UNIVERSITÀ
DEGLI STUDI
DI PADOVA

Sede Amministrativa: Università degli Studi di Padova

Dipartimento di Scienze Statistiche
Corso di Dottorato di Ricerca in Scienze Statistiche
Ciclo XXXII

Likelihood approximation and prediction for large spatial and spatio-temporal datasets using $\mathcal{H}$-matrix approach

Coordinatore del Corso: Prof. Massimiliano Caporin

Supervisore: Prof. Carlo Gaetan

Dottorando: Anastasiia Gorshechnikova

30.09.2019

Abstract

The Gaussian distribution is the most fundamental distribution in statistics. However, many applications of Gaussian random fields (GRFs) are limited by the computational complexity associated to the evaluation of probability density functions. Particularly, large datasets with N irregularly sited spatial (or spatio-temporal) locations are difficult to handle for several applications of GRF such as maximum likelihood estimation (MLE) and kriging prediction. This is due to the fact that computation of the inverse of the dense covariance function requires a computational complexity of $O(N^3)$ floating points operations in spatial or spatio-temporal context. For relatively large N the exact computation becomes infeasible and alternative methods are necessary. Several approaches have been proposed to tackle this problem. Most assume a specific form for the spatial(-temporal) covariance function and use different methods to approximate the resulting covariance matrix. We aim at approximating covariance functions in a format that facilitates the computation of MLE and kriging prediction with very large spatial and spatio-temporal datasets.

For a sufficiently general class of spatial and specific class of spatio-temporal covariance functions, a methodology is developed using a hierarchical matrix approach. Since this method was originally created for the approximation of dense matrices coming from partial differential and integral equations, a theoretical framework is formulated in terms of Stochastic Partial Differential equations (SPDEs). The application of this technique is detailed for covariance functions of GRFs obtained as solutions to SPDEs. The approximation of the covariance matrix in such a low-rank format allows for computation of the matrix-vector products and matrix factorisations in a log-linear computational cost followed by an efficient MLE and kriging prediction. The numerical studies are provided for based on spatial and spatio-temporal datasets and the $\mathcal{H}$-matrix approach is compared with the other methods in terms of computational and statistical efficiency.

Sommario

Tra tutte le distribuzioni probabilistiche in statistica, quella Gaussiana indubbiamente fondamentale. La situazione non cos rosea quando invece si parla di random fields Gaussiani (GRFs), perch la stima della densit ha costi computazionali abbastanza elevati. In particolare, grandi dataset contenenti N posizioni spazio-temporali disposte in maniera irregolare sono molto difficili da trattare in parecchie applicazioni, quali la stima di massima verosimiglianza o la predizione kriging. Questo dovuto al fatto che calcolare l'inversa della matrice di varianza-covarianza richiede una complessit computazionale pari a $O(N^3)$ punti casuali, il che molto dispendioso in un contesto spaziale e spazio-temporale. Per N sufficientemente grande, il calcolo esatto diventa improponibile, rendendo necessari metodi alternativi. Diversi approcci sono stati proposti per ovviare a questo problema. La maggior parte delle soluzioni proposte assume una forma specifica per la funzione di covarianza spaziale (spazio-temporale) e usa metodi differenti per approssimare la matrice di covarianza risultante. Il nostro obiettivo di approssimare le funzioni di covarianza in un formato che faciliti il calcolo della stima di massima verosimiglianza e della predizione kriging in caso di dataset spaziali e spazio-temporali molto grandi.

Per classi sufficientemente generali di funzioni di covarianza spaziali, e per specifiche classi di funzioni di covarianza spazio-temporali, la metodologia proposta si basa sull'uso di matrici gerarchiche. Poich tale metodo fu originariamente sviluppato per matrici dense provenienti da equazioni differenziali e integrali, abbiamo sviluppato un'appropriata struttura teorica per il nostro obiettivo, denominata Stochastic Partial Differential equations (SPDEs), e ci siamo proposti di provvedere un'applicazione di tale approccio, ottenendo funzioni di covarianza di GRFs come soluzioni del nostro metodo. L'approssimazione della matrice di covarianza di rango inferiore permette di calcolare il prodotto matriciale e la sua fattorizzazione con un costo log-lineare, portando a efficienti stime di massima verosimiglianza e predizioni. Presentiamo studi empirici sia su simulazioni spaziali e spazio-temporali, sia un'applicazione su dati reali, e confrontiamo in termini di efficienza statistica e costo computazionale il nostro approccio con altri metodi presenti in letteratura.

To my family

Acknowledgements

First of all, I would like to express my sincere gratitude to the Cariparo Foundation for offering me the opportunity to carry out the research in Padova, financial support of my stay in Italy and one of the most beautiful experiences that I have ever had in my life.

My sincere thanks goes to my supervisor Prof. Carlo Gaetan for the continuous support of my research, for his patience and enthusiasm while working with me, for his positive attitude and passion for our joint work that inspired me to go forward and never give up. His guidance helped me a lot to gain a sufficient knowledge in the subject of this thesis and write the thesis. I could not have imagined having a better Professor for my PhD study.

I would also like to express my gratitude to my family and friends for spiritual support throughout writing this thesis and my life in Italy in general.

Besides mentioned above, my special gratitude goes to the professoriate of the Department of Statistical Sciences of Padova for the valuable and deep knowledge in statistics provided during the first year that allowed to proceed with the research. I thank my group mates for the joint work and restless nights before deadlines during the first year, for the inspiring discussions during these three years and for all the light-hearted pleasure we have had. Particularly, I would like to thank my colleagues and friends Huiting Huang and Ilaria Bussoli for our collaboration and amazing friendship throughout the years.

Nevertheless, I am also grateful to the Secretariat of PhD Course in Statistics, particularly to Patrizia Piacentini, for the the ability to easily arrange and support all my research activities abroad, participation in the conferences, workshops and schools.

In addition, I would like to thank Prof. Steffen Börm for the valuable knowledge in the subject of this work given during the attendance of the Winter School in Kiel.

Last but not the least, I am grateful to all the members of my committee for their patience, valuable comments and questions with respect to my work.

Contents

List of Figures

List of Tables

Introduction

”It’s better to solve the right problem approximately than to solve the wrong problem exactly”

John W.Tukey

Overview

The advancement of technology is generating a growing availability of observations of diverse nature which attracts the attention of data analysts. The amount of data produced every day from various sources is enormous which opens the door to a wide variety of problems. Large data sets are also common in environmental sciences where data are often observed at a large number of spatial locations and at different temporal intervals. This includes: weather forecasts, wind speed, atmospheric carbon dioxide measurements and many others. The amount of available data has grown such that standard computer memory is unable to manage such large volumes. In association to this amount of data, computational and modelling challenges arise which was labeled by Banerjee *et al.* (2008) as ”big N problem”. Due to this problem, standard approaches in statistics become infeasible for the large datasets.

Compared with other commonly-used distributions, the Gaussian distribution results in fitting data with the flexible shape controlled by the mean and variance. Because of its flexibility and ease of implementation in different computer environments, it has many important applications in mathematics, computer science and the natural sciences. However, as was noted by Stein (2008), the exact computation of the likelihood of a Gaussian Random Field observed at N irregularly sited locations generally requires $O(N^3)$ floating point operations and $O(N^2)$ memory. For example, while a sample size of 10^3 is no longer a challenge, whereas a size of 10^4 is already out of reach with classical procedures.

The problem is further aggravated in case of inference with respect to spatio-temporal random effects due to the presence of the additional time dimension. New algorithms are required for spatio-temporal statistical modelling and inference. They have a large number of practical applications in studying phenomena that occur in both space and time.

Therefore, nowadays this computational problem is at the center of attention in the spatial and spatio-temporal statistical communities and alternative methods are required to make a statistical inference. In general two possible strategies exist: simplification of the model and simplification of the fitting method. The first group of methods includes: 1) work of Lindgren *et al.* (2011) that applied a Gaussian Markov Random Field (GMRF) to the GRF obtained as a solution to a Stochastic Partial Differential Equation (SPDE), 2) Cressie and Johannesson (2008) with the low-rank approximation of the GRF after preliminary chosen basis functions, 3) Kaufman *et al.* (2008) with tapering the covariance function with sparse correlation matrix obtaining a positive definite function with compact support. The second group deals with the approximation of the likelihood by pseudolikelihoods as in Lindsay (1988), Eidsvik *et al.* (2014) and others. All the aforementioned approaches require a computational complexity from $O(N \log N)$ to $O(N^2)$ number of floating point operations. Additionally, some sacrifice the accuracy of the statistical inference.

In this thesis we present an approach based on the approximation of covariance functions by hierarchical matrices (or shortly $\mathcal{H}$-matrices). This method involves the partitioning of a matrix into sub-blocks according to a binary cluster tree and specific conditions. Some blocks of the matrix are then further approximated by low-rank matrices with the rank or error of the approximation chosen beforehand. The low-rank structure of the blocks of the approximated covariance matrix results in a reduced cost for many matrix operations, such as matrix-vector multiplications and matrix inversions. This means that a log-linear computational cost of the maximum likelihood estimation (MLE) and kriging prediction can easily be obtained. In addition, we may easily find a trade-off between the computational cost of the procedure and statistical efficiency based on the chosen error of approximation in the low-rank blocks. However, errors in the $\mathcal{H}$-matrix approximation may destroy the symmetry and positive definiteness properties of the symmetric positive definite covariance matrix. Throughout this thesis we also address this problem.

The objectives of this thesis also include a review of the most celebrated existing methods in the literature providing their pros and cons. Focusing on the numerical analysis, the method of $\mathcal{H}$-matrix was exploited by Litvinenko *et al.* (2019) for MLE estimation. We extend his work in terms of providing a general theoretical framework for the application of this method in spatial statistics, deriving new regularity conditions, performing kriging prediction and comparing this technique with other existing methods. Moreover, we extend this method to the application in the spatio-temporal context.

Since the $\mathcal{H}$-matrix method originally was considered for the approximation of dense matrices obtained from the discretization of Partial Differential Equations (PDEs), we formulate our main theory in terms of Stochastic Partial Differential Equations (SPDEs). We further study the impact of the range of the covariance function on the approximation by the $\mathcal{H}$-methods. We determine the relation of $\mathcal{H}$-matrix to physically driven PDEs and how it can be adapted to the stochastic framework. Once the link is established and the conditions of the existence and uniqueness of solutions to some SPDEs are obtained, we obtain the connection between a slow-growing spectral measure and covariance regularity condition. We apply this method to the covariance function of GRFs which are derived from SPDEs and evaluate the possibility to extend to a spatio-temporal SPDEs. Lastly, we derive the regularity conditions for other types of spatial and spatio-temporal covariance functions.

In the application part we provide numerical studies with simulated and real data application for spatial and spatio-temporal datasets. We compare the $\mathcal{H}$-matrix approach with other methods such as covariance tapering, composite likelihoods and fixed rank kriging in terms of computational and statistical efficiencies.

Main contributions of the thesis

The main contributions of this thesis can be summarised as follows

- explore the existing methods addressed to tackle the "big N" computational problem for statistical inference, briefly summarising the advantages and disadvantages of all these methods;

- examine and provide to the readers a complete picture of the $\mathcal{H}$-matrix method;

- formulate a main theoretical framework in terms of deterministic PDEs and stochastic PDEs with the derivation of the required regularity conditions for $\mathcal{H}$-matrices;

- extend the approach for the application to a different class of spatial or spatio-temporal covariance functions not related to SPDEs;

- adjust the existing regularity condition of the $\mathcal{H}$-matrix approach for a range;

- obtain the asymptotical properties of the approximate with the $\mathcal{H}$-matrix estimators;
- with the numerical studies on the simulated and real datasets, perform likelihood estimation and kriging predictions with the use of $\mathcal{H}$-matrices and provide the comparison with other methods, such as fixed rank kriging (FRK) and covariance tapering in both spatial and spatio-temporal contexts;

The original idea of this thesis is to reduce the computational cost of the MLE estimation and kriging prediction through approximating specifically chosen blocks of covariance functions in a low-rank format. The dissertation consists of five chapters and starts with stating the problem in Chapter 1. The major part of Chapter 1 focuses on the overview of the most known methods in both spatial and spatio-temporal context. The existing methods are divided into two categories: simplification of the model and simplification of the fitting method. The first group of methods consists of Fixed Rank Kriging approach developed by Cressie and Johannesson (2008) and Banerjee *et al.* (2008), GMRFs proposed by Lindgren *et al.* (2011) and covariance tapering by Kaufman *et al.* (2008). As part of the second group we consider pseudolikelihoods described by Lindsay (1988), Eidsvik *et al.* (2014) and reviewed in details by Varin *et al.* (2011). We focus on the advantages and disadvantages of each of the methods since some are used for comparison purposes in the application part of this thesis.

The theoretical formulation of the hierarchical matrices approach, exploited in the thesis as a tool for the fast approximation of the covariance matrices, is detailed in Chapter 2. We initiate the Chapter by formulating the main idea of the $\mathcal{H}$-matrix approach in term of matrix compressions. All the main steps of the implementation of this technique are given throughout Chapter 2. It includes the construction of the binary cluster tree and block cluster tree which lead to the discrete structures of the approximated matrix. In addition, we provide details of partition of spatial or spatio-temporal points which is based on the construction of the bounding boxes around the points, clustering technique and on the estimation of the distances between the boxes. Apart from the chosen clustering technique, it is required to specify the conditions for the appropriate partition of the points. This gives rise to the admissibility and asymptotic smoothness conditions which are also given in Chapter 2. We end this Chapter by reviewing some works of the statistics community that exploited this approach to speed up computations.

Chapter 3 provides a theoretical formulation of covariances functions arising from SPDEs and the regularity conditions which are to be satisfied for their approximation by $\mathcal{H}$-matrices. We firstly introduce the main deterministic mathematical tools which are needed to comprehend the link between $\mathcal{H}$-matrices and the SPDE approach. It consists of introducing the distribution theory, Schwartz space, tempered distributions and pseudodifferential equations. We give the definition of slow-growing measures and exploit the results obtained by Vergara *et al.* (2018) of the existence and uniqueness of solutions to some SPDEs. We after present the framework of Generalized Random Distributions and relate the theory of PDEs to geostatistics. We introduce the stochastic version of the deterministic tools in the context of the mean-square theory, where the main characteristics are defined by the mean and covariance structures. We also demonstrate how the theory of Generalized Random Fields (GeRF) can be described within the pseudodifferential operators and their kernels which are related to covariance functions. We show how a slow-growing measure relates to the covariance regularity which is required for the application of $\mathcal{H}$-matrices. We also discuss some spatio-temporal covariance functions obtained within the SPDE approach.

Chapter 4 describes maximum likelihood estimation with the $\mathcal{H}$-matrices in the same way as discussed in Litvinenko *et al.* (2019). Along with the maximum likelihood estimation, Chapter considers kriging prediction which was not exposed in Litvinenko *et al.* (2019). The asymptotical properties of the approximate with the $\mathcal{H}$-matrix estimators are also derived in this Chapter. We conclude Chapter with the reformulation of the main condition required for the application of $\mathcal{H}$-matrices in spatial framework which was not discussed before in the literature.

The results of simulation studies with different sample sizes are given in Chapter 5. We perform the analysis of $\mathcal{H}$-matrices approach are compare it with the covariance tapering in terms of computational and statistical efficiencies due to the similar concept of both approaches. Moreover, we implement the $\mathcal{H}$-matrix method for another class of spatio-temporal covariance functions which is not related to SPDEs. We conclude Chapter by performing a real spatial data application and comparing performance with covariance tapering and fixed rank kriging. The main conclusions of this thesis are discussed in the last part of this work, summarizing the obtained results and presenting possible future directions of the research.

Chapter 1

Literature review

1.1 Statement of the problem

Consider N observations $\boldsymbol{Z} = (Z(x_1), \ldots, Z(x_N))^T$ from a Gaussian Random Field (GRF) $\{Z(x)\}$ defined over a domain indexed by x, where x denotes either a spatial $x : s \in \mathbb{R}^d$ or spatio-temporal domain of observations $x : (s, t) \in \mathbb{R}^d \times \mathbb{R}$.

From a mathematical point of view, it is correct to consider a point in the $\mathbb{R}^{d+1}$ dimension. In spatial statistics we keep the physicists viewpoint and do not consider spatial and temporal dimension in the same way due to the major differences between the spatial and temporal coordinates. Namely, that the time axis is ordered compared to the spatial one. Therefore, an observation is considered as a point on $\mathbb{R}^d \times \mathbb{R}$, where $\mathbb{R}^d$ is the d-dimensional Euclidean space and $\mathbb{R}$ is the time dimension.

In what follows, we assume that the spatial (spatio-temporal) process $Z(x)$ satisfies the regularity condition, $\mathrm{Var}(Z(x)) < \infty$ for all $x \in \mathbb{R}^d \times \mathbb{R}$. Then we can define the mean function as $\mu(x) \equiv \mathbb{E}(Z(x))$ and the covariance function as $c(x_i, x_j) \equiv \mathrm{cov}(Z(x_i), Z(x_j))$ for $i, j = 1, \ldots, N$.

The covariance function $c : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ must be a positive-definite function, such that for any $x_1, \ldots, x_N$, any real weights $\lambda_1, \ldots, \lambda_N$ and any positive integer N, the covariance must satisfy

$$\sum_{j=1}^{N} \sum_{i=1}^{N} \lambda_i \lambda_j c(x_i, x_j) \geq 0$$

for certain functions c.

In the rest of this work, it is assumed that $Z(x)$ is a second-order stationary, i.e. it has a constant mean and its covariance function depends only on the gap $x_i - x_j$ of the variables, namely $\mathrm{cov}(Z(x_i), Z(x_j)) = c(x_i - x_j)$ for $i, j = 1, \ldots, N$. Since in this work we mainly deal with the computational challenges arising when inverting the covariance matrices, without loss of generality we consider a zero-mean GRF (or a detrended process).

To ensure positive definiteness, one often specifies the covariance function c to belong to a parametric family with positive definite members. That is

$$\operatorname{cov}(Z(x), Z(x+a)) = c(a),$$

where $a = (h, u) \in \mathbb{R}^d \times \mathbb{R}$ denotes the spatio-temporal lag.

Consider the stationary spatio-temporal covariance function c. Assume that c is continuous and that its spectral distribution function possesses a positive spectral density. The following statements are equivalent:

- $c : \mathbb{R}^d \times \mathbb{R} \to \mathbb{R}$ is a positive definite function;
- According to Bochner's theorem, c is the Fourier transform of a symmetric non-negative measure μ on $\mathbb{R}^d \times \mathbb{R}$, that is

$$c(a) = \frac{1}{(2\pi)^{d/2}} \int e^{-ia^T\xi} d\mu(\xi), \quad \text{for } \forall a \in \mathbb{R}^d \times \mathbb{R} \tag{1.1}$$

Any continuous covariance function admits the spectral representation (1.1).

Considering parametric covariance function with the vector of the unknown p- dimensional parameters $\theta \in \Theta \subseteq \mathbb{R}^p$, the covariance function $c(x) := c(x; \theta)$ depends on unknown parameter θ. We make a statistical inference with respect to θ based on the Gaussian log-likelihood which can be written as follows

$$L(\theta) = -\frac{N}{2} \log 2\pi - \frac{1}{2} \log \det(C_Z) - \frac{1}{2} \boldsymbol{Z}^\top C_Z^{-1} \boldsymbol{Z}, \tag{1.2}$$

where N is the sample size, C_Z is the covariance matrix of $\boldsymbol{Z}$ and $\det(C_Z)$ is the determinant of C_Z.

As can be seen from (1.2), to make an inference regarding the unknown parameter θ the exact computation of the log-likelihood requires a computation of $\det(C_Z)$ as well as the inverse of the covariance matrix C_Z^{-1} which both require $O(N^3)$ operations. See Figure 1.1 as an example of a large spatial dataset with the size of over 2000000 atmospheric carbon dioxide measurements, ppm, (taken from the AIRS dataset[1]) collected on the globe.

[1] available at *https://cran.r-project.org/web/packages/FRK/index.html*

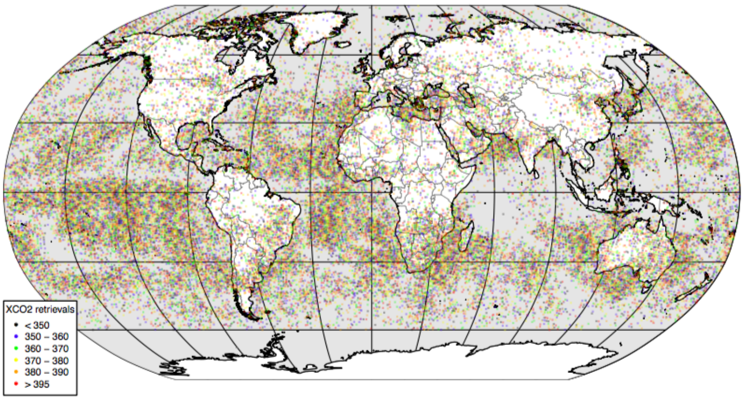


FIGURE 1.1: Satellite carbon dioxide measurements in ppm with irregular allocation

Therefore, the algorithms reducing the cost of the estimation and prediction of GRFs characterized by spatial or (and) spatio-temporal covariance functions are in great demand. Since the main difficulties arise from dealing with the inverse of the covariance matrix, there exist two possible strategies to tackle the 'big N' problem:

1. simplification of the model:
    - approximate GRF with a Gaussian Markov Random Field (GMRF) proposed by Lindgren *et al.* (2011) in the spatial context, the algorithm requires roughly $O(N \log N)$ operations;
    - approximate GRF with low-rank methods proposed by Cressie and Johannesson (2008) that yield $O(N)$ operations;
    - tapering the covariance matrix suggested by Kaufman *et al.* (2008)
2. simplification of the fitting method but keeping the model: for example, pseudo-likelihoods such as composite likelihood.

We do not consider separable structures of the space-time covariance functions as in the large majority of the studied phenomena interest lies in the interaction of space and time which gives rise to so-called non-separable models.

In the following sections we start the discussion by introducing methods proposed in the spatial context with respect to both strategies, and after we describe the non-separable models developed by Jones and Zhang (1997), Cressie and Huang (1999), Gneiting (2002), Ma (2003), Stein (2005) and others.

1.1.1 Approaches proposed to simplify the model

We firstly consider the SPDE based GMRF models described in Lindgren *et al.* (2011) which gain a great deal of computational efficiency introducing the sparse pattern in the precision matrix Q_Z, i.e. the inverse of the Matérn covariance matrix C_Z, which is one of the fundamental covariances in spatial statistics. Denoting by $||\cdot||$ the Euclidean distance, the Matérn covariance function of GRF $Z(x)$

$$c(x) = \frac{\sigma^2}{\Gamma(\nu)2^{\nu-1}}(\kappa||x||)^\nu K_\nu(\kappa||x||), \tag{1.3}$$

where ν is a smoothness parameter, K_ν is the modified Bessel function of second kind of order $\nu > 0$, Γ is the Gamma function, $\kappa = \varphi^{-1}$ is a scale parameter and σ^2 is the marginal variance.

As known from Whittle (1963), a GRF with the Matérn covariance is the solution of the Stochastic Partial Differential equation (SPDE)

$$(\kappa^2 - \triangle)^{\alpha/2} Z(x) = W(x), \quad x \in \mathbb{R}^d, \quad \alpha = \nu + d/2, \tag{1.4}$$

where $W(x)$ is a spatial Gaussian white noise with unit variance, $\kappa > 0$ is a scaling parameter, $\triangle$ is the Laplacian of the dimension d

$$\triangle = \frac{\partial^2}{\partial x_1^2} + \cdots + \frac{\partial^2}{\partial x_d^2} \tag{1.5}$$

and $(\kappa^2 - \triangle)^{\alpha/2}$ is the partial differential operator that will be described in details in the theoretical part of this thesis.

The mean differentiability (or smoothness) of the GRF is determined by the parameter ν which leads to the different covariance functions. For example, $\nu = 0.5$ leads to the exponential type of covariance function (see Figure 1.2), whereas with $\nu \to \infty$, the Matérn covariance converges to the squared exponential (or Gaussian) covariance function.

The basic idea of SPDE based GMRF models starts with the decomposition of $Z(x)$ into the linear combination of the basis functions $b_k(x)$, with $x \in \mathbb{R}^d$

$$Z(x) = \sum_{k=1}^{m} b_k(x)\beta_k, \tag{1.6}$$

where β_k are the Gaussian weights, $b_k(x)$ is the set of non-orthogonal piecewise linear basis functions on the triangulated domain with m denoting the number of vertices in

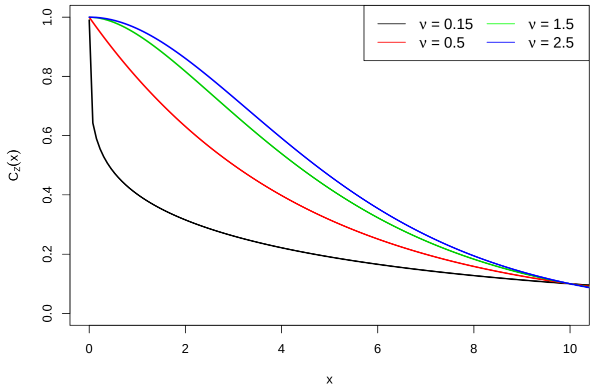


FIGURE 1.2: Matérn covariance function with different ν

the triangulation. The joint distribution of the weights β_k is defined through the Finite Element (FE) construction. The Markov property then can be seen by considering the indexed process as a solution to the SPDE and exploiting it to construct a projection to the finite element representation, where the weights are normally distributed with the sparse precision matrix. This leads to a sparse representation of the inverse of covariance matrix C_Z. The intuition can be seen on the Figure 1.3, where the full conditionals $p(x_i|x_{-i})$, $i = 1, \ldots, N$, only depend on a set of neighbours ∂i to each site i. Elements in the precision matrix of a Gaussian Markov random field are non-zero only for neighbours and diagonal elements, i.e. $Q_{ij} \neq 0 \iff i \in \partial j \cup j$.

Therefore, for two-dimensional GMRFs, this method (implemented in *R*-package *INLA* by Lindgren and Rue (2015)) leads to a $O(N^{3/2})$ cost due to the obtained sparsity in the precision matrix Q_Z.

However, the Markov property holds only for the integer values of the power α in (1.4). In particular, it is therefore not directly applicable to the special case of exponential covariance with $\nu = 1/2$ on $\mathbb{R}^2$, where $\alpha = 3/2$. As will be shown in Chapter 3, *the main benefit of our method over GMRFs is that our application is not restricted to specific values of* α. Bolin *et al.* (2017) extended the application to the fractional powers of the pseudodifferential operator. Some papers concerning SPDE framework include Bolin (2014), Sigrist *et al.* (2015) and others.

Another approach was proposed by Kaufman *et al.* (2008) and called ‘covariance tapering’. In this method, the covariance matrices are ‘tapered’ or multiplied element-wise by a sparse correlation matrix which results in another positive definite function with a compact support. With a reason to believe that distant pairs of observations are independent, we can model this structure using a compactly supported covariance function which is zero after some threshold (i.e spatial lag).

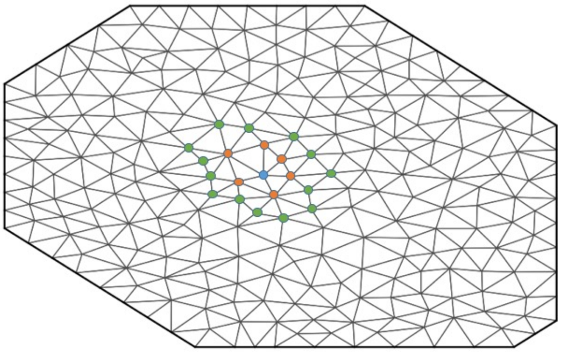


FIGURE 1.3: GMRF representation on the triangulated domain

The tapered covariance matrix takes the form $C_T = C_Z \odot T(\delta)$, where $T(\delta)_{ij} = K(||x_i - x_j||; \delta)$ and taper (or cut-off distance) is denoted by δ. The '$\odot$' notation refers to the element wise matrix product and called the Hadamard product.

Covariance tapering sets small values of the covariance to zero obtaining a positive definite sparse covariance matrix. Thus, the product of the original covariance function and a tapering function $K(v; \delta)$, an isotropic correlation function that is identically 0 whenever $v \geq \delta$.

As an example, see Figure 1.4 with the Wendland correlation function

$$K(v; \delta) = \left(1 - \frac{v}{\delta}\right)_+^4 \left(\frac{4v}{\delta} + 1\right)_+,$$

where $v = ||x_i - x_j||/\delta$ is the spatial lag with $x_i, x_j \in \mathbb{R}^d$. Small values of δ correspond to more severe tapering and for $\delta \to \infty$ we get the ML estimator.

Therefore, the tapered likelihood can be written as

$$L(\theta) = -\frac{N}{2} \log 2\pi - \frac{1}{2} \log \det(C_T) - \frac{1}{2} \boldsymbol{Z}^T C_T^{-1} \boldsymbol{Z}, \tag{1.7}$$

where $C_T = C_Z \odot T(\delta)$.

Furrer *et al.* (2006) proved the asymptotic optimality for the prediction under tapering. The benefit of the approach is that it can be adapted to any covariance functions (compared, for example, to the GMRFs) and the sparsity pattern in the covariance structure reduces the computational cost exploiting sparse algorithms. However, the covariance tapering method may not be effective in accounting for spatial dependence with long range dependence thereby sacrificing some precision as was mentioned in Sang and Huang (2012). In addition, it is not straightforward how to choose the distance to taper off. The implementation of this method is based on the *R*-package *spam* with the sparse algorithms and will be used in the application part of the thesis.

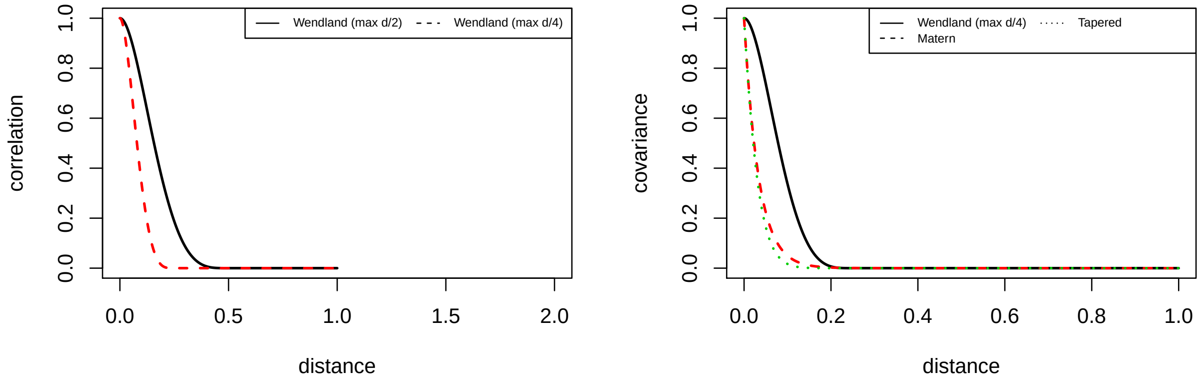


FIGURE 1.4: Wendland correlation function with different δ (left) and result of tapering (right)

Fixed Rank Kriging or FRK of Cressie and Johannesson (2008) or Gaussian Predictive Process of Banerjee *et al.* (2008) are in the group of the low-rank approximation methods. In short, these methods help to perform the exact computations on a lowered rank or simplified version of the field, thus reducing the size of the matrices.

The Gaussian Random Field $Z(x)$ in FRK is modelled as

$$Z(x) = \eta(x) + \epsilon(x), \quad x \in \mathbb{R}^d, \tag{1.8}$$

where $\{\eta(x) : x \in \mathbb{R}^d\}$ is the spatial process and $\epsilon(x)$ is a spatial white noise (or fine-scale process) with mean 0 and $\text{var}[\epsilon(s)] = \zeta^2 v(s) \in (0, \infty)$ for $\zeta^2 > 0$ and a known $v(\cdot)$.

In this approach a spatially correlated mean-zero random process $\eta(x)$ is decomposed using a linear combination of spatial basis functions as in (1.6) with random weights. In other words, the FRK method captures the scales of spatial dependence through a set of m (not necessarily orthogonal) basis functions, $\boldsymbol{B} = (B_1(s), \dots, B_m(s))^T$, where m is fixed. Thus, $\eta(x) = \boldsymbol{B}^T(x)\boldsymbol{\beta}$, where $\boldsymbol{\beta}$ is a m-dimensional vector. The covariance structure is further imposed on $\boldsymbol{\beta}$, i.e. $\text{var}(\boldsymbol{\beta}) = G$ and $G \in \mathbb{R}^{m \times m}$. Therefore, $C_Z = \boldsymbol{BGB}^T + \zeta^2\boldsymbol{V}$, where $\boldsymbol{V}$ is a diagonal matrix with entries given by the measurement error variances, i.e. $\boldsymbol{V} = \text{diag}\{v(s_1), \dots, v(s_N)\}$.

The key point of such a representation is that the number of the basis functions is much less than the number of observations, i.e $m << N$ as depicted in the Figure 1.5[2]. Therefore, the inverse of the reduced covariance matrix of the basis function weights is computed saving the computational time.

[2] The picture is obtained from Katzfuss and Cressie (2011)

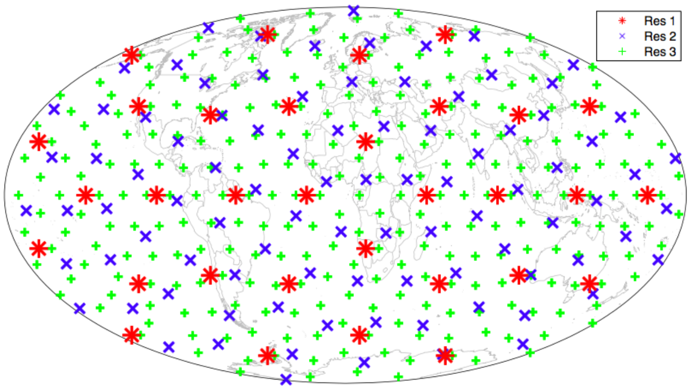


Figure 1.5: Locations of the basis functions with the three resolutions

To capture a small-scale variation, the newest implementation of this method was recently introduced in the R-package *FRK* of Zammit-Mangion and Cressie (2017) and will be exploited in the Chapter 5 of this thesis. However, as was stated by Stein (2014), the low-rank models perform poorly when neighbouring observations are strongly correlated and the spatial signal is stronger than the noise.

1.1.2 Approaches proposed to simplify the fitting method

Consider a statistical model $\{f(z,\theta), z \in \mathbb{R}^m\}$, a set of measurable events $\{\mathcal{A}_k, k = 1,\ldots,K\}$ and the associated likelihoods $L_k(\theta, z) = f(z \in \mathcal{A}_k, \theta)$. Then, following Lindsay (1988) a composite likelihood (L_{comp}) is the weighted product of the likelihoods corresponding to each single event

$$L_{\text{comp}}(\theta, z) = \prod_{k=1}^{K} L_k(\theta, z)^{\rho_k},$$

where $\{\rho_k, k = 1,\ldots,K\}$ are positive weights.

One of the first preudolikelihoods was proposed by Besag (1974). This pseudolikelihood is the product of the conditional densities of a single observation given its neighbours

$$L_{\text{cond}}(\theta, z) = \prod_{i=1}^{m} f(z_i|z_{\partial i,\theta}),$$

where ∂i is the neighbour of i.

In general, composite likelihood methods are based on the idea to subset the data as, for example, in the Figure 1.6 and find the product of the joint densities of the partitioned data. The simplest composite marginal likelihood is the pseudolikelihood constructed under independence assumptions which allows the inference only on marginal parameters

$$L_{\text{ind}}(\theta, z) = \prod_{i=1}^{m} f(z_i, \theta).$$

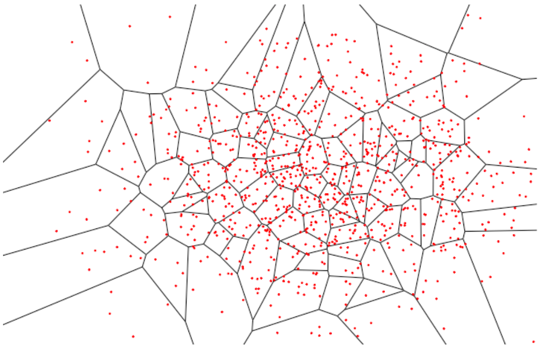


Figure 1.6: Split of the spatial domain

Composite likelihood based estimation and prediction do not involve the large $N \times N$ matrix, and reduce the computational complexity from $O(N^3)$ to roughly $O(N^2)$ floating operations.

The block composite likelihood method substitutes the original likelihood by a composite likelihood that exploits the spatial blocks resulting in a likelihood function that requires much less computational effort. The block CL as was described by Eidsvik *et al.* (2014) can be obtained through the joint density of the adjacent spatial blocks. For the general overview of most of the composite likelihood methods see Varin *et al.* (2011).

Let $z_i = z(x_i)$ be the observation of process Z at location x_i. Vecchia (1988) proposed to approximate the full likelihood with the composite conditional likelihood

$$L_{\text{CC}}(\theta, z) = f(z_1, \theta) \prod_{i=2}^{m} f(z_i | \mathcal{B}_i, \theta),$$

where $\mathcal{B}_i$ is a subset of $\{z_{i-1}, \ldots, z_1\}$ with the size chosen to gain a computational efficiency. Standard approach of Vecchia (1988) restricts $\mathcal{B}_i$ to a number of neighbours of z_i.

Unlike the low-rank methods, the asymptotic properties of the CLs are well defined and these methods allow for the parallel splitting of the job. However, the strategy for the subsetting or blocking of observations, selecting conditioning sets depends on the spatial locations and underlying correlation model which is difficult to determine in advance. The questions arising in the search of the optimal strategy are: How to define the suitable partition of the domain?, How to maximize the number of blocks?, How many observations should be in one block?, How to retain statistical efficiency while minimizing the computational cost?. However, the guideline can be partially formulated after some preliminary analysis of the considered random field.

1.1.3 Spatio-temporal covariance functions

Models that are continuous in space and time are often explored within a Stochastic Partial Differential Equation or SPDE framework. Among them are works of Whittle (1963), Heine (1955), Jones and Zhang (1997), Sigrist *et al.* (2015) and others.

The function (1.3) is one of the possible covariance functions which is a stationary solution to partial differential equation (1.4). This theoretical result was obtained by Whittle (1963) who developed a general framework where stationary Random Functions are related to the SPDEs. In addition, he also presented other examples such as spatio-temporal models related to diffusion equations with damping.

Lindgren *et al.* (2011) provided a general theory applied in the spatial context with a brief mentioning of a possible extension to the spatio-temporal context.

Heine (1955) demonstrated stationary covariance functions as the solutions to some SPDEs involving hyperbolic, parabolic and other types of second order differential operators in the two dimensions. For non-separable space-time processes $Z(s,t)$, the SPDE for a one dimensional spatial process is of the parabolic type

$$\left(\frac{\partial^2}{\partial s^2} - c\frac{\partial}{\partial t} - \kappa\right) Z(s,t) = W(s,t), \tag{1.9}$$

where $W(s,t)$ is spatio-temporal white noise in $\mathbb{R} \times \mathbb{R}$, $\kappa > 0$ is the damping parameter and $c > 0$ is a positive constant. The equation (1.9) can be seen as a heat conduction equation or diffusion equation. The covariance derived by Heine (1955) from this parabolic equation has the following form

$$C_Z(s,t) = e^{-\kappa|s|}\mathrm{Erfc}\left(\beta\sqrt{|t|} - \frac{\kappa|s|}{2\beta\sqrt{|t|}}\right) + e^{\kappa|s|}\mathrm{Erfc}\left(\beta\sqrt{|t|} + \frac{\kappa|s|}{2\beta\sqrt{|t|}}\right), \quad (s,t) \in (\mathbb{R}\times\mathbb{R}) \tag{1.10}$$

where κ, β are suitable coefficients associated with a parabolic type partial differential equation (1.9), $\beta = \kappa/\sqrt{c}$ and Erfc is the error function such that $\mathrm{Erfc}(a) = \frac{2}{\sqrt{\pi}}\int_a^\infty e^{-v^2}dv$ if $a \geq 0$ and $\mathrm{Erfc}(-a) = 2 - \mathrm{Erfc}(a)$ if $a < 0$.

Whittle (1962) commented on the necessity to include the loss term $\kappa > 0$ in (1.9) for the spatial dimension $d \leq 2$ to have a finite variance of the process $Z(s,t)$. He demonstrates that for $d > 1$, the diffusion mechanism does not smooth the process sufficiently and the spectrum decays too slowly at infinity. Since this is a consequence of the irregularity of the input (not smooth enough), it was suggested to include the autocorrelation in space or time of the random input.

Sigrist *et al.* (2015) studied a stochastic form of the advection-diffusion equation with damping. They used a linear combination of deterministic spatial functions (which are Fourier functions with the spatial wave numbers) with random coefficients that evolve dynamically for $s \in \mathbb{R}^d$

$$\left(\frac{\partial}{\partial t} + \mu^T \nabla - \nabla \Sigma \nabla + \kappa\right) Z(s,t) = \epsilon(s,t), \tag{1.11}$$

where $\nabla = \left(\frac{\partial}{\partial x}, \frac{\partial}{\partial y}\right)^T$ is the gradient operator for $s = (x,y)^T$, Σ is the identity matrix, $\nabla \Sigma \nabla$ is a diffusion term that incorporates anisotropy, $\epsilon(s,t)$ is a Gaussian process that is temporally white and spatially coloured and $\mu^T \nabla$ models transport effect or so called advection term. Their approach requires discretization in time for application to discrete time series but describe many types of natural phenomena.

Another class of spatio-temporal covariance functions was proposed by Cressie and Huang (1999) who started their discussion from the fact that separable space-time covariance functions do not take into account spatio-temporal interactions. They defined non-separable parametric spatio-temporal covariance functions based on the invertion of their spectral densities in the space-time framework.

They obtained valid spatio-temporal stationary covariance models $\text{cov}(Z(s,t), Z(s+h, t+u)) = C(h,u)$ with spatio-temporal lag (h,u) by selecting two functions, $\rho(\omega, u)$ and $k(\omega)$, that satisfy two conditions

1. for each ω, $\rho(\omega, \cdot)$ is a continuous autocorrelation function on $\mathbb{R}$ that satisfies $\int_{\mathbb{R}^d} \rho(\omega, u) du < \infty$ and $k(\omega) > 0$;

2. $\int_{\mathbb{R}^d} k(\omega) d\omega < \infty$.

and allow the integral

$$C(h,u) = \int_{\mathbb{R}^d} e^{i(h'w)} q(\omega, u) d\omega, \quad q(\omega,u) = \rho(\omega,u) k(\omega) \tag{1.12}$$

to be analytically evaluated with $\omega \in \mathbb{R}^d$.

The limitation of this class of space-time covariances is that it requires integrability and an analytically closed-form of the Fourier integral which sometimes is difficult to derive.

For this reason, Gneiting (2002) extended this approach and formulated more general class of space-time covariances based on the completely monotone functions and Bernstein functions.

A function $f(v)$ is called completely monotone over (v_1, v_2), where $-\infty \leq v_1 < v_2 \leq +\infty$, if it has derivatives $f^{(k)}$ of all orders and

$$(-1)^k f^{(k)}(v) \geq 0, \quad \text{for } v_1 < v < v_2, \quad k \geq 0. \tag{1.13}$$

This class forms a valid non-separable parametric class of space-time covariance functions which is based on the relationship between completely monotone functions and Laplace transforms of the finite and non-negative measures on $\mathbb{R}^+$. From Berstein theorem stated in Feller (1957), a completely monotone function can be represented

$$f(v) = \int_0^\infty e^{-rv} dF(r), \quad v > 0$$

with non-decreasing cumulative distribution function F.

The Berstein function is a positive function $\psi(v)$ for $v > 0$ with a completely monotone derivative. Thus, the theorem formulated by Gneiting (2002) can be stated as follows. Let $f(v)$ for $v \geq 0$, be a completely monotone function and let $\psi(v)$ for $v \geq 0$ be a Berstein function, then the space-time covariance function has the following form

$$C(h, u) = \frac{\sigma^2}{\psi(|u|^2)^{d/2}} f\left(\frac{||h||^2}{\psi(|u|^2)}\right), \quad (h, u) \in \mathbb{R}^d \times \mathbb{R}. \tag{1.14}$$

However, Kent *et al.* (2011) showed that in certain circumstances the Gneiting model possesses a dimple. For a fixed spatial lag the temporal covariance is not a decreasing function of the temporal lag which leads to an undesirable behaviour.

An alternative procedure for generating non-separable space-time covariance functions was provided by Stein (2005) and is based on the spatio-temporal spectral density $g(\omega, \tau)$ of the following form

$$g(\omega, \tau) = (c_1(a_1^2 + ||\omega||^2)^{\alpha_1} + c_2(a_2^2 + \tau^2)^{\alpha_2})^{-\nu} \tag{1.15}$$

with c_1, $c_2 > 0$, $a_1^2 + a_2^2 > 0$, α_1, α_2 are positive integers that control the spatial and temporal smoothness of the paths of the random function and $d_1/(\alpha_1 \nu) + d_2/(\alpha_2 \nu) < 2$. For example, Jones and Zhang (1997) derived the space-time covariance with the closed-form solution for $d_1 = 2$ and $\alpha_2 = \nu = d_2 = 1$.

The spectral density of the form (1.15) is formulated to avoid any kind of discontinuity that can be presented in some models and explained by the lack of differentiability property away from the origin. However, the closed-form of the covariance function obtained from (1.15) is not known in general.

In the theoretical part of this thesis, we consider covariance functions derived from an elliptic type of SPDEs as in (1.4) and Stein type spatio-temporal covariance functions based on the spectral density (1.15).

Among other space-time covariances functions not considered here are anisotropic covariance functions of Porcu *et al.* (2006), product-sum models of De Iaco *et al.* (2001) and others.

Chapter 2

Hierarchical matrices

2.1 Introduction

In this chapter we detail a theoretical formulation of the hierarchical matrices (or shortly $\mathcal{H}$-matrices) approach used in the thesis as a tool for the fast approximation of covariance matrices. To get insight into the hierarchical matrices approach, we firstly make a distinction between sparse and data sparse matrices which form the basis of the considered technique. To obtain data sparsity, hierarchical matrices rely on the low-rank representation which makes most of the matrix operations feasible for the large data. Since all the matrix operations and factorizations lead to the log-linear time of the computation, this method attracted attention and nowadays it is widely applied in statistics. Therefore, $\mathcal{H}$-matrices approach can be applied to tackle a computational problem in spatial statistics. The application of the $\mathcal{H}$-matrices in spatial statistics is discussed in section 2.3.

The first part of this Chapter aims to describe the main steps for the implementation of this technique. Particularly, we describe the construction of the binary cluster tree and block cluster tree which lead to the discrete structures of the approximated matrix. Partition of considered domain, as described in sections 2.2.1 and 2.2.2, is based on the construction of the bounding boxes around the points, chosen clustering technique and on the estimation of the distances between the boxes. Apart from the chosen clustering technique, it is required to specify the conditions for the appropriate partition of the points. This gives rise to so called admissibility condition defined in section 2.2.3. The standard asymptotic smoothness condition required in order to guarantee the fast decay of the eigenvalues of the underlying function, is described in section 2.2.4 and exposed in the theory that relates $\mathcal{H}$-matrices to spatial statistics in Chapter 3.

Section 2.2.5 provides the low-rank techniques that can be used for the approximation of those blocks of the matrix which satisfy the admissibility condition and derive this condition for specific types of spatial or spatio-temporal covariance functions.

We conclude Chapter 2 with a short literature regarding the application of the $\mathcal{H}$-matrices in spatial statistics, including works of Ambikasaran *et al.* (2013), Ballani and Kressner (2015), the paper of Litvinenko *et al.* (2019) which is closely related to our work.

2.1.1 Low-rank matrices

The main idea of the thesis is to apply $\mathcal{H}$-matrices approach for the approximation of a covariance matrix in order to reduce the computational cost of the ML estimation. The starting point here is to understand the benefits obtained in terms of the matrix operations. Therefore, before introducing the concept of the $\mathcal{H}$-matrices method, we discuss a low-rank algebra and provide the distinction between sparse and data-sparse matrices that form the basis of this approach.

A matrix $E \in \mathbb{R}^{N \times N}$ is called sparse if number of non-zero entries is much smaller than the total number of elements N^2. The covariance tapering method discussed in section 1.1.1 can serve as an example of sparse matrices application. However, the sparsity of the matrix can be easily destroyed after some matrix operations such as matrix inversion. This problem can be solved by considering the sparse approximations to the exact inverse through the sparse algorithms. By contrast, if most of the elements are non-zero, then the matrix is considered as a dense.

A matrix $E \in \mathbb{R}^{N \times N}$ is called data-sparse if it can be represented by the number of elements k called rank which is smaller than N. For example, low-rank k approximation algorithms such as singular value decomposition (SVD) or cross approximation techniques refer to this category. By storing just low-rank factors, the number of the elements can be reduced up to $2Nk$ which leads to great computational savings.

We are now ready to state the first definition of $\mathcal{H}$-matrices. An $\mathcal{H}$-matrix is an efficient data-sparse representation of the dense matrices. The idea behind $\mathcal{H}$-matrices is to use low-rank approximation of the blocks of the dense matrix which are located far from the diagonal entries. The specific partition of the matrix is obtained following the appropriate conditions which will be discussed in the following sections.

Let $C \in \mathbb{R}^{N \times N}$ be any matrix. As known, the row (column) rank of the matrix is the amount of the linearly independent row (column) vectors. Therefore, if a matrix C has rank k, then with any vector $x \in \mathbb{R}^{N \times 1}$, C spans a k-dimensional subspace. If a matrix $C \in \mathbb{R}^{N \times N}$ has rank k, then it can be written as the product of two factors, as

$C = AB^T$, where $A \in \mathbb{R}^{N\times k}$ and $B \in \mathbb{R}^{N\times k}$ are in full-matrix representation, i.e.

$$\begin{pmatrix} c_{11} & c_{12} & c_{13} & \dots c_{1N} \\ c_{21} & c_{22} & c_{23} & \dots c_{2N} \\ c_{31} & c_{32} & c_{33} & \dots c_{3N} \\ \vdots & \ddots & \ddots & \vdots \\ c_{N1} & c_{N2} & c_{N3} & \dots c_{NN} \end{pmatrix} = \begin{pmatrix} a_{11} & a_{12} & \dots a_{1k} \\ a_{21} & a_{22} & \dots a_{2k} \\ \vdots & \ddots & \vdots \\ a_{N1} & a_{N2} & \dots a_{Nk} \end{pmatrix} \times \begin{pmatrix} b_{11} & b_{12} & \dots b_{1N} \\ \vdots & \ddots & \vdots \\ b_{k1} & b_{k2} & \dots b_{kN} \end{pmatrix} \tag{2.1}$$

Thus, the lower the k, the less information is given by the matrix C and the number of the elements to store is reduced up to $2Nk$ as was stated by Ambikasaran *et al.* (2013) compared to a full matrix with N^2 number of elements. Such low-rank representation of a matrix C can be obtained by several methods, including analytic as well as algebraic techniques that will be described later. Following the terminology of $\mathcal{H}$-matrices approach, we state the Definition 2.1.

Definition 2.1. A matrix $C \in \mathbb{R}^{N\times N}$ is called a low-rank k matrix if it is given in a factorised form

$$C = AB^T, \quad A \in \mathbb{R}^{N\times k}, \quad B \in \mathbb{R}^{N\times k}, \quad k, N \in \mathbb{N}_0,$$

where A, B are full matrices.

The main idea of the low-rank matrix representation as in (2.1) is that all the matrix operations can be performed efficiently. For example, the amount of operations required for a matrix-vector multiplication Ax, where $x \in \mathbb{R}^N$, in a full matrix representation is $2N^2 - N$, whereas for a low-rank matrix it is reduced to $2Nk - N - k$ with k smaller than N.

2.2 Construction of the $\mathcal{H}$-matrices

The $\mathcal{H}$-matrices aim to find a low-rank decomposition of matrix blocks which are not data sparse. To get such a representation, the underlying function should follow some specific conditions which will be detailed in sections 2.2.3 and 2.2.4. After a low-rank decomposition is obtained, a binary cluster tree and block clusters are constructed that result in a hierarchical structure of the matrix (see Figure 2.1 as an example of $\mathcal{H}$-matrix representation for the exponential type of covariance function). These all are crucial steps required in order to compress data and perform matrix operations in a linear cost. We elaborate on each of the steps in details in the next sections.

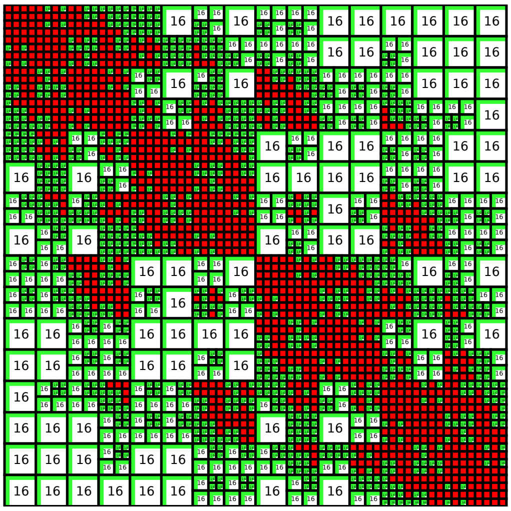


FIGURE 2.1: Hierarchical matrix representation with number of data sites $N = 8000$ and a block rank $k = 16$; green and red blocks represent low-rank and dense blocks respectively

2.2.1 $\mathcal{H}$-clusters

To obtain the structure of a covariance matrix as in Figure 2.1 with the green blocks $\tilde{C}^k_{\text{block}}$ approximated in a low-rank k format, we firstly assign to each data location $x_i \in \mathbb{R}^d$ an index $i \in I$ from the index set $I \subset \mathbb{N}$. The hierarchical structure in Figure 2.1 is obtained by partitioning the index set I into subsets or, equivalently, associated data locations x_i into clusters. This is required in order to obtain matrix blocks which further can be factored, such that a low-rank block $\tilde{C}^k_{\text{block}}$ is characterised by the rank k, where $k << N$.

In this section we provide the definitions and discuss the construction of the hierarchical cluster tree required for the $\mathcal{H}$-matrix technique.

Let I denote the index set associated with data locations x_i for $i \in I$ and $i = 1, \ldots, N$. The data sites should not be necessarily ordered [1]. The $\mathcal{H}$-matrix method is based on the hierarchical partitioning of I following a binary cluster tree structure. Namely, the partition of an index set I creates the sequence $M^l_I = \{I^l_1, \ldots, I^l_{2^l}\}$ on each level l with $l = 0, \ldots, L-1$ of the binary tree, where L is a total number of levels or so called *depth*.

Definition 2.2. We can define

- *root* of the tree on the first level, i.e. $l = 0$, so that $M^0_I = \{I^0_1\} = I$, where I is the entire set that contain indices $i = 1, \ldots, N$ for all the data sites;

[1] The method of hierarchical matrices does not require a special ordering of the indices because it creates its own ordering of indices in I

- a *partition* of I is a set of *disjoint* and non-empty subsets whose union is the entire set I, i.e. $I_q^l = \bigcup\limits_{j=s}^{s+1} I_j^{l+1}$, where $s = 2^q - 1$ and $q = 1, \dots, 2^l$, so that index set on a level l is the *union* of the index sets on the level $(l+1)$ down the tree;
- I_q^l is also called a *cluster*.

The example of the cluster tree partitioning can be seen on the Figure 2.2.

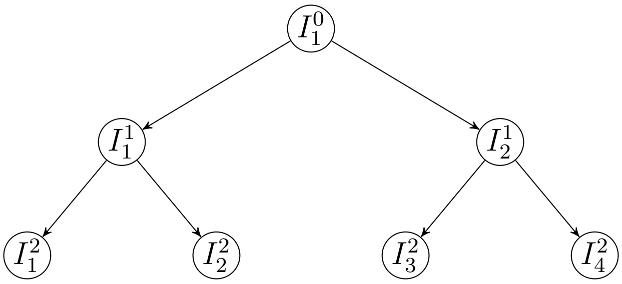


FIGURE 2.2: An example of the partition of the cluster tree

The hierarchical partitioning starts with the full index set $\{I_1^0\} = I$, which is defined as a root of the cluster binary tree (vertex of the cluster tree on the Figure 2.2). A suitable technique is further applied in order to find a disjoint partition of the index set I and use this partition to create index sets on the next levels l.

The hierarchical partitioning of the index set I associated with data locations x_i may be constructed through recursive algorithm. Let $x_{i,w} \in \mathbb{R}^d$ for $i \in I$, where $I = \{1, \dots, N\}$ is the index set for spatial data locations and $w = 1, \dots, d$. For a subset $\sigma \subset I$, we define an axis-parallel box $Q_\sigma \subset \mathbb{R}^d$. This box is termed a *bounding box* of σ

$$Q_\sigma = [a_1, b_1] \times \cdots \times [a_d, b_d], \tag{2.2}$$

where $a_w = \min_{i \in \sigma}\{x_{i,w}\}$ and $b_w = \max_{i \in \sigma}\{x_{i,w}\}$ and contains the data points x_i for $i \in \sigma$.

Then an index set I_q^l can be partitioned into sets I_{2q-1}^{l+1}, I_{2q}^{l+1} according to the following cluster algorithm:

- define the coordinate direction $w \in (1, \dots, d)$ of maximal extent along which the index set I_q^l will be divided, such that $(w :\ \text{argmax}|b_w - a_w|)$ where a_w, b_w define the extent of the bounding box $Q_{I_q^l}$ of I_q^l along the spatial direction w;
- split the box perpendicular to this direction into two subdomains by setting $c_w = (a_w + b_w)/2$

$$I_{2q-1}^{l+1} = \{i \in I_q^l | x_{i,w} < c_w\}, \quad I_{2q}^{l+1} = \{i \in I_q^l | x_{i,w} \geq c_w\}.$$

This procedure is applied recursively to all the clusters (or index sets) until an index set has size $N < N_{\min}$, where $N_{\min}$ is so called *leafsize* and can be stated in advance.

2.2.2 $\mathcal{H}$-block clusters

In the previous section we described the concept of hierarchical partitioning of the index set I associated with the data locations. We now define the hierarchical block partitioning which leads to a final representation of hierarchically partitioned covariance matrix $\tilde{C}$ as in the Figure 2.1.

Let $I, J \subset \mathbb{N}$ now denote finite row and column index sets[2] respectively. From now on, we denote the underlying covariance function $c(\cdot)$ by $c(x_i, x_j)$ to emphasize the dependence of the covariance on two data locations $x_i, x_j \in \mathbb{R}^d$. Then a full covariance matrix with the data sites x_i, x_j on the level $l = 0$ (root level of the cluster tree) is $C \in \mathbb{R}^{I \times J}$ for $i \in I$ and $j \in J$, where the number of indices in each of I and J is the number of data sites N.

As in the previous section we can associate the row and column partitions M_I^l and M_J^l, such that a *block partition* can be defined as $M_{I\times J}^l = \{b_1^l, \ldots, b_{2^{2l}}^l\}$, where a block $b_q^l \subset I \times J$ and $l = 0, \ldots r$ with $q = 1, \ldots, 2^{2l}$, where r is the minimum between the depths L_I and L_J.

Definition 2.3. We give the definition of the hierarchical block tree with the root $I \times J$ required for the construction of the matrices $\mathbb{R}^{I\times J}$ in the $\mathcal{H}$-format

- $M_{I\times J}^l$ is a hierarchical partition of $I \times J$;
- for every block b_q^l, we can define $\sigma \subset I$ and $\tau \subset J$, such that $b_q^l = \sigma \times \tau$;
- *block cluster* tree is characterised by the partition $M_{I\times J} = \{b_1, \ldots, b_{2^{2l}}\}$ of $I \times J$, where each block $b_q \in \bigcup\limits_{l\in\{0,\ldots,L-1\}} M_{I\times J}^l$.

Since a block tree is a special cluster tree for $I \times J$, it corresponds to a disjoint partition of matrix entries which, in turn, leads to the block structure of matrix. Following definition 1.14 of Grasedyck and Hackbusch (2003) reformulated in our framework, the covariance matrix C approximated with the $\mathcal{H}$-method with a block $\sigma\times\tau$ can be defined as

$$\{C \in \mathbb{R}^{I\times J} \mid \forall\ \sigma \times \tau \subset I \times J : \operatorname{rank}(C_{\sigma\times\tau}) \leq k \text{ or } N(\sigma) \leq N_{\min} \text{ or } N(\tau) \leq N_{\min}\}$$

[2] 'row' and 'column' are defined with a reference to matrix

where k is the rank of a block, $N_{\min}$ is the leafsize, $N(\sigma)$ (or $N(\tau)$) is a number of the elements $\sigma \subset I$ (or $\tau \subset J$).

The partition of the block cluster tree must follow so called admissibility condition which we describe in details in the next section. If $b_q \in M_{I\times J}$ is admissible, the partition of this block stops, otherwise it is continued by recursive algorithm. Considering the example given in the previous section (Figure 2.2), the root block ($l = 0$) can be denoted as $I_1^0 \times I_1^0$. The blocks $I_q^l \times I_{q'}^l$ that are not admissible are recursively partitioned, i.e., $I_q^l \times I_{q'}^l$ is partitioned into the set

$$\{I_{2q-1}^{l+1} \times I_{2q'-1}^{l+1}, I_{2q}^{l+1} \times I_{2q'-1}^{l+1}, I_{2q-1}^{l+1} \times I_{2q'}^{l+1}, I_{2q}^{l+1} \times I_{2q'}^{l+1}\}$$

on each level l.

The example of the block partition of $\mathcal{H}$-matrix with level $l = 0, \ldots, L$, where $L = 2$, based on the cluster tree with spatial data sites can be seen in the Figure 2.3. We notice that the spatial sites x_i are permuted based on their 'proximity' and are assigned new indices on each level l. A covariance matrix C has a $\mathcal{H}$-matrix representation if for $\forall \sigma\times\tau$ with $N(\sigma) \le N_{\min}$ or $N(\tau) \le N_{\min}$, a matrix block $C_{\sigma\times\tau}$ admits a full representation (red-coloured blocks) and low-rank k representation for the other leaves (green-coloured blocks).

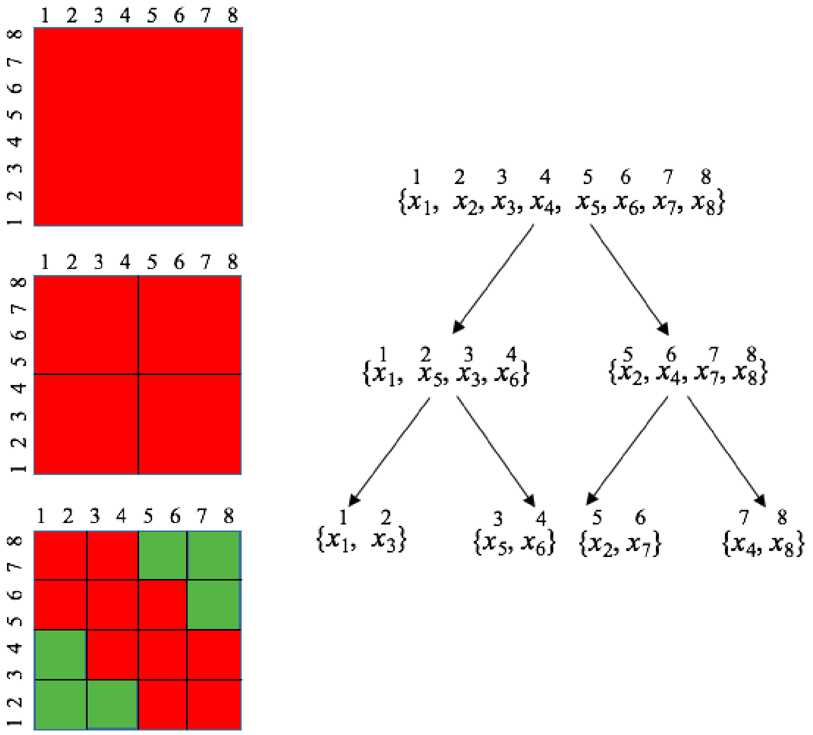


FIGURE 2.3: Block partition (on the left) of the $\mathcal{H}$-matrix according to the cluster tree (on the right) with $l = 0, \ldots, 2$

As an example of $\mathcal{H}$-matrices application, we consider the algorithm of a matrix-vector multiplication. Let covariance matrix $\tilde{C}$ be approximated with $\mathcal{H}$-matrix approach. Then the matrix-vector multiplication $y = y + \tilde{C}x$, with $x \in \mathbb{R}^J$, $y \in \mathbb{R}^I$ can be accomplished by the function $MV(\tilde{C}, I \times J, x, y)$ in the Algorithm 1.

Algorithm 1 Matrix-vector multiplication $\tilde{C}x$

```
Function $MV(\tilde{C}, \sigma \times \tau, x, y)$
if $\sigma \times \tau$ is not a leaf block then
    subsdivide block: for each $\sigma' \times \tau' \subset \sigma \times \tau$ do $MV(\tilde{C}, \sigma' \times \tau', x, y)$
else
    full or low-rank $k$ block: $y_\sigma = y_\sigma + \tilde{C}_{\sigma \times \tau} x_\tau$
end if
```

The complexity estimates for all the operations involving $\mathcal{H}$-matrices can be found in Hackbusch (2015).

To conclude this section, we formulate a new definition of the $\mathcal{H}$-matrices. The $\mathcal{H}$-matrix is a matrix whose block index set has been hierarchically partitioned and whose resulting matrix blocks are given in the factored form.

2.2.3 Admissibility condition

One of the main ideas of $\mathcal{H}$-matrix method is to find a low-rank k representation of separate blocks of the matrix. The admissibility condition is required to define which block $\sigma \times \tau \subset I \times J$ can be approximated by a rank-k matrix. Therefore, we describe this condition in details.

Let I, J be two finite index sets and $M_{I \times J}$ is a partition of $I \times J$ as was defined in the previous section. In general, the admissibility condition can be defined as a function

$$\mathrm{Adm} : M_{I \times J} \to \{\mathrm{true}, \mathrm{false}\},$$

so that a block $b_k : \sigma \times \tau$ is called *admissible* if $\mathrm{Adm}(b_k) = \mathrm{true}$.

Definition 2.4. Let $x_i \in \mathbb{R}^d$ for $i \in I$ be a set of data locations for an index set I. For $\sigma \subset I$ and $\tau \subset J$, a cluster pair (σ, τ) is considered admissible if

$$\min\{\mathrm{diam}(Q_\sigma), \mathrm{diam}(Q_\tau)\} \leq \eta \ \mathrm{dist}(Q_\sigma, Q_\tau) \tag{2.3}$$

with some $\eta > 0$, the bounding boxes $Q_\sigma = [a_1, b_1] \times \cdots \times [a_d, b_d]$, $a_w = \min_{i \in \sigma}\{x_{i,w}\}$ and $b_w = \max_{i \in \sigma}\{x_{i,w}\}$ for $w = 1, \ldots, d$, and $Q_\tau = [c_1, d_1] \times \cdots \times [c_d, d_d]$.

We define

$$\operatorname{diam}(A) = \max_{x,x' \in A} ||x - x'||, \quad \operatorname{dist}(A, B) = \min_{x \in A, x' \in B} ||x - x'||,$$

i.e. the diameter and distance of bounding boxes are computed with respect to the Euclidean norm.

The intuition behind the Definition 2.4 can be stated as follows. The error $|c(x_i, x_j) - \tilde{c}(x_i, x_j)|$ of a low-rank approximation of the covariance function $c(x_i, x_j)$ converges exponentially fast in k provided that the condition (2.3) is satisfied. In other words, the speed of convergence can be defined by the equation (2.3).

As was discussed in Hackbusch (2015), the smaller the parameter η, the more favourable is the admissibility property. In some applications, there is no upper bound. As soon as η is fixed or if we do not want to specify its value explicitly, it is suggested to leave its default value which is $\eta = 1$. Litvinenko *et al.* (2019) exploited the value of $\eta = 2$, but without expanding on the choice.

The condition (2.3) will be justified and explained with its practical application on the completely monotone functions in the section 2.3.1, where we also expose in details the origin of the parameter η.

We now briefly discuss a new procedure for the application of $\mathcal{H}$-matrices in the space-time domain $\mathbb{R}^d \times \mathbb{R}$. The procedure starts with the defining the spatial I and temporal U index sets in the same way as was described in section 2.2.1. We then construct the spatial bounding boxes Q_σ with $\sigma \subset I$ and Q_τ with $\tau \subset J$ as was defined in the previous section. For the temporal dimension, we obtain the time interval P_γ with $\gamma \subset U$ and P_δ with $\delta \subset E$ which are further partitioned and the ordered nature of the time axis is preserved.

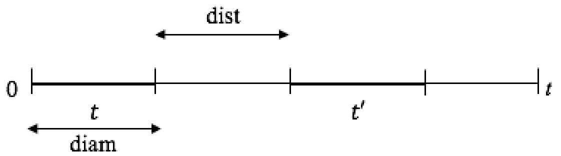


Figure 2.4: Partition of the time axis

On each level l of space-time partition we obtain clusters with Cartesian products of spatial bounding boxes Q_σ and temporal intervals $[a, b]$, i.e. $Q_\sigma \times [a, b]$. A block tree with space-time clusters is then constructed based on the new admissibility condition between space-time blocks $Q_\sigma \times [a, b]$ with $[a, b] \in P_\gamma$ and $Q_\tau \times [a', b']$ with $[a', b'] \in P_\delta$. However, this condition should be adapted specifically to each space-time function separately and will be derived for some cases in section 2.3.1. See Figure 2.4 for the partition of the time axis. The blocks that are inadmissible and not smaller than $N_{\min}$ are recursively partitioned, i.e. the partition stops when a block becomes admissible or $N_{\min}$ is achieved as in the spatial case.

2.2.4 Asymptotic smoothness condition

Another formulation of the $\mathcal{H}$-matrices can be stated as follows. If the covariance function is differentiable and can achieve the analytic form away from the diagonal, it allows for its separable approximation by the $\mathcal{H}$-method. It worth to mention that the matrix C resulting from a covariance function $c(\cdot)$ is neither a sparse matrix nor a data sparse (section 2.1). To find a data sparse representation of some blocks of the covariance matrix, their low-rank decomposition can be exploited. According to Hackbusch (2015), to admit such a representation it is necessary that the underlying functions satisfy so called 'asymptotic smoothness condition'.

The benefit of such a condition is that it guarantees an 'automatic' fulfilment of the admissibility condition discussed in the previous section. In other words, there is no need to explicitly derive the condition (2.3) provided that the asymptotic smoothness condition is satisfied. However, not all the functions can satisfy this condition. Therefore, this thesis aims to find out whether this condition can be satisfied by some spatial (or spatio-temporal) covariance functions. If this condition is satisfied, then it guarantees a successful approximation of the covariance matrices by the $\mathcal{H}$-matrices.

We define a d-dimensional multi-index notation $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_d)$ of non-negative integers. For the multi-index $\alpha \in \mathbb{N}_0^d$ sum of the components or absolute value can be written as $|\alpha| = \alpha_1 + \alpha_2 + \cdots + \alpha_d$ and higher-order partial derivatives as $\partial^\alpha = \partial_1^{\alpha_1} \partial_2^{\alpha_2} \ldots \partial_d^{\alpha_d}$, where $\partial_i^{\alpha_i} = \partial^{\alpha_i} / \partial x_i^{\alpha_i}$ of the dimension d.

We now state the Definition 2.5 of the asymptotically smooth functions given by Hackbusch (2015) and reformulated in terms of covariance functions.

Definition 2.5. Let $X_i, X_j \subset \mathbb{R}^d$ be subsets such that the function $c(x_i, x_j)$ is defined and arbitrarily often differentiable for all spatial locations $x_i \in X_i$ and $x_j \in X_j$ with $x_i \neq x_j$ for $i, j = 1, \ldots, N$. Then the function $c(x_i, x_j)$ is asymptotically smooth if there exist constants $p_1, p_2 \in \mathbb{R}^+$, such that for all multi-indices $\alpha \in N_0^d$, one has

$$|\partial_x^\alpha c(x_i, x_j)| \leq p_1 |\alpha|! p_2^{|\alpha|} (||x_i - x_j||)^{-|\alpha|} \tag{2.4}$$

for all $x_i \neq x_j$.

The factor $p_2^{|\alpha|}$ allows for a change of the growth behaviour. The derivatives tend to 0 as $||x_i - x_j|| \to \infty$. The details of the condition (2.4) are also given in Hackbusch (2015).

Asymptotically smooth and admissibility conditions are related in such a way that the functions that are asymptotically smooth allow for degenerate approximations of the functions on pairs of domains satisfying the admissibility condition (2.3), such that as in Grasedyck and Hackbusch (2003)

$$\max_{(x_i,x_j)\in Q_\sigma \times Q_\tau} |c(x_i, x_j) - \tilde{c}(x_i, x_j)| = \mathcal{O}(p_{\sigma,\tau}^{k^{1/d}}) \tag{2.5}$$

for a block $\tau \times \sigma \subset I \times J$ and the constant $p_{\sigma,\tau} < 1$ depends on the ratio of their distance and diameter in the admissibility condition (2.3). Therefore, the condition (2.4) is required to ensure $p_{\sigma,\tau} < 1$. In this way, we guarantee an exponential convergence with respect to the rank k. The implication of the function being asymptotically smooth is that the blocks of the corresponding matrix located away from the diagonal have exponentially decaying singular values and are well-approximated by the low-rank matrices.

At first sight, the condition (2.4) seems to be restrictive. However, as will be seen from the Chapter 3, this condition is satisfied by the covariance functions derived from the stochastic partial differential equations. In Chapter 5 of this thesis we consider the condition (2.4) for the approximation of the Matérn covariance function (1.3) by $\mathcal{H}$-matrices.

2.2.5 Low-rank approximation methods of admissible blocks

In this section we focus on the block of the covariance matrix C^k_{block} and aim to find a low-rank approximation $C_{\text{block}} \approx AB^\top$, so that the block of the covariance can be written

$$C^k_{ij}|_{\text{block}} \approx \sum_{\nu=1}^{k} A_{i\nu} B_{j\nu},$$

where $A = (a_{i\nu}) \in \mathbb{R}^{\tau \times k}$ and $B = (b_{j\nu}) \in \mathbb{R}^{\sigma \times k}$. Such a representation can be attained for a block of the matrix that satisfies the admissibility condition (2.3).

Definition 2.6. A function f is said to be degenerate (or separable) if it can be written as a sum of terms, each being a product of a function a of x_i and a function b of x_j.

Therefore, if the underlying covariance function $c(x_i, x_j)$ from the matrix C allows for such a separable approximation, then the block matrix yields a low-rank approximation

$$\tilde{c}^k_{\text{block}}(x_i, x_j) = \sum_{\nu=1}^{k} a_\nu(x_i) b_\nu(x_j), \quad x_i \in Q_\sigma, x_j \in Q_\tau \tag{2.6}$$

on a subset $Q_\sigma \times Q_\tau \subset \mathbb{R}^d \times \mathbb{R}^d$, where k is the rank of matrix.

We now discuss two analytical methods to construct separable expansions $\tilde{c}^k(x_i, x_j)$ of the covariance function $c(x_i, x_j)$: (1) Taylor expansion and (2) Polynomial interpolation. We will also briefly expand on the algebraic method known as an adaptive cross approximation.

If the covariance function $c(x_i, x_j)$ allows for a separable approximation $\tilde{c}(x_i, x_j)$ as in (2.6) on a subset $Q_\sigma \times Q_\tau \subset \mathbb{R}^d \times \mathbb{R}^d$, then the corresponding matrix block, where $x_i \in Q_\sigma$ and $x_j \in Q_\tau$, has at most rank k.

We further consider $c(x_i, x_j)$ as a function of $x_i \in Q_\sigma$ for a fixed $x_j \in Q_\tau$ which can be written through the sum of Taylor expansion with respect to x_i around a point of expansion x_{0i}, which is the midpoint of Q_σ

$$c(x_i, x_j) = \sum_{|\alpha|=0}^{p} (x_i - x_{0i})^\alpha \frac{1}{\alpha!} \partial^\alpha_{x_i} c(x_{0i}, x_j) + R_k, \tag{2.7}$$

where the remainder term R_k depends on the rank k and the rate of convergence (2.5). The rank of expansion k is the number $\alpha \in \mathbb{N}^d$ such that $|\alpha| < p = \binom{p+d-1}{d}$. Thus the approximation $\tilde{c}^k(x_i, x_j)$ yields

$$\tilde{c}^k(x_i, x_j) = \sum_{|\alpha|=0}^{p} \underbrace{(x_i - x_{0i})^\alpha \frac{1}{\alpha!}}_{a(x_i)} \underbrace{\partial^\alpha_{x_i} c(x_{0i}, x_j)}_{b(x_j)}. \tag{2.8}$$

As a common practice the point of expansion x_{0i} is chosen to be a center of a bounding box Q_σ (respectively of Q_τ with respect to x_{0j}). As can be seen, the Taylor expansion (2.7) requires many computations of the multidimensional derivatives which are difficult to find and, thus, works only if the derivatives of a function can be evaluated efficiently. In order to avoid this restriction, in this work we exploit Lagrange polynomial interpolation technique.

For the low-rank approximation of the off-diagonal blocks Lagrange interpolation can be applied with the basis consisting of polynomials L_α. Let $\hat{x}_\alpha = (\hat{x}_{n_1,1}, \ldots, \hat{x}_{n_d,d}) \in \mathbb{R}^d$, $\alpha = (\alpha_1, \ldots, \alpha_d) \in \{1, \ldots, p\}^d$ be p^d points on a tensor grid which are transformed to the intervals within a bounding box Q_σ

$$\hat{x}_{e,f} = \frac{a_e + b_e}{2} + \frac{b_e - a_e}{2} \cos\left(\frac{2f-1}{2p}\pi\right), \tag{2.9}$$

where $e \in \{1, \ldots, d\}$ and $f \in \{1, \ldots, p\}$. Therefore, for $d = 1$, Lagrange polynomials in $[a_e, b_e]$

$$L_{e,f}(x) = \prod_{r \in \{1,\ldots,p\} \setminus \{f\}} \frac{x - \hat{x}_{e,r}}{\hat{x}_{e,j} - \hat{x}_{e,r}} \tag{2.10}$$

and the multivariate interpolation problem can be solved by the tensor product of univariate Lagrange interpolation

$$L_\alpha(x) = L_{1,n_1}(x_1) \cdot \cdots \cdot L_{d,n_d}(x_d) \tag{2.11}$$

and $L_\alpha(\hat{x}_r) = \delta_{\alpha,r}$ for $|r| < p$ (Kronecker delta function). Thus, the Lagrange interpolation which approximates a function at the point $\hat{x}_\alpha$ can be formulated as follows

$$\tilde{c}^k(x_i, x_j) = \sum_{|\alpha|<p} \underbrace{L_\alpha(x_i)}_{a(x_i)} \underbrace{c(\hat{x}_\alpha, x_j)}_{b(x_j)}, \tag{2.12}$$

where the rank $k = p^d$.

One of the most important decompositions of the matrix is the singular value decomposition (SVD). A k-order SVD representation in Hackbusch (2015) of a matrix $C \in \mathbb{R}^{M \times N}$ can be writen as $C_k = U\Lambda_k V^T = \sum_{i=1}^k \lambda_i u_i v_i^T$, where the first singular values $k = \min(M, N)$ are given in the decreasing order, i.e. $\lambda_1, \lambda_2, \ldots, \lambda_k$ and $U \in \mathbb{R}^{M \times M}$ and $V \in \mathbb{R}^{N \times N}$. The faster the decay, the more weight is given by the first singular values.

A fast decay of the singular values of the matrix is required to guarantee the convergence of the error of the low-rank approximation. This leads to the fact that the most weight is concentrated on the diagonal and more blocks are given in the low-rank k form which entails the lower computational cost. Then the best approximation of a matrix by the $\mathcal{H}$-method can be obtained by preserving the biggest first k singular values.

As known, SVD is a computationally expensive operation. Therefore, as another approximation technique we refer to the algebraic approach which is called a rank-k adaptive cross approximation (ACA). As was discussed in the beginning of this chapter, any matrix $C \in \mathbb{R}^{N \times N}$ can be represented as $C = AB^T$, where $A \in \mathbb{R}^{N \times k}$ and $B \in \mathbb{R}^{N \times k}$. This form of the matrix of the rank k can be obtained through the ACA algorithm which computes vectors a_s and b_s that form $\tilde{C}^k = \sum_{s=1}^k a_s b_s^T$ such that $||C - \tilde{C}^k||_F \leq \varepsilon$. The

Algorithm 2 Adaptive cross-approximation $C \approx AB^T$

```
while threshold is not achieved do
    $p = 0$;
    Choose a pivot element $c_{i_l j_l}$ which is not close to 0;
    Compute $i_l$ row and $j_l$ column of $C$;
    Set $a_l = c_{,j_l}$ and $b_l = c_{i_l,}^T / c_{i_l j_l}$
    Perform $C = C - a_l b_l^T$;
    $p = p + 1$;
end while
```

approximation is terminated when some threshold ε is achieved. See Algorithm 2 for the details.

Remark. In general there exist two different $\mathcal{H}$-matrix approximation strategies. The *fixed rank* strategy is the method where each sub-block has maximal rank k. However, all the blocks are different and, thus, the accuracy of the approximation in each block cannot be controlled. The *adaptive rank* strategy is based on the idea to represent each block with absolute accuracy (in the spectral norm) which is smaller than ε or better (smaller). The adaptive rank strategy is useful when the accuracy in each block is crucial.

2.3 $\mathcal{H}$-matrices in spatial statistics

The first and basic idea of this thesis is to apply the $\mathcal{H}$-matrices to the covariance matrices in spatial statistics framework. We aim to approximate C_Z by a matrix $\tilde{C}_Z$ which can be stored in a data-sparse format so that the computational cost of the matrix operations is reduced.

As discussed in the previous section, the application of the $\mathcal{H}$-matrices in the spatial context can be motivated by the fact that all the matrix operations and factorizations lead to the $O(k^b N \log^b N)$ floating point operations, where k is the rank, N is a number of observations and $b = 1, 2$ as mentioned by Litvinenko *et al.* (2019). Matrix-vector operations can be also computed at linear cost which is widely applied in computation of the maximum likelihood estimation (MLE) which is one the most important goals in statistics.

For example, maximum likelihood estimation (1.2) can serve as one of the examples of the $\mathcal{H}$-matrices application in statistics and it was exploited in Litvinenko *et al.* (2019). As known, computation of the maximum likelihood can be composed of the Cholesky decomposition that can reduce the computational cost from $O(N^3)$ to $O(N^2)$, followed by the corresponding solver using the vector of observations $\boldsymbol{Z}$. Moreover, the determinant can be found by simply computing the product of diagonal entries of the Cholesky factor Λ. According to Litvinenko *et al.* (2019), the Cholesky factorization with the $\mathcal{H}$-matrices application results in the computational cost of $O(k^2 N \log^2 N)$, thereby reducing the total cost of the MLE. Very recent application of the $\mathcal{H}$-matrices was performed with respect to the first and second-order derivatives of the likelihood by Geoga *et al.* (2019).

Within the Bayesian geostatistical framework, the major problem which arises is to make an inference of the underlying process when a large amount of the measurements becomes available. This inference involves the infeasible computation of the inverse of the dense matrices. As was discussed in Ambikasaran *et al.* (2013), the estimation of the posterior mean and probability distribution requires the computation of the inverse of the covariance with the cost $O(m^2n + mN^2 + mpN)$, where N is the number of measurements and m is a set of the unknowns with $m >> N$. To reduce the computational cost they describe the fast algorithms. Among them are the Fast Fourier Transform (FFT) method which is applied to the regular grid to obtain a specific Toeplitz structure, Fast Multipole method (FMM) proposed by Greengard and Rokhlin (1987). FMM relies on the analytic expression for the multipole expansion of the kernel function and is similar to the $\mathcal{H}$-matrices approach. At last, they consider the $\mathcal{H}$-matrices technique and apply this method to the approximation of P.

Originally $\mathcal{H}$-matrix technique was developed for the approximation of dense matrices coming from partial differential and integral equations in details described in Hackbusch (2015). Particularly, this method is widely applied to the discretized operator whenever the underlying kernel function is asymptotically smooth. As known from Whittle (1963), the Matérn covariance (1.3) is the solution of stochastic partial differential equation with the elliptic operator (1.4) and can be considered as a Green function. The link between the Green function and the $\mathcal{H}$-matrices will be discussed in Chapter 3 of this thesis.

Other applications of the $\mathcal{H}$-matrices include Ballani and Kressner (2015) that exploited $\mathcal{H}$-matrices approach to find a maximum log-likelihood estimator of a dense covariance matrix inverse from p-variate Gaussian distribution and Li *et al.* (2014) that considered Kalman filter variant powered by $\mathcal{H}$-matrices.

In addition, there exist other variants of the hierarchical matrices that can be found in the literature, for example HODLR matrices Ambikasaran *et al.* (2013) and recursive hierarchical matrices of Chen and Stein (2017). In this thesis we exploit one of the first $\mathcal{H}$-matrices proposed by Hackbusch (2015).

To conclude, we summarise the main steps required to find an $\mathcal{H}$-representation to a covariance matrix.

1. Check the asymptotic smoothness condition of the underlying covariance function $c(x)$ which leads to the effective low-rank approximation $\tilde{C}^k_{\text{block}}$ of specific blocks of a covariance matrix C. This condition has link to the partial differential equations and can be adapted to the stochastic framework. This will be discussed in Chapter 3.

2. Formulate a general admissibility condition which dictates the partition of the covariance matrix into the blocks of the dense (red-coloured) and low-rank (green-coloured) structures (see Figure 2.1 as an example of such partition). The blocks C_{block} of a covariance matrix which satisfy the admissibility condition (i.e. green blocks) are approximated in a low-rank format, i.e. we obtain $\tilde{C}^k_{\text{block}}$. In Chapter 5 we will adapt this condition to the spatial statistics.

3. Obtain a suitable hierarchical partitioning of sets of indices which are assigned to all locations $x_i \in \mathbb{R}^d$ for $i = 1, \ldots, N$, following a binary cluster tree structure; define the hierarchical block partitioning (or block clusters) as the Cartesian products of the index sets which form the blocks of a covariance matrix. These are crucial steps required in order to compress data and perform matrix operations in a linear cost.

2.3.1 Admissibility condition for covariance functions

In this section we aim to derive the admissibility conditions, required for low-rank k representation of a block, for several classes of covariance functions. The asymptotic smoothness condition discussed in section 2.2.4 is required in order to guarantee an exponential convergence of the error of approximation of the function by low-rank techniques $|c(x_i, x_j) - \tilde{c}(x_i, x_j)|$, so that $p_{\sigma,\tau} < 1$ in (2.5). However, we can start from the 'root' and find the Taylor series expansion of a covariance function which, however, requires a computation of the derivatives. The new admissibility condition can be then obtained in order to guarantee the exponentially convergence of error $|c(x_i, x_j) - \tilde{c}(x_i, x_j)|$. We now will provide some examples in order to comprehend a link between admissibility condition and low-rank approximation, particularly Taylor series expansion.

We begin by considering completely monotone functions as in (1.13) given in section 1.1.3. Typical examples of such functions are given in the Table 2.1.

As an example, we consider a completely monotone function with $x_i, x_j \in \mathbb{R}$

$$f(x_i, x_j) = (1 + |x_i - x_j|)^{-1},$$

from the Table 2.1 with $c, \gamma, \nu = 1$, where $x_i \in Q_\sigma$, $x_j \in Q_\tau$, $\sigma \subset I, \tau \subset J$.

Let disjoint intervals be defined as $Q_\tau : [a, b]$, $Q_\sigma : [c, d]$, $Q_\tau \times Q_\sigma$ be a subdomain with the property $b < c$. The derivative of order α is of the following form

$$\partial^\alpha_{x_i} f(x_i, x_j) = (-1)^\alpha \alpha! (1 + |x_i - x_j|)^{-\alpha-1}.$$

Then, in the radius of convergence, the Taylor series of $f(x_1, x_j)$ in $x_{0i} = (a+b)/2 \in Q_\sigma$ is $f(x_i, x_j) = \sum_{\alpha=0}^\infty \frac{1}{\alpha!} \partial^\alpha_x f(x_{0i}, x_j)(x_i - x_{0i})^\alpha$ with the remainder

$$f(x_i, x_j) - \tilde{f}(x_i, x_j) = \sum_{\alpha=k}^\infty \frac{1}{\alpha!} \partial^\alpha_x f(x_{0i}, x_j)(x_i - x_{0i})^\alpha,$$

such that

$$\begin{aligned} \left| \sum_{\alpha=k}^\infty \frac{1}{\alpha!} \partial^\alpha_x f(x_{0i}, x_j)(x_i - x_{0i})^\alpha \right| &= \left| \sum_{\alpha=k}^\infty (-1)^\alpha \frac{\alpha!}{\alpha!} (x_i - x_{0i})^\alpha (1 + |x_{0i} - x_j|)^{-\alpha-1} \right| \leq \\ &\leq \left| \sum_{\alpha=k}^\infty (-1)^\alpha \left(\frac{x_i - x_{0i}}{x_{0i} - x_j} \right)^\alpha \right| \end{aligned} \tag{2.13}$$

Let $x_i \in [a, b]$, $a < b$ and $x_j \in [c, d]$. Then from (2.13)

$$\begin{aligned} \sum_{\alpha=k}^\infty \left| \frac{x_i - x_{0i}}{x_{0i} - x_j} \right|^\alpha &\leq \sum_{\alpha=k}^\infty \left| \frac{|x_{0i} - a|}{|x_{0i} - a| + |c - b|} \right|^\alpha \\ &= \left(1 + \frac{|x_{oi} - a|}{|c - b|} \right) \left(1 + \frac{|c - b|}{|x_{oi} - a|} \right)^{-k}, \end{aligned}$$

from where, since $1 + \frac{|c-b|}{|x_{0i} - a|} > 1$, the radius of convergence covers the whole interval $[a, b]$.

If $c \to b$ then the estimate for the remainder tends to infinity and the remainder can

Function	Parameters
$f(x) = \exp(-cx^\gamma)$	$c > 0, \quad 0 < \gamma \leq 1$
$f(x) = (2^{\nu-1}\Gamma(\nu))^{-1}(cx^{1/2})^\nu K_\nu(cx^{1/2})$	$c > 0, \quad \nu > 0$
$f(x) = (1 + cx^\gamma)^{-\nu}$	$c > 0, \quad 0 < \gamma \leq 1, \quad \nu > 0$
$f(x) = 2^\nu(\exp(cx^{1/2}) + \exp(-cx^{1/2}))^\nu$	$c > 0, \quad \nu > 0$
$f(x) = \exp(-c_1 x) - \exp(-c_2 x)/x$	$-c_2 < -c_1 \leq 0$
$f(x) = (ax^\alpha + 1)^{-\beta}$	$a > 0, \quad 0 < \alpha \leq 1, \quad 0 \leq \beta \leq 1$

TABLE 2.1: Examples of the completely monotone functions

diverge. However, if we replace the condition $b < c$, i.e., the disjointness of the intervals, by the stronger admissibility condition (2.3), where

$$\eta = \left(1 + \frac{|x_{oi} - a|}{|c - b|}\right), \tag{2.14}$$

which is simply a constant independent of the order k, then a uniform bound for the approximation error is independent of the intervals as long as the admissibility condition is fulfilled.

We conclude, that the remainder of the truncated Taylor series, $|f(x_i, x_j) - \tilde{f}(x_i, x_j)|$, converges for any $k \in \mathbb{N}$ in the interval Q_τ if and only if $\operatorname{diam}(Q_\sigma) \leq \ \operatorname{dist}(Q_\sigma, Q_\tau)$, where $\operatorname{diam}(Q_\sigma) = |x_{0i} - a|$ and $\operatorname{dist}(Q_\sigma, Q_\tau) = |c - b|$. Therefore, we obtain uniform bound independently of the chosen intervals and the error decays exponentially with respect to the order k. Therefore, the error is small if the admissibility condition is satisfied and the diameter of the clusters is controlled.

We formulate now the Theorem 2.1 and give the general proof of the convergence of the Taylor series of completely monotone functions required to satisfy (2.5).

Theorem 2.1. *Any completely monotone in $0 < x < b$ function $f(x)$ can be represented by a convergent Taylor series.*

Proof. Consider Taylor series of the function $f(x)$ with the exact remainder around point $b > 0$

$$f(x) = \sum_{k=0}^{N} f^{(k)}(b) \frac{(x - b)^k}{k!} + R_N(x) \tag{2.15}$$

or

$$f(x) = T_N(x) + R_N(x).$$

We note that each term in (2.15) is positive for $x \in (0, b)$ due to the definition (1.13) and the fact that the polynomials $(x - b)^k$ have alternating signs. It means that partial sum $T_N(x)$ creates a monotone increasing sequence in the interval $(0, b)$. We need to define whether the remainder converges. The remainder formula for Taylor series expansion

$$R_N(x) = \frac{f^{(N+1)}(c)}{(N + 1)!}(x - b)^{(N+1)}$$

for some $c \in (0, b)$. We have $R_N(x) = f(x) - T_N(x) \geq 0$, thus the partial sums and the Taylor series converge in $(0, b)$. Since $f^{(N+2)}(x) < 0$, then $|f^{(N+1)}(x)|$ are decreasing functions. Thus if, for example, we consider x in the interval $(3b/4, b)$ with a point of expansion $b/2$, we have

$$\begin{aligned} R_N(x) &= \frac{f^{(N+1)}(c)}{(N+1)!}(x-b)^{(N+1)} \\ &\le \frac{|f^{(N+1)}(\frac{a+b}{2})|}{(N+1)!}(b/4)^{(N+1)}, \end{aligned}$$

where the quantity on the right converges to 0 with increasing N.

It follows that the series converges to $f(x)$ for $x \in (3b/4, b)$ and since b is arbitrary, $f(x)$ is analytic in a neighbourhood of the real axis. This completes the proof. □

As was mentioned, the fast convergence of the Taylor series guarantees the small error of the approximation of $|c(x_i, x_j) - \tilde{c}(x_i, x_j)|$ and, thus, the effective low-rank representation of a block of the matrix that satisfy the admissibility condition.

Another innovation of this thesis is the implementation of the $\mathcal{H}$-matrices in the spatio-temporal setting. Therefore, we now aim to find the admissibility condition based on its analytical derivation for spatio-temporal covariance functions for Gneiting (2002) and Ma (2003) type of spatio-temporal covariance functions.

We firstly discuss a spatio-temporal covariance function provided in Ma (2003). From a second order stochastic partial differential diffusion equation (1.9) Heine (1955) derived the space-time covariance function (1.10). However, as was mentioned before, Whittle (1962) noticed non-regular behaviour of the spatial variance derived from (1.9). Ma (2003) derived the space-time covariance function replacing the spatial (temporal) component of (1.10) by the spatial (temporal) variogram and obtained the following form

$$C(s,t) = e^{-\alpha|t|}\text{erfc}\left(\sqrt{\gamma(s)} - \frac{\alpha|t|}{\sqrt{\gamma(s)}}\right) + e^{\alpha|t|}\text{erfc}\left(\sqrt{\gamma(s)} + \frac{\alpha|t|}{\sqrt{\gamma(s)}}\right), \quad s,t \in \mathbb{R}^d \times \mathbb{R} \tag{2.16}$$

where $\gamma(s)$ is intrinsically stationary variogram on $\mathbb{R}^d$ and $\alpha > 0$.

As was discussed in Stein (2005), the derivative of the order $m \neq 0$ in space and k in time of (2.16) has the following form

$$D^{(m_l,k)}C(s,t) = \exp\left\{ -\gamma(s) - \frac{\alpha^2 t^2}{4\gamma(s)}\right\} \sum_l \lambda_l t^{\beta_l} \gamma(s)^{-\alpha_l - \frac{1}{2}} \prod_l \left\{D^{m_{lp}}\gamma(s)\right\}^{\delta_{lp}}, \tag{2.17}$$

where α_l, β_l and δ_{lp} are non-negative integers and $l = 1, \ldots, d$. According to the derived asymptotic smoothness condition of Iske *et al.* (2017), the last term of the right part in (2.17)

$$|D^{m_{lp}}\gamma(s)| \le c|m_{lp}|^b ||s||^{-|m_{lp}|-2a}$$

with the constants c, a and $b \in \mathbb{R}$.

Since we are interested in the asymptotic behaviour of the derivative (2.17), then as $t \to \infty$, $s \to \infty$ ($\gamma(s) \to 1$), the derivative with respect to t and s is a fast decaying function. Then for the convergence of the Taylor series of the covariance function (2.16) with the derivative (2.17), the following space-time condition can be obtained in $\mathbb{R}^d \times \mathbb{R}$

$$\operatorname{diam}\{Q_\sigma\} \cdot \operatorname{diam}\{P_\gamma\} \leq \eta \operatorname{dist}\{Q_\sigma, Q_\tau\} \cdot \operatorname{dist}\{P_\gamma, P_\delta\},$$

where η is a fixed parameter, data sites $s \in Q_\sigma \subset \mathbb{R}^d$, $s' \in Q_\tau \subset \mathbb{R}^d$ and timestamps $t \in P_\gamma \subset \mathbb{R}$ and $t' \in P_\delta \subset \mathbb{R}$, with P_γ, P_δ defined as in section 2.2.3.

A space-time admissibility condition for Gneiting class of spatio-temporal covariance functions (1.14) can be also obtained based on its analytical derivation. For example, for the specific type of (1.14), which is

$$C(s,t) = \frac{\sigma^2}{(|t-t'|/a+1)^{d/2}} \exp\left(-\frac{||s-s'||^2}{c(|t-t'|/a+1)}\right), \tag{2.18}$$

the Taylor series of (2.18) around the point of expansion $(s_0, t_0) \in \mathbb{R}^d \times \mathbb{R}$ converges if $|t - t_0| < |t' - t_0|/a + 1$. Therefore, the admissibility condition can take the form $\min\{\operatorname{diam}(P_\gamma), \operatorname{diam}(P_\delta)\} < \eta \operatorname{dist}(P_\gamma, P_\delta)$ for $t \in P_\gamma \subset \mathbb{R}$ and $t' \in P_\delta \subset \mathbb{R}$.

Since the derivation of aforementioned conditions could be cumbersome, next Chapter provides a general framework for the application of $\mathcal{H}$-matrices to covariance functions from GRF obtained as a solution to the Stochastic Partial Differential Equations (SPDEs).

2.3.2 Preserving symmetry and positivity of covariance

The technique of hierarchical matrices is not exact. The factorizations of all admissible blocks of an $\mathcal{H}$-matrix with the low-rank technique, such as ACA algorithm discussed in the section 2.2.5, may entail a symmetry loss, i.e., a symmetric covariance matrix C can be approximated by a non-symmetric $\tilde{C}$. However, this problem is easily solved by computing only the upper triangular part of an $\tilde{C}$ and using its transpose for the respective blocks in the lower triangular part.

While the problem of symmetry loss can be easily dealt with, the situation is different for positive definiteness property of a covariance. If C is positive definite, then $\tilde{C}$ will eventually be also positive definite since $\tilde{C} \to C$ when error $\varepsilon \to 0$. However, too small ε are impractical for the applications. As was discussed in the section 2.2.5, the error can also depend on the choice of the local rank k. Since the accuracy improves exponentially with increasing k, a moderate choice of k suffices to reach a given tolerance. However, the computational cost increases with k. It is cheaper to perform the calculations with small k with the lowered accuracy.

The method of hierarchical matrices is rather robust, see Bebendorf and Hackbusch (2007) for the detailed discussion. However, with the approximation by $\mathcal{H}$-matrices, the error can propagate and perturb the eigenvalues of the resulting matrix. If the smallest eigenvalue is close to the origin compared with the rounding accuracy ε, the result of these operations might become indefinite. Therefore, the Cholesky decomposition may fail. In fact, it is also a common situation for the exact covariances when the sampling locations are too close which leads to numerical singularities.

Litvinenko *et al.* (2019) also discussed this problem and proposed to use a block $\mathcal{H}$-Cholesky algorithm or use LDL^T factorization instead of Cholesky LL^T. Alternatively, one could simply add a small constant to the diagonal of $\tilde{C}$, albeit sacrificing approximation accuracy for the sake of positive definiteness. In the paper of Bebendorf and Hackbusch (2007), it is also proposed to modify the values on the diagonal. Then it can be guaranteed that a positive definite input matrix remains positive definite after approximations.

To conclude, it needs to be decided whether it is worth to pursue positive definiteness at the expense of accuracy and/or computational efficiency.

Chapter 3

Stochastic Partial Differential equations approach

3.1 Introduction

In Chapter 2 we provided the examples of analytically derived admissibility conditions for the specifically given non-separable spatial (spatio-temporal) covariance functions required for a fast convergence of the error of low-rank approximation of a covariance matrix blocks. The goal of this Chapter is to find a general regularity condition that is satisfied by a more general class of covariances functions.

Since $\mathcal{H}$-matrices method originates in Partial Differential Equations (PDEs) theory (as was mentioned before), the first part of the Chapter provides a description of the deterministic terms and tools involved into the PDEs theory. Sections 3.1.3 and 3.1.4 focus on generalized functions that are, in fact, objects acting on the test-functions that belong to the Schwartz space. Linear functionals over the Schwartz space or so called tempered distributions are detailed in section 3.1.5. Most of the distributions described in this Chapter do not refer to probabilistic distributions until we clearly state it.

We describe pseudodifferential operators defined through the Fourier transform and symbol functions which are measurable polynomially bounded functions. We also define convolutions with fast-decreasing distributions and provide an exchange formula. Finally we state the kernel representation theorem concerned with the representation of a pseudodifferential operator through its kernel and provide the link between kernels of pseudodifferential operators in deterministic framework and covariance functions derived from the Stochastic Partial Differential equations (SPDEs). Particularly, we exploit the results of Vergara *et al.* (2018) who found out that the existence of a stationary solution to a SPDE is equivalent to a slow-growing behaviour requirement of the multiplication between the squared-norm of the reciprocal of the symbol function and the spectral measure of the source term. As a source term, we take White noise which is defined as 'fundamental' case in Vergara *et al.* (2018).

We give general conditions under which $\mathcal{H}$-matrices technique can be successfully exploited in the SPDE framework for the spatial and spatio-temporal contexts. Namely, *we associate growth of the random spectral measure at infinity with the regularity of the covariance function which is required for the approximation by $\mathcal{H}$-methods.* Based on this, the bounds of the derivatives of the covariance functions coming from SPDEs can be easily obtained which in turn, correspond to the asymptotic smoothness condition defined in section 2.2.4 of Chapter 2.

3.1.1 The SPDE approach

The first theoretical result was obtained by Whittle (1963), where field with Matérn covariance was obtained as solution to a SPDE (1.4). Lindgren *et al.* (2011) then represented a GRF with Matérn covariance function as a Gaussian Markov Random Field (GMRF) through the link with SPDEs. There are multiple factors that encourage a spatial statistician to exploit SPDE approach. This includes a possibility to model effectively a wide class of natural phenomena which are difficult to take into account with the standard approaches in geostatistics. These phenomena include a wind speed, a weather forecast or a wave propagation with the parameters such as velocity vector, diffusivity coefficient or anisotropic diffusivity matrix as, for example, in (1.11).

The literature that we are interested in within this context includes Heine (1955), Whittle (1962), Sigrist *et al.* (2015) and Jones and Zhang (1997) discussed in section 1.1.3. A theoretical result of Vergara *et al.* (2018) which gives an explicit link between the covariance structures and its corresponding SPDEs is of main interest in this Chapter. We aim to demonstrate how distribution theory can be used to obtain the solutions of SPDEs through the application of the Fourier transform. Particularly, the first part of Chapter focuses on the deterministic tools involving partial differential equations. This will be needed to comprehend a link between the SPDEs and $\mathcal{H}$-matrices approach described in Chapter 2.

3.1.2 Notations and terminology

Let μ denote the Lebesgue measure on $\mathbb{R}^d$ and $L^p(\mathbb{R}^d)$ for $p \in [1, \infty)$ the set of complex valued Borel measurable functions on $\mathbb{R}^d$ which are absolutely summable with respect to Lebesgue measure. Then $L^p(\mathbb{R}^d)$ for $p \in [1, \infty)$ is a Banach space which is equipped with the following norm

$$||f||_{L^p(\mathbb{R}^d)} = \left(\int_{\mathbb{R}^d} |f(x)|^p d\mu(x) \right)^{\frac{1}{p}}. \tag{3.1}$$

In particular, $L^1(\mathbb{R}^d)$ is the space of absolutely integrable functions on $\mathbb{R}^d$ with the norm $||f||_{L^1(\mathbb{R}^d)} = \int_{\mathbb{R}^d} |f(x)|d\mu(x)$. If $p = \infty$, then $f \in L^\infty(\mathbb{R}^d)$ if it is measurable and bounded, i.e. if

$$||f||_{L^\infty(\mathbb{R}^d)} = \sup_{x\in\mathbb{R}^d} |f(x)| \tag{3.2}$$

is finite. Therefore, there exists a small positive p, such that $|f(x)| \leq p$ for all $x \in \mathbb{R}^d$.

The Riemann-Lebesgue Theorem says that if $f \in L^1(\mathbb{R}^d)$, then the Fourier transform $f(\xi)$ is a continuous function on $\mathbb{R}^d$ that vanishes at infinity, i.e $f(\xi) \to 0$ as $||\xi|| \to \infty$ for $\xi \in \mathbb{R}^d$.

For $p = 2$ in (3.1) it is a Hilbert space H with the inner product

$$\langle f, g\rangle_H = \int_{\mathbb{R}^d} f(x)\bar{g}(x)d\mu(x).$$

In this work we define a d-dimensional multi-index notation $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_d)$ of non-negative integers. For the multi-index $\alpha \in \mathbb{N}_0^d$ sum of the components or absolute value can be written as

$$|\alpha| = \alpha_1 + \alpha_2 + \cdots + \alpha_d \tag{3.3}$$

and higher-order partial derivatives as

$$\partial^\alpha = \partial_1^{\alpha_1} \partial_2^{\alpha_2} \ldots \partial_d^{\alpha_d},$$

where $\partial_i^{\alpha_i} = \partial^{\alpha_i}/\partial x_i^{\alpha_i}$ of the dimension d.

A complex function f defined on $\mathbb{R}^d$, is said to be a k-times continuously differentiable or C^k function if the first k derivatives with $|\alpha| \leq k$ exist and continuous with α defined in (3.3).

Definition 3.1. A function f in $\mathbb{R}^d$ is said to be a C^∞ function or infinitely differentiable function if it is a C^k function for all integers $k \in [0, \infty)$. The C^k functions in $\mathbb{R}^d$ $(0 \leq k \leq \infty)$ form a vector space over the field of complex numbers, which we shall denote by $C^\infty(\mathbb{R}^d)$.

3.1.3 Generalized functions

The focus of this section is on the objects that act over the smooth functions or so called test-functions ϕ. These objects are called distributions [1] or generalized functions and serve as linear and continuous functionals for every function of a certain space. The basic idea is that with the actions on a set of test functions ϕ, we can reconstruct the original function f.

For example, to any integrable real function $f(x)$ we can associate the linear functional T defined as

$$\phi(x) \to T(\phi) = \int_{\mathbb{R}^d} f(x)\phi(x)dx,$$

where f and ϕ are both integrable and f is allowed to grow at infinity arbitrarily fast. Then it becomes necessary that ϕ decays at infinity arbitrarily fast or vanishes outside some bounded set. For example, we can consider analytic functions with a high order of decay at infinity, or functions with compact support which are sufficiently smooth.

Within this approach, the operations on distributions are well defined which could not be a case for the underlying functions. For example, it is possible to find derivatives of a discontinuous function in case it is considered as a distribution. In the next section we give the description of the space of rapidly decreasing functions $S(\mathbb{R}^d)$ or Schwartz space.

3.1.4 Schwartz space

The aim of this section is to study the theory of distributions and some of its applications which was introduced by Schwartz (1966) whose goal was to define the derivative of a large class of objects. Our interest in this theory is motivated by the fact that the Schwartz theory has a direct connection to the $\mathcal{H}$-matrices approach and it will be demonstrated later in this Chapter.

We define $S(\mathbb{R}^d)$ as the space of all infinitely differentiable smooth, rapidly decreasing functions ϕ equipped with the Schwartz topology described also in Rudin (1973), such that

$$\sup_{x\in\mathbb{R}^d} |x^\beta(\partial^\alpha\phi)(x)| < \infty$$

for all multi-indices $\alpha, \beta \geq 0$, where x^β denotes the real number determined by $x^\beta = x_1^{\beta_1}x_2^{\beta_2}\dots x_d^{\beta_d}$. The norm on the Schwartz space $S(\mathbb{R}^d)$ can be defined as

$$\max_{|\alpha|+|\beta|\leq k} ||\varphi||_k = \sup_{x\in\mathbb{R}^d} |x^\beta\partial^\alpha\varphi(x)|.$$

[1] Here and after ‘distributions’ do not refer to the probabilistic distributions until clearly stated

Definition 3.2. A smooth function ϕ is in the Schwartz space $S(\mathbb{R}^d)$ if $\phi \in C^\infty(\mathbb{R}^d)$ and for all α and $N \geq 0$, there is a constant p_α such that $\forall x \in \mathbb{R}^d$

$$|\partial^\alpha \phi(x)| \leq p_\alpha (1 + ||x||^2)^{-N}. \tag{3.4}$$

The elements of the Schwartz space are the complex-valued functions ϕ which are defined and infinitely differentiable in $\mathbb{R}^d$. The growth (or rather, decrease) of ϕ is regulated at infinity, such that all their derivatives tend to zero at infinity faster than any power of $||x||^{-1}$.

This space is important in the theory of distributions and will be widely exploited in our framework in connection with the Fourier transformation. Obviously $S(\mathbb{R}^d) \subset C^\infty(\mathbb{R}^d)$ where the inclusions are strict. In particular, $S(\mathbb{R}^d) \subset L^1(\mathbb{R}^d)$, thus, the main benefit of the Schwartz space is that the Fourier transform performs properly in $S(\mathbb{R}^d)$. This will also lead to a new space of distributions known as the space of tempered distributions that we will discuss in the next section.

Definition 3.3. Let $\{\phi_n(x)\}_{n=1}^\infty$ be a sequence of functions in the Schwartz space $S(\mathbb{R}^d)$. We say that the sequence converges to 0 in $S(\mathbb{R}^d)$ if for every $k \in Z^+$, $\alpha \in \mathbb{N}_0^d$, the sequence $\{x^k D^\alpha \phi_n(x)\}_{n=1}^\infty$ converges to 0 uniformly.

Proposition 3.1. *If $\phi \in S(\mathbb{R}^d)$, then $x^\beta \partial^\alpha \phi \in S(\mathbb{R}^d)$.*

Proof. We denote $N_s(\phi) = \sup_{|c|,|g| \leq s}(x^c \partial^g \phi)$. The proof can be obtained from the fact that

$$N_p(x^\beta \partial^\alpha \phi) = \sup_{|c|,|g| \leq p} (x^c \partial^g (x^\beta \partial^\alpha \phi)) \leq p N_{p+q}(\phi)$$

with $|\alpha|, |\beta| \leq q$ and $q \geq 0$. □

For example, the function $\phi(x) = e^{-||x||^2}$ belongs to the Schwartz space $S(\mathbb{R}^d)$. This can be proved in the following way. Since $\partial^\alpha \phi = P_\alpha(x) e^{-||x||^2}$, where $|P_\alpha|$ is the polynomial bounded by $p_\alpha ||x||^k$. It is known that $|e^{-||x||^2}| \leq |1/(1 + ||x||^k/(k/2)!)|$. Hence, we have $\sup_{x \in \mathbb{R}^d} ||x||^\beta |\partial^\alpha \phi| \leq p_\alpha (k/2)!$. Thus, $f(x) = e^{-||x||^2}$ and all of its derivatives are rapidly decreasing in the Schwartz space and belong to $S(\mathbb{R}^d)$.

3.1.5 Tempered distributions

The dual space defined through a continuous linear functional, described in section 3.1.3, over the Schwartz space $S(\mathbb{R}^d)$ is denoted by $S'(\mathbb{R}^d)$ and called the space of tempered distributions. The term tempered also means that this is space of distributions of a slow growth. Following Treves (1967) we state the definition of tempered distributions.

Definition 3.4. A mapping $T : S(\mathbb{R}^d) \to \mathbb{C}$ is called a tempered distribution if it is linear such that for any m, k the following holds

$$T(m\phi + k\zeta) = mT(\phi) + kT(\zeta), \quad \text{for } \phi \in S(\mathbb{R}^d), \zeta \in S(\mathbb{R}^d)$$

and it is continuous so that

$$\{T(\phi_n)\} \to T(\phi) \in \mathbb{C}, \quad \text{for } \phi_n \to \phi \in S(\mathbb{R}^d).$$

Here and after through the angle brackets we define the functional $\langle f, \phi \rangle = \int f(x)\phi(x)dx$ for any $\phi \in S(\mathbb{R}^d)$. The scalar product $\langle f, \phi \rangle$ denotes the value of a functional $f \in S'(\mathbb{R}^d)$ at the test function $\phi \in S(\mathbb{R}^d)$, i.e.

$$\langle f, \phi \rangle = f(\phi), \quad \phi \in S(\mathbb{R}^d).$$

As an example, a Radon measure μ is called tempered if μ defines a tempered distribution. For example, any L^p-function or complex Radon measure of polynomial growth is a tempered Radon measure. Let $f : (\mathbb{R}^d) \to \mathbb{C}$ be a measurable function such that

$$\int_{\mathbb{R}^d} \frac{|f(x)|}{(1 + ||x||^2)^N} dx < \infty$$

for some integer $N > 0$. Then the function f is called a tempered function as was also defined in Wong (2014).

Theorem 3.1. *Let f be a tempered function defined on $\mathbb{R}^d$. Then the linear functional T_f on $S(\mathbb{R}^d)$ defined by*

$$T_f(\phi) = \int_{\mathbb{R}^d} f(x)\phi(x)dx, \quad \text{for all } \phi \in S(\mathbb{R}^d)$$

is a tempered distribution. For the proof see Proposition 4.5 of Wong (2014).

If $f : \mathbb{R}^d \to \mathbb{C}$ is a measurable function such that $(x)^{-N} f(x) \in L^1(\mathbb{R}^d)$ for some $N \in \mathbb{N}$, then f is identified with a tempered distribution $T_f \in S'(\mathbb{R}^d)$ defined by

$$\langle T_f, \phi \rangle = \int_{\mathbb{R}^d} f(x)\phi(x)dx, \quad \phi \in S(\mathbb{R}^d).$$

Definition 3.5. Let (T_n) be a sequence of tempered distributions. We say that (T_n) tends to T in $S'(\mathbb{R}^d)$, when for any function $\phi \in S(\mathbb{R}^d)$, it holds that $\langle T_n, \phi \rangle \to \langle T, \phi \rangle$.

As an example, the Schwartz functions are infinitely smooth and locally integrable, so each of them defines a tempered distribution. Another example is a Dirac delta function which is in fact the distribution centered at the point x of $\mathbb{R}^d$ and defined by

$$\delta(\phi)(x) = \delta_x(\phi) = \phi(x).$$

In other words, $\delta_x(\phi)$ associates to the test function $\phi(x)$ its value at the point x. The functional δ_x is well defined and linear. It is also continuous since for a sequence $\phi_n \to \phi$, $|\delta_x(\phi_n) - \delta_x(\phi)| = |\phi_n(x) - \phi(x)|$ converges to 0 as $n \to \infty$, ϕ_n converges uniformly and pointwise to ϕ. In addition, δ_x is well defined for $\phi \in S(\mathbb{R}^d)$ and it is therefore a tempered distribution.

One of the most important properties of distributions is that they can be differentiated infinitely many times (but in a generalized sense). The differential operator $\partial^\alpha : S'(\mathbb{R}^d) \to S'(\mathbb{R}^d)$ is defined as

$$\langle \partial^\alpha u, \phi \rangle = (-1)^{|\alpha|} \langle u, \partial^\alpha \phi \rangle, \quad \text{for } u \in S'(\mathbb{R}^d), \phi \in S(\mathbb{R}^d).$$

In general, every tempered distribution is the derivative of large order of a polynomially bounded continuous function (Reed and Simon (1972)).

3.1.6 Fourier transform

Let $f \in L^1(\mathbb{R}^d)$. The Fourier transform $\mathcal{F}f(\cdot)$ is defined as

$$\mathcal{F}f(\xi) = \frac{1}{(2\pi)^{d/2}} \int_{\mathbb{R}^d} e^{-i\xi^T x} f(x) dx \tag{3.5}$$

for $\xi \in \mathbb{R}^d$. And the inverse Fourier transform

$$\mathcal{F}^{-1}f(\xi) = \frac{1}{(2\pi)^{d/2}} \int_{\mathbb{R}^d} e^{i\xi^T x} f(x) dx. \tag{3.6}$$

It is known that the Fourier transform defines the map, i.e. $\mathcal{F} : L^1(\mathbb{R}^d) \to L^\infty(\mathbb{R}^d)$ and from the Lebesgue dominated convergence theorem, this map is continuous.

The classical property of the Fourier transformation with respect to kth derivative $f^{(k)}$ is given by

$$\mathcal{F}f^{(k)}(\xi) = (i\xi)^k \mathcal{F}f(\xi) \tag{3.7}$$

and will be widely used throughout this Chapter.

We also define the Fourier transform in the Schwartz space described in the previous section. Let $\phi \in S(\mathbb{R}^d)$, then the Fourier transform

$$\mathcal{F}\phi(\xi) = \frac{1}{(2\pi)^{d/2}} \int_{\mathbb{R}^d} e^{-i\xi^T x} \phi(x) dx$$

for $\xi \in \mathbb{R}^d$. And the inverse Fourier transform

$$\mathcal{F}^{-1}\phi(\xi) = \frac{1}{(2\pi)^{d/2}} \int_{\mathbb{R}^d} e^{i\xi^T x} \phi(x) dx.$$

The Fourier transform over the space of tempered distributions $S'(\mathbb{R}^d)$ (see Definition 3.4) can be defined as follows

$$\langle \mathcal{F}(u), \phi \rangle = \langle u, \mathcal{F}(\phi) \rangle, \quad \text{for } u \in S'(\mathbb{R}^d), \phi \in S(\mathbb{R}^d).$$

As was stated in Treves (1967), the Fourier transform $\mathcal{F}$ maps $S(\mathbb{R}^d)$ to $S(\mathbb{R}^d)$ and this map is continuous. In addition, the Fourier transform defines a map from $S'(\mathbb{R}^d)$ to $S'(\mathbb{R}^d)$.

3.1.7 Space of convoluters and multiplicators

Consider the operator R and τ_h defined as $(R\zeta)(x) = \zeta(-x)$ and $(\tau_h \zeta)(x) = \zeta(x-h)$. We provide the definition of the convolution between two functions $\phi, \zeta \in S(\mathbb{R}^d)$. Since

$$(\tau_x R\zeta)(y) = (R\zeta)(y - x) = \zeta(x - y),$$

then

$$(\phi * \zeta)(x) = \int_{\mathbb{R}^d} \phi(y)\zeta(x - y)dy = \int_{\mathbb{R}^d} \phi(y)(\tau_x R\zeta)(y)dy.$$

Therefore, for the distributions we can write

$$(\phi * \zeta)(x) = \phi(\tau_x R\zeta),$$

where $\tau_x R\phi \in S(\mathbb{R}^d)$ and $\tau_x, R : S(\mathbb{R}^d) \to S(\mathbb{R}^d)$ are continuous.

For $T \in S'(\mathbb{R}^d)$ we have

$$\delta * T = T,$$

where δ is a Dirac distribution defined in section 3.1.4. This follows from the fact that $\langle T * \delta, \zeta \rangle = \langle T, \zeta * \delta \rangle$, but $(\zeta * \delta)(x) = \langle \delta_y, \tau_x R\zeta(y) \rangle = \zeta(x)$.

Also $\delta_a * T = \tau_\alpha T$, i.e. translation of the distribution $T \in S'(\mathbb{R}^d)$ by $x + a$. In addition, it holds that $\tau_\alpha(S * T) = \tau_\alpha(S) * T = \tau_\alpha(T) * S$.

Definition 3.6. Let space of multiplicators or shortly $O_M(\mathbb{R}^d)$ denote the set of infinitely differentiable functions on $\mathbb{R}^d$ which together with their derivatives are polynomially bounded. If $f \in O_M(\mathbb{R}^d)$, then f is $C^\infty(\mathbb{R}^d)$ and for each multi-index $\alpha \in \mathbb{N}_0^d$, there is an N and p depending on α such that

$$|(\partial^\alpha f)(x)| \leq p_\alpha[1 + ||x||^2]^N, \tag{3.8}$$

where $O_M(\mathbb{R}^d) \subset S'(\mathbb{R}^d)$ can be also defined as the space of functions slowly increasing at infinity, as well as each of its derivatives, whereas functions of Schwartz space formulated in Definition 3.2 are fast decreasing smooth functions.

If $f \in O_M(\mathbb{R}^d)$ and $\phi \in S(\mathbb{R}^d)$, then $f\phi \in S(\mathbb{R}^d)$. The multiplication of $f \in O_M(\mathbb{R}^d)$ with $T \in S'(\mathbb{R}^d)$ is defined as a distribution

$$\langle fT, \phi \rangle = \langle T, f\phi \rangle$$

and belongs to $S'(\mathbb{R}^d)$. For the details see Treves (1967).

Definition 3.7. The space of the convoluters of tempered distributions or $O_c(\mathbb{R}^d)$ is the space of distributions $T \in S'(\mathbb{R}^d)$ defined with the following property that for $\forall q$: $x^q T$ is a bounded distribution or finite sum of derivatives of continuous and bounded functions.

If $S \in O_c(\mathbb{R}^d)$ and $\phi \in S(\mathbb{R}^d)$, then $S * \phi \in S(\mathbb{R}^d)$, and $\phi \to S * \phi$ is a continuous linear operator from $S(\mathbb{R}^d)$ to $S(\mathbb{R}^d)$ as defined in Treves (1967).

Another important property is that Fourier transform $\mathcal{F}$ exchanges the space of convoluters $O_c(\mathbb{R}^d)$ with the space of multiplicators $O_M(\mathbb{R}^d)$ which is stated in the Theorem 3.2.

Theorem 3.2. *Fourier Transformation defines a linear map $\mathcal{F} : O_c(\mathbb{R}^d) \to O_M(\mathbb{R}^d)$ and $\mathcal{F} : O_M(\mathbb{R}^d) \to O_c(\mathbb{R}^d)$.*

For the proof see Theorem 30.3 of Treves (1967).

Theorem 3.3. *Let $\alpha \in O_M(\mathbb{R}^d)$, $T \in S'(\mathbb{R}^d)$ and $S \in O_c(\mathbb{R}^d)$, then*

$$\mathcal{F}(\alpha T) = (2\pi)^{\frac{d}{2}} \mathcal{F}\alpha * \mathcal{F}T, \tag{3.9}$$

$$\mathcal{F}(S * T) = (2\pi)^{-\frac{d}{2}} \mathcal{F}(S)\mathcal{F}(T). \tag{3.10}$$

For the proof see Theorem 30.4 of Treves (1967). We further state the regularity theorem for distributions.

Theorem 3.4. *Let $T \in S'(\mathbb{R}^d)$. Then T is the derivative of a large power of the polynomial f, i.e. $T = D^\alpha f$ for $\alpha \in \mathbb{N}_0^d$ and*

$$T(\phi) = \int (-1)^{|\alpha|} f(x)(D^\alpha \phi)(x) dx$$

for all $\phi \in S(\mathbb{R}^d)$.

Therefore the growth of a tempered distribution $T \in S'(\mathbb{R}^d)$ is related to the decrease restrictions imposed on functions $\phi \in S(\mathbb{R}^d)$. The proof can be found in Reed and Simon (1972).

3.1.8 Pseudodifferential Partial Differential equations

This section provides the definition of pseudodifferential operators which are important when treating differential and integral operators. It is common that any pseudodifferential operator can be characterized by so-called *symbol function* as described in Taylor (1991). Therefore, in this section we give the definition of symbol functions that are associated with pseudodifferential operators, such that any operations on the pseudodifferential operators can be defined through its symbol functions. Symbols are based on the use of the Fourier transformations (3.5) and its inverse (3.6) and, thus, are also known as Fourier multipliers.

We begin by recalling the definition of a linear partial differential operator $\mathcal{L}(x, D)$ on $\mathbb{R}^d$ given by

$$\mathcal{L}(x, D) = \sum_{|\alpha| \leq r} a_\alpha(x) D_x^\alpha, \tag{3.11}$$

which is a polynomial in the derivatives D_x^α with the constants $a_\alpha(x)$ defined on $\mathbb{R}^d$ and r is a positive integer. If we replace D_x^α in (3.11) by the monomial ξ^α in $\mathbb{R}^d$, then we obtain a symbol function $\sigma(x, \xi) : \mathbb{R}^d \to C$, i.e.

$$\sigma(x, \xi) = \sum_{|\alpha| \leq r} a_\alpha(x) ||\xi||^\alpha \tag{3.12}$$

of the operator (3.11) which is a polynomial in the phase variable $\xi \in \mathbb{R}^d$ with the constants $a_\alpha(x)$ that depend on the space variable x.

The natural question which arises now is how differential operator $\mathcal{L}(x, D)$ relates to its symbol function $\sigma(x, \xi)$. Thus, following the property stated in (3.7), we represent a pseudodifferential operator $\mathcal{L}(x,D)$ on the function $\phi(x) \in S(\mathbb{R}^d)$ in terms of its symbol function in the following way

$$\begin{aligned}(\mathcal{L}(x,D)\phi)(x) &= \sum_{|\alpha|\leq r} a_\alpha(x)(D^\alpha\phi)(x)\\ &= \sum_{|\alpha|\leq r} a_\alpha(x)\mathcal{F}^{-1}(\mathcal{F}(D^\alpha\phi))(x)\\ &= \sum_{|\alpha|\leq r} a_\alpha(x)\mathcal{F}^{-1}((i\xi)^\alpha(\mathcal{F}\phi))(x)\\ &= \sum_{|\alpha|\leq r} a_\alpha(x)\frac{1}{(2\pi)^{d/2}}\int_{\mathbb{R}^d}(i\xi)^\alpha e^{ix\cdot\xi}\mathcal{F}\phi(\xi)d\xi\\ &= \frac{1}{(2\pi)^{d/2}}\int_{\mathbb{R}^d} e^{ix\cdot\xi}\sigma(x,\xi)\mathcal{F}\phi(\xi)d\xi, \end{aligned} \tag{3.13}$$

where $\mathcal{F}\phi(\xi)$ is the Fourier transform of ϕ with the map $S(\mathbb{R}^d) \to S(\mathbb{R}^d)$, i.e. $\mathcal{F}\phi(\xi) = \mathcal{F}_{x\to\xi}\phi(x)$.

Therefore, we have represented the partial differential operator $\mathcal{L}(x,D)$ in terms of its symbol by means of the Fourier transform. It we replace the symbol $\sigma(x,\xi)$ by a more general symbol function $p(x,\xi)$ which is no longer polynomial in ξ, then we obtain so called pseudodifferential operator.

Definition 3.8. Let $p(x,\xi) : \mathbb{R}^d \to \mathbb{C}$ be a symbol function. Then the pseudodifferential operator $\mathcal{L}(x,D)$ associated to $p(x,\xi)$ is defined by

$$(\mathcal{L}_p(x,D)\phi)(x) = (2\pi)^{-\frac{d}{2}}\int_{\mathbb{R}^d} e^{ix\cdot\xi}p(x,\xi)\mathcal{F}\phi(\xi)d\xi \tag{3.14}$$

and denoted by $\mathcal{L}_p$.

Hence application of $\mathcal{L}_p$ to ϕ acts as multiplication of $\mathcal{F}\phi(\xi)$ by the symbol $p(x,\xi)$. This means that, if $\mathcal{L}_p$ is bounded and invertible operator, the inversion of $\mathcal{L}_p$ corresponds to the multiplication of $p(x,\xi)^{-1}$ on the Fourier transformed functions.

To get tractable class of operators, it is necessary to impose certain conditions on the functions $p(x,\xi)$ which gives rise to the definition of a symbol class.

Definition 3.9. For $r \in \mathbb{R}$ we define the symbol class T^r that consists of the set of infinitely differentiable smooth functions $p \in C^\infty(\mathbb{R}^d \times \mathbb{R}^d)$ with the constant coefficients such that for multi-indices α, $\beta \in \mathbb{N}_0^d$ there exists a positive constant g_α and the following holds

$$\left|\left(\frac{\partial}{\partial\xi}\right)^\alpha p(x,\xi)\right| \leq g_\alpha(1+||\xi||^2)^{\frac{1}{2}(r-|\alpha|)} \tag{3.15}$$

for any $\xi \in \mathbb{R}^d$.

We call these elements as symbols of order r.

A differential operator $\mathcal{L}_p$ is called elliptic, if $p(x,\xi) \neq 0$ for $(x,\xi) \in \mathbb{R}^d \times \mathbb{R}^d \backslash 0$. The symbol $p(x,\xi) \in T^r$ is elliptic of order r, if there exists some $R \geq 0$, such that $p(x,\xi)$ is invertible for all (x,ξ) with $||\xi|| \geq R$ and

$$|p(x,\xi)| \leq c(1+||\xi||^2)^{\frac{r}{2}}, \quad {}^{\xi\in\mathbb{R}^d}. \tag{3.16}$$

If $\phi \in S(\mathbb{R}^d)$, then $\mathcal{F}\phi \in S(\mathbb{R}^d)$ and therefore $p(x,\xi)\mathcal{F}\phi(\xi) \in S(\mathbb{R}^d)$ with respect to ξ for every fixed $x \in \mathbb{R}^d$. Therefore, the integral in (3.13) exists and $\mathcal{L}(x,D)\phi$ is well-defined.

Let $p(x,\xi) = \sum_{|\alpha|\leq r} a_\alpha(x)\xi^\alpha$ be a polynomial in ξ of order r with smooth coefficients $a_\alpha(x) \in C^\infty(\mathbb{R}^d)$. Then $p \in T^r$ and

$$\mathcal{L}_p(x,D)\phi = \sum_{|\alpha|\leq r} a_\alpha(x) D_x^\alpha \phi$$

for every $\phi \in S(\mathbb{R}^d)$. Therefore every linear differential operator with smooth coefficients is a pseudodifferential operator.

As an example, the Laplacian $\triangle = \partial_1^2 + \cdots + \partial_d^2$ is a pseudodifferential operator with the symbol $-||\xi||^2 = -\xi_1^2 - \cdots - \xi_d^2$.

3.1.9 Tensor product

In this section we introduce the definition of tensor product between the functions which can be extended to the case of distributions defined through the tensor products with the test-functions $\phi \in S(\mathbb{R}^d)$. This was firstly described by Schwartz (1966) before he announced the kernel theorem which will be discussed in the following section.

Definition 3.10. The tensor product between two functions $f : \mathbb{R}^d \to C$ and $g : \mathbb{R}^n \to C$ is the function $(f \otimes g) : \mathbb{R}^d \times \mathbb{R}^n \to C$ defined as

$$(f \otimes g)(x,y) = f(x)g(y)$$

for all $x \in \mathbb{R}^d$ and $y \in \mathbb{R}^n$.

Let two distributions $T_1 \in S'(\mathbb{R}^d)$ and $T_2 \in S'(\mathbb{R}^n)$. The tensor product between two distributions is a tempered distribution $T_1 \otimes T_2 \in S'(\mathbb{R}^d \times \mathbb{R}^n)$ on the dn-dimensional space such that

$$\langle \phi \otimes \zeta, T_1 \otimes T_2 \rangle = \langle \phi, T_1 \rangle \langle \zeta, T_2 \rangle$$

for all $\phi \in S(\mathbb{R}^d)$ and $\zeta \in S(\mathbb{R}^n)$.

Let $\mathcal{L} : S(\mathbb{R}^d) \to S(\mathbb{R}^d)$ be a continuous and linear operator over the Schwartz space. The adjoint operator $\mathcal{L}^*$ is the linear operator over the space of tempered distributions $\mathcal{L}^* : S'(\mathbb{R}^d) \to S'(\mathbb{R}^d)$ defined through

$$\langle \mathcal{L}^* T, \phi \rangle = \langle T, \mathcal{L}\phi \rangle,$$

which means that the adjoint operator perform the composition denoted by the symbol $\circ$, that is $\mathcal{L}^* T = T \circ \mathcal{L}$ for every $T \in S'(\mathbb{R}^d)$.

In addition, the operator is called self-adjoint if it satisfies the following

$$\langle \phi, \mathcal{L}\zeta \rangle = \langle \mathcal{L}\phi, \zeta \rangle$$

for all functions $\phi, \zeta \in S(\mathbb{R}^d)$.

Definition 3.11. We denote by $\mathcal{L}_1 : S'(\mathbb{R}^d) \to S'(\mathbb{R}^d)$ and $\mathcal{L}_2 : S'(\mathbb{R}^n) \to S'(\mathbb{R}^n)$ two linear and continuous operators. Then the tensor product $\mathcal{L}_1 \otimes \mathcal{L}_2 : S'(\mathbb{R}^d \times \mathbb{R}^n) \to S'(\mathbb{R}^d \times \mathbb{R}^n)$ is linear and continuous operator such that

$$(\mathcal{L} \otimes \mathcal{L})(\phi) = (I_d \otimes \mathcal{L}_2)(\mathcal{L}_1 \otimes I_n)(\phi)$$

for a test function $\phi \in S(\mathbb{R}^d \times \mathbb{R}^n)$ and identity operators I_d and I_n.

For two test-functions $\phi \in S(\mathbb{R}^d)$, $\zeta \in S(\mathbb{R}^n)$ and linear operators $\mathcal{L}_1$, $\mathcal{L}_2$, the tensor product can be defined as follows $(\mathcal{L}_1 \otimes \mathcal{L}_2)(\zeta \otimes \phi) = \mathcal{L}_1 \zeta \otimes \mathcal{L}_2 \phi$.

3.1.10 Kernel representation of pseudodifferential operator

We now formulate the Theorem 4.1 which was stated by Schwartz (1966).

Theorem 3.5. *Every continuous linear map $\mathcal{K}$ of the space of test-functions $S(\mathbb{R}^d)$ in some variable x into the space $S'(\mathbb{R}^d)$ of distributions in a second variable y is given by an unique distribution $k \in S'(\mathbb{R}^d \times \mathbb{R}^d)$ in both variables x and y, such that*

$$\langle \mathcal{K}\phi, \zeta \rangle_{S(\mathbb{R}^d)} = \langle k, \phi \otimes \zeta \rangle_{S(\mathbb{R}^d \times \mathbb{R}^d)}, \quad \textit{for all } \phi, \zeta \in S(\mathbb{R}^d), \tag{3.17}$$

where $(\phi \otimes \zeta)(x, y) = \phi(x)\zeta(y)$ and $\langle f, \phi \rangle_{S(\mathbb{R}^d)} = f(\phi)$ is the duality product of $f \in S'(\mathbb{R}^d)$ and $\phi \in S(\mathbb{R}^d)$.

The analog of this theorem called Nuclear Theorem can be found in Reed and Simon (1972) and Treves (1967). The result of the theorem can also be expressed in the integral form

$$\mathcal{K}\phi = \int_{\mathbb{R}^d} k(x,y)\phi(x)dx, \quad \text{for } \phi \in S(\mathbb{R}^d), \zeta \in S(\mathbb{R}^d),$$

where $k \in S'(\mathbb{R}^d \times \mathbb{R}^d)$ is called kernel of the operator $\mathcal{K}$.

It follows that this result can be applied for pseudo-differential operators since $\mathcal{L}_p(x,D) : S(\mathbb{R}^d) \to S(\mathbb{R}^d) \subset S'(\mathbb{R}^d)$ for the symbol function $p(x,\xi) \in T^r$ of the arbitrary order $r \in \mathbb{R}$ as in (3.15).

The pseudodifferential operator given in (3.13) can be rewritten through the integration in y and then in ξ as follows

$$\begin{aligned}(\mathcal{L}(x,D)\phi)(x) &= \int_{\mathbb{R}^d}\int_{\mathbb{R}^d} e^{i(x-y)\cdot\xi} p(x,\xi)\phi(y)dyd\xi \\ &= \int_{\mathbb{R}^d}\int_{\mathbb{R}^d} e^{ih\cdot\xi} p(x,\xi)\tau_h\phi(x)dhd\xi \\ &= \int_{\mathbb{R}^d} k'(x,h)\phi(x-h)dh \\ &= \int_{\mathbb{R}^d} k'(x,x-y)\phi(y)dy\end{aligned}$$

by Fubini theorem for all $\phi \in S(\mathbb{R}^d)$ and $h = x - y$, where $k'(x,h) = \mathcal{F}^{-1}_{\xi\to h} p(x,\xi)$ is continuous. We further provide the bounds of the Schwartz kernel of pseudodifferential operator which will be required for the successful application of the $\mathcal{H}$-matrices.

Proposition 3.2. *Let symbol function $p(x,\xi)$ belong to the symbol class T^r as defined in* (3.15)*. Then a kernel $k(x,h) \in C^\infty$ is a smooth function which rapidly decreases with all the derivatives as $||h|| \to \infty$, i.e. it satisfies*

$$|\partial_h^\beta k(x,h)| \le c||h||^{-k} \tag{3.18}$$

for all $x \in \mathbb{R}^d$, a constant c and $k > r + d + |\beta|$.

Proof. Since kernel $k(x,h)$ (as was mentioned above) is derived through the Fourier transform of a symbol function $p(x,\xi)$, therefore it is straightforward to define the following link

$$(2\pi i h)^\alpha \partial_h^\beta k(x,h) = \mathcal{F}^{-1}_{\xi\to h}(\partial_\xi^\alpha[(2\pi i\xi)^\beta p(x,\xi)])(h). \tag{3.19}$$

We note that $(2\pi i\xi)^\beta p(x,\xi) \in T^{r+|\beta|}$, so that it is a symbol function of order $r + |\beta|$. Therefore, according to the Definition 3.9 we may write

$$\left|\partial_\xi^\alpha[(2\pi i\xi)^\beta p(x,\xi)]\right| \le c_{\alpha\beta}(1 + ||\xi||^2)^{\frac{1}{2}(r+|\beta|-\alpha)}. \tag{3.20}$$

We then conclude that $\partial_\xi^\alpha[(2\pi i\xi)^\beta p(x,\xi)]$ is in $L^1(\mathbb{R}^d)$ for ξ if $r+|\beta|<\alpha-d$ or $\alpha>r+d+|\beta|$ according to the integrability criterion with respect to ξ.

From the section 3.1.6 we know that the Fourier transform defines the map $\mathcal{F}^{-1}$: $L^1(\mathbb{R}^d)\to L^\infty(\mathbb{R}^d)$. Thus, $\mathcal{F}^{-1}_{\xi\to h}(\partial_\xi^\alpha[(2\pi i\xi)^\beta p(\cdot,\xi)])(h)$ in (3.19) is in $L^\infty(\mathbb{R}^d)$, where the norm of the function is given by the essential supremum of the function (3.20) as in (3.2). Therefore, the left part of (3.19) is bounded by the constant $c_{\alpha\beta}$ and

$$\partial_h^\beta k(x,h)\leq c_{\alpha\beta}||h||^{-\alpha}$$

for $\alpha>r+d+|\beta|$ and since derivatives of $k(x,h)$ with respect to x are not involved, this implies (3.18). □

As the order $r\to-\infty$, the kernel $k(x,h)\in C^\infty$ becomes smoother and decrease rapidly with all the derivatives as $||h||\to\infty$. Recall the definition of the asymptotically smooth condition given in (2.4). Comparing (2.4) with the bounds obtained in (3.18), we conclude that the bounds of the derivatives of the Schwartz kernel correspond to the bounds stated in the condition (2.4).

Finally, the kernel $k\in S'(\mathbb{R}^d\times\mathbb{R}^d)$ does not have to be continuous as was already discussed in the beginning of this Chapter. For example, if symbol function $p(x,\xi)=1$, then we have

$$\langle k,\phi(x)\zeta(y)\rangle_{S(\mathbb{R}^d\times\mathbb{R}^d)}=\int_{\mathbb{R}^d}\phi(x)\zeta(x)dx, \tag{3.21}$$

so that $k(x,y)=\delta_0(x-y)$ and $(\mathcal{L}(x,D)\phi)(x)=\delta_0*\phi(x)=\phi(x)$.

The second part of the Chapter will focus on the PDEs formulated in the stochastic framework with the Generalized Random Fields involved.

3.2 Stochastic framework

The second part of this Chapter focuses on the interpretation of the deterministic PDEs in the stochastic framework and random generalised functions defined through the mean and covariance structure. Within the stochastic framework we aim to interpret a variable as the realization of a random process that satisfies a Stochastic Partial Differential Equation (SPDE). We consider a stationary real random function over $\mathbb{R}^d$ which is a family of squared-integrable real random variables $Z=\{Z(x),x\in\mathbb{R}^d\}$.

The mean function is $m(x)=\mathbb{E}(Z(x))$ and the covariance is defined as $C_Z(x_1,x_2)=\mathrm{Cov}(Z(x_1),Z(x_2))$ as was stated in Chapter 1. From now on, we define Z as a stationary process with a constant mean and a covariance function that depends only on the spatial lag, i.e $\mathrm{Cov}(Z(x_1), Z(x_2)) = c(||x_1 - x_2||)$. In this section we explore mean and covariance functions derived from SPDEs which can be defined as PDEs with the random objects involved.

Recall that the covariance function should be a positive definite function (see Chapter 2) and recall the Bochner's theorem given in (1.1): a continuous function $c : \mathbb{R}^d \to \mathbb{R}$ is positive definite if and only if it is a Fourier trasform of a positive finite measure μ

$$c(x) = (2\pi)^{-d/2} \int_{\mathbb{R}^d} e^{-ix^T\xi} d\mu(\xi), \quad \text{for } \forall x \in \mathbb{R}^d.$$

The measure μ that satisfies $c = \mathcal{F}(\mu_Z)$ is called the spectral measure of the process Z. Using this relation and dominated convergence theorem it can be concluded that c is N times continuously differentiable if and only if the measure $||\xi||^N d\mu(\xi)$ is finite.

For example, the spectral measure of the Matérn covariance is

$$d\mu(\xi) = \frac{d\xi}{(2\pi)^{d/2}(\kappa^2 + ||\xi||^2)^\alpha},$$

where the parameter α controls the regularity of the Matérn covariance and κ is defined as in (1.4).

3.2.1 Generalized Random Fields

The notion of Generalized Random Fields (GeRF) has been introduced in the 50s by Gelfand and Shilov (1968). The GeRF is a collection of real random variables that is indexed by a class of functions and that satisfies linearity and some regularity conditions. The generalized random function in this context is characterised by a generalized mean and covariance distributions, and the linearity and continuity properties are interpreted in a mean-square sense.

As was mentioned before, in this work we exploit the Schwartz space $S(\mathbb{R}^d)$ of the test-functions ϕ. Therefore, in our context the GeRFs are defined as stochastic processes indexed by a space of test-functions. Any mean-square tempered random distribution over $\mathbb{R}^d$ is a real and continuous linear operator from $S(\mathbb{R}^d) \to L^2(\Omega, \mathcal{A}, \mathbb{P})$, i.e. Z serves as a linear functional over the test functions $\phi \in S(\mathbb{R}^d)$. Thus as was defined in Vergara *et al.* (2018) we can write $\langle Z, \phi \rangle = Z(\phi)$ which will be characterised by mean $m_Z \in S'(\mathbb{R}^d)$

$$\langle m_Z, \phi \rangle = m(\phi) = \mathbb{E}(\langle Z, \phi \rangle)$$

for any $\phi \in S(\mathbb{R}^d)$. Define $k : S(\mathbb{R}^d) \times S(\mathbb{R}^d) \to \mathbb{C}$ such that

$$k(\phi, \zeta) = \mathrm{Cov}(\langle Z, \phi \rangle, \langle Z, \zeta \rangle)$$

for any $\phi, \zeta \in S(\mathbb{R}^d)$. A kernel k is a positive definite if $k(\phi, \phi) \geq 0$ for all $\phi \in S(\mathbb{R}^d)$. Since $\phi \to k(\phi, \zeta)$ is continuous linear functional in $S'(\mathbb{R}^d)$ and $\zeta \to k(\phi, \zeta)$ is continuous linear functional in $S'(\mathbb{R}^d)$, by Theorem 4.1 there exists unique tempered distribution $C_Z \in S'(\mathbb{R}^d \times \mathbb{R}^d)$ such that

$$C_Z(\phi \otimes \zeta) = k(\phi, \zeta) = \mathrm{Cov}(\langle Z, \phi \rangle, \langle Z, \zeta \rangle). \tag{3.22}$$

Definition 3.12. The measure μ is a slow-growing random measure as was stated by Vergara *et al.* (2018) if there exists a strictly positive polynomial p such that the measure $\frac{1}{p}\mu$ is finite, i.e. if there is $N \in \mathbb{N}$ such that $\int_{\mathbb{R}^d} (1 + ||x||^2)^{-N} d\mu(x) < \infty$, since for every polynomial p, there exist $N \in \mathbb{N}$ such that $|p(x)| \leq k(1 + ||x||^2)^N$ for any positive constant k and $x \in \mathbb{R}^d$. The set of all slow-growing complex measures over $\mathbb{R}^d$ is denoted by $\mathcal{M}_{SG}(\mathbb{R}^d)$.

Therefore, a GeRF Z is a slow-growing random measure if and only if its covariance distribution $C_Z \in M_{SG}(\mathbb{R}^d \times \mathbb{R}^d)$, i.e a slow-growing measure. We may also define a measure μ over $\mathbb{R}^d$ indexed by the set of $\phi \in S(\mathbb{R}^d)$ as a linear functional through the integral $\langle \mu, \phi \rangle$ with the mean and covariance structures written as follows

$$\mathbb{E}(\langle \mu, \phi \rangle) = \langle m_\mu, \phi \rangle, \quad \mathrm{Cov}(\langle \mu, \phi \rangle, \langle \mu, \phi \rangle) = \langle C_\mu, \phi \otimes \phi \rangle. \tag{3.23}$$

Finally, we can consider a linear continuous operator $\mathcal{L} : S'(\mathbb{R}^d) \to S'(\mathbb{R}^d)$ and Z as a real GeRF, then

$$\langle \mathcal{L}Z, \phi \rangle = \langle Z, \mathcal{L}^* \phi \rangle$$

for any $\phi \in S(\mathbb{R}^d)$ and the adjoint operator $\mathcal{L}^*$. This is a well-defined operation since $\mathcal{L}^* \phi \in S(\mathbb{R}^d)$ and $\mathcal{L}Z$ serves as a linear map from $S(\mathbb{R}^d)$ to $L^2(\Omega, \mathcal{A}, \mathbb{P})$.

3.2.2 Stochastic Partial Differential equations

We consider a Stochastic Partial Differential equation with the pseudodifferential operator characterised by a symbol function $p(\cdot, \xi) : \mathbb{R}^d \to \mathbb{C}$ which is polynomially bounded measurable function defined as in (3.12)

$$\mathcal{L}_p Z = U, \tag{3.24}$$

where the solution $Z = \{Z(x), x \in \mathbb{R}^d\}$ and the source term U are GeRFs described in the previous sections and $\mathcal{L} : S'(\mathbb{R}^d) \to S'(\mathbb{R}^d)$ is linear functional acting over the space of tempered distributions. The source U in SPDE serves as a random noise which brings some instability (it also could be some boundary condition which are commonly considered in physical framework).

We make use of the result of Vergara *et al.* (2018) who defined the criteria for a suitable integrability condition between the symbol function and spectral measure of the source term. The Theorem 1 of Vergara (2018) establishes the conditions on the symbol function $p(\cdot, \xi)$ and U under which the SPDE (3.24) has a stationary solution within the tempered distribution framework. We state this result in the form of the Theorem 3.6 given below.

Theorem 3.6. *Let U be a real stationary GeRF defined with spectral measure μ_U, $p(\cdot, \xi) : \mathbb{R}^d \to C$ be a symbol function and pseudodifferential operator $\mathcal{L}_p$ be defined through $p(\cdot, \xi)$. Then there exists a real stationary GeRF solution to* (3.24) *if and only if there exists $N \in \mathbb{N}$ such that*

$$\int_{\mathbb{R}^d} \frac{d\mu_U(\xi)}{|p(\xi)|^2(1 + ||\xi||^2)^N} < \infty.$$

The measure

$$d\mu_Z(\xi) = |p(\xi)|^{-2} d\mu_U(\xi)$$

is a spectral measure and any stationary GeRF with this spectral measure solves (3.24)*. Moreover, μ_Z is the unique solution if and only if $|p(\xi)| > 0$.*

For the proof see Theorem 1 of Vergara *et al.* (2018).

The Proposition 1 of Vergara *et al.* (2018) is stated in the Theorem 3.7.

Theorem 3.7. *Let Z be a real stationary GeRF on $\mathbb{R}^d$ with spectral measure μ_Z, and let p be a symbol function over $\mathbb{R}^d$. Then $\mathcal{L}_p Z$ in* (3.24)*, is a real stationary GeRF with spectral measure $\mu_{\mathcal{L}_p Z} = |p(\xi)|^2 \mu_Z$ and its covariance distribution is $c_{\mathcal{L}_p Z} = \mathcal{L}_{|p|^2} c_Z$.*

From the Theorem 3.7, a spectral measure μ_Z of potential stationary solution to (3.24) must satisfy

$$|p(\xi)|^2 \mu_Z = \mu_U. \tag{3.25}$$

It is possible to associate the regular condition for covariance functions and random spectral measure: the slower the growth rate of the random spectral measure μ_Z at infinity, the more regular the covariance function $c(\cdot)$ is.

Let $B(D)$ be the Borel sets of a bounded domain $D \subset \mathbb{R}^d$, then W is the White noise if random measure $W(A) : A \in B(D)$ is a Gaussian random variable with zero mean and covariance $C_W = \mathbb{E}[W(A)W(B)] = |A \cap B|$, that denotes the Lebesgue's measure of A and $W(A \cup B) = W(A) + W(B)$ for disjoint A and $B \in B(D)$. Hence, W is a stationary GeRF with stationary covariance distribution $C_W = \delta_0$, which is not defined as a function as was already discussed in section 3.1.10, thus it can be interpreted as a GeRF.

Note that due to $\mu_W = \mathcal{F}(C_W)$, it follows that the spectral measure of the White Noise

$$d\mu_W = (2\pi)^{-\frac{d}{2}} d\xi. \tag{3.26}$$

A particular case of (3.24) that we consider in this thesis

$$\mathcal{L}_p Z = W, \tag{3.27}$$

where the White noise W is taken as a source noise defined in the right part of the equation. From Theorem 3.6 and (3.26) it follows that there are stationary solutions of (3.27) if and only if $|p(\xi)|^{-2} d\xi$ defines a slow-growing measure and $d\mu_Z(\xi) = (2\pi)^{-\frac{d}{2}} |p(\xi)|^{-2} d\xi$.

We discussed deterministic tools defined in the framework of the partial differential equations and how distribution theory can be used to obtain the solutions of SPDEs through the application of the Fourier transform. We further summarise all the material presented up to now in the following section, where we derive the regularity condition for covariance functions obtained from SPDEs thereby defining the link to the $\mathcal{H}$-matrices approach.

3.2.3 Application of $\mathcal{H}$-matrices to SPDEs

As was defined in Fasshauer (2012), the inverse of an elliptic partial differential operator $\mathcal{L}$ can be written in the integral form by means of the kernel function or so called Green's function as a Schwartz kernel. This statement comes from the Schwartz Theorem 4.1 stated in section 3.1.10. In particular, following the Theorem 3.7, the equation (3.27) formulated in the second order sense leads to

$$\mathcal{L}_{|p|^2} c = \delta, \tag{3.28}$$

where $\delta \in S'(\mathbb{R}^d)$ is the Dirac measure in 0 which follows from (3.21), so that the covariance function in (3.28) can be considered as a Green's (or kernel) function of the operator $\mathcal{L}_{|p|^2}$ (see (3.17)).

As was stated in the Chapter 2 of this thesis, the $\mathcal{H}$-matrices can approximate well a covariance matrix provided that the derivative of the underlying covariance function with respect to the spatial or spatio-temporal coordinates is a fast decaying function, i.e. if the asymptotically smoothness condition (2.4) is satisfied.

For a kernel function of pseudo-differential operator which is characterised by the bounded symbol function, this fact was proved in the Proposition 4.2 of the section 3.1.10. In this section we make use of the results obtained in Vergara *et al.* (2018) and refer to the Proposition 4.2 that formulates the bounds of the kernels derived from the deterministic PDEs. To find the regularity condition of the covariance function required for the application of $\mathcal{H}$-matrices and obtained as solution to the SPDEs as in (3.24), we formulate the following proposition.

Proposition 3.3. *Let $\mathcal{L} : S'(\mathbb{R}^d) \to S'(\mathbb{R}^d)$ be a linear continuous and boundedly invertible pseudo-differential operator defined with the symbol function $p(x, \xi)$, such that $1/p(x, \xi) \in T^l$ for $l \in \mathbb{R}$. Z is a real stationary GeRF with the spectral measure μ_Z and W is the White noise as in* (3.27)*. Then the covariance function c under the White noise W for the solution of the SPDE is regular and for $k > 2l + d + |\beta|$ satisfies the following bounds*

$$|\partial_h^\beta c(h)| \leq s_\beta ||h||^{-k}, \tag{3.29}$$

i.e. it is infinitely differentiable and smooth so that the derivative of c with multi-index β and a constant s_β is the decreasing function as the spatial lag $h \to \infty$.

Proof. We consider the SPDE as in (3.27)

$$\mathcal{L}_p Z = W,$$

where W is the White noise as a source term. As was mentioned in the previous section, there exists a stationary solution of (3.27) if and only if the density $|p(\xi)|^{-2} d\xi$ defines a slowly growing measure. If it holds, then

$$c = \mathcal{F}(\mu_Z) = \frac{1}{(2\pi)^{d/2}} \int_{\mathbb{R}^d} |p(\xi)|^{-2} d\xi,$$

where $|p(\xi)|^{-2} \in O_M(\mathbb{R}^d)$, i.e. space of multiplicators of the Schwartz space given in the Definition 3.6 (due to the regularity and boundedness conditions for $p(\xi)^{-1}$). According to the Theorem 3.2, if $|p(\xi)|^{-2} \in O_M(\mathbb{R}^d)$, the Fourier transform exchanges the space of multiplicators with the space of the fast decreasing functions $O_C(\mathbb{R}^d)$. Therefore, $c \in O_C(\mathbb{R}^d)$ and is regular.

We now have to derive the exact form of the regularity condition for c. The symbol function $|p(\xi)|^{-2}$ here belongs to the class T^r for $r = 2l$. Then according to the Proposition 4.2, we conclude that the covariance function $c(h)$ with the spatial lag h obtained as a solution to (3.27) is smooth, i.e. in the Schwartz space $S(\mathbb{R}^d)$ and the derivatives satisfy the bounds as $||h|| \to \infty$

$$|\partial_h^\beta c_Z(h)| \leq s_\beta ||h||^{-k},$$

if $k > 2l + d + |\beta|$, hence the bounds (3.29) are obtained. □

In the next chapter, we provide some examples of the covariance functions obtained as solution to SPDEs. In particular, we derive the bounds obtained for the Matérn covariance function (1.3) with the application of the Proposition 3.3.

3.2.4 Application to spatio-temporal models

We firstly derive the bounds obtained for the Matérn covariance function. Recall the SPDE formulated in section 1.1.1 that has the following form

$$(\kappa^2 - \triangle)^{\alpha/2} Z = W, \quad \alpha = \nu + d/2, \tag{3.30}$$

where W is a spatial Gaussian white noise with unit variance, $\kappa > 0$ is a scaling parameter, $\triangle$ is the Laplacian of the dimension d defined in (1.5) and $(\kappa^2 - \triangle)^{\alpha/2}$ is the partial pseudo-differential operator. The Matérn covariance function with the spatial lag $h \in \mathbb{R}^d$ is

$$c(h) = \frac{\sigma^2}{\Gamma(\nu) 2^{\nu-1}} (\kappa ||h||)^\nu K_\nu(\kappa ||h||),$$

where ν defines the smoothness of the random field.

We consider SPDE formulated in the second order sense in (3.29). The symbol function $p(x, \xi)$ which corresponds to the pseudo-differential operator $\mathcal{L} = (\kappa^2 - \triangle)^{\alpha/2}$ has the following form

$$p(x, \xi) = (1 + \frac{1}{\kappa^2} ||\xi||^2)^{\frac{\alpha}{2}}, \tag{3.31}$$

which does not depend on x due to the stationarity of the underlying Gaussian process and belongs to the symbol class T^r with $r = \nu + d/2$ (since $\alpha = \nu + d/2$). Therefore, we omit x in the notation and define $p(\cdot, \xi)$.

Since $c = \mathcal{F}(\mu_Z) = \mathcal{F}((2\pi)^{-\frac{d}{2}} |p(\xi)|^{-2})$ and $p(\cdot, \xi)^{-1} \in T^l$ for $l = -\nu - d/2$, given the Proposition 3.3 with the bounds (3.29), we conclude that the Matérn covariance function is infinitely differentiable and admits the following form of the bounds for its derivative with respect to the spatial lag h as $||h|| \to \infty$

$$|\partial_h^\beta c(h)| \leq s_\beta ||h||^{-k} \tag{3.32}$$

for $k > -2\nu + |\beta|$.

The value $\nu = \infty$ in (3.30) corresponds to a Gaussian covariance model that describes a very smooth, infinitely differentiable field since its spectral measure has a density which decreases faster than any polynomial. In fact this covariance function belongs to $S(\mathbb{R}^d)$ and was shown as the example in section 3.1.4.

Recall the definition of the asymptotically smooth condition given in (2.4). Comparing (2.4) with the bounds obtained in (3.32), we conclude that it is possible to apply the $\mathcal{H}$-matrix method to the Matérn covariance function obtained as a solution of SPDE (3.30). Particularly, we can apply the H-matrix approach to approximate the Matérn covariance function with any ν, including exponential form for $\nu = 0.5$, for which the Markov property does not hold due to the fractional power in (3.30). In this case the approximation method of Lindgren *et al.* (2011) is not effective.

We next move to the application of $\mathcal{H}$-matrix to some spatio-temporal SPDEs. Recall the Stein model for generating non-separable space-time covariance functions based on the spectral density $g(\omega, \tau)$ stated in (1.15). As was defined in Vergara *et al.* (2018) the spectral measure over $\mathbb{R}^d \times \mathbb{R}$, where the variable $\omega \in \mathbb{R}^d$ denotes a variable of the spatial domain and $\tau \in \mathbb{R}$ a variable of the temporal domain, has the following form

$$d\mu_Z(\omega, \tau) = (2\pi)^{(d+1)/2}(c_1(a_1^2 + ||\omega||^2)^{\alpha_1} + c_2(a_2^2 + \tau^2)^{\alpha_2})^{-\nu} d\xi d\omega$$

with c_1, $c_2 > 0$, $a_1^2 + a_2^2 > 0$, α_1, α_2 are positive integers and $d_1/(\alpha_1\nu) + d_2/(\alpha_2\nu) < 2$. It is a measure with the density which is inverse of a positive and polynomially bounded continuous function. Therefore, with the White noise as a source noise, the spatio-temporal symbol function

$$p(\omega, \tau) = (2\pi)^{(d+1)/2}(c_1(a_1^2 + ||\omega||^2)^{\alpha_1} + c_2(a_2^2 + \tau^2)^{\alpha_2})^{\nu/2}$$

is smooth with polynomially bounded derivatives of all orders and has a fast decreasing behaviour. Therefore, this spatio-temporal covariance function is also regular and the $\mathcal{H}$-matrix approach can be applied.

Chapter 4

Estimation and Prediction with $\mathcal{H}$-matrices

4.1 Maximum Likelihood estimation

As was described in section 2.3, the $\mathcal{H}$-matrix technique can be used to approximate the Gaussian likelihood function (1.2).

As known, the likelihood can be evaluated through the Cholesky decomposition, i.e decomposition of a positive-definite matrix into the product of a lower triangular matrix and its conjugate transpose as $C_Z = \Lambda\Lambda^T$. This can reduce the computational cost of (1.2) from $O(N^3)$ to $O(N^2)$, followed by the corresponding solver using the vector of observations $\boldsymbol{Z}$. Moreover, the determinant can be found by simply computing the product of diagonal entries of the Cholesky factor Λ.

However, it is still of cubic cost to perform the Cholesky factorization (see Table 4.1). The advantage of the Cholesky factorization with the $\mathcal{H}$-matrices is that it results in the computational cost of $O(k^2 N \log^2 N)$ thereby reducing the total cost of the MLE.

In what follows, we denote the $\mathcal{H}$-matrix approximation of the covariance matrix by $\tilde{C}$ and approximation of the Cholesky factor by $\tilde{\Lambda}$. We aim to approximate a covariance matrix by the $\mathcal{H}$-method and perform a fast approximated Cholesky decomposition, i.e $\tilde{C} = \tilde{\Lambda}\tilde{\Lambda}^T$. However, to be able to perform approximate Cholesky decomposition, the positive definiteness property should be preserved following the steps discussed in the section 2.3.2. In the application part of this thesis, we will get back to this discussion.

The third term in the likelihood (1.2) can be defined through the Cholesky factor as follows

$$\boldsymbol{Z}^T C \boldsymbol{Z} = \boldsymbol{Z}^T(\tilde{\Lambda}\tilde{\Lambda}^T)^{-1}\boldsymbol{Z} = \boldsymbol{U}^T\boldsymbol{U},$$

where N observations $\boldsymbol{Z} = (Z(x_1), \ldots, Z(x_N))^T$ are from a Gaussian Random Field

(GRF) $\{Z(x)\}$ defined over a domain indexed by x, $\boldsymbol{U}$ is a solution of the linear system $\tilde{\Lambda}\boldsymbol{U} = \boldsymbol{Z}$ (also referred to as backsolving) which is composed of the matrix-vector multiplications with a linear cost due to $\mathcal{H}$-matrix technique.

Let $\tilde{\lambda}_i$ be diagonal elements of the $\mathcal{H}$-Cholesky factor $\tilde{\Lambda}$, then

$$\log\det\,(C) = \log\det\,\tilde{\Lambda}\tilde{\Lambda}^T = \log\det\left(\prod_{i=1}^{N}\tilde{\lambda}_i^2\right) = 2\sum_{i=1}^{N}\log\tilde{\lambda}_i.$$

As we defined in section 1.1, the exact Gaussian log-likelihood has the form

$$L(\theta) = -\frac{N}{2}\log 2\pi - \frac{1}{2}\log\det(C) - \frac{1}{2}\boldsymbol{Z}^\top C^{-1}\boldsymbol{Z},$$

then the $\mathcal{H}$-approximation of the exact log-likelihood $L(\theta)$ is defined by $\tilde{L}(\theta, k)$ and can be reformulated as follows

$$\tilde{L}(\theta, k) = -\frac{n}{2}\log 2\pi - \sum_{i=1}^{n}\log\tilde{\lambda}_i - \frac{1}{2}\boldsymbol{U}^\top\boldsymbol{U}, \tag{4.1}$$

where an $\mathcal{H}$-matrix approximation of the Cholesky factor $\tilde{\Lambda}$ is defined through the maximal rank k chosen after preliminary analysis. The gain in the computational cost is illustrated in the Table 4.1.

We now show that the maximum CL estimators with the $\mathcal{H}$-matrices are consistent. Therefore, the aim of the following part is to define whether an approximate computation of the MLE or likelihood with the $\mathcal{H}$-matrices is sufficient to retain the asymptotics under two simulation settings: increasing and fixed domain asymptotics.

As the number of observations N grows, there exist two main asymptotic frameworks in spatial statistics: fixed domain asymptotics described by Stein (1999) and increasing domain asymptotics from Mardia and Marshall (1984). Under fixed domain asymptotics, the sampling region is fixed and bounded, and with increasing N the observations become dense in this window. On the other hand, within increasing domain asymptotics a minimum spacing between the observation points is inserted which leads to an infinite observation domain.

Operation	$\tilde{L}(\theta, k)$	$L(\theta)$
MV multiplication	$O(kN\,log^2N)$	$O(N^2)$
Cholesky decomposition	$O(kN\,log^2N)$	$O(N^3)$

TABLE 4.1: A single MLE evaluation with $\mathcal{H}$-matrices

Under the first setup not all the parameters of the Matérn covariance can be estimated consistently. This fact was confirmed by the celebrated paper of Zhang (2004) who, using equivalence of the probability measures, demonstrated that the range parameter κ in (1.3) can not be estimated consistently. Thus, under the fixed domain asymptotics method, Zhang (2004) established that the maximum likelihood estimate of $\sigma^2\kappa^{2\nu}$ is strongly consistent when $d \leq 3$.

Consistency. We consider a Gaussian random field $\{Z(x) : x \in X\}$ with covariance function $C(\theta)$ with $\theta \in \Theta$, Θ is a compact set of $\mathbb{R}^p$. Z is defined over an arbitrary lattice which is not necessarily non-regular. An increasing sequence of finite subsets of S_N as $N \to \infty$ is considered. The true unknown value of the parameter θ_0 is an interior point of Θ.

Let the error matrix be defined as $E_N = C_N - \tilde{C}_N$, where $\tilde{C}_N$ is the approximated in the $\mathcal{H}$-format covariance matrix. Then $\tilde{C}_N^{-1}E_N = \tilde{C}_N^{-1}C_N - I_N$, where I_N is the identity matrix, and a spectral radius $\rho(\tilde{C}_N^{-1}E_N) = \rho(\tilde{C}_N^{-1}C_N - I_N) < \epsilon_N$, where we exploited the fact that the spectral radius is dominated by any operator norm.

In addition, we refer to the following assumptions on the covariance function C_N and a data vector $\boldsymbol{Z} = \boldsymbol{z}$:

(A1) $||\boldsymbol{z}_N||_2 \leq p$ for some small positive p;

(A2) $||C_N^{-1}|| \leq c_N$. However, as was mentioned by Litvinenko *et al.* (2019), this assumption is strong and depends on the parameters of the covariance function and the chosen rank k;

(A3) the error $\epsilon_N < 1$, i.e.

$$\rho(\tilde{C}_N^{-1}E_N) = \rho(\tilde{C}_N^{-1}C_N - I_N) < \epsilon_N < 1. \tag{4.2}$$

Theorem 4.1. *Assume conditions (A1)-(A3). Then $\tilde{\theta}_N$ is consistent.*

Proof. As known, $\hat{\theta}_N \to \theta_0$ a.s. with $N \to \infty$ provided that there exists a deterministic asymptotic criterion function $L(\theta)$ such that $N^{-1}L_N(\theta) \to L(\theta)$ a.s. uniformly with respect to $\theta \in \Theta$ with $\theta_0 = \text{argmax } L(\theta)$ and $\hat{\theta}_N = \text{argmax } L_N(\theta)/N$. Then, if $\tilde{\theta}_N$ is the approximated with the $\mathcal{H}$-matrix estimator such that $\tilde{\theta}_N \approx \hat{\theta}_N$, then $\tilde{\theta}_N$ is consistent.

From (4.2), Ballani and Kressner (2015) obtained

$$|\log\det(C_N) - \log\det(\tilde{C}_N)| \leq -N\log(1-\epsilon_N), \tag{4.3}$$

Thus, using the result obtained by Litvinenko *et al.* (2019), (4.2) and (4.3), for a data vector $\boldsymbol{z}$

$$\begin{aligned}\left|\frac{\tilde{L}_N(\theta)}{N}-\frac{L_N(\theta)}{N}\right| &= \frac{1}{2}\frac{1}{N}\log\frac{\det(C_N)}{\det(\tilde{C}_N)}-\frac{1}{2}\frac{1}{N}|\boldsymbol{z}_N^T(C_N^{-1}-\tilde{C}_N^{-1})\boldsymbol{z}_N| \\ &\le -\frac{1}{2}\log(1-\epsilon_N)-\frac{1}{2}\frac{1}{N}|\boldsymbol{z}_N^T(I_N-\tilde{C}_N^{-1}C_N)C_N^{-1}\boldsymbol{z}_N| \\ &\le \frac{1}{2}\epsilon_N+\frac{1}{2}\frac{||\boldsymbol{z}_N||_2^2}{N}||C_N^{-1}||_2\epsilon_N \\ &\le \frac{1}{2}\epsilon_N+\frac{1}{2}\frac{||\boldsymbol{z}_N||_2^2}{N}c_N\epsilon_N, \end{aligned} \tag{4.4}$$

from where we conclude that for $\epsilon_N < 1$ as $N \to \infty$, $L(\tilde{\theta}_N), \approx L(\hat{\theta}_N)$ and, therefore, $\tilde{\theta}_N$ is consistent. □

Following the same steps as Kaufman *et al.* (2008), we now consider the fixed domain asymptotics.

Theorem 4.2. *Let $C = \sigma^2\Sigma(\varphi)$, where $\Sigma(\varphi)$ is known correlation function and the variance parameter σ^2 is unknown. In addition, let the likelihood approximation* (4.1) *be based on the observations in a set of finite subsets S_N for $N = (1, \ldots, \infty)$. Then $\hat{\tilde{\sigma}}_N^2 \to \sigma^2$ almost surely as $N \to \infty$.*

Proof. Let $\Sigma_N = \Lambda_N\Lambda_N^T$, then $\frac{1}{\sigma}\Lambda_N^{-1}Z_N \sim \mathcal{N}(0, I_N)$ and the $\mathcal{H}$-approximated $\hat{\tilde{\sigma}}_N^2$ is as follows

$$\begin{aligned}\hat{\tilde{\sigma}}_N^2 &= \frac{1}{N}Z_N^T\tilde{\Sigma}_N^{-1}Z_N = \frac{1}{N}X_N^T(\sigma\Lambda_N)^T\tilde{\Sigma}_N^{-1}(\sigma\Lambda_N)X_N = \\ &= \frac{1}{N}X_N^T(\sigma\Lambda_N)^T(\tilde{\Lambda}_N^T)^{-1}\tilde{\Lambda}_N^{-1}(\sigma\Lambda_N)X_N = \\ &= \frac{\sigma^2}{N}X_N^T\Lambda_N^T(\tilde{\Lambda}_N^T)^{-1}\tilde{\Lambda}_N^{-1}\Lambda_N X_N,\end{aligned}$$

where $X_N \sim \mathcal{N}(0, I_N)$.

Since $\Lambda_N = \tilde{\Lambda}_N + E_N = \tilde{\Lambda}_N(I + \tilde{\Lambda}_N^{-1}E_N)$, where $E_N \in \mathbb{R}^{N\times N}$ is the error matrix, then

$$\begin{aligned}\hat{\tilde{\sigma}}_N^2 &= \frac{\sigma^2}{N}X_N^T(I+\tilde{\Lambda}_N^{-1}E_N)^T\tilde{\Lambda}_N^T(\tilde{\Lambda}_N^T)^{-1}\tilde{\Lambda}_N^{-1}\tilde{\Lambda}_N(I+\tilde{\Lambda}_N^{-1}E_N)X_N \\ &= \frac{\sigma^2}{N}X_N^T(I+\tilde{\Lambda}_N^{-1}E_N)^T(I+\tilde{\Lambda}_N^{-1}E_N)X_N \\ &\approx \frac{\sigma^2}{N}\sum_{i=1}^{N}\lambda_{N,i}^2\chi_i^2,\end{aligned}$$

where $\lambda_{N,i}$ is the eigenvalue of $(I + \tilde{\Lambda}_N^{-1}E_N)$ and χ_i^2 are iid χ_1^2 random variables.

As was discussed in Kaufman *et al.* (2008) in the proof of the theorem 3, $\hat{\tilde{\sigma}}_N^2 \to \sigma^2$ almost surely as $N \to \infty$ if $\sup_N(\frac{1}{N}\sum_{i=1}^N |\lambda_{N,i}|^q)^{1/q} < \infty$ for some $q = (1, \dots, \infty)$. From (4.2) we conclude

$$\begin{aligned} \sup\left(\frac{1}{N}\sum_{i=1}^{N}|\lambda_{N,i}(I+\tilde{\Lambda}_N^{-1}E_N)|^q\right)^{1/q} &\leq \left(\frac{1}{N}N|\lambda_{N,\max}(I+\tilde{\Lambda}_N^{-1}E_N)|^q\right)^{1/q} \\ &\leq (\rho(I+\tilde{\Lambda}_N^{-1}E_N)^q)^{1/q} \\ &< 1-\epsilon_N < \infty, \end{aligned}$$

where ρ is the spectral radius and we used the fact that $\rho(\tilde{\Lambda}_n^{-1}E_N) < \epsilon_N$ for $\epsilon_N < 1$, i.e spectral radius is dominated by any operator norm, then the condition is satisfied. Then $\hat{\tilde{\sigma}}_N^2 \to \sigma^2$ almost surely as $N \to \infty$. □

The results for $\hat{\tilde{\sigma}}_n^2/\varphi^{*2\nu} \to \sigma^2/\varphi^{2\nu}$ for the fixed range parameter φ^* can be easily obtained in analogy with the proof of Corollary 1 of Kaufman *et al.* (2008).

4.2 Kriging prediction

In geostatistics the standard approach, called kriging, is based on the principle of minimum mean squared prediction error. Consider a zero-mean Gaussian Random Field $\{Z(x) : x \in \mathbb{R}^d\}$ which is characterised by covariance function $C_Z(x_i, x_j)$, for $x_i, x_j \in \mathbb{R}^d$. With the inference on the process $Z(x)$, the best linear unbiased prediction (BLUP), as stated in Cressie (1993), at an unobserved location x_0 is defined as follows

$$Z(x_0) = \boldsymbol{c}(x_0)^\top C_Z^{-1}\boldsymbol{Z}, \tag{4.5}$$

where $\boldsymbol{c}(x_0) = [c(x_0, x_1), \dots, c(x_0, x_N)]'$ is covariance vector formed based on a new location x_0 and $C_Z = C(x_i, x_j)$. The mean squared prediction error $\mathrm{MSPE}(x_0, C_Z)$ has the form

$$\mathrm{MSPE}(x_0, C_Z) = C(x_0, x_0) - c(x_0)^T C_Z^{-1} c(x_0).$$

Following Furrer *et al.* (2006), if the BLUP is calculated under a different covariance function K_Z, the mean-squared prediction error has the form

$$\mathrm{MSPE}(x_0, K_Z) = C(x_0, x_0) - 2\tilde{\boldsymbol{c}}^T(x_0)K_Z^{-1}\boldsymbol{c}(x_0) + \tilde{\boldsymbol{c}}^T(x_0)K_Z^{-1}CK_Z^{-1}\tilde{\boldsymbol{c}}(x_0).$$

For the $\mathcal{H}$-matrix approach and kriging prediction we can proceed in the same way as with the likelihood. Firstly, we substitute C_Z in (4.3) by the approximated by $\mathcal{H}$-method covariance $\tilde{C}_Z$ and then find the Cholesky factorisation. Then a simple kriging prediction for a location x_0 using the estimated covariance function with $\hat{\theta}$ in (4.1) will take the following form

$$\tilde{Z}(x_0) = \tilde{\boldsymbol{c}}(x_0)^\top \tilde{C}_Z^{-1} \boldsymbol{Z}, \tag{4.6}$$

where $\tilde{\boldsymbol{c}}(x_0) = [\tilde{c}(x_0, x_1), \ldots, \tilde{c}(x_0, x_N)]'$ is the $\mathcal{H}$-matrix approximation of the corresponding covariance vector. We note that (4.6) is again based on the matrix-vector multiplications which lead to the log-linear cost computation.

In general, the weights $c(x_0)^T C_Z^{-1}$ in (4.5) are close to zero for observations whose locations are far from x_0. The $\mathcal{H}$-matrix approach is based on the idea to 'diminish' the importance of the distant from x_0 points while preserving the correlation of x_0 with closely located observations. Taking into account these two remarks, we believe that the substitution C_Z in (4.5) by the approximated with the $\mathcal{H}$-method covariance $\tilde{C}_Z$ leads to an asymptotically optimal mean squared error. However, to find the conditions under which the asymptotic mean squared error of the predictions using the $\mathcal{H}$-covariance converges to the minimal error is out of the scope of this thesis.

4.3 Admissibility condition in practice

As was mentioned in section 2.2.1, the minimum number of elements or so called leafsize is required in order to stop the partitioning of the binary cluster tree. Therefore, in this section we aim at defining this number based on the statistical analysis. In addition, we focus on the correction of the admissibility condition (2.3) defined in section 2.2.3. Particularly, we analyse the influence of the range parameter on the admissibility condition.

We begin by setting up the hypothesis, that the data sites are uniformly distributed throughout the domain of interest $A \subseteq \mathbb{R}^2$. To define the leafsize in each block we adopt the following condition. Denoting I a cluster of interest, $I \subseteq A$, we find the expected number $\bar{N}$ of locations in I with the radius r as follows

$$\begin{aligned} \bar{N} &\approx \sum_{i=1}^{N} \mathbb{E}[1_I(x_i, y_i)] \\ &= \sum_{i=1}^{N} \Pr[(X_i, Y_i) \in I] = N \cdot \pi r^2 / |A|, \end{aligned} \tag{4.7}$$

where $1_I(x_i, y_i)$ is the indicator function, N is the total number of data locations in the sample, $|A|$ is the dimension of domain A and πr^2 represents the area of the cluster I. The radius r corresponds to the required cut-off distance which can be chosen according to the preliminary estimate of the spatial dependence, i.e. based on the value of correlation between sites.

The last part of this section focuses on adjusting the admissibility condition (2.3) stated in the section 2.2.3. Particularly, we aim to adapt the parameter η which was left unexplained in the framework of spatial statistics.

As was mentioned in the section 2.2.3, the parameter is often considered at its default value, i.e. $\eta = 1$. However, since the admissibility condition (2.3) depends on the distance which is usually scaled by the range parameter $\varphi = \eta/\kappa$ in spatial statistics, we find it reasonable to adjust this condition and parameter η for the range parameter.

The range parameter flattens a covariance function $c(x_i, x_j)$ with increasing φ by scaling the distance $||x||$. Therefore, we obtain a new condition

$$\min\{\mathrm{diam}(Q_\sigma), \mathrm{diam}(Q_\tau)\} \leq \frac{\eta}{\varphi}\mathrm{dist}(Q_\sigma, Q_\tau), \tag{4.8}$$

where $Q_\sigma \subset \mathbb{R}^d$, $\sigma \subset I$ and $Q_\tau \subset \mathbb{R}^d$ with $\tau \subset J$.

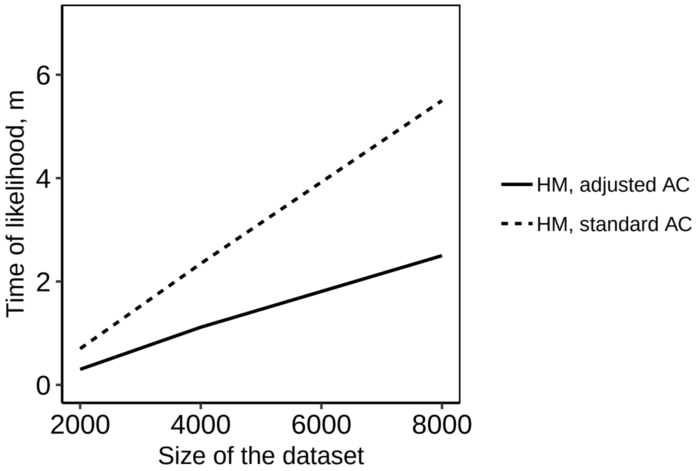


Figure 4.1: Time of the ML estimation with scaled (solid) for the range parameter φ and standard (dashed) admissibility condition (AC) with different sample size

In this simulation study we consider increasing domain asymptotics method and perform the experiments on the perturbed grid of spatial locations. For the different sample size of $N_k = \{2000, 4000, 8000\}$ points chosen without replacement, we simulated $L = 100$ realizations of zero-mean GRF with mean 0 and Matérn covariance with the true parameters $\theta = (\sigma^2, \varphi) = (1, 0.1)$ with the fixed $\nu = 0.5$ in (1.3) that corresponds to the exponential type of covariance function.

As a result of the ML estimation with scaling the distance of exponential covariance by the range parameter, the computational efficiency is doubled with respect to the standard admissibility condition. See Figure 4.1 for the comparison of the computational time for the MLE with standard and scaled admissibility conditions.

We assume that the sharp behaviour of covariance function (small range parameter) will lead to nearly diagonal and thus well-conditioned matrices. Therefore, with respect to the $\mathcal{H}$-matrices, a small value of the range parameter leads to the fast convergence rates and respectively computations.

Chapter 5

Numerical Results

5.1 Covariance tapering and composite likelihood

Following the main idea of the $\mathcal{H}$-matrix approximation, the certain elements of the covariance matrix C_Z can be defined through the low-rank factors. More precisely, the underlying covariance function follows an exact representation $c(x, x')$, for $||x - x'|| \leq a$, where $a > 0$ and $x, x' \in \mathbb{R}^d$, otherwise it can be written through the factors a, b, i.e. $\tilde{c}(x, x') = \sum_{j=1}^{k} a_j(x) b_j(x')$ for $||x - x'|| > 0$ and rank k. As regards covariance tapering approach, if we have reason to believe that distant pairs of observations are independent, then this structure is modelled using a compactly supported covariance function.

Taking into account this comparison, it can be concluded $\mathcal{H}$-matrix approximation and covariance tapering approach are based on quite similar idea. Therefore, the main purpose of this section is to compare the performance of these two methods based on computational and statistical efficiency. In addition, we provide the comparison with the weighted composite likelihood following the idea of Bevilacqua *et al.* (2012).

As was already described in section 1.1.1, in the covariance tapering approach some elements of the covariance matrix C_Z are set to zero after element-wise multiplication by a correlation matrix, i.e $C_T = C_Z \odot T(\delta)$, where $T(\delta)$ is the taper with the cut-off distance δ. The Wendland correlation function, discussed in the section 1.1.1, is chosen as follows (see Figure 1.4)

$$K(||x - x'||; \delta) = \left(1 - \frac{||x - x'||}{\delta}\right)_+^4 \left(\frac{4||x - x'||}{\delta} + 1\right)_+. \tag{5.1}$$

Kaufman *et al.* (2008) proposed two approximations of the log-likelihood, namely

$$L_{1T}(\theta) = -\frac{N}{2}\log 2\pi - \frac{1}{2}\log\det[C_Z \odot T(\delta)] - \frac{1}{2}\boldsymbol{Z}^\top [C_Z \odot T(\delta)]^{-1}\boldsymbol{Z} \tag{5.2}$$

and

$$L_{2T}(\theta) = -\frac{N}{2}\log 2\pi - \frac{1}{2}\log\det[C_Z \odot T(\delta)] - \frac{1}{2}\boldsymbol{Z}^\top([C_Z \odot T(\delta)]^{-1} \odot T(\delta))\boldsymbol{Z}. \quad (5.3)$$

The first approximation (5.2) is computationally more efficient but the corresponding 'score' function for θ

$$\frac{\partial}{\partial\theta}L_{1T} = \frac{1}{2}\big(\boldsymbol{Z}^T[C_Z \odot T(\delta)]^{-1}C_i[C_Z \odot T(\delta)]^{-1}\boldsymbol{Z} - \mathrm{tr}([C_Z \odot T(\delta]))^{-1}C_i)\big),$$

where $C_i = \partial C_Z(\boldsymbol{\theta})/\partial\theta_i$, is biased or $\mathbb{E}_{\theta_0}\big(\frac{\partial}{\partial\theta}L_{1T}\big) \neq 0$, i.e. the expectation is not null at $\theta = \theta_0$, where $\theta_0 \in \Theta$ denotes the true parameter value.

Therefore, the second tapered likelihood (5.3) was proposed where both the model covariance matrix and the sample covariance matrix are tapered which leads to the unbiased derivative of (5.3).

Since (4.1) also entails a biased score function, i.e $\frac{1}{2}\big(\boldsymbol{Z}^T\tilde{C}_Z^{-1}C_i\tilde{C}_Z^{-1}\boldsymbol{Z} - \mathrm{tr}(\tilde{C}_Z^{-1}C_{iZ})\big)$, in this experiment we use (4.1) with $\mathcal{H}$-matrix approximated covariance $\tilde{C}_Z$ and (5.2) with tapered covariance C_T.

The problem of covariance tapering for interpolation of large spatial datasets was discussed by Furrer *et al.* (2006). The best linear unbiased prediction at an unobserved location x_0 follows from (4.5). By replacing the covariance matrix C_Z by the tapered version C_T, the linear system with the weights in (4.5) can be solved efficiently, i.e. $\boldsymbol{c}(x_0)^T C_T^{-1}$. By Furrer *et al.* (2006) it was shown that the asymptotic mean squared error of the predictions using the tapered covariance converges to the minimal error.

We now shortly describe the method proposed by Bevilacqua *et al.* (2012), Bevilacqua *et al.* (2012) and called weighted composite likelihood. Given a space-time realization $\{Z(x_i,t_l)\}$, $l = 1,\dots,T$, $i = 1,\dots,N$ with $x_i \in \mathbb{R}^d$, $t \in \mathbb{R}$, we consider the Gaussian density of a pair $Z(x_i,t_l)$ and $Z(x_j,t_k)$. Then the pairwise log-likelihood

$$l_{\text{pair}}(\theta, z) = \sum_{i,j,k\in D} \log[f_Z(z_{il}, z_{jk})]w_{ijlk}, \quad (5.4)$$

where

$$D = \begin{cases} l = 1,\dots,T, i = 1,\dots,N, & k = l,\dots,T \\ j = i+1,\dots,N, & \text{if } l = k \\ j = 1,\dots,N, & \text{if } l > k \end{cases}$$

and w_{ijlk} are non-negative weights defined as

$$w_{ijlk} = \begin{cases} 1, & \text{if } ||x_i - x_j|| < d_x, |t_l - t_k| < d_t \\ 0, & \text{otherwise} \end{cases}, \tag{5.5}$$

for a fixed spatiotemporal lag (d_x, d_t).

The pairwise likelihood estimator $\theta \in \Theta \subset \mathbb{R}^p$ is obtained maximizing (5.4) with respect to θ.

The goal of this Chapter is to recover the true values of the parameters (σ^2, φ) with the approximated in the $\mathcal{H}$-format Matérn covariance. We compare the performance of two methods: $\mathcal{H}$-matrices and covariance tapering in the spatial setting under the increasing domain asymptotics setup. In the section 5.2.2, the comparison is also performed with the weighted composite likelihood in the spatio-temporal framework on the simulated GRF with space-time Gneiting type of covariance function (1.14). Chapter is concluded with the real data application.

5.2 Simulated data

The simulation study, considered in this section, is concerned with the increasing domain asymptotics setup. Taking advantage of the fact that $\mathcal{H}$-matrix representations are easily applied to irregularly allocated data sites, in this Chapter we perform the experiments on the randomly perturbed grid of spatial locations. For the data locations we follow the simulation setup described in Kaufman *et al.* (2008). Particularly, a regular grid with increments 0.03 is constructed over

$$W_k = [0, 2^{(k+2)/2}] \times [0, 2^{(k+2)/2}], \quad k = 0, \dots, 2.$$

To obtain irregular allocation of the data sites and avoid numerical instabilities, the regular grid points were perturbed by adding a uniform random value on $[-0.01, 0.01]$. With this setup, each data location is at least 0.01 units distant from its neighbours. As an example, the regular grid is shown on the left part of the Figure 5.1, perturbed grid is given on the right side.

For the different sample size of $N_k = \{2000, 4000, 8000\}$ points with $k = 0, \dots, 2$ chosen without replacement, we simulated $L = 100$ realizations of zero-mean GRF with mean 0 and Matérn covariance with the true parameters $\theta = (\sigma^2, \varphi, \nu, \tau^2) = (1, 0.1, 0.5, 0.1)$, where we fixed the smoothness parameter $\nu = 0.5$ in (1.3) that corresponds to the exponential type of covariance function and nugget parameter which is

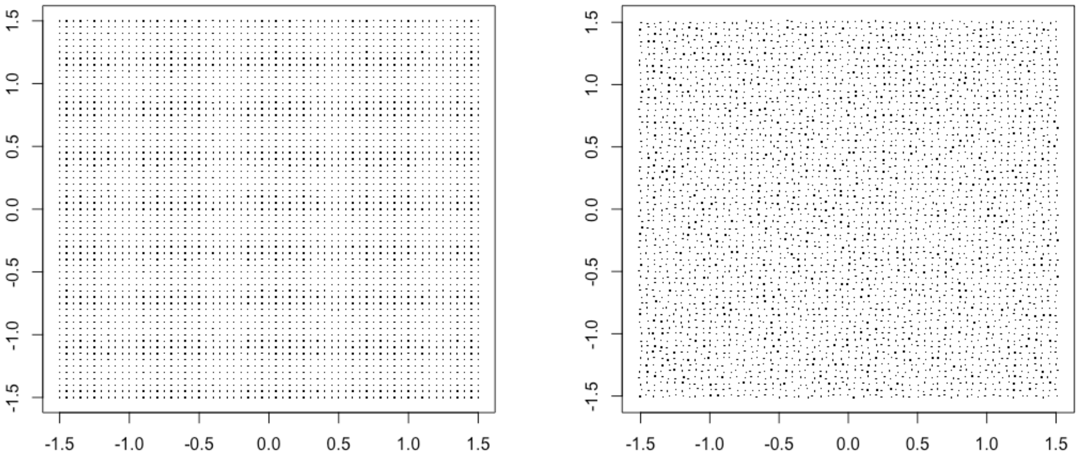


FIGURE 5.1: Regular (left) and perturbed (right) grid

denoted by τ^2. The nugget parameter $\tau^2 = 0.1$ was added to the diagonal in order to preserve the positive definiteness property as was suggested by Litvinenko *et al.* (2019) and discussed in the section 2.3.2.

To check the predictive performance with the increasing N_k, we divided the simulated data into a training dataset chosen at random and a validation dataset containing the remaining part, i.e $M = \{200, 400, 800\}$ observations respectively.

5.2.1 Results: spatial framework

For $\mathcal{H}$-matrix approach we implemented C code, exploiting the *h2lib* library[1] for the construction of the cluster, block cluster trees and $\mathcal{H}$-matrices. The minimal number of the elements at which partition in a block stops was chosen following the condition (4.5) stated in section 4.3. The spatial lag $h_{\min}$ was selected to guarantee the minimum value of correlation 0.5 between the observations in one block.

Instead of the standard admissibility condition defined in (2.3), the adapted for spatial statistics condition (4.8) was exploited. In addition, for the low-rank approximation of the off-diagonal blocks we used adaptive cross-approximation described in the Algorithm 2 of the section 2.2.5.

As practical range we set $\varphi = 0.1$ due to consistency of φ over the spatial domain to increasing domain framework. Because we keep d as fixed, increasing k and consequently the number N_k of observations, the percentage of nonzero elements in the resulting tapered covariance matrix decreases. By varying the practical range $\delta = \{0.15, 0.3, 0.5\}$, the percentage of non-zero elements p in the tapered covariance matrix increases, for

[1] Steffen Borm, Scientific Computing Group, Kiel University

example for $N_1 = 2000$: $\{p_1 = 0.2, p_2 = 0.5, p_3 = 1.5\}$, $N_2 = 4000$: $\{p_1 = 0.15, p_2 = 0.38, p_3 = 1.33\}$, $N_3 = 8000$: $\{p_1 = 0.1, p_2 = 0.27, p_3 = 1.12\}$.

For the $\mathcal{H}$-matrices we control the compression ratio q which is defined as the ratio between the sizes of a compressed (hierarchical matrix) $\tilde{C}$ and original matrix C

$$q = \frac{\text{size}(\tilde{C})}{\text{size}(C)}, \tag{5.6}$$

where size of the matrix $A \in \mathbb{R}^{N \times N}$ is defined by the number of rows N and columns N that it contains. The reduced size of the hierarchical matrix is explained by low-rank blocks $\tilde{A}|_{\text{block}} \in \mathbb{R}^{N \times k}$ (denoted by green colour in Figure 2.1), where k is the rank.

To be consistent with the percentage of non-zero entries in tapered covariance, the parameter η in the improved adjusted for the range condition (4.8), was varied between $\eta = (2, 1.4, 1)$. That means that the compression ratio (5.6) decreases as η/φ gets bigger. For example, for $N_1 = 2000$: we obtained $\{q_1 = 0.3, q_2 = 0.8, q_3 = 1.9\}$, expressed in percentages, $N_2 = 4000$: $\{q_1 = 0.25, q_2 = 0.48, q_3 = 1.56\}$, $N_3 = 8000$: $\{q_1 = 0.2, q_2 = 0.37, q_3 = 1.31\}$, so that it is quite comparable with the changing density of the tapered covariance.

For evaluating (5.2), the implementation[2] of Kaufman *et al.* (2008) and sparse matrix implementation with the *R* package *spam* Furrer *et al.* (2010) was exploited. In this case the factorization is computed once. Afterwards, by using pointers in C, the structure of the matrix is passed and numerical computations are performed by functions of the sparse library. With the repeated iterations it can be evaluated efficiently.

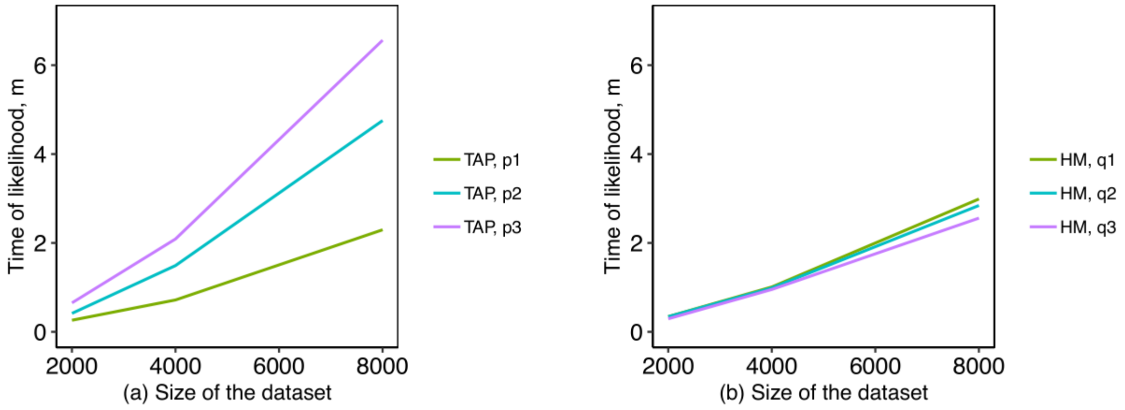


FIGURE 5.2: Time (min) of the likelihood computation with different p, q for (a) Tapering and (b) $\mathcal{H}$-matrices approach

[2] available at http://www.image.ucar.edu/Data/precip_tapering/

Figure 5.2 and 5.3 show the computational time for maximum likelihood estimation[3] and kriging prediction given different values of the density of the tapered matrix p and compression ratio q for the $\mathcal{H}$-matrix method.

It can be seen that the $\mathcal{H}$-matrix approach implements the ML estimation and kriging faster and the difference between two methods is more evident when number of observation N_k increases. The experiments were performed on 2.4 GHz processor with 8 GB of memory running Mac OS X 10.11.

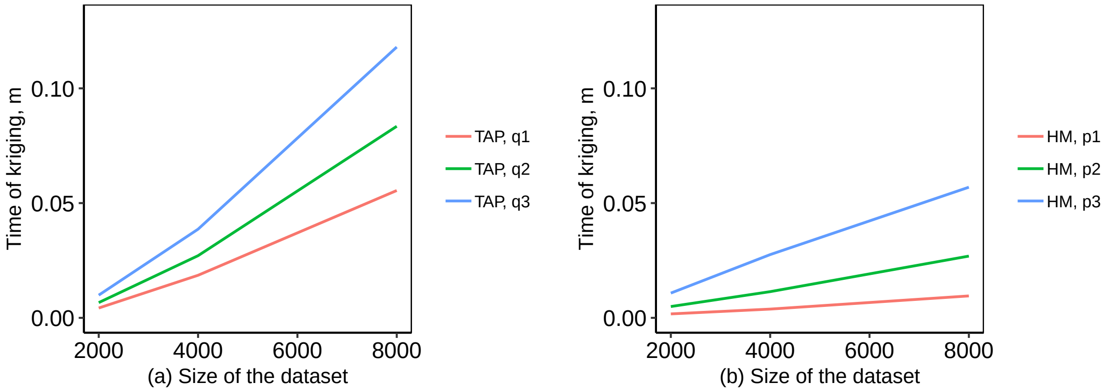


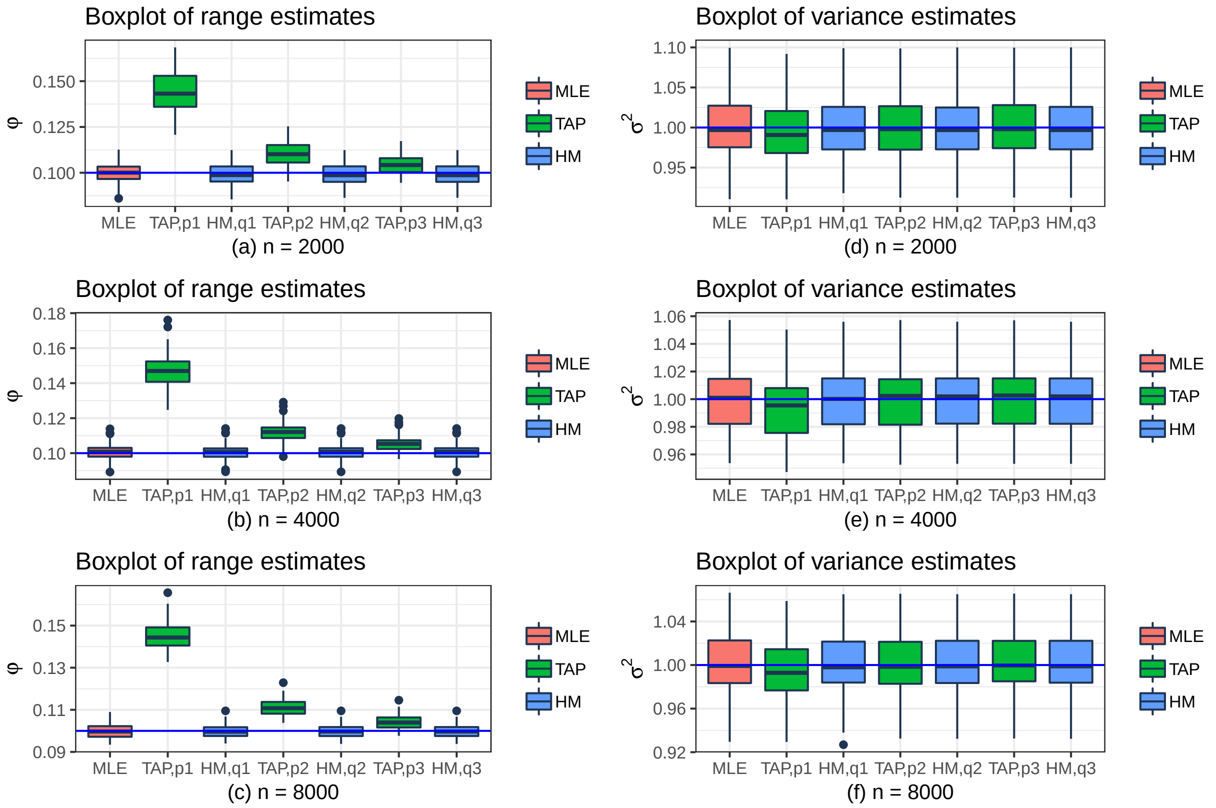
FIGURE 5.3: Time (min) of the kriging prediction for $N_1 = 2000$: $\{p_1 = 0.2, p_2 = 0.5, p_3 = 1.5\}$, $N_2 = 4000$: $\{p_1 = 0.15, p_2 = 0.38, p_3 = 1.33\}$, $N_3 = 8000$: $\{p_1 = 0.1, p_2 = 0.27, p_3 = 1.12\}$ for (a) Tapering and (b) $\mathcal{H}$-matrices approach with $N_1 = 2000$: $\{q_1 = 0.3, q_2 = 0.8, q_3 = 1.9\}$, $N_2 = 4000$: $\{q_1 = 0.25, q_2 = 0.48, q_3 = 1.56\}$, $N_3 = 8000$: $\{q_1 = 0.2, q_2 = 0.37, q_3 = 1.31\}$

Figure 5.4 shows boxplots of the estimates of the range φ and variance σ^2 parameters with the both methods, including the exact ML estimation. The horizontal line indicates the true values of the estimates ($\varphi = 0.1, \sigma^2 = 1$). As taper δ decreases, the biases in the one-taper estimates increase. In fact the bias higher for δ smaller compared to the true correlation range of the process. In contrast, we see negligible bias in the $\mathcal{H}$-matrices estimates. The difference in variance estimates with both methods is almost indistinguishable.

It can be seen that the application of $\mathcal{H}$-matrices approach for ML estimation resulted in good statistical efficiency, even with a small compression ratio q defined in (5.6), which in turn results in a gain of computational efficiency.

To compare the predictive performance of both methods we compute the measure, termed Root-Mean-Squared Prediction Error (RMSPE). With the increasing N_k we divided the simulated data into a training dataset chosen at random and a validation dataset containing the remaining part, i.e $M = \{200, 400, 800\}$ observations respectively.

[3] Nelder-Mead method was used for evaluating the likelihood with the absolute tolerance 10^{-3} for both methods



(a) n = 2000

(d) n = 2000

(b) n = 4000

(e) n = 4000

(c) n = 8000

(f) n = 8000

FIGURE 5.4: Boxplots of sampled estimates (a)-(c) $\hat{\varphi}$ and (d)-(f) $\hat{\sigma}^2$ with the horizontal line of the true estimates ($\varphi = 0.1, \sigma^2 = 1$) under the exact maximum likelihood estimation (MLE), covariance tapering (TAP) and $\mathcal{H}$-matrices (HM)

The set of the predicted locations for each M is denoted as D_M^* with each new location $x_0 \in D_M^* \subset \mathbb{R}^d$.

If $\tilde{Z}(x_0, l)$ denote the model-A predictor, where $Z(x_0, l)$ is the lth simulated process evaluated at a new location x_0 and $A =$TAP, HM. Then the model-A predictor RMSPE for the lth simulation is

$$\text{RMSPE}_A(l) = \sqrt{\sum_{x_0 \in D_M^*} \left(\tilde{Z}(x_0, l) - Z(x_0, l)\right)^2}, \quad l = 1, \dots, L. \tag{5.7}$$

We then consider a measure of relative skill (RS), relative to HM:

$$\text{RS}(N) = \text{RMSPE}_{\text{HM}}(l)/\text{RMSPE}_{\text{TAP}}(l), \quad \text{for } l = 1, \dots L$$

for the different $N = \{2000, 4000, 8000\}$.

Hence, as can be seen from the Figure 5.5 for different sample size N_k and various density and correlation ratio p and q respectively, $RS(N) < 1$ mainly. Therefore, $\mathcal{H}$-matrices approach has a better predictive accuracy.

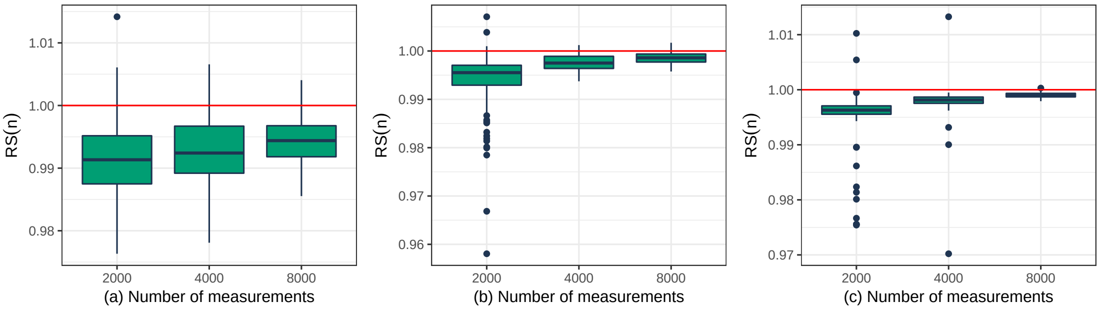


FIGURE 5.5: Boxplots of RS(N) for (a) $N_1 = 2000 : p_1 = 0.2, q_1 = 0.3$, $N_2 = 4000 : p_1 = 0.15, q_1 = 0.25$, $N_3 = 8000 : p_1 = 0.1, q_1 = 0.2$, (b) $N_1 = 2000 : p_2 = 0.5, q_2 = 0.8$, $N_2 = 4000 : p_2 = 0.38, q_2 = 0.48$, $N_3 = 8000 : p_2 = 0.27, q_2 = 0.37$, (c) $N_1 = 2000 : p_3 = 1.5, q_3 = 1.9$, $N_2 = 4000 : p_3 = 1.33, q_3 = 1.56$, $N_3 = 8000 : p_3 = 1.12, q_3 = 1.31$

5.2.2 Results: spatio-temporal framework

In this section we evaluate the performance of the $\mathcal{H}$-method for Gneiting class of spatio-temporal covariance function based on the analytical derivation of the admissibility condition provided in the section 2.2.3 of Chapter 2. The performance is compared with the covariance tapering approach and the weighted composite likelihood method.

We simulated $NT = 16000$ irregularly located spatio-temporal sites on a space-time domain $S \times T$ with 80 spatial locations with $S \subset \mathbb{R}^d : [0, 20] \times [0, 20]$ and 200 temporal instants with $T \subset \mathbb{R} : [0, 20]$. The selected locations were then divided into the training 14400 and test 1600 locations sets. The simulation setup for spatio-temporal framework reflects the scenario described for the spatial domain in the section 5.2.

A realization from a space-time GRF $Z(s, t)$ for 100 random samples is obtained by considering the non-separable Gneiting space-time covariance function

$$C(h, u) = \frac{\sigma^2}{(20|u|^{2\alpha}/a + 1)} \exp\left(- \frac{3||h||}{b(20|u|^{2\alpha}/a + 1)^{\eta/2}} \right) \tag{5.5}$$

with the spatial $h = s - s'$ and temporal $u = t - t'$ lags, such that the Cholesky factor Λ was multiplied on a Gaussian random vector $W \sim N(0, I)$, $C_Z = \Lambda\Lambda^T$. In this numerical study, the smoothness $\alpha = 0.5$, space-time interaction $\eta = 0.5$ parameters and a small nugget $\tau^2 = 0.1$ were fixed.

We aimed to estimate temporal a and spatial b range parameters under two simulation settings: 1) large-scale dependence: $a = 10$, $b = 20$; 2) small-scale dependence: $a = 5$, $b = 10$.

The simplest weights (5.5) for the weighted composite likelihood (WCL) approach can be chosen as $w_{ij} = 1$, if $||x_i - x_j|| \leq d_x$ and $|t_i - t_j| \leq d_t$, and 0 otherwise. As was mentioned by Bevilacqua *et al.* (2012), this choice has evident computational advantages and can also improve the statistical efficiency. The spatiotemporal lag was fixed at the value $\{d_x = 15, d_t = 5\}$ for the first setup and $\{d_x = 10, d_t = 5\}$ for setup 2.

For simulation, ML estimation and kriging prediction with the covariance tapering (TAP) and WCL approaches we exploited a new *R*-package *GeoModels* (see https:// vmoprojs.github.io /GeoModels-page/). As space-time compactly supported function, separable space-time Wendland correlation model was chosen, that is easily obtained as a product of a spatial $K_s(||x-x'||;\delta)$ and a temporal correlation model $K_t(|t-t'|;\tau)$

$$K(||x-x'||,|t-t'|;\delta,\tau) = K_s(||x-x'||;\delta)K_t(|t-t'|;\tau),$$

where under the two simulation settings the spatial and temporal cut-off distances were selected as $(\delta_1, \delta_2) = (10, 15)$ and $(\tau_1, \tau_2) = (5, 5)$ for compactly supported correlation (5.1) in spatial and temporal coordinates respectively.

To obtain a $\mathcal{H}$-matrix (HM) representation of the spatio-temporal covariance (5.5), we implemented a new space-time code in C. In addition, the compression ratio (5.6) was chosen to be consistent with the density parameter of TAP approach, i.e. $\text{size}(\tilde{C})/\text{size}(C)$ is 15% for setup (1) and 10% for (2). The space-time regularity condition was adjusted for the spatial and temporal ranges as was discussed in section 4.3.

Mean estimated values of the covariance parameters and other model parameters are demonstrated in Table 5.1. Under both parameter settings, the HM approach clearly outperformed TAP in terms of computational time. However, the WCL method led to a faster computation compared to the HM method. Since the time of the computation provided in Table 5.1 for HM approach is still not optimised, then finding a way to improve the speed of convergence in future work could be fruitful.

	Method	$\hat{a}$	$\hat{b}$	RMSPE	Time (lik), min	Time (kr), min
Setup 1	*HM*	**10.08**	**20.1**	**0.42**	24.07	**1.16**
	TAP	10.13	19.88	0.49	71.1	3.1
	WCL	10.33	20.12	0.71	**10.2**	15.1
Setup 2	*HM*	**5.09**	**10.05**	**0.32**	16.12	**0.85**
	TAP	5.11	9.87	0.35	53.21	2.13
	WCL	5.12	10.1	0.62	**5.21**	14.13

Table 5.1: The mean estimated values of the parameters for Gneiting space-time covariance. The results are based on 100 runs of simulations.

The predictive performance of HM and TAP methods with the use of RMSPE provided in the equation (5.7) is quite similar. Note that, the convergence is slower for the larger spatial and temporal scales dependence and the predictive performance decreases when the range increases for all the methods. As was noted in the section 4.3, for $\mathcal{H}$-matrix approach there exists an influence of the scaling distance parameters a and b on the time of computation. To conclude, a preferred choice of the method depends on whether it is worth to reduce the computational time at the expense of accuracy.

5.3 Real data application

In this section we apply $\mathcal{H}$-matrix approach (HM), Covariance Tapering (TAP) and Fixed Rank Kriging (FRK) methods on the real dataset. The method of WCL was not included in the analysis due to its infeasibility for the kriging prediction for a large dataset.

The US National Aeronautics and Space Administration (NASA) launched the Aqua satellite on May 04 2002 with several instruments on board including the Atmospheric Infrared Sounder (*AIRS*). *AIRS* retrieves column-averaged CO2, denoted *XCO2* (with particular sensitivity in the mid-troposphere), amongst other geophysical quantities we use *XCO2* measurements taken between May 01 2003 and May 03 2003. These data are a subset of those available with FRK. The map of the data sites can be seen on the Figure 5.6.

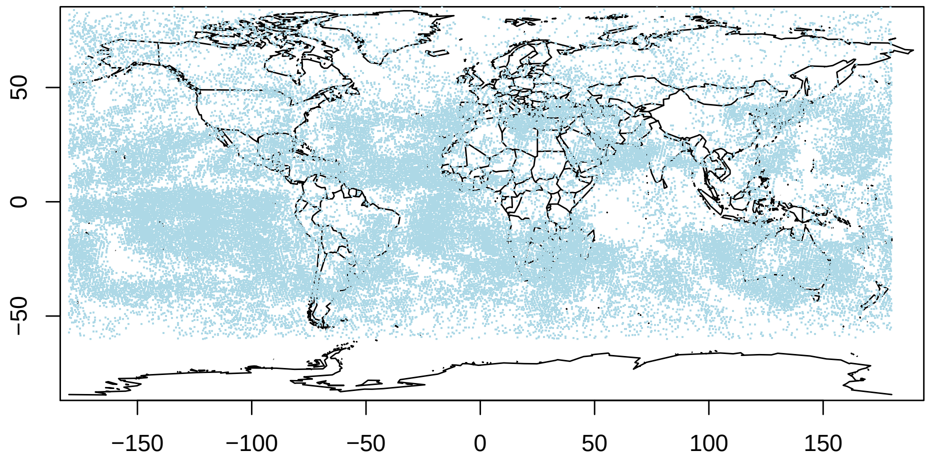


FIGURE 5.6: *AIRS*: map of the sites

We compare *HM*, *TAP* and *FRK* with the application on the 3-day *AIRS* dataset ($N = 43059$ measurements). We divide the data into a training dataset of 30000 observations chosen at random and a validation dataset containing the remaining observations.

As can be seen from the plots depicted on the Figure 5.7, CO_2 mole fraction has a latitudinal gradient (lat), therefore we use latitude as a covariate and consider process $Y(x)$

$$Y(x) = \boldsymbol{t}^T\alpha + Z(x),$$

where $Z(x)$ is the spatial process with the Matérn covariance function $C_Z + \tau^2 I$, where the measurement error (or so called nugget effect) is represented by $\tau^2 I$ for $i = j$, $\boldsymbol{t} = (t(x_1), \ldots, t(x_N))^T$ is a vector of a covariate (lat) and the coefficients $\alpha = (\alpha_1, \ldots \alpha_p)$ are unknown. We fit a linear model to the latitude as the covariate. After fitting, we noticed a strong departure from Gaussianity. Therefore, this real dataset is a scenario when the model is potentially misspecified due to a probable non-stationarity or non-Gaussian errors.

With the obtained OLS estimate for α, the data are detrended, i.e. $\tilde{Z}(x) = Y(x) - \boldsymbol{t}^T\hat{\alpha}$. Then we estimate the parameters $\theta = (\varphi, \sigma^2)$ and nugget τ^2 of the Matérn covariance function (1.3) with $\nu = 0.5$ (exponential type). Then the plug-in predictor using the estimated covariance function with ML parameters $\hat{\theta}$

$$\tilde{Y}(x_0) = t(x_0)^T\hat{\alpha} + \boldsymbol{c}(x_0)^T\tilde{C}^{-1}(\boldsymbol{Y} - \boldsymbol{t}^T\hat{\alpha}),$$

$$\hat{\alpha} = (\boldsymbol{t}^T\tilde{C}^{-1}\boldsymbol{t})^{-1}\boldsymbol{t}^T\tilde{C}^{-1}\boldsymbol{Y},$$

where $\tilde{C}$ is the covariance matrix approximated under FRK, TAP and HM methods, $\boldsymbol{t}$ is the unit vector and $t(x_0) = 1$.

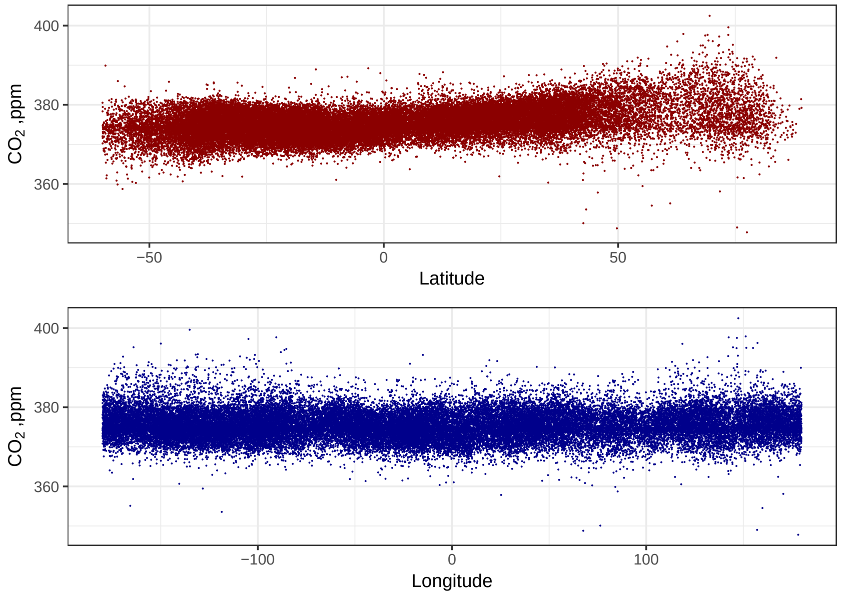


FIGURE 5.7: *AIRS*: CO_2 mole fraction on latitude (top) and longitude (bottom)

Cressie (1993) pointed out that estimating the variogram at a particular lag may be obtained as estimation of the location of the square differences $[Z(x) - Z(x+h)]^2$ or as second order moment $[Z(x) - Z(x+h)]$. In order to use available knowledge of robust location estimation, Cressie and Hawkins (1980) obtained fourth roots of squared differences, yielding robust estimators. Thus, to obtain initial values for the parameters $\theta_0 = (\varphi_0, \sigma_0^2)$ and nugget τ_0^2 of the Matérn covariance function, we exploited robust variogram estimator described by Cressie and Hawkins (1980). The initial parameters were found to be $(\varphi_0, \sigma_0^2, \tau_0^2) = (5.14, 7.8, 4.05)$.

For this application we used the newest implementation of the FRK method in the *R*-package *FRK* of Zammit-Mangion and Cressie (2017) which is based on the construction of a spatial random effects (SRE) model on a fine-resolution discretised spatial domain termed as basic areal units[4]. Moreover, for FRK application we selected irregular allocation of basis functions from (1.6) of the exponential type and domain plane with three resolutions. For the TAP method we used the *R*-package *spam* with the Wendland type of correlation function as in (3.32) with the taper range $\delta = 5$ based on the preliminary estimate of φ_0.

Method	RMSPE	Time (lik), min	Time (kr), min
HM	**3.09**	**18.6**	**1.6**
TAP	3.12	98	3.1
FRK	3.14	24	2.32

Table 5.2: The results on the *AIRS* dataset application with the best values denoted by the bold font

For the $\mathcal{H}$-matrices approach we used admissibility condition adjusted for the preliminary estimate of the range parameter φ_0 discussed in 4.3 and $\eta = 2$. The minimal size of the dense blocks $n_{min} = 80$ was chosen satisfying the condition (4.7).

The estimated parameters with the HM method are $(\hat{\varphi}, \hat{\sigma}^2, \hat{\tau}^2) = (6.2, 7.1, 8.1)$. The results are shown in the Table 5.2 which confirm the better performance of the HM method. Therefore, we conclude that $\mathcal{H}$-matrices approach leads to the computational and statistical advantages over FRK and TAP methods.

[4] available at *https://cran.r-project.org/web/packages/FRK/index.html*

Chapter 6

Conclusions

6.1 Concluding remarks

This research aimed to address one of the most common challenges arising in spatial statistics, such as the "big N problem" that emerges when the number of spatial data sites is large and, thus, the statistical inference is limited by the computational complexity of the likelihood evaluation. The problem is further increased by including the additional time dimension. In this work we analysed the existing methods to tackle the computational problem, such as fitting Gaussian Markov Random Field to GRF obtained as a solution to Stochastic Partial Differential Equation, Fixed Rank Kriging, Covariance Tapering. Despite that the application of aforementioned methods can result in reducing the computational cost, they often sacrifice statistical efficiency. This fact motivated the use of a relatively new method based on $\mathcal{H}$-matrices for the approximation of some classes of covariance functions.

In order to get insight into the $\mathcal{H}$-method, we adapted the description of this approach to the framework of spatial statistics in Chapter 2. For the appropriate partition of spatial points and effective low-rank approximation of specific blocks of covariance functions, we described the standard admissibility and asymptotic smoothness conditions required for the successful application of the $\mathcal{H}$-method. We also provided examples of analytically derived admissibility conditions for concrete types of covariance functions. Since such a computation can be tedious, we aimed to find a general condition under which this method was suitable.

Since the $\mathcal{H}$-matrix approach was originally created for the approximation of dense matrices coming from partial differential equations, the methodology was formulated in terms of Stochastic Partial Differential equations in Chapter 3. The relation of the regularity condition to physically driven PDEs was obtained through the bounds of the inverse of the pseudodifferential operator, so-called Schwartz kernel. With this link, we introduced the stochastic version of the deterministic tools in the context of the mean-square theory, where the main characteristics are defined by the mean and covariance structures. We demonstrated how a slow-growing measure relates to the covariance regularity which is required for the application of $\mathcal{H}$-matrices. We also discussed some spatio-temporal covariance functions obtained within the SPDE approach.

In Chapter 4 of this thesis we applied $\mathcal{H}$-matrix method to the Matérn covariance function of GRFs obtained as solutions to the SPDE. We also discussed the asymptotical properties of the estimators obtained with the $\mathcal{H}$-matrix technique. In addition, the properties of $\mathcal{H}$-matrices were adapted to the spatial statistics framework. Particularly, the regularity condition was scaled for the range parameter that appeared to influence the approximation of covariances by the $\mathcal{H}$-method.

The results of simulation studies with different sample size were given in Chapter 5, where we provided numerical studies with simulated and real data application for spatial and spatio-temporal datasets. The $\mathcal{H}$-matrix approach was then compared with the other methods, such as covariance tapering, fixed rank kriging and weighted composite likelihood in terms of computational and statistical efficiencies. Namely, the results of likelihood evaluation, based on real dataset of over 40000 measurements, showed fourfold superiority compared to Covariance Tapering in terms of computational time. In addition, the application of the $\mathcal{H}$-matrix approach for ML estimation resulted in a sufficient statistical efficiency even with the small compression ratio. This, in turn, entailed a gain of computational efficiency compared to the other methods. It worth to mention that the estimates obtained with the $\mathcal{H}$-matrix approach are close to ML estimates.

We conclude that the application of $\mathcal{H}$-matrices for evaluating the likelihood and performing kriging prediction allowed to preserve a balance between computational advantages and statistical efficiency.

6.2 Future directions of research

A possible drawback of the $\mathcal{H}$-matrix approach is the complexity of the underlying algorithm. In addition, the software library used for hierarchical matrices application within this thesis is available only in C programming language. However, there are other software libraries that are freely available in C, C++, Matlab, Python, that implement the hierarchical matrices approach. Therefore, as a future direction we consider the development of the R-package that implements maximum likelihood estimation and kriging prediction with the $\mathcal{H}$-matrix approximation.

In this thesis we mainly concentrated on the exact forms of the covariance functions obtained from the solutions to SPDEs. However, the $\mathcal{H}$-matrix method can be applied to the matrices obtained after Finite Element construction and triangulation of the considered domain in the same way as in Lindgren *et al.* (2011). In other words, we can exploit $\mathcal{H}$-matrices in order to solve the stochastic partial differential equation using the Finite Element method that gives a Matérn random field.

In addition, in this work the theory was adapted for the stationary case scenario. Future work may include considering non-stationary settings. It must be straightforward to adapt the asymptotic smoothness condition for anisotropic models. A directionality can be introduced through a distance measure different than the standard Euclidean distance and, in fact, the asymptotic smoothness condition that accounts for the directionality was already introduced in Hackbusch (2015). However, the more complicated non-stationary structures could be the directions of future work.

There is a growing interest in developing covariance functions for processes on the surface of a sphere. Thus, another challenge that was not addressed in this thesis is the extensions to a spherical framework. For this purpose, we can reinterpret the SPDE to be defined on $\mathbb{S}^2$ as surface embedded on $\mathbb{R}^3$ and apply $\mathcal{H}$-matrices for the discretisation of SPDEs on the spherical domain, since the solution remains a Matérn field.

Bibliography

Ambikasaran, S., Saibaba, A. K., Darve, E. F. and Kitanidis, P. K. (2013) Fast algorithms for bayesian inversion. In *Computational Challenges in the Geosciences*, pp. 101–142. Springer.

Ballani, J. and Kressner, D. (2015) Sparse inverse covariance estimation with hierarchical matrices.

Banerjee, S., Gelfand, A. E., Finley, A. O. and Sang, H. (2008) Gaussian predictive process models for large spatial data sets. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **70**(4), 825–848.

Bebendorf, M. and Hackbusch, W. (2007) Stabilized rounded addition of hierarchical matrices. *Numerical Linear Algebra with Applications* **14**(5), 407–423.

Bevilacqua, M., Gaetan, C., Mateu, J. and Porcu, E. (2012) Estimating space and space-time covariance functions for large data sets: a weighted composite likelihood approach. *Journal of the American Statistical Association* **107**(497), 268–280.

Bolin, D. (2014) Spatial matérn fields driven by non-gaussian noise. *Scandinavian Journal of Statistics* **41**(3), 557–579.

Bolin, D., Kirchner, K. and Kovács, M. (2017) Numerical solution of fractional elliptic stochastic pdes with spatial white noise. *arXiv preprint arXiv:1705.06565* .

Chen, J. and Stein, M. L. (2017) Linear-cost covariance functions for gaussian random fields. *arXiv preprint arXiv:1711.05895* .

Cressie, N. (1993) *Statistics for spatial data.* Wiley Classics Library.

Cressie, N. and Hawkins, D. M. (1980) Robust estimation of the variogram: I. *Journal of the International Association for Mathematical Geology* **12**(2), 115–125.

Cressie, N. and Huang, H.-C. (1999) Classes of nonseparable, spatio-temporal stationary covariance functions. *Journal of the American Statistical Association* **94**(448), 1330–1340.

Cressie, N. and Johannesson, G. (2008) Fixed rank kriging for very large spatial data sets. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **70**(1), 209–226.

De Iaco, S., Myers, D. E. and Posa, D. (2001) Space–time analysis using a general product–sum model. *Statistics & Probability Letters* **52**(1), 21–28.

Eidsvik, J., Shaby, B. A., Reich, B. J., Wheeler, M. and Niemi, J. (2014) Estimation and prediction in spatial models with block composite likelihoods. *Journal of Computational and Graphical Statistics* **23**(2), 295–315.

Fasshauer, G. E. (2012) Green's functions: Taking another look at kernel approximation, radial basis functions, and splines. *Approximation Theory XIII: San Antonio 2010* pp. 37–63.

Feller, W. (1957) An introduction to probability theory and its applications. *Wiley, New York* .

Furrer, R., Genton, M. G. and Nychka, D. (2006) *Journal of Computational and Graphical Statistics* **15**(3), 502–523.

Furrer, R., Sain, S. R. *et al.* (2010) spam: A sparse matrix r package with emphasis on mcmc methods for gaussian markov random fields. *Journal of Statistical Software* **36**(10), 1–25.

Gelfand, I. and Shilov, G. (1968) *Generalized functions. Vol. 2, Spaces of fundamental and generalized functions.* Academic Press, New York-London.

Geoga, C. J., Anitescu, M. and Stein, M. L. (2019) Scalable gaussian process computations using hierarchical matrices. *Journal of Computational and Graphical Statistics* pp. 1–11.

Gneiting, T. (2002) Nonseparable, stationary covariance functions for space–time data. *Journal of the American Statistical Association* **97**(458), 590–600.

Grasedyck, L. and Hackbusch, W. (2003) Construction and arithmetics of h-matrices. *Computing* **70**(4), 295–334.

Greengard, L. and Rokhlin, V. (1987) A fast algorithm for particle simulations. *Journal of computational physics* **73**(2), 325–348.

Hackbusch, W. (2015) *Hierarchical matrices: algorithms and analysis.* Volume 49. Springer, Heidelberg.

Heine, V. (1955) Models for two-dimensional stationary stochastic processes. *Biometrika* **42**(1-2), 170–178.

Iske, A., Borne, S. L. and Wende, M. (2017) Hierarchical matrix approximation for kernel-based scattered data interpolation. *SIAM Journal on Scientific Computing* **39**(5), A2287–A2316.

Jones, R. H. and Zhang, Y. (1997) Models for continuous stationary space-time processes. In *Modelling longitudinal and spatially correlated data*, pp. 289–298. Springer.

Katzfuss, M. and Cressie, N. (2011) Tutorial on fixed rank kriging (frk) of co2 data. *Department of Statistics, The Ohio State University, Columbus* .

Kaufman, C. G., Schervish, M. J. and Nychka, D. W. (2008) Covariance tapering for likelihood-based estimation in large spatial data sets. *Journal of the American Statistical Association* **103**(484), 1545–1555.

Kent, J. T., Mohammadzadeh, M. and Mosammam, A. M. (2011) The dimple in gneiting's spatial-temporal covariance model. *Biometrika* **98**(2), 489–494.

Li, J. Y., Ambikasaran, S., Darve, E. F. and Kitanidis, P. K. (2014) A kalman filter powered by h2-matrices for quasi-continuous data assimilation problems. *Water Resources Research* **50**(5), 3734–3749.

Lindgren, F. and Rue, H. (2015) Bayesian spatial modelling with r-inla. *Journal of Statistical Software* **63**(19), 1–25.

Lindgren, F., Rue, H. and Lindström, J. (2011) An explicit link between gaussian fields and gaussian markov random fields: the stochastic partial differential equation approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **73**(4), 423–498.

Lindsay, B. G. (1988) Composite likelihood methods. *Contemporary mathematics* **80**(1), 221–239.

Litvinenko, A., Sun, Y., Genton, M. G. and Keyes, D. E. (2019) Likelihood approximation with hierarchical matrices for large spatial datasets. *Computational Statistics & Data Analysis* **137**, 115–132.

Ma, C. (2003) Families of spatio-temporal stationary covariance models. *Journal of Statistical Planning and Inference* **116**(2), 489–501.

Mardia, K. V. and Marshall, R. J. (1984) Maximum likelihood estimation of models for residual covariance in spatial regression. *Biometrika* **71**(1), 135–146.

Porcu, E., Gregori, P. and Mateu, J. (2006) Nonseparable stationary anisotropic space–time covariance functions. *Stochastic Environmental Research and Risk Assessment* **21**(2), 113–122.

Reed, M. and Simon, B. (1972) *Methods of modern mathematical physics. Functional analysis*. Volume 1. Elsevier.

Rudin, W. (1973) Functional analysis, mcgraw-hill series in higher mathematics .

Sang, H. and Huang, J. Z. (2012) A full scale approximation of covariance functions for large spatial data sets. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **74**(1), 111–132.

Schwartz, L. (1966) *Théorie des distributions*. Volume 2. Hermann Paris.

Sigrist, F., Künsch, H. R. and Stahel, W. A. (2015) Stochastic partial differential equation based modelling of large space–time data sets. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **77**(1), 3–33.

Stein, M. L. (1999) *Interpolation of spatial data: some theory for kriging*. New York: Springer.

Stein, M. L. (2005) Space–time covariance functions. *Journal of the American Statistical Association* **100**(469), 310–321.

Stein, M. L. (2008) A modeling approach for large spatial datasets. *Journal of the Korean Statistical Society* **37**(1), 3–10.

Stein, M. L. (2014) Limitations on low rank approximations for covariance matrices of spatial data. *Spatial Statistics* **8**, 1–19.

Taylor, M. E. (1991) *Pseudodifferential operators and nonlinear PDE*. Volume 100. Progress in Mathematics.

Treves, F. (1967) *Topological Vector Spaces, Distributions and Kernels: Pure and Applied Mathematics*. Volume 25. Elsevier.

Varin, C., Reid, N. and Firth, D. (2011) An overview of composite likelihood methods. *Statistica Sinica* pp. 5–42.

Vecchia, A. V. (1988) Estimation and model identification for continuous spatial processes. *Journal of the Royal Statistical Society: Series B (Methodological)* **50**(2), 297–312.

Vergara, R. C., Allard, D. and Desassis, N. (2018) A general framework for spde-based stationary random fields. *arXiv preprint arXiv:1806.04999* .

Whittle, P. (1962) Topographic correlation, power-law covariance functions, and diffusion. *Biometrika* **49**(3-4), 305–314.

Whittle, P. (1963) Stochastic-processes in several dimensions. *Bulletin of the International Statistical Institute* **40**(2), 974–994.

Wong, M. (2014) *An introduction to pseudo-differential operators*. Volume 6. World Scientific Publishing Company.

Zammit-Mangion, A. and Cressie, N. (2017) Frk: An r package for spatial and spatio-temporal prediction with large datasets. *arXiv preprint arXiv:1705.08105* .

Zhang, H. (2004) Inconsistent estimation and asymptotically equal interpolations in model-based geostatistics. *Journal of the American Statistical Association* **99**(465), 250–261.

Anastasiia Gorshechnikova

CURRICULUM VITAE

Contact Information

University of Padova
Department of Statistics
via Cesare Battisti, 241-243
35121 Padova. Italy.

Tel. +39 331 920 9402
e-mail: anastasiia.gorshechnikova@studenti.unipd.it

Current Position

Since October 2016; (expected completion: September 2019/ March 2020)
PhD Student in Statistical Sciences, University of Padova.
Thesis title: "Likelihood approximation and prediction for large spatial and spatio-temporal datasets using $\mathcal{H}$-matrix approach"
Supervisor: Prof. Carlo Gaetan

Research interests

- Computational statistics
- Spatial statistics
- Spatio-temporal statistics

Education

September 2013 - July 2015
Master degree in Applied Mathematics and Computer Science.
Ufa State Aviation Technical University, Faculty of Computer Science and Robotics
Title of dissertation: "The analysis of subjective student scores based on hierarchical and panel regression modelling"
Supervisor: Prof. Irina Lackman
Final mark: with honors

September 2008 - June 2013
Engineer diploma in Telecommunication Systems.
Ufa State Aviation Technical University, Faculty of Telecommunication Systems
Title of dissertation: "Wireless network IEEE 802.11n planning and configuration at the railway station of Ufa"
Supervisor: Prof. Rustem Sultanov
Final mark: with honors

Visiting periods

February 2018 - March 2018
Winter School on Hierarchical Matrices
University of Trento,
Kiel, Germany.
Supervisor: Prof. Steffen Börm

June 2017
StartUp Research workshop
Certosa di Pontignano,
Siena, Italy
Supervisor: Prof. Marian Scott

February 2016 - July 2016
ERASMUS+ Exchange Program
University of Trento,
Trento, Italy.
Supervisor: Prof. Yannis Velegrakis

Further education

December 2015
TOEFL IBT test;
Score: 92 at proficient level C2

February 2012 - March 2013
CCNA Exploration (Network Fundamentals, Routing Protocols and Concepts, LAN Switching and Wireless, Accessing the WAN)
Cisco Networking Academy, Ufa, Russia
Instructor: Rustem Sultanov

Work experience

March 2013 - February 2016
Telecommunications Engineer
Bashtel, Ufa, Russia.

Awards and Scholarship

October 2015
ERASMUS+ Exchange Program grant holder

June 2014
Second place diploma for participation at the National Youth conference;

May 2014
First place for participation at the Scientific Youth Conference "Mathematical methods and models"

March 2014
Winner of the Academician Nikolay Fedorenko International Scientific Foundation of Economic Research (Moscow, Russian Academy of Sciences)

Computer skills

- Good knowledge of C and C++;
- Basic knowledge of Python (basic level) and Spark
- Good knowledge of R and Stata;

Language skills

Russian: native; English: fluent; Italian: moderate; French: basic.

Publications

Articles in journals

Bertarelli, G., Corbella, A., Di Iorio, J., Gorshechnikova, A., Scott, M. (2017, June). Curve Clustering for Brain Functional Activity and Synchronization. In START UP RESEARCH (pp. 75-90). Springer, Cham.

Conference presentations

Gorshechnikova A., Gaetan C., 2019. Likelihood approximation and prediction for large spatial and spatio-temporal datasets using $\mathcal{H}$-matrix approach (poster presentation); *RSS International Conference*, Belfast, Northern Ireland, September 2-5, 2019

Gorshechnikova A., Gaetan C., 2019. Likelihood approximation and prediction for large spatial and spatio-temporal datasets using $\mathcal{H}$-matrix approach (poster presentation); *The European Courses in Advanced Statistics on Statistical Analysis for Space-TimeData (ECAS2019)*, Lisbon, Portugal, July 15-17, 2019

Gorshechnikova A., Gaetan C., 2019. Likelihood approximation and prediction for large spatial and spatio-temporal datasets using $\mathcal{H}$-matrix approach (oral presentation); *Towards Spatial Data Science*, Sitges, Spain, July 10-13, 2019

Gorshechnikova A., Gaetan C., 2018. Fast Approximation of Covariance Functions Using a Hierarchical Matrices Approach (oral presentation); *CFE-CMStatistics: 11th International Conference on Computational and Methodological Statistics*, Pisa, Italy, December 14-16, 2018

Gorshechnikova A., Gaetan C., 2018. Fast Approximation of Covariance Functions Using a Hierarchical Matrices Approach (oral presentation) *RESSTE and the EcoStat network: workshop on the Theory and Practice of SPDE models and INLA*, Avignon, France, November 7-9, 2018

Other Interests

Travelling, swimming, fantasy books

References

Prof. Carlo Gaetan
Ca' Foscari University of Venice
Dipartimento di Scienze Ambientali, Informatica e Statistica - DAIS
Via Torino, 155, I-30172 Mestre (VE) ITALY
Phone: +39 041 234 8404