporal kernel is often chosen as exponential while the spatial as Gaussian, but not realistic in all contexts ⟶ need to develop a parametric method allowing any kind of kernel, to enhance flexibility, adaptability, and accuracy (subsection 5.4.4).

Concepts and Methodologies used in this thesis

- **Simulation:** clustering simulation algorithm ⟶ directly exploits the cluster structure of the data (subsection 5.3.2).
- **Estimation and Inference:** ERM-inspired least squares loss approach (subsection 5.4.2).
- **FaDIn:** Staerman et al. (2023) inference method for temporal Hawkes processes, based on three key points: *(i)* finite-support kernels, *(ii)* discretization of time, *(iii)* precomputations ⟶ flexible, efficient and accurate parametric estimation (Section 5.6).

Real-world Applications and Datasets used in this thesis.

- **Applications:** seismological and criminal data: clustering and triggering patterns ⟶ MSTHP well-suited (Section 5.5).
- **Datasets:** seismic activity in California (nce, 2014) and reported burglaries in Chicago.

6

A Fast Method for Parametric Inference in Space-Time Hawkes Models

Contents

6.1 Introduction

In this chapter we develop a new methodology for general parametric inference for space-time Hawkes processes. This contribution answers to our **Research Questions 4 and 5** (see Section 1.2 in Chapter 1):

How to learn from a multivariate spatio-temporal Hawkes process, despite the modeling and numerical challenges posed by parametric STHP's complexity? How to accurately model real-world situations, where space-time interactions occur and where a latency between aftershocks may be observed, by means of Hawkes processes?

Our algorithm is based on key concepts similar to the ones used by Staerman et al. (2023) in the context of temporal events, which we extend to spatio-temporal events. Our algorithm solves the modeling and numerical challenges posed by parametric STHP's complexity, and makes it possible to account for space-time interactions/heterogeneity. The method consists of a fast ℓ_2 gradient-based solver, which provides flexibility in the choice of kernels, and is made computationally efficient based on the following three ideas:

1. First, we model the intensity function with bounded support parametric kernels. We show that this choice offers significant advantages regarding computational complexity, allowing the use of discrete convolution and fast Fourier transform.

2. Second, our method leverages a discretization of the space-time domain, which can be seen as a hyperparameter to tune, whose selection depends on the underlying sampling precision of the data and data availability. Combined with the first key component, it allows any choice of temporal and spatial kernels, providing flexibility in the modeling.

3. Third, our approach resorts to extensive use of precomputations, allowing for efficient implementation of the gradient-based inference procedure, with optimization steps independent of the number of events.

In this work, we show that the discretization has a low impact on the statistical performance of our estimator, and demonstrate the flexibility and efficiency of the proposed method on simulated and real-world datasets.

The chapter is organized as follows. In Section 6.2 we detail the key components of our approach. Next, in Section 6.3 we present the inference approach for STHPs, which consists of a fast ℓ_2 gradient-based solver, and discuss its numerical advantages. Section 6.4 provides theoretical guarantees on the bias induced by discretization. In Section 6.5 we investigate the performance of the methodology promoted on simulated data from an empirical point of view, while in Section 6.6 we present the experimental results obtained using our method based on two real data, earthquake and burglary, for illustration purposes. Finally, some concluding remarks are collected in Section 6.7. The proofs of the main results are provided in Appendix C.

Contributions. The main motivations for this work are the limitations of existing space-time Hawkes models, due to simplifying assumptions that heavily limit the possibility of applying them to real data. Our contributions to overcome these limitations are:

- We develop a fast method to infer kernel parameters for space-time Hawkes models. The method allows incorporating any parametric kernels for the triggering function.

- We show the accuracy of our approach on simulated data.

- Finally, the advantages of our inference method are proved on real earthquake and burglary data. The fully reproducible experiments are available on GitHub.

6.2 Key Components

In this section, we outline the statistical approach promoted in this chapter. Building upon recent work by Staerman et al. (2023), we develop a parametric inference framework that allows the estimation of any spatio-temporal kernel to reflect the underlying excitation of the process. In particular, our approach shows linear scalability as a function of the number of events, greatly improving current methodologies. Our inference procedure relies on three fundamental principles: *(i)* discretization, *(ii)* finite support kernels for both spatial and temporal components of our processes, and *(iii)* precomputation terms.

Recall the usual setting presented in Section 5.2 in Chapter 5. Let $T \in \mathbb{R}_+$ be a stopping time and $[0, T]$ the resulting observation period, and let $\mathcal{S} \subset \mathbb{R}^2$ be a compact set of the space domain containing the locations of the observed events until time T. Let $D \in \mathbb{N}^*$ be the number of different types of events of the MSTHP. Then, an MSTHP realization consists of D sets of distinct events: $\mathcal{H}_T^i = \left\{ u_n^i = (x_n^i, y_n^i, t_n^i), (x_n^i, y_n^i) \in \mathcal{S}, t_n^i \in [0, T] \right\}, \forall i \in \{1, \dots, D\}$ occurring in continuous space-time, with an associated time t_n^i and a location (x_n^i, y_n^i). The process's behavior is fully characterized by its D intensity functions, which rely on the time and location of past events. The conditional intensity function of the i-th process has the following form:

$$\lambda_i(x, y, t | \mathcal{H}_t) = \boldsymbol{\mu}_i + \sum_{j=1}^{D} \sum_{u_n^j \in \mathcal{H}_t^j} \boldsymbol{\alpha}_{ij}\, g_{ij}(x - x_n^j, y - y_n^j, t - t_n^j), \tag{6.1}$$

where $\boldsymbol{\mu}_i > 0$ is the baseline (or background) parameter, $\boldsymbol{\alpha}_{ij} > 0$ is the excitation scaling parameter, and $g_{ij} : \mathcal{S} \times [0, T] \mapsto \mathbb{R}_+$ is the spatio-temporal kernel (or excitation/triggering function). We refer the reader to subsection 5.2.2 in Chapter 5 for further details on MSTHPs.

For simplicity, we assume that the spatial domain is a rectangle, *i.e.*, of the form $\mathcal{S} = \mathcal{X} \times \mathcal{Y}$ with $\mathcal{X} = [-S_{\mathcal{X}}, S_{\mathcal{X}}]$ and $\mathcal{Y} = [-S_{\mathcal{Y}}, S_{\mathcal{Y}}]$. Our approach does not require the upper and lower limits to be identical or symmetrical with respect to zero for it to work.

6.2.1 Convolutional Writing

For all $u = ((x, y), t) \in \mathcal{S} \times [0, T]$ and for all $s \in [0, T]$, let $z_s^i(u) = \sum_{u_n^i \in \mathcal{H}_s^i} \delta_{u_n^i}(u)$ be the sum of Dirac functions of event occurrences u_n^i, such that $z_s^i(u) = 1$ if $u \in \mathcal{H}_s^i$ and 0 otherwise. The intensity function in Equation (6.1) can be reformulated as a convolution between the kernel g_{ij} and z_t^j:

$$\forall u \in \mathcal{S} \times [0, T], \quad \lambda_i(u) = \boldsymbol{\mu}_i + \sum_{j=1}^{D} \boldsymbol{\alpha}_{ij}\, (g_{ij} * z_t^j)(u). \tag{6.2}$$

6.2.2 Finite Support Kernels

We consider the spatio-temporal kernels to be of finite length. Let $W_{\mathcal{X}}, W_{\mathcal{Y}}$ and W_T be the length of spatial and temporal supports. We assume that $\forall (x, y, t) \notin [-W_{\mathcal{X}}, W_{\mathcal{X}}] \times [-W_{\mathcal{Y}}, W_{\mathcal{Y}}] \times [0, W_T]$, $g_{ij}(x, y, t) = 0$. Thus, any event $u_n^i = (x_n^i, y_n^i, t_n^i)$ may induce a new

event only in

$$[x_n^i - W_{\mathcal{X}}, x_n^i + W_{\mathcal{X}}] \times [y_n^i - W_{\mathcal{Y}}, y_n^i + W_{\mathcal{Y}}] \times [t_n^i, t_n^i + W_T] \ .$$

In combination with Equation (6.2), the main advantage of this assumption is to make it possible to leverage discrete convolution and fast Fourier transform for efficient intensity computation. Indeed, unlike previous parametric methods, this approach enables the computation of the conditional intensity function even for long-duration processes, thus reducing the computational burden.

6.2.3 Discretization

Discretization has been successfully used in Hawkes processes (Kirchner, 2017) and recently in spatio-temporal processes (Sheen et al., 2022). Given the spatial support as a compact set $[-S_{\mathcal{X}}, S_{\mathcal{X}}] \times [-S_{\mathcal{Y}}, S_{\mathcal{Y}}] \subset \mathbb{R}^2$, we propose to define a three-dimensional regular grid $\mathcal{G} = \mathcal{G}_{\mathcal{X}} \times \mathcal{G}_{\mathcal{Y}} \times \mathcal{G}_T$ such that $\mathcal{G}_{\mathcal{X}} = \{-S_{\mathcal{X}}, -S_{\mathcal{X}} + \Delta_{\mathcal{X}}, \cdots, -S_{\mathcal{X}} + G_{\mathcal{X}}\Delta_{\mathcal{X}}\}$, $\mathcal{G}_{\mathcal{Y}} = \{-S_{\mathcal{Y}}, -S_{\mathcal{Y}} + \Delta_{\mathcal{Y}}, \cdots, -S_{\mathcal{Y}} + G_{\mathcal{Y}}\Delta_{\mathcal{Y}}\}$ and $\mathcal{G}_T = \{0, \Delta_T, \cdots, G_T\Delta_T\}$ with $G_T\Delta_T = T$, $G_{\mathcal{X}}\Delta_{\mathcal{X}} = 2S_{\mathcal{X}}$, $G_{\mathcal{Y}}\Delta_{\mathcal{Y}} = 2S_{\mathcal{Y}}$, and $\Delta_{\mathcal{X}}, \Delta_{\mathcal{Y}}, \Delta_T > 0$ are the stepsizes of the spatial and temporal grids. Further, we project the observed events on these grids and define $\widetilde{\mathcal{H}}_T^i$ as the projected space-time stamps of $\mathcal{H}_T^i$. Given $v = (v_x, v_y, v_t) \in [\![0, G_{\mathcal{X}}]\!] \times [\![0, G_{\mathcal{Y}}]\!] \times [\![0, G_T]\!]$, we define the vector versions $g_{ij}^{\Delta}[v] = g_{ij}(v\Delta)$ of the kernels, and the sparse vector of events:

$$z_t^j[v] = \#\{(x_n^j, y_n^j, t_n^j) : \ |x_n^j - (-S_{\mathcal{X}} + v_x\Delta_{\mathcal{X}})| \leq \frac{\Delta_{\mathcal{X}}}{2}, |y_n^j - (-S_{\mathcal{Y}} + v_y\Delta_{\mathcal{Y}})| \leq \frac{\Delta_{\mathcal{Y}}}{2}, |t_n^j - v_t\Delta_T| \leq \frac{\Delta_T}{2}\},$$

that reflects the number of events projected at the position v on the grid $\mathcal{G}$. With these notations, we can rewrite the intensity function of the i^{th} process of our discretized STHP relying on discrete convolution such that for any $v = (v_x, v_y, v_t) \in [\![0, G_{\mathcal{X}}]\!] \times [\![0, G_{\mathcal{Y}}]\!] \times [\![0, G_T]\!]$, we have:

$$\tilde{\lambda}_i[v] = \boldsymbol{\mu}_i + \sum_{j=1}^{D} \boldsymbol{\alpha}_{ij} (g_{ij}^{\Delta} * z_{v_t\Delta_T}^j)[v] = \boldsymbol{\mu}_i + \sum_{j=1}^{D} \sum_{\tau_x=1}^{L_{\mathcal{X}}} \sum_{\tau_y=1}^{L_{\mathcal{Y}}} \sum_{\tau_t=1}^{L_T} \boldsymbol{\alpha}_{ij}\, g_{ij}^{\Delta}[\tau]\, z_{v_t\Delta_T}^j[v - \kappa],$$

where $\tau = (\tau_x, \tau_y, \tau_t)$, $\kappa = \Big(\tau_x - l_{\mathcal{X}}, \tau_y - l_{\mathcal{Y}}, \tau_t\Big)$, with $l_{\mathcal{X}} = \Big\lfloor L_{\mathcal{X}}/2 \Big\rfloor + 1$, $l_{\mathcal{Y}} = \Big\lfloor L_{\mathcal{Y}}/2 \Big\rfloor + 1$. Define $L_T = \Big\lfloor W_T/\Delta_T \Big\rfloor + 1$ the number of points on the discretized temporal support, and $L_{\mathcal{X}} = \Big\lfloor 2W_{\mathcal{X}}/\Delta_{\mathcal{X}} \Big\rfloor + 1$, $L_{\mathcal{Y}} = \Big\lfloor 2W_{\mathcal{Y}}/\Delta_{\mathcal{Y}} \Big\rfloor + 1$ the number of points on each component of the discretized spatial support.

Thus, by combining the discretization step with the finite support kernels assumption, the triggering component in the conditional intensity function can be reformulated in a discretized manner. This means that instead of summing over all past events, we now sum over a finite number $L_{\mathcal{X}} \times L_{\mathcal{Y}} \times L_T$ of grid elements.

Furthermore, in Section 6.4, we provide theoretical guarantees concerning the influence of discretization on parameter estimation.

6.3 Efficient Inference with Empirical Risk Minimization

In contrast to existing spatio-temporal literature, we leverage the overlooked ERM-inspired ℓ_2 loss, which we extend to the spatial domain. Although this loss benefits from advantageous precomputation terms, we propose an approximation for computationally intensive terms to accelerate the inference.

6.3.1 ERM-inspired Least Squares Loss

While the spatio-temporal literature focuses on the negative log-likelihood minimization to infer Hawkes parameters, we decide to focus on the ERM-inspired least squares loss (Reynaud-Bouret and Rivoirard, 2010; Reynaud-Bouret et al., 2014; Bacry et al., 2020) only used in classical temporal Hawkes process so far. In contrast to the log-likelihood, the least squares loss disentangles the computation dependency in the number of events from the optimization procedure. Indeed, it involves precomputation terms, that summarize the offset information of events thanks to the absence of logarithm in the right part of the loss. Let $\mathcal{H}_T = \{\mathcal{H}_T^i\}_{i=1}^D$ be a set of observed spatio-temporal events. Assuming a class of spatio-temporal parametric kernels parametrized by η_{ij}, the objective is to find $\boldsymbol{\theta} = \{\boldsymbol{\mu}_i, \boldsymbol{\alpha}_{ij}, \eta_{ij}\}_{i,j}$ that minimizes:

$$\mathcal{L}(\boldsymbol{\theta}, \mathcal{H}_T) = \sum_{i=1}^{D} \left(\int_0^T \int_{\mathcal{S}} \lambda_i(x,y,t)^2 \, \mathrm{d}x\mathrm{d}y\mathrm{d}t - 2 \sum_{u_n^i \in \mathcal{H}_T^i} \lambda_i(x_n^i, y_n^i, t_n^i) \right). \tag{6.3}$$

Given the core components of our method described in Section 6.2, the objective is then to minimize the discretized ℓ_2 loss defined by:

$$\mathcal{L}_{\mathcal{G}}(\boldsymbol{\theta}, \widetilde{\mathcal{H}}_T) = \sum_{i=1}^{D} \left(\Delta_{\mathcal{X}} \Delta_{\mathcal{Y}} \Delta_T \sum_{v_x=0}^{G_{\mathcal{X}}} \sum_{v_y=0}^{G_{\mathcal{Y}}} \sum_{v_t=0}^{G_T} \left(\tilde{\lambda}_i[v_x, v_y, v_t] \right)^2 - 2 \sum_{\tilde{u}_n^i \in \widetilde{\mathcal{H}}_T^i} \tilde{\lambda}_i \left[\frac{\tilde{x}_n^i}{\Delta_{\mathcal{X}}}, \frac{\tilde{y}_n^i}{\Delta_{\mathcal{Y}}}, \frac{\tilde{t}_n^i}{\Delta_T} \right] \right).$$

This approximates the integral in Equation (6.3) by a sum on the grid $\mathcal{G}$ after projecting the space-time stamps of $\mathcal{H}_T$ on it.

6.3.2 Precomputations

By developing and rearranging the terms in the discretized loss above, one can see some constants that do not depend on $\boldsymbol{\theta}$ and thus can be precomputed:

$$\begin{aligned}
\mathcal{L}_{\mathcal{G}}(\boldsymbol{\theta}, \widetilde{\mathcal{H}}_T) =& (T+\Delta_T)(2S_{\mathcal{X}}+\Delta_{\mathcal{X}})(2S_{\mathcal{Y}}+\Delta_{\mathcal{Y}}) \sum_{i=1}^{D} \boldsymbol{\mu}_i^2 \\
&+ 2\Delta_{\mathcal{X}}\Delta_{\mathcal{Y}}\Delta_T \sum_{i=1}^{D} \boldsymbol{\mu}_i \sum_{j=1}^{D} \sum_{\tau_x=1}^{L_{\mathcal{X}}} \sum_{\tau_y=1}^{L_{\mathcal{Y}}} \sum_{\tau_t=1}^{L_T} \boldsymbol{\alpha}_{ij}\, g_{ij}^{\Delta}[\tau]\, \Phi_j(\tau; G) \\
&+ \Delta_{\mathcal{X}}\Delta_{\mathcal{Y}}\Delta_T \sum_{i,j,k=1}^{D} \sum_{\tau_x,\tau_x'=1}^{L_{\mathcal{X}}} \sum_{\tau_y,\tau_y'=1}^{L_{\mathcal{Y}}} \sum_{\tau_t,\tau_t'=1}^{L_T} \boldsymbol{\alpha}_{ij}\, \boldsymbol{\alpha}_{ik}\, g_{ij}^{\Delta}[\tau]\, g_{ik}^{\Delta}\left[\tau'\right] \Psi_{j,k}(\tau, \tau'; G) \\
&- 2 \sum_{i=1}^{D} \left(N_T^i \boldsymbol{\mu}_i + \sum_{j=1}^{D} \sum_{\tau_x=1}^{L_{\mathcal{X}}} \sum_{\tau_y=1}^{L_{\mathcal{Y}}} \sum_{\tau_t=1}^{L_T} \boldsymbol{\alpha}_{ij}\, g_{ij}^{\Delta}[\tau]\, \Phi_j(\tau; \widetilde{\mathcal{H}}_T^i) \right),
\end{aligned}$$

where the following terms can be precomputed:

- $\Phi_j(\tau; G) = \sum_{v_x=0}^{G_{\mathcal{X}}} \sum_{v_y=0}^{G_{\mathcal{Y}}} \sum_{v_t=0}^{G_T} z^j_{v_t\Delta_T}[v-\kappa]$,
- $\Phi_j(\tau; \widetilde{\mathcal{H}}_T^i) = \sum_{\tilde{u}_n^i \in \widetilde{\mathcal{H}}_T^i} z^j_{\tilde{t}_n^i}\left[\frac{\tilde{u}_n^i}{\Delta} - \kappa\right]$,
- $\Psi_{j,k}(\tau, \tau'; G) = \sum_{v_x=0}^{G_{\mathcal{X}}} \sum_{v_y=0}^{G_{\mathcal{Y}}} \sum_{v_t=0}^{G_T} z^j_{v_t\Delta_T}[v-\kappa]\, z^k_{v_t\Delta_T}\left[v-\kappa'\right]$,

with $\kappa = \left(\tau_x - l_{\mathcal{X}}, \tau_y - l_{\mathcal{Y}}, \tau_t\right)$ and $G = (G_{\mathcal{X}}, G_{\mathcal{X}}, G_T)$. $\Phi_j(\tau; G)$ defines the total number of events of the j-th process by removing a part of the grid of size κ. $\Psi_{j,k}(\tau, \tau'; G)$ denotes how many events of the j-th process with a lag κ are matching the events of the k-th process with a lag κ'. $\Phi_j(\tau; \widetilde{\mathcal{H}}_T^i)$ assess how many events in the j-th process are at the same position than events of the i-th process with a lag κ. As these three terms do not depend on the set of parameters, they can be precomputed at initialization and used at each step of the optimization procedure. Let $\overline{G} = G_{\mathcal{X}} G_{\mathcal{Y}} G_T$ be the total number of element on the grid $\mathcal{G}$ and $\overline{L} = L_{\mathcal{X}} L_{\mathcal{Y}} L_T$ be the total number of discretization points of the kernels g_{ij}. The term $\Psi_{j,k}(\tau, \tau'; G)$ is the bottleneck of these precomputations and requires $O(\overline{G})$ for each tuples (τ, τ') and (j,k). Thus, it leads to a total computational complexity of $O(D^2 \overline{L}^2 \overline{G})$. This may be limiting in the choices of the discretization steps $\Delta_{\mathcal{X}}, \Delta_{\mathcal{Y}}$ and Δ_T, driving the user to take them not too small and then inducing discretization bias in the results of the solver.

6.3.3 Approximation of Ψ

The precomputation terms are computed only once, but in the spatio-temporal setting, they may suffer from a computational burden. The bottleneck is the $\overline{L}^2$ presence in the computational complexity of $\Psi_{i,j}(\cdot\,;\cdot)$. Here, we provide an approximation of $\Psi_{i,j}$, denoted by $\widetilde{\Psi}_{i,j}$, to alleviate this computation challenge. Precisely, we propose

$$\widetilde{\Psi}_{j,k}(\tau;G) = \sum_{v_x=0}^{G_\mathcal{X}} \sum_{v_y=0}^{G_\mathcal{Y}} \sum_{v_t=0}^{G_T} z^j_{v_t\Delta_T}\Big[v_x, v_y, v_t\Big] z^k_{v_t\Delta_T}\Big[v_x-\tau_x, v_y-\tau_y, v_t-\tau_t\Big].$$

The $\widetilde{\Psi}_{j,k}(\tau;G)$ is the number of events in the discretized j-th process that have events with a lag τ in the discretized k-th process. Thus, it is evaluated in $\tau-\tau'$ in the third term of the loss $\mathcal{L}_\mathcal{G}$. The quadratic complexity in $\overline{L}$ is then removed and makes the computation of $\widetilde{\Psi}_{j,k}$ of order $O(\overline{LG})$, which is linear with the grid discretization and the kernel grids and comparable to the computation complexity of $\Phi_j(\cdot;\cdot)$. The loss of information in this approximation lies in the boarding effects of the grid, which are small if the domain size is large in front of the kernel support.

6.3.4 Gradient-based Optimization

The inference procedure employs gradient descent to minimize the ℓ_2 loss function $\mathcal{L}_\mathcal{G}$. Our approach design enables the utilization of flexible parametric kernels for both temporal and spatial patterns. It efficiently computes exact gradients for each kernel parameter, assuming the kernel is both differentiable and possesses finite support. Consequently, gradient-based optimization methods can be applied without constraints, in stark contrast to the EM algorithm, widely used in the literature, necessitating a closed-form solution to nullify the gradient – a challenge with numerous kernels. It's worth noting that this problem typically entails non-convexity, potentially leading to convergence towards local minima.

The gradients of the proposed loss w.r.t. each set of parameters are given below. Let $\tau = (\tau_x, \tau_y, \tau_t)$ be a vector on the grid $\mathcal{G}$.

Baselines. The gradient of the loss with respect to the constant background for all $m \in [\![1, D]\!]$ is

$$\begin{aligned}\frac{\partial \mathcal{L}_\mathcal{G}(\boldsymbol{\theta}, \widetilde{\mathcal{H}}_T)}{\partial \boldsymbol{\mu}_m} &= 2(T+\Delta_T)(2S_\mathcal{X}+\Delta_\mathcal{X})(2S_\mathcal{Y}+\Delta_\mathcal{Y})\boldsymbol{\mu}_m - 2N_T^m \\ &\quad + 2\Delta_\mathcal{X}\Delta_\mathcal{Y}\Delta_T \sum_{k=1}^{D}\sum_{\tau_x=1}^{L_\mathcal{X}}\sum_{\tau_y=1}^{L_\mathcal{Y}}\sum_{\tau_t=1}^{L_T} \boldsymbol{\alpha}_{mk}\, g^\Delta_{mk}[\tau]\,\Phi_k(\tau;G).\end{aligned}$$

Excitation scaling parameters. The gradient of the loss with respect to $\boldsymbol{\alpha}_{m,l}$ for all $(m,l) \in [\![1, D]\!]^2$ is

$$\begin{aligned}\frac{\partial \mathcal{L}_\mathcal{G}(\boldsymbol{\theta}, \widetilde{\mathcal{H}}_T)}{\partial \boldsymbol{\alpha}_{m,l}} &= 2\Delta_\mathcal{X}\Delta_\mathcal{Y}\Delta_T \boldsymbol{\mu}_m \sum_{\tau_x=1}^{L_\mathcal{X}}\sum_{\tau_y=1}^{L_\mathcal{Y}}\sum_{\tau_t=1}^{L_T} g^\Delta_{ml}[\tau]\,\Phi_l(\tau;G) - 2\sum_{\tau_x=1}^{L_\mathcal{X}}\sum_{\tau_y=1}^{L_\mathcal{Y}}\sum_{\tau_t=1}^{L_T} g^\Delta_{ml}[\tau]\,\Phi_l(\tau;\widetilde{\mathcal{H}}_T^m) \\ &\quad + 2\Delta_\mathcal{X}\Delta_\mathcal{Y}\Delta_T \sum_{k=1}^{D}\sum_{\tau_x,\tau'_x=1}^{L_\mathcal{X}}\sum_{\tau_y,\tau'_y=1}^{L_\mathcal{Y}}\sum_{\tau_t,\tau'_t=1}^{L_T} \boldsymbol{\alpha}_{mk}\, g^\Delta_{ml}[\tau]\, g^\Delta_{mk}\Big[\tau'\Big]\,\Psi_{l,k}(\tau,\tau';G).\end{aligned}$$

Kernel parameters. The gradient of the loss with respect to the parameter of the kernel for all $(m,l) \in [\![1, D]\!]^2$ is

$$\frac{\partial \mathcal{L}_{\mathcal{G}}(\boldsymbol{\theta}, \widetilde{\mathcal{H}}_T)}{\partial \boldsymbol{\eta}_{m,l}} = 2\Delta_{\mathcal{X}}\Delta_{\mathcal{Y}}\Delta_T \boldsymbol{\mu}_m \sum_{\tau_x=1}^{L_{\mathcal{X}}} \sum_{\tau_y=1}^{L_{\mathcal{Y}}} \sum_{\tau_t=1}^{L_T} \boldsymbol{\alpha}_{ml} \frac{\partial g^{\Delta}_{ml}[\tau]}{\partial \boldsymbol{\eta}_{m,l}} \Phi_l(\tau; G)$$
$$- 2 \sum_{\tau_x=1}^{L_{\mathcal{X}}} \sum_{\tau_y=1}^{L_{\mathcal{Y}}} \sum_{\tau_t=1}^{L_T} \boldsymbol{\alpha}_{ml} \frac{\partial g^{\Delta}_{ml}[\tau]}{\partial \boldsymbol{\eta}_{m,l}} \Phi_l(\tau; \widetilde{\mathcal{H}}^m_T)$$
$$+ 2\Delta_{\mathcal{X}}\Delta_{\mathcal{Y}}\Delta_T \sum_{k=1}^{D} \sum_{\tau_x, \tau'_x=1}^{L_{\mathcal{S}}} \sum_{\tau_y, \tau'_y=1}^{L_{\mathcal{S}}} \sum_{\tau_t, \tau'_t=1}^{L_T} \boldsymbol{\alpha}_{ml}\, \boldsymbol{\alpha}_{mk} \frac{\partial g^{\Delta}_{ml}[\tau]}{\partial \boldsymbol{\eta}_{m,l}} g^{\Delta}_{mk}\left[\tau'\right] \Psi_{l,k}(\tau, \tau'; G).$$

6.4 On the Bias of Spatio-temporal Discretization – Theoretical Guarantees

Discretization introduces a perturbation in the loss value. In this section, we assess the impact of this perturbation on parameter estimation as $\Delta_{\mathcal{X}}, \Delta_{\mathcal{Y}}$ and Δ_T approach 0. Throughout this section, we consider a set of events $\mathcal{H}_T$ stemming from a spatio-temporal Hawkes process with intensity functions expressed in the parametric form $\lambda_i(\cdot; \boldsymbol{\theta}^*)$, with $\boldsymbol{\theta}^* = \{\boldsymbol{\mu}^*_i, \boldsymbol{\alpha}^*_{ij}, \boldsymbol{\eta}^*_{ij}\}_{i,j}$. It is important to note that if the intensity of the process $\mathcal{H}_T$ does not belong to the parametric family $\lambda_i(\cdot; \boldsymbol{\theta})$, then $\boldsymbol{\theta}^*$ is defined as the best approximation of its intensity function in the ℓ_2 sense. The objective of the inference process is to estimate the parameters $\boldsymbol{\theta}^*$.

When working with the projected set of events $\widetilde{\mathcal{H}}_T$, the original tuple (x^i_n, y^i_n, t^i_n) is replaced with its projection on the grid $\mathcal{G}$, denoted as $\tilde{x}^i_n = x^i_n + \delta^i_{x,n}, \tilde{y}^i_n = y^i_n + \delta^i_{y,n}, \tilde{t}^i_n = t^i_n + \delta^i_{t,n}$. Here, $\delta^i_{x,n}$ is uniformly distributed over the interval $[-\Delta_{\mathcal{X}}/2, \Delta_{\mathcal{X}}/2]$, $\delta^i_{y,n}$ over $[-\Delta_{\mathcal{Y}}/2, \Delta_{\mathcal{Y}}/2]$ and $\delta^i_{t,n}$ over $[-\Delta_T/2, \Delta_T/2]$. We define the discrete estimator $\widehat{\boldsymbol{\theta}}_\Delta$ as $\widehat{\boldsymbol{\theta}}_\Delta = \arg\min_{\boldsymbol{\theta}} \mathcal{L}_{\mathcal{G}}(\boldsymbol{\theta}, \widetilde{\mathcal{H}}_T)$, the set of parameters minimizing the discrete loss. The error induced by $\widehat{\boldsymbol{\theta}}_\Delta$ can be upper-bounded as follows:

$$\|\widehat{\boldsymbol{\theta}}_\Delta - \boldsymbol{\theta}^*\| \leq \underbrace{\|\widehat{\boldsymbol{\theta}}_c - \boldsymbol{\theta}^*\|}_{(1)} + \underbrace{\|\widehat{\boldsymbol{\theta}}_\Delta - \widehat{\boldsymbol{\theta}}_c\|}_{(2)}, \tag{6.4}$$

where $\widehat{\boldsymbol{\theta}}_c = \arg\min_{\boldsymbol{\theta}} \mathcal{L}\left(\boldsymbol{\theta}, \mathcal{H}_T\right)$ is the reference estimator for $\boldsymbol{\theta}^*$ based on the standard ℓ_2 estimator for continuous spatio-temporal Hawkes processes. This decomposition involves the statistical error (1) and the bias error induced by the discretization (2). The statistical term (1) measures the deviation of the parameters obtained by minimizing the ℓ_2 continuous loss from the true parameters, given a finite amount of data. In contrast, the term (2) represents the discretization bias induced by minimizing the discrete loss instead of the continuous one.

In the following proposition, we focus on the discretization error (2), which relates to the computational trade-off offered by our method. Before stating our proposition, we need further assumption on the discretized grid, implying that no event collapses on the same grid element.

Assumption 6.1. *Suppose for any $i, j \in [\![1, D]\!]$, we have $\Delta_{\mathcal{X}} < \min_{x^i_n, x^j_m \in \mathcal{H}_T} |x^i_n - x^j_m|$, $\Delta_{\mathcal{Y}} < \min_{y^i_n, y^j_m \in \mathcal{H}_T} |y^i_n - y^j_m|$ and $\Delta_T < \min_{t^i_n, t^j_m \in \mathcal{H}_T} |t^i_n - t^j_m|$.*

We now study the perturbation of the loss due to discretization.

Proposition 6.2. *Let $\mathcal{H}_T$ and $\widetilde{\mathcal{H}}_T$ be respectively a set of events (drawn from a spatio-temporal Hawkes process) and its discretized version on the grid $\mathcal{G}$ with stepsize $\Delta = (\Delta_\mathcal{X}, \Delta_\mathcal{Y}, \Delta_T)$. Suppose Assumption 6.1 to be satisfied. Thus, for any $v = (v_x, v_y, v_t)$, it holds:*

$$\widetilde{\lambda}_i[v] = \lambda_i(v\Delta) - \sum_{j=1}^{p} \sum_{u_m^j \in \mathcal{H}_{v\Delta}^j} \delta_m^j . \nabla_u g_{ij}(v\Delta - u_m^j) + O\left(\|\Delta\|^2\right),$$

and

$$\mathcal{L}(\theta, \widetilde{\mathcal{H}}_T) \leq \mathcal{L}(\theta, \mathcal{H}_T) + \|\Delta\| \sum_{i=1} C(\lambda_i) + 2 \sum_{i,j} \sum_{\substack{u_n^i \in \mathcal{H}_T^i \\ u_m^j \in \mathcal{H}_T^j}} \left(\delta_m^j - \delta_n^i\right) . \nabla_u g_{ij}\left(u_n^i - u_m^j\right) + O\left(\|\Delta\|^2\right),$$

where $C(\lambda_i)$ is a constant depending only on the regularity of λ_i.

The technical proof is provided in Appendix C, Section C.1. The first result follows directly from the Taylor expansion of the intensity for the kernels. For the loss, the initial perturbation term, $\|\Delta\| \sum_{i=1} C(\lambda_i)$, arises from approximating the integral with a finite Euler sum by the generalization of the Koksma-Hlawka inequality for piecewise smooth functions (Brandolini et al., 2013), while the second term stems from the perturbation of the intensity. This proposition demonstrates that, as the norm of the discretization steps $\|\Delta\|$ approaches 0, the perturbed intensity and the ℓ_2 loss serve as accurate estimates of their continuous counterparts. We now proceed to quantify the discretization error (2) as $\|\Delta\|$ goes to 0.

Proposition 6.3. *Suppose the assumption in Proposition 6.2 is satisfied. Then, if the estimators $\widehat{\theta}_c = \arg\min_\theta \mathcal{L}(\theta, \mathcal{H}_T)$ and $\widehat{\theta}_\Delta = \arg\min_\theta \mathcal{L}_\mathcal{G}(\theta, \widetilde{\mathcal{H}}_T)$ are uniquely defined, $\widehat{\theta}_\Delta$ converges to $\widehat{\theta}_c$ as $\|\Delta\| \to 0$. Moreover, if $\mathcal{L}$ is C^2 and its hessian $\nabla^2 \mathcal{L}\left(\widehat{\theta}_c\right)$ is positive definite with $\varepsilon > 0$ its smallest eigenvalue, then $\left\|\widehat{\theta}_\Delta - \widehat{\theta}_c\right\| \leq \frac{\max\{\|\Delta\|, \|\Delta\|_\infty\}}{\varepsilon}\, \omega\left(\widehat{\theta}_\Delta\right)$, with $\omega\left(\widehat{\theta}_\Delta\right) = O(1)$.*

The technical proof is provided in Appendix C, Section C.2. In contrast to the bound in Staerman et al. (2023) that does not consider the spatial components, the asymptotic rates we obtained depend on $\Delta_\mathcal{X}, \Delta_\mathcal{Y}$ and Δ_T. It shows that they all must go towards zero to have θ_Δ converging to the continuous estimates. When $\Delta_\mathcal{X}, \Delta_\mathcal{Y}$ and Δ_T go to zero, the provided rate is linear w.r.t. the stepsize grid parameters, allowing fast convergence. However, it also shows that if one of the grids is not refined enough, it may deteriorate the performance of the discretized estimator. Another important remark is that the rate only depends on the sum of the discretization stepsize, and does involve their product. This means that we can use discretization of the same order for each dimension without having incurring a large degradation of the statistical efficiency of the estimator.

6.5 Numerical Experiments

In this section, we present several numerical experiments on synthetic data. First, we analyze the influence of the discretization bias on the accuracy of the method as well as the computation time of the proposed approach. Then, we assess its statistical accuracy with various values for the stopping time T and the set of locations $\mathcal{S}$. Finally, we investigate the accuracy and the computation time of the developed precomputation approximation. The codes used for this study are publicly available on GitHub.

6.5.1 On the Bias of Spatio-temporal Discretization – Numerical Assessment

To study the estimation error induced by the discretization step, we consider a one-dimensional STHP with intensity function described as in Equation (6.1). We simulate events according to the Immigration-Birth algorithm (Møller and Rasmussen, 2005) with a stopping time T and a square spatial support with bound $\mathcal{S}$. The excitation kernel g is defined as a time-space separated kernel such that $g(x,y,t) = h(x,y)f(t)$. We set the baseline parameter $\mu = 0.5$ and the scaling excitation factor $\alpha = 0.6$. To illustrate the flexibility of our proposed method, we run several experiments with varying temporal and spatial kernel functions. The considered spatial kernel functions, with finite support $[-1,1]^2$, are the following:

- The *truncated Gaussian* kernel:

$$h(x,y;m,\sigma) \propto \exp\left(-\frac{(x-m_1)^2+(y-m_2)^2}{2\sigma^2}\right)\mathbb{I}\{(x,y)\in[-1,1]^2\}, \tag{6.5}$$

  where the mean is $m = (m_1, m_2)$.

- The *truncated Inverse Power Law* kernel:

$$h(x,y;m,d) = \left(1+\frac{(x-m_1)^2+(y-m_2)^2}{d}\right)^{-3/2}\mathbb{I}\{(x,y)\in[-1,1]^2\}, \tag{6.6}$$

  where $m = (m_1, m_2)$.

For the temporal kernel function, we run the experiments with the following functions, with finite support $[0,1]$:

- The *Kumaraswamy* density function:

$$f(t;a,b) = abt^{a-1}(1-t^a)^{b-1}\mathbb{I}\{0\le t\le 1\}. \tag{6.7}$$

- The *truncated Gaussian* kernel:

$$f(t;m_T,\sigma_T) \propto \exp\left(-\frac{(t-m_T)^2}{2\sigma_T^2}\right)\mathbb{I}\{0\le t\le W_T\}. \tag{6.8}$$

- The *truncated Exponential* kernel:

$$f(t;\lambda) \propto \lambda\exp\big(-\lambda t\big)\mathbb{I}\{0\le t\le W_T\}. \tag{6.9}$$

We run four experiments and report their results in Figures 6.1, 6.2, and 6.3.

Experiment 1. For the first experiment, the spatial kernel is defined as a truncated Gaussian (6.5) with mean $m = (0,0)$ and standard deviation $\sigma = 0.1$, and the temporal triggering function is chosen as the Kumaraswamy (6.7) with $a = 2$ and $b = 2$. Hence, the set of parameters to estimate is $\theta^* = (\mu, \alpha, m, \sigma, a, b)$. We compute the estimates of θ^* for varying stepsizes $\Delta = (\Delta_\mathcal{X}, \Delta_\mathcal{Y}, \Delta_T)$ of the spatial and temporal grids. To highlight Proposition 6.3, we set equal refinement of the grid w.r.t. each modality, i.e., $\Delta_T = \Delta_\mathcal{X} = \Delta_\mathcal{Y} \in [0.5, 0.05]$. The experiments are computed for multiple ending time $T \in \{10, 100\}$ and spatial bounds $\mathcal{S} \in \{10, 20\}$. Our estimates $\hat{\theta}$, obtained by applying our approach, are compared to θ^*. Precisely, the median (over 100 runs) and the 25%-75% quantiles of the ℓ_2 estimation error $\|\hat{\theta} - \theta^*\|$ are displayed in Figure 6.1 (left). The associated computation time is depicted in Figure 6.1 (right).

One can observe that the estimation error goes towards zero as Δ decreases and supports the theoretical rates obtained in Proposition 6.3. In addition, when T and $\mathcal{S}$ increases, i.e. the number of events increases, the error diminishes. The computation time is efficient according to the setting size and grows as T and $\mathcal{S}$ increase. As expected, the spatial bound adds more computation than the temporal one.

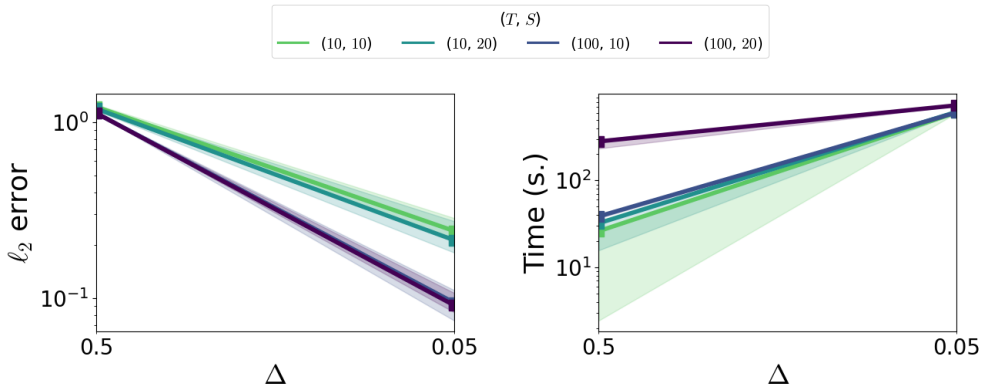


Figure 6.1: Median and 25%-75% quantiles of the ℓ_2-norm between true and estimated parameters (left), and computational time (right) with respect to Δ, for various T and $\mathcal{S}$ (with truncated Gaussian spatial kernel and Kumaraswamy temporal kernel).

Experiment 2. We now select the truncated Inverse Power Law function (6.6) for the spatial kernel and the Kumaraswamy temporal kernel function (6.7). The set of parameters to estimate is $\theta^* = (\mu, \alpha, m, \sigma, a, b)$. In Figure 6.2, we show the median (over 10 runs) and the 25%-75% quantiles of the ℓ_2 estimation error, for various values of T, $\mathcal{S}$ and Δ (with the same values as above).

One can observe that the ℓ_2 error tends toward zero as Δ decreases, as for the Gaussian spatial kernel result in Figure 6.1.

Experiment 3. Furthermore, we investigate the popular case of the truncated Exponential kernel (6.9) for the triggering temporal function, with decay $\lambda = 1$. The spatial kernel is the truncated Gaussian kernel (6.5). Hence, the parameters to estimate are $\theta^* = (\mu, \alpha, m, \sigma, \lambda)$.

Experiment 4. Finally, we select a truncated Gaussian spatial (6.5) and temporal (6.8) kernels, with $m = (0,0)$, $\sigma = 0.1$, $m_T = 0.5$, and $\sigma_T = 0.1$. The set of parameters is $\theta^* = (\mu, \alpha, m, \sigma, m_T, \sigma_T)$.

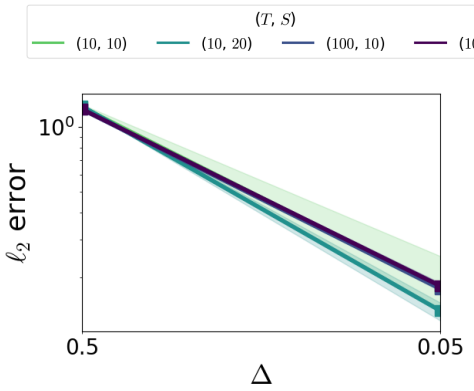


Figure 6.2: Median and 25%-75% quantiles of the ℓ_2-norm between true and estimated parameters with respect to Δ, for various T and $\mathcal{S}$ (with truncated Inverse Power Law spatial kernel and Kumaraswamy temporal kernel).

For the last two experiments, we apply the same procedures as above, and display the results in Figure 6.3: the median (over 100 runs) and the 25%-75% quantiles of the ℓ_2 estimation error $\|\hat{\theta} - \theta^*\|$ are given for Experiment 3 (left) and Experiment 4 (right).

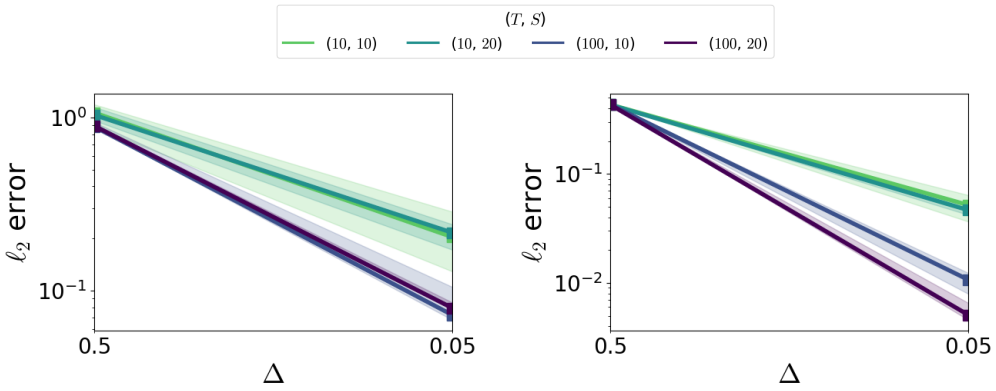


Figure 6.3: Median and 25%-75% quantiles of the ℓ_2-norm between true and estimated parameters, with respect to Δ, for various T and $\mathcal{S}$ (with truncated Gaussian spatial kernel and truncated Exponential (left), and truncated Gaussian (right) temporal kernels).

We observe that, for both settings, the error decreases as the stepsizes increase. Furthermore, the error also decreases with respect to the values of T and $\mathcal{S}$.

We notice that the ℓ_2 error is smaller in the case of a truncated Gaussian temporal kernel. This remark and the observations made for Figure 6.2 support our claims: the method is efficient and flexible. Thus, our approach can be well-suited for applications to real-world data, where the events do not immediately trigger more events and where the triggering structure does not necessarily follow a Gaussian function for the spatial domain.

For the last two experiments, we give some details about each parameter estimation separately.

Details about parameter estimation. In addition, we display the ℓ_2 error for each parameter separately in Figure 6.4 for the Kumaraswamy (Experiment 1) and in Figure 6.5 for the truncated Exponential (Experiment 3) temporal kernels.

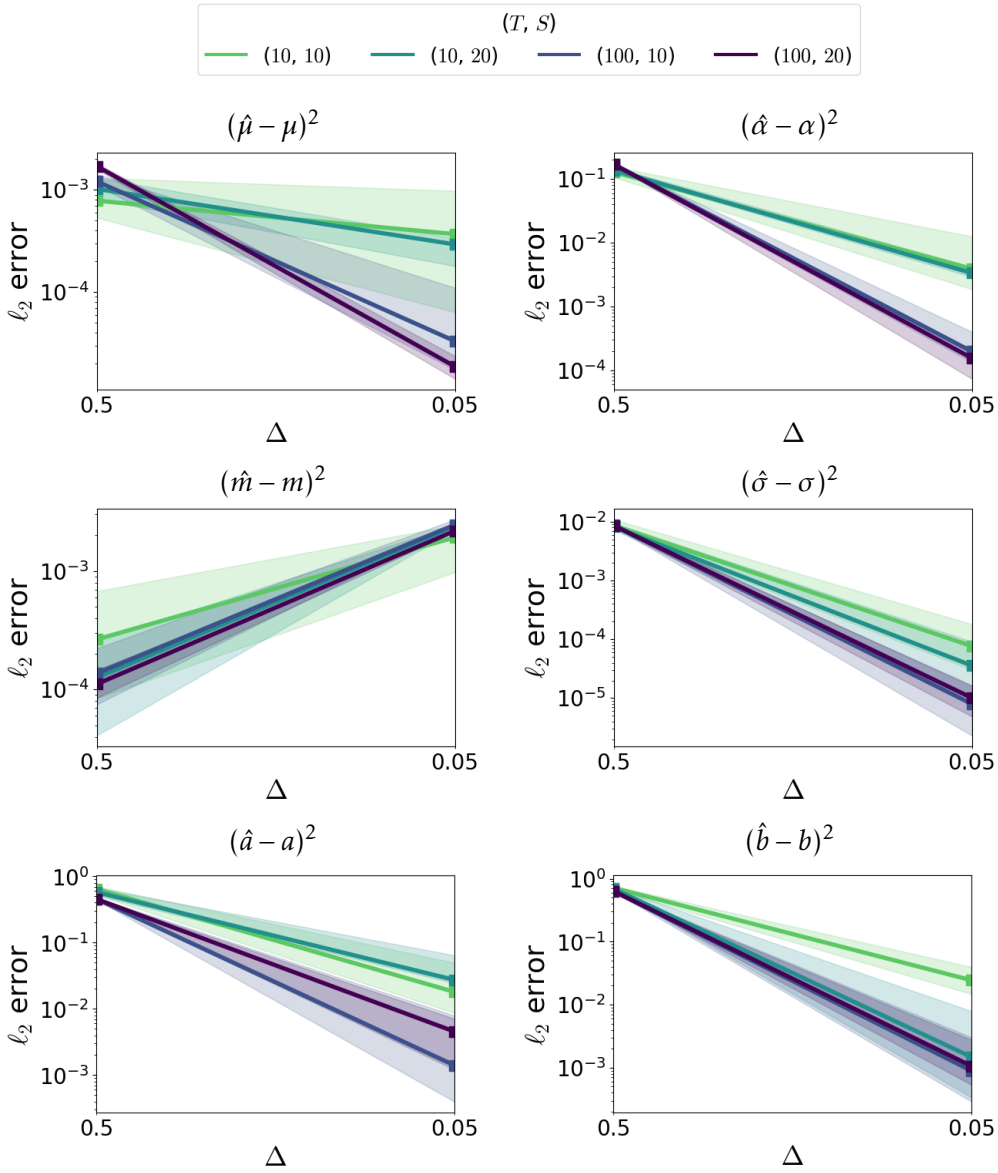


Figure 6.4: Square error on parameters for the Kumaraswamy temporal kernel, as a function of T, $\mathcal{S}$ and Δ.

6.5.2 On the Statistical Error

The statistical error of a STHP is challenging to assess theoretically. To that end, we investigate the statistical error returned by the parameters estimations of our approach based on the values of the ending time T and of the spatial bounds $\mathcal{S}$, assuming a square spatial support.

We simulate a one dimensional STHP with a truncated Gaussian spatial kernel defined as in (6.5), with $m = (0, 0)$ and $\sigma = 0.1$, and a truncated Gaussian temporal kernel (6.8) with mean $m_T = 0.5$, standard deviation $\sigma_T = 0.1$ and finite support length $W_T = 1$.

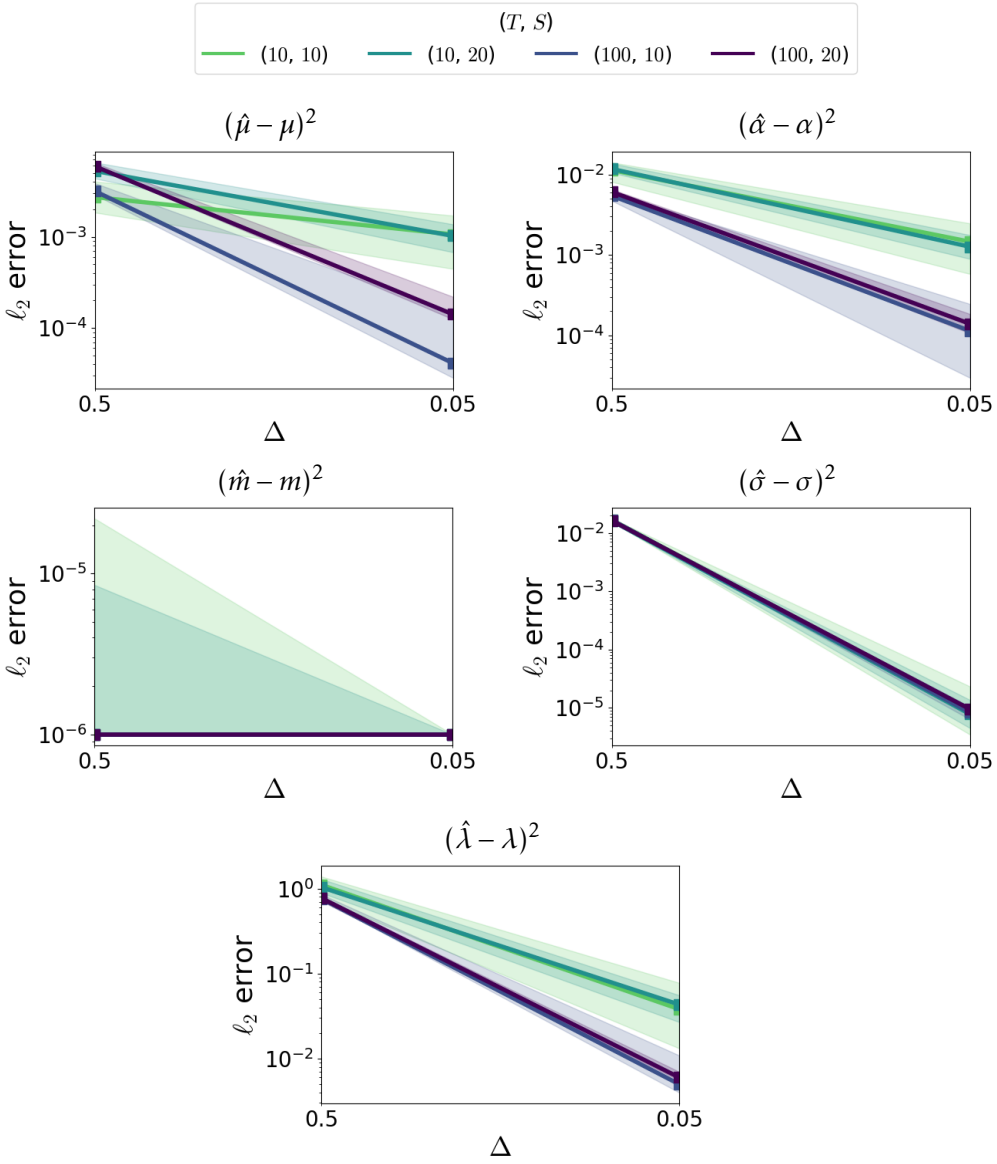


Figure 6.5: Square error on parameters for the truncated Exponential temporal kernel, as a function of T, $\mathcal{S}$ and Δ.

The set of parameters to estimate is $\theta^* = (\mu, \alpha, m, \sigma, m_T, \sigma_T)$. Events are simulated with varying end time and spatial bounds, i.e., $T \in [10, 1000]$ and $\mathcal{S} \in \{10, 20\}$. We compute our proposed approach by fixing $\Delta = (0.1, 0.1, 0.1)$ since we are no longer interested in the discretization bias. We report the median (over 100 runs) and the 25%-75% quantiles of the ℓ_2 estimation error $\|\hat{\theta} - \theta^*\|$ in Figure 6.6 (left), alongside with the computation time with respect to T and $\mathcal{S}$ (right).

We observe that the ℓ_2-norm goes towards zero as T increases. We can see that the spatial bound value has influence on the variance of the error but the convergence is identical w.r.t. the median.

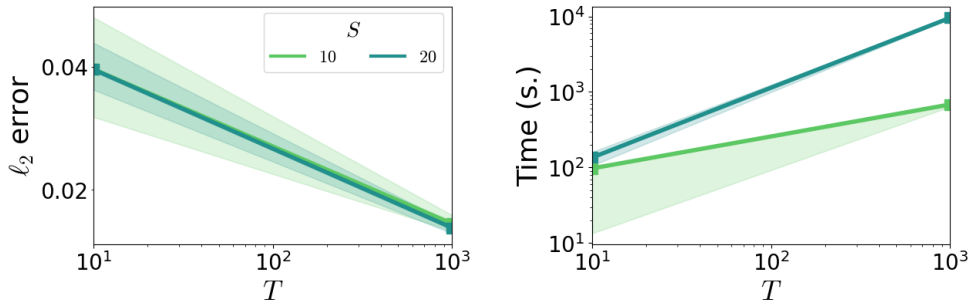


Figure 6.6: Median and 25%-75% quantiles of the ℓ_2-norm between true and estimated parameters (left), and computational time with respect to T (right), for various $\mathcal{S}$ (with truncated Gaussian spatial and temporal kernels).

6.5.3 Approximation of the Bottleneck Precomputation Term

The experiment in this part supports the choice of the approximation of Ψ discussed in subsection 6.3.3. We simulate data with the same setting than in subsection 6.5.2. In order to show the relevance of the chosen approximation, we computed the true precomputation, denoted by Ψ^*, and the approximated one $\widetilde{\Psi}$ for various sizes of T and $\mathcal{S}$. Note that, due to the computational burden from Ψ, the values of T and $\mathcal{S}$ are small. Here, we assess the relative approximation error between Ψ^* and $\widetilde{\Psi}$ with two metrics: the 1-norm $\|\cdot\|_1$ and the Frobenius norm $\|\cdot\|_F$ between tensors. The results are reported in Table 6.1 together with their computation time.

Table 6.1: 1-norm and Frobenius norm (upper) of the difference between the true Ψ^* and approximated $\widetilde{\Psi}$, and their computation time in seconds (lower), for various T and $\mathcal{S}$.

$(T,\mathcal{S})$	$(5,5)$	$(10,10)$	$(50,10)$
$\|\Psi^* - \widetilde{\Psi}\|_1$	0.118	0.039	0.022
$\|\Psi^* - \widetilde{\Psi}\|_F$	0.162	0.062	0.041
Time Ψ^* (s.)	582	4952	25343
Time $\widetilde{\Psi}$ (s.)	0.18	7.5	45.7

The results validate the approximation choice for Ψ: as T and $\mathcal{S}$ grow, the norms tend to 0 with a linear rate. As expected, the computational time for the true Ψ explodes, while it remains feasible for the approximated version.

6.6 Applications to Real Data

To show the flexibility of the approach we propose, we introduce three space-time non-separable kernels, with finite support $[-1,1]^2 \times [0,1]$:

- A function from the class of non-separable spatio-temporal functions proposed in Cressie and Huang (1999):

$$g(x,y,t;a,b,c,\sigma) = \sigma^2 \exp\Big(-at - b^2(x^2+y^2) - ct(x^2+y^2)\Big)\mathbb{I}\{(x,y,t) \in [-1,1]^2 \times [0,1]\}, \tag{6.10}$$

  where a and b are non-negative scaling parameters of time and space (respectively), $c > 0$ is a space-time interaction parameter, and $\sigma^2 > 0$.

- A spatio-temporal function from the Gneiting (2002) class:

$$g(x,y,t;b,r) = \frac{1}{1+bt} \exp\left(-\frac{r(x^2+y^2)}{1+bt}\right)\mathbb{I}\{(x,y,t) \in [-1,1]^2 \times [0,1]\}, \tag{6.11}$$

  where b and r are non-negative scaling parameters of time and space (respectively).

- The space-time non-separable kernel proposed in Zhu et al. (2021):

$$g(x,y,t;\beta,m,\sigma) = \frac{\exp(-\beta t)}{t} \exp\left(-\frac{(x-m)^2+(y-m)^2}{2\sigma^2 t}\right)\mathbb{I}\{(x,y,t) \in [-1,1]^2 \times [0,1]\}, \tag{6.12}$$

  where $\beta > 0$ controls the temporal decay rate and m and σ^2 are the mean and variance parameters.

In this section, we present experiments on two real-world datasets: (1) seismic activity in California and (2) burglaries in Chicago.

6.6.1 Seismic Activity in California

The *Northern California Earthquake Data Center* [1] (NCEDC; nce, 2014) provides time series datasets, collecting information such as location and timing of seismic events in California (see Figure 6.7). Time series data for seismic regions reveal highly complex dependence structures, which can be found between events and between neighboring regions (Ogata, 1999, 1998; Vere-Jones, 1995). The first proposed method to study earthquake occurrences is the *Epidemic Type Aftershock Sequence* (ETAS; Ogata, 1988) model, which only relies on the timing of seismic events and their magnitude, ignoring the spatial dimension. Hawkes processes are well-suited to model earthquake occurrences (Musmeci and Vere-Jones, 1992) due to their self-exciting nature in space and time: an earthquake can trigger further replica in a given period and spatial neighborhood. These triggered events, often referenced as 'aftershocks', can in turn trigger other events. A space-time clustering form is generally observed when studying seismic datasets. We refer the reader to Sections 5.1, 5.2, and 5.5 in Chapter 5 and the examples within for more details about Hawkes processes modeling for seismic data. Actual models usually assume space-time separated kernels with Gaussian density for the space dimension and an exponential density for the time dimension, and thus limit the modeling power of such processes (Schoenberg, 2003; Veen and Schoenberg, 2008; Zhuang, 2011; Fox et al., 2016).

[1] https://ncedc.org/

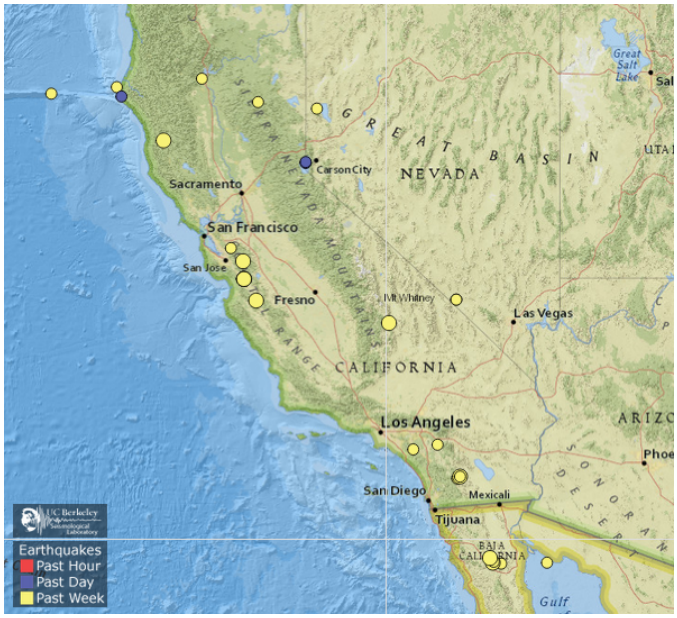


Figure 6.7: Earthquake occurrences from 13 July to the 20 July 2024, in California (USA). The yellow dots represent events occurred from 13/07/2024 to 19/07/2024, the blue dots depict events occurred on 20/07/2024, and the size of each dot gives information about the earthquake's magnitude.
Source: Berkeley Seismology Lab (https://earthquakes.berkeley.edu/seismo.real.time.map.html).

We carry out experiments on three different datasets extracted from NCEDC database, each with a different clustering structure and behavior, and different time lags. The first dataset `1987-1989' contains 605 events, the second dataset `2003-2014' includes 2439 events, and the third dataset `1967-2003' counts 14644 events, defined as seismic records with a magnitude larger than 3.0. Each event is defined by its time and location (no other information is used in our experiments).

We apply the following non-separable and separable kernels:

- Non-separable kernels: the Cressie and Huang (1999) defined in Equation (6.10) and noted NS1, and the Gneiting (2002) defined in Equation (6.11) and noted NS2.

- Separable kernels with various spatial and temporal components: the truncated Gaussian (TG) defined in (6.5) and inverse Power-Law (POW) see (6.6) as spatial kernels and TG, exponential (EXP) and Kumaraswamy (KUM) as temporal ones (defined in Equations (6.8), (6.9) and (6.7), respectively). We denote the overall triggering kernel as SPACE + TIME.

We set $\Delta = (0.05, 0.05, 0.05)$, $W_{\mathcal{X}} = 1$, $W_{\mathcal{Y}} = 1$ and $W_T = 1$, and apply our solver with varying kernel types on the three datasets. For each dataset, we use 80% of the events for training and 20% for testing. We use a discretized version of the Negative Log-Likelihood (NLL) metric, a standard tool to compare models in the point process lit-

erature (Shchur et al., 2021), defined as:

$$NLL_{\mathcal{G}}(\theta, \tilde{\mathcal{H}}_T) = \sum_{i=1}^{D} \left(\Delta_{\mathcal{X}} \Delta_{\mathcal{Y}} \Delta_T \sum_{v_x=0}^{G_{\mathcal{X}}} \sum_{v_y=0}^{G_{\mathcal{Y}}} \sum_{v_t=0}^{G_T} \tilde{\lambda}_i[v_x, v_y, v_t] - \sum_{\tilde{u}_n^i \in \tilde{\mathcal{H}}_T^i} \log \left(\tilde{\lambda}_i \left[\frac{\tilde{x}_n^i}{\Delta_{\mathcal{X}}}, \frac{\tilde{y}_n^i}{\Delta_{\mathcal{Y}}}, \frac{\tilde{t}_n^i}{\Delta_T} \right] \right) \right). \tag{6.13}$$

Table 6.2: NLL values on test sets of various extracted earthquake datasets with several triggering (separable and non-separable) kernels. The best NLL is in bold and the second best is underlined.

Setting	1987 - 1989	2003 - 2014	1967 - 2003
TG + TG	2.77	1.76	0.72
TG + EXP	3.25	2.14	0.65
TG + KUM	2.98	2.66	0.57
POW + TG	2.11	**1.04**	**0.18**
POW + EXP	**1.72**	1.57	<u>0.20</u>
POW + KUM	<u>2.06</u>	<u>1.50</u>	0.29
NS1	3.77	2.68	0.88
NS2	3.77	2.67	0.87

We first apply our solver to the training set in order to estimate the parameters of each model and then compute the NLL on the testing set using the estimated parameters. In Table 6.2, we report the NLL values for different kernels: the best NLL value is in bold while the second best is underlined. The best performance is achieved with the POW + TG model for two datasets, and with the POW + EXP model for one dataset. Furthermore, all three models using the POW spatial function outperform the standard TG + EXP model (as well as the other two models with TG spatial function) for all three datasets. Thus, we achieve better performance when going beyond the traditional exponential and Gaussian kernels. Therefore, our results confirm the limiting modeling power of the models used so far and the advantages of our approach.

We also notice that the two models with a non-separable kernel function (NS1 from Cressie and Huang, 1999 and NS2 from Gneiting, 2002) have a lower performance compared to all other models. A possible reason for this result is that, since the time window is large and the timing of the events are generally far from one another, the space-time interaction may be more difficult to capture.

6.6.2 Burglary in Chicago

The *Chicago Crime Dataset* [2] comprises reported crimes in the City of Chicago from 2001 to the present. The dataset gathers different type of incidents: theft, criminal damage, robbery, burglary, etc. Here, we focus on burglaries (see Figure 6.8).

[2] https://data.cityofchicago.org/Public-Safety/Crimes-2001-to-Present/ijzp-q8t2

Studies on urban burglaries have shown that it is generally possible to observe a clustering structure in such data. This is related to the concept of 'near-repeat victimization' (Mohler et al., 2011), where following a burglary in a neighborhood, the probability of another occurring in the same vicinity within the following days increases. The first Hawkes models to study crimes such as burglaries were motivated by the similarity between the observed self-exciting patterns of crimes and earthquakes (Mohler et al., 2011), see Section 5.5 for further details. However, the models that are currently used in criminology suffer from the same limitations as in seismology, namely the restriction to space-time separable kernels and the traditional Gaussian/exponential form assumption for the spatial/temporal kernels (Mohler et al., 2011; Mohler, 2014; Zhu and Xie, 2022; D'Angelo et al., 2022).

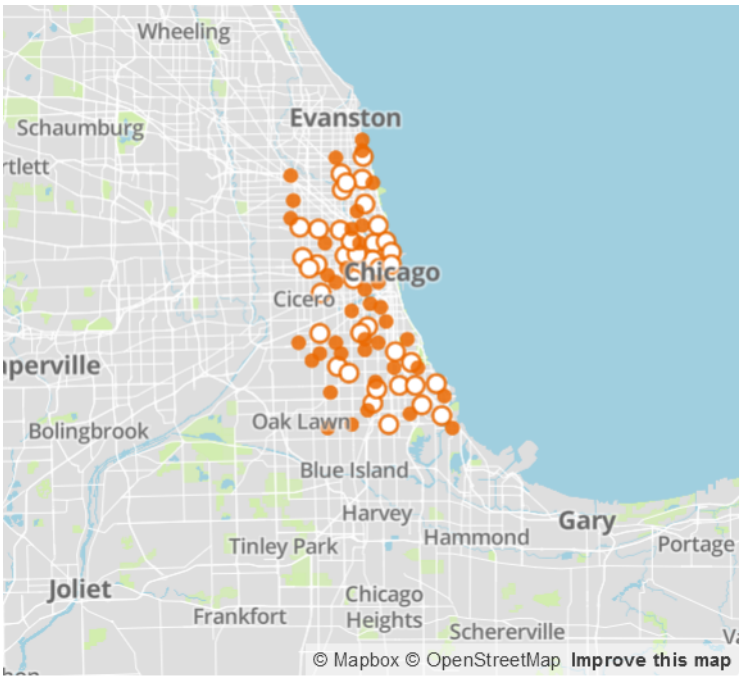


Figure 6.8: Burglary occurrences from 6 July to 13 July 2024, in Chicago (USA). The orange dots represent one crime, while the white dots two or more crimes.
Source: Chicago Data Portal (https://data.cityofchicago.org/stories/s/Crimes-2001-to-present-Dashboard/5cd6-ry5g).

We conduct experiments using three distinct datasets from the Chicago Crime database, each with a different clustering structure and time lag. The first dataset, labeled `2008', includes 4 233 events, the second dataset `2002-2004' contains 23 167 events, while the third dataset `2002-2006' counts 43 822 events. We focus on burglaries occurring in apartments or residences and we collect only the time and location of each event.

We apply the same non-separable and separable kernels as in the previous experiments in subsection 6.6.1, with the addition of the space-time non-separable kernel in Equation (6.12), noted NS3. We set $\Delta = (0.05, 0.05, 0.05)$, $W_{\mathcal{X}} = 1$, $W_{\mathcal{Y}} = 1$ and $W_T = 1$. We use 80% of the data to train our solver and 20% for testing, using the same discretized version of the NLL in (6.13). The NLL values on the test sets for each model are shown in Table 6.3. The best performance is achieved by the non-separable models NS1 and NS2 across all three datasets, closely followed by the non-separable NS3

model. These results prove the strength of our proposed approach in the presence of space-time interactions within the data. Additionally, separable kernels also yield good results, particularly with the TG spatial function.

A possible reason behind the difference between the performance of space-time non-separable kernels in the earthquake and in the burglary applications is the following. For the crime dataset, we have access to a large number of observations across the city and new events are generally observed for each day included in the sample. Indeed, as shown in Figure 6.8, the number of burglaries in a week is particularly high and events tend to concentrate in some neighborhoods (particularly in the areas with white dots, representing two or more events). Thus, with a large number of events that occur closely to one another both in time and space, it is possible to identify a potential spatio-temporal dependence pattern. On the other hand, in the earthquake dataset, there are less observations for a same period of time: in 2 years, from 1987 to 1989, 605 earthquakes were reported, while 23 167 burglaries occurred in the city of Chicago from 2002 to 2004. Furthermore, Figure 6.7 shows that the number of seismic events in one week is lower than the number of burglaries that occur over the same period of time, and the spatial domain of interest is considerably larger (a state, v.s. a city). This means that, for this specific earthquake dataset, no observation is recorded for many time periods and many locations, which arguably makes space-time interactions less relevant for this dataset.

Table 6.3: NLL values on test sets of various extracted burglary datasets with several triggering (separable and non-separable) kernels. The best NLL is in bold and the second best is underlined.

Setting	2008	2002 - 2004	2002 - 2006
TG + TG	-0.24	0.26	0.51
TG + EXP	-0.24	0.38	0.60
TG + KUM	-0.23	0.35	0.54
POW + TG	0.54	1.04	1.10
POW + EXP	1.27	1.03	1.08
POW + KUM	0.83	0.86	0.91
NS1	<u>-0.37</u>	<u>-0.43</u>	<u>-0.28</u>
NS2	**-0.95**	**-0.49**	**-0.31**
NS3	**-0.95**	**-0.49**	**-0.31**

Our numerical experiments with two real-world datasets show the advantages of our method's flexibility. For various applications and types of observed data, our approach achieves accurate estimations due to its ability to infer general parametric kernels and capture space-time interactions.

6.7 Conclusion

Motivated by the growing demand for spatio-temporal data modeling in various fields, we introduce a novel, flexible and efficient approach to infer any parametric kernels in the context of space-time Hawkes processes that are not necessarily independent of space-time. This new approach allows us to overcome our **Challenges 4 and 5** (see Section 1.2 in Chapter 1). Based on ERM-inspired least squares loss for point processes, our framework overcomes the significant computational difficulties inherent in fitting such complex models. It relies on the use of finite support kernels, a discretization scheme and precomputations. After formally proving that the discretization error is minimal, we empirically investigate the statistical efficiency of our approach. As the precomputation terms are expansive, we propose a computationally efficient approximation and show that the error is negligible. Finally, we demonstrate the value of using different kernels to model both earthquake aftershocks and near-repeat victimization patterns (for burglaries), which is possible thanks to the numerical and statistical advantages of our method.

However, our numerical experiments on synthetic data show that the choice of the discretization steps Δ is limited. Indeed, lower values will result in excessively long computational times and may cause memory issues. Therefore, it could be of interest to explore further approximations for the precomputation terms (like in subsection 6.3.3 for Ψ) in order to allow sampling a more precised refinement of the discretization grid, resulting in more accurate predictions.

Key points of this Chapter.

Key Components of our method developed in this Chapter

- **Finite Support Kernels:** the spatio-temporal kernels are assumed of finite length (subsection 6.2.2) ⟶ allows **efficient** intensity computations.
- **Discretization:** we propose a three-dimensional regular grid, with projected events (subsection 6.2.3) ⟶ **reduces the computational burden**.
- **ERM-inspired Least Squares Loss:** see subsec. 5.4.2, Chap. 5 for a presentation ⟶ involves precomputation terms (subsection 6.3.1).
- **Precomputations:** constants independent of the parameters appear in the discretized loss (subsec. 6.3.2) ⟶ greatly **reduces** the computational cost and allows **efficient** gradient-based optimization (subsec. 6.3.4).

Accuracy and Efficiency of our method developed in this Chapter

- **Theoretical guarantees** and **Numerical assessment** on the discretization bias (Prop. 6.3 and subsec. 6.5.1) ⟶ **efficient and flexible** method.
- Numerical assessment of the statistical error (subsec. 6.5.2) ⟶ **accurate** method.

Challenges overcome in this Chapter

- **General Parametric Model:** our approach provides **flexibility in the choice of the kernels**: the results of our numerical experiments for various spatial and temporal kernels (Section 6.5) show this advantage.
- **Space-time Non-separability:** our method **enables** accounting for **space-time interactions/heterogeneity** ⟶ adapted for real-world situations.
- **Efficiency:** based on the key concepts explained in Section 6.2, our method **solves** the modeling and numerical **challenges** posed by parametric models' complexity.
- **Applications to Real Data:** applications to seismic activity in California and to reported burglaries in Chicago ⟶ the results in Section 6.6 prove the advantages of our proposed approach.

Publication

- Emilia Siviero, Guillaume Staerman, Stephan Clémençon, Thomas Moreau. Flexible Parametric Inference for Space-time Hawkes Processes. *arXiv preprint arXiv:2406.06849*, 2024.

7

Conclusion and Perspectives

In this thesis, we investigate different methods to learn from spatial data, taking into account their strong dependence structure. The two main difficulties encountered when dealing with spatial data are: (1) how to account for the strong spatial (or spatio-temporal) dependence structure of the underlying phenomenon of interest, (2) how to develop approaches that can allow one to make accurate inference when a single realization of the phenomenon is observed at a finite number of spatial (or spatio-temporal) points.

The general goal of our work is to answer **Research Questions 1** (see Section 1.2 in Chapter 1) by providing statistical guarantees for prediction methods and developing new, efficient and accurate methods to learn, model, and predict from spatial data:

How to learn from spatial data that presents a strong dependence structure? How does the dependence structure of the observed phenomenon affect the performance of the algorithms?

Our work is divided into two parts. The first part concerns geostatistical data, where a random field is observed at n spatial locations, and the goal is to predict the values of the random field at each unobserved spatial locations. This case is covered in Chapters 3 and 4. The second part deals with point patterns data, where the observation points are considered as events of a point process of interest. In Chapter 6, we present a method for the inference of spatio-temporal point processes.

In the following, we first recall our contributions, present the main limitations of our work, and discuss future lines of research.

Part I: Statistical Learning for Spatial Data

The first part of this thesis, Chapters 3 and 4, aimed at providing theoretical guarantees for the simple Kriging problem, an interpolation method of Geostatistics, answering our **Research Questions 2 and 3**

How accurate is the empirical covariance estimator, based on a finite number of observations on a regular grid and with one unique realization? What is the non-asymptotic behavior of the Kriging predictor when the dependence structure is unknown and with a finite number of observations? To what extent the Kriging weights depend on the accuracy of the covariance function estimation and on the location of samples?

Contributions. Let us recall the main contributions presented in Chapters 3 and 4:

- We derive, under appropriate assumptions, non-asymptotic bounds for the accuracy of the non-parametric covariance estimator for second-order stationary Gaussian processes which covariance function, unknown in practice, is assumed to be isotropic.

- We evaluate the generalization capacity of the empirical Kriging predictor at all unobserved locations within the spatial domain by deriving non-asymptotic tail bounds for the global excess risk associated with the Kriging method, under appropriate conditions.

- We conduct numerous numerical experiments on simulated data using various covariance models, some of which meet the assumptions above, while others do not. Our numerical experiments support our theoretical results for the valid covariance models.

- We show the efficiency of the empirical Kriging predictor through numerical experiments on real meteorological data. Our findings validate the theoretical guarantees and indicate that applying the non-parametric empirical Kriging prediction method can result in robust performance and flexibility.

- The code to reproduce the results is available on GitHub.

Limitations and Future Works. Since our goal is to give a general theoretical framework of the simple Kriging method, simplifying assumptions are employed. As highlighted in the proofs of the main results and in the numerical experiments (see Sections 3.4 and 4.4), even in the simplest framework, the analysis is far from straightforward. In the sketch of proof in subsection 4.3.2 (and in the more detailed proofs in Appendix B), the importance of certain hypotheses is underlined by showing how some hypotheses are necessary for completing the steps of the proofs (see also Figure 4.1 in Chapter 4). In a near future, our objective is to find which assumptions can be relaxed and with which consequences on the learning bounds. We present alternative statistical frameworks for the Kriging problem and discuss possible avenues to relax some of the assumptions involved in our analysis in order to extend our main results to a more general framework.

Alternative Kriging Frameworks:

Let us recall that in our work, we observe a centered Gaussian random field $\mathbf{X}$ at d spatial locations $s_1, \cdots, s_d$ and the goal is to predict its values at all unobserved locations $s \in \mathcal{S}$ in order to compute a complete map. To do so, since the dependence structure of $\mathbf{X}$ is unknown in practice, we estimate the covariance function $c(\cdot)$ by means of a non-parametric estimator $\widehat{c}(\cdot)$, based on a training dataset $\mathbf{X}'$ (a single realization of $\mathbf{X}$) observed at $n \geq 1$ spatial locations $\sigma_1, \cdots, \sigma_n$ of $\mathcal{S}$. To ensure a successful empirical estimation, we assume that the random field is second-order stationary with isotropic covariance function (Assumption 3.2).

The assumptions that the mean of the random field is known and the second-order stationarity assumptions, may not always hold in real situations. Thus, these limitations could be subject of further work aiming at extending our results to an even more general framework.

However, in most practical situations, the mean is unknown. *Ordinary Kriging* is an interpolation method that does not require any knowledge of the mean and is suitable for these cases (see *e.g.* Section 3 in Chiles and Delfiner, 1999). Furthermore, one can relax the second-order stationarity assumption that is classically made and formulate instead a weaker stationarity assumption in such a way that only the variance is assumed to be constant over the spatial domain, while the mean can differ in a deterministic way and so present a spatial trend: this alternative framework is called *Universal Kriging*. Another framework is *Cokriging*, an interpolation method that uses additional observed variables, often correlated with each other and with the variable of interest, to improve the precision of the interpolation.

Irregular Grids:

The non-parametric covariance function estimator is the following

$$\widehat{c}(h) = \frac{1}{n_h} \sum_{(\sigma_i,\sigma_j)\in N(h)} \mathbf{X}'_{\sigma_i}\mathbf{X}'_{\sigma_j},$$

where $N(h) = \left\{(\sigma_i,\sigma_j), \|\sigma_i - \sigma_j\| = h, (i,j) \in [\![1,n]\!]^2\right\}$ is the set of pairs of sites that are at distance h from one another and $n_h = |N(h)|$ denotes its cardinality. In order to obtain an unbiased estimator, we assume that the observations $\sigma_1, \cdots, \sigma_n$ form a regular grid of $\mathcal{S}$.

However, in some practical applications, this assumption does not hold. Thus, extending our results to observations on irregular grids is important and will be the subject of further research.

Unavoidably, the covariance estimation is affected by the spatial configuration of the observations. Therefore, the study of uncertainty effects, on the Kriging predictor, induced by the sampling setting are of major interest (Bardossy, 1988; Müller and Zimmerman, 1999; Wang et al., 2020). In future research, we may consider other types of observation grids, like irregular ones, implying additional technical difficulties, for example when controlling the spectrum of the covariance matrix (see the proof of Proposition 3.9 in Chapter 3). This may also result in an estimation bias for the covariance or may lead to defining different sets $N_\varepsilon(h)$ of neighbors like the set of pairs that are at a distance more or less h (with error $\varepsilon > 0$), see the corresponding paragraph in subsection 2.1.4 in Chapter 2. For a first look at the results that one can obtain within this alternative framework, please refer to subsection 4.6.3 in Chapter 4, where we considered irregular locations under preferential sampling.

Alternative Asymptotic Settings:

Furthermore, in order to have a robust estimation, we make two additional assumptions. First, we assume that new observations are collected in the same fixed spatial domain $\mathcal{S}$, thus adopting an in-fill asymptotic setting. Then, we assume that the true covariance function of $\mathbf{X}$ is equal to zero after a given lag, as stated in Assumption 3.3. Indeed, these two assumptions allows us to ensure that enough elements for the computation of the empirical covariance function $\widehat{c}(\cdot)$ are available (*i.e.* that for each lag h on which the covariance is estimated, n_h is large enough).

As mentioned in subsection 2.1.1 (Chapter 2) and in Remark 4.5 (Chapter 4), other statistical frameworks for Kriging have already been studied. For example, existing studies (Mardia and Marshall, 1984; Sherman and Carlstein, 1994) have proposed

adopting the 'increasing domain' asymptotic (also referred to as the *out-fill* setting), where the spatial domain $\mathcal{S}$ under study becomes wider and wider as the number of observations n grows and a minimum distance between neighboring sampling locations is assumed. A possible avenue for future work includes the consideration of the out-fill setup for investigating the generalization capacity of Kriging predictors over a wider and wider spatial domain. Other researchers (Hall and Patil, 1994; Lahiri, 1999; Lahiri et al., 1999; Putter and Young, 2001) have also considered a hybrid setting, where a combination of in-fill and increasing domain asymptotics point of view is taken, often assuming that both the size of the spatial domain of the observations and the number of observations in each of its subsets grow with n.

Relaxation of Assumption 3.3:

We made Assumption 3.3 as it greatly simplifies our argument, making it more understandable. To relax it would require handling the decay rate of the covariance function. Furthermore, as highlighted in the numerical experiments, even if the Gaussian kernel fails to satisfy Assumption 3.3, the empirical results encourage us to generalize our theoretical analysis to a more general framework, by relaxing this assumption. This also holds for the two additional covariance models that do not satisfy Assumption 3.3, and even tend to zero less quickly than the Gaussian model, namely the exponential and the Matern (when $\nu_m = 3/2$) models.

Relaxation of Assumption 3.10:

Under the aforementioned assumptions, in Chapter 3 we derive Poisson tail bounds for the empirical covariance function for all observed lags h of the sampled regular grid $\sigma_1, \cdots, \sigma_n$. However, since the objective is to predict the values of $\mathbf{X}$ at all unobserved spatial location, we need to extend the covariance function estimation for all lags h. Based on a piecewise constant estimator, the empirical covariance function is extrapolated at unobserved lags. Therefore, to extend the previous theoretical results for unobserved lags, we assume that the covariance function is of class $\mathcal{C}^1$ with gradient bounded by $Q < +\infty$, as stipulated by Assumption 3.10. Thus, non-asymptotic bounds for the covariance function estimation at all lags are obtained in Corollary 3.11, with a term depending on Q.

Another possible extension of our work is to replace Assumption 3.10 by more restrictive regularity assumptions (*e.g.* the mapping $h \in [0,\ 1 - 2^{-j_1}] \mapsto c(h)$ can be assumed of class $\mathcal{C}^2$). Thus, under alternative smoothness hypotheses, the accuracy of other non-parametric estimation techniques can be established, inducing a possibly different bias term in Corollary 3.11, and thus in the excess risk bound in Theorem 4.8.

Part II: Heterogeneity in Space-Time Data – Hawkes models

Contributions. In the second part of this thesis, we develop a new method for the inference of space-time Hawkes processes, establishing the following contributions:

- In Chapter 6, we develop a fast method for estimating kernel parameters in spatio-temporal Hawkes models. Our method supports the integration of various parametric kernels for the triggering function: (1) for space-time separable kernel, it allows for various spatial and temporal kernels, going beyond conventional Gaussian and exponential forms, (2) it makes it possible to use space-time

non-separable kernels, offering a diverse range of options for kernels that capture the space-time interactions present in the data. Thus, our method enhances the precision and adaptability when dealing with complex dependencies in real data.

- We conduct various numerical experiments on simulated data. Our results show the flexibility of our approach: our method accurately performs with different kernels for the spatial and the temporal triggering functions. Furthermore, the numerical experiments show that our method overcomes the computational time challenges posed by parametric STHPs.
- Finally, the advantages of our inference method are proved on both earthquake and crime data. By allowing any parametric kernel for the (space-time separable) triggering function, our approach provides better insights into seismological datasets. Furthermore, criminal data, such as burglaries, tend to exhibit a space-time dependence structure due to the near-repeat victimization pattern observed within such data. The possibility of relying on space-time non-separable kernels thus proves valuable in real-world cases.
- The implementation of our approach is available on GitHub.

These contributions provide answers to the **Research Questions 4 and 5** (see Section 1.2 in Chapter 1):

How to learn from a multivariate spatio-temporal Hawkes process, despite the modeling and numerical challenges posed by parametric STHP's complexity? How to accurately model real-world situations, where space-time interactions occur and where a latency between aftershocks may be observed, by means of Hawkes processes?

Limitations and Future Works. We recall the key concepts of our approach and give new perspectives to investigate in future work. Let $\mathcal{S} \times [0,T]$ be the observation set, with $T \in \mathbb{R}_+$ a stopping time and $\mathcal{S} = \mathcal{X} \times \mathcal{Y} \subset \mathbb{R}^2$ a compact set that contains the locations of the observed events up to time T. We consider a multivariate spatio-temporal Hawkes process, which behavior is entirely characterized by its intensity functions $\forall i \in \{1, \cdots, D\}$

$$\lambda_i(x,y,t|\mathcal{H}_t) = \boldsymbol{\mu}_i + \sum_{j=1}^{D} \sum_{u_n^j \in \mathcal{H}_t^j} \boldsymbol{\alpha}_{ij}\, g_{ij}(x - x_n^j, y - y_n^j, t - t_n^j),$$

where $D \in \mathbb{N}^*$ is the dimension of the process and $\mathcal{H}_T^i$ is the set of (space-time) events for the i-th process. The conditional intensity is composed of a baseline parameter $\boldsymbol{\mu}_i > 0$, an excitation scaling parameter $\boldsymbol{\alpha}_{ij} > 0$, and the spatio-temporal kernel $g_{ij} : \mathcal{S} \times [0,T] \mapsto \mathbb{R}_+$ with parameters $\boldsymbol{\eta}_{ij}$.

Non-constant Baseline:

We assume that the baseline parameter is constant for each process $i \in \{1, \cdots, D\}$, which implies that the background rate of events is uniform across all spatial locations and times. However, in some real-world applications, such as criminology, the

background rate may depend on the spatial location or the time of each event (Zhuang and Mateu, 2019; D'Angelo et al., 2022).

In future work, it may be valuable to include baselines that vary based on spatial or temporal dimensions, or both. For burglary predictions, the baseline could vary spatially according to neighborhood characteristics (such as residential, industrial, etc.), and temporally because of seasonal factors. For instance, burglaries are more common in residential areas and during holiday seasons. Therefore, incorporating a non-constant baseline may improve the accuracy of burglary predictions within a city.

Marked Spatio-temporal Hawkes Processes:

Spatio-temporal Hawkes models provide powerful tools to deal with data that exhibit self-exciting and clustering behavior, such as earthquake data. Another important feature of earthquakes is their magnitude: generally, a mainshock of high magnitude is more likely to trigger subsequent earthquakes than a mainshock of low magnitude. Therefore, taking the magnitude of each event into account could provide additional valuable information for accurate modeling and prediction. Marked processes allow the incorporation of such information by associating each event with a mark.

An interesting avenue for future work is to extend our approach to marked spatio-temporal Hawkes processes. For example, we could define the following conditional intensity for marked spatio-temporal Hawkes processes (Mohler, 2014):

$$\lambda_i(x,y,t,M|\mathcal{H}_t) = \boldsymbol{\mu}_i + \sum_{j=1}^{D} \sum_{u_n^j \in \mathcal{H}_t^j} \boldsymbol{\alpha}_{ij}\, g_{ij}(x-x_n^j, y-y_n^j, t-t_n^j, M-M_n^j),$$

where M_n^j is the mark of the event $u_n^j = (x_n^j, y_n^j, t_n^j, M_n^j)$ of the j-th process. Usually, the kernel is supposed to be marked-separable (Mohler, 2014), so that it can be written as $g_{ij}(x-x_n^j, y-y_n^j, t-t_n^j, M-M_n^j) = g^{(1)}(x-x_n^j, y-y_n^j, t-t_n^j)\, g^{(2)}(M-M_n^j)$. Another promising avenue is to extend the work in Mohler (2014), where marks are used to describe the type of crime for each observed event. This could be done by combining our flexible approach with their marked model. Our method could easily be extended to the case of separable marks, since it will imply only an additional parameter.

Irregular Grid Discretization:

Our work, inspired from the approach proposed in Staerman et al. (2023) for temporal Hawkes processes, relies on three key ingredients. The first ingredient consists of discretizing the spatio-temporal domain of observations into a three-dimensional regular grid. Next, we assume that the kernel functions are of finite length. By incorporating these initial two aspects, the conditional intensity λ_i can be rewritten by replacing the triggering function with its discretized version. This implies that the sum over past events is replaced by a sum over a finite number of grid elements, which depends on the stepsizes of the grids and the size of the finite support of the kernels. By deriving a discretized version of the least squares loss, we identify some constants that do not depend on the model parameters. This leads us to the third key component, which is the precomputation of these terms, reducing the computational cost and allowing our approach to be efficient.

Instead of defining a three-dimensional regular grid, another option could be to define irregular discretization grids, designed according to the particular phenomenon under study or to the prior knowledge of the spatial domain. The definition of the ir-

regular grid must inevitably adhere to specific assumptions that need to be precisely formulated. This will inevitably increase the computational burden: since the irregular grid will not be characterized by a regular stepsize, the sum over past events could not be replaced by a sum over a simple finite number of grid elements. We will need to control the number of grid elements within the finite support of the kernels, which is not straightforward.

Non-separability in Marked Processes:

As shown by the numerical experiments, a major advantage of our method is that it makes it feasible to use any parametric kernel for the triggering function, including the class of space-time non-separable kernels. Another interesting future line of research is to consider marked spatio-temporal Hawkes processes described above and design a method that could handle non-separable marked kernels.

The non-separability characteristic of kernels can be explored also for marked non-separability, thus accounting for space-time and marks interactions (see *e.g.* Schoenberg, 2004; Díaz-Avalos et al., 2014 and other references in subsection 5.4.3 in Chapter 5). Indeed, in the case of earthquakes, the epicenters (*i.e.* the spatial locations of the mainshocks with high magnitude) are generally spatially close from one another, implying a possible dependence between the mark and the spatial location of the events. Therefore, new kernels must be designed to incorporate space-time and mark interactions.

Possible avenues for Future Work.

Part I – Statistical Learning for Spatial Data

- **Alternative asymptotic settings:** out-fill (the spatial domain becomes wider and wider) or hybrid (combining in-fill and out-fill) asymptotic settings.
- **Alternative Kriging frameworks:** ordinary Kriging, universal Kriging.
- **Relaxation of Assumptions:** relax Assumptions 3.3 and 3.10.
- **Irregular grids:** the regular grid of observations could be relaxed to more general ones, which may result in an estimation bias or lead to defining different sets of neighbors.

Part II – Heterogeneity in Space-Time Data – Hawkes Models

- **Spatially or temporally varying baseline:** incorporate baseline depending on spatial or temporal dimensions into our approach.
- **Marked STHP:** extend the efficient approach to marked space-time Hawkes processes, which may deliver valuable additional information for earthquake and burglary predictions.
- **Irregular grid for the discretization:** the discretization grid could be designed according to the specificities of the phenomenon or of the spatial domain.
- **Space-time and mark interactions:** design and incorporation of space-time and mark non-separable kernels.

8

Introduction en Fançais

Contents

8.1 Statistiques Spatiales

En apprentissage automatique, la théorie repose généralement sur le caractère indépendant et identiquement distribué (i.i.d.) des données. En d'autres termes, on suppose que les observations d'un phénomène sont recueillies de manière uniforme, chaque observation étant indépendante des autres. Cela permet d'utiliser des méthodes des statistiques classiques pour construire un modèle précis et robuste et de prédire de nouveaux phénomènes à partir de ces observations. Sous des hypothèses faibles, une solide théorie probabiliste (Devroye et al., 1996; Boucheron et al., 2013), garantissant la capacité de généralisation des règles prédictives apprises empiriquement, renforce ces techniques d'apprentissage automatique.

L'hypothèse d'indépendance s'avère particulièrement pratique. Elle rend l'apprentissage automatique flexible, facilement implémentable, et donc un outil performant avec des algorithmes efficaces. Ces dernières années, une variété de techniques d'apprentissage statistique – y compris les méthodes de boosting, les machines à vecteurs de support, les réseaux de neurones, entre autres – ont été développées avec succès pour effectuer diverses tâches telles que la classification, la régression ou le clustering.

Example 8.1. *(Reconnaissance d'images) La reconnaissance d'images, une branche de la vision par ordinateur, se concentre sur le développement d'algorithmes et de modèles capables d'interpréter et de catégoriser des données visuelles issues d'images. Elle consiste à extraire des caractéristiques et des motifs au sein des images pour reconnaître des objets ou des scènes. La reconnaissance d'images peut englober plusieurs domaines :*

- *Classification d'images : classifier les images en catégories, par exemple en images de chiens ou de chats. Chaque image peut être considérée comme une observation indépendante, en supposant que les images proviennent d'instances différentes et que les caractéristiques extraites de ces images sont identiquement distribuées dans l'ensemble des données.*

- *Détection d'objets : l'objectif est de détecter et localiser des objets dans une image. Si le but est de localiser plusieurs instances du même objet (par exemple des voitures), chaque tâche pour un objet spécifique peut être traitée comme i.i.d., en supposant que les caractéristiques des différentes instances du même objet sont identiquement distribuées.*

Cependant, dans les tâches de reconnaissance d'images telles que la détection d'objets, comprendre le contexte et les relations entre des objets ou des régions dans une image devient crucial. Les objets peuvent avoir des interactions complexes qui ne peuvent pas être modélisées efficacement sous l'hypothèse i.i.d.

La validité de l'apprentissage statistique reste principalement limitée au cas de données d'entraînement i.i.d. Parallèlement, des progrès spectaculaires ont été réalisés dans la collecte, la gestion et le stockage de larges ensembles de données pour des applications scientifiques, médicales ou commerciales, reposant sur des technologies modernes telles que l'imagerie satellitaire ou la tomographie géophysique. Ces données ont tendance à présenter des structures de dépendance complexes, remettant en question l'hypothèse i.i.d.

Nous sommes de plus en plus confrontés à des situations où les données sont de nature spatiale et présentent une forte structure de dépendance. Dans le contexte des données spatiales, des dépendances existent dans toutes les directions. En particulier, les points de données qui sont spatialement proches les uns des autres ont tendance à présenter une corrélation, et la dépendance s'affaiblit à mesure que la distance entre les données augmente.

Example 8.2. *(Météorologie) En météorologie, le but est généralement de comprendre et prédire les phénomènes météorologiques, les tendances climatiques, et les phénomènes atmosphériques. La météorologie joue un rôle crucial dans divers secteurs tels que l'agriculture, le transport et l'énergie. Une des caractéristiques principales des données météorologiques est, bien évidemment, leur nature spatiale. La variabilité spatiale des données météorologiques est influencée par des facteurs comme la géographie, la topographie, ou la proximité des masses d'eau. Les données météorologiques présentent donc une forte structure de dépendance, essentielle tant pour les tâches d'observation que de modélisation. Les dépendances spatiales sont évidentes dans certains cas : les lieux proches ont tendance à avoir des conditions météorologiques similaires (par exemple dans le cas des précipitations, Goovaerts, 2000) et montrent des variations progressives le long des régions géographiques.*

Les premiers modèles adaptés aux données dépendantes sont apparus dans le cas des séries temporelles (Box and Jenkins, 1970; Steinwart and Christmann, 2008; Steinwart et al., 2009; Kuznetsov and Mohri, 2014; Hanneke, 2017; Clémençon et al., 2019). Ces modèles supposent que les observations, identiquement distribuées et se produisant à intervalles de temps réguliers, présentent une dépendance, et que cette dépendance repose sur le flux unidirectionnel du temps. Cela implique que la modélisation dans les études temporelles est causale.

Contrairement au cas des séries temporelles qui reçoit une attention croissante (Steinwart and Christmann, 2009; Kuznetsov and Mohri, 2014), celui des données spatiales est en revanche moins étudié dans la littérature de l'apprentissage statistique. Comme dans le cas des données temporelles, les statistiques spatiales diffèrent des statistiques classiques par les caractéristiques de dépendance des observations. Cependant, les

modèles spatiaux se détachent des modèles temporels de deux façons principales : ils doivent être plus flexibles, car il n'y a pas d'équivalence au concept de passé, présent et futur dans les contextes spatiaux ; et, ils doivent tenir compte de la position spatiale des données recueillies, une information cruciale pour comprendre le phénomène étudié.

Introduisons maintenant le modèle spatial général. Soit $s \in \mathbb{R}^p$ une localisation des données, où $p \in \mathbb{N}^*$. Supposons que $\mathbf{X}_s$ soit une quantité aléatoire. Supposons aussi que s varie sur un ensemble d'indices $\mathcal{S} \subset \mathbb{R}^p$, de manière à générer le processus aléatoire :

$$\mathbf{X} = \left\{\mathbf{X}_s, s \in \mathcal{S}\right\}. \tag{8.1}$$

Nous distinguons X la variable étudiée, également appelée variable régionalisée par Matheron (1965), et $\mathbf{X}$ la modélisation de X par un champ aléatoire. Avec cette notation, X est une réalisation de $\mathbf{X}$.

Les hypothèses sur $\mathcal{S}$ peuvent varier : il peut s'agir d'un sous-ensemble fixe (non aléatoire) de $\mathbb{R}^p$, ou d'un ensemble aléatoire (ce qui implique que $\mathcal{S}$ peut varier d'une réalisation à l'autre). Les études en statistiques spatiales peuvent être divisées en trois catégories, selon la nature de l'ensemble d'indices $\mathcal{S}$ (Cressie, 1993, Chapitre 1) :

1. **Données géostatistiques :** $\mathcal{S}$ est un sous-ensemble fixe de $\mathbb{R}^p$ ($\mathcal{S}$ est continu) et $\{\mathbf{X}_s, s \in \mathcal{S}\}$ est un vecteur aléatoire à la localisation $s \in \mathcal{S}$. On suppose que le champ aléatoire est observé en n points fixes $\{s_1, s_2, \cdots, s_n\}$. Les observations peuvent être soit échantillonnées aléatoirement sur $\mathcal{S}$, soit sélectionnées sur une grille régulière. La géostatistique traite des tâches telles que la modélisation, la prédiction (appelée krigeage) sur un site non observé s, et la construction d'une carte complète du champ aléatoire sur l'ensemble du domaine $\mathcal{S}$.

   Example 8.3. *(Exploitation minière) La géostatistique a émergé comme une étude interdisciplinaire impliquant à la fois l'ingénierie minière et les statistiques. Les méthodes précédentes employées dans les mines utilisaient souvent des histogrammes des teneurs en minerai, en se concentrant uniquement sur le taux de ces échantillons, négligeant ainsi la position spatiale des observations. Pourtant, la localisation spatiale ainsi que les schémas possibles (comme le regroupement) dans le gisement sont des informations précieuses dans les opérations minières. Matheron (1963) a proposé la géostatistique comme une nouvelle approche pour estimer les teneurs en minerai et les réserves de minerai dans les opérations minières. Basé sur un ensemble d'observations sur le gisement minier, il a développé une méthode de prédiction qui prend en compte la position spatiale des échantillons, ainsi que la structure de dépendance des teneurs en minerai. Voir l'Exemple 2.1 dans le Chapitre 2 pour plus de détails sur la méthodologie de la géostatistique et ses applications dans les mines.*

2. **Données latticielles :** $\mathcal{S}$ est une collection fixe de points dénombrables de $\mathbb{R}^p$ ($\mathcal{S}$ est discret) et $\mathbf{X}_s$ est un vecteur aléatoire à la localisation $s \in \mathcal{S}$. Les données sont liées à des unités ou régions spatiales, formant ainsi un réseau. Dans les données en réseau, on peut s'intéresser à l'étude de la corrélation spatiale, la prédiction, ou par exemple l'analyse et la restauration d'images.

   Example 8.4. *(Analyse d'images) Dans le contexte de l'analyse d'images et de la restauration d'images, la méthodologie des données en réseau s'applique, car les images peuvent être vues comme une grille de pixels, chaque pixel représentant une unité*

spatiale. Les méthodes de données en réseau peuvent aider à comprendre la structure spatiale et les dépendances au sein de l'image, et à mesurer la corrélation spatiale (ce qui est informatif pour comprendre comment la valeur du pixel dans une image est corrélée avec ses voisins). Voir par exemple Cressie, 1993, Section 7.4 et Besag (1974); Ripley (2005) pour plus de détails.

3. **Données ponctuelles :** $\mathcal{S}$ est un processus ponctuel dans $\mathbb{R}^p$ ($\mathcal{S}$ est aléatoire) et $\mathbf{X}_s$ est un vecteur aléatoire à la localisation $s \in \mathcal{S}$. Ici, les emplacements de données des observations $\{s_1, s_2, \cdots, s_n\}$ et le nombre d'observations n sont aléatoires. Dans ce cas, les emplacements de données contiennent l'information désirée, et les points d'observation sont considérés comme des événements d'un processus ponctuel. Le statisticien cherchera à capturer un schéma dans les données, tel qu'un regroupement, un déclenchement, une régularité, ou de l'aléatoire complet.

   Example 8.5. *(Sismologie) En sismologie, des schémas de regroupement apparaissent typiquement : certaines régions sont particulièrement affectées par les tremblements de terre, tandis que d'autres peuvent ne jamais en subir. Ogata (1988) a introduit les processus ponctuels pour étudier les occurrences d'un tremblement de terre dans une région donnée. Les événements présentent aussi un schéma de déclenchement : un tremblement de terre peut en déclencher d'autres, appelés répliques. Par conséquent, pour des raisons de sécurité, il est essentiel d'identifier les schémas de regroupement pour déterminer les régions les plus touchées et de comprendre les relations de déclenchement pour prévenir les dégâts futurs. Il est également possible d'utiliser un processus marqué, où la marque peut représenter la magnitude du tremblement de terre. Voir les sous-sections 5.1 et 5.2 pour plus de détails sur les études sismologiques en statistiques spatiales.*

Dans cette thèse, nous choisissons de nous concentrer sur deux catégories de données spatiales : les données géostatistiques et les données ponctuelles.

8.2 Motivations

Dans cette section, nous présentons les questions de recherche qui ont motivé cette thèse ainsi que les défis qui en découlent.

Le principal objectif de cette thèse est de développer des méthodes permettant de prendre en compte la forte structure de dépendance des données spatiales, en se basant sur une observation du phénomène en un nombre fini de localisations spatiales, afin de modéliser, prédire et apprendre des données spatiales. La thèse se divise en deux grandes parties : la première concerne une méthode géostatistique et vise à fournir des garanties théoriques pour cette méthode de prédiction, tandis que la seconde partie se concentre sur la conception d'une nouvelle approche pour une catégorie spécifique de processus spatio-temporels. Bien que la géostatistique et les processus ponctuels soient deux catégories distinctes de statistiques spatiales, différant par leur hypothèse sur le domaine spatial, elles partagent en partie des motivations et des difficultés communes.

Domaines d'application. La plupart des éléments qui nous entourent possèdent une dimension spatiale. Cela inclut les phénomènes naturels, comme le climat et les catastrophes naturelles, ainsi que les infrastructures humaines, telles que les puits d'eau

et la planification urbaine. Tous sont fortement influencés par des facteurs spatiaux. Par exemple, à petite échelle, les villes géographiquement proches tendent à connaître des conditions climatiques similaires. De même, l'organisation optimale des rues et de la circulation, cruciale pour le bon fonctionnement d'une ville, nécessite de prendre en compte les interdépendances du trafic entre différents quartiers de la ville.

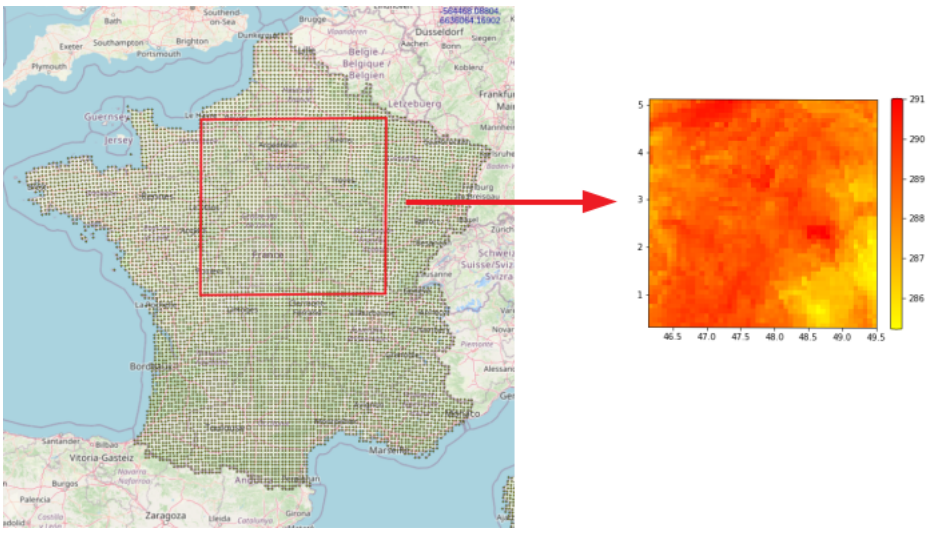


Figure 8.1: Carte de la France avec une grille carrée échantillonnée (à gauche) ; carte des températures (en Kelvin) de la grille carrée échantillonnée, 2 juin 2005 (à droite). Source : DRIAS (https://drias-prod.meteo.fr/okapi/accueil/okapiWebDrias/).

Prenons l'exemple de la Figure 8.1 qui montre la température moyenne journalière en France (exprimée en Kelvin). À droite, nous voyons une carte colorée des températures observées en chaque point dans la grille échantillonnée de gauche. Une zone jaune indique des valeurs de température légèrement plus basses, tandis que des zones rouges indiquent des températures plus élevées (la différence entre ces températures est relativement faible). On remarque des changements progressifs de la température sur tout le domaine spatial ainsi que la présence de zones avec des valeurs similaires. Modéliser correctement ces interdépendances de température est essentiel pour améliorer les prévisions météorologiques, qui sont cruciales pour anticiper des situations extrêmes telles que des tempêtes, des cyclones ou des périodes de sécheresse sévère. Prenons maintenant l'exemple de la Figure 8.5. Cette carte montre tous les séismes enregistrés dans la région volcanique des Champs Phlégréens (à l'ouest de Naples, Italie) de janvier à juillet 2024. On constate une concentration importante de l'activité sismique dans cette zone, principalement le long de la côte près du volcan dormant d'Agnano. L'occurrence de potentiels tremblements de terre est associée à un phénomène volcanique cyclique causant des phases de soulèvement et de subsidence du sol. C'est précisément durant les phases de soulèvement qu'une augmentation de l'activité sismique est observable. Bien que la majorité de ces séismes soient de faible magnitude (en comparaison avec le séisme qui a eu lieu en Turquie et en Syrie en février 2023, voir Figure 8.4), leur forte concentration dans cette région pourrait indiquer une possibilité d'éruption imminente ou de séismes plus puissants. Il est donc crucial de prédire avec précision les futurs événements sismiques en tenant compte des données historiques et de la distribution spatiale.

Nous n'avons mentionné que deux exemples de données spatiales (prévisions météorologiques et prédiction des tremblements de terre). Il existe de nombreux autres domaines d'application, dont certains sont (brièvement) abordés dans cette thèse, comme les mines, l'hydrologie, l'écologie, l'épidémiologie, la finance et la criminologie.

Violation de l'hypothèse d'indépendance i.i.d. Les techniques statistiques classiques supposent généralement que les observations d'un phénomène sont indépendantes et identiquement distribuées. Cependant, en raison de la présence d'une structure de dépendance dans le cadre des données spatiales, l'hypothèse d'observations i.i.d. n'est pas satisfaite. De nouvelles méthodes et des résultats théoriques doivent donc être établis dans ce contexte. La principale difficulté pour apprendre à partir de données spatiales est d'obtenir des informations sur la structure de dépendance sous-jacente, afin qu'elle puisse être prise en compte lors de la modélisation et de la prédiction à partir de ces données.

Cela nous mène à notre première question de recherche et au défi correspondant, qui constituent les objectifs généraux de cette thèse.

Questions de Recherche 1 : Comment apprendre à partir de données spatiales présentant une **forte structure de dépendance** ? Comment la structure de dépendance du phénomène observé affecte-t-elle la **performance** des algorithmes ?

Défi 1 : Fournir des garanties statistiques pour les méthodes utilisées pour prédire les données spatiales. Développer des méthodes nouvelles, efficaces et précises pour prédire à partir de données spatiales.

Comme mentionné précédemment, la première partie de cette thèse concerne les données géostatistiques et vise à contribuer à relever le premier défi des données spatiales en fournissant des garanties statistiques pour les méthodes de prédiction spatiale.

Données géostatistiques avec une unique réalisation du phénomène. Apprendre à partir de données géostatistiques implique deux principaux défis. Le premier, déjà mentionné, est la présence d'une forte structure de dépendance dans les données. Le second est le fait que, généralement, une seule réalisation du phénomène est disponible. Par exemple, un événement naturel spécifique, tel qu'une tempête, ne se produit qu'une seule fois, et aucune autre réalisation indépendante de celui-ci ne peut être observée. D'autres exemples incluent le coût économique élevé de la collecte de données et la possible dégradation de l'environnement. C'est le cas dans le jeu de données hydrogéologiques présenté à la Figure 8.2. L'hydrogéologie vise à évaluer la qualité des eaux souterraines (voir l'Exemple 2.2 dans le Chapitre 2 pour plus de détails) en fonction des caractéristiques de l'eau, telles que le niveau de pH, la conductivité de l'eau et la température. Pour cela, des observations sont collectées sur une région spatiale, ici dans le département de La Guajira en Colombie, et des mesures spécifiques sont effectuées. Cependant, cette procédure implique un coût économique significatif.

En géostatistique, un phénomène est modélisé par un champ aléatoire, supposé observé en un nombre fini de localisations sur le domaine spatial $\mathcal{S} \subset \mathbb{R}^p$. Les caractéristiques de dépendance des données sont modélisées par la fonction de covariance du champ aléatoire. Notre cadre est le suivant : nous sommes intéressés par la prédiction des valeurs aux emplacements spatiaux $s \in \mathcal{S}$ d'un champ aléatoire $\mathbf{X}$, observé

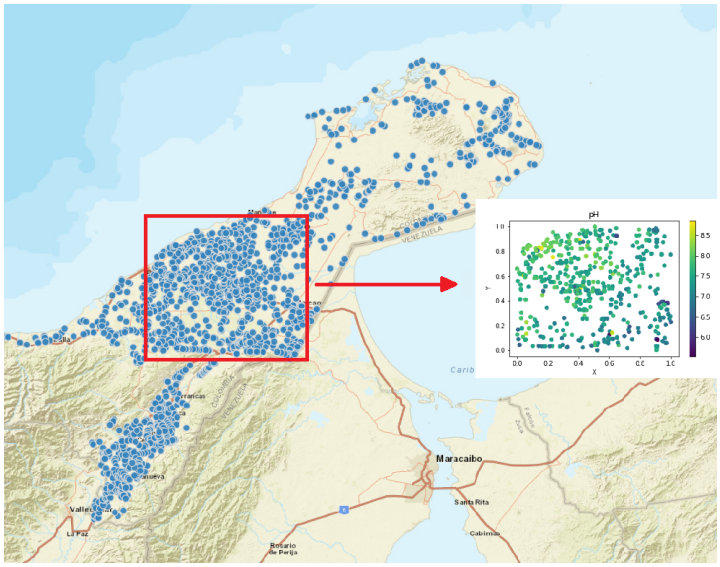


Figure 8.2: Carte hydrogéologique du département de La Guajira (Colombie) en 2016. Chaque point sur cette carte représente un plan d'eau. À droite, la carte des échantillons représentant les valeurs de pH à chaque point spatial.
Source : Servicio Geológico Colombiano (https://datos.sgc.gov.co/). Carte créée en 2016 par le Groupe des Eaux Souterraines du Service Géologique Colombien.

sur un ensemble de localisations fixes $\{s_1, \ldots, s_d\}$. Le champ aléatoire est supposé être un processus aléatoire gaussien stationnaire d'ordre deux avec une fonction de covariance isotrope. Ces hypothèses sont souvent formulées en géostatistique car elles assurent une bonne approche fréquentiste. La méthode d'interpolation résultante est appelée *Kriging*, ou Krigeage (Matheron, 1962), et vise à construire un prédicteur $\widehat{\mathbf{X}}_s$ de $\mathbf{X}_s$, défini comme une combinaison linéaire des observations. Les poids du Krigeage dépendent de la fonction de covariance du champ aléatoire.

Estimation non-paramétrique de la covariance. Lorsqu'on travaille avec des données réelles, la fonction de covariance est inconnue. À partir d'un ensemble d'entraînement, défini comme une unique réalisation $\mathbf{X}'$ de $\mathbf{X}$ observée en n localisations $\{\sigma_1, \ldots, \sigma_n\}$, la fonction de covariance peut être estimée. Des résultats précédants concernant l'estimation de la fonction de covariance ont été développés, soit dans une perspective asymptotique (Stein, 1999), soit par une approche paramétrique (Zimmerman, 1989; Zimmerman and Cressie, 1992). En revanche, nous sommes intéressés par le comportement pour un échantillon fini de l'estimateur non-paramétrique de la covariance, dans un cadre asymptotique dit *in-fill* (c'est-à-dire en supposant que de nouvelles observations apparaissent dans le même domaine spatial fixe, qui devient de plus en plus dense).

Cela nous amène à notre deuxième question de recherche et au défi qui lui est associé, visant à définir la précision de l'estimation non-paramétrique de la covariance.

Question de Recherche 2 : Quelle est la **précision** de l'estimateur empirique de la covariance, basé sur un **nombre fini d'observations** sur une grille régulière et une **unique réalisation** ?

Défi 2 : Obtenir des bornes non asymptotiques pour l'**estimateur non-paramétrique de la fonction de covariance**, dans le cadre asymptotique in-fill.

Le Défi 2 peut être considéré comme un défi intermédiaire pour répondre à notre Question de Recherche 3 (voir ci-dessous), à savoir établir des garanties non asymptotiques pour la méthode de Krigeage. En effet, comme expliqué ci-dessous, la précision de la méthode de Krigeage de prédiction dépend de la qualité de l'estimation de la covariance. Ainsi, il est primordial de calculer une estimation précise de la fonction de covariance et d'identifier les potentiels effets d'incertitude de cette estimation sur le prédicteur de Krigeage.

Garanties non asymptotiques pour la méthode de Krigeage empirique. Lorsque la structure de dépendance du champ aléatoire est connue, la méthode de Krigeage est optimale (nous l'appelons *Krigeage théorique*). Cependant, dans le cas d'une fonction de covariance inconnue, la capacité de généralisation de la méthode résultante (nous l'appelons *Krigeage empirique*) reste à établir. L'objectif est de développer un nouveau cadre théorique offrant des garanties non asymptotiques pour les prédictions par Krigeage simple empirique. Les garanties de généralisation du prédicteur empirique sont fournies sous la forme d'une borne sur l'excès de risque global. Ce risque est défini comme l'écart global entre les erreurs de prédiction des prédicteurs de Krigeage théorique et empirique.

Questions de Recherche 3 : Quel est le comportement **non asymptotique** du prédicteur de Krigeage lorsque la structure de dépendance est **inconnue** et avec un nombre fini d'observations ? Dans quelle mesure les poids de Krigeage dépendent-ils de la précision de l'estimation de la fonction de covariance et de la localisation des échantillons ?

Défi 3 : Établir des bornes **non asymptotiques** pour **l'excès de risque global** de la méthode de Krigeage. Ces résultats théoriques doivent dépendre de l'estimation de la fonction de covariance et du choix du cadre d'échantillonnage.

La seconde partie de cette thèse concerne les données ponctuelles, et plus précisément les processus de Hawkes spatio-temporels. Dans ce contexte, les observations sont considérées comme des événements d'un processus. Les processus de Hawkes trouvent des applications dans divers domaines, tels que l'étude des catastrophes naturelles, comme expliqué ci-dessous.

Prédiction de séismes et évaluation des risques. En 2023, le nombre total de décès dus aux séismes s'élevait à 62 451, selon le rapport annuel du CRED (Centre de recherche sur l'épidémiologie des catastrophes), soit presque le double de la moyenne des vingt dernières années (voir Figure 8.3). Ce nombre était particulièrement élevé cette année-là en raison du séisme survenu en Turquie et en Syrie en février 2023, qui a

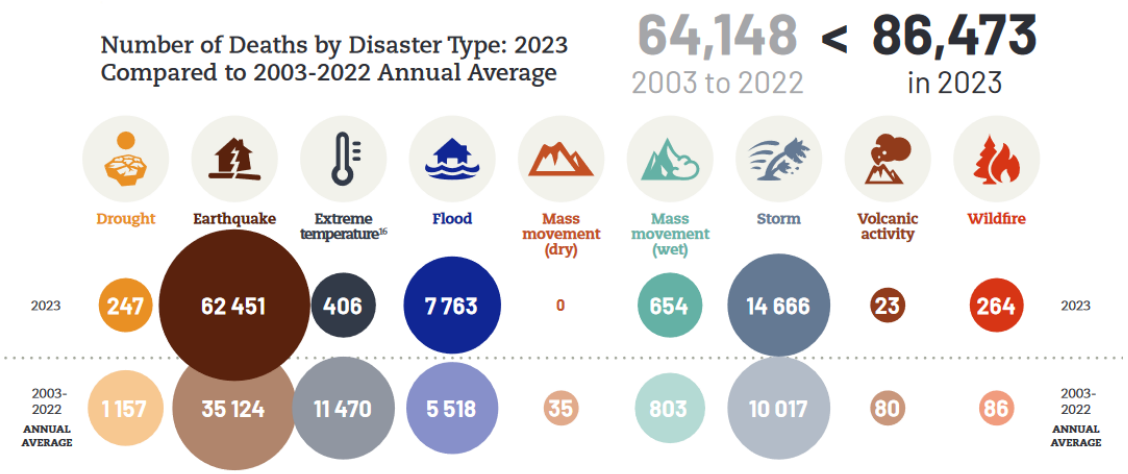


Figure 8.3: Nombre de décès par type de catastrophe : comparaison entre le nombre en 2023 et la moyenne annuelle de 2003 à 2022.
Source : EM-DAT (https://www.emdat.be/), rapport annuel du CRED, *2023 Disasters in numbers*.

été, comme l'écrit le CRED, « *l'événement le plus catastrophique de l'année en termes de mortalité et de dommages économiques, représentant les deux tiers des décès totaux* ». Le 6 février 2023, une séquence de séismes a frappé le sud-est de la Turquie, à la frontière avec la Syrie. Un premier séisme majeur de magnitude 7,8 a frappé près de la ville de Gaziantep, suivi de répliques de moindre magnitude, touchant toutes les régions environnantes (voir Figure 8.4). Cette région est fréquemment sujette à des événements sismiques. Il est donc urgent de disposer de modèles robustes et précis pour prédire l'activité sismique et améliorer l'évaluation des risques dans les régions fortement impactées par les séismes. Une autre région de ce type est les Champs Phlégréens, dont l'activité sismique en 2024 est représentée dans la Figure 8.5 (où la taille d'un point représente la magnitude de l'événement et la couleur indique la période de temps à laquelle il s'est produit).

En observant les Figures 8.4 et 8.5, on constate clairement un comportement de regroupement des événements. En effet, dans la Figure 8.5, les événements sont regroupés dans les dimensions temporelle et spatiale : les points de même couleur sont rassemblés dans une même région spatiale. Cette observation révèle les caractéristiques de déclenchement des séismes. En effet, un premier séisme majeur de forte magnitude (appelé séisme principal) peut déclencher une nouvelle occurrence, généralement de moindre magnitude (appelée réplique). Cet effet de déclenchement se manifeste principalement par un motif de regroupement, c'est-à-dire que l'apparition de nouvelles occurrences se produit dans une fenêtre temporelle spécifique et dans un certain voisinage spatial de l'épicentre initial. Les comportements de déclenchement et de regroupement d'un phénomène sismique sont donc cruciaux pour comprendre l'activité sismique sous-jacente et améliorer la prédiction des événements futurs.

Processus de Hawkes spatio-temporels. Parmi les processus ponctuels, les modèles de Hawkes (Hawkes, 1971) ont récemment suscité beaucoup d'attention, car ils prennent en compte de manière très flexible la nature auto-excitante des événements

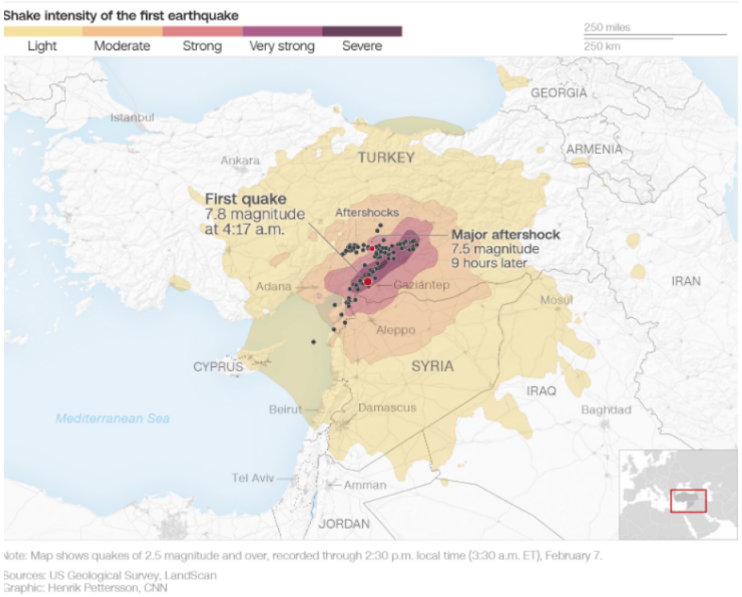


Figure 8.4: Le séisme en Turquie et en Syrie, le 6 février 2023. Les épicentres sont en rouge, et l'échelle de couleurs montre l'intensité de la secousse du premier séisme dans la région.
Source : USGS (https://www.usgs.gov/), United States Geological Survey, LandScan.

observés, l'interaction spatio-temporelle et l'anisotropie spatiale. Avec des fonctions d'intensité bien choisies, la probabilité d'occurrence d'événements futurs sur une période donnée augmente avec ces processus ponctuels (Reinhart, 2018). Vere-Jones (1970) et Ogata (1988) ont introduit ces processus en sismologie en raison du comportement de déclenchement des séismes. En effet, les modèles de *Séquence de Répliques de Type Épidémique* (ETAS, Epidemic-Type Aftershock Sequence) sont bien adaptés à la modélisation des activités sismiques, car ils impliquent que chaque séisme peut initier des répliques, lesquelles peuvent à leur tour engendrer d'autres répliques, aboutissant à une réaction en chaîne d'activité sismique. Ainsi, les processus de Hawkes s'avèrent être des outils puissants pour les données présentant une nature auto-excitante. Cependant, les premiers modèles de Hawkes étaient purement temporels (Ogata, 1988), négligeant ainsi la dimension spatiale du phénomène. En effet, comme observé dans la Figure 8.5, les dynamiques complexes des séismes montrent des regroupements à la fois spatiaux et temporels. Dans cette thèse, nous explorons les processus de Hawkes spatio-temporels (STHPs, Space-Time Hawkes Processes) pour prendre en compte les dépendances spatio-temporelles entre les événements.

Les propriétés auto-excitatrices des processus de Hawkes spatio-temporels expliquent pourquoi ils sont de plus en plus utilisés dans de nombreux domaines nécessitant une analyse spatio-temporelle, tels que l'épidémiologie (Holbrook et al., 2022; Kresin et al., 2022; Rambhatla et al., 2022; Dong et al., 2023), la criminologie (Mohler et al., 2011; Mohler, 2014; D'Angelo et al., 2022; Zhu and Xie, 2022) et la sismologie (Ogata, 1998; Musmeci and Vere-Jones, 1992; Kwon et al., 2023), par exemple. Le principal défi méthodologique consiste alors à concevoir des techniques d'inférence efficaces pour ajuster les modèles de Hawkes à des ensembles de données spatio-temporels.

Question de Recherche 4 : Comment **apprendre d'un processus de Hawkes spatio-temporel multivarié**, malgré les **défis** de modélisation et les défis numériques posés par la complexité des STHP paramétriques ?

Défi 4 : Développer une **nouvelle méthode efficace et flexible** pour l'inférence paramétrique pour les processus de Hawkes spatio-temporels, consistant en un solveur rapide basé sur le gradient ℓ_2.

Pour des raisons de simplicité et de calcul, la plupart des méthodes précédentes sont limitées aux noyaux séparables dans le temps et l'espace, où le noyau temporel est souvent choisi comme exponentiel et l'influence spatiale est modélisée par un noyau Gaussien (Mohler, 2014; Yuan et al., 2019; Ilhan and Kozat, 2020).

Noyaux paramétriques généraux. Le noyau temporel exponentiel généralement assumé, bien qu'il apporte une efficacité de calcul, implique des limitations majeures dans les situations réelles, car il suppose qu'un événement déclenche immédiatement un événement futur. Cependant, dans le cas des séismes, cette hypothèse est généralement invalide. Par exemple, lors de l'activité sismique ayant touché la Turquie et la Syrie en 2023, un premier séisme principal (de magnitude 7,8) a eu lieu vers 4 h du matin. Un deuxième séisme principal (de magnitude 7,5) est survenu 9 heures plus tard, vers 13 h. Les épicentres de ces deux séismes sont spatialement proches, comme le montre la Figure 8.4. Dans ce cas, le noyau exponentiel ne convient pas, car une latence est observée entre les deux séismes majeurs. À l'inverse, le 20 mai 2024, plusieurs séismes ont frappé la région des Champs Phlégréens entre 19 h 51 et 21 h 55 (de magnitudes comprises entre 3,1 et 4,4), impliquant une influence plus immédiate. Ainsi, selon plusieurs facteurs (tels que les plaques tectoniques sous-jacentes de la région, la présence d'un volcan, etc.), les comportements temporels de déclenchement et de regroupement des séismes peuvent varier d'une région à l'autre. De plus, pour la dimension spatiale, la dispersion des répliques dans la Figure 8.5 ne semble pas suivre une distribution Gaussienne.

Interactions spatio-temporelles. La séparabilité spatio-temporelle du noyau d'un processus de Hawkes est une hypothèse courante (voir par exemple Mohler, 2014; Yuan et al., 2019; Ilhan and Kozat, 2020). En effet, elle apporte de la simplicité, car elle implique que le noyau est un produit d'influences spatiales et temporelles qui peuvent être modélisées séparément. Cependant, lorsqu'on traite de phénomènes naturels tels que les séismes, une interaction spatio-temporelle peut généralement être observée.

Ces deux limites des approches précédentes motivent la nécessité d'une nouvelle méthode efficace et flexible pour modéliser les processus de Hawkes spatio-temporels. Cette nouvelle méthode doit être adaptée à des noyaux paramétriques généraux et à des noyaux non séparables spatio-temporellement, permettant une meilleure prédiction basée sur les caractéristiques du domaine spatial ou du phénomène étudié.

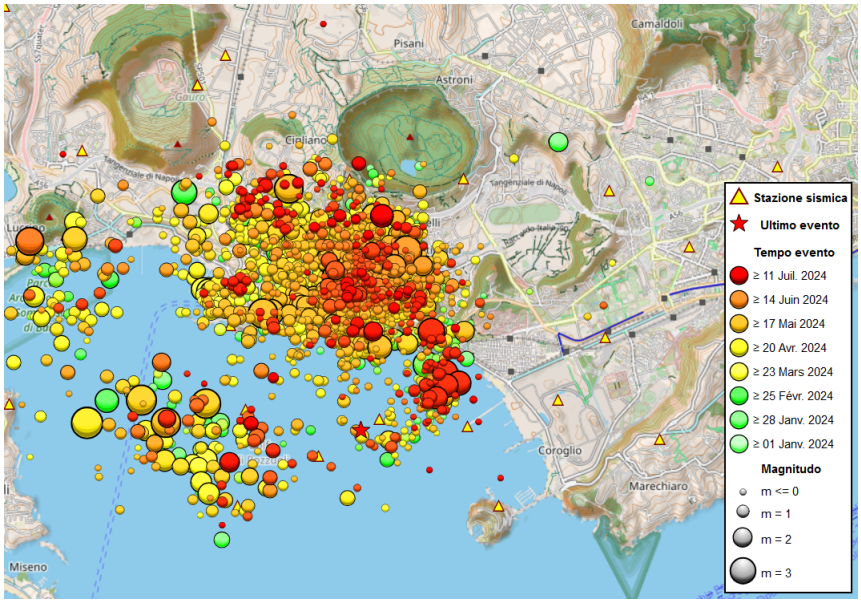


Figure 8.5: Occurrences de séismes en 2024, dans les Champs Phlégréens, à l'ouest de Naples (Italie). L'échelle de couleurs représente le moment de chaque événement et la taille du point, sa magnitude.
Source : INGV (https://terremoti.ov.ingv.it/gossip/flegrei/2024/), Institut National de Géophysique et de Volcanologie.

Question de Recherche 5 : Comment **modéliser avec précision des situations réelles**, où des **interactions spatio-temporelles** se produisent et où une latence entre les répliques peut être observée, à l'aide des processus de Hawkes?

Défi 5 : Adapter la méthode paramétrique de manière à permettre **tout type de noyaux** et à estimer les paramètres d'un processus de Hawkes non séparable spatio-temporellement, offrant ainsi **flexibilité et précision** dans la modélisation des **dépendances complexes dans des ensembles de données réels**.

Méthode paramétrique flexible et efficace. Nous développons une méthode paramétrique rapide qui permet l'utilisation de tout type de noyaux et de noyaux non séparables spatio-temporellement. La méthode dérivée s'inspire des travaux de Staerman et al. (2023) pour les processus de Hawkes temporels, en étendant la méthode pour capturer les interactions spatio-temporelles.

8.3 Contributions

Pour surmonter les **défis** décrits ci-dessus et répondre à nos **questions de recherche**, nos contributions sont les suivantes (voir le tableau 8.1 pour un résumé de ces contributions).

Partie I. La première partie de cette thèse vise à contribuer à la conception et à l'étude de méthodes d'apprentissage statistique appliquées aux données spatiales, en explorant le problème de Krigeage. L'objectif du Krigeage est de prédire les valeurs d'un

champ aléatoire $\mathbf{X} = \{\mathbf{X}_s, s \in \mathcal{S}\}$, $\mathcal{S} \subset \mathbb{R}^2$, en toutes localisations non observées dans $\mathcal{S}$, en se basant sur un nombre fini $d \geq 1$ d'observations $\mathbf{X}(\mathbf{s}_d) := (\mathbf{X}_{s_1}, \ldots, \mathbf{X}_{s_d})$, avec $\mathbf{s}_d = (s_1, \ldots, s_d)$. Pour cet ensemble d'observations, on pose $\Sigma(\mathbf{s}_d) = Var(\mathbf{X}(\mathbf{s}_d))$ la matrice de covariance et $\mathbf{c}_d(s) = (Cov(\mathbf{X}_s, \mathbf{X}_{s_1}), \ldots, Cov(\mathbf{X}_s, \mathbf{X}_{s_d}))$ le vecteur de covariance. Le but est de construire une carte prédictive $f(s) = f_{\Lambda_d}(s, \mathbf{X}(\mathbf{s}_d)) = \lambda_1(s)\mathbf{X}_{s_1} + \ldots + \lambda_d(s)\mathbf{X}_{s_d}$ linéaire en $\mathbf{X}(\mathbf{s}_d)$ qui minimise l'erreur quadratique moyenne intégrée

$$L_{\mathcal{S}}(f_{\Lambda_d}) = \mathbb{E}_{\mathbf{X}}\left[\int_{s\in\mathcal{S}} \left(f_{\Lambda_d}(s, \mathbf{X}(\mathbf{s}_d)) - \mathbf{X}_s\right)^2 ds\right],$$

où $\Lambda_d : s \in \mathcal{S} \mapsto (\lambda_1(s), \ldots, \lambda_d(s))$ est une fonction mesurable à valeurs dans $\mathbb{R}^d$. Lorsque la vraie fonction de covariance $c(\cdot)$ de $\mathbf{X}$ est connue et que la matrice $\Sigma(\mathbf{s}_d)$ est définie positive, le prédicteur de Krigeage $f_{\Lambda_d^*}(s, \mathbf{X}(\mathbf{s}_d)) = \mathbf{X}(\mathbf{s}_d)^\top \Sigma(\mathbf{s}_d)^{-1}\mathbf{c}_d(s)$ atteint une performance optimale. Désignons l'erreur minimale globale par $L_{\mathcal{S}}^* := L_{\mathcal{S}}(f_{\Lambda_d^*})$ et les poids de Krigeage optimaux $\Lambda_d^*(s) = \Sigma(\mathbf{s}_d)^{-1}\mathbf{c}_d(s)$. Cependant, cette optimalité n'est pas toujours réalisable en pratique, car la vraie structure de covariance des données réelles reste inconnue. Ainsi, sur la base d'un ensemble de données d'entraînement $\mathbf{X}'$, défini comme une unique réalisation de $\mathbf{X}$ observée à $n \geq 1$ localisations spatiales $\sigma_1, \ldots, \sigma_n$ formant une grille dyadique régulière, une estimation empirique $\widehat{c}(\cdot)$ de la fonction de covariance peut être obtenue. À partir de $\widehat{c}(\cdot)$, on peut calculer les estimateurs empiriques $\widehat{\Sigma}(\mathbf{s}_d)$ et $\widehat{\mathbf{c}}_d(s)$, de $\Sigma(\mathbf{s}_d)$ et $\mathbf{c}_d(s)$ respectivement. En remplaçant $\Sigma(\mathbf{s}_d)^{-1}$ et $\mathbf{c}_d(s)$ par leurs estimateurs, un contrepartie empirique naturelle de Λ_d^* est construite par la méthode du *plug-in* et une version empirique du prédicteur de Krigeage est

$$f_{\widehat{\Lambda}_d}(s, \mathbf{X}(\mathbf{s}_d)) = \mathbf{X}(\mathbf{s}_d)^\top \widehat{\Sigma}(\mathbf{s}_d)^{-1}\widehat{\mathbf{c}}_d(s).$$

Considérant le Krigeage dual comme un problème de régression ridge à noyau. Nous montrons que le prédicteur optimal f_{Λ_d} a la même forme qu'un régresseur ridge à noyau (Kernel Ridge Regression), où on remplace la matrice de Gram pour la régression par la vraie matrice de covariance de $\mathbf{X}(\mathbf{s}_d)$ (voir Chapitre 4).

Notre objectif est maintenant de fournir des garanties théoriques pour le prédicteur de Krigeage empirique sous la forme de bornes non asymptotiques pour l'excès de risque global $L_{\mathcal{S}}(f_{\widehat{\Lambda}_d}) - L_{\mathcal{S}}^*$. Comme le prédicteur empirique $f_{\widehat{\Lambda}_d}$ dépend de l'estimation de la fonction de covariance $\widehat{c}(\cdot)$, notre premier objectif est d'évaluer la précision de cette estimation.

Bornes non asymptotiques pour l'estimation de la fonction de covariance. En géostatistique, lorsque le champ aléatoire est stationnaire (d'ordre deux), on utilise le semi-variogramme $\gamma(\cdot)$ pour caractériser la structure de dépendance spatiale des observations. La relation entre la fonction de covariance isotrope et le semi-variogramme est donnée par l'équation suivante : $\gamma(h) = c(0) - c(h)$. Nous étendons cette relation à leurs estimateurs basés sur les observations $\mathbf{X}'_{\sigma_1}, \cdots, \mathbf{X}'_{\sigma_n}$: $\widehat{\gamma}(h) = \widehat{c}_h(0) - \widehat{c}(h)$. Dans le Chapitre 3, sous l'hypothèse que $\mathbf{X}$ est un champ aléatoire gaussien d'ordre deux stationnaire avec une fonction de covariance isotrope, nous identifions d'abord la distribution des estimateurs non paramétriques $\widehat{\gamma}(h)$ et $\widehat{c}_h(0)$, qui est donnée par une somme pondérée de variables aléatoires χ^2. Sous des conditions appropriées, nous établissons des bornes de queues de type Poisson pour ces estimateurs, en nous basant sur de nouveaux résultats de concentration pour les variables Gamma et χ^2 (Bercu et al., 2015; Wang and Ma, 2020). Ces bornes sont établies uniquement pour les distances observées h de la grille régulière échantillonnée $\sigma_1, \cdots, \sigma_n$. Grâce à la relation

entre les estimateurs, des bornes correspondantes peuvent également être dérivées pour l'estimation de la fonction de covariance. Enfin, en supposant que $c(\cdot)$ est de classe $\mathcal{C}^1$ avec un gradient borné par une constante $Q < +\infty$, nous étendons les bornes précédentes à toutes les distances du domaine spatial supposé borné. Ces contributions nous permettent de répondre à notre **Question de Recherche 2**.

Ensuite, grâce aux contributions ci-dessus, nous analysons l'impact de la précision de l'estimation de la fonction de covariance sur la performance du prédicteur empirique de Krigeage.

Garanties statistiques pour la méthode de Krigeage. Nous fournissons d'abord des bornes non asymptotiques pour la précision de l'estimation de la matrice de covariance et de la matrice de précision ($\widehat{\Sigma}(\mathbf{s}_d)$ et $\widehat{\Sigma}(\mathbf{s}_d)^{-1}$ respectivement) dans le Chapitre 4. Ces bornes découlent des résultats précédents pour l'estimation de la fonction de covariance $\widehat{c}(\cdot)$, sous une hypothèse supplémentaire sur les valeurs propres de $\Sigma(\mathbf{s}_d)$. Ensuite, dans le Chapitre 4, nous évaluons la capacité de généralisation du prédicteur empirique de Krigeage sur tous les sites non observés du domaine spatial, en dérivant des bornes non asymptotiques pour l'excès de risque global de la méthode de Krigeage. Le résultat final est fourni par le Théorème 4.8, où des bornes sur la vitesse d'apprentissage d'ordre $O_{\mathbb{P}}(1/\sqrt{n})$ sont établies pour le prédicteur empirique, sous des conditions appropriées. Notre résultat principal est le suivant :

Pour tout $\delta \in (0,1)$, nous avons avec une probabilité d'au moins $1-\delta$:

$$L_{\mathcal{S}}(f_{\widehat{\Lambda}_d}) - L^*_{\mathcal{S}} \leq C_6\, d^2 \sqrt{\log(4n/\delta)/n} + C_6'\, d^2\, Q/(\sqrt{n}-1),$$

dès que $n \geq C_6'' \log(4n/\delta)$, où C_6, C_6' et C_6'' sont des constantes positives.

Ce résultat nous permet de répondre à notre **Question de Recherche 3**.

Expériences numériques. Les résultats théoriques, ainsi que le rôle joué par les conditions techniques requises pour les établir, sont illustrés dans le Chapitre 4 par diverses expériences numériques sur des données simulées. Nous réalisons les expériences pour différents modèles de covariance, dont certains satisfont toutes les conditions requises pour nos résultats tandis que d'autres ne les respectent pas. Nous répétons les expériences pour différentes tailles de grilles d'observations. Nos expériences numériques corroborent pleinement nos résultats théoriques pour tous les modèles de covariance satisfaisant les hypothèses. En outre, nous explorons d'autres extensions possibles de notre travail en fournissant des expériences supplémentaires abordant les cas suivants : (1) les d points d'observation sont extraits de configurations différentes de la procédure uniforme aléatoire, (2) le cas des modèles de covariance anisotropes, (3) le cas des grilles irrégulières pour l'échantillon d'apprentissage. Les résultats de nos expériences montrent que la méthode de prédiction est robuste en cas de légères violations des hypothèses ci-dessus.

Applications aux données réelles. Nous illustrons la force et les avantages du prédicteur de Krigeage empirique par des expériences numériques sur des données météorologiques réelles dans le Chapitre 4. Le jeu de données DRIAS fournit la température moyenne quotidienne en France, observée sur une grille régulière (voir Figure 8.1). Un prédicteur de Krigeage paramétrique, construit à l'aide d'une fonction de covariance paramétrique, ainsi que le prédicteur de Krigeage non paramétrique, sont appliqués à ces données. Nos résultats corroborent les garanties théoriques établies et démontrent qu'une application directe de la méthode de prédiction de Krigeage em-

pirique peut donner des performances solides et une meilleure flexibilité par rapport à une méthode paramétrique.

Code. Nos expériences sont entièrement reproductibles et peuvent être répliquées avec les codes disponibles sur GitHub[1].

Partie II. La deuxième partie de cette thèse vise à concevoir une nouvelle méthode d'inférence pour les processus de Hawkes spatio-temporels multivariés. La caractéristique principale d'un processus de Hawkes est qu'il prend en compte la nature auto-excitante du phénomène sous-jacent. Soit $T \in \mathbb{R}_+$ un temps d'arrêt, et considérons $[0,T]$ la période d'observation résultante. De plus, soit $\mathcal{S} = \mathcal{X} \times \mathcal{Y} \subset \mathbb{R}^2$ un ensemble compact au sein du domaine spatial qui contient les emplacements des événements observés jusqu'au temps T. Soit $D \in \mathbb{N}^*$ la dimension du processus de Hawkes spatio-temporel multivarié. Une réalisation consiste en D ensembles d'événements distincts : $\mathcal{H}_T^i = \left\{ u_n^i = (x_n^i, y_n^i, t_n^i), (x_n^i, y_n^i) \in \mathcal{S}, t_n^i \in [0,T] \right\}, \forall i \in \{1,\dots,D\}$ se produisant dans un espace-temps continu, avec un temps associé t_n^i et une localisation (x_n^i, y_n^i). Le comportement du processus est entièrement décrit par ses D fonctions d'intensité, qui dépendent des temps et des emplacements des événements passés. La fonction d'intensité conditionnelle pour le i-ème processus est :

$$\lambda_i(x,y,t|\mathcal{H}_t) = \boldsymbol{\mu}_i + \sum_{j=1}^{D} \sum_{u_n^j \in \mathcal{H}_t^j} \boldsymbol{\alpha}_{ij}\, g_{ij}(x - x_n^j, y - y_n^j, t - t_n^j),$$

où $\boldsymbol{\mu}_i > 0$ est le paramètre de base, $\boldsymbol{\alpha}_{ij} > 0$ est le paramètre de mise à l'échelle de l'excitation, et $g_{ij} : \mathcal{S} \times [0,T] \mapsto \mathbb{R}+$ est le noyau spatio-temporel avec des paramètres $\boldsymbol{\eta}_{ij}$. Notez que nous utilisons la même notation que pour les poids du prédicteur de Krigeage dans la Partie I pour respecter les notations habituelles de ces domaines.

L'objectif de notre méthode est de pouvoir inférer les paramètres pour : (1) tous les noyaux paramétriques, y compris (2) les noyaux non séparables en espace-temps. Notre travail s'inspire de l'approche proposée par Staerman et al. (2023) pour les processus de Hawkes temporels, dont la procédure repose sur trois idées clés que nous étendons aux données spatio-temporelles. Le premier concept est que le domaine spatio-temporel des observations est discrétisé en une grille régulière tridimensionnelle et les observations sont projetées dessus. Ensuite, nous supposons que les fonctions de noyau sont à support fini. En combinant ces deux premières idées, le noyau dans l'intensité conditionnelle λ_i peut être remplacé par une version discrétisée, remplaçant ainsi la somme sur les événements passés par une somme sur un nombre fini d'éléments de la grille. Ensuite, nous nous concentrons sur la perte des moindres carrés et en dérivons une version discrétisée

$$\mathcal{L}_{\mathcal{G}}(\boldsymbol{\theta}, \widetilde{\mathcal{H}}_T) = \sum_{i=1}^{D} \left(\Delta_{\mathcal{X}} \Delta_{\mathcal{Y}} \Delta_T \sum_{v_x=0}^{G_{\mathcal{X}}} \sum_{v_y=0}^{G_{\mathcal{Y}}} \sum_{v_t=0}^{G_T} \left(\tilde{\lambda}_i[v_x, v_y, v_t] \right)^2 - 2 \sum_{\tilde{u}_n^i \in \widetilde{\mathcal{H}}_T^i} \tilde{\lambda}_i \left[\frac{\tilde{x}_n^i}{\Delta_{\mathcal{X}}}, \frac{\tilde{y}_n^i}{\Delta_{\mathcal{Y}}}, \frac{\tilde{t}_n^i}{\Delta_T} \right] \right),$$

où $(\Delta_{\mathcal{X}}, \Delta_{\mathcal{Y}}, \Delta_T)$ sont les pas de discrétisation de la grille tridimensionnelle, $(G_{\mathcal{X}}, G_{\mathcal{Y}}, G_T)$ sont les tailles des grilles discrétisées, et $\widetilde{\mathcal{H}}_T^i$ est l'ensemble des évènements projetés de $\mathcal{H}_T^i$. Cela nous amène à la troisième composante clé de notre approche, qui est

[1] https://github.com/EmiliaSiv/Simple-Kriging-Code

l'identification de termes de pré-calculs (qui ne dépendent pas du paramètre $\theta = \{\mu_i, \alpha_{ij}, \eta_{ij}\}_{i,j}$) dans la fonction de perte discrétisée. Grâce à ces termes de pré-calculs, notre approche est efficace et permet une inférence rapide.

En combinant ces trois idées clés, nous concevons une méthode répondant à nos **Questions de Recherche 4 et 5**, définies dans la Section 8.2. Nous fournissons également des garanties théoriques sur le biais induit par la discrétisation, montrant son faible impact sur la précision de l'estimation des paramètres.

Méthode efficace et flexible pour l'inférence paramétrique dans les modèles de Hawkes spatio-temporels. Dans le Chapitre 6, nous développons une méthode rapide pour inférer les paramètres des noyaux dans les modèles de Hawkes spatio-temporels. La méthode que nous concevons permet d'incorporer n'importe quel noyau paramétrique pour la fonction de déclenchement, s'étendant au-delà des formes traditionnelles gaussiennes et exponentielles. De plus, pour mieux s'adapter aux données réelles, l'approche prend également en compte les interactions spatio-temporelles, en s'étendant au cas de noyaux non séparables en espace-temps. Ces deux innovations améliorent la précision et la flexibilité dans la modélisation des dépendances complexes dans les ensembles de données du monde réel.

Expériences numériques. Nous montrons les avantages de notre approche par différentes expériences sur des données simulées dans le Chapitre 6. Tout d'abord, nous étudions l'impact du pas de discrétisation sur la précision de la méthode en répétant les expériences pour différentes valeurs de $(\Delta_x, \Delta_y, \Delta_T)$. Nos résultats montrent que l'erreur d'estimation tend vers zéro à mesure que les pas diminuent simultanément, soutenant nos résultats théoriques sur la discrétisation. Ensuite, nos expériences, réalisées pour différents temps de fin T et limites spatiales $\mathcal{S}$, prouvent l'exactitude de la méthode. Le temps de calcul par rapport au pas de discrétisation et par rapport à $(\mathcal{S}, T)$ est également étudié, prouvant l'efficacité de notre méthode. Enfin, toutes les expériences sont réalisées avec des noyaux spatiaux et temporels variés, montrant ainsi la flexibilité de notre approche.

Applications aux données réelles. Les avantages de notre méthode d'inférence sont également prouvés en l'appliquant à deux ensembles de données du monde réel dans le Chapitre 6 : (1) des données réelles sur les tremblements de terre, basées sur le jeu de données *Northern California Earthquake Data Center* [2] (NCEDC; nce, 2014) et (2) des données de cambriolages provenant du *Chicago Crime Dataset* [3]. En effet, les deux ensembles de données réelles enfreignent les deux conditions supposées par la majorité des approches précédentes. En général, un tremblement de terre ne déclenche pas immédiatement des répliques (donc le noyau temporel exponentiel n'est pas adapté) et les effets de déclenchement peuvent varier selon les différentes directions spatiales (ce qui implique que le noyau spatial gaussien ne reflète pas le processus sous-jacent). De plus, les événements de cambriolage présentent des dépendances spatio-temporelles, en raison du motif de "victimisation proche répétée" (Johnson, 2008) : les cambrioleurs ciblent souvent le même quartier à plusieurs reprises dans un court laps de temps.

Code. L'implémentation de notre approche est disponible sur GitHub[4].

[2] https://ncedc.org/

[3] https://data.cityofchicago.org/Public-Safety/Crimes-2001-to-Present/ijzp-q8t2

[4] https://github.com/EmiliaSiv/Flexible-Parametric-Inference-for-Space-Time-Hawkes-Processes

Table 8.1: Résumé des contributions.

Chapitres	Contributions
Chapitre 3	• Garanties théoriques pour l'estimation non paramétrique de la fonction de covariance, sous des conditions appropriées. • Expériences numériques sur des données simulées qui valident l'utilisation des hypothèses.
Chapitre 4	• Garanties statistiques pour la méthode de Krigeage. • Expériences numériques sur des données simulées qui vérifient nos résultats théoriques. • Applications sur des données météorologiques réelles.
Chapitre 6	• Nouvelle méthode efficace et flexible pour l'inférence de processus de Hawkes spatio-temporels. • Expériences numériques sur des données simulées qui prouvent la performance et la flexibilité de notre méthode. • Applications sur des données réelles de tremblements de terre et de cambriolages.

8.4 Plan de la thèse

La Partie I se concentre sur une approche d'apprentissage statistique du Krigeage simple et développe un nouveau cadre théorique offrant des garanties non asymptotiques pour les règles empiriques de Krigeage simple. L'objectif principal de cette première partie est de surmonter les défis posés par les caractéristiques des données spatiales, principalement la présence d'une forte structure de dépendance et l'observation d'une unique réalisation du phénomène étudié.

- Le Chapitre 2 fournit les bases nécessaires pour étudier les données spatiales à l'aide d'outils géostatistiques et présente les résultats fondamentaux de l'apprentissage statistique, en mettant l'accent sur le principe de minimisation du risque empirique.

- Dans le Chapitre 3, nous proposons des bornes non asymptotiques pour l'estimation non paramétrique de la fonction de covariance d'un champ aléatoire basée sur un échantillon fini d'observations et une unique réalisation du phénomène.

- Les résultats finaux sont présentés dans le Chapitre 4, où des bornes sur les taux d'apprentissage sont obtenues pour le prédicteur empirique de Krigeage simple.

La Partie II est consacrée aux processus de Hawkes et à l'étude des données spatio-temporelles issues de jeux de données réels présentant de l'hétérogénéité. L'objectif de cette partie est de développer une méthode efficace et flexible pour l'inférence paramétrique dans les modèles spatio-temporels de Hawkes.

- Le Chapitre 5 expose les bases sur les processus ponctuels utiles pour notre approche et met en lumière les avantages de développer une telle méthode en investiguant des exemples réels où l'hétérogénéité et les interactions spatio-temporelles sont observés.
- Le Chapitre 6 introduit une nouvelle approche, flexible et efficace, pour inférer tout noyau paramétrique dans le contexte des processus de Hawkes spatio-temporels.

Le Chapitre 7 fournit une conclusion globale et propose une vue d'ensemble des perspectives et des futures pistes de recherche découlant des travaux développés dans cette thèse.

8.5 Publications

Les contributions présentées ici ont donné lieu à la publication et au préprint suivants:

- (Siviero et al., 2024a) Emilia Siviero, Emilie Chautru, Stephan Clémençon. A Statistical Learning View of Simple Kriging. In *TEST, vol. 33, no 1*, pages 271-296, 2024. Reproduit dans les Chapitres 3 et 4.
- (Siviero et al., 2024b) Emilia Siviero, Guillaume Staerman, Stephan Clémençon, Thomas Moreau. Flexible Parametric Inference for Space-time Hawkes Processes. *arXiv preprint arXiv:2406.06849*, 2024. Reproduit dans le Chapitre 6.

Les publications ont été présentées dans les conférences et séminaires suivants :

- Juillet 2022 : Présentation poster de l'article 'A Statistical Learning View of Simple Kriging', lors de la conférence française sur l'apprentissage automatique (CAp 2022), Vannes (France).
- Août 2022 : Présentation orale de l'article 'A Statistical Learning View of Simple Kriging', lors de la conférence internationale sur les statistiques computationnelles (COMPSTAT 2022), Bologne (Italie).
- Mars 2023 : Présentation orale de l'article 'A Statistical Learning View of Simple Kriging', au séminaire de l'équipe MIND, Inria, Palaiseau (France).
- Août 2024 : Présentation orale de l'article 'Flexible Parametric Inference for Space-time Hawkes Processes', lors de la conférence internationale sur les statistiques computationnelles (COMPSTAT 2024), Giessen (Allemagne).

Appendix

A Appendices for Chapter 3

The Appendix is organized as follows: first, we present the auxiliary results used in the technical proofs, then the proofs of the main results of Chapter 3 are detailed at length.

A.1 Auxiliary Results

Here and throughout, given a Hermitian matrix A of size $n \times n$, denote $\xi_n(A) \leq \cdots \leq \xi_1(A)$ its eigenvalues (arranged in decreasing order) and $Rank(A)$ its rank.

Auxiliary Result for the Proof of Proposition 3.7

The following lemma is used in the proof of Proposition 3.7 and allows us to generalize the proof for the distributions of both the semi-variogram and the variance estimators.

Lemma .6. *Let $X \sim \mathcal{N}(0, \Sigma)$ be a centered Gaussian random field with positive definite covariance function and R a symmetric and positive semi-definite matrix of size $n \times n$, such that $Rank(R) \leq r$ (where r is a strictly positive integer). Then, we have:*

$$^tXRX \sim \sum_{i=1}^{r} \xi_i(R\Sigma)\chi_i^2,$$

where the χ_i^2's are independent χ^2 random variables with one degree of freedom and the $\xi_i(R\Sigma)$'s are the r (strictly positive) eigenvalues of $R\Sigma$.

Proof. Thanks to the assumptions, the covariance matrix Σ is symmetric and positive definite. Thus, using a well-known result of matrix algebra (see *e.g.* Chapter 21 in Harville, 1998), define the square root of Σ, a symmetric and positive definite matrix by $\Sigma^{1/2}$ such that $\Sigma = \Sigma^{1/2}\Sigma^{1/2}$. The square root matrix is invertible and its inverse $\Sigma^{-1/2}$ is symmetric and positive definite. Let $Y = \Sigma^{-1/2}X \sim \mathcal{N}(0, \mathbf{I}_n)$, where $\mathbf{I}_n$ is the $n \times n$ identity matrix. Then

$$\begin{aligned} ^tXRX &=^t X^t(\Sigma^{1/2}\Sigma^{-1/2})R(\Sigma^{1/2}\Sigma^{-1/2})X \\ &=^t (\Sigma^{-1/2}X)^t\Sigma^{1/2}R\Sigma^{1/2}(\Sigma^{-1/2}X) =^t YTY, \end{aligned}$$

where $T =^t \Sigma^{1/2}R\Sigma^{1/2}$. Since R is symmetric and positive semi-definite, T is also symmetric and positive semi-definite. Furthermore,

$$R\Sigma = \Sigma^{-1/2}\Sigma^{1/2}R\Sigma^{1/2}\Sigma^{1/2} = \Sigma^{-1/2}T\Sigma^{1/2},$$

which implies that the matrices T and $R\Sigma$ are similar and have the same eigenvalues (Harville, 1998, Chapter 21). Thanks to the spectral decomposition, there exists an orthogonal matrix P and a diagonal matrix $D = Diag((t_i)_{i\in\{1,\cdots,n\}})$ of the eigenvalues of T, such that $T =^t PDP$. The eigenvalues t_i's are positive as T is positive semi-definite. Recall that the rank of the matrix R is upper bounded by a positive integer r and this implies that the rank of $R\Sigma$ is upper bounded by r too. Thus, denote $\xi_i(R\Sigma)$ the r positive eigenvalues of $R\Sigma$. Furthermore

$$^tXRX =^t Y^tPDPY =^t ZDZ = \sum_{i=1}^{n}\sum_{j=1}^{n} Z_iZ_jD_{i,j} = \sum_{i=1}^{r} Z_i^2\xi_i(R\Sigma),$$

where $Z_i \sim \mathcal{N}(0, \mathbf{I}_n)$ are independent Gaussian random variables. Finally, notice that Z_i^2 are r χ^2 random variables with one degree of freedom, which concludes the proof. ■

Bounds on Largest Eigenvalues

Here, we present several results useful for Proposition 3.9. These results are used for bounding the eigenvalues that appear in the weighted sums of χ^2 random variables in the distributions of both the semi-variogram and the variance estimators.

Eigenvalues of the Product of Positive Semi-definite Hermitian Matrices. We first present inequalities for the eigenvalues of the product of positive semi-definite Hermitian matrices (the proof can be found in Wang and Zhang, 1992; Xi and Zhang, 2019).

Proposition .7. *Let A and B two positive semi-definite $n \times n$ Hermitian matrices. Denote $\xi_n(A) \leq \cdots \leq \xi_1(A)$ and $\xi_n(B) \leq \cdots \leq \xi_1(B)$ the eigenvalues of A and B respectively. Let $k > 1$. Then, for $1 \leq i_1 < \cdots < i_k \leq n$,*

$$\sum_{t=1}^{k} \xi_{i_t}(A)\xi_{n-t+1}(B) \leq \sum_{t=1}^{k} \xi_{i_t}(AB) \leq \sum_{t=1}^{k} \xi_{i_t}(A)\xi_t(B). \tag{2}$$

Bounds on the Largest Eigenvalue of Laplacian Matrices. We give a bound on the largest eigenvalue $\xi_1(L)$ of the Laplacian matrix L of a graph in terms of the maximum degree of its vertices.

Proposition .8. *Let $G = (V, E)$ a graph with maximum degree $d_{\max} = \max_{v \in V} deg(v)$ and L the Laplacian matrix of G. Then $\xi_1(L) \leq 2d_{\max}$.*

The proof essentially relies on the following propositions.

Proposition .9. ***(Spielman, 2012, Lemma 3.4.1)*** *Let $G = (V, E)$ a graph with maximum degree $d_{\max} = \max_{v \in V} deg(v)$. Let D the degree matrix of G (defined as the diagonal matrix with entries $D_{ii} = deg(v_i)$, $\forall i \in \{1, \ldots, |V|\}$) and A the adjacency matrix of G (with entries $A_{ij} = \mathbb{I}\{(v_i, v_j) \in E\}$, $\forall i, j \in \{1, \ldots, |V|\}$). Then $\xi_1(D) = d_{\max}$ and $\xi_1(A) \leq d_{\max}$.*

The next proposition is a well-known result, often called Weyl's inequality (see *e.g.* Horn and Johnson, 2012, Theorem 4.3.1, for a proof of the result).

Proposition .10. *Let A and B be two Hermitian matrices. Then*

$$\xi_1(A - B) \leq \xi_1(A) + \xi_1(B).$$

Combining these results with the definition of the Laplacian matrix as $L = D - A$, where D is the degree matrix and A the adjacency matrix of a graph, we have the wanted result in Proposition .8.

Bounds on the Largest Eigenvalue of the Covariance Matrix of a Stationary Random Field. Now, we present a result on the bounded eigenvalues of the covariance matrix for a stationary random field. This result derives from the application of Bochner's Theorem (Stein, 1999, Chapter 2), combined with the assumed bounds on the spectral density.

Lemma .11. *Let $(X_s)_{s\in\mathbb{R}^2}$ be a stationary (in the second-order sense) process with spectral density Φ and covariance matrix Σ. Suppose that Assumption 3.8 is fulfilled. Thus, the eigenvalues $\xi(\Sigma)$ of the covariance matrix are bounded as follows:*

$$\exists c > 0, c' > 0,\ cm \leq \xi(\Sigma) \leq c'M,$$

where m and M are given in Assumption 3.8.

Upper Bound on the Variances of the Semi-Variogram and Variance Estimators

We present a well-known result from Cressie (1993) for the variance of the semi-variogram estimator.

Proposition .12. ***(Cressie, 1993, Section 2.4)*** *Variances of the semi-variogram estimator $\widehat{\gamma}(h)$ for a fixed h are $O(1/n)$.*

Proof. Notice that, under the Gaussian and the intrinsic assumptions, we have $Var((X_{s+h}-X_s)^2) = 2(2\gamma(h))^2$. Then

$$\begin{aligned} Var(\widehat{\gamma}(h)) &= \left(\frac{1}{2n_h}\right)^2 \sum_{(s_i,s_j)\in N(h)} Var((X_{s_i}-X_{s_j})^2) \\ &= \left(\frac{1}{2n_h}\right)^2 \sum_{(s_i,s_j)\in N(h)} 2(2\gamma(h))^2 = \frac{2\gamma(h)^2}{n_h}, \end{aligned}$$

which gives the desired result. ■

Furthermore, it's easy to see that the variance of the covariance estimator $\widehat{c}(h)$ for a fixed h is $O(1/n)$. This implies, thanks to the relationship between the semi-variogram estimator and the covariance estimator in Equation (3.5) in Chapter 3, that the variance of $\widehat{c}_h(0)$ is also $O(1/n)$. Thus, we obtain

$$\exists c > 0,\ Var\left(\widehat{\gamma}(h)\right) \leq \frac{c}{n} \tag{3}$$

and

$$\exists c' > 0,\ Var\left(\widehat{c}_h(0)\right) \leq \frac{c'}{n}. \tag{4}$$

Gamma Random Variables

To avoid any ambiguity, we give the definition of a Gamma random variable and a proposition on the relationship between Gamma and χ^2 random variables.

Definition .13. *The density function of $Z \sim \Gamma(\alpha,\beta)$ a Gamma random variable with shape parameter $\alpha \in \mathbb{R}_+$ and rate parameter $\beta \in \mathbb{R}_+$ is*

$$f_Z(z) = \frac{\beta^\alpha}{\Gamma(\alpha)} z^{\alpha-1} e^{-z\beta},\ \forall z > 0,$$

where Γ is the Gamma function. The mean of a Gamma random variable is: $\mathbb{E}[Z] = \frac{\alpha}{\beta}$.

Proposition .14. *If $Z \sim \chi^2_k$ and $c > 0$ then $cZ \sim \Gamma\left(\frac{k}{2}, \frac{1}{2c}\right)$.*

Extension of Tail Bound Inequalities for the Semi-Variogram and Variance Estimators

As a first preliminary result, we give an upper bound on the total number of distinct observable distances $h \in \mathcal{H}_n$ on the regular grid of size n. The idea is the following. Let n_x be the number of columns/rows, such that $n = n_x \times n_x$. Then, we have n_x^2 possible combinations between all pairs of points location, at which we may withdraw n_x values (the locations that are on the diagonal of the grid). Since the process is assumed to be isotropic, we may divide this value by 2. Finally, we add n_x for the diagonal and have the following result: $|\mathcal{H}_n| = \frac{n_x(n_x+1)}{2}$. Then, we obtain the following upper bound, since $n = n_x^2$:

$$|\mathcal{H}_n| \leq n. \tag{5}$$

Then, we present a corollary to Proposition 3.9, that extends the results on tail bound inequalities for the semi-variogram estimator and the variance estimator.

Corollary .15. *Suppose that Assumptions 3.1–3.8 are fulfilled. Let $k > 0$,*

$$\mathbb{P}\left(\max_{h\in\mathcal{H}_n}\left|\widehat{\gamma}(h)-\gamma(h)\right| \geq k\right) \leq 2ne^{-C_1' nk^2}, \text{ whenever } k \leq \frac{C_1}{C_1'},$$

and

$$\mathbb{P}\left(\max_{h\in\mathcal{H}_n}\left|\widehat{c_h}(0)-c(0)\right| \geq k\right) \leq 2ne^{-C_2' nk^2}, \text{ whenever } k \leq \frac{C_2}{C_2'},$$

where C_i and C_i', $i \in \{1, 2\}$, are positive constants depending on j_1, m and M solely (given in Proposition 3.9).

Proof. We give the proof for the semi-variogram estimator. The proof for the variance estimator follows the same steps, replacing the constants C_1 and C_1' by C_2 and C_2'. Notice that

$$\forall k > 0,\ \mathbb{P}\left(\max_{h\in\mathcal{H}_n}\left|\widehat{\gamma}(h)-\gamma(h)\right| \geq k\right) = \mathbb{P}\left(\bigcup_{h\in\mathcal{H}_n}\left\{\left|\widehat{\gamma}(h)-\gamma(h)\right| \geq k\right\}\right).$$

Thanks to the Union Bound (or Boole's Inequality) and then applying the result in Proposition 3.9, we have

$$\begin{aligned}
\mathbb{P}\left(\bigcup_{h\in\mathcal{H}_n}\left\{\left|\widehat{\gamma}(h)-\gamma(h)\right| \geq k\right\}\right) &\leq \sum_{h\in\mathcal{H}_n}\mathbb{P}\left(\left|\widehat{\gamma}(h)-\gamma(h)\right| \geq k\right) \\
&\leq \sum_{h\in\mathcal{H}_n}\left(e^{-C_1 nk} + e^{-C_1' nk^2}\right) \\
&\leq |\mathcal{H}_n| 2\max\left\{e^{-C_1 nk}, e^{-C_1' nk^2}\right\}.
\end{aligned}$$

Then, if we take $k \leq \frac{C_1}{C_1'}$, the maximum is obtained for $e^{-C_1' nk^2}$ and, combining this with the result on the cardinality of $\mathcal{H}_n$ in (5), one gets

$$|\mathcal{H}_n| 2\max\left\{e^{-C_1 nk}, e^{-C_1' nk^2}\right\} \leq 2ne^{-C_1' nk^2},$$

which concludes the proof. ■

A.2 Technical Proofs

Proof of Lemma 3.6

For any strictly positive $h \in \mathcal{H}_n$, $\exists (d,q) \in \mathbb{N}^* \times \mathbb{N}^*$ s.t. $h = \left(\sqrt{d^2+q^2}\right)2^{-J}$. Since the random field X is isotropic (Assumption 3.2), we have the following bound $n_h \geq 4(n_x - d)(n_x - q)$, where $n = n_x^2$ (n_x represents the number of columns/rows of the square grid). Then, let $h' = h2^J = \sqrt{d^2+q^2} > 0$. Since $d \leq h'$ and $q \leq h'$, we have $n_h \geq 4(n_x - h')^2 \geq 4(n_x^2 - 2n_x h') = 4(n - 2n_x h')$. Furthermore, under Assumption 3.3, we are only interested in distances s.t. $h < \sqrt{2} - 2^{-j_1}$. This implies that, for n large enough (condition given by $n > (\sqrt{2} - 2^{-j_1})^2$), we finally obtain:

$$\forall h \in \mathcal{H}_n, \quad n_h > \nu n, \tag{6}$$

where ν is a positive constant depending on j_1 only.

Proof of Proposition 3.7

The proof of Proposition 3.7 essentially relies on Lemma .6, simultaneously applied to the estimators $\widehat{\gamma}(h)$ and $\widehat{c_h}(0)$. Refer to Appendix A.1 for its presentation and proof. We first study the semi-variogram estimator $\widehat{\gamma}(h)$ and then the variance estimator $\widehat{c_h}(0)$. The goal is to define a matrix $L(n,h)$ such that

$$\widehat{\gamma}(h) = {}^t\mathbf{X}'(\sigma_n)\frac{1}{n_h}L(n,h)\mathbf{X}'(\sigma_n).$$

Notice that: $\forall h > 0$,

$$\begin{aligned}
\widehat{\gamma}(h) &= \frac{1}{2n_h}\sum_{i=1}^{n}\sum_{j=1}^{n}\left(\mathbf{X}'^2_{\sigma_i} + \mathbf{X}'^2_{\sigma_j} - 2\mathbf{X}'_{\sigma_i}\mathbf{X}'_{\sigma_j}\right)\mathbb{I}\{(\sigma_i,\sigma_j) \in N(h)\} \\
&= \frac{1}{n_h}\sum_{i=1}^{n}\mathbf{X}'^2_{\sigma_i}n_h(i) - \frac{1}{n_h}\sum_{i=1}^{n}\sum_{j=1}^{n}\mathbf{X}'_{\sigma_i}\mathbf{X}'_{\sigma_j}\mathbb{I}\{(\sigma_i,\sigma_j) \in N(h)\} \\
&= \frac{1}{n_h}\sum_{i=1}^{n}\mathbf{X}'^2_{\sigma_i}n_h(i) - \frac{1}{n_h}\sum_{i=1}^{n}\mathbf{X}'^2_{\sigma_i}\mathbb{I}\{(\sigma_i,\sigma_i) \in N(h)\} \\
&\quad - \frac{1}{n_h}\sum_{i=1}^{n}\sum_{j=1,j\neq i}^{n}\mathbf{X}'_{\sigma_i}\mathbf{X}'_{\sigma_j}\mathbb{I}\{(\sigma_i,\sigma_j) \in N(h)\} \\
&= \frac{1}{n_h}\left(\sum_{i=1}^{n}\mathbf{X}'^2_{\sigma_i}n_h(i) - \sum_{i=1}^{n}\sum_{j=1,j\neq i}^{n}\mathbf{X}'_{\sigma_i}\mathbf{X}'_{\sigma_j}\mathbb{I}\{(\sigma_i,\sigma_j) \in N(h)\}\right),
\end{aligned}$$

where $\mathbb{I}\{(\sigma_i,\sigma_i) \in N(h)\} = 0$ (since $h > 0$). Then, let $L(n,h)$ the matrix with entries $L_{i,j}(n,h) = -\mathbb{I}\{(\sigma_i,\sigma_j) \in N(h)\}$ if $i \neq j$ and $L_{i,i}(n,h) = n_h(i)$.

Remark .16. *For a fixed $h \in \mathcal{H}_n$, define $G_h = (V_h, E_h)$ the graph described by the regular grid, where the set of vertices V_h is the set of the n observations' locations and E_h is the set of the n_h edges that are defined by the pairs of locations that are at distance h. Then, $L(n,h)$ is the Laplacian matrix of G_h, equal to $D(n,h) - A(n,h)$ where $D(n,h)$ is the diagonal matrix of the degrees of the vertices of the graph and $A(n,h)$ is the adjacency matrix. Thanks to the Gershgorin Circle Theorem (Shi, 2007), the Laplacian matrix is positive semi-definite, which implies that all its eigenvalues are nonnegative.*

Thanks to the remark above and the isotropy assumption (Assumption 3.2), the matrix $L(n,h)$ is symmetric and positive semi-definite. Based on Cressie (1993), it is possible to rewrite the matrix as $L(n,h) = \frac{1}{2}{}^tQ(n,h)Q(n,h)$, where $Q(n,h) \in \mathbb{R}^{n_h \times n}$ is a matrix whose entries are only $-1, 0$ and 1. The idea is to let $(u_l)_{l \leq n_h}$ the elements of $N(h)$, such that $\forall l \in [\![1, n_h]\!]$, $\exists (i,j) \in [\![1, n_h]\!]^2$, $u_l = (u_l^1, u_l^2) = (\sigma_i, \sigma_j)$, where $\|\sigma_i - \sigma_j\| = \|h\|$. Thus, $u_l^1 = \sigma_i$ is equivalent to the fact that there exists $j \in [\![1, n_h]\!]$ such that $(\sigma_i, \sigma_j) \in N(h)$. Furthermore, $\mathbb{I}\{(\sigma_i, \sigma_j) \in N(h)\} = 1$ is equivalent to the fact that there exists $l \in [\![1, n_h]\!]$ such that $u_l = (\sigma_i, \sigma_j)$. Then, let

$$\forall l \in [\![1, n_h]\!], \forall i \in \{1, \dots, n\}, q_{li} = Q_{,i}(n,h) = \begin{cases} 1\,, \text{ if } u_l^1 = \sigma_i \text{ and } u_l^2 \neq \sigma_i \\ -1\,, \text{ if } u_l^2 = \sigma_i \text{ and } u_l^1 \neq \sigma_i \\ 0\,, \text{otherwise} \end{cases}$$

Then, $Rank(Q(n,h)) \leq n_h$, which implies $Rank(L(n,h)) \leq n_h$. Hence, it is possible to apply Lemma .6 to $\widehat{\gamma}(h)$ with the matrix $L(n,h)$ and the random field $\mathbf{X}$ with positive definite covariance matrix Σ_n

$$\widehat{\gamma}(h) \sim \frac{1}{n_h} \sum_{i=1}^{n_h} \ell_i(h) \chi_i^2,$$

where the χ_i^2's are independent χ^2 random variables with one degree of freedom and the $\ell_i(h)$'s are the n_h (strictly positive) eigenvalues of $L(n,h)\Sigma_n$.

For the variance estimator $\widehat{c_h}(0)$, the proof follows the same idea. Let $D(n,h)$ the diagonal matrix with entries $D_{i,i}(n,h) = n_h(i)$. We notice that $D(n,h)$ is the degree matrix of the graph G_h described by the regular grid for a fixed $h \in \mathcal{H}_n$ (see Remark .16). Then, it's clear that

$$\widehat{c_h}(0) =^t \mathbf{X}(\sigma_n) \frac{1}{n_h} D(n,h) \mathbf{X}(\sigma_n),$$

and that the matrix $D(n,h)$ is symmetric and positive semi-definite. Furthermore, one may see that $Rank(D(n,h)) \leq n_h$. Indeed, the diagonal elements are $n_h(i)$, that is, for a fixed location point σ_i, the number of points that are at distance h from σ_i. The total sum of these $n_h(i)$ over all the grid locations σ_i is equal to n_h. Thus, the extreme case is when all the $n_h(i)$'s are equal to 1, which implies that exactly n_h elements on the diagonal are non zero and in this case the rank of $D(n,h)$ is equal to n_h. It is possible to apply Lemma .6 to $\widehat{c_h}(0)$ with the matrix $D(n,h)$ and the random field X with positive definite covariance matrix Σ_n

$$\widehat{c_h}(0) \sim \frac{1}{n_h} \sum_{i=1}^{n_h} \rho_i(h) \chi_i^2,$$

where the χ_i^2's are independent χ^2 random variables with one degree of freedom and the $\rho_i(h)$'s are the n_h (strictly positive) eigenvalues of $D(n,h)\Sigma_n$.

Proof of Proposition 3.9

Since we are interested only on which variables the constants in the final results depend on, we let, in the proof and in the corresponding preliminary results, c and c' as positive constants that are not always the same, but that depend on variables such

as j_1, m and M. The proofs of the Poisson tail bounds for the deviations for both the semi-variogram and the variance estimators are structured as follows: firstly, thanks to Proposition 3.7, the distributions of both estimators are known and these can be seen as the sum of independent Gamma random variables; secondly, we deduce exponential inequalities for these tail bounds (from Bercu et al., 2015 and Wang and Ma, 2020); then, using the previous preliminary results, we can bound the largest eigenvalues involved in the distributions of the estimators, and finally, using the lower bound on n_h given in Lemma 3.6, we conclude the proof.

In the first part of the proof, we deal with the semi-variogram estimator $\widehat{\gamma}(h)$. Let $t > 0$ and $h \in \mathcal{H}_n$, $\mathbb{P}\Big(\big|\widehat{\gamma}(h) - \gamma(h)\big| \geq t\Big) = \mathbb{P}\Big(\widehat{\gamma}(h) \geq \gamma(h) + t\Big) + \mathbb{P}\Big(\widehat{\gamma}(h) \leq \gamma(h) - t\Big)$. Let $\mu_1 = \mathbb{E}[\widehat{\gamma}(h)] = \gamma(h)$ (since $\widehat{\gamma}(h)$ is unbiased). Recall that, thanks to Proposition 3.7: $\widehat{\gamma}(h) \sim \frac{1}{n_h}\sum_{i=1}^{n_h} \ell_i(h)\chi_i^2$. Since the eigenvalues $\ell_i(h)$, $\forall i \in \{1,\dots,n_h\}$ of $L(n,h)\Sigma_n$ are non negatives, from the relationship between Gamma and χ^2 random variables (see Proposition .14), $\frac{1}{n_h}\ell_i(h)\chi_i^2 \sim \Gamma\Big(\frac{1}{2}, \frac{n_h}{2\ell_i(h)}\Big)$. This implies that $\widehat{\gamma}(h)$ can be seen as the sum of n_h independent Gamma variables with parameters $\alpha_i = \frac{1}{2}$ and $\beta_i(h) = \frac{n_h}{2\ell_i(h)}$, $\forall i \in \{1,\dots,n\}$. Let $\beta_*(h) = \min_{i \leq n_h}\Big\{\beta_i(h)\Big\} = \frac{n_h}{2\ell_{\max}(h)}$, where $\ell_{\max}(h) = \max_{i \leq n_h} \ell_i(h)$. We first study the term $\mathbb{P}\Big(\widehat{\gamma}(h) \leq \mu_1 - t\Big)$. Using the result from Bercu et al. (2015) (see Theorem 2.24 in subsection 2.2.4, Chapter 2), with $x = \frac{t}{\mu_1} \in]0,1[$, since $\mu_1 - t \geq 0$, as Gamma variables are positive random variables:

$$\mathbb{P}\Big(\widehat{\gamma}(h) \leq \mu_1 - t\Big) \leq \exp\left(-\frac{t^2}{2V_1^2}\right),$$

where $V_1^2 = Var(\widehat{\gamma}(h))$ is the variance of the semi-variogram estimator. Furthermore, from the upper bound on the variance of the semi-variogram estimator given above (see Equation (3)), we have

$$\forall t > 0, \quad \mathbb{P}\Big(\widehat{\gamma}(h) \leq \gamma(h) - t\Big) \leq \exp\Big(-C_1' n t^2\Big), \tag{7}$$

where C_1' is a positive constant depending on j_1 only.

Now, we study the term $\mathbb{P}\Big(\widehat{\gamma}(h) \geq \mu_1 + t\Big)$. Thanks to a slight modification of the result in Wang and Ma (2020) (see Corollary 2.26 in subsection 2.2.4, Chapter 2, for more details) for k independent variables $Z_i \sim \Gamma(u_i, v_i)$, let $\mu_Z = \sum_{i=1}^{k} \mathbb{E}[Z_i]$ and $v_* = \min v_i$

$$\forall z \geq 1, \quad \mathbb{P}\left(\frac{1}{k}\sum_{i=1}^{k}\Big(Z_i - \mathbb{E}[Z_i]\Big) \geq z\mu_Z\right) \leq \exp\Big(-v_*\mu_Z\,(kz - \log(1+kz))\Big).$$

Thus,

$$\mathbb{P}\Big(\widehat{\gamma}(h) \geq \mu_1 + t\Big) \leq \exp\left(-\beta_*(h)\mu_1\left(\frac{t}{\mu_1} - \log\left(1 + \frac{t}{\mu_1}\right)\right)\right).$$

We need an upper bound on the largest eigenvalue of the matrix $L(n,h)\Sigma_n$. First, we use the result presented in Proposition .7, which is derived from some previous works on inequalities for the eigenvalues of the product of positive semi-definite Hermitian

matrices (see *e.g.* Wang and Zhang, 1992; Xi and Zhang, 2019). This result allows us to split the study of the upper bound in two: on one hand the largest eigenvalue of the matrix $L(n,h)$, on the other hand the largest eigenvalue of the covariance matrix Σ_n. Using Proposition .7

$$\xi_1(L(n,h))\xi_n(\Sigma_n) \leq \ell_{\max}(h) \leq \xi_1(L(n,h))\xi_1(\Sigma_n).$$

As defined, $L(n,h)$ is the Laplacian matrix of the graph described by the regular grid, for a fixed $h \in \mathcal{H}_n$ (see Remark .16). Thus, we refer to Proposition .8 above for the result on the bound of the largest eigenvalue of a Laplacian matrix. Furthermore, since the number n of observations is finite and for $h \in \mathcal{H}_n$ the number n_h of pairs at distance h in the regular grid is also finite, the maximum degree $d_{\max}$ of the corresponding graph G_h is also always finite and we have

$$\exists c > 0,\ \xi_1(L(n,h)) \leq c. \tag{8}$$

Now, we deal with the largest eigenvalue of the covariance matrix. Lemma .11, under Assumption 3.8, gives the following bound:

$$\exists\, c' > 0,\ \xi_1(\Sigma_n) \leq c'M \tag{9}$$

Thus, let c_1 a positive constant, such that we have the upper bound on the largest eigenvalue

$$\ell_{\max}(h) \leq c_1 M. \tag{10}$$

This implies the bound on the minimum value of the parameters $\beta_i(h)$

$$\exists\, C > 0,\ \beta_*(h) \geq C n_h,$$

where C depends on M. Combining this result with the lower bound on n_h in Lemma 3.6, we have

$$\forall t > 0, \quad \mathbb{P}\Big(\widehat{\gamma}(h) \geq \gamma(h) + t\Big) \leq \exp\Big(-C_1 n t\Big), \tag{11}$$

where C_1 is a positive constant depending on j_1 and M only, which concludes the proof for the semi-variogram estimator tail bounds.

In the second part of the proof, we study the variance estimator $\widehat{c}_h(0)$, with similar steps as for the previous result on the semi-variogram estimator. Let $t > 0$ and $h \in \mathcal{H}_n$, $\mathbb{P}\Big(\big|\widehat{c}_h(0) - c(0)\big| \geq t\Big) = \mathbb{P}\Big(\widehat{c}_h(0) \geq c(0) + t\Big) + \mathbb{P}\Big(\widehat{c}_h(0) \leq c(0) - t\Big)$. Let $\mu_2 = \mathbb{E}[\widehat{c}_h(0)] = c(0)$ (since $\widehat{c}_h(0)$ is unbiased). Recall that, thanks to Proposition 3.7: $\widehat{c}_h(0) \sim \frac{1}{n_h}\sum_{i=1}^{n_h} \rho_i(h)\chi_i^2$. Since the eigenvalues $\rho_i(h), \forall i \in \{1,\dots,n_h\}$ of $D(n,h)\Sigma_n$ are non negatives, from the relationship between Gamma and χ^2 random variables, $\frac{1}{n_h}\rho_i(h)\chi_i^2 \sim \Gamma\Big(\frac{1}{2}, \frac{n_h}{2\rho_i(h)}\Big)$. This implies that $\widehat{c}_h(0)$ can be seen as the sum of n_h independent Gamma random variables with parameters $a_i = \frac{1}{2}$ and $b_i(h) = \frac{n_h}{2\rho_i(h)}$, $\forall i \in \{1,\dots,n\}$. Let $b_*(h) = \min_{i \leq n_h}\Big\{b_i(h)\Big\} = \frac{n_h}{2\rho_{\max}(h)}$, where $\rho_{\max}(h) = \max_{i \leq n_h} \rho_i(h)$. We first study the term $\mathbb{P}\Big(\widehat{c}_h(0) \leq \mu_2 - t\Big)$. Using the result from Bercu et al. (2015) in Theorem 2.24, Chapter 2, with $x = \frac{t}{\mu_2} \in]0,1[$:

$$\mathbb{P}\Big(\widehat{c}_h(0) \leq \mu_2 - t\Big) \leq \exp\left(-\frac{t^2}{2V_2^2}\right),$$

where $V_2^2 = Var(\widehat{c}_h(0))$. From the result given above (see Equation (4)), we have

$$\forall t > 0, \quad \mathbb{P}\Big(\widehat{c}_h(0) \leq c(0) - t\Big) \leq \exp\Big(-C_2' n t^2\Big), \tag{12}$$

where C_2' is a positive constant depending on j_1 only.

Now, we study the term $\mathbb{P}\Big(\widehat{c}_h(0) \geq \mu_2 + t\Big)$. Combining the bound for the product of matrices given in Proposition .7, the bound on the largest eigenvalue of the matrix of the degrees of a graph (see Proposition .9) and the bound on the largest eigenvalue of a covariance matrix in Equation (9), we have an upper bound on the largest eigenvalue of the matrix $D(n,h)\Sigma_n$:

$$\rho_{\max}(h) \leq c_2 M. \tag{13}$$

Thus, using the same argumentation as for the semi-variogram estimator

$$\forall t > 0, \quad \mathbb{P}\Big(\widehat{c}_h(0) \geq c(0) + t\Big) \leq \exp\Big(-C_2 n t\Big), \tag{14}$$

where C_2 is a positive constant depending on j_1 and M only, which concludes the proof.

Proof of Corollary 3.11

For the proof, we shall use both preliminary results in Appendix A.1: an upper bound on the total number of distinct observable distances and Corollary .15, which proof, that simply follows from Proposition 3.9, is given in Appendix A.1. In a first place, we study $\Big|\widehat{c}(h) - c(h)\Big|$ for all $h \geq 0$. Thanks to the definition of the covariance function estimation at unobserved lags by mean of the 1-NN estimator (see Section 3.2), for any distance h, let $h_o \in \mathcal{H}_n$ the observable distance that is the 1-NN of h and such that: $\widehat{c}(h) = \widehat{c}(h_o)$. Then,

$$\Big|\widehat{c}(h) - c(h)\Big| = \Big|\widehat{c}(h) - c(h) + c(h_o) - c(h_o)\Big| \leq \Big|\widehat{c}(h_o) - c(h_o)\Big| + \Big|c(h_o) - c(h)\Big|$$

Applying the mean value (or finite increment) inequality, combined with Assumption 3.10, we have

$$\Big|c(h_o) - c(h)\Big| \leq Q\|h - h_o\| \leq \frac{Q}{\sqrt{n} - 1},$$

since $\forall h \geq 0$, $\|h - h_o\| \leq 1/(\sqrt{n} - 1)$ (see Section 3.2). From the relationship between the covariance and the semi-variogram functions and the relationship for their estimators in (3.5), we have

$$\Big|\widehat{c}(h_o) - c(h_o)\Big| \leq \Big|\widehat{c}_{h_o}(0) - c(0)\Big| + \Big|\widehat{\gamma}(h_o) - \gamma(h_o)\Big|.$$

Then, we have:

$$\sup_{h \geq 0} \Big|\widehat{c}(h) - c(h)\Big| \leq \max_{h \in \mathcal{H}_n} \Big|\widehat{c}_h(0) - c(0)\Big| + \max_{h \in \mathcal{H}_n} \Big|\widehat{\gamma}(h) - \gamma(h)\Big| + \frac{Q}{\sqrt{n} - 1}.$$

This yields: $\forall t > 0$,

$$\begin{aligned}
&\mathbb{P}\left(\sup_{h\geq 0}\left|\widehat{c}(h)-c(h)\right|\geq t\right)\\
&\quad\leq \mathbb{P}\left(\max_{h\in\mathcal{H}_n}\left|\widehat{c}_h(0)-c(0)\right|+\max_{h\in\mathcal{H}_n}\left|\widehat{\gamma}(h)-\gamma(h)\right|+Q/(\sqrt{n}-1)\geq t\right)\\
&\quad\leq \mathbb{P}\left(\max_{h\in\mathcal{H}_n}\left|\widehat{c}_h(0)-c(0)\right|\geq \frac{1}{2}\Big(t-Q/(\sqrt{n}-1)\Big)\right)\\
&\qquad+\mathbb{P}\left(\max_{h\in\mathcal{H}_n}\left|\widehat{\gamma}(h)-\gamma(h)\right|\geq \frac{1}{2}\Big(t-Q/(\sqrt{n}-1)\Big)\right).
\end{aligned}$$

Then, we can apply the result in Corollary .15 for both estimators with $k=\Big(t-Q/(\sqrt{n}-1)\Big)/2$ and we obtain

$$\forall t>0,\ \mathbb{P}\left(\sup_{h\geq 0}\left|\widehat{c}(h)-c(h)\right|\geq t\right)\leq 2ne^{-C_2'nk^2}+2ne^{-C_1'nk^2},$$

as soon as $k\leq \min\left\{C_1/C_1',\ C_2/C_2'\right\}=C'_{min}$. Furthermore, we have

$$2ne^{-C_2'nk^2}+2ne^{-C_1'nk^2}\leq 4n\max\left\{e^{-C_2'nk^2},e^{-C_1'nk^2}\right\}=4ne^{-C_{min}nk^2},$$

where $C_{min}=\min\{C_1',C_2'\}$. Finally, let $\delta\in(0,1)$, such that $\delta=4ne^{-C_{min}nk^2}$ with $k=\frac{1}{2}\Big(t-Q/(\sqrt{n}-1)\Big)$. Thus, by a simple calculation, this implies that there exists a positive constant $C_3=2/\sqrt{C_{min}}$ depending on j_1, m and M solely such that $t=C_3\sqrt{\log\Big(4n/\delta\Big)/n}+Q/(\sqrt{n}-1)$. Furthermore, going back to the condition on the variable k, by a straightforward computation, we have

$$k=\frac{1}{2}\left(t-\frac{Q}{\sqrt{n}-1}\right)\leq C'_{min}\Longleftrightarrow n\geq C_3'\log\left(\frac{4n}{\delta}\right),$$

where C_3' is a positive constant depending on j_1, m and M solely. Thus,

$$\mathbb{P}\left(\sup_{h\geq 0}\left|\widehat{c}(h)-c(h)\right|\leq C_3\sqrt{\log\Big(4n/\delta\Big)/n}+Q/(\sqrt{n}-1)\right)\geq 1-\delta,$$

as soon as $n\geq C_3'\log\left(\frac{4n}{\delta}\right)$.

B Appendices for Chapter 4

In this Appendix are gathered all the technical proofs of the results stated in Chapter 4.

B.1 Auxiliary Result for the Proof of Proposition 4.7

We present an auxiliary result from Wedin, 1973, Theorem 4.1 (see also Staerman et al., 2021, Lemma 5) used in the proof of Proposition 4.7.

Theorem .17. *Let A and B be two invertible matrices of size $d \times d$. Then it holds:*

$$|||A^{-1} - B^{-1}||| \leq |||A^{-1}|||\,|||B^{-1}|||\,|||A - B|||. \tag{15}$$

B.2 Proof of Proposition 4.7

Proof of Assertion *(i)*

First, recall that the max norm and the operator norm are equivalent (since any norms in a given finite-dimensional vector space are equivalent and that the space of the squared matrices of size d is a finite-dimensional vector space):

$$|||\widehat{\Sigma}(\mathbf{s}_d) - \Sigma(\mathbf{s}_d)||| \leq d\,\|\widehat{\Sigma}(\mathbf{s}_d) - \Sigma(\mathbf{s}_d)\|_\infty, \tag{16}$$

where $\|A\|_\infty = \max_{i,j\in\{1,\cdots,d\}} |A_{ij}|$ is the max norm for any squared matrix A of size d. By the definition of the estimated and the true covariance matrices, notice that

$$\begin{aligned}\|\widehat{\Sigma}(\mathbf{s}_d) - \Sigma(\mathbf{s}_d)\|_\infty &= \max_{i,j\in\{1,\cdots,d\}} \left|\widehat{c}(\|s_i - s_j\|) - c(\|s_i - s_j\|)\right| \\ &\leq \sup_{h\geq 0} \left|\widehat{c}(h) - c(h)\right|.\end{aligned}$$

Then, applying the non-asymptotic bound in Corollary 3.11, we have the wanted result.

Proof of Assertion *(ii)*

Thanks to the result in the previous assertion, where the operator norm of the difference $\widehat{\Sigma}(\mathbf{s}_d) - \Sigma(\mathbf{s}_d)$ is bounded with high probability, we can deduce that the eigenvalues of $\widehat{\Sigma}(\mathbf{s}_d)$ have near values to the eigenvalues of $\Sigma(\mathbf{s}_d)$. Recall that the eigenvalues of $\Sigma(\mathbf{s}_d)$ are assumed to be bounded by $\underline{m}$ and $\overline{M}$, two positive constants (see Assumption 4.6). Thus, with high probability, the spectrum of $\widehat{\Sigma}(\mathbf{s}_d)$ is also bounded by $\underline{m} > 0$ and $\overline{M} > 0$. Finally, we can deduce that $\widehat{\Sigma}(\mathbf{s}_d)$ is invertible with high probability. The first step of the proof is to apply the result from Theorem .17. Indeed

$$|||\widehat{\Sigma}(\mathbf{s}_d)^{-1} - \Sigma(\mathbf{s}_d)^{-1}||| \leq |||\Sigma(\mathbf{s}_d)^{-1}|||\;|||\widehat{\Sigma}(\mathbf{s}_d)^{-1}|||\;|||\widehat{\Sigma}(\mathbf{s}_d) - \Sigma(\mathbf{s}_d)|||. \tag{17}$$

First notice that under Assumption 3.5, $\Sigma(\mathbf{s}_d)$ is always positive definite and invertible, so all its eigenvalues are strictly positive. Furthermore, we know that the operator norm of a symmetric positive definite matrix is equal to the largest eigenvalue of the matrix: $|||\Sigma(\mathbf{s}_d)^{-1}||| = \max_{i\in\{1,\cdots,d\}} \xi_i\left(\Sigma(\mathbf{s}_d)^{-1}\right) = \xi_d\left(\Sigma(\mathbf{s}_d)\right)^{-1}$. Finally, one has:

$$|||\Sigma(\mathbf{s}_d)^{-1}||| \leq \underline{m}^{-1}, \tag{18}$$

where $\underline{m} > 0$ is the lower bound of the spectrum of $\Sigma(\mathbf{s}_d)$. As a consequence of Assertion *(i)*, the eigenvalues of $\widehat{\Sigma}(\mathbf{s}_d)$ are also bounded and bounded away from 0, with high probability. Using the same argumentation as above, one has, with high probability:

$$\forall \delta \in (0,1),\ \mathbb{P}\left(|||\widehat{\Sigma}(\mathbf{s}_d)^{-1}||| \leq \underline{m}^{-1}\right) \geq 1-\delta. \tag{19}$$

Thus, going back to the inequality (17)

$$\forall \delta \in (0,1),\ \mathbb{P}\left(|||\widehat{\Sigma}(\mathbf{s}_d)^{-1} - \Sigma(\mathbf{s}_d)^{-1}||| \leq (\underline{m}^{-1})^2\, |||\widehat{\Sigma}(\mathbf{s}_d) - \Sigma(\mathbf{s}_d)|||\right) \geq 1-\delta.$$

Combining all the previous results and the accuracy $|||\widehat{\Sigma}(\mathbf{s}_d) - \Sigma(\mathbf{s}_d)|||$ of the covariance matrix estimator (described in a non-asymptotic fashion by the bound given in Assertion *(i)*), one can deduce the result in Assertion *(ii)*.

B.3 Proof of Theorem 4.8

Proof of Assertion *(i)*

First, notice that

$$\begin{aligned}
\|\widehat{\Lambda}_d(s) - \Lambda_d^*(s)\| &= \|\widehat{\Sigma}(\mathbf{s}_d)^{-1}\widehat{\mathbf{c}}_d(s) - \Sigma(\mathbf{s}_d)^{-1}\mathbf{c}_d(s)\| \\
&= \|\Sigma(\mathbf{s}_d)^{-1}\left(\widehat{\mathbf{c}}_d(s) - \mathbf{c}_d(s)\right) + \left(\widehat{\Sigma}(\mathbf{s}_d)^{-1} - \Sigma(\mathbf{s}_d)^{-1}\right)\widehat{\mathbf{c}}_d(s)\| \\
&\leq |||\Sigma(\mathbf{s}_d)^{-1}|||\, \|\widehat{\mathbf{c}}_d(s) - \mathbf{c}_d(s)\| + |||\widehat{\Sigma}(\mathbf{s}_d)^{-1} - \Sigma(\mathbf{s}_d)^{-1}|||\, \|\widehat{\mathbf{c}}_d(s)\|,
\end{aligned}$$

and taking the supremum over the domain $\mathcal{S}$

$$\begin{aligned}
\sup_{s\in\mathcal{S}} \|\widehat{\Lambda}_d(s) - \Lambda_d^*(s)\| \leq\ & |||\Sigma(\mathbf{s}_d)^{-1}|||\, \sup_{s\in\mathcal{S}} \|\widehat{\mathbf{c}}_d(s) - \mathbf{c}_d(s)\| \\
& + |||\widehat{\Sigma}(\mathbf{s}_d)^{-1} - \Sigma(\mathbf{s}_d)^{-1}|||\, \sup_{s\in\mathcal{S}} \|\widehat{\mathbf{c}}_d(s)\|,
\end{aligned}$$

Firstly, for the accuracy of the covariance vector estimator, since the max norm and the Euclidean norm are equivalent, one has

$$\begin{aligned}
\sup_{s\in\mathcal{S}} \|\widehat{\mathbf{c}}_d(s) - \mathbf{c}_d(s)\| &\leq \sqrt{d}\, \sup_{s\in\mathcal{S}} \|\widehat{\mathbf{c}}_d(s) - \mathbf{c}_d(s)\|_\infty \\
&= \sqrt{d}\, \sup_{s\in\mathcal{S}} \max_{i\in\{1,\cdots,d\}} |\widehat{c}(\|s - s_i\|) - c(\|s - s_i\|)| \leq \sqrt{d}\, \sup_{h\geq 0} |\widehat{c}(h) - c(h)|,
\end{aligned}$$

which allows using Corollary 3.11 to control in a non-asymptotic fashion the supremum over all positive lags of the error estimation of the covariance function. Thus, one obtains the following bound for any $\delta \in (0,1)$, with probability at least $1-\delta$

$$\sup_{s\in\mathcal{S}} \|\widehat{\mathbf{c}}_d(s) - \mathbf{c}_d(s)\| \leq C_3\, \sqrt{d}\, \sqrt{\log(4n/\delta)/n} + \sqrt{d}\, Q/(\sqrt{n} - 1), \tag{20}$$

as soon as $n \geq C_3' \log(4n/\delta)$. As above, from the link between the max norm and the Euclidean norm, one has

$$\sup_{s\in\mathcal{S}} \|\widehat{\mathbf{c}}_d(s)\| \leq \sqrt{d}\, \sup_{s\in\mathcal{S}} \|\widehat{\mathbf{c}}_d(s)\|_\infty \leq \sqrt{d}\, \sup_{h\geq 0} |\widehat{c}(h)|.$$

Furthermore, as a consequence of the result in Corollary 3.11, we can deduce that $\widehat{c}(h)$ is close to $c(h)$ for any lag $h \geq 0$, with high probability. Thus, under Assumption 3.10, one has the bound

$$\forall \delta \in (0,1), \mathbb{P}\left(\sup_{s \in \mathcal{S}} \|\widehat{\mathbf{c}}_d(s)\| \leq \sqrt{d} B \right) \geq 1 - \delta. \tag{21}$$

Lastly, notice that the last two terms have been studied in previous proofs: $|||\Sigma(\mathbf{s}_d)^{-1}|||$ is upper bounded by $\underline{m}^{-1}$ (see Equation (18) in the proof of Proposition 4.7 Assertion *(ii)*) ; and $|||\widehat{\Sigma}(\mathbf{s}_d)^{-1} - \Sigma(\mathbf{s}_d)^{-1}|||$ is bounded in a non-asymptotic fashion in Proposition 4.7 Assertion *(ii)*. Thus, combining all the previous results, one can deduce the wanted non-asymptotic bound.

Proof of Assertion *(ii)*

As announced in subsection 4.3.2, with probability one, the excess of integrated quadratic risk can be written as follows:

$$\begin{aligned}
L_{\mathcal{S}}(f_{\widehat{\Lambda}_d}) - L^*_{\mathcal{S}} = & \\
& \int_{s \in \mathcal{S}} \left({}^t\widehat{\Lambda}_d(s) \Sigma(\mathbf{s}_d) \widehat{\Lambda}_d(s) - {}^t\Lambda^*_d(s) \Sigma(\mathbf{s}_d) \Lambda^*_d(s) - 2\, {}^t\mathbf{c}_d(s) \left(\widehat{\Lambda}_d(s) - \Lambda^*_d(s) \right) \right) ds \\
& = \int_{s \in \mathcal{S}} {}^t\left(\widehat{\Lambda}_d(s) - \Lambda^*_d(s) \right) \Sigma(\mathbf{s}_d) \widehat{\Lambda}_d(s) + {}^t\Lambda^*_d(s) \Sigma(\mathbf{s}_d) \left(\widehat{\Lambda}_d(s) - \Lambda^*_d(s) \right) \\
& \qquad + 2\, {}^t\mathbf{c}_d(s) \left(\Lambda^*_d(s) - \widehat{\Lambda}_d(s) \right) ds
\end{aligned}$$

Notice that

$$\begin{aligned}
{}^t\left(\widehat{\Lambda}_d(s) - \Lambda^*_d(s) \right) \Sigma(\mathbf{s}_d) \widehat{\Lambda}_d(s) &= \langle \widehat{\Lambda}_d(s) - \Lambda^*_d(s), \Sigma(\mathbf{s}_d) \widehat{\Lambda}_d(s) \rangle \\
&\leq \|\widehat{\Lambda}_d(s) - \Lambda^*_d(s)\| \, \|\Sigma(\mathbf{s}_d) \widehat{\Lambda}_d(s)\| \\
&\leq \|\widehat{\Lambda}_d(s) - \Lambda^*_d(s)\| \, |||\Sigma(\mathbf{s}_d)||| \, |||\widehat{\Sigma}(\mathbf{s}_d)^{-1}||| \, \|\widehat{\mathbf{c}}_d(s)\|.
\end{aligned}$$

Following the same idea, we have

$$\begin{aligned}
& {}^t\Lambda^*_d(s) \Sigma(\mathbf{s}_d) \left(\widehat{\Lambda}_d(s) - \Lambda^*_d(s) \right) \\
& \qquad \leq |||\Sigma(\mathbf{s}_d)^{-1}||| \, \|\mathbf{c}_d(s)\| \, |||\Sigma(\mathbf{s}_d)||| \, \|\widehat{\Lambda}_d(s) - \Lambda^*_d(s)\|,
\end{aligned}$$

and

$${}^t\mathbf{c}_d(s) \left(\Lambda^*_d(s) - \widehat{\Lambda}_d(s) \right) \leq \|\mathbf{c}_d(s)\| \, \|\Lambda^*_d(s) - \widehat{\Lambda}_d(s)\|.$$

Thus, taking the supremum over the domain $\mathcal{S}$, the term under study has now become

$$\begin{aligned}
& \sup_{s \in \mathcal{S}} \|\widehat{\Lambda}_d(s) - \Lambda^*_d(s)\| \, |||\Sigma(\mathbf{s}_d)||| \, |||\widehat{\Sigma}(\mathbf{s}_d)^{-1}||| \sup_{s \in \mathcal{S}} \|\widehat{\mathbf{c}}_d(s)\| \\
& \qquad + \sup_{s \in \mathcal{S}} \|\widehat{\Lambda}_d(s) - \Lambda^*_d(s)\| \, |||\Sigma(\mathbf{s}_d)||| \, |||\Sigma(\mathbf{s}_d)^{-1}||| \sup_{s \in \mathcal{S}} \|\mathbf{c}_d(s)\| \\
& \qquad + 2 \sup_{s \in \mathcal{S}} \|\widehat{\Lambda}_d(s) - \Lambda^*_d(s)\| \sup_{s \in \mathcal{S}} \|\mathbf{c}_d(s)\|.
\end{aligned}$$

Notice that some of the terms have already been studied in previous results and their proofs: a non-asymptotic bound for $\sup_{s\in\mathcal{S}} \|\widehat{\Lambda}_d(s) - \Lambda_d^*(s)\|$ is given by Theorem 4.8 Assertion *(i)* ; the operator norm of the precision matrix estimator is upper bounded by $\underline{m}^{-1}$, with high probability, in Equation (19) (see the proof of Proposition 4.7 Assertion *(ii)*) ; the supremum over all domain $\mathcal{S}$ of the Euclidean norm of the covariance vector estimator is upper bounded with high probability by $\sqrt{d}B$ (see Equation (21), in the proof of Assertion *(i)*) ; and $|||\Sigma(\mathbf{s}_d)^{-1}|||$ is upper bounded by $\underline{m}^{-1}$ (see Equation (18) in the proof of Proposition 4.7 Assertion *(ii)*). Furthermore, from Assumption 4.6, $|||\Sigma(\mathbf{s}_d)||| \leq \overline{M}$. Finally, for the last term defined as the supremum over the domain $\mathcal{S}$ of the Euclidean norm of the covariance vector, using the link between the max norm and the Euclidean norm, and Assumption 3.10

$$\sup_{s\in\mathcal{S}} \|\mathbf{c}_d(s)\| \leq \sqrt{d} \sup_{s\in\mathcal{S}} \|\mathbf{c}_d(s)\|_\infty \leq \sqrt{d} \sup_{h\geq 0} |c(h)| \leq \sqrt{d}B.$$

Since the domain $\mathcal{S}$ is bounded, combining all these results allows us to conclude.

C Appendices for Chapter 6

C.1 Proof of Proposition 6.2

Denote by $\delta^i_{x,n}, \delta^i_{y,n}, \delta^i_{t,n}$ the spread between the original events and the n-th projected event of the i-th process, i.e. $\tilde{x}^i_n = x^i_n + \delta^i_{x,n}$, $\tilde{y}^i_n = y^i_n + \delta^i_{y,n}$, $\tilde{t}^i_n = t^i_n + \delta^i_{t,n}$ with $\delta^i_{x,n} \in [-\Delta_{\mathcal{X}}/2, \Delta_{\mathcal{X}}/2]$, $\delta^i_{y,n} \in [-\Delta_{\mathcal{Y}}/2, \Delta_{\mathcal{Y}}/2]$ and $\delta^i_{t,n} \in [-\Delta_T/2, \Delta_T/2]$. Recall that for any $v = (v_x, v_y, v_t) \in [\![0, G_{\mathcal{X}}]\!] \times [\![0, G_{\mathcal{Y}}]\!] \times [\![0, G_T]\!]$, we have

$$\lambda_i(v_x\Delta_{\mathcal{X}}, v_y\Delta_{\mathcal{Y}}, v_t\Delta_T) = \mu_i + \sum_{j=1}^{p} \sum_{u^j_m \in \mathcal{H}^j_T} g_{ij}(v_x\Delta_{\mathcal{X}} - x^j_m, v_y\Delta_{\mathcal{Y}} - y^j_m, v_t\Delta_T - t^j_m).$$

By defining $\Delta = (\Delta_{\mathcal{X}}, \Delta_{\mathcal{Y}}, \Delta_T)$, $\delta^i_n = (\delta^i_{x,n}, \delta^i_{y,n}, \delta^i_{t,n})$, $\forall\ i \in [\![1, D]\!]$ and $\forall\ 1 \le n \le N^i_T$, the vector of the intensity function on the grid is given by

$$\begin{aligned}
\tilde{\lambda}_i[v] &= \mu_i + \sum_{j=1}^{D} \sum_{\tilde{u}^j_m \in \tilde{\mathcal{H}}^j_{v\Delta}} g_{ij}(v\Delta - \tilde{u}^j_m) \\
&= \mu_i + \sum_{j=1}^{D} \sum_{u^j_m \in \mathcal{H}^j_{v\Delta}} g_{ij}(v\Delta - u^j_m - \delta^j_m) \qquad (22)
\end{aligned}$$

where 22 holds because $\Delta_{\mathcal{X}} < \min_{x^i_n, x^j_m \in \mathcal{H}_T} |x^i_n - x^j_m|$, $\Delta_{\mathcal{Y}} < \min_{y^i_n, y^j_m \in \mathcal{H}_T} |y^i_n - y^j_m|$, and $\Delta_T < \min_{t^i_n, t^j_m \in \mathcal{H}_T} |t^i_n - t^j_m|$, which ensures that no event collapses on the same bin of the grid and that $|\widetilde{\mathcal{H}}^j_{v\Delta}| = |\mathcal{H}^j_{v\Delta}|$, where $|\cdot|$ denotes the cardinal of a set. Note that this hypothesis also implies that the intensity function is smooth for all points on the grid $\mathcal{G}$. Applying the first-order Taylor expansion to the kernels g_{ij} in $v\Delta - u^j_m$ and bounding the perturbation δ^j_m by Δ yields the first result of the proposition.

For the perturbation of the discrete loss, we have

$$\begin{aligned}
\mathcal{L}_{\mathcal{G}}(\boldsymbol{\theta}, \widetilde{\mathcal{H}}_T) &= \sum_{i=1}^{D} \left(\Delta_{\mathcal{X}}\Delta_{\mathcal{Y}}\Delta_T \sum_{v_x=0}^{G_{\mathcal{X}}} \sum_{v_y=0}^{G_{\mathcal{Y}}} \sum_{v_t=0}^{G_T} \left(\tilde{\lambda}_i[v]\right)^2 - 2 \sum_{\tilde{u}^i_n \in \widetilde{\mathcal{H}}^i_T} \tilde{\lambda}_i\left[\frac{\tilde{u}^i_n}{\Delta}\right] \right) \\
&= \mathcal{L}(\theta, \mathcal{H}_T) + \sum_{i=1}^{D} \Bigg(\underbrace{\Delta_{\mathcal{X}}\Delta_{\mathcal{Y}}\Delta_T \sum_{v_x=0}^{G_{\mathcal{X}}} \sum_{v_y=0}^{G_{\mathcal{Y}}} \sum_{v_t=0}^{G_T} \tilde{\lambda}_i[v]^2 - \int_0^T \int_{\mathcal{S}} \lambda_i(x,y,t)^2 \mathrm{d}x\mathrm{d}y\mathrm{d}t}_{(*)} \\
&\qquad \underbrace{- 2 \sum_{u^i_n \in \mathcal{H}^i_T} \left(\widetilde{\lambda}_i\left[\frac{\widetilde{u}^i_n}{\Delta}\right] - \lambda_i\left(u^i_n\right) \right)}_{(**)} \Bigg),
\end{aligned}$$

where $\frac{\tilde{u}_n^i}{\Delta}$ is the division term by term of these three-dimensional vectors. The first term $(*)$ is the error of a Riemann approximation of the integral. We can use the generalization of the Koksma-Hlawka inequality (Brandolini et al., 2013) for piece-wise smooth functions on compact set of $\mathbb{R}^d$:

$$\left|\Delta_{\mathcal{X}}\Delta_{\mathcal{Y}}\Delta_T \sum_{v_x=0}^{G_{\mathcal{X}}}\sum_{v_y=0}^{G_{\mathcal{Y}}}\sum_{v_t=0}^{G_T}\tilde{\lambda}_i[v]^2 - \int_0^T\int_{\mathcal{S}}\lambda_i(x,y,t)^2\mathrm{d}x\mathrm{d}y\mathrm{d}t\right| \leq C(\lambda_i)\|\Delta\|, \tag{23}$$

where $\|\Delta\|$ comes from the maximal distance on the uniform spatio-temporal grid and $C(\lambda_i)$ is a constant that depends on the regularity of λ_i, see Theorem 1 in Brandolini et al. (2013).

For the second term $(**)$, we re-use the expression from (22) but use a Taylor expansion in $u_n^i - u_m^j$. The perturbation becomes $\delta_m^j - \delta_n^i$,

$$\sum_{u_n^i\in\mathcal{H}_T^i}\left(\widetilde{\lambda}_i\left[\frac{\widetilde{u}_n^i}{\Delta}\right] - \lambda_i\left(u_n^i\right)\right) = \sum_{j=1}^{D}\sum_{\substack{u_n^i\in\mathcal{H}_T^i\\ u_m^j\in\mathcal{H}_T^j}}\left(\delta_n^i - \delta_m^j\right)\nabla_u g_{ij}\left(u_n^i - u_m^j\right) + O\left(\|\Delta\|^2\right). \tag{24}$$

Summing (23) and (24) concludes the proof.

C.2 Proof of Proposition 6.3

We consider the two estimators $\widehat{\boldsymbol{\theta}}_\Delta = \arg\min_{\boldsymbol{\theta}}\ \mathcal{L}_{\mathcal{G}}(\boldsymbol{\theta}, \widetilde{\mathcal{H}}_T)$ and $\widehat{\boldsymbol{\theta}}_c = \arg\min_{\boldsymbol{\theta}}\ \mathcal{L}(\boldsymbol{\theta}, \mathcal{H}_T)$. With the loss approximation from Proposition 6.2, we have a point-wise convergence of $\mathcal{L}_{\mathcal{G}}(\boldsymbol{\theta}, \widetilde{\mathcal{H}}_T)$ towards $\mathcal{L}(\boldsymbol{\theta}, \mathcal{H}_T)$ for all $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ as $\|\Delta\|$ goes to 0. By continuity of $\mathcal{L}_{\mathcal{G}}(\boldsymbol{\theta}, \widetilde{\mathcal{H}}_T)$, we have that the limit of $\widehat{\boldsymbol{\theta}}_\Delta$ when $\|\Delta\|$ goes to 0 exists and is equal to $\widehat{\boldsymbol{\theta}}_c$. This proves that the discretized estimator converges to the continuous one as $\|\Delta\|$ decreases.

The Karush-Kuhn-Tucker conditions imply that:

$$\nabla_{\boldsymbol{\theta}}\mathcal{L}_{\mathcal{G}}\left(\widehat{\boldsymbol{\theta}}_\Delta, \widetilde{\mathcal{H}}_T\right) = 0 \qquad \text{and} \qquad \nabla_{\boldsymbol{\theta}}\mathcal{L}\left(\widehat{\boldsymbol{\theta}}_c, \mathcal{H}_T\right) = 0. \tag{25}$$

Using the approximation from (23) and (24), one gets in the limit of small $\|\Delta\|$:

$$\begin{aligned}\nabla_{\boldsymbol{\theta}}\mathcal{L}_{\mathcal{G}}(\widehat{\boldsymbol{\theta}}_\Delta, \widetilde{\mathcal{H}}_T) \geq &\nabla_{\boldsymbol{\theta}}\mathcal{L}(\widehat{\boldsymbol{\theta}}_\Delta, \mathcal{H}_T) - \|\Delta\|\sum_i \nabla_{\boldsymbol{\theta}}C(\lambda_i)\\ &+ 2\sum_{i,j}\sum_{\substack{u_n^i\in\mathcal{H}_T^i\\ u_m^j\in\mathcal{H}_T^j}}\left(\delta_m^j - \delta_n^i\right).\nabla_{\boldsymbol{\theta}}\nabla_u g_{ij}\left(u_n^i - u_m^j\right) + O\left(\|\Delta\|^2\right).\end{aligned}$$

Combining this with Equation (25), we get:

$$\begin{aligned}\nabla_{\boldsymbol{\theta}}\mathcal{L}(\widehat{\boldsymbol{\theta}}_\Delta, \mathcal{H}_T) \leq &\|\Delta\|\sum_i \nabla_{\boldsymbol{\theta}}C(\lambda_i)\\ &+ 2\sum_{i,j}\sum_{\substack{u_n^i\in\mathcal{H}_T^i\\ u_m^j\in\mathcal{H}_T^j}}\left(\delta_n^i - \delta_m^j\right).\nabla_{\boldsymbol{\theta}}\nabla_u g_{ij}\left(u_n^i - u_m^j\right) + O\left(\|\Delta\|^2\right).\end{aligned}$$

Thus, we have

$$\begin{aligned}\left\|\nabla_\theta \mathcal{L}(\widehat{\theta}_\Delta, \mathcal{H}_T) - \nabla_\theta \mathcal{L}\left(\widehat{\theta}_c, \mathcal{H}_T\right)\right\| &\leq \left\| 2 \sum_{i,j} \sum_{\substack{u_n^i \in \mathcal{H}_T^i \\ u_m^j \in \mathcal{H}_T^j}} \left(\delta_n^i - \delta_m^j\right) . \nabla_\theta \nabla_u g_{ij}\left(u_n^i - u_m^j\right) \right. \\ &\qquad \left. + \|\Delta\| \sum_i \nabla_\theta C(\lambda_i) \right\| + O\left(\|\Delta\|^2\right) \\ &\leq \max\{\|\Delta\|, \|\Delta\|_\infty\}\ \omega(\widehat{\theta}_\Delta),\end{aligned}$$

where $\omega(\theta) = \left\| 2 \sum_{i,j} \sum_{\substack{u_n^i \in \mathcal{H}_T^i \\ u_m^j \in \mathcal{H}_T^j}} \langle \mathbf{1}, \nabla_\theta \nabla_u g_{ij}\left(u_n^i - u_m^j\right)\rangle + \sum_i \nabla_\theta C(\lambda_i) \right\|$ with $\mathbf{1}$ a three-dimensional vector of one. This function is a $O(1)$. Using the hypothesis that the hessian $\nabla^2_\theta \mathcal{L}(\widehat{\theta}_c, \mathcal{H}_T)$ exists and is positive definite with smallest eigenvalue ε, we have:

$$\begin{aligned}\varepsilon \left\|\widehat{\theta}_\Delta - \widehat{\theta}_c\right\|^2 &\leq \left\|\nabla_\theta \mathcal{L}\left(\widehat{\theta}_\Delta, \mathcal{H}_T\right) - \nabla_\theta \mathcal{L}\left(\widehat{\theta}_c, \mathcal{H}_T\right)\right\|^2 \\ \textit{i.e.} \qquad \varepsilon \left\|\widehat{\theta}_\Delta - \widehat{\theta}_c\right\|^2 &\leq \frac{\max\{\|\Delta\|, \|\Delta\|_\infty\}}{\varepsilon}\ \omega\left(\widehat{\theta}_\Delta\right).\end{aligned}$$

This concludes the proof.

Bibliography

Northern California Earthquake Data Center, NCEDC Dataset. *UC Berkeley Seismological Laboratory*, 2014. doi: doi:10.7932/NCEDC. pages 28, 124, 129, 146, 178

R. Adamczak, A. Litvak, A. Pajor, and N. Tomczak-Jaegermann. Quantitative estimates of the convergence of the empirical covariance matrix in log-concave ensembles. *Journal of the American Mathematical Society*, 23(2):535–561, 2010. page 58

M. Armstrong. Improving the estimation and modelling of the variogram. In *Geostatistics for natural resources characterization*, pages 1–19. Springer, 1984. doi: 10.1007/978-94-009-3699-7_1. page 57

M. Armstrong. *Experimental Variograms in Basic linear geostatistics*. Springer Science & Business Media, 1998. doi: 10.1007/978-3-642-58727-6_4. page 50

E. Bacry, I. Mastromatteo, and J.-F. Muzy. Hawkes Processes in Finance. *Market Microstructure and Liquidity*, 01(01):1550005, 2015. doi: 10.1142/S2382626615500057. page 111

E. Bacry, M. Bompaire, S. Gaïffas, and J.-F. Muzy. Sparse and low-rank multivariate Hawkes processes. *Journal of Machine Learning Research*, 21(50):1–32, 2020. pages 122, 135

A. Bardossy. Notes on the robustness of the Kriging system. *Mathematical geology*, 20 (3):189–203, 1988. pages 57, 156

L. Bel. Non Parametric Variogram Estimator. Application to Air Pollution Data. In X. Sanchez-Vila, J. Carrera, and J. J. Gómez-Hernández, editors, *geoENV IV — Geostatistics for Environmental Applications*, pages 29–40, Dordrecht, 2004. Springer Netherlands. ISBN 978-1-4020-2115-2. pages 37, 42

B. Bercu, B. Delyon, and E. Rio. *Concentration Inequalities for Sums and Martingales*. Springer, Cham, 2015. doi: 10.1007/978-3-319-22099-4. pages 25, 63, 67, 73, 74, 175, 189, 190

J. Besag. Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society: Series B (Methodological)*, 36(2):192–225, 1974. doi: https://doi.org/10.1111/j.2517-6161.1974.tb00999.x. pages 16, 166

M. Bompaire. *Machine learning based on Hawkes processes and stochastic optimization*. Theses, Université Paris Saclay (COmUE), July 2019. pages 115, 118

S. Boucheron, O. Bousquet, and G. Lugosi. Theory of classification: A survey of some recent advances. *ESAIM: probability and statistics*, 9:323–375, 2005. pages 59, 60

S. Boucheron, G. Lugosi, and P. Massart. *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press, London, U.K., 2013. pages 14, 62, 163

O. Bousquet, S. Boucheron, and G. Lugosi. Introduction to statistical learning theory. In *Summer school on machine learning*, pages 169–207. Springer, 2003. page 59

G. Box and G. Jenkins. *Time Series Analysis: Forecasting and Control*. Holden-Day series in time series analysis and digital processing. Holden-Day, 1970. ISBN 9780816210947. pages 15, 164

L. Brandolini, L. Colzani, G. Gigante, and G. Travaglini. On the Koksma–Hlawka Inequality. *Journal of Complexity*, 29(2):158–172, 2013. pages 139, 198

P. Brockwell and R. Davis. *Time Series: Theory and Methods*. Springer, New York, NY, 1987. pages 69, 73

P. Brooker. A parametric study of robustness of Kriging variance as a function of range and relative nugget effect for a spherical semivariogram. *Mathematical Geology*, 18: 477–488, 1986. page 57

X. Chang and M. L. Stein. Decorrelation property of discrete wavelet transform under fixed-domain asymptotics. *IEEE transactions on information theory*, 59(12):8001–8013, 2013. pages 37, 69

S. Cheng, D. J. Eck, and F. W. Crawford. Estimating the size of a hidden finite set: Large-sample behavior of estimators. *Statistics Surveys*, 14:1 – 31, 2020. doi: 10.1214/19-SS127. page 37

W.-H. Chiang, X. Liu, and G. Mohler. Hawkes process modeling of covid-19 with mobility leading indicators and spatial covariates. *International journal of forecasting*, 38(2):505–520, 2022. page 126

J.-P. Chiles and P. Delfiner. *Geostatistics: Modeling Spatial Uncertainty*. Wiley, New York, 1999. ISBN 0471083151 9780471083153. pages 50, 53, 55, 56, 84, 97, 156

E. Choi and P. Hall. Nonparametric approach to analysis of space-time data on earthquake occurrences. *Journal of Computational and Graphical Statistics*, 8(4):733–748, 1999. page 121

S. Clémençon, G. Ciolek, and P. Bertail. Statistical Learning Based on Markovian Data: Maximal Deviation Inequalities and Learning Rates. *The Annals of Mathematics and Artificial Intelligence*, 2019. pages 15, 164

N. Cressie. *Statistics for Spatial Data*, pages 1–26. John Wiley and Sons, Ltd, New York, NY, 1993. ISBN 9781119115151. doi: 10.1002/9781119115151.ch1. pages 16, 37, 38, 42, 44, 45, 48, 53, 55, 56, 68, 69, 71, 73, 97, 165, 166, 185, 188

N. Cressie and D. Hawkins. Robust estimation of the variogram. *Mathematical Geology*, 12:115–125, 1980. doi: 10.1007/BF01035243. page 48

N. Cressie and H.-C. Huang. Classes of Nonseparable, Spatio-Temporal Stationary Covariance Functions. *Journal of the American Statistical Association*, 94(448):1330–1340, 1999. ISSN 01621459. pages 123, 146, 147, 148

N. Cressie and D. L. Zimmerman. On the stability of the geostatistical method. *Mathematical Geology*, 24:45–59, 1992. pages 44, 57, 71

D. J. Daley and D. Vere-Jones. *An introduction to the theory of point processes: volume I: elementary theory and methods*. Springer, 2003. pages 110, 111, 116, 118, 120, 122

N. Desassis, D. Renard, H. Beucher, S. Petiteau, and X. Freulon. A pairwise likelihood approach for the empirical estimation of the underlyingvariograms in the plurigaussian models. *arXiv preprint arXiv:1510.02668*, 2015. page 53

L. Devroye, L. Györfi, and G. Lugosi. *A Probabilistic Theory of Pattern Recognition*, volume 31. Springer Science & Business Media, New York, NY, 1996. pages 14, 60, 62, 86, 163

P. Diamond and M. Armstrong. Robustness of variograms and conditioning of Kriging matrices. *Journal of the International Association for Mathematical Geology*, 16(8):809–822, 1984. page 57

P. J. Diggle. *Statistical analysis of spatial and spatio-temporal point patterns*. CRC press, 2013. page 110

P. J. Diggle, A. G. Chetwynd, R. Häggkvist, and S. E. Morris. Second-order analysis of space-time clustering. *Statistical methods in medical research*, 4(2):124–136, 1995. page 121

Z. Dong, S. Zhu, Y. Xie, J. Mateu, and F. J. Rodríguez-Cortés. 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, 2023. pages 23, 42, 118, 125, 172

D. L. Donoho, S. G. Mallat, R. von Sachs, et al. Estimating covariances of locally stationary processes: rates of convergence of best basis methods. In *Proceedings IEEE, TFTS-96*, 1996. page 37

C. Díaz-Avalos, P. Juan, and J. Mateu. Significance tests for covariate-dependent trends in inhomogeneous spatio-temporal point processes. *Stochastic Environmental Research and Risk Assessment*, 28, 03 2014. doi: 10.1007/s00477-013-0775-1. pages 123, 160

N. D'Angelo, D. Payares, G. Adelfio, and J. Mateu. Self-exciting point process modelling of crimes on linear networks. *Statistical Modelling*, 0(0):1471082X221094146, 2022. doi: 10.1177/1471082X221094146. pages 23, 118, 125, 149, 159, 172

S. Elogne, O. Perrin, and C. Thomas-Agnan. Non Parametric Estimation of Smooth Stationary Covariance Functions by Interpolation Methods. *Statistical Inference for Stochastic Processes*, 11(2):177–205, June 2008. doi: 10.1007/s11203-007-9014-z. page 49

P. Embrechts, T. Liniger, and L. Lin. Multivariate Hawkes processes: an application to financial data. *Journal of Applied Probability*, 48(A):367–378, 2011. page 118

E. W. Fox, F. P. Schoenberg, and J. S. Gordon. Spatially inhomogeneous background rate estimators and uncertainty quantification for nonparametric Hawkes point process models of earthquake occurrences. *The Annals of Applied Statistics*, 10(3): 1725–1756, 2016. doi: 10.1214/16-AOAS957. page 146

C. Gaetan and X. Guyon. *Spatial Statistics and Modeling*. Springer Series in Statistics. Springer New York, New York, NY, 2009. ISBN 9780387922577. pages 37, 38, 45, 52, 53, 55, 73, 84

M. Genton. Highly robust variogram estimation. *Mathematical Geosciences*, 30(2):213–221, Jan. 1998. ISSN 1874-8961. doi: 10.1023/A:1021728614555. page 48

T. Gneiting. Nonseparable, stationary covariance functions for space-time data. *Journal of the American Statistical Association*, 97(458):590–600, 2002. ISSN 01621459. pages 123, 126, 146, 147, 148

B. I. Golubov. On Abel—Poisson Type and Riesz Means. *Analysis Mathematica*, 7(3): 161–184, Sep 1981. doi: 10.1007/BF01908520. page 75

J. A. González, F. J. Rodríguez-Cortés, O. Cronie, and J. Mateu. Spatio-temporal point process statistics: a review. *Spatial Statistics*, 18:505–544, 2016. page 123

P. Goovaerts. Geostatistical approaches for incorporating elevation into the spatial interpolation of rainfall. *Journal of Hydrology*, 228(1):113–129, 2000. ISSN 0022-1694. doi: https://doi.org/10.1016/S0022-1694(00)00144-X. pages 15, 164

Y. Gratton. Le krigeage: la méthode optimale d'interpolation spatiale. *Les articles de l'Institut d'Analyse Géographique*, 1(4), 2002. page 56

L. Györfi, M. Kohler, A. Krzyzak, and H. Walk. *A Distribution-Free Theory of Nonparametric Regression*. Springer, New York, NY, 2002. pages 60, 62

P. Hall and P. Patil. Properties of Nonparametric Estimators of Autocovariance for Stationary Random Fields. *Probability Theory and Related Fields*, 99(3):399–424, Sep 1994. doi: 10.1007/BF01199899. pages 36, 37, 49, 84, 157

P. Hall, N. I. Fisher, B. Hoffmann, et al. On the Nonparametric Estimation of Covariance Functions. *The Annals of Statistics*, 22(4):2115–2134, 1994. pages 37, 49, 74

S. Hanneke. Learning Whenever Learning is Possible: Universal Learning under General Stochastic Processes. *arXiv:1706.01418*, 2017. pages 15, 164

D. A. Harville. Matrix Algebra From a Statistician's Perspective. *Technometrics*, 40(2): 164–164, 1998. doi: 10.1080/00401706.1998.10485214. page 183

A. G. Hawkes. Point Spectra of Some Mutually Exciting Point Processes. *Journal of the Royal Statistical Society: Series B (Methodological)*, 33(3):438–443, 1971. ISSN 0035-9246. doi: 10.1111/j.2517-6161.1971.tb01530.x. pages 22, 113, 115, 171

A. G. Hawkes and D. Oakes. A cluster process representation of a self-exciting process. *Journal of Applied Probability*, 11(3):493–503, 1974. ISSN 00219002. pages 115, 120

A. J. Holbrook, X. Ji, and M. A. Suchard. From viral evolution to spatial contagion: a biologically modulated Hawkes model. *Bioinformatics*, 38(7):1846–1856, 2022. pages 23, 125, 126, 172

R. A. Horn and C. R. Johnson. *Matrix Analysis*. Cambridge university press, Cambridge, U.K., 2012. page 184

F. Ilhan and S. S. Kozat. Modeling of spatio-temporal Hawkes processes with randomized kernels. *IEEE Transactions on Signal Processing*, 68:4946–4958, 2020. pages 23, 118, 120, 123, 173

H. K. Im, M. L. Stein, and Z. Zhu. Semiparametric estimation of spectral density with irregular observations. *Journal of the American Statistical Association*, 102(478): 726–735, 2007. page 53

S. D. Johnson. Repeat burglary victimisation: a tale of two theories. *Journal of Experimental Criminology*, 4:215–240, 2008. pages 28, 123, 125, 178

M. Kanagawa, P. Hennig, D. Sejdinovic, and B. K. Sriperumbudur. Gaussian Processes and Kernel Methods: A Review on Connections and Equivalences. *arXiv preprint arXiv:1807.02582*, 2018. page 83

M. Kanevski, V. Timonin, and A. Pozdnukhov. *Machine Learning for Spatial Environmental Data: Theory, Applications, and Software*. Environmental sciences research report. CRC Press, 2009. ISBN 9781439808085. pages 37, 59, 83

M. Kirchner. An estimation procedure for the Hawkes process. *Quantitative Finance*, 17(4):571–595, 2017. page 134

C. Kresin, F. P. Schoenberg, and G. Mohler. Comparison of Hawkes and seir models for the spread of covid-19. *Advances and Applications in Statistics*, 74:83–106, 2022. pages 23, 111, 125, 172

D. G. Krige. A Statistical Approach to Some Basic Mine Valuation Problems on the Witwatersrand. *Journal of the Southern African Institute of Mining and Metallurgy*, 52 (6):119–139, 1951. pages 35, 81

D. Kurisu. Nonparametric regression for locally stationary random fields under stochastic sampling design. *Bernoulli*, 28(2):1250–1275, 2022. page 42

V. Kuznetsov and M. Mohri. Generalization Bounds for Time Series Prediction with Non-stationary Processes. In *Proceedings of ALT'14*, 2014. pages 15, 164

J. Kwon, Y. Zheng, and M. Jun. Flexible spatio-temporal Hawkes process models for earthquake occurrences. *Spatial Statistics*, 54:100728, 2023. pages 23, 118, 120, 121, 122, 123, 172

S. Lahiri. Asymptotic Distribution of the Empirical Spatial Cumulative Distribution Function Predictor and Prediction Bands based on a Subsampling Method. *Probability Theory and Related Fields*, 114(1):55–84, 1999. doi: 10.1007/s004400050221. pages 36, 157

S. N. Lahiri, M. S. Kaiser, N. Cressie, and N.-J. Hsu. Prediction of Spatial Cumulative Distribution Functions using Subsampling. *Journal of the American Statistical Association*, 94(445):86–97, 1999. doi: 10.1080/01621459.1999.10473821. pages 36, 157

C. Lantuéjoul. Ergodicity and integral range. *Journal of Microscopy*, 161(3):387–403, 1991. doi: 10.1111/j.1365-2818.1991.tb03099.x. page 42

P. J. Laub, T. Taimre, and P. K. Pollett. Hawkes processes. *arXiv preprint arXiv:1507.02822*, 2015. page 115

G. Lecué and S. Mendelson. Learning subgaussian classes: Upper and minimax bounds. *arXiv preprint arXiv:1305.4825*, 2013. page 61

E. Lewis and G. Mohler. A nonparametric EM algorithm for multiscale Hawkes processes. *Journal of nonparametric statistics*, 1(1):1–20, 2011. pages 115, 121, 122

P. W. Lewis and G. S. Shedler. Simulation of nonhomogeneous Poisson processes by thinning. *Naval research logistics quarterly*, 26(3):403–413, 1979. page 119

A. Loukas. How Close Are the Eigenvectors of the Sample and Actual Covariance Matrices? In *International Conference on Machine Learning*, pages 2228–2237. PMLR, 2017. page 58

G. Lugosi. Pattern classification and learning theory. In *Principles of nonparametric learning*, pages 1–56. Springer, 2002. pages 59, 60, 62, 63

G. Lugosi and S. Mendelson. Risk Minimization by Median-of-means Tournaments. *arXiv preprint arXiv:1608.00757*, 2016. page 83

B. Marchant and R. Lark. Estimating variogram uncertainty. *Mathematical Geology*, 36:867–898, 01 2004. doi: 10.1023/B:MATG.0000048797.08986.a7. page 58

K. V. Mardia and R. J. Marshall. Maximum Likelihood Estimation of Models for Residual Covariance in Spatial Regression. *Biometrika*, 71(1):135–146, 04 1984. ISSN 0006-3444. doi: 10.1093/biomet/71.1.135. pages 36, 156

D. Marsan and O. Lengline. Extending earthquakes' reach through cascading. *Science*, 319(5866):1076–1079, 2008. pages 118, 122

G. Matheron. *Traité de Géostatistique Appliquée. Tome 1*. Number 14 in Mémoires du BRGM. Tecnip, Paris, 1962. pages 20, 35, 45, 47, 48, 49, 51, 53, 54, 65, 72, 81, 169

G. Matheron. Principles of geostatistics. *Economic Geology*, 58(8):1246–1266, 12 1963. ISSN 0361-0128. doi: 10.2113/gsecongeo.58.8.1246. pages 16, 35, 165

G. Matheron. Les variables régionalisées et leur estimation: une application de la théorie de fonctions aléatoires aux sciences de la nature. *(No Title)*, 1965. pages 16, 57, 165

A. McBratney and R. Webster. The design of optimal sampling schemes for local estimation and mapping of regionalized variables—ii: program and examples. *Computers & Geosciences*, 7(4):335–365, 1981. pages 58, 69

S. Meyer and L. Held. Power-law models for infectious disease spread. *The Annals of Applied Statistics*, 8(3):1612 – 1639, 2014. doi: 10.1214/14-AOAS743. page 111

S. Meyer, J. Elias, and M. Höhle. A space–time conditional intensity model for invasive meningococcal disease occurrence. *Biometrics*, 68(2):607–616, 2012. page 114

S. A. Mingoti and G. Rosa. A note on robust and non-robust variogram estimators. *Rem: Revista Escola de Minas*, 61:87 – 95, 03 2008. ISSN 0370-4467. page 48

G. Mohler. Marked point process hotspot maps for homicide and gun crime prediction in chicago. *International Journal of Forecasting*, 30(3):491–497, 2014. pages 23, 118, 123, 125, 149, 159, 172, 173

G. O. Mohler, M. B. Short, P. J. Brantingham, F. P. Schoenberg, and G. E. Tita. Self-exciting point process modeling of crime. *Journal of the American Statistical Association*, 106(493):100–108, 2011. pages 23, 118, 125, 149, 172

J. Møller and J. G. Rasmussen. Perfect simulation of Hawkes processes. *Advances in applied probability*, 37(3):629–646, 2005. pages 120, 140

W. G. Müller and D. Zimmerman. Optimal Designs for Variogram Estimation. *Environmetrics*, 10:23–37, 1999. page 156

F. Musmeci and D. Vere-Jones. A space-time clustering model for historical earthquakes. *Annals of the Institute of Statistical Mathematics*, 44:1–11, 1992. pages 23, 118, 124, 146, 172

S. Müller and L. Schüler. Geostat-framework/gstools. zenodo., 2020. pages 54, 75

S. Nadarajah and T. Pogány. On the distribution of the product of correlated normal random variables. *Comptes Rendus Mathematique*, 354, 10 2015. doi: 10.1016/j.crma.2015.10.019. pages 44, 73

Y. Ogata. Statistical models for earthquake occurrences and residual analysis for point processes. *Journal of the American Statistical association*, 83(401):9–27, 1988. pages 17, 22, 111, 113, 115, 116, 118, 123, 146, 166, 172

Y. Ogata. Space-time point-process models for earthquake occurrences. *Annals of the Institute of Statistical Mathematics*, 50:379–402, 1998. pages 23, 119, 123, 124, 146, 172

Y. Ogata. *Seismicity Analysis through Point-process Modeling: A Review*, pages 471–507. Birkhäuser Basel, Basel, 1999. ISBN 978-3-0348-8677-2. doi: 10.1007/978-3-0348-8677-2_14. pages 111, 146

E. Pardo-Igúzquiza and R. A. Olea. Varboot: A spatial bootstrap program for semivariogram uncertainty assessment. *Computers & Geosciences*, 41:188–198, 2012. page 53

H. Putter and A. G. Young. On the Effect of Covariance Function Estimation on the Accuracy of Kriging Predictors. *Bernoulli*, 7(3):421–438, 06 2001. pages 36, 57, 157

H. Qiao, M. C. Hucumenoglu, and P. Pal. Compressive Kriging Using Multi-Dimensional Generalized Nested Sampling. In *2018 52nd Asilomar Conference on Signals, Systems, and Computers*, pages 84–88, 2018. doi: 10.1109/ACSSC.2018.8645258. page 84

S. Rambhatla, S. Zeighami, K. Shahabi, C. Shahabi, and Y. Liu. Toward accurate spatiotemporal covid-19 risk scores using high-resolution real-world mobility data. *ACM Transactions on Spatial Algorithms and Systems (TSAS)*, 8(2):1–30, 2022. pages 23, 125, 172

J. H. Ratcliffe and G. F. Rengert. Near-repeat patterns in philadelphia shootings. *Security Journal*, 21:58–76, 2008. page 125

A. Reinhart. A review of self-exciting spatio-temporal point processes and their applications. *Statistical Science*, 33(3):299–318, 2018. pages 22, 110, 116, 118, 120, 121, 122, 123, 172

P. Reynaud-Bouret and V. Rivoirard. Near optimal thresholding estimation of a poisson intensity on the real line. *Electronic journal of statistics*, 4:172–238, 2010. pages 122, 135

P. Reynaud-Bouret, V. Rivoirard, F. Grammont, and C. Tuleau-Malot. Goodness-of-fit tests and nonparametric adaptive estimation for spike train analysis. *The Journal of Mathematical Neuroscience*, 4(1):1–41, 2014. pages 122, 135

B. D. Ripley. *Spatial statistics*. John Wiley & Sons, 2005. pages 16, 166

G. Robinson. A role for variograms. *Australian Journal of Statistics*, 32(3):327–335, 1990. doi: https://doi.org/10.1111/j.1467-842X.1990.tb01028.x. page 44

P. J. Rousseeuw and C. Croux. Explicit scale estimators with high breakdown point. *L1-Statistical analysis and related methods*, 1:77–92, 1992. page 48

M. Rudelson. Random vectors in the isotropic position. *Journal of Functional Analysis*, 164(1):60–72, 1999. page 58

D. Russo. Design of an optimal sampling network for estimating the variogram. *Soil Science Society of America Journal*, 48(4):708–716, 1984. doi: https://doi.org/10.2136/sssaj1984.03615995004800040003x. pages 58, 69

P. D. Sampson and P. Guttorp. Nonparametric estimation of nonstationary spatial covariance structure. *Journal of the American Statistical Association*, 87(417):108–119, 1992. page 42

F. P. Schoenberg. Multidimensional residual analysis of point process models for earthquake occurrences. *Journal of the American Statistical Association*, 98(464):789–795, 2003. page 146

F. P. Schoenberg. Testing separability in spatial-temporal marked point processes. *Biometrics*, 60(2):471–481, 2004. ISSN 0006341X, 15410420. pages 118, 123, 160

F. P. Schoenberg, M. Hoffmann, and R. J. Harrigan. A recursive point process model for infectious diseases. *Annals of the Institute of Statistical Mathematics*, 71:1271–1287, 2019. page 114

O. Shchur, A. C. Türkmen, T. Januschowski, and S. Günnemann. Neural temporal point processes: A review. *arXiv preprint arXiv:2104.03528*, 2021. page 148

H. Sheen, X. Zhu, and Y. Xie. Tensor kernel recovery for discrete spatio-temporal Hawkes processes. *IEEE Transactions on Signal Processing*, 70:5859–5870, 2022. page 134

M. Sherman and E. Carlstein. Nonparametric Estimation of the Moments of a General Statistic Computed from Spatial Data. *Journal of the American Statistical Association*, 89(426):496–500, 1994. doi: 10.1080/01621459.1994.10476773. pages 36, 156

L. Shi. Bounds on the (Laplacian) Spectral Radius of Graphs. *Linear algebra and its applications*, 422(2-3):755–770, 2007. page 187

E. Siviero, E. Chautru, and S. Clémençon. A statistical learning view of simple Kriging. *TEST*, 33(1):271–296, 2024a. pages 29, 180

E. Siviero, G. Staerman, S. Clémençon, and T. Moreau. Flexible parametric inference for space-time Hawkes processes. *arXiv preprint arXiv:2406.06849*, 2024b. pages 29, 180

D. A. Spielman. Spectral Graph Theory, 2012. Lecture 3, The Adjacency Matrix and the nth Eigenvalue. page 184

G. Staerman, P. Mozharovskyi, and S. Clémençon. Affine-Invariant Integrated Rank-Weighted Depth: Definition, Properties and Finite Sample Analysis. *arXiv preprint arXiv:2106.11068*, 2021. page 193

G. Staerman, C. Allain, A. Gramfort, and T. Moreau. Fadin: Fast discretized inference for Hawkes processes with general parametric kernels. In *International Conference on Machine Learning*, pages 32575–32597. PMLR, 2023. pages 24, 27, 110, 123, 124, 126, 127, 128, 129, 132, 133, 139, 159, 174, 177

M. L. Stein. Asymptotically efficient prediction of a random field with a misspecified covariance function. *The Annals of Statistics*, 16(1):55–63, 1988. page 36

M. L. Stein. Fixed-domain asymptotics for spatial periodograms. *Journal of the American Statistical Association*, 90(432):1277–1288, 1995. page 36

M. L. Stein. *Interpolation of Spatial Data*. Springer Series in Statistics. Springer-Verlag, New York, 1999. ISBN 0-387-98629-4. doi: 10.1007/978-1-4612-1494-6. Some theory for Kriging. pages 20, 37, 43, 56, 57, 64, 68, 69, 169, 184

I. Steinwart and A. Christmann. *Support Vector Machines*. Springer Science & Business Media, New York, NY, 2008. pages 15, 62, 164

I. Steinwart and A. Christmann. Fast Learning from non-i.i.d. Observations. *Advances in Neural Information Processing Systems*, 22:1768–1776, 2009. pages 15, 164

I. Steinwart, D. Hush, and C. Scovel. Learning from Dependent Observations. *Journal of Multivariate Analysis*, 100(1):175–194, 2009. pages 15, 164

A. Tikhonov. On the stability of inverse problems. *Dokl. Akad. Nauk SSSR*, 39(5): 195–198, 1943. page 85

E. Todini. Influence of parameter estimation uncertainty in Kriging. *Hydrology and Earth System Sciences*, 5, 06 2001. doi: 10.5194/hess-5-215-2001. page 57

A. Veen and F. P. Schoenberg. Estimation of space–time branching process models in seismology using an em–type algorithm. *Journal of the American Statistical Association*, 103(482):614–624, 2008. pages 122, 146

D. Vere-Jones. Stochastic models for earthquake occurrence. *Journal of the Royal Statistical Society: Series B (Methodological)*, 32(1):1–45, 1970. pages 22, 111, 172

D. Vere-Jones. Forecasting earthquakes and earthquake risk. *International Journal of Forecasting*, 11(4):503–538, 1995. page 146

D. Vere-Jones. Some models and procedures for space-time point processes. *Environmental and Ecological Statistics*, 16:173–195, 2009. page 110

R. Vershynin. How Close is the Sample Covariance Matrix to the Actual Covariance Matrix? *Journal of Theoretical Probability*, 25(3):655–686, 2012. page 58

B. Wang and F. Zhang. Some Inequalities for the Eigenvalues of the Product of Positive Semidefinite Hermitian Matrices. *Linear algebra and its applications*, 160:113–118, 1992. doi: 10.1016/0024-3795(92)90442-D. pages 184, 190

L. Wang and T. Ma. Tail Bounds for Sum of Gamma Variables and Related Inferences. *Communications in Statistics - Theory and Methods*, 0(0):1–10, 2020. doi: 10.1080/03610926.2020.1756329. pages 25, 63, 67, 73, 74, 175, 189

W. Wang, R. Tuo, and C. Jeff Wu. On prediction properties of Kriging: Uniform error bounds and robustness. *Journal of the American Statistical Association*, 115(530): 920–930, 2020. pages 58, 69, 156

P.-Å. Wedin. Perturbation Theory for Pseudo-Inverses. *BIT Numerical Mathematics*, 13 (2):217–232, 1973. pages 86, 193

J. Worrall, B. Raiha, W. Paul, and M. Kerrie. Fifty years later: new directions in Hawkes processes. *SORT-Statistics and Operations Research Transactions*, 46(1):3–38, Jun. 2022. doi: 10.2436/20.8080.02.116. page 110

B.-Y. Xi and F. Zhang. Inequalities for Selected Eigenvalues of the Product of Matrices. *arXiv preprint arXiv:1905.03821*, 2019. doi: 10.1090/proc/14529. pages 184, 190

A. M. Yaglom. *Correlation Theory of Stationary and Related Random Functions, Volume I: Basic Results*, volume 131. Springer, 1987. pages 43, 50

S. Yakowitz and F. Szidarovszky. A comparison of Kriging with nonparametric regression methods. *Journal of Multivariate Analysis*, 16(1):21–53, 1985. page 36

E. A. Yfantis, G. T. Flatman, and J. V. Behar. Efficiency of Kriging estimation for square, triangular, and hexagonal grids. *Mathematical Geology*, 19(3):183–205, 1987. page 58

B. Yuan, H. Li, A. L. Bertozzi, P. J. Brantingham, and M. A. Porter. Multivariate spatiotemporal Hawkes processes and network reconstruction. *SIAM Journal on Mathematics of Data Science*, 1(2):356–382, 2019. pages 23, 118, 120, 121, 123, 173

B. Yuan, F. P. Schoenberg, and A. L. Bertozzi. Fast estimation of multivariate spatiotemporal Hawkes processes and network reconstruction. *Annals of the Institute of Statistical Mathematics*, pages 1–26, 2021. page 118

S. Zhu and Y. Xie. Spatiotemporal-textual point processes for crime linkage detection. *The Annals of Applied Statistics*, 16(2):1151–1170, 2022. pages 23, 118, 125, 149, 172

S. Zhu, S. Li, Z. Peng, and Y. Xie. Imitation learning of neural spatio-temporal point processes. *IEEE Transactions on Knowledge and Data Engineering*, 34(11):5391–5402, 2021. page 146

J. Zhuang. Next-day earthquake forecasts for the japan region generated by the etas model. *Earth, planets and space*, 63:207–216, 2011. page 146

J. Zhuang and J. Mateu. A semiparametric spatiotemporal Hawkes-type point process model with periodic background for crime data. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 182(3):919–942, 2019. pages 118, 159

J. Zhuang, Y. Ogata, and D. Vere-Jones. Stochastic declustering of space-time earthquake occurrences. *Journal of the American Statistical Association*, 97(458):369–380, 2002. page 122

J. Zhuang, Y. Ogata, and D. Vere-Jones. Analyzing earthquake clustering features by using stochastic reconstruction. *Journal of Geophysical Research: Solid Earth*, 109 (B5), 2004. page 120

D. Zimmerman and N. Cressie. Mean Squared Prediction Error in the Spatial Linear Model with Estimated Covariance Parameters. *Annals of the Institute of Statistical Mathematics*, 44:27–43, 02 1992. doi: 10.1007/BF00048668. pages 20, 169

D. L. Zimmerman. Computationally Exploitable Structure of Covariance Matrices and Generalized Covariance Matrices in Spatial Models. *Journal of Statistical Computation and Simulation*, 32(1-2):1–15, 1989. doi: 10.1080/00949658908811149. pages 20, 169

ECOLE
DOCTORALE
DE MATHEMATIQUES
HADAMARD

Titre : Apprentissage Statistique pour les données Spatiales: théorie et algorithmes

Mots clés : apprentissage statistique, données spatiales, structure de dépendance, krigeage, fonction de covariance

Résumé : À l'époque des grandes données, l'accès à des ensembles de données massives, présentant une structure de dépendance spatiale possiblement complexe, augmente de plus en plus. Dans cette thèse, notre objectif est de surmonter les enjeux liés à la structure de dépendance des données spatiales (et spatio-temporelles).

En un premier temps, nous analysons le *Krigeage simple*, problème clé en Géostatistique, en adoptant le point de vue de l'apprentissage statistique, *i.e.* en effectuant une analyse prédictive non paramétrique à partir d'un échantillon fini. Dans ce contexte, la théorie probabiliste standard de l'apprentissage statistique ne s'applique pas directement. De nouvelles garanties sur la capacité de généralisation du prédicteur par Krigeage doivent être établies. Étant donné une réalisation d'un champ aléatoire de covariance inconnue, observé en un nombre fini de sites du domaine spatial, l'objectif est de prédire les valeurs inconnues du champ aléatoire à n'importe quel point du domaine, tout en minimisant le risque quadratique. En raison du caractère non indépendant et non identiquement distribué des données d'apprentissage, déterminer la capacité de généralisation des minimiseurs de risque empiriques est un défi complexe. Dans la première partie de cette thèse, nous présentons des bornes non asymptotiques pour l'excès de risque d'une règle prédictive *plug-in* imitant le vrai minimiseur. Ces bornes sont établies pour des processus gaussiens stationnaires avec une fonction de covariance isotrope, observés lors de la phase d'apprentissage à des emplacements formant une grille régulière. Nos résultats théoriques, ainsi que le rôle joué par les conditions techniques requises pour les définir, sont illustrés par diverses expériences numériques, sur des données simulées ainsi que sur des données réelles, et ouvrent, nous l'espérons, la voie à de nouveaux développements dans l'apprentissage statistique basé sur des données spatiales.

En un second temps, nous nous concentrons sur les processus de Hawkes spatio-temporels. De nombreux ensembles de données spatio-temporelles, en sociologie, épidémiologie ou sismologie, par exemple, présentent des caractéristiques d'auto-excitation: les événements ont tendance à se regrouper ou à déclencher une série d'événements successifs, ou encore les deux à la fois. Dans ce contexte, les processus de Hawkes spatio-temporels se révèlent être un outil puissant grâce à leur capacité à capturer ces comportements avec précision. Cependant, traiter efficacement le grand volume de données actuellement disponible s'avère difficile. La deuxième partie de cette thèse vise à développer une technique d'inférence paramétrique rapide et flexible pour obtenir les paramètres des fonctions noyaux impliquées dans la fonction d'intensité d'un processus de Hawkes spatio-temporel. Notre approche statistique combine trois ingrédients clés : (1) nous considérons des fonctions noyaux à support, (2) le domaine spatio-temporel est discrétisé de manière appropriée, et (3) des calculs préalables (approximatifs) sont utilisés. La technique d'inférence que nous proposons consiste en un solveur rapide et statistiquement précis. En complément de la description des aspects algorithmiques, des expériences numériques ont été menées sur des données spatio-temporelles, tant synthétiques que réelles, apportant des preuves empiriques solides de la pertinence de la méthodologie proposée.

Title : Statistical Learning for Spatial data: theory and algorithms

Keywords : statistical learning, spatial data, dependence structure, kriging, covariance function

Abstract : In the Big Data era, massive datasets exhibiting a possibly complex spatial dependence structure are becoming increasingly available. In this thesis, we aim at developing approaches to efficiently exploit the dependence structure of spatial (and spatio-temporal) data.

We first analyze the *simple Kriging* task, the flagship problem in Geostatistics, from a statistical learning perspective, *i.e.* by carrying out a non-parametric finite-sample predictive analysis. In this context, the standard probabilistic theory of statistical learning does not apply directly and theoretical guarantees of the generalization capacity of the Kriging predictive rule learned from spatial data are left to be established. Given a finite number of values taken by a realization of a square integrable random field, with unknown covariance structure, the goal is to predict the unknown values that the random field takes at any other location in the spatial domain with minimum quadratic risk. Establishing the generalization capacity of empirical risk minimizer is far from straightforward, due to the non independent and identically distributed nature of the training data involved in the learning procedure. In the first part of this thesis, non-asymptotic bounds are proved for the excess risk of a *plug-in* predictive rule mimicking the true minimizer in the case of isotropic stationary Gaussian processes, observed at locations forming a regular grid in the learning stage. These theoretical results, as well as the role played by the technical conditions required to establish them, are illustrated by various numerical experiments, on simulated data and on real-world datasets, and may hopefully pave the way for further developments in statistical learning based on spatial data.

In the second part of this thesis, we focus on space-time Hawkes processes. Many modern spatio-temporal data sets, in sociology, epidemiology or seismology, for example, exhibit self-exciting characteristics, with simultaneous triggering and clustering behaviors, that a suitable spatio-temporal Hawkes process can accurately capture. However, dealing efficiently with the high volumes of data now available is challenging. We aim at developing a fast and flexible parametric inference technique to recover the parameters of the kernel functions involved in the intensity function of a spatio-temporal Hawkes process based on such data. Our statistical approach combines three key ingredients: (1) kernels with finite support are considered, (2) the space-time domain is appropriately discretized, and (3) (approximate) precomputations are used. The inference technique we propose consists of a fast and statistically accurate solver. In addition to describing the algorithmic aspects, numerical experiments have been carried out on synthetic and real spatio-temporal data, providing solid empirical evidence of the relevance of the proposed methodology.

Institut Polytechnique de Paris
91120 Palaiseau, France