Computational Methods for Large Spatio-temporal Datasets and Functional Data Ranking

Dissertation

Huang Huang

In Partial Fulfillment of the Requirements

For the Degree of

Doctor of Philosophy

King Abdullah University of Science and Technology
Thuwal, Kingdom of Saudi Arabia

June, 2017
The dissertation of Huang Huang is approved by the examination committee

Committee Chairperson: Ying Sun
Committee Members: Mohamed-Slim Alouini, Marc G. Genton, David E. Keyes
ABSTRACT

Computational Methods for Large Spatio-temporal Datasets and Functional Data Ranking
Huang Huang

This thesis focuses on two topics, computational methods for large spatial datasets and functional data ranking. Both are tackling the challenges of big and high-dimensional data.

The first topic is motivated by the prohibitive computational burden in fitting Gaussian process models to large and irregularly spaced spatial datasets. Various approximation methods have been introduced to reduce the computational cost, but many rely on unrealistic assumptions about the process and retaining statistical efficiency remains an issue. We propose a new scheme to approximate the maximum likelihood estimator and the kriging predictor when the exact computation is infeasible. The proposed method provides different types of hierarchical low-rank approximations that are both computationally and statistically efficient. We explore the improvement of the approximation theoretically and investigate the performance by simulations. For real applications, we analyze a soil moisture dataset with 2 million measurements with the hierarchical low-rank approximation and apply the proposed fast kriging to fill gaps for satellite images.

The second topic is motivated by rank-based outlier detection methods for functional data. Compared to magnitude outliers, it is more challenging to detect shape outliers as they are often masked among samples. We develop a new notion of functional data depth by taking the integration of a univariate depth function. Having a form of the integrated depth, it shares many desirable features. Furthermore, the
novel formation leads to a useful decomposition for detecting both shape and magnitude outliers. Our simulation studies show the proposed outlier detection procedure outperforms competitors in various outlier models. We also illustrate our methodology using real datasets of curves, images, and video frames. Finally, we introduce the functional data ranking technique to spatio-temporal statistics for visualizing and assessing covariance properties, such as separability and full symmetry. We formulate test functions as functions of temporal lags for each pair of spatial locations and develop a rank-based testing procedure induced by functional data depth for assessing these properties. The method is illustrated using simulated data from widely used spatio-temporal covariance models, as well as real datasets from weather stations and climate model outputs.
ACKNOWLEDGEMENTS

I would like to thank my supervisor Prof. Ying Sun for her guidance and valuable discussions for my research, inspiring me to try out all the possible ideas. Sincere thanks to Prof. Mohamed-Slim Alouini, Prof. Marc G. Genton, and Prof. David E. Keyes, for being my committee members and teaching me useful graduate courses. I am also grateful to my friends for their support and help. Last but not least, I would like to thank KAUST for providing me such a good opportunity to work on interesting research topics.
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Examination Committee Page</td>
<td>2</td>
</tr>
<tr>
<td>Copyright</td>
<td>3</td>
</tr>
<tr>
<td>Abstract</td>
<td>4</td>
</tr>
<tr>
<td>Acknowledgements</td>
<td>6</td>
</tr>
<tr>
<td>List of Figures</td>
<td>11</td>
</tr>
<tr>
<td>List of Tables</td>
<td>15</td>
</tr>
<tr>
<td>1 Introduction</td>
<td>16</td>
</tr>
<tr>
<td>1.1 Motivation and objective</td>
<td>16</td>
</tr>
<tr>
<td>1.2 Thesis structure</td>
<td>16</td>
</tr>
<tr>
<td>2 Hierarchical Low-Rank Approximation of Likelihoods for Large Spatial Datasets</td>
<td>19</td>
</tr>
<tr>
<td>2.1 Introduction</td>
<td>19</td>
</tr>
<tr>
<td>2.2 Methodology</td>
<td>21</td>
</tr>
<tr>
<td>2.2.1 Approximating likelihoods</td>
<td>21</td>
</tr>
<tr>
<td>2.2.2 Hierarchical low-rank representation</td>
<td>23</td>
</tr>
<tr>
<td>2.2.3 Hierarchical low-rank approximation method</td>
<td>24</td>
</tr>
<tr>
<td>2.2.4 Assessing model quality</td>
<td>26</td>
</tr>
<tr>
<td>2.2.5 Computational complexity and parallelization</td>
<td>28</td>
</tr>
<tr>
<td>2.3 Numerical study</td>
<td>29</td>
</tr>
<tr>
<td>2.3.1 Design setup</td>
<td>29</td>
</tr>
<tr>
<td>2.3.2 Dependence level</td>
<td>30</td>
</tr>
<tr>
<td>2.3.3 Smoothness level</td>
<td>31</td>
</tr>
<tr>
<td>2.3.4 Noise level</td>
<td>31</td>
</tr>
<tr>
<td>2.3.5 Sample size and rank</td>
<td>31</td>
</tr>
<tr>
<td>2.4 Simulation study</td>
<td>33</td>
</tr>
</tbody>
</table>
5 Fast Kriging with Hierarchical Low-Rank Approximation for Large Spatial Datasets

5.1 Introduction .............................................. 83
5.2 Kriging .................................................. 85
5.3 HLR kriging .............................................. 87
5.4 Numerical studies ........................................ 88
5.5 Application to satellite images ......................... 90
5.6 Discussion and future work ............................ 92

6 Conclusion .................................................. 94

References ................................................... 97

Appendices .................................................... 106
## LIST OF FIGURES

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1</td>
<td>A random field where $n = 5$ locations have observations.</td>
<td>26</td>
</tr>
<tr>
<td>2.2</td>
<td>Four panels showing the Kullback–Leibler divergence against rank with 900 locations in IND (black), NN (green), SUM (orange), NNSUM (blue), and HLR (red) methods. The corresponding parameters are indicated in the titles.</td>
<td>32</td>
</tr>
<tr>
<td>2.3</td>
<td>Two panels showing the Kullback–Leibler divergence against rank with 2500 locations in IND (black), NN (green), SUM (orange), NNSUM (blue), and HLR (red) methods. The corresponding parameters are indicated in the titles.</td>
<td>33</td>
</tr>
<tr>
<td>2.4</td>
<td>Two panels showing the boxplot of parameter estimates and mean squared error. The solid line is a reference for the true parameter value, and the dash line with squares ( - - □ - - ) is the corresponding root mean squared error of the 100 times of estimates in each method. Left: illustration for estimated $\alpha$; Right: illustration for estimated $\beta$.</td>
<td>34</td>
</tr>
<tr>
<td>2.5</td>
<td>Soil moisture (unit: percentage) at the top layer of the Mississippi basin, U.S.A. on January 1, 2004.</td>
<td>35</td>
</tr>
<tr>
<td>2.6</td>
<td>Left: the histogram of the transformed residuals; Right: the image plot of the empirical variogram at different distances and along different directions.</td>
<td>37</td>
</tr>
<tr>
<td>3.1</td>
<td>Visualization tools: (a) Observations with a detected magnitude outlier (red) and shape outliers (orange). (b) Shape outlyingness plot with points in orange corresponding to shape outliers. (c) Functional boxplot using TVD after removing detected shape outliers.</td>
<td>49</td>
</tr>
<tr>
<td>3.2</td>
<td>Top panel: 100 curves (grey) from Model 1 with one of different outliers from Model 2 (red), Model 3 (orange), Model 4 (blue), Model 5 (green) and Model 6 (black), respectively. Bottom panel: 100 curves (grey) contaminated by one outlier from Model 7 (purple).</td>
<td>51</td>
</tr>
</tbody>
</table>
3.3 The central regions constructed by MBD (left), ED (middle), and TVD (right) for Model 3 (top), Model 4 (middle) and Model 7 (bottom). The central region is displayed by the pink polygon. The solid black line in the middle is the median curve, which has the largest depth value. All the grey lines in the background are the dataset curves.

3.4 The sea surface temperatures in a Niño zone from July 1982 to June 2016. (a) The original data. (b) The boxplot of MSS. (c) The functional boxplot with TVD.

3.5 Runs of the sea surface height model. (a) The median. (b) One of the shape outliers. (c) One of the magnitude outliers.

3.6 The video frames. (a) The median. (b) One of the shape outliers. (c) One of the magnitude outliers.

4.1 Top: functional boxplots of the separability test functions for the Gneiting model $C^G(h, u; \beta)$ with different values of $\beta$. The magenta area is the central region, the black line is the median, and the red lines are outliers. Bottom: contour plots of the covariances for the Gneiting model $C^G(h, u; \beta)$ with different values of $\beta$.

4.2 Top: functional boxplots of the separability test functions for the Cressie and Huang separable model $C^{CH}_\text{sep}(h, u)$ and non-separable model $C^{CH}_\text{nsep}(h, u)$. The magenta area is the central region, the black line is the median, and the red lines are outliers. Bottom: contour plots of the covariances for $C^{CH}_\text{sep}(h, u)$ and $C^{CH}_\text{nsep}(h, u)$.

4.3 Top: functional boxplots of the separability test functions for the De Cesare model $C^C(h, u; k)$ with different values of $k$. The magenta area is the central region, the black line is the median, and the red lines are outliers. Bottom: contour plots of covariances for $C^C(h, u; k)$ with different values of $k$.

4.4 Top: functional boxplots of the symmetry test function for the Gneiting model $C^{G'}(h, u; \lambda)$ with different values of $\lambda$. The magenta area is the central region, the black line is the median, and the red lines are outliers. Bottom: contour plots of covariances for $C^{G'}(h, u; \lambda)$ with different values of $k$ where $h_2$ is fixed to 0.
4.5 Functional boxplots of the estimated separability test function for data generated from the Gneiting model $C^G(h, u; \beta)$ with $\beta = 0, 0.5, \text{and } 1$. The magenta area is the central region, the black line is the median, the red lines are outliers, and the green dash line is the reference for separability in theory.

4.6 Functional boxplot of the estimated separability test function for data generated from Cressie-Huang models $C_{nsep}^{CH}$ and $C_{sep}^{CH}$. The magenta area is the central region, the black line is the median, the red lines are outliers, and the green dash line is the reference for separability in theory.

4.7 Functional boxplots of the estimated symmetry test function for data generated from the Gneiting model $C^G(h, u; \lambda)$ with $\lambda = 0, 0.05, \text{and } 0.1$. The magenta area is the central region, the black line is the median, the red lines are outliers, and the green dash line is the reference for separability in theory.

4.8 Ten locations for observing wind speed in a broad (left) and zoomed (right) view. The green points are grouped in the coastal class and the orange points are grouped in the inland class.

4.9 The estimated separability test functions for wind speed in the northwestern U.S.A. during summer and winter, in coastal and inland regions. The p-values of the rank-based test for separability are indicated in the title of each case.

4.10 The estimated symmetry test functions for wind speed in the northwestern U.S.A. during summer and winter, in coastal and inland regions. The p-values of the rank-based test for symmetry are indicated in the title of each case.

4.11 Study area in the North Atlantic Ocean, where the white dots indicate the spatial locations from the model outputs.

4.12 Mean of the sea surface temperature (unit: K) during 1976–1979 at each of the 16 locations.

4.13 The estimated separability test functions for daily surface temperature with different combinations of GCMs and RCMs. The p-values of the rank-based test for separability are indicated in the title of each case.

4.14 The estimated separability test functions for daily wind speed with different combinations of GCMs and RCMs.
4.15 The estimated symmetry test functions for daily wind speed with different combinations of GCMs and RCMs. . . . . . . . . . . . . . . . . 81

5.1 Four panels showing the exact kriging weights (black) and HLR kriging weights (red) for 100 locations with rank \( r = 2 \). The blue line is the absolute difference between the exact and HLR kriging weights. . . . 90

5.2 Four panels showing the exact kriging weights (black) and HLR kriging weights (red) for 100 locations with rank \( r = 10 \). The blue line is the absolute difference between the exact and HLR kriging weights. . . . 91

5.3 An illustration of Landsat 7 satellite image with gaps. . . . . . . . . . 92

5.4 The Landsat 7 satellite image with gaps for the study area (top-left). Image with gap filled by the HLR kriging with rank \( r = 2 \) (top-right) and \( r = 10 \) (bottom-left), and the exact kriging (bottom-right). . . . 93
# LIST OF TABLES

<table>
<thead>
<tr>
<th>Table</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1</td>
<td>Description of the five methods used in the numerical study</td>
<td>30</td>
</tr>
<tr>
<td>2.2</td>
<td>Parameter estimation results</td>
<td>38</td>
</tr>
<tr>
<td>3.1</td>
<td>Results of outlier detection using different methods for different models.</td>
<td>54</td>
</tr>
<tr>
<td></td>
<td>The values are the means in the 500 experiments in each case, and the values in brackets are the corresponding standard deviations.</td>
<td></td>
</tr>
<tr>
<td>4.1</td>
<td>Notations of the chosen covariance models.</td>
<td>64</td>
</tr>
<tr>
<td>4.2</td>
<td>Type I error and power for simulated data from the Gneiting model ( C^G(h, u; \beta) ) with different values of ( \beta ) in 100 experiments.</td>
<td>72</td>
</tr>
<tr>
<td>4.3</td>
<td>Proportion of rejections for simulated data from the Cressie-Huang separable and non-separable models in 100 experiments.</td>
<td>73</td>
</tr>
<tr>
<td>4.4</td>
<td>Type I error and power for simulated data from the Gneiting model ( C^{G'}(h, u; \lambda) ) with different values of ( \lambda ) in 100 experiments.</td>
<td>75</td>
</tr>
</tbody>
</table>
Chapter 1

Introduction

1.1 Motivation and objective

The development of modern technology has facilitated the collection of large-scale datasets, either from observations or simulations. Statistical methodology often faces tremendous challenges of big and high-dimensional data.

For example, in spatial statistics, due to the large number of observations, statistical inference on high-resolution spatial datasets suffers from the prohibitive burdens of computations and storage; in functional data analysis, where data are infinite-dimensional objects, simple extensions of univariate statistical methods to the functional setting do not properly capture features of functional data.

The aim of this thesis is to tackle two types of problems. One is to develop statistically and computationally efficient methods for fitting Gaussian process models to large and irregularly spaced spatial datasets and provide fast prediction algorithms. The other is to develop new functional data ranking methods for outlier detection, especially for detecting shape outliers, and exploit the functional data ranking technique for visualizing and assessing spatio-temporal covariance properties.

1.2 Thesis structure

In Chapter 2 we focus on large and irregularly spaced spatial datasets in the fields of climate and environmental sciences. To model such datasets, the widely used Gaussian process models in spatial statistics face tremendous challenges due to the
prohibitive computational burden. Various approximation methods have been introduced to reduce the computational cost. However, most of them rely on unrealistic assumptions about the underlying process and retaining statistical efficiency remains an issue. We develop a new approximation scheme for maximum likelihood estimation. We show how the composite likelihood method can be adapted to provide different types of hierarchical low-rank approximations that are both computationally and statistically efficient. The improvement of the proposed method is explored theoretically; the performance is investigated by numerical and simulation studies; and the practicality is illustrated through applying our methods to 2 million measurements of soil moisture in the area of the Mississippi River basin, which facilitates a better understanding of the climate variability.

In nonparametric statistics, there has been extensive work on data depth-based methods for multivariate data and high-dimensional data. Recent developments have moved to infinite-dimensional objects such as functional data. In Chapter 3, we propose a notion of depth, the total variation depth, for functional data, which has many desirable features and is well suited for outlier detection. The proposed depth is of the form of an integral of a univariate depth function. We show that the novel formation of the total variation depth leads to a useful decomposition associated with shape and magnitude outlyingness of functional data. Compared to magnitude outliers, shape outliers are often masked among the rest of samples and more difficult to identify. We then further develop an effective procedure and visualization tools for detecting both types of outliers, while naturally accounting for the correlation in functional data. The outlier detection performance is investigated through simulations under various outlier models. In the end, the proposed methodology is demonstrated using real datasets of curves, images, and video frames.

In Chapter 4, we bring the functional data ranking technique to spatio-temporal statistics. Spatio-temporal covariances are very important in modeling geostatistical
data and are often used to describe the spatio-temporal variability of the underlying random process. For second-order stationary processes, there exist subclasses of covariance functions by assuming simpler spatio-temporal dependence structure, such as separability and full symmetry. However, it is challenging to visualize and assess such properties from spatio-temporal observations. We propose a functional data analysis approach by constructing test functions with cross-covariances of each pair of the time series observed from spatial locations. These test functions are functions in temporal lag and summarize the property of separability or symmetry for given spatial pairs. We then apply the functional boxplot to visualize the functional median and the variability of the test functions, where the extent of departure from zero at all temporal lags indicates the degree of non-separability or asymmetry. We also develop a rank-based nonparametric testing procedure for assessing the significance of the non-separability or asymmetry. The performance of the proposed methods is examined by simulation for various commonly used spatio-temporal covariance models. For applications, we demonstrate our method on different types of real datasets, including weather station data and climate model outputs.

In Chapter 5, we study another important problem in geostatistics—prediction or interpolation for large spatial datasets. The best linear unbiased predictor, or kriging predictor, requires solving a linear system. The computation often becomes prohibitive for large and irregularly spaced spatial datasets. We propose a hierarchical low-rank kriging for fast prediction using the hierarchical low-rank approximation framework developed in Chapter 2. We examine the approximation quality by numerical studies and apply the algorithms for filling gaps in satellite images.

In Chapter 6, we summarize the contribution of our work and point out how each of our proposed methods could be improved further for more complicated problems.
Chapter 2

Hierarchical Low-Rank Approximation of Likelihoods for Large Spatial Datasets

2.1 Introduction

Soil moisture is a key factor in climate systems, which has a significant impact on hydrological processes, runoff generations and drought developments. To understand its spatial variability and predict values at unsampled locations, Gaussian process models are widely used [1], where likelihood based methods are appropriate for model fitting. However, it generally requires \( O(n^3) \) computations and \( O(n^2) \) memory for \( n \) irregularly spaced locations [2]. Similar to other climate variables, many satellite-based or numerical model generated soil moisture datasets have nearly a global coverage with high spatial resolutions, so that the exact computation of Gaussian likelihood becomes prohibitive. There are various existing methods, many of which were discussed by Sun et al. [3]. For example, covariance tapering [4, 5, 6] assumes a compactly supported covariance function, which leads to a sparse covariance matrix; low-rank models, including space-time Kalman filtering [7], low-rank splines [8], moving averages [9], predictive processes [10] and fixed rank kriging [11], make use of a latent process with a lower dimension where the resulting covariance matrix has a low-rank representation; and Markov random field models [12, 13, 14, 15] exploit fast-approximated conditional distributions assuming conditional independence with the precision matrix being sparse. These methods use models that may allow exact computations to reduce computations and/or storage, and each has its strength and weakness.
For instance, Stein [16] studied the properties of the covariance tapers and showed that covariance tapering sometimes performs even worse than assuming independent blocks in the covariance; Stein [17] discussed the limitations on the low-rank approximations; and Markov models depend on the observation locations, and realignment to a much finer grid with missing values is required for irregular locations [2]. Recently developed methods include the nearest-neighbor Gaussian process model [18], which is used as a sparsity-inducing prior within a Bayesian hierarchical modeling framework, the multiresolution Gaussian process model [19], which constructs basis functions using compactly supported correlation function on different level of grids, equivalent kriging [20], which uses an equivalent kernel to approximate the kriging weight function when a nontrivial nugget exists, and multi-level restricted Gaussian maximum likelihood method [21], for estimating the covariance function parameters using contrasts.

An alternative way to reduce computations is via likelihood and score equation approximations. Vecchia [22] first proposed to approximate the likelihood using the composite likelihood method, where the conditional densities were calculated by choosing only a subset of the complete conditioning set. Stein et al. [23] adapted this method for restricted maximum likelihoods approximation. Instead of approximating the likelihood itself, Sun and Stein [2] proposed new unbiased estimating equations for score equation approximation, where the sparse precision matrix approximation is constructed by a similar method. In these approximation methods, the exact likelihood and the score equations can be obtained by using the complete conditioning set to calculate each conditional density. It was shown that the approximation quality or the statistical efficiency depends on the selected size of the subset. It is common that the subset is still inadequate by considering the largest possible number of nearest neighbors, which motivates this work.

In this chapter, we propose a generalized hierarchical low-rank method for likeli-
hood approximation. The proposed method utilizes low-rank approximations hierarchically, which does not lead to a low-rank covariance matrix approximation. Therefore, it is different from the predictive process method \cite{10}, where the covariance matrix is approximated by a low-rank representation. Furthermore, the proposed method contains the independent blocks \cite{16} and nearest neighbors \cite{2} approaches as special cases. The improvement of the proposed method is explored theoretically and the performance is investigated by numerical and simulation studies. We show that the hierarchical low-rank approximation significantly improves the statistical efficiency of the most commonly used methods while retaining the computational efficiency, especially when the size of conditional subsets is restricted by the computational capacity, which is always the case for real datasets. For illustrations, our method is applied to a large real-world spatial dataset of soil moisture in the Mississippi River basin, U.S.A., to facilitate a better understanding of the hydrological process and climate variability. Our method is able to fit a Gaussian process model to 2 million measurements with fast computations, making it practical and attractive for very large datasets.

2.2 Methodology

2.2.1 Approximating likelihoods

Let \( \{ Z(s) : s \in D \subset \mathbb{R}^d \} \) be a stationary isotropic Gaussian Process in a domain \( D \) in the \( d \)-dimensional Euclidean space, and typically \( d = 2 \). We assume the mean of the process is zero for simplicity and the covariance function has a parametric form \( C(h; \theta) = \text{Cov}\{Z(s), z(s')\} \), where \( h = \|s - s'\| \) and \( \theta \) is the parameter vector of length \( p \). Suppose that data are observed at \( n \) irregularly spaced locations \( s_1, \ldots, s_n \), then,

\[
Z = (Z_1, \ldots, Z_n)^T \sim N(0, \Sigma(\theta)),
\]
where \( Z_i = Z(s_i), \) \( i = 1, \ldots, n \), and \( \Sigma(\theta) \) is the variance-covariance matrix with the \((i, j)\)-th element \( C(\|s_i - s_j\|; \theta) \). For simplicity, \( \theta \) is omitted in notations hereinafter unless clarification is needed.

The Maximum Likelihood Estimation (MLE) can be obtained by maximizing the log-likelihood,

\[
\ell(\theta \mid Z) = \log \{f(Z \mid \theta)\} = -\frac{1}{2} \log(|\Sigma|) - \frac{1}{2} Z^T \Sigma^{-1} Z - \frac{n}{2} \log(2\pi),
\]

where \( f \) is the multivariate normal density. In practice, if the mean of \( Z \) is a vector that depends linearly on unknown parameters, the restricted maximum likelihood estimate should be employed \[23\].

When computations become prohibitive, one way to approximate the likelihood is through log-conditional densities,

\[
\ell(\theta \mid Z) = \log \{f(Z_1 \mid \theta)\} + \sum_{j=1}^{n-1} \log \{f(Z_{j+1} \mid Z_j, \theta)\},
\]

where \( Z_j = (Z_1, \ldots, Z_j)^T \), for \( 1 \leq j \leq n - 1 \), indicating all the “past” observations of \( Z_{j+1} \). Since,

\[
\text{Cov} \left( \begin{array}{c} Z_j \\ Z_{j+1} \end{array} \right) = \begin{pmatrix} \Sigma_{jj} & \sigma_j \\ \sigma_j^T & \sigma_{j+1,j+1} \end{pmatrix},
\]

it is easy to show that for \( j = 1, \ldots, n - 1 \),

\[
\log \{f(Z_{j+1} \mid Z_j)\} = -\frac{1}{2} \left\{ \frac{(Z_{j+1} - \sigma_j^T \Sigma_{jj}^{-1} Z_j)^2}{\sigma_{j+1,j+1} - \sigma_j^T \Sigma_{jj}^{-1} \sigma_j} + \log \left( \sigma_{j+1,j+1} - \sigma_j^T \Sigma_{jj}^{-1} \sigma_j \right) + \log(2\pi) \right\}
\]

(2.1)

which is the log-density of \( W_j = b_j^T Z \), where \( b_j = (-\sigma_j^T \Sigma_{jj}^{-1}, 1, 0, \ldots, 0)^T \). It can be shown that \( W_j \)s are independent and \( W_j \sim N(0, V_j) \), where \( V_j = b_j^T \Sigma b_j \) \[23\]. Sun and Stein \[2\] further showed that the precision matrix is \( \Sigma^{-1} = \sum_{j=0}^{n-1} b_j b_j^T / V_j \), where \( b_0 = (1, 0, \ldots, 0)^T \) and \( V_0 = b_0^T \Sigma b_0 \).
More generally, $Z_{j+1}$ can be defined as a vector which is usually more computationally efficient, and the corresponding $b_j = (-\sigma_j^T \Sigma_{jj}^{-1} I, 0, \ldots, 0)^T$, where $I$ is an identity matrix of size $j$.

However, for a large $j$, it is computationally expensive to evaluate $\Sigma_{jj}^{-1} \sigma_j$. Vecchia \cite{Vecchia1993} proposed approximating each conditional density by only conditioning on a subset $Z_{j+1}$ consisting of $r \ll j$ nearest neighbors. The same approach is used by Stein et al. \cite{Stein1992} for approximating the restricted maximum likelihood estimate. Sun and Stein \cite{Sun2007} also used the subset of nearest neighbors to approximate the precision matrix for score equation approximation.

In this chapter, we propose a generalized framework that allows to approximate these conditional densities hierarchically using a low-rank representation. Although we implement our algorithm for application in \cite{2.5} with $Z_{j+1}$ being a vector, we present and illustrate our methodology assuming $Z_{j+1}$ is scalar for simplicity.

### 2.2.2 Hierarchical low-rank representation

Motivated by the nearest neighbors method, where only $r \ll j$ nearest neighbors are selected to approximate $\Sigma_{jj}^{-1} \sigma_j$ for a large $j$ in Equation (2.1), we propose a general approximation framework for $j > r$ using a low-rank representation.

Denote $\Sigma_{jj}^{-1} \sigma_j$ by $x_j$, or $\Sigma_{jj} x_j = \sigma_j$. We propose to approximate $x_j$ by a low-rank representation $\hat{x}_j = A_{j,r} \tilde{x}_j$, where $\tilde{x}_j$ is a vector of length $r$ and $A_{j,r}$ is a $j \times r$ matrix. Then, instead of solving $\Sigma_j x_j = \sigma_j$, we minimize the norm $\|\Sigma_{jj} A_{j,r} \tilde{x}_j - \sigma_j\|_{\Sigma_{jj}^{-1}} = (\Sigma_{jj} A_{j,r} \tilde{x}_j - \sigma_j)^T (\Sigma_{jj}^{-1})(\Sigma_{jj} A_{j,r} \tilde{x}_j - \sigma_j)$ or equivalently solve $A_{j,r}^T \Sigma_{jj} A_{j,r} \tilde{x}_j = A_{j,r}^T \sigma_j$. Therefore, $x_j$ is approximated by,

$$\hat{x}_j = A_{j,r} \tilde{x}_j = A_{j,r} (A_{j,r}^T \Sigma_{jj} A_{j,r})^{-1} A_{j,r}^T \sigma_j,$$

which only involves a linear solve of dimension $r$. In this framework, we approximate
for each $j > r$ hierarchically by a low-rank representation, which includes many commonly used strategies as special cases with different choices of $A_{j,r}$. The following are some examples:

**Example 1. Independent Blocks Method (IND).** In this method, no correlation between “past” points and the “current” point is considered. Namely, $A_{j,r}$ is a $0$ matrix; however, $Z_{j+1}$ is a vector of length $r$ here for fair comparison to other methods in terms of computation.

**Example 2. Nearest Neighbors Method (NN).** Choose $r$ nearest neighbors of $Z_{j+1}$ from $Z_j$. The corresponding $A_{j,r}$ is of $j \times r$ dimensions, where each column consists of only one element $1$ at the $k$-th row if $Z_k$ is selected from $Z_j$ and zero otherwise.

**Example 3. Nearest Neighboring Sets Method (SUM).** Choose $r$ nearest neighboring sets of $Z_{j+1}$, where each set contains $m > 1$ neighbors and a total of $mr \ll j$ neighbors are selected from $Z_j$. The matrix $A_{j,r}$ is specified as a $j \times r$ matrix with each column having $m$ elements of $1$, indicating the sum of the $m$ selected neighbors are considered. In this way, more neighbors are included while the computational cost remains the same.

**Example 4. Nearest Neighbors and Nearest Neighboring Sets Method (NNSUM).** Combine Examples 2 and 3, where $r_1$ columns of $A_{j,r}$ are constructed as in Example 2, and $r - r_1$ are built as in Example 3. In this way, we use the exact information from the $r_1$ nearest neighbors and consider $r - r_1$ nearest neighboring sets with a total number of $r_1 + m(r - r_1)$ selected nearest neighbors.

### 2.2.3 Hierarchical low-rank approximation method

In this section, we propose a generalized Hierarchical Low-Rank Approximation Method (HLR). In Equation (2.2), the matrix $A_{j,r}$ is a 0-1 matrix. The $r \times r$ matrix $A_{j,r}^T \Sigma_{jj} A_{j,r}$ only extracts the corresponding rows or columns of $\Sigma_{jj}$. Now suppose we select $mr$
nearest neighbors of $Z_{j+1}$, and the corresponding $A_{j, mr}$ is of size $j \times mr$. To retain the same computational costs associated with rank $r$, we propose the following approximation,

$$A_{j, mr}^T \Sigma_{jj} A_{j, mr} \approx P_j L_j P_j^T + \epsilon_j^2 I_{mr},$$

(2.3)

where $L_j$ is a positive definite matrix of dimension $r \times r$, $P_j$ is a $mr \times r$ matrix consisting of $r$ basis functions, $I_{mr}$ is the identity matrix of size $mr$, and $\epsilon_j^2$ accounts for the fine-scale variability, or the reminders of the low-rank approximation so that the approximated matrix is invertible. By the Sherman–Morrison–Woodbury formula,

$$(P_j L_j P_j^T + \epsilon_j^2 I_{mr})^{-1} = \epsilon_j^{-2} I_{mr} - \epsilon_j^{-4} P_j (L_j^{-1} + \epsilon_j^{-2} P_j^T P_j)^{-1} P_j^T,$$

(2.4)

then $(A_{j, mr}^T \Sigma_{jj} A_{j, mr})^{-1}$ in Equation (2.2) can be approximated by only inverting an $r \times r$ matrix $L_j$.

This approach is similar to the predictive process [10] and fixed rank kriging [11]. However both methods approximate the covariance function by a low-rank representation while the low-rank approximation is done for each $j > r$ hierarchy in our method, and the resulting approximated covariance is no longer low-rank. For the choice of $P_j$, Cressie and Johannesson [11] discussed several popular options. In this thesis, we use $P_j$ consisting of eigenvectors associated with the first $r$ eigenvalues of $A_{j, mr}^T \Sigma_{jj} A_{j, mr}$, and $\epsilon_j$ is then chosen to be the $(r + 1)$-th eigenvalue. Specifically, suppose $A_{j, mr}^T \Sigma_{jj} A_{j, mr} = \sum_{i=1}^{mr} \lambda_i u_i u_i^T$, where $\lambda_1 \geq \cdots \geq \lambda_{mr}$ are the eigenvalues and $u_i$s are the corresponding eigenvectors. Let $P_j = (u_1, \ldots, u_r)$ be the $mr \times r$ matrix with orthogonal columns. By choosing $\epsilon_j = \lambda_{r+1}$ and $L_j = \text{diag}(\lambda_1 - \lambda_{r+1}, \ldots, \lambda_r - \lambda_{r+1})$, it is easy to derive that the low-rank approximation error is $\sum_{i=r+2}^{mr} (\lambda_i - \lambda_{r+1}) u_i u_i^T$. Although it requires extra time to obtain these eigenvalues and eigenvectors, the computation is relatively cheap because the matrix size $mr$ is generally small and we only need to compute $r + 1$ leading eigenvalues and $r$ leading eigenvectors without
solving the entire eigen-decomposition.

To help comprehend, Figure 2.1 illustrates the methods described in §2.2.2 and §2.2.3 for $n = 5$ observations $\mathbf{Z} = (Z_1, \ldots, Z_5)^T$. Let $r = 2$, then IND considers 3 independent blocks, and $f(\mathbf{Z})$ is approximated by $f(Z_5)f(Z_4, Z_3)f(Z_2, Z_1)$. For the other four methods, the conditional density is required to calculate in each hierarchy. For instance, in hierarchy $j = 5$, NN approximates the conditional density $f(Z_5 \mid Z_4, \ldots, Z_1)$ by $f(Z_5 \mid Z_4, Z_3)$; SUM by $f(Z_5 \mid Z_4 + Z_3, Z_2 + Z_1)$; NNSUM by $f(Z_5 \mid Z_4, Z_3 + Z_2)$; and HLR by $f(Z_5 \mid a_{14}Z_4 + a_{13}Z_3 + a_{12}Z_2 + a_{11}Z_1, a_{24}Z_4 + a_{23}Z_3 + a_{22}Z_2 + a_{21}Z_1)$, where $a_{ij}$'s are determined by the low-rank approximation.

2.2.4 Assessing model quality

There are various ways to measure the performance of approximation methods, including the Kullback–Leibler divergence, the Godambe information matrix, and the Frobenius norm.

The Kullback–Leibler divergence computes the divergence of the approximated from the exact distributions. For the zero-mean Gaussian process, the Kullback–

Figure 2.1: A random field where $n = 5$ locations have observations.
Leibler divergence has the closed form,

\[ D_{\text{K-L}}(N_e\|N_a) = \frac{1}{2} \left\{ \text{tr}(\Sigma_a^{-1}\Sigma_e) + \log(|\Sigma_a|) - \log(|\Sigma_e|) - n \right\}, \]

where \( N_e \) and \( N_a \) stand for the exact and the approximated distributions, respectively, \( \Sigma_e \) and \( \Sigma_a \) are the corresponding covariance matrices, and \( n \) is the dimension of the distribution.

The Godambe information matrix gives the asymptotic variances and covariances for the estimated parameters in the Gaussian process, as used by Kaufman et al. [5] and Sun and Stein [2]. The Frobenius norm is another way to think about this problem. However, it is a matrix norm and does not penalize the positive definiteness of a covariance matrix [17].

For our numerical and simulation studies in §2.3, we choose the Kullback–Leibler divergence and the Godambe Information matrix to assess the quality of the approximation. Because the results in terms of showing the different performances are similar, we only present the results of Kullback–Leibler divergence. It will be shown numerically that the Kullback–Leibler divergence of HLR is always the smallest when the rank \( r \) is small. This is due to the fact that for sufficiently large \( j \), \( j > r \), HLR provides a better approximation in Equation (2.2) by including more neighbors than the nearest neighbors method. In the following Theorem 1, we have proved that HLR always gives the smallest error in approximating the covariance matrix in Equation (2.2) in terms of matrix norm for each sufficiently large \( j \). While our numerical study agrees with this result, it also provides some insights on why HLR outperforms other methods with the smallest Kullback-Leibler divergence for the setting we have considered. Let \( V_{jj}^N \) be the \( r \times r \) matrix defined by \( A_{j,r}^T \Sigma_{jj} A_{j,r} \) in Equation (2.2) using NN and let \( V_{jj}^H = P_j L_j P_j^T + \epsilon_j^2 I_{mr} \) be the \( mr \times mr \) matrix for approximating \( A_{j,mr}^T \Sigma_{jj} A_{j,mr} \) in Equation (2.3) by HLR, where \( P_j \) consists of eigenvectors. The fol-
lowing theorem shows that the approximation to \( \Sigma_{jj} \) induced by \( V_{jj}^H \) is better than that induced by \( V_{jj}^N \) in terms of the Frobenius norm.

**Theorem 1.** Let \( \lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_{mr} > 0 \) be the eigenvalues of \( A_{j,mr}^T \Sigma_{jj} A_{j,mr} \). If \( \epsilon_j^2 \) in Equation (2.3) satisfies \( \epsilon_j^2 < (\lambda_r + \lambda_{mr})/2 \), we have,

\[
\| A_{j,mr} V_{jj}^H A_{j,mr}^T - \Sigma_{jj} \|_F \leq \| A_{j,r} V_{jj}^N A_{j,r}^T - \Sigma_{jj} \|_F,
\]

where \( \| \cdot \|_F \) means the Frobenius norm.

The proof is shown in the Appendix A. Similar results hold for the comparison between HLR and SUM, or NNSUM.

### 2.2.5 Computational complexity and parallelization

For the proposed HLR, we need to execute a linear solve of dimension \( r \), which requires \( O(\min(j,r)^3) \) computation in Equation (2.4) for each hierarchy \( j = 1, \ldots, n - 1 \) assuming that the direct method is employed. Then the total computational cost is \( O(r^3n) \) for likelihood approximation per value. When \( r \ll n^{2/3} \), the computational cost is much smaller than \( O(n^3) \), which is required by the Cholesky decomposition.

Another advantage of the proposed HLR is that it does not need to store any \( n \times n \) covariance matrix explicitly, which generally requires \( O(n^2) \) memory. In each hierarchy, our proposed method only requires the covariance matrix of the chosen neighboring points to be stored, so that it uses a much smaller amount of memory than storing the entire \( n \times n \) matrix.

In practice, the computation time can be reduced further by choosing \( Z_{j+1} \) as a vector of an appropriate size due to the fact that it leads to a smaller number of hierarchies that need to be evaluated while the increased computation in each hierarchy is comparable small. It is also worth noting that our approach can be parallelized easily because the computation of each hierarchy is independent of each
2.3 Numerical study

2.3.1 Design setup

In the numerical study in this section and the following simulation study in §2.4, we focus on irregularly spaced data with an unstructured covariance matrix [2]. The observations are generated at the locations $n^{-1/2}(r - 0.5 + X_{r\ell}, \ell - 0.5 + Y_{r\ell})$ for $r, \ell \in \{1, \ldots, n^{1/2}\}$, where $n$ is the number of locations, and $X_{r\ell}$s and $Y_{r\ell}$s are independent and identically distributed, uniform on $(-0.4, 0.4)$. The advantage of this design is that it is irregular, and we can guarantee that no two locations are too close.

Here, we study the performances of different approximation methods proposed in §2.2.2 and §2.2.3 in different settings. We consider a zero-mean Gaussian process model with Matérn covariance function possibly with a nugget,

$$C(h; \alpha, \beta, \nu, \tau^2) = \alpha \{(2\nu)^{1/2}h/\beta\}^\nu K_\nu\{(2\nu)^{1/2}h/\beta\}/\{\Gamma(\nu)2^{\nu-1}\} + \tau^2 1(h = 0), \tag{2.5}$$

where $K_\nu(\cdot)$ is the modified Bessel function of the second kind of order $\nu$, $\Gamma(\cdot)$ is the gamma function, $1(\cdot)$ is the indicator function, $h \geq 0$ is the distance between two locations, $\alpha > 0$ is the sill parameter, $\beta > 0$ is the range parameter, $\nu > 0$ is the smoothness parameter, and $\tau^2$ is the nugget effect.

For $n$ irregularly spaced locations, the description of the five methods considered is shown in Table 2.1.

In §2.3.2–2.3.4, we present the Kullback–Leibler divergence calculated from different settings for the five methods with $\alpha$ fixed at 1 and $n = 900$. In §2.3.5, we discuss the effect of sample size $n$ and the rank $r$. 
Table 2.1: Description of the five methods used in the numerical study.

<table>
<thead>
<tr>
<th>Method</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>IND</td>
<td>Divide the locations into ( \lceil n/r \rceil ) blocks, each of which contains ( r ) points. ( \lceil n/r \rceil ) means the largest integer that is no larger than ( n/r ).</td>
</tr>
<tr>
<td>NN</td>
<td>A number of ( r ) nearest neighbors are selected to construct ( A_{j,r} ).</td>
</tr>
<tr>
<td>SUM</td>
<td>A number of ( r ) nearest neighboring sets are selected and each set has 2 locations. Then a total number of ( 2r ) nearest neighbors are used to construct ( A_{j,r} ).</td>
</tr>
<tr>
<td>NNSUM</td>
<td>A number of ( \lceil r/2 \rceil ) nearest neighbors are first selected, then the following ( 2(r - \lceil r/2 \rceil) ) nearest neighbors are divided into ( r - \lceil r/2 \rceil ) sets of size 2.</td>
</tr>
<tr>
<td>HLR</td>
<td>A number of ( 2r ) nearest neighbors are considered, where ( L_j ) is a ( r \times r ) diagonal matrix with elements corresponding to the ( r ) leading eigenvalues. ( P ) consists of the ( r ) corresponding eigenvectors.</td>
</tr>
</tbody>
</table>

2.3.2 Dependence level

In the Matérn model in equation (2.5), the range parameter \( \beta \) controls the dependence of the process. In this section, we consider different \( \beta \). Given \( \nu = 0.5 \), which corresponds to an exponential covariance function and \( \tau^2 = 0.15 \), the top-left and top-right panels of Figure 2.2 show the Kullback–Leibler divergence for \( \beta = 0.1 \), which means a weaker dependence, and \( \beta = 0.5 \), which indicates a stronger dependence, as the rank \( r \) increases from 2 to 8. We can see that the HLR approximation is always the best with the smallest Kullback–Leibler divergence, and SUM and NNSUM win against NN for \( r \) up to 3 when \( \beta = 0.1 \), while when \( \beta = 0.5 \), the improvement of SUM and NNSUM exists up to \( r = 8 \). It implies that when a strong correlation is present, a small number of nearest neighbors is not adequate to provide a good approximation of the conditional density. It is also worth noting that the range of \( r/n \) in this study is from 0.22% to 0.89%. For very large \( n \), and \( r \ll n \), the improvement from HLR, SUM or NNSUM approaches can be substantial.
2.3.3 Smoothness level

In the Matérn covariance function, a larger $\nu$ indicates a smoother process. In this section, we fix $\beta = 0.1$ and $\tau^2 = 0$. We consider two smoothness levels with $\nu = 0.5$ and $\nu = 1$, which correspond to the exponential and Whittle covariance functions, respectively. The top-left and bottom-left panels of Figure 2.2 show the Kullback–Leibler divergence. Similarly, the HLR approach outperforms the other methods. For the smoother process when $\nu = 1$, the improvement of NNSUM and SUM over NN for small ranks decreases, and all the methods need a large $r$ to achieve similar performances as $\nu = 0.5$.

2.3.4 Noise level

The nugget effect can be viewed as measurement errors or the micro-structure in the underlying process. In this section, we consider different $\tau^2$. Given $\beta = 0.1$ and $\nu = 0.5$, the bottom-right and top-left panels of Figure 2.2 show the Kullback–Leibler divergence for $\tau^2 = 0$ and $\tau^2 = 0.15$. In both cases, the HLR approach still provides the best approximation. We can see that SUM, NNSUM and HLR give better approximations when the process is noisy or with a larger $\tau^2$. And if the rank $r$ is limited to a small number, SUM or NNSUM can improve NN for noisy processes.

2.3.5 Sample size and rank

In this section, we explore the effect of sample size given the rank $r$ or the ratio of $r/n$. Figure 2.3 shows the results for a similar design as in the first row of Figure 2.2 but with $n = 2500$. Comparing Figure 2.3 to the first row of Figure 2.2, we can see that for a given process, a larger number of locations does require larger ranks to achieve a similar approximation quality. When $r$ is fixed, NN is often not adequate, especially for large $n$, and SUM, NNSUM, and HLR can improve the approximation by including more neighbors.
Figure 2.2: Four panels showing the Kullback–Leibler divergence against rank with 900 locations in IND (black), NN (green), SUM (orange), NNSUM (blue), and HLR (red) methods. The corresponding parameters are indicated in the titles.

Although it is not realistic for a large dataset, we also investigate a situation where NN is adequate to provide a good approximation at rank $r$, and then compare the Kullback–Leibler divergence for NN at $r+1$ and NNSUM with the same first $r$ nearest neighbors and one additional set containing the next 2 nearest neighbors. We find that for $\alpha = 1, \beta = 0.5, \nu = 0.5, \tau^2 = 0$ and $n = 900$, NN with rank $r + 1 = 51$ gives a Kullback–Leibler divergence as $6.7 \times 10^{-3}$ and NNSUM reduces Kullback–Leibler divergence by 1%.
2.4 Simulation study

In §2.3, we calculated the Kullback–Leibler divergence at the true parameter values. In this section, we aim to test our estimation procedures with replications. To save computational time, only a relatively small number of locations is considered. We generate $n = 900$ observations with parameters $\alpha = 1, \beta = 0.25, \nu = 0.5$ and $\tau^2 = 0.15$. We run the optimization for $\alpha, \beta$ while fixing $\tau^2$ and $\nu$ at the true value and obtain the estimates of $\alpha, \beta$ by maximizing the approximated likelihoods with $r = 2$. We repeat the estimates procedure 100 times and the boxplots of $\alpha$ and $\beta$ are shown in Fig. 2.4. Compared to the exact maximum likelihood estimates, the boxplots show that all the approximation methods perform reasonably well. Nevertheless, we can see that the estimates obtained by the hierarchical low-rank approximation method have the smallest trimmed root mean squared errors (TRMSE) among all the approximation methods. The TRMSE for a given parameter $\theta$ is defined as follows.

![Figure 2.3: Two panels showing the Kullback–Leibler divergence against rank with 2500 locations in IND (black), NN (green), SUM (orange), NNSUM (blue), and HLR (red) methods. The corresponding parameters are indicated in the titles.](image)
by only keeping the central 50 estimates out of the 100 simulations:

$$\text{TRMSE}(\theta) = \left( \frac{\sum_{i=26}^{75} (\theta - \hat{\theta}_{[i]})^2}{100} \right)^{0.5},$$

where $\theta$ is the true parameter value, and $\hat{\theta}_{[i]}$ denotes the $i$-th largest estimate in the 100 simulations.

Figure 2.4: Two panels showing the boxplot of parameter estimates and mean squared error. The solid line is a reference for the true parameter value, and the dash line with squares (- - □ - -) is the corresponding root mean squared error of the 100 times of estimates in each method. Left: illustration for estimated $\alpha$; Right: illustration for estimated $\beta$.

2.5 Application

In this section, we apply our method to modeling soil moisture, a key factor in evaluating the state of the hydrological process, including runoff generation and drought development. Soil moisture information is also valuable to a wide range of applications, such as early warning of flood and drought, irrigation management, crop yield prediction, and weather pattern forecasting. Because soil moisture controls the energy exchange between the land and the atmosphere through evaporation and plant transpiration, it has been shown that better characterization of soil moisture in weather
prediction models can lead to significant improvements on temperature and precipitation forecasting. As a result, the development of better statistical models for soil moisture plays an important role in understanding its spatial variability. Furthermore, many other environmental data, such as temperature, pressure, and humidity, often share similar data structure as the soil moisture that we have analyzed, and their spatial variability is also of great interest. Therefore, our analyses can be used as an illustration to model and make inferences on such high-resolution irregularly spaced spatial datasets.

2.5.1 Dataset description

We consider high-resolution daily soil moisture data at the top layer of the Mississippi basin, U.S.A., on January 1, 2004 [24]. The spatial resolution is of 0.0083 degrees. The grid consists of $1830 \times 1329 = 2,432,070$ locations with 2,153,888 observations and 278,182 missing values. The illustration of the data is shown in Figure 2.5.

![Soil moisture map](image_url)

Figure 2.5: Soil moisture (unit: percentage) at the top layer of the Mississippi basin, U.S.A. on January 1, 2004.
We know that a one-degree difference in latitude along any longitude line is equivalent to 111 km; however, the distance of one-degree difference in longitude depends on the corresponding latitude. As the range of the latitude in this region is relatively small, for simplicity, we use the distance of one-degree difference in longitude at the center location of the region to represent all others, which is 87.5 km; namely, in this region, 1° in latitude is 111 km and 1° in longitude is 87.5 km.

To understand the structure of the day’s soil moisture, we fit a Gaussian process model with a Matérn covariance function. From all the locations, we randomly pick \( n = 2,000,000 \) points, which are irregularly spaced, to train our model. To assess the quality of our model, the fitted models can be used to predict the left-out observations.

### 2.5.2 Estimation and prediction

To use a Gaussian process model, we first fit a linear model to the longitude and latitude as the covariates to the soil moisture. After fitting, we find the negatively skewed residuals, hence we apply a logarithm transformation with some shift. The histogram of the transformed residual is shown in the left panel of Figure 2.6 which does not show strong departure from Gaussianity. To examine the isotropy of this process, we calculate the directional empirical variograms as illustrated in the right panel of Figure 2.6. We see the variograms on the circle with the same radius to the origin have similar values, suggesting that it is reasonable to assume an isotropic model.

Let \( Z(s) \) denote the transformed residual and the region \( D \) be the set of the selected locations, then the proposed Gaussian process model here is \( \{Z(s) : s \in D \subset \mathbb{R}^d \} \sim \text{GP}(0, C(h; \theta)) \). We choose three different covariance functions: the exponential, which has the smoothness parameter \( \nu = 0.5 \); the Whittle, which has \( \nu = 1 \); and the Matérn covariance function, which has an unknown \( \nu \). The formula is given in Equation (2.5). Given that the 2,000,000 observations follow \( Z \sim N(0, \Sigma(\theta)) \), \( \Sigma(\theta) \)
is the 2 million by 2 million variance-covariance matrix, obtained from the chosen
 covariance function. We use NN and HLR with rank \( r = 60 \) to get the approximated
likelihood and then obtain the parameter estimates. The results are shown in Ta-
ble 2.2. The Matérn covariance model is more flexible by allowing to estimate \( \nu \). The
estimated \( \nu \) in the Matérn covariance model by both methods is smaller than 0.5,
and the estimated \( \beta \) has the largest value. It suggests a rougher process with a larger
dependence range compared to the estimated exponential covariance model. The last
row of Table 2.2 shows the values of log-likelihood per observation. For each given
covariance model, the likelihood with parameters estimated by HLR is always larger
than that by NN. Among different covariance models, the likelihood with Matérn
covariance is the largest.

The size of the problem in this application is in the millions, a dataset which is far
beyond the ability of classic analysis methods. However, NN and HLR can evaluate
the approximated likelihood at each iteration in the optimization procedure within
5 and 14 minutes, respectively. The fast computation makes it highly practical for
applying the proposed methods to a large real-world spatial dataset problem. The
experiment is performed with the Intel Xeon E5-2680 v3@2.50GHz processor.
Table 2.2: Parameter estimation results.

<table>
<thead>
<tr>
<th></th>
<th>Nearest neighbors</th>
<th>Hierarchical low-rank approximation</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Exponential</td>
<td>Whittle</td>
</tr>
<tr>
<td>$\hat{\alpha}$</td>
<td>1.0073</td>
<td>0.9787</td>
</tr>
<tr>
<td>$\hat{\beta}$ (km)</td>
<td>21.6115</td>
<td>5.9316</td>
</tr>
<tr>
<td>$\hat{\tau}^2$</td>
<td>0.0107</td>
<td>0.0013</td>
</tr>
<tr>
<td>$\hat{\nu}$</td>
<td>0.5000</td>
<td>1.0000</td>
</tr>
<tr>
<td>$\ell/n$</td>
<td>-0.1042</td>
<td>-0.1417</td>
</tr>
</tbody>
</table>

$\hat{\alpha}$, $\hat{\beta}$, $\hat{\tau}^2$, $\hat{\nu}$ are the estimated parameter values, $n$ is the sample size, and $\ell$ is the log-likelihood.

Next, we use the fitted Matérn model by HLR to predict soil moisture at 1000 out of the left-out locations by kriging, which is known to provide the Best Linear Unbiased Predictor (BLUP) as well as the prediction standard errors [12]. However, the problem here is of size $n = 2,000,000$, hence kriging cannot be employed directly, because it involves a linear solve of size $n$ [4]. In fact, the proposed methods in this chapter can be adopted for approximating kriging equations as well. But for the purpose of validating the fitted model, we explore the exact computation method by treating the irregularly spaced data as observations on a finer regular grid with missing values. The resulting covariance matrix has a block Toeplitz Toeplitz block structure, which can be embedded in a block circulant circulant block matrix [25]. Then kriging can be done by fast Fourier transformation. More details can be found in Chan and Ng [26]. The mean squared prediction errors over the 1000 validation locations is $4.23 \times 10^{-5}$, which is notably small.

2.6 Discussion

The implementation in this chapter was done with a single-thread program, however as aforementioned in §2.2.5, computation in each hierarchy can be parallelized, which would reduce the computation time dramatically and make applications even more practical. The proposed method can be also extended to more complicated settings.
For example, although the rank was fixed to the same in each hierarchy, it can be chosen flexibly in accordance with the number of “past” observations that are involved in the hierarchy, which, we believe, would give a better approximation.
Chapter 3

Total Variation Depth and Its Decomposition for Outlier Detection in Functional Data

3.1 Introduction

Functional data, realizations of a one-dimensional stochastic process in the form of functions, are observed and collected with increasing frequency across research fields, including meteorology, neuroscience, environmental science, and engineering. Functional Data Analysis (FDA) considers the continuity of functions and various parametric and nonparametric methods can be found in Ferraty and Vieu [27] and Ramsay et al. [28]. In recent years, extensive developments have extended typical techniques of FDA to the analysis of more complicated functional objectives. Besides model-based methods, Exploratory Data Analysis (EDA) has been extended to functional data as well.

Sun and Genton [29] proposed the functional boxplot as an informative visualization tool for functional data, and Genton et al. [30] extended it to image data. Similar to the classical boxplot, if we simply extend the univariate ranking to the functional setting, the features of functional data cannot be captured. Data depth is a widely used concept in multivariate and functional data ranking. The general requirement for a data depth notion is the so-called “center-outwards” ordering, which means the natural center of the functional data should have the largest depth value and the depth decreases as the data approach outwards. Liu et al. [31] reviewed many popular depth notions for multivariate data. Some examples in the nonparametric
framework include half-space depth [32], simplicial depth [33], projection depth [34], and spatial depth [35]. For univariate functional data depth, (modified) band depth [36], integrated data depth [37], half-region depth [38], and extremal depth [39] have been proposed depending on the desired emphases. Nowadays, the study of multivariate functional data also arouses extensive interest due to its practical applications.

Berrendero et al. [40] studied the daily temperature functions on different surfaces, and Sangalli et al. [41] and Pigoli and Sangalli [42] analyzed multivariate functional medical data. To provide a valid depth of multivariate functional data, Ieva and Paganoni [43] derived depth measures for multivariate functional data from averaging univariate functional data depth, and Claeskens et al. [44] gave a generalization to multivariate functional data depths by averaging a multivariate depth function over the time points with a weight function. In particular, López-Pintado et al. [45] proposed and studied the simplicial band depth for multivariate functional data, which is an extension of the univariate functional band depth.

Data-depth-based methods provide many attractive tools in solving problems related to classification, clustering, and outlier detection. For example, Jörnsten [46], Ghosh and Chaudhuri [47] and Dutta and Ghosh [48] used various depth notions to classify or cluster multivariate data. López-Pintado and Romo [49] and Cuevas et al. [50] extended the classification problem to functional data. Outlier detection is another challenging problem, especially for functional data, because there is no clear definition of functional outliers. Roughly speaking, functional outliers can be categorized into two types: magnitude and shape outliers. Magnitude outliers have very deviated values in some dimensions, while shape outliers have a different shape compared to the vast majority; however, no data value may deviate too much from the center. Dang and Serfling [51] introduced nonparametric multivariate outlier identifiers based on various multivariate depths. For functional data, Febrero et al. [52] used a cutoff, which is determined by a bootstrap, for the functional data depth
to detect outliers. Hyndman and Shang [53] proposed using the first two robust principal components to construct a bagplot [54], or a highest density region plot [55] to detect outliers. However, the first two principal components often do not adequately describe the variabilities in functional data. In the functional boxplot proposed by Sun and Genton [29], outliers were detected by the 1.5 times the 50% central region rule, which means that any departure of an observation from the fence, that is, the inflation of the 50% central region by 1.5 times, makes it an outlier. The 50% central region is the envelope of the first 50% of curves that have the largest depth values, and the factor 1.5 can be adjusted according to the distribution [56]. Good performance has been shown for various types of outliers, especially for magnitude outliers. However, for shape outliers, two issues arise: outliers may appear among the first 50% of curves with the largest depth values and outliers that are masked in the fence can never be detected, even though their depth values are small. Recently, Arribas-Gil and Romo [57] proposed the outliergram, where the relationship between modified band depth [36] and modified epigraph index [38] is studied to identify shape outliers. Hubert et al. [58] proposed functional bagdistance and skewness-adjusted projection depth to detect various outliers for multivariate functional data.

In this chapter, we propose a notion of data depth, the total variation depth, and develop an effective procedure to detect both magnitude and shape outliers using attractive features of the total variation depth. We were motivated by the drawback of the modified band depth. Typically, functional data depth is constructed via averaging the pointwise depth. The modified band depth can be viewed as such an average for temporal curves, however, it does not take the time order or temporal correlations into account. In contrast, our proposed total variation depth allows for a meaningful decomposition that considers the correlation of adjacent dimensions and can be used for outlier detection. Although the total variation depth is essentially of the form of an integral of a univariate simplicial depth function, this decomposition
is achieved by a novel formation of the simplicial depth pairwisely. In our proposed outlier detection procedure, we show that combined with the functional boxplot, we are able to detect both magnitude and shape outliers. We also develop informative visualization tools for easy interpretation of the outlyingness. The outlier detection performance has been examined through simulation studies. The visualization tool is described and illustrated through three kinds of real-data applications: curves, images and video frames.

3.2 Methodology

3.2.1 Total variation depth

Let $X$ be a real-valued stochastic process on $\mathcal{T}$ with distribution $F_X$, where $\mathcal{T}$ is an interval in $\mathbb{R}$. We propose the total variation depth for a given function $f$ w.r.t. $F_X$. We use $f$ to denote a function, and $f(t)$ to denote the functional value at a given $t$. First, we define the pointwise total variation depth for a given $t$. Let $R_f(t) = 1\{X(t) \leq f(t)\}$, where $1$ is the indicator function. It is easy to see that $p_f(t) = \mathbb{E}\{R_f(t)\} = \mathbb{P}\{X(t) \leq f(t)\}$ is associated with the relative position of $f(t)$ w.r.t. $X(t)$. For instance, if $f^*(t)$ is the true median, then $p_{f^*}(t) = 1/2$. For a given function $f(t)$ at each fixed $t$, we introduce the pointwise total variation depth of $f(t)$ by $D_f(t) = \text{Var}\{R_f(t)\} = p_f(t)\{1-p_f(t)\}$, which is equivalent to the univariate simplicial depth, however, with a different characterization. For a fixed $t$, we know $D_f(t)$ is maximized at the center when $p_f(t) = 1/2$, or simply the univariate median. Next, we define the functional Total Variation Depth (TVD) for the given function $f(t)$ on $\mathcal{T}$ by

$$\text{TVD}(f) = \int_{\mathcal{T}} w(t)D_f(t)dt,$$

where $w(t)$ is a weight function defined on $\mathcal{T}$. Although the proposed depth is of the form of the integrated depth [37], with the choice of simplicial depth as the univariate
depth function, we will see later in §3.2.2 that the novel formulation of \( D_f(t) \) as the variance of \( R_f(t) \) leads to useful decomposition associated with shape and magnitude properties of the functions. There are many ways to choose the weight function, and we now provide two examples for the choice of \( w(t) \). If we let \( w(t) \) be a constant as \( w(t) \equiv 1/|\mathcal{T}| \), then it can be shown that the TVD is also equivalent to the modified band depth [36]. Another example is to let \( w(t) \) be proportional to the vertical variability at different time points proposed by Claeskens et al. [44]. The motivation is to put more emphases on the order of curves in the regions with large amplitude variability, which may lead to a better separation in the resulting sample depth values. We have considered both and notice that these two choices of weight functions perform similarly in the outlier detection procedure we proposed. Therefore, we choose the constant weight to present our results in this chapter.

### 3.2.2 Properties of the total variation depth

Zuo and Serfling [34] studied the key properties of a valid multivariate data depth, and Claeskens et al. [44] extended them to a functional setting, including affine invariance, maximality at the center, monotonicity relative to the deepest point, and vanishing at infinity. Denote the total variation depth of \( f \) w.r.t. \( F_X \) by \( \text{TVD}(f, F_X) \). Having a form of the integrated depth [37], the TVD enjoys the following properties immediately:

- **Affine invariance.** \( \text{TVD}(f, F_X) = \text{TVD}(af + g, F_{aX+g}) \) for any \( a \in \mathbb{R}\backslash\{0\} \) and any function \( f \) and \( g \) on \( \mathcal{T} \).

- **Maximality at the center.** If at each time point \( t \in \mathcal{T} \), the distribution \( F_X(t) \) has a uniquely defined center \( f^*(t) \), then \( \text{TVD}(f^*, F_X) = \sup_{f \in C(\mathcal{T})} \text{TVD}(f, F_X) \), where \( C(\mathcal{T}) \) denotes the set of all continuous functions on \( \mathcal{T} \).
• **Monotonicity relative to the deepest point.** If $f^*$ is the function such that $\text{TVD}(f^*, F_X) = \sup_{f \in C(T)} \text{TVD}(f, F_X)$, then $\text{TVD}(f, F_X) \leq \text{TVD}(f^* + \alpha(f - f^*), F_X)$, for any $f$ on $T$ and $\alpha \in [0, 1]$.

• **Vanishing at infinity.** $\text{TVD}(f, F_X) \to 0$ as $\|f(t)\| \to \infty$ for almost all time points $t$ in $T$.

The consistency of the sample integrated depth was first investigated by Fraiman and Muniz [37]. Then, Nagy et al. [59] studied further, provided the consistency properties of the integrated depth with predetermined weight function, and pointed out the challenges to prove the consistency when the weight function depends on the distribution. Following the arguments in Nagy et al. [59], one can show that the sample TVD with the constant weight is a consistent estimator of the TVD.

Next, we show in Theorem 2 that the proposed pointwise total variation depth can be decomposed into the pointwise magnitude similarity and pointwise shape similarity, which has important practical implications for outlier detection problems. The result in Theorem 2 follows from the law of the total variance.

**Theorem 2 (Decomposition of TVD).** Let $s, t$ be two time points where $s = t - \Delta$ provided $s, t \in T$. The pointwise total variation depth has the following decomposition:

$$D_f(t) = \text{Var}\{R_f(t)\} = \text{Var}[\mathbb{E}\{R_f(t) \mid R_f(s)\}] + \mathbb{E}[\text{Var}\{R_f(t) \mid R_f(s)\}].$$

The decomposition implies that the total variance of $R_f(t)$ can be decomposed into $\text{Var}[\mathbb{E}\{R_f(t) \mid R_f(s)\}]$, the variability that is explained by $R_f(s)$, and $\mathbb{E}[\text{Var}\{R_f(t) \mid R_f(s)\}]$, the variability that is independent of $R_f(s)$. We call $\text{Var}[\mathbb{E}\{R_f(t) \mid R_f(s)\}]$ the shape component and $\mathbb{E}[\text{Var}\{R_f(t) \mid R_f(s)\}]$ the magnitude component. If the shape component of a sample curve dominates the total variation depth, it indicates that the curve has a similar shape compared to the majority. Notice that the shape of the function $f$ is characterized by considering a span of time $\Delta$, where $1/\Delta$ is
the sampling frequency. In practice, \((t - \Delta, t)\) denote two consecutive discrete time points. Now we give the definition of the shape similarity for a given function using the ratio of the shape component to the total variation depth.

**Definition 1.** For any function \(f\) on \(T\), the Shape Similarity (SS) of \(f\) w.r.t. the distribution \(F_X\) is defined as

\[
SS(f) = \int_T v(t; \Delta) S_f(t; \Delta) dt,
\]

where \(v(t; \Delta)\) is a weight function, and \(S_f(t; \Delta)\) is given by

\[
S_f(t; \Delta) = \begin{cases} 
\frac{\text{Var}(\mathbb{E}[R_f(t) \mid R_f(t-\Delta)])}{D_f(t)} & , \quad D_f(t) \neq 0 \\
1 & , \quad D_f(t) = 0 
\end{cases}
\]

Now we discuss the choice of the weight function \(v(t; \Delta)\). Notice that at a given \(t\), \(R_f(t) = 1\{X(t) \leq f(t)\}\) is an indicator function. If \(X(t) \leq f(t)\), \(R_f(t) = 1\) no matter how large the value of \(f(t)\) is. To account for the value of \(f(t)\), we choose the weight function \(v(t; \Delta)\) to be the normalized changes in \(f(t)\) on \(T\). More precisely, \(v(t; \Delta)\) is given by \(v(t; \Delta) = | f(t) - f(t-\Delta) | / \int_T | f(t) - f(t-\Delta) | \). Then, more weights are assigned to the time intervals where \(f(t)\) has larger changes in magnitude. This choice of \(v(t; \Delta)\) is useful in practice when sample curves are only different in magnitude within short time intervals.

In Definition 1, smaller values of the shape similarity are associated with larger shape outlyingness. However, we notice that for outlying pairs \((f(t - \Delta), f(t))\) with a small value of the shape component \(\text{Var}(\mathbb{E}[R_f(t) \mid R_f(t-\Delta)])\), \(S_f(t; \Delta)\) may not be small enough, if \(D_f(t)\) in the denominator is too small. To better reflect the shape outlyingness via the shape similarity, we shift \((f(t - \Delta), f(t))\) to the center, such that \((\tilde{f}(t - \Delta), \tilde{f}(t)) = (f(t - \Delta), f(t)) - \delta_t\), where \(\delta_t = f(t) - \text{median}\{X(t)\}\), and then define the modified shape similarity using the shifted pairs as follows:
Definition 2. For any function $f$ on $\mathcal{T}$, the Modified Shape Similarity (MSS) of $f$ w.r.t. the distribution $F_X$ is defined as

$$MSS(f) = \int_\mathcal{T} v(t; \Delta)S_f(t; \Delta)dt.$$ 

Similar to the shape similarity, the modified shape similarity characterizes shape outlyingness of $f$ via each pair of function values, but it is a stronger indicator by shifting each pair to the center when calculating $S_f(t; \Delta)$. Arribas-Gil and Romo [57] used a similar idea for shape outlier detection, whereas the entire extreme curve at all time points is shifted to the center by the same amplitude.

In real applications, we observe functional data at discretized time points and only have samples from the true distribution. We replace the cumulative density function $F_X$ by its empirical version and obtain the sample total variation depth and the sample modified shape similarity. All the estimations use sample proportions to approximate probabilities, and the details are provided in the Appendix B.

3.2.3 Outlier detection rule and visualization

To obtain robust inferences, outlier detection is often necessary; however, the procedure is challenging for functional data, because the characterization of functional data in infinite dimensions and appropriate outlier detection rules are needed. There are many proposed ways to detect functional outliers, some of which have been discussed in §3.1, and here, we propose a new outlier detection rule by making good use of the decomposition property of the proposed total variation depth. Suppose we observe $n$ sample curves, then the outlier detection procedure is summarized as follows:

1. Estimate TVD and MSS for each curve as described in §3.2.1 and §3.2.2.

2. Draw a classical boxplot for the $n$ values of MSS and detect outliers by the $F$ times interquartile range empirical rule, where the factor $F$ can be adjusted by
users, depending on the distributions. For example, Hubert and Vandervieren [60] discussed the adjustment for skewed distributions and provide a more robust rule for outliers if the distribution is not symmetric. In the simulation study and applications, we choose $F = 3$ following the conservative outlier detection rule in a classical boxplot. Curves with MSS values below the lower fence in the boxplot are identified as shape outliers.

3. Remove detected shape outliers and draw a functional boxplot using the total variation depth to detect all the magnitude outliers by the 1.5 times of the 50% (w.r.t. the number of original observation before removing shape outliers) central region rule in the functional boxplot [29]. The factor 1.5 can be adjusted using bootstrap methods [56].

It is noteworthy that using the functional boxplot along with the boxplot of MSS enables us to detect both magnitude outliers and shape outliers with small oscillations. We call the boxplot of MSS the shape outlyingness plot because the proposed MSS is a good indicator of shape outlyingness. It is useful especially in detecting shape outliers without a significant magnitude deviation.

Next, we propose a set of informative visualization tools following one outlier detection procedure. Figure 3.1 (a) shows three examples of simulated datasets containing different types of outliers. The model details will be introduced in §3.3.1. The grey curves are the non-outlying observations and the black curve is the functional median with the largest TVD. The highlighted orange curves are the shape outliers detected by the shape outlyingness plot, as shown in Figure 3.1 (b), where the points in orange correspond to shape outliers. The red curve is a magnitude outlier detected by the functional boxplot constructed in Step 3 of the outlier detection procedure after removing the shape outliers. Finally, the functional boxplot in Figure 3.1 (c) displays the median, the 50% central region, the maximal and minimal envelopes, and the magnitude outlier.
3.3 Simulation Study

3.3.1 Outlier models

In this section, we choose a similar simulation design to that by Sun and Genton [29] and Narisetty and Nair [39], but we also introduce new outlier models. We study seven models in total, where Model 1 is the base model with no outliers compared to Model 2 – 6, and Model 7 concerns another kind of contamination with a different base model. The contamination ratio for outliers in Model 2 – 7 is $\epsilon = 10\%$. For all the models, $\mathcal{T}$ is set to be $\mathcal{T} = [0, 1]$.

**Model 1:** $X_i(t) = 4t + e_i(t)$, for $i = 1, \ldots, n$, where $e_i(t)$ is a zero-mean Gaussian process with covariance function $c(s, t) = \exp\{-|s - t|\}$, for $s, t \in \mathcal{T}$.

**Model 2:** $X_i(t) = 4t + e_i(t) + 6c_i\sigma_i$, for $i = 1, \ldots, n$ where $c_i \sim \text{Bernoulli}(\rho)$ and $\sigma_i$ takes values of 1 and $-1$ with probability 1/2, respectively, and $e_i(t)$ has the same definition as before.

**Model 3:** $X_i(t) = 4t + e_i(t) + 6c_i\sigma_i$, if $t \geq T_i$, and $X_i(t) = 4t + e_i(t)$, if $t < T_i$, for $i = 1, \ldots, n$, where $T_i \sim \text{Unif}([0, 1])$, and $e_i(t)$, $c_i$ and $\sigma_i$ have the same definition as before.
as before.

**Model 4:** $X_i(t) = 4t + e_i(t) + 6c_i\sigma_i$, if $T_i \leq t \leq T_i + l$, and $X_i(t) = 4t + e_i(t)$ otherwise, for $i = 1, \ldots, n$, where $l = 0.08, T_i \sim \text{Unif}([0, 1 - l])$, and $e_i(t), c_i$ and $\sigma_i$ have the same definition as before.

**Model 5:** $X_i(t) = 4t + (1 - c_i)e_i(t) + c_i\tilde{e}_i(t)$, for $i = 1, \ldots, n$, where $c_i$ and $e_i(t)$ have the same definition as before, and $\tilde{e}_i(t)$ is another zero-mean Gaussian process with a different covariance function $\tilde{c}(s, t) = 8\exp\{-|s - t|^{0.1}\}$.

**Model 6:** $X_i(t) = 4t + e_i(t) + c_i(0.5\sin(40\pi t))$, for $i = 1, \ldots, n$, where $c_i$ and $e_i(t)$ have the same definition as before.

**Model 7:** $X_i(t) = 2\sin(15\pi x + 2c_i) + e_i(t)$, and the corresponding base model in Model 7 is $X_i(t) = 2\sin(15\pi x + 2c_i) + e_i(t)$, for $i = 1, \ldots, n$, where $c_i$ and $e_i(t)$ have the same definition as before.

Then, Model 2 contains symmetric magnitude outliers with a shift; Model 3 makes the magnitude outliers deviate starting from a random time point; Model 4 generates magnitude outliers that have peaks lasting for a short time period; Model 5 has outliers that have a different temporal covariance, so that the outliers have a different shape as well as a larger variance; Model 6 considers shape outliers by adding an oscillating function, where the oscillation is frequent in time but close to the majority in magnitude; and Model 7 introduces shape outliers with a phase shift. The illustration of generated outliers from Model 2 – 7 is shown in Figure 3.2. For the following simulation studies, we generate $n = 100$ curves taking values on an equally spaced grid of $[0, 1]$ with $p = 50$ time points for each model.

### 3.3.2 Central region

In this section, we study how the 50% central region is affected by different choices of depth notions in the given outlier models. A good 50% central region is supposed to be compact and not affected by outliers. We compare our depth notion with the Modified
Figure 3.2: Top panel: 100 curves (grey) from Model 1 with one of different outliers from Model 2 (red), Model 3 (orange), Model 4 (blue), Model 5 (green) and Model 6 (black), respectively. Bottom panel: 100 curves (grey) contaminated by one outlier from Model 7 (purple).

Band Depth (MBD) \cite{36} and the Extremal Depth (ED) \cite{39}. When using TVD, the 50% central region is constructed by 50% of the deepest curves after removing the detected shape outliers, as described in §3.2.3. The resulting central regions clearly differ by choosing different depth notions in Models 3 and 4. From Figure 3.3, we see that the central region constructed using MBD is contaminated by the outliers in Models 3 and 4, and even contains sudden peaks, while ED leads to compact central regions because of its emphasis on extremal properties. The reason MBD fails for Models 3 and 4 is because it only accounts for averaged magnitude outlyingness, and
for outliers that only deviate in a short time period as in Models 3 and 4, MBD may assign large depth values, so that these outliers fall into the first half of the deepest curves. By definition, TVD also considers the averaged magnitude outlyingness as MBD. However, by removing the detected shape outliers first, the central region remains compact. All other models show depth notions leading to compact central regions, and we only pick the results for Model 7 in the bottom of Figure 3.3. Overall, the performance of our proposed method is appealing in constructing the 50% central region.

### 3.3.3 Outlier detection

In this section, we use our outlier detection rule described in §3.2.3 to detect outliers. Sun and Genton [29] examined the outlier detection performance of the functional boxplot using the modified band depth and 1.5 times of the 50% central region rule. Later on, Narisetty and Nair [39] compared the outlier detection performance by replacing MBD with their proposed ED in the functional boxplot. Arribas-Gil and Romo [57] proposed the outliergram for shape outlier detection, and magnitude outliers were still detected by the functional boxplot using the modified band depth. All of these methods use the functional boxplot, but with different choices of depth, or they exploit other tools for detecting shape outliers. We now compare the outlier detection performance of the functional boxplot using MBD and ED, the functional boxplot together with the outliergram (OG+MBD), and our proposed method (TVD+MSS), where TVD is used in the functional boxplot for magnitude outliers and shape outlier are detected by MSS.

In the experiment, we assess the performance by the True Positive Rate (TPR), which is the ratio of the number of correctly detected outliers by the number of true outliers, and the False Positive Rate (FPR), which is the ratio of the number of wrongly detected outliers by the number of true non-outliers. A larger TPR means
Figure 3.3: The central regions constructed by MBD (left), ED (middle), and TVD (right) for Model 3 (top), Model 4 (middle) and Model 7 (bottom). The central region is displayed by the pink polygon. The solid black line in the middle is the median curve, which has the largest depth value. All the grey lines in the background are the dataset curves.
more outliers are correctly detected, and a smaller FPR means fewer non-outliers are falsely detected as outliers. The results are shown in Table 3.1 for 500 experiments in each case, with mean values and standard deviations (in brackets). We see that our proposed outlier detection method is very satisfactory for all models, especially for Models 3 - 7. Even for Model 4, which ED favors, our detection method successfully detects all of the outliers with a TPR of 100%, and it outperforms the functional boxplot with ED, for which the TPR is 86.62%. For Model 6, where shape outliers only have small oscillations, all the other methods fail with very low TPRs, but our method still gains high accuracy. For shape outliers that also show outlyingness in magnitude as in Models 3, 5 and 7, the outliergram correctly detects most of the outliers, while our method has similar or higher TPRs, but much lower FPRs. For magnitude outliers in Model 2, all the methods perform well. For Model 1, where no outliers exist, all the methods except the outliergram retain a low FPR. In fact, a relatively high FPR of the outliergram is observed for all the models, indicating that the outliergram tends to falsely detect too many outliers.

Table 3.1: Results of outlier detection using different methods for different models. The values are the means in the 500 experiments in each case, and the values in brackets are the corresponding standard deviations.

<table>
<thead>
<tr>
<th>Model</th>
<th>MBD</th>
<th>ED</th>
<th>OG+MBD</th>
<th>TVD+MSS</th>
</tr>
</thead>
<tbody>
<tr>
<td>FPR</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Model 1</td>
<td>0.07(0.26)</td>
<td>0.03(0.18)</td>
<td>5.27(2.35)</td>
<td>0.07(0.26)</td>
</tr>
<tr>
<td>Model 2</td>
<td>98.87(3.35)</td>
<td>98.5(4.16)</td>
<td>98.87(3.35)</td>
<td>98.88(3.39)</td>
</tr>
<tr>
<td>FPR</td>
<td>0.04(0.21)</td>
<td>0.02(0.18)</td>
<td><strong>4.9(2.5)</strong></td>
<td>0.04(0.22)</td>
</tr>
<tr>
<td>Model 3</td>
<td>82.19(14.87)</td>
<td>86.15(13.73)</td>
<td>89.09(11.69)</td>
<td>98.08(4.68)</td>
</tr>
<tr>
<td>TPR</td>
<td>0.05(0.25)</td>
<td>0.02(0.14)</td>
<td><strong>3.3(2.07)</strong></td>
<td>0.25(0.55)</td>
</tr>
<tr>
<td>FPR</td>
<td>47.86(20.01)</td>
<td><strong>86.62(12.27)</strong></td>
<td>75.85(18.14)</td>
<td><strong>100(0)</strong></td>
</tr>
<tr>
<td>Model 4</td>
<td>0.06(0.27)</td>
<td>0.02(0.16)</td>
<td>3.22(2.05)</td>
<td>0.04(0.22)</td>
</tr>
<tr>
<td>TPR</td>
<td>84.27(14.14)</td>
<td>82.1(14.02)</td>
<td>99.99(0.26)</td>
<td>100(0)</td>
</tr>
<tr>
<td>FPR</td>
<td>0.03(0.2)</td>
<td>0.01(0.12)</td>
<td><strong>2.13(1.73)</strong></td>
<td>0.03(0.19)</td>
</tr>
<tr>
<td>Model 5</td>
<td>0.06(0.74)</td>
<td>0.03(0.55)</td>
<td>6.74(8.61)</td>
<td><strong>99.84(1.99)</strong></td>
</tr>
<tr>
<td>TPR</td>
<td>0.04(0.22)</td>
<td>0.01(0.12)</td>
<td>4.82(2.23)</td>
<td>0.03(0.19)</td>
</tr>
<tr>
<td>FPR</td>
<td>44.18(24.34)</td>
<td>34.53(22.36)</td>
<td>99.89(2.53)</td>
<td>100(0)</td>
</tr>
<tr>
<td>Model 6</td>
<td>0.03(0.18)</td>
<td>0.01(0.11)</td>
<td><strong>2.16(1.77)</strong></td>
<td>0.03(0.19)</td>
</tr>
</tbody>
</table>
3.4 Applications

In this section, we apply our proposed outlier detection procedure and visualization tools to three applications: the sea surface temperature in Niño zones, the sea surface height in the Red Sea, and the surveillance video of a meeting room. The three applications cover examples of datasets consisting of curves, images, and video frames.

3.4.1 Sea surface temperature

The El Niño southern oscillation (ENSO), irregular cycles of warm and cold temperatures in the eastern tropical Pacific Ocean, has a global impact on climate and weather patterns, including temperature, rainfall, and wind pressure. There are several indices to measure ENSO, one of which is the sea surface temperature in the Niño regions. The warm phase of ENSO is referred as El Niño, and the cold phase is called La Niña. For example, the years 1982–1983, 1997–1998, and 2015–2016, were reported as strong El Niño years, because the sea surface temperature anomalies were significantly larger for several consecutive seasons.

![Figure 3.4: The sea surface temperatures in a Niño zone from July 1982 to June 2016.](image)

(a) The original data. (b) The boxplot of MSS. (c) The functional boxplot with TVD.

In this application, we use our method on the monthly sea surface temperature in
one of the Niño zones (0 – 10° South and 90 – 80° West) from July 1982 to June 2016. In Figure 3.4 (a), each curve represents a yearly sea surface temperature from July to next June. There are 35 curves in total. We see in Figure 3.4 (b) that there are no shape outliers, and two magnitude outliers have been detected, which are the years 1982–1983 and 1997–1998. The outliers coincide with the strong El Niño events. The upper envelop shown as the upper blue line in Figure 3.4 (c), partly comprises July and August of 1983 (shown by the green line in Figure 3.4 (a)), the extension of the El Niño year 1982–1983, and September to December of 2015 (shown by the purple line in Figure 3.4 (a)), agreeing with the El Niño year 2015–2016. The median is the year 1989–1990 shown as the black line in Figure 3.4 (a), when no obvious El Niño event occurred.

3.4.2 Sea surface height

Sea surface height plays an important role in understanding ocean currents. Geophysical scientists run simulations with different initial conditions to generate ensembles. To understand the variability in these model runs, we apply our method to 50 simulated sea surface temperatures of the Red Sea on January 1, 2016, where each observation is an image with 25,808 valid values. We first order the observations along the axis of the Red Sea in a zigzag fashion and obtain the one-dimensional functional data. As an example, one shape outlier is shown in Figure 3.5 (b). After removing the two shape outliers, the functional boxplot is applied using the total variation depth. One of the seven magnitude outliers is shown in Figure 3.5 (c), and the median is shown in Figure 3.5 (a). We see that compared to the median, the shape outlier has a different pattern in the northern part of the Red Sea, and the magnitude outlier has extreme heights in some places in the middle part of the Red Sea.
Figure 3.5: Runs of the sea surface height model. (a) The median. (b) One of the shape outliers. (c) One of the magnitude outliers.

3.4.3 Surveillance video

Video datasets consist of a sequence of video frames, each of which can be treated as one functional observation. Hubert et al. [61] illustrated their outlier detection method using a beach surveillance video. We explore a video of a meeting room, filmed by Li et al. [62]. In this video, there was nobody in the meeting room at the beginning, and only a curtain was moving with the wind. Later, a person came in, stayed for a while, and left, three times. The goal is to detect those video frames, where either the curtain is moving too far away, or the person is in the meeting room.

This video was filmed for 94 seconds, and consists of 2,964 frames. We first equally sub-sampled 1,280 out of 20,480 pixels in each video frame to capture the main features of the image. Then, the selected pixels are ordered by rows in a zigzag fashion to obtain the one-dimensional data vector. After applying our outlier
detection method, we get 306 shape outliers and 110 magnitude outliers. The median in Figure 3.6 (a) shows the most representative frame during the 94 seconds. It indicates the typical still position of the curtain in the video with nobody in the meeting room. Figure 3.6 (b) is one example of the shape outliers, where we see the person in the meeting room, making this frame show a quite different pattern among the others. Figure 3.6 (c) is one example of the detected magnitude outliers, where we see the curtain moved far away from the median. The detected outlying frames cover the periods in the video where the person was inside the meeting room.

Figure 3.6: The video frames. (a) The median. (b) One of the shape outliers. (c) One of the magnitude outliers.

3.5 Discussion

In this chapter, we proposed a depth notion, TVD, for functional data. We illustrated that this model has many attractive properties and in particular, we highlight its decomposition as especially useful for detecting shape outliers of the most challenging type. We proposed MSS via the decomposition as an indicator of the shape outlyingness and constructed the shape outlyingness plot to detect shape outliers. Together with the functional boxplot, our proposed outlier detection procedure can detect both magnitude and shape outliers. Good performances of outlier detection were demonstrated through simulation studies with various types of outlier models.
Moreover, we also developed a set of visualization tools for functional data, where we display the original observations with informative summary statistics, such as the median and the central region, and highlight the detected magnitude and shape outliers, along with the corresponding shape outlyingness plot and the functional boxplot.

Another advantage of the proposed total variation depth is that it can be extended to multivariate functional data by properly defining the indicator function \( R_f(t) = 1\{X(t) \leq f(t)\} \). For example, for a multivariate function \( f(t) \) that takes values in \( \mathbb{R}^p \), let \( R_f(t) = 1\{ f(t) \in S(t) \} \), where \( S(t) \) is a properly defined set for the multivariate stochastic process \( X \) at time \( t \). One naive choice would be \( S(t) = \{ X(t) : \|X(t)\| \leq \|f(t)\| \} \). However, further research is needed for the properties and outlier detection performances of the multivariate total variation depth.
Chapter 4

Visualization and Assessment of Spatio-temporal Covariance Properties

4.1 Introduction

Spatio-temporal covariance functions play an important role in parametrically modeling geostatistical data, particularly for Gaussian random fields. Spatio-temporal models often have complex structures and rely on various assumptions to simplify the model and reduce the computational burden \[63\], such as full symmetry and separability \[64, 65, 66, 67\]. Specifically, let \( Z(s, t) \) be a second-order stationary random process at spatial location \( s \in \mathbb{R}^d \), and temporal point \( t \in \mathbb{R} \). The covariance function \( C(h, u) = \text{Cov}\{Z(s, t), Z(s+h, t+u)\} \) is positive definite and only depends on the spatial lag \( h \) and temporal lag \( u \). Then, the covariance function is called fully symmetric if \( C(h, u) = C(-h, u) = C(h, -u) \), and separable if \( C(h, u) = C(h, 0)C(0, u)/C(0, 0) \), for any spatial lag \( h \) and temporal lag \( u \). One can show that separability implies full symmetry. The separable covariance model is a product of purely spatial and temporal covariances, ignoring the interaction between space and time; the symmetric covariance model acknowledges the interaction but assumes that the cross-covariances of temporal lags with opposite signs are identical at two particular spatial locations. Although these models with simplified structures are easier to fit, they have limitations and may not be realistic in practical applications. In separable models, even small spatial changes can lead to large changes in the correlations, causing a lack of smoothness \[64\], and full symmetry is often violated when data are influenced by
dynamic processes with a prevailing flow direction [65].

Therefore, it has been an active research area to develop more flexible spatio-temporal covariance models motivated by real applications. Jones and Zhang [68] developed families of spectral densities that lead to non-separable covariance models without a closed form. Cressie and Huang [66] proposed families of non-separable covariance functions based on the Fourier transform of non-negative, finite measures with explicit expressions, but these are limited to classes with known analytical solutions of Fourier integrals. Later, Gneiting [65] extended this approach to more general classes with Fourier-free implementations. All these models are non-separable but fully symmetric. Stein [64] generated covariance functions that are isotropic but not fully symmetric by taking derivatives of fully symmetric models. Gneiting et al. [63] proposed anisotropic, asymmetric models by adding a compactly supported Lagrangian correlation function to a fully symmetric covariance model, where the coefficient of the Lagrangian component controls the extent of asymmetry.

In real data analysis, before choosing a covariance function from the existing models, it is necessary to assess the properties of separability and symmetry either by Exploratory Data Analysis (EDA) or through formal hypothesis tests. However, it is challenging to visualize and assess such properties from spatio-temporal observations. Brown et al. [69] determined the separability from the closeness to zero of the maximum likelihood estimates of blurring parameters associated with separability, without considering the uncertainty. Shitan and Brockwell [70] developed an asymptotic test for separability, but it is restricted to spatial autoregressive processes. Fuentes [71] presented a test of separability in the spectral domain using a simple two-factor analysis of variance. Regarding likelihood-based tests, Mitchell et al. [72] proposed a likelihood ratio test of separability for replicated multivariate data that examines whether the covariance matrix is a Kronecker product of two matrices of smaller dimensions. Mitchell et al. [73] extended this approach to test the
separability of spatio-temporal covariances by partitioning the observations into approximate replicates. After Scaccia and Martin [74] presented tests of symmetry and separability for lattice processes using the periodogram, Li et al. [75] developed a unified testing framework for both separability and symmetry by constructing contrasts of covariances at selected spatio-temporal lags.

Existing methods mostly build separability and symmetry tests based on given spatial and temporal lag sets, and work on visualization is sparse. In this chapter, we propose formulating separability and symmetry test functions as functional data, and provide a visualization tool to illustrate these properties. These test functions are functions of temporal lag and constructed from the cross-covariances of each pair of the time series observed at spatial locations. With the obtained functional data of test functions, we develop rank-based testing procedures that are model-free with data depth-based functional data ranking techniques. Because the test statistics are rank-based, the results are more robust to outliers, which may come from erroneous measurements of the variables under study or poor estimates of the covariances due to limited sample size.

The rest of the chapter is organized as follows. In §4.2 we introduce separability and symmetry test functions, which are used to visualize the covariance property of separability and symmetry, and provide the detailed rank-based testing procedure. In §4.3 we illustrate the visualization tools for simulated data and demonstrate the rank-based test for various spatio-temporal covariance models. In §4.4 we apply our methods to two real datasets, wind speed from monitoring stations, and surface temperatures and wind speed from numerical model outputs, to study their covariance structures.
4.2 Methodology

4.2.1 Non-separable covariance functions

Gneiting [65] introduced the following rich family of non-separable spatio-temporal covariance functions, which includes separable ones as a special case,

\[
C(h, u) = \frac{\sigma^2}{(a|u|^{2\alpha} + 1)^\beta} \exp \left\{ - \frac{c\|h\|^{2\gamma}}{(a|u|^{2\alpha} + 1)^{\beta\gamma}} \right\},
\]

where \(\sigma^2\) is the variance, \(a, c \geq 0\) determine the temporal and spatial range, \(\alpha, \gamma \in [0, 1]\) control the temporal and spatial smoothness, and \(\beta \in [0, 1]\) is the spatio-temporal interaction. The covariance function is separable when \(\beta = 0\), and a larger \(\beta\) is associated with a more non-separable model.

Cressie and Huang [66] also proposed a non-separable model,

\[
C(h, u) = \frac{\sigma^2(a|u| + 1)}{(a|u| + 1)^2 + b^2\|h\|^2}^{3/2},
\]

and a possible corresponding separable model was discussed in Mitchell et al. [73] as,

\[
C(h, u) = \frac{\sigma^2}{(a|u| + 1)^2(b^2\|h\|^2 + 1)^{3/2}}.
\]

Later, Rodrigues and Diggle [76] further defined the properties of positive and negative non-separability and showed that the two aforementioned models are both positively non-separable. To investigate models of negative non-separability, we consider the one from De Cesare et al. [77],

\[
C(h, u) = k_1 C_s(h)C_t(u) + k_2 C_s(h) + k_3 C_t(u),
\]

where \(k_1 > 0, k_2, k_3 \geq 0\), and \(C_s(\cdot)\) and \(C_t(\cdot)\) are valid spatial and temporal covariance functions, respectively.
Table 4.1 summarizes the specific models used to illustrate our methods, after we plugged in the appropriate parameter values, chose suitable building covariance functions, and applied some trivial transformations, as necessary.

Table 4.1: Notations of the chosen covariance models.

<table>
<thead>
<tr>
<th>Model</th>
<th>Notation</th>
<th>Formula</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gneiting</td>
<td>$C^G(h, u; \beta)$</td>
<td>$\frac{1}{0.5</td>
</tr>
<tr>
<td>Cressie and Huang separable</td>
<td>$C^{CH}_{sep}(h, u)$</td>
<td>$\frac{1}{(0.5</td>
</tr>
<tr>
<td>Cressie and Huang non-separable</td>
<td>$C^{CH}_{nsep}(h, u)$</td>
<td>$\frac{1}{((0.5</td>
</tr>
<tr>
<td>De Cesare</td>
<td>$C^C(h, u; k)$</td>
<td>$\frac{1}{3}{\exp(-h)\exp(-u) + \exp(-u) + k\exp(-h)}$</td>
</tr>
</tbody>
</table>

4.2.2 Visualization for separability

Noting that separable covariance functions require $C(h, u) = C(h, 0)C(0, u)/C(0, 0)$, for any spatial lag $h$ and temporal lag $u$, or equivalently, $C(h, u)/C(h, 0)$ remains a constant $C(0, u)/C(0, 0)$ for any given $h$. Therefore, we introduce the separability test function defined below.

**Definition 1.** Given a valid spatio-temporal covariance function $C(h, u)$, the separability test function $f_h(u)$ is a function of temporal lag $u$ for any spatial lag $h$, defined as

$$f_h(u) = C(h, u)/C(h, 0) - C(0, u)/C(0, 0).$$

By Definition 1, $f_h(u)$ is 0 for any $u$ and $h$ if $C(h, u)$ is separable, and moves away from zero for non-separable models. Suppose $f_{h_1}(u), \ldots, f_{h_n}(u)$ are a set of test functions for pairs of locations with spatial lags $h_1, \ldots, h_n$. We use functional boxplots [29] to visualize these separability test functions. The motivation is that functional boxplots can characterize the most representative functional realization, or the functional median, and remove the outliers, which are often involved in the
datasets, especially when we are going to use sample covariances to estimate the test functions. Functional boxplots order the functional observations by their functional data depths, identify the observation with the largest depth value as the functional median and construct the 50% central region that covers half of the data with largest depth values. Outliers are detected by the 1.5 times the 50% central region empirical rule. Alternatively, contour plots are the traditional way to visualize the pattern of covariances with respect to different spatial and temporal lags. We show next that the functional boxplots provide a much better visualization and interpretation of the separability than the contour plots.

Firstly, we use functional boxplots of the separability test functions and contour plots of the covariances to visualize the Gneiting model $C^G(h, u; \beta)$ for different values of $\beta$. The spatial locations are on a $4 \times 4$ regular grid within the unit square $[0, 1] \times [0, 1]$. The top panels in Figure 4.1 show the functional boxplots applied to the separability test functions for the Gneiting model $C^G(h, u; \beta)$ with different values of $\beta$. Each curve is associated with a specific pair of spatial locations. The central region (magenta area) and the median (black curve) move away from zero as $\beta$ increases. The bottom panels in Figure 4.1 show the contours of the covariances for different spatial and temporal lags. In the separable case, the covariances at non-zero temporal lags should decay more rapidly as the spatial lag increases. We do see this trend in the contour plots; however, the separability is much more obvious in the functional boxplot.

Similarly, we draw the same plots to compare the Cressie and Huang separable model $C^{CH}_{sep}(h, u)$ and the Cressie and Huang non-separable model $C^{CH}_{nsep}(h, u)$, and the De Cesare model $C^{C}(h, u; k)$ for different values of $k$. The results are shown in Figures 4.2 and 4.3. The functional boxplots clearly show the separability or non-separability of each test function, and, if non-separable, the positive or negative non-separability. While the contour plots show different patterns and the separable
Figure 4.1: Top: functional boxplots of the separability test functions for the Gneiting model $C^G(h, u; \beta)$ with different values of $\beta$. The magenta area is the central region, the black line is the median, and the red lines are outliers. Bottom: contour plots of the covariances for the Gneiting model $C^G(h, u; \beta)$ with different values of $\beta$. Covariance decays much faster than non-separable covariances at non-zero temporal lags along an increasing spatial lag, the indication for separability is still unclear.

4.2.3 Visualization and rank-based testing procedure for real data

In §4.2.1–4.2.2, we discussed the visualization of true spatio-temporal covariance functions, and we saw that the functional boxplots produce intuitive indications of separability in the test functions. However, the true covariance model is unknown in practice, and we must use sample covariance estimators to estimate the test functions without assuming any parametric model to fit. For each pair of spatial locations, we use the two time series to estimate all the components in the separability test function $f_h(u)$ by sample covariances. In other words, if we have observations $Z(s, t)$ at $p$ time
Figure 4.2: Top: functional boxplots of the separability test functions for the Cressie and Huang separable model $C^{CH}_{sep}(h, u)$ and non-separable model $C^{CH}_{nsep}(h, u)$. The magenta area is the central region, the black line is the median, and the red lines are outliers. Bottom: contour plots of the covariances for $C^{CH}_{sep}(h, u)$ and $C^{CH}_{nsep}(h, u)$.

points 1, \ldots, p, at two locations $s_1$ and $s_2$ with a spatial lag $h = s_2 - s_1$, then the test function is estimated by

$$\hat{f}_h(u) = \hat{C}(h, u) / \hat{C}(h, 0) - \hat{C}(0, u) / \hat{C}(0, 0), \quad u = 0, \ldots, p - 1,$$

where

$$\hat{C}(h, u) = \frac{\sum_{i=1}^{p-u} \left\{ Z(s_1, i) - \frac{1}{p-u} \sum_{j=1}^{p-u} Z(s_1, j) \right\} \left\{ Z(s_1 + h, i + u) - \frac{1}{p-u} \sum_{j=1}^{p-u} Z(s_1 + h, j + u) \right\}}{p-u},$$

and other quantities are calculated similarly.

With these estimated separability test functions, we also develop a rank-based test for the separability hypothesis. López-Pintado and Romo \cite{36} proposed a test for functional data to decide whether two sets of samples are from the same distribu-
Figure 4.3: Top: functional boxplots of the separability test functions for the De Cesare model $C_C(h, u; k)$ with different values of $k$. The magenta area is the central region, the black line is the median, and the red lines are outliers. Bottom: contour plots of covariances for $C_C(h, u; k)$ with different values of $k$.

It can be viewed as a functional data generalization of the univariate Wilcoxon rank test. The main idea is to see whether the positions of the samples from the two sets are similar, with respect to a reference distribution. We follow a similar procedure to test the null hypothesis, separability, by comparing the given observations to a simulated dataset under $H_0$: separability. In this rank-based test, two additional datasets are required. One follows the distribution under the null hypothesis $H_0$, and the other is a reference dataset following the distribution under the null hypothesis $H_0$ or the same distribution as the observations. We propose to generate these two datasets under $H_0$ by simulating data from a constructed separable covariance function $\hat{C}^{H_0}$, for which we estimate the covariance at observed spatial and temporal lags by sample covariances, respectively, and let $\hat{C}^{H_0}(h, u) = \hat{C}(h, 0)\hat{C}(0, u)$. After we get all estimated separability test functions, the hypothesis test can be performed to com-
pare the two populations. More details can be found in the paper by López-Pintado and Romo [36], and we briefly explain the procedure as follows:

Step 1 Estimate the test functions for the observations, denoted by \( \hat{f}_{h_i}(u) \), \( i = 1, \ldots, n \).

Step 2 Generate two sets of spatio-temporal datasets from the constructed separable covariance function \( C^{H_0}(h, u) \) and estimate the test functions, denoted by \( \hat{f}_{h_j}^0(u) \), \( j = 1, \ldots, m \) and \( \hat{f}_{h_l}^R(u) \), \( l = 1, \ldots, r \).

Step 3 For each \( i \), combine \( \hat{f}_{h_i}(u) \) and all \( \hat{f}_{h_l}^R(u) \) for \( l = 1, \ldots, r \), and compute their functional Band Depth (BD) and Modified Band Depth (MBD) [36]. Order these functions by their BD values, and use MBD to separate ties. Let \( r_i \) be the proportion of functions \( \hat{f}_{h_l}^R(u) \) with smaller depth values than \( \hat{f}_{h_i}(u) \).

Step 4 Repeat Step 3 for each \( \hat{f}_{h_j}^0(u) \), and let \( r_j' \) be the proportion of functions \( \hat{f}_{h_l}^R(u) \) with smaller depth values than \( \hat{f}_{h_i}(u) \).

Step 5 Order \( r_1, \ldots, r_n, r_1', \ldots, r_m' \) from smallest to largest, and assume the ranks of \( r_1, \ldots, r_n \) are \( q_1, \ldots, q_n \). Then the test statistics is \( W = \sum_{i=1}^{n} q_i \).

Critical values can be determined from the asymptotic distribution of the test statistics, which is the sum of \( n \) numbers that are randomly chosen from 1, \ldots, \( n + m \). In practice, however, we found that the distribution with a limited sample size is not close enough to the asymptotic one. Hence, we use the bootstrap to calculate the critical values. Under \( H_0 \), we generate \( b \) sets of spatio-temporal datasets from \( \hat{C}^{H_0}(h, u) \). We repeat Steps 1–5 and obtain the test statistics for each dataset, which provides an approximate distribution under the null hypothesis. We examine the performance of this technique through simulations in §4.3, where we use \( b = 100 \).

4.2.4 Symmetric covariance functions and visualization

Gneiting et al. [63] proposed the following family of asymmetric covariance models
that include symmetric ones as a special case,

\[
C(h; u) = \frac{(1 - \lambda)}{a|u|^{2\alpha} + 1} \exp \left\{ -\frac{c\|h\|}{(a|u|^{2\alpha} + 1)^{\beta/2}} \right\} + \lambda \left( 1 - \frac{1}{2\nu} |h_1 - vu| \right)_+ ,
\]

where \((\cdot)_+ = \max(0, \cdot)\), \(h_1\) is the first component of the spatial lag \(h\), \(\lambda\) determines the symmetry, \(\nu\) is the velocity along the direction of \(h_1\), \(a\) and \(c\) correspond to the temporal and spatial ranges, \(\alpha\) corresponds to smoothness, and \(\beta\) corresponds to the separability. After plugging in appropriate parameter values, we get the following covariance model to study symmetry, referred as the Gneiting model \(C_{G'}(h, u; \lambda)\).

\[
C_{G'}(h, u; \lambda) = \frac{1 - \lambda}{0.2|u| + 1} \exp(-\|h\|) + \lambda \left( 1 - \frac{1}{2} |h_1 - 0.2u| \right)_+ .
\]

We know that a fully symmetric covariance function satisfies \(C(h, u) = C(h, -u)\). Analogously to Definition 1, we introduce the symmetry test functions with definition below.

**Definition 2.** Given a valid spatio-temporal covariance model \(C(h, u)\), the symmetry test function \(g_h(u)\) is a function of temporal lag \(u\) for any spatial lag \(h\), defined as

\[
g_h(u) = C(h, u) - C(h, -u).
\]

It is easy to see that \(g_h(u)\) remains zero when the underlying covariance model is fully symmetric. An illustration is given in Figure 4.4 for the Gneiting model \(C_{G'}\) with different values of \(\lambda\), where we see the symmetry test function \(g_h(u)\) moves away from zero as \(\lambda\) increases. Contour plots of the covariances are also shown in the bottom panels of Figure 4.4. Since the covariance relies on the temporal lag \(u\) and both the first and second components \(h_1, h_2\) of the spatial lag \(h\), we fix \(h_2 = 0\), and draw the contours for different values of \(u\) and \(h_1\). It is easy to observe that the covariances are clearly asymmetric when \(\lambda > 0\). For real data, the estimated
symmetry test functions are used in the functional boxplots, and a similar rank-based testing procedure described in §4.2.3 is developed to test the significance of the asymmetry.

Figure 4.4: Top: functional boxplots of the symmetry test function for the Gneiting model $C_G'(h, u; \lambda)$ with different values of $\lambda$. The magenta area is the central region, the black line is the median, and the red lines are outliers. Bottom: contour plots of covariances for $C_G'(h, u; \lambda)$ with different values of $k$ where $h_2$ is fixed to 0.

4.3 Simulation

4.3.1 Simulation design

Spatio-temporal data collected from monitoring sites are often sparse in space but dense in time. In this simulation, we consider data from $4 \times 4 = 16$ regularly spaced locations in the unit square $[0, 1] \times [0, 1]$, and $p = 2,000$ (a comparable number to many real applications) equally spaced time points.

Noting that the correlation between observations with temporal lags greater than 100 is negligible in each model, we generate our spatio-temporal data sequentially in
time. Let $Z_i$ be a matrix of size $16 \times 100$, indicating the spatio-temporal observations from time point $(i-1) \times 100 + 1$ to $i \times 100$ at the 16 locations, for $i = 1, \ldots, 20$. We first generate $Z_1$, and then generate $Z_i$, $i = 2, \ldots, 20$, only conditioning on $Z_{i-1}$. Thus the entire generated spatio-temporal data is $Z = (Z_1, Z_2, \ldots, Z_{20})$.

4.3.2 Visualization and assessment for separability

The Gneiting model $C^G(h, u; \beta)$ with five different values of $\beta = 0, 0.25, 0.5, 0.75,$ and $1$ is used to generate observations $Z$. Three examples of the estimated separability test functions are shown in Figure 4.5. We see that, in the functional boxplot, the distance between the median (black line) and zero (green dashes) indicates the degree of separability. As $\beta$ increases, the majority of the separability test functions move away from zero. This gives us a clear visualization of the separability.

![Figure 4.5: Functional boxplots of the estimated separability test function for data generated from the Gneiting model $C^G(h, u; \beta)$ with $\beta = 0, 0.5,$ and $1$. The magenta area is the central region, the black line is the median, the red lines are outliers, and the green dash line is the reference for separability in theory.](image)

<table>
<thead>
<tr>
<th>$\beta$</th>
<th>0</th>
<th>0.5</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>-0.3</td>
<td>-0.1</td>
<td>0.1</td>
</tr>
<tr>
<td>0.25</td>
<td>-0.3</td>
<td>-0.1</td>
<td>0.1</td>
</tr>
<tr>
<td>0.5</td>
<td>-0.3</td>
<td>-0.1</td>
<td>0.1</td>
</tr>
<tr>
<td>0.75</td>
<td>-0.3</td>
<td>-0.1</td>
<td>0.1</td>
</tr>
<tr>
<td>1</td>
<td>-0.3</td>
<td>-0.1</td>
<td>0.1</td>
</tr>
</tbody>
</table>

Table 4.2: Type I error and power for simulated data from the Gneiting model $C^G(h, u; \beta)$ with different values of $\beta$ in 100 experiments.

<table>
<thead>
<tr>
<th>Nominal Level</th>
<th>Type I error</th>
<th>Power</th>
</tr>
</thead>
<tbody>
<tr>
<td>Proportion of rejections</td>
<td>$\alpha$</td>
<td>$\beta = 0$</td>
</tr>
<tr>
<td>$5%$</td>
<td>0.04</td>
<td>0.26</td>
</tr>
<tr>
<td>$1%$</td>
<td>0.02</td>
<td>0.10</td>
</tr>
</tbody>
</table>
To formally test the separability and examine the performance of the proposed rank-based testing procedure in §4.2.3, we show the Type I error and power for simulated data over 100 simulations. The results are shown in Table 4.2. We see the Type I error is close to the nominal level of the hypothesis test, and the power for non-separable models increases with an increasing $\beta$. The power is quite close to one when $\beta$ approaches 1.

In addition, we also take a look at the Cressie-Huang models $C_{CH}^{sep}$ and $C_{CH}^{nsep}$. Estimated separability test functions from one simulated dataset are shown in Figure 4.6. We see the majority of the separability test functions from the non-separable model move away from zero. Similarly, a rank-based test is used to examine the performance; the results are shown in Table 4.3.

Figure 4.6: Functional boxplot of the estimated separability test function for data generated from Cressie-Huang models $C_{CH}^{sep}$ and $C_{CH}^{nsep}$. The magenta area is the central region, the black line is the median, the red lines are outliers, and the green dash line is the reference for separability in theory.

<table>
<thead>
<tr>
<th>Nominal Level</th>
<th>Model</th>
<th>$C_{CH}^{sep}$</th>
<th>$C_{CH}^{nsep}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Proportion of rejections</td>
<td>$5%$</td>
<td>$0.04$</td>
<td>$0.97$</td>
</tr>
<tr>
<td></td>
<td>$1%$</td>
<td>$0.00$</td>
<td>$0.87$</td>
</tr>
</tbody>
</table>
4.3.3 Visualization and assessment for symmetry

We choose five different values of $\lambda = 0, 0.025, 0.05, 0.075, \text{ and } 0.1$ in the Gneiting model $C^G'(h, u; \lambda)$ to generate observations $Z$. The reason for choosing relatively small values for $\lambda$ is that the power of the rank-based test is already strong enough for small values of $\lambda$. Three examples of the estimated symmetry test functions are shown in Figure 4.7. Since the chosen values of $\lambda$ are very small, the deviation of the symmetry test functions from zero in our most extreme case when $\lambda = 0.1$ is still subtle; see Figure 4.4. Although the functional medians of the estimated test functions do not show obvious changes for the different values of $\lambda$, we can still see that the majority of the symmetry test functions move away from zero as $\lambda$ increases. Furthermore, the results of the rank-based test in Table 4.4 show that the power is very strong, even when $\lambda$ is as small as 0.01, and the Type I errors are all close to the nominal levels.

![Functional boxplots of the estimated symmetry test function for data generated from the Gneiting model $C^G'(h, u; \lambda)$ with $\lambda = 0, 0.05, \text{ and } 0.1$. The magenta area is the central region, the black line is the median, the red lines are outliers, and the green dash line is the reference for separability in theory.](image-url)
Table 4.4: Type I error and power for simulated data from the Gneiting model $C^G(h, u; \lambda)$ with different values of $\lambda$ in 100 experiments.

<table>
<thead>
<tr>
<th>Nominal Level</th>
<th>Type I error</th>
<th>Power</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\alpha$</td>
<td>$\lambda = 0$</td>
<td>$\lambda = 0.025$</td>
</tr>
<tr>
<td>Proportion of rejections</td>
<td>5%</td>
<td>0.04</td>
</tr>
<tr>
<td>1%</td>
<td>0.00</td>
<td>0.01</td>
</tr>
</tbody>
</table>

4.4 Applications

4.4.1 Wind speed

Wind speed is an important atmospheric variable in weather forecasting and many other environmental applications. One source of wind speed data is from monitoring stations. In this section, we use our proposed method to visualize and assess properties of covariances in the wind speed from ten monitoring stations in the northwestern U.S.A. The locations are shown in Figure 4.8 where we divide them into two regions. Coastal sites are in green, and inland sites are in orange. We analyze the hourly data observed at each station from 2012 to 2013 and visualize the spatio-temporal covariance structure from two seasons, summer (June to August) and winter (December to February).

Figure 4.8: Ten locations for observing wind speed in a broad (left) and zoomed (right) view. The green points are grouped in the coastal class and the orange points are grouped in the inland class.

The plots of the estimated separability test functions are shown in Figure 4.9.
and the p-value results of the rank-based test are given in the titles. We see the separability test functions are away from zero in all cases, and all the p-values for testing separability are much smaller than 0.05, indicating rejection of separability at $\alpha = 0.05$. When computing the p-value, we choose 10,000 replicates from the bootstrap to approximate the distribution of the test statistics under the null hypothesis. Since the covariance functions for all the four cases are significantly non-separable, we next investigate their properties of symmetry. The symmetry test functions as well as the p-values of the test for different cases are shown in Figure 4.10. The functional boxplot for the coastal region in winter suggests the strongest symmetry, and its corresponding p-value is indeed the largest, while p-values for the other cases are much smaller than 0.05. One possible reason might be that there is no prevailing wind direction during the winter in the coastal region.

Figure 4.9: The estimated separability test functions for wind speed in the northwestern U.S.A. during summer and winter, in coastal and inland regions. The p-values of the rank-based test for separability are indicated in the title of each case.
Figure 4.10: The estimated symmetry test functions for wind speed in the northwestern U.S.A. during summer and winter, in coastal and inland regions. The p-values of the rank-based test for symmetry are indicated in the title of each case.

4.4.2 Climate model outputs

In climate studies, numerical models are used to simulate many important variables by solving a set of dynamic equations. Two types of models are available: a General Circulation Model (GCM) is used to describe the dynamic system of the entire earth, with a global coverage but a comparatively low resolution; however, a Regional Climate Model (RCM) with GCM outputs for boundary conditions is used to simulate locally high-resolution data. Different combinations of GCMs and RCMs may lead to different results. In this section, we apply our visualization and assessment tools to data based on four combinations of two GCMs, GFDL and HADCM3, and two RCMs, ECP2 and HRM3, which are provided by the North American Regional Climate Change Assessment Program (NARCCAP) [78]. More details about these numerical models can be found at http://www.narccap.ucar.edu/ We focus on a small region, which consists of 16 spatial locations in the North Atlantic Ocean as
Figure 4.11: Study area in the North Atlantic Ocean, where the white dots indicate the spatial locations from the model outputs.

shown in Figure 4.11 and study the daily surface temperature and wind speed from 1976 to 1979.

We first look at the mean of the surface temperature across the four years at each location, which is shown in Figure 4.12. We see different GCMs as boundary conditions give different patterns of surface temperatures, while different RCMs show similar results. We then apply our visualization and assessment tools to the covariance structures of these four GCM and RCM combinations. To eliminate seasonality, we remove the monthly mean from the daily temperatures. The estimated separability test functions and the p-values of the rank-based test for separability are shown in Figure 4.13. With all the p-values greater than 0.05, there is no significant evidence that the covariances of these four cases are non-separable at the 5% significance level. However, the output from GFDL and HRM3 shows stronger non-separability and the
Figure 4.12: Mean of the sea surface temperature (unit: K) during 1976–1979 at each of the 16 locations.

separability is rejected at the 10% significance level with a p-value 0.09.

Next, we look at the daily wind speed in the same area. The mean wind speeds at these locations are all close to zero and do not show much difference. Again, we remove the monthly mean from the daily wind speed and then apply our visualization and assessment methods. The estimated separability test functions are shown in Figure 4.14, where the medians are away from zero. These test functions show large variability in the functional boxplot, and all the p-values from the rank-based tests are as small as zero, indicating strong non-separability. To further investigate the symmetry, we plot the estimated symmetry test functions; the illustration is given in Figure 4.15. All the symmetry test functions are also very far from zero, and we find that the p-values of the test for symmetry are zero. Therefore, we conclude that the covariance of wind speed is neither separable nor symmetric.

In this application, the covariance of the daily surface temperatures tends to be separable and thus also symmetric; for daily wind speed, the covariance is non-separable and asymmetric. Our analysis suggests that, for the region we have considered, the daily surface temperatures produced by the climate models do not show
Figure 4.13: The estimated separability test functions for daily surface temperature with different combinations of GCMs and RCMs. The p-values of the rank-based test for separability are indicated in the title of each case.

significant space-time interaction, whereas the daily wind speeds clearly show asymmetric space-time interaction.

### 4.5 Discussion

We presented a functional data analysis approach to visualizing and assessing spatio-temporal covariance properties. The proposed method is suitable for visualizing the separability and the symmetry of a stationary spatio-temporal process. The p-value calculated in the rank-based test can be used to measure the degree of separability or symmetry. We illustrated our methods using various classes of spatio-temporal covariance models, and our simulations demonstrated the good performance of our proposed methods. In the practical applications, we applied our method to temperature and wind speed data from either monitoring sites or climate model outputs, and illustrated how to interpret the separability and symmetry in these different scenarios.
Figure 4.14: The estimated separability test functions for daily wind speed with different combinations of GCMs and RCMs.

Figure 4.15: The estimated symmetry test functions for daily wind speed with different combinations of GCMs and RCMs.
However, we have only considered spatio-temporal datasets collected from a small number of spatial locations. Because we need to estimate the test functions for each pair of locations, computation will become an issue as the number of locations grows. One possible solution is to divide the region of interest into smaller subregions, and estimate all the test functions separately for each of these subregions. Then, the obtained test functions can be combined for the visualization and the rank-based test. This would significantly reduce the number of test functions, and each small subregion is more likely to be stationary. However, future research is required to assess the performance of this approach.

Acknowledgement

We wish to thank North American Regional Climate Change Assessment Program (NARCCAP) for providing the data used in this thesis. NARCCAP is funded by the National Science Foundation (NSF), the U.S. Department of Energy (DoE), the National Oceanic and Atmospheric Administration (NOAA), and the U.S. Environmental Protection Agency Office of Research and Development (EPA).
Chapter 5

Fast Kriging with Hierarchical Low-Rank Approximation for Large Spatial Datasets

5.1 Introduction

In spatial statistics, with a fitted spatial model, one can provide predictions at unsampled locations, or impute missing values for spatial datasets. Such spatial interpolation problems find applications in a wide range of scientific research areas. Bradley et al. [79] reviewed different prediction methods and classified them as deterministic and stochastic predictors. Deterministic approaches do not consider the stochastic component in the underlying process. For example, spline predictors [80] solve a minimization problem with a penalized least-square objective function, accounting for both the goodness-of-fit and the smoothness; negative-exponential distance-weighting methods [12] lead to a predictor as a weighted average of the observations, where the weight of each observation is assigned according to its distance to the prediction location, particularly, exponentially decayed as distance increases. In contrast, stochastic methods derive predictors using statistical models, such as kriging [12, 81], fixed rank kriging [82], modified predictive process [10], integrated nested Laplace approximations based on stochastic partial differential equations [83], and lattice kriging [19]. The kriging predictor is the best unbiased linear predictor which minimizes the variance of the mean squared prediction errors.

The standard implementation of kriging requires solving a linear system of the covariance matrix of the observations. However, the size of the linear system depends
on the number of locations, and the computation becomes prohibitive when data are available at a large number of irregularly spaced locations. In geostatistics or remote sensing, such large spatial datasets are not unusual. Moreover, computationally efficient kriging algorithms are also needed in Bayesian inference for large spatial datasets, such as sampling from one of the full conditional distributions in Gibbs sampling. When the exact computation for prediction is not feasible, one needs approximation methods. Fixed rank kriging \cite{82} uses a low-rank approximation by choosing a small number of basis functions to characterize the variability of the random process. Then the computation is significantly reduced when the rank is much smaller than the number of locations. The predictive process method \cite{10} shares a similar idea, but uses a latent predictive process on a small number of knots over the spatial domain under study for fast computations in the Bayesian hierarchical model. Unlike the fixed rank kriging which implies a low-rank covariance matrix, Markov models assume conditionally independent observations, leading to a sparse precision matrix that allows for fast computations. Gaussian Markov random field approximation has been exploited to achieve fast Bayesian prediction for latent Gaussian random field models through integrated nested Laplace approximations \cite{83}. Prediction with tapered covariance function was proposed by \cite{84}, where the covariance function is multiplied by a compactly supported positive definite function, and efficient sparse matrix algorithms are used to solve a sparse approximate linear system. The lattice kriging method \cite{19} also takes advantage of sparse matrix algorithms, but applies different numbers of the compactly supported basis functions on each level, leading to a multiresolution representation of the spatial process.

In this chapter, we develop a fast kriging algorithm for large and irregularly spaced spatial datasets using the hierarchical low-rank framework introduced in Chapter 2. For observations on a regular grid, the possibility of exact computations should be explored first. As explained in \cite{2,5,2} when solving the linear system, the matrix-vector
multiplication is best done by embedding the covariance matrix of block Toeplitz and Toeplitz block in a circulant matrix and using fast Fourier transformation for the computation of the product. We describe the HLR kriging in §5.3 and examine the approximation quality numerically in §5.4. The application of the HLR kriging for satellite data gap-filling problems is presented in §5.5. We make a summary and point out research directions for further studies in §5.6.

5.2 Kriging

Suppose the random field \{Z(s) : s ∈ D ⊂ \mathbb{R}^d\} in the \(d\)-dimensional Euclidean space (typically \(d = 2\)) has a representation as follows,

\[
Z(s) = m(s)^T \beta + \epsilon(s),
\]

where \(m(s)\) is a known \(p\)-dimensional covariate vector, \(\beta\) is the unknown \(p\)-dimensional regression coefficient vector, and \(\epsilon(s)\) is a zero-mean spatial random process. Although, \(\epsilon(s)\) can be further divided to describe different scales of variations including nugget effects, here, for simplicity, we do not differential them and use a single process \(\epsilon(s)\) only. Suppose \(\epsilon(s)\) is a second-order stationary process with mean zero, and a parametric covariance function \(C(\cdot; \theta)\), where \(\theta\) is the parameter vector of the covariance function. Then, the covariance between \(Z(s_i)\) and \(Z(s_j)\) is \(C(s_i - s_j; \theta)\).

For simplicity, we present the simple Kriging without considering the mean \(m(s)^T \beta\). Suppose \(Z(s)\) has mean zero and is observed at \(n\) locations \(s_1, \ldots, s_n\), and we are interested in predicting \(Z(s)\) at location \(s_0\). Kriging, or the Best Linear Unbiased
Predictor (BLUP), at $s_0$ is $\widehat{Z}(s_0) = \sigma^T \Sigma^{-1} Z$, where

$$Z = \{Z(s_1), \ldots, Z(s_n)\}^T,$$

$$\Sigma = \text{Cov}(Z, Z^T), \quad \text{i.e.} \quad \Sigma_{(i,j)} = \text{Cov}\{Z(s_i), Z(s_j)\},$$

$$\sigma = \text{Cov}\{Z, Z(s_0)\}, \quad \text{i.e.} \quad \sigma_{(i)} = \text{Cov}\{Z(s_i), Z(s_0)\}.$$

We briefly show how the BLUP is derived. Since kriging is a linear predictor, we write $\widehat{Z}(s_0)$ as a linear combination of all the observations $Z(s_i)$, or,

$$\widehat{Z}(s_0) = \sum_{i=1}^{n} \lambda_i Z(s_i),$$

where we call all the values of $\lambda_i$ the kriging weights and denote $\lambda = (\lambda_1, \ldots, \lambda_n)^T$.

Since we do not consider the mean $m(s)^T \beta$ at the moment, the unbiasedness is guaranteed naturally due to the zero mean of $\epsilon(s)$. Then, the BLUP requires to minimize the variance of prediction error $\text{Var}\{\widehat{Z}(s_0) - Z(s_0)\}$. Because

$$\text{Var}\{\widehat{Z}(s_0) - Z(s_0)\} = \text{Var}\{\widehat{Z}(s_0)\} + \text{Var}\{Z(s_0)\} - 2\text{Cov}\{\widehat{Z}(s_0), Z(s_0)\}$$

$$= \text{Var}\{\lambda^T Z\} + \text{Var}\{Z(s_0)\} - 2\text{Cov}\{\lambda^T Z, Z(s_0)\}$$

$$= \lambda^T \Sigma \lambda + \text{Var}\{\epsilon(s_0)\} - 2\lambda^T \sigma,$$

where $\text{Var}\{\epsilon(s_0)\}$ does not depend on $\lambda$, we know the kriging weights are the solution of the optimization problem as follows,

$$\widehat{\lambda} = \arg\min_{\lambda \in \mathbb{R}^n} \{\lambda^T \Sigma \lambda - 2\lambda^T \sigma\}.$$

One can easily show that $\widehat{\lambda} = \Sigma^{-1} \sigma$.

For universal kriging, where $m(s)$ is not zero, we denote $M = \{m(s_1), \ldots, m(s_n)\}^T,$
and then it can be demonstrated that the kriging weights have the form as follows,

$$\hat{\lambda} = \{\Sigma^{-1} - \Sigma^{-1}M(M^T\Sigma^{-1}M)^{-1}M^T\Sigma^{-1}\}\sigma + \Sigma^{-1}M(M^T\Sigma^{-1}M)^{-1}m(s_0).$$

It is derived from solving a similar optimization problem for minimizing the variance of the prediction error but with a constraint for unbiasedness. More details can be seen in the book by Cressie [12].

For the simple kriging predictor $\hat{Z}(s_0) = \sigma^T\Sigma^{-1}Z$, the computational cost is $O(n^3)$ for calculating the kriging weights $\lambda = \Sigma^{-1}\sigma$ with Cholesky decomposition, and efficient approximation methods are needed for large spatial datasets.

5.3 HLR kriging

In this section, we propose the HLR kriging by approximating the kriging weights using the HLR representation of $\Sigma^{-1}$ for large and irregularly spaced spatial datasets.

Recall the notations from Chapter 2. Let $s_1, \ldots, s_n$ be the locations ordered according to their coordinates. For $j = 1, \ldots, n - 1$, we denote

$$Z_j = \{Z(s_1), \ldots, Z(s_j)\}^T,$$
$$\Sigma_{jj} = \text{Cov}(Z_j, Z_j),$$
$$\sigma_j = \text{Cov}\{Z_j, Z(s_{j+1})\}.$$ 

Then $Z_j$ is a $j$-dimensional vector of observation values from locations $s_1$ to $s_j$. $\Sigma_{jj}$ is a $j \times j$ matrix of covariances for the first $j$ observations. $\sigma_j$ is a $j$-dimensional vector of covariances between the first $j$ observations and the $(j + 1)$-th observation. Denote the $n$-dimensional vectors $b_j = (-\sigma_j^T\Sigma_{jj}^{-1}1, 0, \ldots, 0)^T$, for $j = 1, \ldots, n - 1$ and $b_0 = (1, 0, \ldots, 0)^T$. One can show that the precision matrix is

$$\Sigma^{-1} = \sum_{j=0}^{n-1} b_j b_j^T/(b_j^T \Sigma b_j).$$
For large $j$, the computation of $\Sigma_{jj}^{-1}$ in $b_j$ becomes expensive. When $j > r$, where $r$ is the maximum rank considered in $j$-th approximation step, the methods proposed in Chapter 2 use the following approximation,

$$\hat{b}_j = \left\{ -\left\{ A_{j,r}^T \Sigma_{jj} A_{j,r} \right\}^{-1} A_{j,r}^T \sigma_j \right\}^T, 1, 0, \ldots, 0]^T, j = r + 1, \ldots, n - 1.$$

The kriging weights can be computed accordingly with different choices of $A_{j,r}$ associated with approximation methods proposed in §2.2.2 and §2.2.3.

Here, we propose the HLR kriging by the HLR approximation introduced in §2.2.3 with

$$\hat{b}_j = \left\{ -\left\{ A_{j,mr}^T \Sigma_{jj} A_{j,mr} \right\}^{-1} A_{j,mr}^T \sigma_j \right\}^T, 1, 0, \ldots, 0]^T, j = r + 1, \ldots, n - 1,$$

where $m$ is a small number greater than 1 to include more neighbors. In addition, $A_{j,mr}^T \Sigma_{jj} A_{j,mr}$ is approximated by $r$ basis functions with $A_{j,mr}^T \Sigma_{jj} A_{j,mr} \approx P_j L_j P_j^T + \epsilon_j^2 I_{mr}$, where $P_j$ is an $mr \times r$ matrix generated by the $r$ basis functions, $L_j$ is an $r \times r$ positive definite matrix, $I_{mr}$ is the $mr \times mr$ identity matrix, and $\epsilon_j^2$ is the nugget. Then, the inverse of $A_{j,mr}^T \Sigma_{jj} A_{j,mr}$ is computed by

$$(A_{j,mr}^T \Sigma_{jj} A_{j,mr})^{-1} = \epsilon_j^{-2} I_{mr} - \epsilon_j^{-4} P_j (L_j^{-1} + \epsilon_j^{-2} P_j^T P_j)^{-1} P_j^T.$$

The computational complexity becomes $O(r^3 n)$, and this HLR kriging can significantly ease the computational burden when the rank, $r$, is much smaller than the number of locations, $n$.

### 5.4 Numerical studies

In this section, we examine the quality of the HLR kriging weight approximation for different spatial processes. We use the same simulation design as in §2.3. We first
generate irregularly spaced locations on the unit square $[0, 1] \times [0, 1]$ at $n^{-1/2}(r - 0.5 + X_{r\ell}, \ell - 0.5 + Y_{r\ell})$ for $r, \ell \in \{1, \ldots, n^{1/2}\}$, where $X_{r\ell}$s and $Y_{r\ell}$s are independent and identically distributed, uniform on $(-0.4, 0.4)$. We assume a zero-mean process with a Matérn covariance function in Equation (2.5) as shown below,

$$C(h; \alpha, \beta, \nu, \tau^2) = \alpha \{(2\nu)^{1/2}h/\beta\}^\nu K_\nu\{(2\nu)^{1/2}h/\beta\}/\{\Gamma(\nu)2^{\nu-1}\} + \tau^2 \mathbb{1}(h = 0),$$

where $\beta$ is the range parameter, $\nu$ is the smoothness parameter, and $\tau^2$ is the nugget effect.

We use a small design with $n = 100$ locations to illustrate the approximation quality for different choices of $\beta$, $\nu$, and $\tau^2$. Suppose the prediction location is $s_0 = (0.5, 0.5)$. We compute the approximated kriging weights with $r = 2$ and 10, and compare them with the exact ones. The results are shown in Figures 5.1 and 5.2 where the locations on the axis are ordered according to an increasing distance to $s_0$. The exact kriging weights are decreasing to zero as the site moves far away from $s_0$. When we increase $r$ from 2 to 10, the HLR kriging provides a much better approximation with all the errors close to zero. In the top-left panel of Figure 5.1, the approximation is the best among the four cases. We now discuss the other three cases compared to this one. We see that the approximation becomes worse when $\beta$ increases from 0.1 to 0.5 in the top-right panel of Figure 5.1. This is because the rank $r = 2$ is too small and fails to select enough highly correlated observations in the approximation. When the process is smoother for $\nu = 1$ in the bottom-left panel, the approximation also becomes worse. It suggests that we should increase $r$ and select more neighbors for a smoother process to improve the approximation. For a less noisy process when the nugget is zero in the bottom-right panel, the result turns out to be similar.
Figure 5.1: Four panels showing the exact kriging weights (black) and HLR kriging weights (red) for 100 locations with rank $r = 2$. The blue line is the absolute difference between the exact and HLR kriging weights.

5.5 Application to satellite images

We give a simple application in this section using our proposed HLR kriging. Landsat 7 satellite, launched in 1999, is an important instrument to provide surface images of the earth. However, something went wrong with the Scan Line Corrector in Landsat 7 in 2003, which made the satellite fail to adjust forward motion and led to images presented with gaps from then on. An illustration is given in Figure 5.3 where the image is based on the wave with bandwidth from 0.45 to 0.52 micrometers from the sensor of Landsat 7. For other satellites, although there are no systematic gaps, gaps
can still exist due to several reasons, such as the sensor is blocked by clouds or is temporarily out of work. To fill these gaps, we use kriging as a spatial interpolator. We focus on a region consisting of 50 × 50 pixels so that the exact kriging is feasible. We use the HLR kriging with rank $r = 2$ and 10 for comparisons. The results are shown in Figure 5.4, where we see the three prediction results show similar patterns, indicating a good approximation of the HLR kriging. Moreover, the root mean squared difference of the HLR kriging prediction with $r = 10$ from the exact kriging prediction is around 50% smaller than that of the HLR kriging prediction with $r = 2$. 

Figure 5.2: Four panels showing the exact kriging weights (black) and HLR kriging weights (red) for 100 locations with rank $r = 10$. The blue line is the absolute difference between the exact and HLR kriging weights.
5.6 Discussion and future work

In this chapter, we introduced the HLR kriging to approximate kriging weights for large spatial datasets. The numerical studies illustrated the approximation quality given the true covariance function. Although the HLR Kriging weights provided a good approximation for the covariance functions we have considered, and showed a good prediction map for the gap-filling problem, further studies on the prediction errors using estimated covariances are needed, and efficient algorithms should be implemented for real applications with large spatial datasets.
Figure 5.4: The Landsat 7 satellite image with gaps for the study area (top-left). Image with gap filled by the HLR kriging with rank \( r = 2 \) (top-right) and \( r = 10 \) (bottom-left), and the exact kriging (bottom-right).
Chapter 6

Conclusion

In this thesis, motivated by challenges of big and high-dimensional data, we first studied computational methods for large spatial datasets. Specifically, we proposed a new approximation scheme for fitting Gaussian process models to large and irregularly spaced spatial datasets, as well as providing fast kriging algorithms for spatial prediction and interpolation. Secondly, we investigated data depth-based outlier detection methods for functional data. In particular, we proposed a new notion of functional data depth and developed outlier detection procedures that can detect both magnitude and shape outliers. Finally, we exploited the functional data ranking technique for visualizing and assessing properties of spatio-temporal covariances and developed rank-based tests for testing the separability and symmetry of the spatio-temporal covariance function.

The main contribution of the proposed Hierarchical Low-Rank Approximation Method (HLR) in Chapter 2 is that it provides a unified framework for Gaussian likelihood approximation. By writing the likelihood as a product of conditional densities, the low-rank approximation was applied to each conditional density hierarchically. Therefore, it provides a much more accurate approximation, and the resulting approximated likelihood or covariance is not low-rank anymore. The proposed HLR approximation framework also allows for comparing the statistical efficiency of different types of approximation. Our simulation studies showed that the HLR method significantly improves the statistical efficiency of the commonly used independent block and the nearest neighbor approaches, especially when the maximum rank, $r$,
for which the computation is feasible, is much smaller than the number of observations, \( n \). For an illustration of the presented methods, a real application of massive datasets consisting of 2 million spatial measurements was analyzed, where the model was fitted with the HLR approximation and cross-validation. In our study, we have fixed the rank, \( r \), for each hierarchy. Future research is needed to choose \( r \) adaptively to further improve the approximation quality without a significant increase in computations.

Functional outlier detection is a challenging problem, especially for detecting shape outliers. In Chapter 3, we proposed an effective outlier detection procedure via the decomposition of the developed Total Variation Depth (TVD). Our proposed TVD has a form of integrated depth and was constructed by taking the integration of a univariate depth function. We showed that the novel formulation allows for a useful decomposition of the total variance into shape and magnitude components. Such decomposition considers two adjacent time points of a function, and thus naturally accounts for time order and better characterizes the shape of a function. Informative visualization tools were also developed for presenting detected potential outliers for further investigation. Three applications of different data types, curves, images and video frames, were studied to illustrate our proposed methodology. As discussed in §3.5, the proposed total variation depth can be extended to multivariate functional data, although the shape outlier detection is even more challenging as it involves cross-correlation between any two variables of the multivariate functional data.

Many research questions in the interaction between functional data analysis and spatio-temporal statistics still remain unexplored. In Chapter 4, we brought the functional data ranking technique for visualizing and assessing spatio-temporal covariance properties, separability and symmetry in particular. The test function was estimated as a function of the temporal lag for any two given locations, and the rank-based test induced by functional data depth was developed to test the significance of the
non-separability and asymmetry. The proposed method also provides a visualization tool to compare spatio-temporal patterns of different data products, such as weather station data from different regions, and climate model simulations with different conditions. In spatio-temporal statistics, the covariance function plays a central role in describing the variability of the process. The conventional methods assess its properties only for selected time lags and spatial distances, rather than treating it as a functional quantity and exploiting functional data analysis approaches. In our simulation studies and applications, we have shown the good performance of the proposed methods for various commonly used spatio-temporal models. However, we only analyzed spatio-temporal datasets that are frequent in time but only from a moderate number of locations. For datasets with more spatial locations, further studies are needed because this may cause computational issues and the spatial region may not be stationary. One option is to estimate the test functions locally, and then use the proposed methods to summarize all the test functions from different regions.

Finally, in Chapter 5 we revisited the computational issues for large spatial dataset problems but focused on developing fast kriging algorithms using the HLR approximation scheme proposed in Chapter 2. We examined the approximation quality numerically assuming the covariance function is known. Performances of the proposed HLR kriging when using an estimated covariance function could be investigated further.
REFERENCES


[53] R. J. Hyndman and H. L. Shang, “Rainbow plots, bagplots, and boxplots for


[75] B. Li, M. G. Genton, and M. Sherman, “A nonparametric assessment of prop-


APPENDICES

A  The proof of Theorem 1

Recall that the dimension of $V^H_{jj}$ is $mr \times r$, $V^N_{jj}$ is $r \times r$, $A_{j,r}$ is $j \times r$, $A_{j,mr}$ is $j \times mr$, and $\Sigma_{jj}$ is $j \times j$. Define $B$ to be the $mr \times r$ matrix by keeping the $mr$ selected rows from $A_{j,r}$, or $B = A_{j,mr}^T A_{j,r}$. Let $M$ denote $A_{j,mr}^T \Sigma_{jj} A_{j,mr}$. The proof of Theorem 1 is as follows.

Proof of Theorem 1

With the equality as follows,

$$
\|A_{j,mr} V^H_{jj} A_{j,mr}^T - \Sigma_{jj}\|_F^2 - \|A_{j,r} V^N_{jj} A_{j,r}^T - \Sigma_{jj}\|_F^2
$$

$$
= \|A_{j,mr}^T (A_{j,mr} V^H_{jj} A_{j,mr}^T - \Sigma_{jj}) A_{j,mr}\|_F^2 - \|A_{j,mr}^T (A_{j,r} V^N_{jj} A_{j,r}^T - \Sigma_{jj}) A_{j,mr}\|_F^2
$$

$$
= \|V^H_{jj} - M\|_F^2 - \|BV^N_{jj} B^T - M\|_F^2
$$

it suffices to show,

$$
\|V^H_{jj} - M\|_F \leq \|BV^N_{jj} B^T - M\|_F.
$$

Noting that $V^H_{jj} = P_j L_j P_j^T + \epsilon^2_j I_{mr}$, we have $\|V^H_{jj} - M\|_F = \|P_j L_j P_j^T - (M - \epsilon^2_j I_{mr})\|_F$.

Since $\epsilon^2_j < (\lambda_\gamma + \lambda_{mr})/2$, we know that the eigenvalues of $M - \epsilon^2_j I_{mr}$ satisfy $\lambda_1 - \epsilon^2_j \geq \lambda_2 - \epsilon^2_j \geq \cdots \geq \lambda_r - \epsilon^2_j$ and $|\lambda_r - \epsilon^2_j| \geq \max_{k=r+1}^{mr} (|\lambda_k - \epsilon^2_j|)$. Thus, $|\lambda_1 - \epsilon^2_j| \geq |\lambda_2 - \epsilon^2_j| \geq \cdots \geq |\lambda_r - \epsilon^2_j| \geq \max_{k=r+1}^{mr} (|\lambda_k - \epsilon^2_j|)$. By the construction of $P_j$ and $L_j$,
and Eckart-Young-Mirsky theorem \cite{85, 86}, we know,

\[
\|P_j L_j P_j^T - (M - \epsilon_j^2 I_{mr})\|_F = \inf_{\text{rank}(X) \leq r} \|X - (M - \epsilon_j^2 I_{mr})\|_F.
\]

Noting that the rank of $BV^N_{jj} B^T$ is $r$, we have $\|V_{jj}^H - M\|_F = \|P_j L_j P_j^T - (M - \epsilon_j^2 I_{mr})\|_F \leq \|BV^N_{jj} B^T - (M - \epsilon_j^2 I_{mr})\|_F = \|(BV^N_{jj} B^T - M) - \epsilon_j^2 I_{mr}\|_F$. It is easy to observe that the diagonal elements of $BV^N_{jj} B^T - M$ is non-positive, thus $\|(BV^N_{jj} B^T - M) - \epsilon_j^2 I_{mr}\|_F \leq \|BV^N_{jj} B^T - M\|_F$. Then $\|V_{jj}^H - M\|_F \leq \|BV^N_{jj} B^T - M\|_F$. This completes the proof.
B Estimation of the total variation depth and shape similarity

Recall that the pointwise total variation depth is defined as $D_f(t) = p_f(t)\{1 - p_f(t)\}$, where $p_f(t) = \mathbb{P}\{X(t) \leq f(t)\}$. Given $n$ observations at $m$ time points $X_j(t_i)$, $j = 1, \ldots, n$, $i = 1, \ldots, m$, we estimate $D_f(t_i)$ by $\hat{D}_f(t_i) = \hat{p}_f(t_i)\{1 - \hat{p}_f(t_i)\}$, where $\hat{p}_f(t_i) = \frac{\#\{j : X_j(t_i) \leq f(t_i)\}}{n}$, the proportion of $X_j(t_i)$’s that are below $f(t_i)$. To estimate the shape similarity, which is defined by $S_f(t; \Delta) = \text{Var}(\mathbb{E}[R_f(t) \mid R_f(t - \Delta)]) / D_f(t)$ for $D_f(t) \neq 0$, namely, $p_f(t_i) \neq 1$, we show how to estimate $\text{Var}(\mathbb{E}[R_f(t) \mid R_f(t - \Delta)])$. The population estimation is done on the discrete time points, and at each $t_i, i > 1$, we let $\Delta = t_i - t_{i-1}$. Since

$$\mathbb{E}\{R_f(t_i) \mid R_f(t_{i-1})\} = \mathbb{P}\{R_f(t_i) = 1 \mid R_f(t_{i-1})\} = \mathbb{P}\{X(t_i) \leq f(t_i) \mid R_f(t_{i-1})\},$$

then

$$\text{Var}[\mathbb{E}\{R_f(t_i) \mid R_f(t_{i-1})\}] = \text{Var}[\mathbb{P}\{X(t) \leq f(t_i) \mid R_f(t_{i-1})\}] = \mathbb{E}[\mathbb{P}^2\{X(t_i) \leq f(t_i) \mid R_f(t_{i-1})\}] - \mathbb{E}^2[\mathbb{P}\{X(t_i) \leq f(t_i) \mid R_f(t_{i-1})\}].$$
We estimate the second term $\mathbb{E}^2[\mathbb{P}\{X(t_i) \leq f(t_i) \mid R_f(t_{i-1})\}]$ by $\hat{p}_f^2(t_i)$, and we show the first term can be simplified as follows:

$$
\begin{align*}
\mathbb{E}[\mathbb{P}^2\{X(t_i) \leq f(t_i) \mid R_f(t_{i-1})\}] &= \mathbb{P}\{R_f(t_{i-1}) = 0\}\mathbb{P}^2\{X(t_i) \leq f(t_i) \mid R_f(t_{i-1}) = 0\} \\
&\quad + \mathbb{P}\{R_f(t_{i-1}) = 1\}\mathbb{P}^2\{X(t_i) \leq f(t_i) \mid R_f(t_{i-1}) = 1\} \\
&= \mathbb{P}\{X(t_{i-1}) \leq f(t_{i-1})\}\mathbb{P}^2\{X(t_i) \leq f(t_i) \mid X(t_{i-1}) \leq f(t_{i-1})\} \\
&\quad + \mathbb{P}\{X(t_{i-1}) > f(t_{i-1})\}\mathbb{P}^2\{X(t_i) \leq f(t_i) \mid X(t_{i-1}) > f(t_{i-1})\}.
\end{align*}
$$

Let

$$
\hat{p}_f(t_i, t_{i-1}^-) = \#\{j : X_j(t_i) \leq f(t_i), X_j(t_{i-1}) \leq f(t_{i-1})\}/n,
$$
and

$$
\hat{p}_f(t_i, t_{i-1}^+) = \#\{j : X_j(t_i) \leq f(t_i), X_j(t_{i-1}) > f(t_{i-1})\}/n,
$$
we then estimate $\mathbb{E}[\mathbb{P}^2\{X(t_i) \leq f(t_i) \mid R_f(t_{i-1})\}]$ by

$$
\frac{\hat{p}_f^2(t_i, t_{i-1}^-)/\hat{p}_f(t_{i-1}) + \hat{p}_f^2(t_i, t_{i-1}^+)/\hat{p}_f(t_{i-1})}{\hat{p}_f(t_{i-1})}\left(1 - \hat{p}_f(t_{i-1})\right),
$$
when $p_f(t_{i-1}) \neq 1$. For $\hat{p}(t_{i-1}) = 1$, the second part vanishes. Then, the estimator for $S_f(t_i; t_{i-1})$ is

$$
\hat{S}_f(t_i; t_{i-1}) = \begin{cases} 
1 & \text{, } \hat{p}(t_i) = 1, \\
\frac{\hat{p}_f^2(t_i, t_{i-1}^-)/\hat{p}_f(t_{i-1}) - \hat{p}_f^2(t_i)}{\hat{p}_f(t_{i-1})\left(1 - \hat{p}_f(t_{i-1})\right)} & \text{, } \hat{p}(t_i) \neq 1, \hat{p}(t_{i-1}) = 1, \\
\frac{\hat{p}_f^2(t_i, t_{i-1}^-)}{\hat{p}_f(t_{i-1})} + \frac{\hat{p}_f^2(t_i, t_{i-1}^+)}{1 - \hat{p}_f(t_{i-1})} - \hat{p}_f^2(t_i) & \text{, } \hat{p}(t_i) \neq 1, \hat{p}(t_{i-1}) \neq 1.
\end{cases}
$$

For the weight function $v(t)$ in the shape similarity, which is the normalized dif-
ference in function values, the empirical estimator at time $t_i$ is

$$\hat{v}(t_i) = \frac{|f(t_i) - f(t_{i-1})|}{\sum_{r=2}^{m} |f(t_r) - f(t_{r-1})|}.$$  

Finally, the estimators of the TVD and SS are given by

$$\widehat{\text{TVD}}(f) = \sum_{i=1}^{m} \hat{D}_f(t_i)/m, \quad \widehat{\text{SS}}(f) = \sum_{i=2}^{m} \hat{v}(t_i)\hat{S}_f(t_i; t_{i-1}).$$