The dissertation of Gaurav Agarwal is approved by the examination committee

Committee Chairperson: Prof. Ying Sun
Committee Members: Prof. Hernando Ombao, Prof. Mark Tester, Prof. Xuming He
ABSTRACT

Quantile Function Modeling and Analysis for Multivariate Functional Data
Gaurav Agarwal

Quantile function modeling is a more robust, comprehensive, and flexible method of statistical analysis than the commonly used mean-based methods. More and more data are collected in the form of multivariate, functional, and multivariate functional data, for which many aspects of quantile analysis remain unexplored and challenging. This thesis presents a set of quantile analysis methods for multivariate data and multivariate functional data, with an emphasis on environmental applications, and consists of four significant contributions. Firstly, it proposes bivariate quantile analysis methods that can predict the joint distribution of bivariate response and improve on conventional univariate quantile regression. The proposed robust statistical techniques are applied to examine barley plants grown in saltwater and freshwater conditions providing interesting insights into barley’s responses, informing future crop decisions. Secondly, it proposes modeling and visualization of bivariate functional data to characterize the distribution and detect outliers. The proposed methods provide an informative visualization tool for bivariate functional data and can characterize non-Gaussian, skewed, and heavy-tailed distributions using directional quantile envelopes. The radiosonde wind data application illustrates our proposed quantile analysis methods for visualization, outlier detection, and prediction. However, the directional quantile envelopes are convex by definition. This feature is shared by most existing methods, which is not desirable in nonconvex and multimodal distributions. Thirdly, this challenge is addressed by modeling multivariate functional data for flexible quantile contour estimation and prediction. The estimated contours are flexible in
the sense that they can characterize non-Gaussian and nonconvex marginal distributions. The proposed multivariate quantile function enjoys the theoretical properties of monotonicity, uniqueness, and the consistency of its contours. The proposed methods are applied to air pollution data. Finally, we perform quantile spatial prediction for non-Gaussian spatial data, which often emerges in environmental applications. We introduce a copula-based multiple indicator kriging model, which makes no distributional assumptions on the marginal distribution, thus offers more flexibility. The method performs better than the commonly used variogram approach and Gaussian kriging for spatial prediction in simulations and application to precipitation data.
ACKNOWLEDGEMENTS

I am deeply grateful to my advisor, Prof. Ying Sun, for her unwavering support throughout my Ph.D. at King Abdullah University of Science and Technology (KAUST). She has played a major role in my time at the university, and her constant mentoring has shaped my Ph.D. and helped me to achieve my goals. She inspired me to become a better researcher with her ideas and guidance and develop my full potential.

I express my sincere gratitude to Prof. Hernando Ombao, Prof. Mark Tester, and Prof. Xuming He for being my committee members and for giving their valuable comments and insightful suggestions, which helped me see my research from a wider perspective. I am also thankful to other faculty members, Prof. Marc Genton, Prof. Håvard Rue, and Prof. Raphaël Huser for building my knowledge through courses at KAUST and helping me thrive in my scientific and programming skills. I am grateful to Prof. Linglong Kong and Prof. Judy Wang for their collaboration on exciting research projects.

I would like to extend my gratitude to KAUST for providing me with a life-changing experience through my Ph.D. I would like to appreciate the excellent resources provided by KAUST, which creates an ideal environment for conducting research. I want to thank my colleagues and friends for filling my Ph.D. life with warmth and joy. Finally, I would like to thank my family for being my pillars of strength throughout this wonderful journey.
TABLE OF CONTENTS

Examination Committee Page 2
Copyright 3
Abstract 4
Acknowledgements 6
Table of Contents 7
List of Figures 10
List of Tables 15

1 Introduction 16
  1.1 Background .................................................. 16
  1.2 Motivation .................................................. 18
  1.3 Contributions and outline of the thesis ...................... 20

2 Quantile Function Modeling with Application to Salinity Tolerance Analysis of Plant Data 23
  2.1 Introduction .................................................. 23
  2.2 Methods ..................................................... 26
    2.2.1 Plant Material .......................................... 26
    2.2.2 Field trial ............................................. 27
    2.2.3 Univariate quantile analyses .......................... 27
    2.2.4 Web application for univariate quantile analyses ....... 30
    2.2.5 Bivariate quantile analyses ........................... 31
  2.3 Results and discussion ..................................... 34
    2.3.1 Behaviour of traits in non-saline and saline conditions 34
    2.3.2 Traits affecting salinity tolerance indices ............. 37
    2.3.3 High salt tolerant and high yielding accessions ......... 39
    2.3.4 Bivariate prediction of yield for given traits .......... 41
## 3 Bivariate Functional Quantile Envelopes with Application to Radiosonde Wind Data

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.1 Introduction</td>
<td>47</td>
</tr>
<tr>
<td>3.2 Methodology</td>
<td>52</td>
</tr>
<tr>
<td>3.2.1 Directional quantile envelopes for bivariate functional data</td>
<td>52</td>
</tr>
<tr>
<td>3.2.2 Outlier detection and visualization</td>
<td>56</td>
</tr>
<tr>
<td>3.2.3 Quantile envelope prediction</td>
<td>64</td>
</tr>
<tr>
<td>3.3 Simulation study</td>
<td>66</td>
</tr>
<tr>
<td>3.3.1 Bivariate Gaussian random field</td>
<td>67</td>
</tr>
<tr>
<td>3.3.2 Simulation motivated by radiosonde data</td>
<td>69</td>
</tr>
<tr>
<td>3.4 Applications to radiosonde winds</td>
<td>71</td>
</tr>
<tr>
<td>3.4.1 Prediction of median</td>
<td>71</td>
</tr>
<tr>
<td>3.4.2 Predicting envelope for a given $\tau$</td>
<td>73</td>
</tr>
<tr>
<td>3.5 Discussion</td>
<td>74</td>
</tr>
</tbody>
</table>

## 4 Flexible Quantile Contours for Multivariate Functional Data: Beyond Convexity

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.1 Introduction</td>
<td>77</td>
</tr>
<tr>
<td>4.2 Methodology</td>
<td>81</td>
</tr>
<tr>
<td>4.2.1 Multivariate functional quantile model</td>
<td>81</td>
</tr>
<tr>
<td>4.2.2 Estimation of conditional multivariate functional quantiles</td>
<td>83</td>
</tr>
<tr>
<td>4.2.3 Quantile contour estimation and prediction</td>
<td>86</td>
</tr>
<tr>
<td>4.3 Theoretical properties</td>
<td>88</td>
</tr>
<tr>
<td>4.4 Simulation study</td>
<td>93</td>
</tr>
<tr>
<td>4.4.1 Bivariate functional data with convex margins</td>
<td>93</td>
</tr>
<tr>
<td>4.4.2 Bivariate functional data with nonconvex margins</td>
<td>96</td>
</tr>
<tr>
<td>4.4.3 Trivariate functional data with nonconvex margins</td>
<td>97</td>
</tr>
<tr>
<td>4.5 Applications to air pollution data</td>
<td>99</td>
</tr>
<tr>
<td>4.6 Discussion</td>
<td>102</td>
</tr>
</tbody>
</table>

## 5 Copula-based Multiple Indicator Kriging for non-Gaussian Random Fields

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>5.1 Introduction</td>
<td>103</td>
</tr>
<tr>
<td>5.2 Methodology</td>
<td>107</td>
</tr>
<tr>
<td>5.2.1 Copula-based multiple indicator kriging</td>
<td>108</td>
</tr>
</tbody>
</table>
5.2.2 Spatial modeling using copulas .......................... 110
5.2.3 Spatial probabilistic prediction .......................... 114
5.3 Simulation study ............................................. 118
5.3.1 Evaluation of point prediction ......................... 119
5.3.2 Evaluation of probabilistic prediction ................. 121
5.4 Applications ............................................... 124
5.4.1 Application to precipitation data ................... 124
5.4.2 Application to Meuse data ............................. 128
5.5 Discussion ................................................. 131

6 Concluding Remarks ........................................ 132
6.1 Summary ................................................ 132
6.2 Future research work .................................... 134

References .................................................. 136
LIST OF FIGURES

2.1 Directional quantile envelopes for bivariate distribution of yield (a) For fixed p=0.0125, the inner envelope of the directional quantile lines forms the directional quantile envelope (b) Directional quantile envelope for $p = 2^i/10$, $i = -5, \ldots, 2$. .................................................. 32

2.2 Relationship between plant traits and yield explained through quantile regression modeling: Each panel represents the estimated value of regression coefficient in black dots and the estimated 95% confidence interval in the grey area as a function of quantile level for respective covariate in the model with yield as a response, obtained by fitting a multiple quantile regression model (n=808). The horizontal black lines represent the zero value of the estimated coefficients. The variable condition is a dummy variable and was coded 1 for non-saline conditions and 0 for saline conditions. a*b represents the interaction between covariates a and b. ................................. 36

2.3 Relationship between plant traits in saline conditions and SWP index explained through quantile regression modeling: Each panel represents the estimated value of regression coefficient in black dots and the estimated 95% confidence interval in the grey area as a function of quantile level for each covariate in the model with SWP index as a response, obtained by fitting a multiple quantile regression model (n=404). The horizontal black lines represent the zero value of the estimated coefficients. ................................. 40
2.4 High salt tolerant and high yielding accessions obtained using conditional and marginal quantiles: The top-performing accessions with respect to high tolerance and yield are identified with filled green dots. The dashed blue and red lines represent the fitted lines of quantile regression for quantile levels 0.85 and 0.5, respectively, and the solid blue and red lines are marginal quantiles of level 0.9 and 0.5, respectively. The accessions above conditional quantile level 0.85, but below 0.9 marginal quantile levels are denoted by blue dots, and represent accessions with high salt tolerance but not necessarily with high yield.

2.5 Bivariate regression to predict the joint distribution of yield under non-saline and saline conditions for a given plant trait: The predicted quantile envelopes for three values of grain number per ear: 7, 11, 15 for $p = 0.1, 0.25, p^*$ forming the outer, inner and median envelope respectively. The different colors indicate different values of grain number per ear.

2.6 Relationship between plant traits and stress-weighted performance index (SWP) predicted using bivariate regression: Each of the nine panels represents a scatter plot between a plant trait and SWP. The solid black line represents the predicted values of the index using bivariate regression; the dashed black lines are the lower and upper confidence bands, formed using inner and outer envelopes, respectively, and superimposed is the median regression line in red.

3.1 Each of the eight panels shows a scatter plot of the $u$ and $v$ components of the radiosonde wind data and the estimated extreme quantile envelopes for $\tau = 0.0002$ at a given pressure level. The outliers are marked by red dots.

3.2 The solid contours are the estimated directional quantile envelopes for $\tau = 0.05, 0.0125$, and the dashed contours are the extreme quantile envelopes; the color variation represents the different pressure levels. The red dots are the outlying launches. The solid black line is the estimated functional median. (a) First time period 1962-1986, and (b) second time period 1987-2011.
3.3 The three-dimensional plot represents the $u$ component on the x-axis, the $v$ component on the y-axis, and the pressure level on the z-axis. The grey dots represent the observed $u$ and $v$ components of wind speeds from the radiosonde data, and the black contours represent the estimated directional quantile envelopes for $\tau = 0.05$ during the time period 1987-2011. The different colors of the connected directional quantile envelopes for $\tau = 0.05$ represent different directions.

3.4 Simulated curves from a bivariate Gaussian distribution with zero mean. The curve in red is the shape outlier; the curve in blue is the magnitude outlier.

3.5 Boxplots to detect functional outliers. (a) Magnitude outliers and (b) shape outliers.

3.6 Results of simulations from a bivariate Gaussian process. The four panels represent the models with no outliers, 1% outliers, 5% outliers, and 10% outliers. Each panel shows bar plots of MSPE for the three methods, conditional mean (blue), quantile kriging (green), quantile regression (red), and for three cases of range parameters, 0.1, 0.3, and 0.5.

3.7 The plots show the scatter between the $u$-component and $v$-component of the wind along with the estimated quantile envelope (blue) from the observed data and the predicted quantile envelopes from the method of kriging (green) and quantile regression (red). The left panel shows the plot for pressure level 200, and the right panel represents pressure level 300.

4.1 (a) Predicted contours for simulated data at $t = 0.48$ for bivariate functional data with convex margins are in solid black for quantile levels 0.25, 0.5, and 0.75, with the density contours of the true bivariate normal distribution overlayed with colored lines for quantiles level 0.25 (red), 0.5 (blue), 0.75 (green); and (b) predicted functional quantile curves for the 441 points of the reference distribution $U(0,1)^2$ are shown in grey, and the true mean function is overlayed in red; (c) predicted median curves to estimate the center of the distribution for 100 simulations are in black, and the true center is in red.
4.2 Predicted contours for bivariate functional data with convex margins at $t = 0.48$ using (a) multivariate functional quantile model (4.1) in the chapter for six quantile levels (0.1, 0.26, 0.42, 0.58, 0.74, 0.9) averaged over 100 simulations, and (b) Kong et al.’s method for six depth levels (0.01, 0.09, 0.17, 0.24, 0.32, 0.4).

4.3 Predicted contours for bivariate functional data with nonconvex margins at $t = 0.48$ using (a) multivariate functional quantile model (4.1) for six quantile levels (0.1, 0.26, 0.42, 0.58, 0.74, 0.9), and (b) Kong et al.’s method for six depth levels (0.01, 0.07, 0.13, 0.18, 0.24, 0.3).

4.4 Predicted functional quantile curves for bivariate functional data with nonconvex margins for the 441 points of the reference distribution $U(0,1)^2$ are shown in grey and the true mean function is overlayed in red.

4.5 Predicted quantiles shapes for trivariate functional data with nonconvex margins for six quantile levels (0.1, 0.26, 0.42, 0.58, 0.74, 0.9) at $t = 0.48$.

4.6 (a) Predicted quantile contours for six quantile levels, (0.1, 0.26, 0.42, 0.58, 0.74, 0.9), and (b) predicted functional quantile curves over time for the distribution of PM$_{2.5}$ and geopotential height at 850 hPa from October 2011 to March 2012.

4.7 (a) Predicted quantile contours for six quantile levels, (0.1, 0.26, 0.42, 0.58, 0.74, 0.9), and (b) predicted functional quantile curves over time for the distribution of PM$_{2.5}$ and geopotential height at 850 hPa from April 2012 to November 2012.

5.1 PIT histograms for the predictive distributions of GK, VMIK, and proposed CMIK$_G$ and CMIK$_t$ methods for three cases: skewed ($g = 0.75$, $h = 0$), heavy-tailed ($g = 0$, $h = 0.5$), and both skewed and heavy-tailed ($g = 0.5$, $h = 0.3$).

5.2 The daily average precipitation (mm) for November 2019 at 169 monitoring stations in Spain. (a) Map of the precipitation data; (b) histogram of the precipitation intensities.

5.3 PIT histograms for the predictive distributions of GK, logGK, VMIK, and proposed CMIK$_G$ and CMIK$_t$ methods for the precipitation data.
5.4 Results from the proposed CMIK$_t$ method for the precipitation data.
(a) Map of predicted values, (b) contour map for the probability of exceeding 50mm on average daily during November 2019 in Spain.

5.5 The zinc concentrations (in ppm) at 155 locations in the topsoil of a floodplain along the river Meuse. (a) Map of zinc concentrations; (b) histogram of zinc concentrations.

5.6 Results from the proposed CMIK$_t$ method for the Meuse data. (a) Map of predicted values of zinc concentrations, (b) contour map for the probability of exceeding 500 ppm zinc concentrations.
LIST OF TABLES

2.1 Results of mean regression between plant traits and yield. 37
2.2 Results of mean regression between plant traits in saline conditions and SWP index. 39
3.1 Mean square error from model motivated by real data. 71
3.2 MSPE obtained by comparing the center of the radiosonde wind data distribution using leave-one-out cross-validation over two time periods. 72
3.3 MSPE obtained by comparing envelopes at each direction of the two methods kriging and quantile regression from the observed envelope. 75
5.1 MAD and standard errors (in parentheses) of GK, VMIK, and proposed CMIK\(_C\) and CMIK\(_t\) methods for Tukey \(g\)-and-\(h\) random fields with exponential covariance function and range parameters \(\theta = 0.25, 0.375, 0.5\) with multiple \(g\) and \(h\) values. 120
5.2 Coverage percentages (CP) and average lengths (Length) of prediction confidence intervals. 124
5.3 CRPS and standard errors (in parentheses) of GK, VMIK, and proposed CMIK\(_C\) and CMIK\(_t\) methods with three cases: skewed \((g = 0.75, h = 0)\), heavy-tailed \((g = 0, h = 0.5)\), and both skewed and heavy-tailed \((g = 0.5, h = 0.3)\). 124
5.4 Numerical assessment of the predictive performance of GK, logGK, VMIK, and proposed CMIK\(_C\) and CMIK\(_t\) methods for the precipitation data. 127
5.5 Numerical assessment of the predictive performance of GK, logGK, VMIK, and proposed CMIK\(_C\) and CMIK\(_t\) methods for the Meuse data. 130
Chapter 1

Introduction

1.1 Background

Quantile regression models the entire distribution of response for given covariates and gives a comprehensive analysis of the effects of covariates on the different quantiles of the response. Quantile function modeling is a versatile, robust, and powerful tool for statistical analysis that provides valuable insights into applications of biology, ecology, meteorology, finance, business analytics, and environmental sciences. The concept of quantiles has been extended to multivariate settings by various researchers. It can be used to construct multivariate quantile regression models and establish quantile contours for multivariate data. The estimated quantile contours completely characterize the underlying distribution. However, little work has been done for studying quantiles for functional data. The modeling and estimation of quantiles for multivariate functional data is statistically challenging. The notion of quantiles for multivariate functional data is still an unsolved problem.

In the case of a univariate distribution, quantiles are well-defined because the ordering of observations follows the natural ordering in the real line $\mathbb{R}$. Hence, obtaining quantiles in the univariate case is straightforward. For a given real-valued random variable $Y$ with a distribution function $F$, the $p^\text{th}$ quantile is given by

$$Q(p) = F^{-1}(p) = \inf\{y : F(y) \geq p\} \quad \text{for} \quad 0 < p < 1.$$
distribution, there is no natural ordering of observations, and thus obtaining bivariate quantiles is statistically challenging (Chen and Welsh, 2002; Chaudhuri, 1996a; Serfling, 2002b). Many researchers have developed various ideas for ordering multivariate data analogous to the concepts behind univariate order statistics, and each has its advantages in specific applications. For example, Chaudhuri (1996b) extended notions of quantiles to multivariate data using the geometry based on Euclidean distances, Chakraborty (2003) defined regression quantiles for multivariate data, Kong and Mizera (2012) and Hallin et al. (2010) defined projection-based quantiles to get useful insights into multivariate data. Moreover, data depth, introduced by Tukey (1975), provides a way to order multivariate data. Tukey (1975) ordered multivariate data by defining a useful tool of the depth function, known as the Tukey depth, which provides a center-outward ranking of data points. These depth-based methods follow linear monotonicity axiom, which imposes the convexity of its contours (Liu, 1990; Zuo and Serfling, 2000). Most of the methods in the literature rely on ranking multivariate data using depth functions (Hettmansperger et al., 1992; Breckling et al., 2001; Serfling, 2010).

Data is often available in the form of random curves that may be observed as univariate or multivariate functional data in all areas of science, medicine, and engineering. Functional data analysis concerns data for which each point is a curve, surface, or anything else varying over a continuum. The area of functional data analysis has greatly attracted attention among researchers during the past two decades. Multivariate functional data can be viewed as realizations of multivariate random functions, such as patient’s CD4 cell count and viral load over therapy duration, and observed variables such as temperature, geopotential height, and PM2.5 over time. Fraiman and Pateiro-López (2012) defined projection-based quantiles for high dimensional multivariate data and univariate functional data. For multivariate functional data, Kong et al. (2015) modeled bivariate functional data using directional quantiles,
and the estimated quantile contours for the marginal distribution coincides with the Tukey depth contours. Computing quantiles is essential for data analysis; however, the extension of quantiles to multivariate functional data is quite challenging. In this thesis, I develop a set of robust quantile methods for the analysis of multivariate data and multivariate functional data.

1.2 Motivation

The study of salinity tolerance is vital to study the growth of plants. Salinity tolerance is the performance of plants in harsh saline water conditions compared to the freshwater (Munns and Tester, 2008). Salinity limits the growth of plants around the world due to poor irrigation conditions (Shrivastava and Kumar, 2015). Chapter 2 conducts a salinity tolerance analysis to evaluate the performance of plants using a set of quantile methods. Salinity tolerance indices evaluate plant performance under saline conditions relative to control conditions, so I identify the traits affecting the plants with high values of indices using quantile regression. However, these indices are univariate and result in the reduction of the dimensions of data. Consequently, they might not be sufficient to summarize the relationship between the indices and the covariates. I propose bivariate quantile analysis methods to model the joint distribution of yield from both saline and non-saline conditions to study the effect of plant traits on salinity tolerance.

Functional data are collected in numerous research areas, and there is an increasing need for techniques to analyze functional data. I extend the concepts of directional quantile envelopes of bivariate data to bivariate functional data in Chapter 3 with an application to radiosonde wind data. The directional quantile envelopes characterize the underlying distribution of data and give an idea of the data’s geometry. The global radiosonde archives contain valuable weather data, such as temperature, humidity, wind speed, wind direction, and atmospheric pressure. Being the only direct
measurement of these variables in the upper atmosphere, they are prone to errors. Therefore, the quality control of radiosonde data is essential, for which robust statistical methods are often needed. Among all the variables, the radiosonde winds, which consist of vertical and horizontal wind speeds, are particularly challenging to analyze. The radiosonde wind data is useful in studying the variations in climate and features of global warming. They are used as input in climate models and producing wind forecast errors.

The directional quantile envelopes are convex sets and follow the property of linear monotonicity. Most of the depth concepts defined for ordering multivariate data follow the property of linear monotonicity, which imposes the feature of convexity in the contours generated by them. Hence, these contours are not desirable for distributions with nonconvex margins. Chernozhukov et al. (2017) and Carlier et al. (2016) introduces concepts of multivariate quantiles based on optimal transport that are not restricted by convexity. However, multivariate data methods cannot be directly used for functional data because of the dimensionality issues. The quantile analysis gets more challenging as the data complexity increases. To address these challenges, I propose the estimation of flexible quantile contours for marginal distributions of multivariate functional data in Chapter 4.

Nonparametric methods based on quantiles are suitable for estimating non-Gaussian processes. In spatial statistics, to perform spatial prediction, kriging is the best predictor for Gaussian processes. However, the assumption of Gaussianity is quite strong and is rarely met in practice. Hence, kriging is not optimal for non-Gaussian spatial data, e.g., skewed or heavy-tailed. For non-Gaussian processes, I introduce a copula-based multiple indicator kriging model and propose a semiparametric estimator in Chapter 5. Multiple indicator kriging involves creating multiple sets of binary observations using indicator variables on the process and thresholding by different quantile levels. These multiple sets are used as co-variables for cokriging. The pro-
posed method models the marginal distribution nonparametrically, thus offers more flexibility.

1.3 Contributions and outline of the thesis

This thesis presents a set of robust statistical methods for analyzing several aspects of multivariate and multivariate functional data. Chapter 2 defines a quantile function modeling approach with application to salinity tolerance in plants using univariate and bivariate quantile analysis methods. I develop advanced quantile regression techniques to examine how different barley plant traits impact on yields grown in saline and non-saline conditions. I study the effects of plant traits on the yield of plants under fresh and salt water using quantile regression with covariates such as flowering time, ear number per plant, and grain number per ear. The use of univariate quantile analyses for quantifying yield under both conditions facilitates the identification of traits affecting salinity tolerance and is more informative than mean regression. I also propose a flexible method to identify the top-performing plants in terms of high yield in saline conditions and high salinity tolerance using marginal and conditional quantiles. The bivariate quantile analysis methods model the joint distribution of yield under both conditions for a given covariate and improve on the conventional univariate quantile regression.

Chapter 3 improves on the existing approaches for multivariate data to define methods for multivariate functional data. It defines directional quantile envelopes for bivariate functional data for visualization, outlier detection, and prediction of radiosonde wind data, which is extensively used in climate models and weather predictions. It studies the joint distribution of wind speed and direction along several pressure levels from a station in Denver, Colorado, for the period 1962-2011. The wind profiles of radiosonde data are treated as bivariate functional data across several pressure levels. Since the bivariate distribution of the wind speeds at a given
pressure level is not Gaussian but instead skewed and heavy-tailed, I propose a set of quantile methods based on directional quantile envelopes to characterize the distribution as well as an outlier detection procedure to identify both magnitude and shape outliers. I also propose two methods to predict directional quantile envelopes at unobserved pressure levels using quantile regression and kriging. The proposed methods provide an informative visualization tool for multivariate functional data. The proposed methods are shown to be robust against outliers, skewed and heavy-tailed data using simulation studies.

Chapter 4 introduces multivariate quantiles and flexible quantile contour estimation for multivariate functional data, which can characterize non-Gaussian and nonconvex distributions. It proposes a multivariate functional quantile model, which is a nonparametric, time-varying coefficient model, and it uses basis functions for the estimation and prediction. The functional quantile curves are estimated over time. Computationally, the proposed method is efficient for high dimensions and can handle more than just bivariate functional data. The monotonicity, uniqueness, and consistency of the estimated multivariate quantile function have been established. Consequently, the consistency of the proposed quantile contours is also proved. The proposed method was demonstrated on bivariate and trivariate functional data in the simulation studies, where the estimated contours correctly pick up the shape of the distribution, while the existing depth-based methods failed to do so. The proposed method was applied to air pollution data to study the joint distribution of PM$_{2.5}$ and geopotential height over time in the Northeastern United States. The estimated contours highlight the nonconvex joint distribution, and the functional quantile curves capture the dynamic change across time.

Chapter 5 proposes a quantile spatial prediction method for non-Gaussian processes by introducing a copula-based multiple indicator kriging model. I propose to describe the spatial dependence using copulas to overcome the challenges of the exist-
ing variogram-based approach of multiple indicator kriging by deriving the relationship between indicator covariances and copula functions. The method also provides a convenient way to make both point predictions and interval predictions by predicting the local cumulative distribution function or quantile function at new locations. The predictive performance of the proposed method has been demonstrated using extensive simulation studies, which performs better than the commonly used variogram approach and Gaussian kriging for spatial prediction. The proposed methods are illustrated on Spain’s precipitation data during November 2019, and heavy metal dataset in topsoil along the river Meuse, and obtain probability exceedance maps.
Chapter 2

Quantile Function Modeling with Application to Salinity Tolerance Analysis of Plant Data

2.1 Introduction

Soil salinity is a major abiotic stress that negatively impacts agriculture, as plants grown under saline conditions grow more slowly and have lower yields than plants grown under non-saline conditions (Shrivastava and Kumar, 2015). Therefore, understanding the mechanisms of salinity tolerance in plants is important to improve plant growth and productivity. Plants are able to maintain growth in saline conditions relative to non-saline conditions using a range of mechanisms, where a range of traits can contribute to this maintenance of growth and yield. Munns and Tester (2008) suggested three main traits contributed to salinity tolerance: exclusion of toxic salts from the shoot, tolerance of toxic salts in the shoot that were not excluded from the shoot, and tolerance processes that were independent of shoot salt effects. These considerations have been developed further by Morton et al. (2019) to include a wider range of other physiological traits, focusing in particular on the ability of plants to maintain processes in saline conditions relative to non-saline conditions. The technical approaches that can be taken to measure these traits are detailed in Negrão et al. (2017).

To study salinity tolerance, a typical way is to define salinity tolerance indices, which measure the plant performance in saline conditions relative to non-saline conditions (Munns et al., 2002; Fernandez, 1993; Asadi et al., 2012). These indices are
univariate and result in the reduction of the dimensions of data. As a consequence, a single index might not be sufficient to summarize the relationship between the indices and the covariates. In this chapter, we apply a set of quantile analysis methods and demonstrate the necessity of these methods by studying the dependence of plant traits on salinity tolerance of barley accessions. A conventional statistics tool used to investigate the relationship between a response variable and covariates is the mean regression (Talei et al., 2013; Meng et al., 2017). Mean regression only provides an incomplete picture of the response distribution corresponding to the covariates, just as the mean does by providing an incomplete summary of a single distribution, and not accounting for extreme values in the data (Mosteller and Tukey, 1977; Lè Cook and Manning, 2013).

Quantiles are the values that divide the entire distribution such that a given proportion of values, say $p$, lie below the $p^{th}$ quantile, where $p \in (0, 1)$ (Bland, 2015). For example, median is the 0.5$^{th}$ quantile. The data can be divided into different quantiles, and we can check how the data is behaving for each quantile. On the other hand, mean provides a grand summary of the distribution by computing its average, hence losing information. Mean regression models the average of the distribution of the response variable for given covariates, assuming that the variables behave similarly at the upper and lower tails of the distribution as well as the mean. On the other hand, quantile regression models the entire distribution of the response, given the covariates, and provides a more comprehensive analysis of the effect of the predictors on the response (Mosteller and Tukey, 1977; Young et al., 2008). Quantile regressions are particularly valuable in applications where extremes are imperative, such as agricultural studies for which higher quantiles of yield are critical (Chen, 2005). The regression that involves modeling the conditional mean of the response distribution might obscure the effect of a trait on the tails of the response, whereas quantile regression can reveal those effects. For instance, one particular trait may have
a negligible effect on conditional means but may lower conditional 10\textsuperscript{th} percentiles sharply (Burgette et al., 2011).

Quantile regression has drawn considerable research interest in recent years and is being applied in various fields. Quantile regression is becoming adapted in ecology and environmental sciences (Knight and Ackerly, 2002; Cade and Noon, 2003; Cade, 2017). For instance, in some ecological applications, the approach of quantile regression was used to estimate the upper quantiles of the growth rates of marine phytoplankton as a function of temperature (Bissinger et al., 2008) and to reveal the uncertainty in the relationship between an organism and its habitat at different quantile levels (Cade et al., 2005). It has been used in biology to test the significance of dissolved oxygen concentration at the upper quantiles of body size of deep-sea organisms (McClain and Rex, 2001; Anderson, 2008). Quantile regression has long been used in other disciplines, like business and economic analysis (Dimelis and Louri, 2002; Girma and Görg, 2005; Hung et al., 2010; Chunying, 2011). Methods based on quantile regression have been used in health and medicine and demonstrated how richer inferences could be drawn using quantile regression (Austin et al., 2005; Wei et al., 2006). In this chapter, we are suggesting to extend the application of quantile regression techniques to the field of agriculture and salinity tolerance.

In the case of a univariate distribution, the natural ordering of a variable is the order on real line $\mathbb{R}$. Hence, obtaining quantiles, in that case, is straightforward. However, for a bivariate distribution, there is no natural ordering of observations, and thus obtaining bivariate quantiles is statistically challenging (Chen and Welsh, 2002; Chaudhuri, 1996a; Serfling, 2002b), since we need to consider not only values but also directions. Kong and Mizera (Kong and Mizera, 2012) proposed directional quantiles and directional quantile envelopes to characterize multivariate distribution. Using directional quantile envelopes, we propose a bivariate quantile regression model to predict the behavior of the bivariate response variable jointly for given covariates.
To illustrate the methods we used, we perform a salinity tolerance analysis to evaluate the performance of barley accessions. We focus on the upper tails of the response distribution, as the accessions that are highly salt-tolerant and have a high yield in non-saline conditions are of primary interest. We perform a quantile regression analysis using plant agronomic traits and a salinity tolerance index to identify the traits that affect the accessions with high indices. We also propose a flexible approach to identify accessions with high salinity tolerance along with high yield using conditional and marginal quantiles. We predict the bivariate distribution of plant yield under two different conditions (non-saline and saline), for a given plant trait. Hence, we directly associate salinity tolerance indices with the plant traits to get a detailed analytic understanding of the effects of plant traits on salinity tolerance. The dataset presented in this chapter is used to provide an example of how the quantile analysis methods can be applied to the field of agriculture and salinity tolerance.

The remainder of the chapter is organized as follows. Section 2.2 describes the data obtained from the experiments. It also presents our methodology on univariate and bivariate quantile analyses for analyzing the salinity tolerance of barley accessions. In Section 2.3, we present the results and discuss the effects of plant traits on salinity tolerance. Finally, in Section 2.4, we conclude our findings, which may impact our understanding of salinity tolerance.

2.2 Methods

2.2.1 Plant Material

The plant material consisted of 404 barley accessions from a barley association mapping (AM) population provided by Prof. Robbie Waugh from the James Hutton Institute, United Kingdom. All accessions were 2-row spring barley cultivars.
2.2.2 Field trial

Plants from the AM population were grown at the International Center for Biosaline Agriculture (ICBA, Dubai), over a year, from 2013 to 2014. Plots were irrigated with fresh (1 dS/m; referred to as ‘non-saline’) and saline water (17 dS/m; referred to as ‘saline’). An augmented design was used where salt tolerant check lines (116/2A, 58/1A, and CM72) were added every seven plots on average. Detailed descriptions of the field design and practice are provided in Saade et al. (2016), who grew and studied another population, HEB-25. The following plant traits were recorded under both conditions: flowering time (days), ripening period (days), plant height (cm), ear number per plant, grain number per ear, thousand grain mass (g), dry mass per m$^2$ (g/m$^2$), grain mass per m$^2$ (g/m$^2$, referred to as yield), and harvest index. A detailed description of each trait and how it was measured is provided in Saade et al. (2016). The experiment that generated the raw phenotypic data used in this chapter was originally designed for an association mapping analysis of salinity tolerance in barley.

2.2.3 Univariate quantile analyses

As defined in the book “Quantile Regression” by Koenker (2005), for a given real-valued random variable $Y$ with a distribution function $F$, the $p^{th}$ quantile is given by

$$Q(p) = F^{-1}(p) = \inf\{y : F(y) \geq p\} \quad \text{for} \quad 0 < p < 1.$$ 

If we denote the $p^{th}$ conditional quantile function as $Q_y(p \mid x) = x^T \beta(p)$, the optimization problem of quantile regression can be formulated as

$$\min_{\beta \in \mathbb{R}^p} \sum_{i=1}^{n} \rho_p(y_i - x_i^T \beta),$$

where $\rho_p(u) = u(p - \mathbb{1}(u < 0))$ is the loss function, and $\mathbb{1}(\cdot)$ is an indicator function. The $y_i$’s represent the realizations of the response variable; $x$ is the design matrix with
the first column as the unit vector, and the rest of the columns represent the values of the covariates; $\beta$ is the vector of regression coefficients. Regression coefficients of a quantile regression model are estimated by minimizing the loss function $\rho_p(u)$. We include the saline and non-saline groups in a common model by including a categorical covariate for that condition classification and adding its interactions with all the other covariates. This allows the analysis to not only test and estimate the effects of covariates for the saline and non-saline groups separately, but also provide the possibility of testing and estimating the differences between the estimated effects of each covariate for the non-saline and saline groups. The categorical variable for the classification of non-saline and saline conditions was coded as a dummy variable. The model used for univariate quantile regression is given by

$$y_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} + \alpha_0 D_i + \sum_{j=1}^{p} \alpha_j D_i x_{ij} + \epsilon_i, \quad i = 1, \ldots, n,$$

where $y$ is the response, $\beta_0$, $\alpha_0$, $\beta_j$ and $\alpha_j$, $j = 1, \ldots, p$ are regression coefficients, $x_j$, $j = 1, \ldots, p$ are covariates, $D$ is a dummy variable: $D_i = 1$ if $i \in$ non-saline group and $D_i = 0$ if $i \in$ saline group and $\epsilon$ is random error. Here, $Dx_j$ denotes the interaction terms.

We eliminate the variables harvest index, thousand grain mass (g), and dry mass per m$^2$ from our multiple regression model as they are partial expressions of yield, the response variable, so it might not be useful to study their effects, and they could cause the problem of multicollinearity. After dropping these variables, the variance inflation factors (VIFs) for all the plant traits were less than 2.5, so we consider all other plant traits as covariates in the model. The sample size for the model was $n = 808$ (404 for saline and 404 for non-saline condition). We do not scale the covariates to unit variance as we notice no advantage gained by scaling, and indeed, estimated effects are far more interpretable in their original units. We center the covariates just so that
their mean is 0, and thus the intercept represents the response (yield) at the mean of all predictors. The model is fit using the rq() function of the quantreg package in R for quantile levels ranging from 0.1 to 0.9. Although we are interested in studying the accessions with high yield, we investigate the model for the whole range of quantiles, which allows us to check for the stability in the coefficient value and examine the change in slope as we move from lower quantile to upper quantiles. After fitting the quantile regression model, we plot the estimated values of coefficients and the estimated 95% confidence intervals of the plant traits as a function of the quantile level to examine the relationship between the plant trait and different quantiles of yield. The upper and lower bounds for the estimated quantile regression coefficients are calculated using the rankscore test inversion (Koenker, 1994). This method is suitable in case of small sample sizes (less than 1000). The assumption of independent and identically distributed errors is also relaxed (Koenker and Machado, 1999). The test of significance is determined using the confidence intervals produced by the rank inversion method. If the estimated confidence interval around the observed effect includes 0, then the effect is not statistically significant.

By plotting the estimated regression coefficients along with the estimated confidence interval against the quantile level, we were able to give a complete picture of the relationship between traits and response distribution in both the non-saline and saline conditions separately and also on the differences between the two conditions.

We also performed a quantile regression analysis on salinity tolerance index SWP (stress-weighted performance, Saade et al. (2016)) for different quantile levels where the upper tails of response distribution were of principal interest. SWP is defined as

\[ \text{SWP} = \frac{y_s}{\sqrt{y_c}}, \]

where \( y_s \) denotes the yield under saline conditions and \( y_c \) yield under non-saline
conditions. The salinity tolerance index SWP was set as the response, and the plant traits from saline conditions were set as the covariates and were centered to mean 0. The model for univariate quantile regression for salinity tolerance index is given by

\[ y_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} + \epsilon_i, \quad i = 1, \ldots, n, \]

where \( y \) is the response, \( \beta_0, \beta_j, j = 1, \ldots, p \) are regression coefficients, \( x_j, j = 1, \ldots, p \) are covariates and \( \epsilon \) is the random error. The standard errors were computed using the rankscore inversion test without the assumption of independent and identically distributed errors.

To identify high salt tolerant and high yielding accessions, we make use of conditional and marginal quantiles. We regress yield under non-saline conditions on yield under saline conditions using the model

\[ y_i = \alpha x_i + \epsilon_i, \quad i = 1, \ldots, n, \]

where \( y \) is the response, \( \alpha \) is the regression coefficient, \( x \) is the covariate, and \( \epsilon \) is random error. We fit a univariate quantile regression model to obtain the conditional quantile function \( Q_{y|x}(p_1) \) for quantile level \( p_1 \). We also obtain the marginal quantile function \( Q_y(p_2) \) for quantile level \( p_2 \). The accession lying above the fitted conditional quantile line, i.e., the accessions with positive residuals are highly salt tolerant, while the accessions falling above the marginal quantile of \( y \), will have a high yield in saline conditions, for chosen quantile levels \( p_1 \) and \( p_2 \). By taking the intersection of the two methods, we can identify highly salt tolerant and high yielding accessions.

### 2.2.4 Web application for univariate quantile analyses

The method of univariate quantile regression analysis was implemented in a broader framework, in an open-source online application called MVApp (Julkowska et al.,
2019). The application was built using the Shiny framework of R. This method is available online at mvapp.kaust.edu.sa and is freely and easily accessible. Users can upload their data on the application and choose their response, covariates, treatment, and how they want to subset their data. The results of the analysis can be downloaded as a summary table and as plots.

2.2.5 Bivariate quantile analyses

The goal of the bivariate quantile analysis method is to predict the bivariate response, for a given covariate, by predicting directional quantile envelopes for the bivariate distribution. The notion of a directional quantile envelope was proposed by Kong and Mizera (2012) in 2012. In their approach, they project the bivariate distribution to univariate distribution along a direction \( s \), and obtain the quantiles of the projected distribution, calling them directional quantiles. Consider a normalized direction \( s \), on the unit circle \( S \), the \( p \)th directional quantile of the random vector \( Y \), in direction \( s \), is defined by

\[
Q(p, s) = \inf \{ y : F(s^T y) \geq p \}.
\]

For \( p \in (0, 1/2] \), the \( p \)th directional quantile line is given by the equation \( s^T y = Q(p, s) \) which indicates how directional quantiles divide the data. The \( p \)th directional quantile envelope produced by \( Q(p, s) \) is defined as the intersection,

\[
D(p) = \bigcap_{s \in S} H(s, Q(p, s)),
\]

where \( H(s, q) = \{ y : s^T y \geq q \} \) is the supporting halfspace. These envelopes are closely related to the Tukey depth contours proposed by Tukey in 1975 (Tukey, 1975). They are essentially Tukey depth level sets. The Tukey depth contours completely characterize the empirical distribution, for any multivariate dataset (Struyf and Rousseeuw, 1999).
Figure 2.1: Directional quantile envelopes for bivariate distribution of yield (a) For fixed $p=0.0125$, the inner envelope of the directional quantile lines forms the directional quantile envelope (b) Directional quantile envelope for $p = 2^i/10$, $i = -5, \ldots, 2$.

The construction of the directional quantile envelope is explained in Figure 2.1a. For a fixed $p=0.0125$, directional quantile lines are plotted for 100 directions along a unit circle, with the lines contributing to the construction of the inner envelope as active lines in black color, and the rest of the directional quantile lines in grey color. The polygon formed from the vertices of the intersecting active lines is the directional quantile envelope. In Figure 2.1b, the directional quantile envelopes are shown for several $p = 2^i/10$, $i = -5, \ldots, 2$, which are the estimated contours of the bivariate data.

The directional quantile envelopes for bivariate data are non-empty for $p \leq 1/3$, because of a result known as the centerpoint theorem (Mizera, 2002). The points corresponding to the highest depth are the deepest (Tukey, 1975; Zuo and Serfling,
2000). We obtain the highest value of $p \in (0.33, 0.5)$ for which the envelope is non-empty (for every case) and denote it by $p^*$; we call the envelope corresponding to $p^*$ as the median envelope since the envelope corresponding to the highest $p$ will have the highest depth. We take the average of the vertices of the median envelope to obtain the bivariate median. We then choose two values, 0.1 and 0.25, and call the envelopes corresponding to these values the outer and inner envelopes, respectively.

To construct the $p^{th}$ directional quantile envelopes for a given covariate, we need to obtain the $p^{th}$ directional quantile for the given value of covariate in a subset of all the directions along a unit circle. For each direction $s$, we model the projected distribution $s^Ty$ using a cubic spline function of the given plant trait. Let $y_{si} = s^Ty_i, i = 1, \ldots, n$, we fit the following quantile regression model for quantile $p$

$$y_{si} = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3 + \sum_{j=1}^{K} \delta_j (x_i - k_j)_+^3 + \epsilon_i, \quad i = 1, \ldots, n,$$

where $y$ is the response, $x$ is the covariate, $\beta_0, \beta_1, \beta_2, \beta_3$, and $\{\delta_j\}_{j=1}^{K}$ are regression coefficients, and $\{k_j\}_{j=1}^{K}$ are prespecified set of knots. The number of knots were fixed to 3 and the knots are typically chosen as suitable quantile of $x$. From the fitted model, we obtain the $p^{th}$ directional quantile for a given value of covariate $x$, in direction $s$, denoted by $\hat{Q}(p, s)$. Then the predicted $p^{th}$ directional quantile envelope for a given value of covariate $x$, produced by $\hat{Q}(p, s)$ is defined as the intersection,

$$\hat{D}(p) = \bigcap_{s \in S} H(s, \hat{Q}(p, s)).$$

Hence, we predict the bivariate distribution of yield by predicting the bivariate median for a given covariate. The salinity tolerance indices are functions of yield under both conditions. So the predicted estimates of yield under both conditions from bivariate quantile regression were used to compute the salinity tolerance index SWP. After obtaining the estimates of yield in non-saline and saline conditions, we estimated
SWP as
\[ \text{SWP} = \frac{\hat{y}_s}{\sqrt{\hat{y}_c}}, \]
where \((\hat{y}_s, \hat{y}_c)\) are bivariate quantile regression estimates of yield under saline and non-saline conditions. Hence, for a given plant trait, we obtain SWP, which is capable of identifying top-performing accessions in terms of high yield and high salinity tolerance (Saade et al., 2016), together with its upper and lower bounds obtained from the predicted outer and inner envelopes.

2.3 Results and discussion

2.3.1 Behaviour of traits in non-saline and saline conditions

In this section, we perform a univariate quantile regression under both non-saline and saline conditions using a common model, for different quantile levels, to observe the behavior of the traits on the complete distribution of the response yield. The plot of the results of the fitted quantile regression model is shown in Figure 2.2. Since the categorical variable condition was coded as 0 for saline conditions and 1 for non-saline conditions, the individual estimated effects represent the results for saline conditions, and the interaction terms represent the difference between the estimated effects of each covariate for the accessions with non-saline and saline conditions. We can observe that condition is positively significant for all quantile levels since the estimated confidence interval does not include the horizontal line for zero value of the estimated coefficient (Fig. 2.2g), which means that for an average value of plant traits, the yield in non-saline conditions is significantly greater than the yield in saline conditions. The change in slope at higher quantiles means that the difference is higher for accessions with higher yields.

Besides, the difference in yield under the two conditions can be attributed to the height of accessions, since the interaction term of plant height is significant at
some quantiles (Fig. 2.2h). As the plant height increases, the yield in non-saline conditions decreases, while plant height does not significantly affect the yield in saline conditions. The ripening period positively affects yield in saline conditions for low and mid quantiles but negatively affects yield for higher quantiles (Fig. 2.2d).

We also found that the flowering time seems to have a negative impact on yield in saline conditions, for accessions with mid-level and high yield (Fig. 2.2c). This effect is more substantial for accessions with a high yield than mid-level yield, which can be seen from the slope change, while this observation is not significant for accessions with a low yield. Based on the differences, the negative effects of flowering time at mid and high quantiles are similar in non-saline conditions but have stronger negative effects on yield at lower quantiles compared to saline conditions (Fig. 2.2i). Ear number per plant (Fig. 2.2e), and grain number per ear (Fig. 2.2f) have a significant positive impact on yield in saline conditions. The impact of grain number per ear on yield under saline conditions is more substantial for accessions with high yield. The interaction terms for variables except for plant height and flowering time, are not significant, so there is no significant difference in the estimated effects of ripening period (Fig. 2.2j), ear number per plant (Fig. 2.2k), and grain number per ear (Fig. 2.2l) on yield between non-saline and saline conditions.

The results of a similar framework using mean regression for the same model is shown in Table 2.1. From these results, we can merely comment that, on average, accessions with late flowering time have a lower yield. Quantile regression reveals that this effect is not significant for accessions with low yield (lower quantiles). In addition, a later flowering time affects accessions with high yield more than it does for the accessions with an average yield.

Mean regression also shows that, with an increase in grain number per ear, on average, the yield in saline conditions increases by 6.7 g/m². While with quantile regression, we can observe for accessions with high yield, the increase is nearly 10 g/m².
Figure 2.2: Relationship between plant traits and yield explained through quantile regression modeling: Each panel represents the estimated value of regression coefficient in black dots and the estimated 95% confidence interval in the grey area as a function of quantile level for respective covariate in the model with yield as a response, obtained by fitting a multiple quantile regression model (n=808). The horizontal black lines represent the zero value of the estimated coefficients. The variable condition is a dummy variable and was coded 1 for non-saline conditions and 0 for saline conditions. a*b represents the interaction between covariates a and b.

Therefore, mean regression provides limited opportunity for studying the accessions with extreme yields, which are of utmost agronomic importance, while quantile regression allows us to fine-tune the relationship between a trait and yield at different quantiles.
Table 2.1: Results of mean regression between plant traits and yield.

<table>
<thead>
<tr>
<th>Traits</th>
<th>Coefficient value</th>
<th>Standard error</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>109.9237</td>
<td>3.5151</td>
<td>2e-16 *</td>
</tr>
<tr>
<td>Plant height (cm)</td>
<td>0.3526</td>
<td>0.2329</td>
<td>0.130530</td>
</tr>
<tr>
<td>Flowering time (days)</td>
<td>-1.1282</td>
<td>0.3195</td>
<td>0.000438*</td>
</tr>
<tr>
<td>Ripening period (days)</td>
<td>-0.8952</td>
<td>2.5387</td>
<td>0.724464</td>
</tr>
<tr>
<td>Ear number per plant</td>
<td>10.3050</td>
<td>3.2122</td>
<td>0.001390*</td>
</tr>
<tr>
<td>Grain number per ear</td>
<td>6.7732</td>
<td>0.7903</td>
<td>2e-16 *</td>
</tr>
<tr>
<td>Condition</td>
<td>27.7525</td>
<td>4.2314</td>
<td>9.77e-11*</td>
</tr>
<tr>
<td>Condition × Plant height (cm)</td>
<td>-0.6567</td>
<td>0.2811</td>
<td>0.019716*</td>
</tr>
<tr>
<td>Condition × Flowering time (days)</td>
<td>-0.2733</td>
<td>0.4232</td>
<td>0.518699</td>
</tr>
<tr>
<td>Condition × Ripening period (days)</td>
<td>1.2312</td>
<td>2.9102</td>
<td>0.672357</td>
</tr>
<tr>
<td>Condition × Ear number per plant</td>
<td>-0.5443</td>
<td>3.6711</td>
<td>0.882178</td>
</tr>
<tr>
<td>Condition × Grain number per ear</td>
<td>0.9860</td>
<td>1.0241</td>
<td>0.335920</td>
</tr>
</tbody>
</table>

The three columns represent the estimated coefficient value, standard error and the p-value for the respective covariate obtained by fitting a multiple mean regression model with sample size 808. The significant p-values are marked with a ‘*’, for a significance level of 0.05. The variable condition is a dummy variable and was coded 1 for non-saline conditions and 0 for saline conditions. a × b represents the interaction between a and b.

This univariate quantile analysis is a simple approach that gives a thorough visualization of how the plant traits affect the complete distribution of yield for both conditions and also the difference in the estimated effects between non-saline and saline conditions. A limitation of this study is that the standard errors of the estimates were high due to small sample sizes, and some of the variables might be nonsignificant due to the high uncertainty of the estimated values.

2.3.2 Traits affecting salinity tolerance indices

To study the characteristics of a specific set of response variables that may be important in the context of salinity tolerance, we investigate the tail behavior of the response using quantile regression. One of our goals is to examine the accessions with a high salinity tolerance index. Several salinity tolerance indices have been pre-
viously proposed to identify stress tolerant and high-yielding accessions (Fernandez, 1993; Munns et al., 2002; Saade et al., 2016). Using our dataset, we compute SWP using yield under saline and non-saline conditions and conduct a trait analysis to assess the significance of traits affecting the salinity tolerance. Here, we consider the effect of the traits under saline conditions on SWP, which is considered as the response.

Saade et al. (2016) shows how SWP is better than other salinity tolerance indices (S/C and STI) in selecting accessions that are salt tolerant and have a high yield. We compute SWP using yield under saline and non-saline conditions and perform a quantile regression on SWP using plant traits under saline conditions (n= 404); we then check for the significance of plant traits that affect the salinity tolerance of plants. The results for the quantile regression model on SWP are shown in Figure 2.3.

We observed that late flowering time has a negative impact on salinity tolerance, with the accessions with high SWP being affected the most, and it has no significant impact on accessions with low SWP (Fig. 2.3c). This could be explained by the fact that plants that flower later are more exposed to the heat, and plants with low SWP are already struggling with salt stress. Quantile regression helped us observe that the ripening period is not significant for accessions with median SWP, but it is negatively significant for accessions with high SWP (Fig. 2.3d). It was also observed that ear number per plant (Fig. 2.3e) and grain number per ear (Fig. 2.3f) have a significant positive impact on salinity tolerance index SWP. The effect of grain number per ear is more substantial for accessions with high SWP than with median SWP.

Since SWP is used to differentiate the top-performing accessions from the other accessions based on the order of their values, accessions with high values of the index are of more practical importance. Using the quantile analyses, we study the effects of plant traits on accessions with high salinity tolerance, whereas, from the results of mean regression, we can only comment on accessions with average salinity tolerance,
Table 2.2: Results of mean regression between plant traits in saline conditions and SWP index.

<table>
<thead>
<tr>
<th>Traits in saline conditions</th>
<th>Coefficient value</th>
<th>Standard error</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>7.506834</td>
<td>0.054184</td>
<td>2e-16 *</td>
</tr>
<tr>
<td>Plant height (cm)</td>
<td>0.014669</td>
<td>0.009262</td>
<td>0.114</td>
</tr>
<tr>
<td>Flowering time (days)</td>
<td>-0.050227</td>
<td>0.012705</td>
<td>9.12e-05 *</td>
</tr>
<tr>
<td>Ripening period (days)</td>
<td>-0.065254</td>
<td>0.100945</td>
<td>0.518</td>
</tr>
<tr>
<td>Ear number per plant</td>
<td>0.588800</td>
<td>0.127725</td>
<td>5.44e-06 *</td>
</tr>
<tr>
<td>Grain number per ear</td>
<td>0.291359</td>
<td>0.031424</td>
<td>2e-16 *</td>
</tr>
</tbody>
</table>

The three columns represent the estimated coefficient value, standard error, and the p-value for the respective covariate in the model with SWP as the response, obtained by fitting a multiple mean regression model with sample size 404. The significant p-values are marked with a ‘*’, for a significance level of 0.05.

and therefore do not have any information on accessions with different ranges of salinity tolerance. The results of the mean regression for salinity tolerance indices are shown in Table 2.2. It shows that, on average, the ripening period does not have a significant effect on SWP, but quantile regression revealed that the ripening period is significant for the high quantiles of SWP.

2.3.3 High salt tolerant and high yielding accessions

Saade et al. (2016) showed how SWP outperforms STI in terms of selecting salt-tolerant accessions and how it chooses accessions with higher marginal yield than those chosen by S/C. Here, we propose a flexible approach to classify the observations of a bivariate distribution, using conditional and marginal quantiles. The conditional and marginal quantile levels can be modified as a trade-off between high salt tolerance of accession and high yield. The observations are classified using the intersection of the fitted univariate quantile regression line and marginal quantile line. We consider the distribution of yield under saline and non-saline conditions to illustrate the use of this method. Here, we are interested in accessions with a high stress tolerance together with high yield under non-saline conditions. Accessions lying above the fitted
Figure 2.3: Relationship between plant traits in saline conditions and SWP index explained through quantile regression modeling: Each panel represents the estimated value of regression coefficient in black dots and the estimated 95% confidence interval in the grey area as a function of quantile level for each covariate in the model with SWP index as a response, obtained by fitting a multiple quantile regression model (n=404). The horizontal black lines represent the zero value of the estimated coefficients.

conditional quantile line, for yield under saline conditions conditioned on non-saline conditions, are salt tolerant, while those lying above the marginal quantile of yield under saline conditions have a high yield. We take the intersection of both conditions and obtain the top-performing accessions. This method was applied to yield under
Figure 2.4: High salt tolerant and high yielding accessions obtained using conditional and marginal quantiles: The top-performing accessions with respect to high tolerance and yield are identified with filled green dots. The dashed blue and red lines represent the fitted lines of quantile regression for quantile levels 0.85 and 0.5, respectively, and the solid blue and red lines are marginal quantiles of level 0.9 and 0.5, respectively. The accessions above conditional quantile level 0.85, but below 0.9 marginal quantile levels are denoted by blue dots, and represent accessions with high salt tolerance but not necessarily with high yield.

The salinity tolerance indices depend on yield from both saline and non-saline conditions of barley. The best performing accessions in terms of both salt tolerance and high yield are identified with green circles lying above the 85th conditional and 90th marginal quantiles (Figure 2.4).

2.3.4 Bivariate prediction of yield for given traits

The salinity tolerance indices depend on yield from both saline and non-saline conditions, so to link an index to a plant trait, we need to model the joint distribution of yield for that plant trait. We applied the method of directional quantiles (Kong
and Mizera, 2012) to estimate the empirical distribution of our bivariate data, non-parametrically. Using the directional quantile envelopes (Kong and Mizera, 2012), we defined a way to predict the bivariate vector of yield for a given covariate. We linked the covariates of the yield under saline conditions with the bivariate data. We predicted these envelopes for a given value of the plant trait. Three $p^{th}$ directional quantile envelopes (also known as depth contours) were predicted, corresponding to $p = p^*, 0.25, 0.1$; they were called the median, inner and outer envelope respectively, with $p^*$ being the highest quantile value obtained for a non-empty quantile envelope in $p \in (0, 1/2]$, which has the highest depth. Since the observation corresponding to the largest depth value in the data cloud is the deepest value, it is referred to as multivariate median (Tukey, 1975), we named the envelope corresponding to the highest depth value obtained, the median envelope.

Figure 2.5 shows the predicted envelopes for three values of grain number per ear: 7, 11, 15. These values are random values chosen from lower, median, and upper quantiles of the trait distribution. These envelopes demonstrate the dependence of increasing grain number per ear for $p = 0.1$, $p = 0.25$, $p = p^*$, forming the outer envelope, inner envelope, and the median envelope, respectively. The directional quantile envelopes move upward along the data cloud, showing the dependence on increasing the covariate grain number per ear.

For a given value of grain number per ear, the bivariate distribution of yield was estimated from the bivariate median. SWP is then estimated from the bivariate regression estimates of yield in saline and non-saline conditions. Using the bivariate regression estimates, we can compute the estimate of any stress tolerance index for a given plant trait since they are functions of yield in both conditions. We obtain a comprehensive view of how the salinity tolerance index varied for a given plant trait. Figure 2.6 demonstrates the effect of each plant trait, taken one at a time, on the salinity tolerance index SWP.
Figure 2.5: Bivariate regression to predict the joint distribution of yield under non-saline and saline conditions for a given plant trait: The predicted quantile envelopes for three values of grain number per ear: 7, 11, 15 for \( p = 0.1, 0.25, p^* \) forming the outer, inner and median envelope respectively. The different colors indicate different values of grain number per ear.

The results of the bivariate quantile regression were then compared with those obtained using a conventional quantile regression method. We observed an apparent increase in the value of the index as the grain number per ear (Fig. 2.6h), ear number per plant (Fig. 2.6g), and dry mass per m\(^2\) (Fig. 2.6b) increases. However, the increase was not always linear, and the slope varied with the range of values of
Figure 2.6: Relationship between plant traits and stress-weighted performance index (SWP) predicted using bivariate regression: Each of the nine panels represents a scatter plot between a plant trait and SWP. The solid black line represents the predicted values of the index using bivariate regression; the dashed black lines are the lower and upper confidence bands, formed using inner and outer envelopes, respectively, and superimposed is the median regression line in red.

covariates. The stress-weighted performance index (SWP) increased linearly as the ear number per plant increased to 3; it then stopped increasing, even with the further rise in ear number per plant (Fig. 2.6g). A possible explanation is that the plant can still cope with the salt stress while producing seeds, but up to 3 ear number per
plant, after which producing more seeds comes at the expense of salinity tolerance. SWP remained approximately constant for small values of flowering time, and then it decreased linearly as the flowering time increased (Fig. 2.6e). A possible explanation is that the yield of late-flowering plants grown under saline conditions is also affected by heat as the season progresses.

The study of salinity tolerance is crucial to improve crop yield in salt-affected areas. We provide methods to study the effect of plant traits on salinity tolerance. We show that the quantile analysis methods are advantageous over mean regression methods for studying the relationship between covariates and the entire distribution of response by performing a salinity tolerance analysis. Using quantile analyses, we study traits affecting plants with extreme yields and high salinity tolerance indices.

We propose bivariate quantile analysis methods to study the effect of plant traits on salinity tolerance indices. The added value of using bivariate quantile regression is that it provided the ability to predict the bivariate response for a given value of covariate, whereas the traditional method of quantile regression reduced the response to univariate, by taking the ratio of the yield under saline conditions to the square root of yield under non-saline conditions. As previously mentioned, computing a stress index reduces dimensionality to one; hence, we lose information. In the method we used, we do not need to compute the index before making the prediction; we can find the relationship of the plant traits with any salinity tolerance index once we have predicted the bivariate distribution of yield. Another advantage of using this bivariate quantile regression is that it allows nonlinearities in response by using cubic splines on the covariate. Also, since the yield in two conditions is related, it is favorable to model the joint distribution of yield for a given plant trait instead of modeling merely a univariate function of the two components of yield. With conventional univariate regression, we cannot recover the individual estimates of yield, but with the proposed bivariate quantile regression method, we obtain the estimates of yield in saline and
non-saline conditions for a given plant trait.

2.4 Conclusions

From the different quantile analysis approaches used in this chapter, we can make new observations and find out information that could not be previously obtained from analyses such as those presented in Saade et al. (2016). From quantile analyses, using yield and plant traits under saline and non-saline conditions, we can observe the effects of plant traits on yield. We observe that a late flowering time has a negative impact on yield in saline conditions, for accessions with high yield. From quantile analyses, using the SWP index, we notice that an increase in ear number per plant and grain number per ear increases the salinity tolerance index, and in the case of grain number per ear, the rate of increase is larger for accessions with high yield. On the contrary, a late flowering time decreases the salinity tolerance index for accessions with high yield. The use of conditional and marginal quantiles provides a flexible approach for selecting high yielding and high salinity tolerant accessions. From bivariate quantile analysis methods, we observe that SWP remains stable for earlier flowering times and then starts declining as the flowering time increases. SWP increases with an increase in the ear number per plant, and then stabilizes for higher values without any further increase, while it continuously increases for grain number per ear. These observations are biologically relevant and may impact our understanding of mechanisms of tolerance to salinity. The dataset analyzed during this study is available in the Open Science Framework repository (https://osf.io/wzhe7/).
Chapter 3

Bivariate Functional Quantile Envelopes with Application to Radiosonde Wind Data

3.1 Introduction

Weather data obtained from the atmosphere, beginning three meters above the earth’s surface, is known as weather balloon data or upper air data. A small, expendable instrument known as the radiosonde, which is suspended below a 2-meter wide balloon filled with hydrogen or helium that ascends through the upper-air, collects and transmits that data back to the ground. The sensors on the radiosonde measure vertical profiles of temperature, humidity, atmospheric pressure, and geopotential height. By tracking the position of the radiosonde in flight, information on wind speed and direction is also obtained. The Integrated Global Radiosonde Archive (IGRA) consists of more than 1500 globally distributed radiosonde observations from different time periods, ranging from the 1960s to the present; an overview of the dataset is given in Durre et al. (2006). The National Center for Atmospheric Research (NCAR) Upper Air Database (UADB) contains the longest possible time series of radiosonde data, from the 1920s to the present, at more than 1600 global locations. Analyzing such substantial and complex datasets need a robust analysis. The dataset analyzed in this chapter is available online for free at (https://rda.ucar.edu/datasets/ds370.1/) in the Research Data Archive at NCAR (DSS/CISL/NCAR/UCAR, 2014).

Radiosonde data has been extensively used in weather and climate research activities (Chien and Smith, 1973) and studying the features of global warming (Sherwood
The main use of radiosonde data is to input them into climate models for numerical weather predictions. The radiosonde winds were used as data assimilation products in the National Center for Environmental Prediction reanalyses (Kalnay et al., 1996; Kanamitsu et al., 2002). Sherwood et al. (2008) emphasized the utility of wind data in explaining climatological variations, and Hollingsworth and Lönnberg (1986) used radiosonde data to predict wind forecast errors, which were used as an input to a global data assimilation system. So the quality control of weather data is of vital importance. The lack of quality control of the data causes hindrance to the quality of the analysis.

The wind speed and direction from the radiosonde data are recorded in terms of orthogonal velocity components, i.e., $u$-component and $v$-component, along different pressure levels, where $u$ represents the component of wind speed towards the east, and $v$ indicates the component of wind speed towards the north. In fact, the $u$ and $v$ components are the cartesian coordinates of the wind speed and direction transformed from their polar coordinates. Typically, for radiosonde data, the bivariate distribution of these two components of wind is non-Gaussian, heavy-tailed, skewed, and susceptible to both measurement and random errors. Hence, robust methods are needed to analyze such distributions. Therefore, in our research, we study the bivariate distribution of these two components along different pressure levels. We treat the $u$ and $v$ components of wind as bivariate functions of the pressure level and develop robust methods based on quantiles for the radiosonde wind data analysis.

In practice, quantiles can be used to characterize univariate non-Gaussian distributions nonparametrically. While measures based on the mean are sensitive to outliers, the quantile approach is more robust (Moors et al., 1996). The univariate quantile is well-defined because the ordering of observations is straightforward and follows the natural ordering in the real line $\mathbb{R}$; but there is no such natural basis in the case of multivariate data. It was not until the late 1970s, Barnett (1976) made attempts to
define principles for ordering multivariate data that were analogous to the concepts behind univariate order statistics. Chaudhuri (1996b) extended the concept of quantiles to multiple dimensions using the geometry, based on Euclidean distances, of the multivariate data. Chakraborty (2003) defined regression quantiles for multivariate quantile regression. Moreover, Tukey (1975) ordered multivariate data by defining a useful tool of the depth function, known as the Tukey or halfspace depth, which provides a center-outward ranking of data points. Many notions of data depth have been defined since its first introduction, such as the Oja depth (Oja, 1983), simplicial depth (Liu, 1990), projection depth (Zuo and Serfling, 2000; Zuo, 2003), and spatial depth (Vardi and Zhang, 2000; Serfling, 2002a).

Kong and Mizera (2012) defined directional quantile envelopes as the intersection of the halfspaces generated from directional quantiles, i.e., quantiles of projections along multiple directions obtained simultaneously, and showed that they are essentially halfspace (Tukey) depth contours. Tukey depth contours completely characterize the empirical distribution of multivariate datasets, as shown in Struyf and Rousseeuw (1999) and Kong and Zuo (2010). Thus, the directional quantile envelopes provide valuable insights into the multivariate data and are of vital importance. Hallin et al. (2010) defined multivariate quantiles based on the directional approach of Koenker and Bassett’s traditional regression quantiles (Koenker and Bassett Jr, 1982) and used quantile contour plots to study the properties of underlying multivariate distributions, see also Chakraborty (2001). The previous methods are for multivariate data, and we extend the existing approaches to multivariate functional data.

Data is often available in the form of random curves that may be observed as univariate or multivariate functional data in various fields. For instance, in medical science, it can be the growth curves of patients; in meteorology, weather parameters can be observed along several pressure levels, and in chemometrics, the curves aris-
ing from physics and chemistry experiments. Depth notions have been extended to both univariate functional data (Fraiman and Muniz, 2001; López-Pintado and Romo, 2009, 2011) and multivariate functional data (Claeskens et al., 2014; López-Pintado et al., 2014; Hubert et al., 2015). Although those depth notions for multivariate functional data are convenient to rank multivariate functions from the center outwards so that the central region can be defined, it is not easy to construct meaningful contours marginally. Since for the radiosonde application, the marginal characterization of the bivariate distribution is also of interest, especially when the observations are not viewed as an entire launch across all the pressure levels, we chose to use the quantile envelope for marginal distributions and connect them across pressure levels afterwards.

However, with multivariate functional data, the data complexity increases, and it is even more challenging to visualize the data. For univariate functional data, the functional boxplot (Sun and Genton, 2011) and the adjusted functional boxplot (Sun and Genton, 2012) were defined, which uses the depth to order data and the empirical rule to detect outliers. For multivariate functional data, Hubert et al. (2015) used depth functions and distance measures to detect multivariate functional outliers and Dai and Genton (2019, 2018a,b) used multivariate directional outlyingness for visualization and marking outliers. Kong et al. (2015) defined a bivariate quantile model for functional responses for given covariates of interest and constructed directional quantile envelopes to study the quantile association between responses and covariates. Wei (2008) proposed dynamic bivariate quantile contours that are flexible to incorporate time-dependent and covariate-varying effect. Kong and Mizera (2012) used the extremal quantile envelopes to detect outliers, though not for functional data. The current statistical analysis for radiosonde data has been limited to multivariate analysis. For radiosonde data, Sun et al. (2017) used contour approaches based on the bivariate skew-t distribution to mark potential outliers at each pressure level.
Since the observations can be treated as functional data along the pressure level, we develop a new set of quantile based methods to visualize the data and detect potential outliers.

Here, our primary goal is to estimate and predict bivariate non-Gaussian distributions of bivariate functional data nonparametrically, as well as to detect outliers. We improve on the existing approaches for multivariate data to define methods for multivariate functional data. Due to the increasing need for statistical techniques in functional data analysis, we intend to develop directional quantile envelope methods and an effective visualization tool to provide a graphical representation of the bivariate functional data, demonstrating the features of the data might not be apparent otherwise. The existing methods for multivariate functional data, such as multivariate functional depth notions, can not be directly applied to construct meaningful contours marginally. Motivated by the heavily-tailed and skewed bivariate radiosonde wind profiles obtained at several atmospheric pressure levels, we propose a model for estimating the directional quantile envelopes of bivariate functional data, visualizing the geometry of the distribution, and detecting both magnitude and shape outliers. We also propose robust methods of predicting the directional quantile envelopes at an unobserved pressure level for the radiosonde winds, using the methods of quantile regression and quantile kriging. The proposed methods provide a robust prediction for the center of the distribution over the traditional predictor, the conditional mean.

The rest of the chapter is organized as follows. Section 3.2 describes the details of directional quantile envelopes and the proposed model for bivariate functional data. We discuss the visualization techniques and methods of detecting outliers, and also describe the two methods used to predict the directional quantile envelopes. In Section 3.3, we use two simulation designs to show the robustness of the proposed methods against outliers and skewed data. In Section 3.4, we apply these methods to the radiosonde wind data. Finally, Section 3.5 contains discussions and conclusions.
3.2 Methodology

3.2.1 Directional quantile envelopes for bivariate functional data

In this section, we first review the directional quantile envelopes defined by Kong and Mizera (2012) and discuss their useful properties. Then, we define the model used to estimate and predict the directional quantile envelopes for bivariate functional data. We also discuss the estimation of extreme quantile envelopes for outlier detection.

Directional quantile envelopes

To characterize the distribution of multivariate data with quantiles, Kong and Mizera (2012) defined directional quantile envelopes. For a random vector $X$ in $\mathbb{R}^d$, with distribution $F$, they considered a directional approach to project this vector on a straight line and then obtain the quantiles, hence called them directional quantiles. For a direction $s \in \mathbb{R}^d$, belonging to the unit sphere $S^{d-1} = \{ y \in \mathbb{R}^d : \|y\| = 1 \}$, and quantile level $\tau \in (0, 1)$, the $\tau$th directional quantile is defined as

$$Q_{(\tau, s)}(X) = \inf \{ u : P[s^T X \leq u] \geq \tau \}.$$  

The directional quantile relates to the hyperplane, with the equation of hyperplane $s^T x = Q_{(\tau, s)}(x)$, which leads to the characterization of closed halfspace, which is denoted by $H(s, Q_{(\tau, s)}(x)) = \{ x : s^T x \geq Q_{(\tau, s)}(x) \}$. For a fixed $\tau \in (0, 1/2]$, the directional quantile envelope is generated by taking the intersection of all the halfspaces taken in all possible directions of a subset of $S^{d-1}$. Then, the $\tau$th directional quantile envelope is defined as

$$D(\tau) = \bigcap_{s \in S^{d-1}} H(s, Q_{(\tau, s)}(x)).$$
The directional quantile envelopes are related to the halfspace depth contours, associated with Tukey. The Tukey depth of a point \( x \in \mathbb{R}^d \), with respect to the distribution \( P \) of the vector \( X \), is given as the smallest probability obtained such that any closed halfspace contains \( x \), i.e.,

\[
D_{\text{Tukey}}(x | P) = \inf_{u \in S^{d-1}} P(u^T x \geq u^T X).
\]

The directional quantile envelopes are essentially Tukey depth level sets, which means contours of equal depth. Since the depth contours characterize the entire underlying multivariate distribution, the directional quantile envelopes describe the empirical distribution of the multivariate data. In other words, these quantile envelopes give a geometrical representation of multivariate data with a distribution \( F \).

In our proposal, we deal with functional data, where the radiosonde wind is treated as a continuous function in pressure level, and for a given pressure level, the two components of the wind have a bivariate distribution marginally. Although we use the method in Kong and Mizera (2012) to construct directional quantile envelopes, the functional aspect allows us to connect or predict those quantiles continuously along pressure levels.

**Directional quantile envelope models for bivariate functional data**

We extend the concept of directional quantile envelopes to the bivariate functional data setup. In practical situations, functional data is usually generated in a discretized form, and is denoted by \((X(t_1), \ldots, X(t_m))\). We introduce a model that we will use to estimate the directional quantile envelopes for bivariate functional data. We consider the bivariate functional data, \( X(t) = (X_1(t), X_2(t))^T \), where \( t = t_1, \ldots, t_m \), belong to an interval \( T \) in \( \mathbb{R} \). For a given \( t \) and fixed \( \tau \in (0, 1/2] \), we model the directional
quantile in the direction \( \mathbf{s} \) as

\[
Q_{(\tau, \mathbf{s})}(\mathbf{X}(t)) = \beta_{0(\tau, \mathbf{s})} + b_{(\tau, \mathbf{s})}(t),
\]  

(3.1)

where \( b_{(\tau, \mathbf{s})}(t) \) is the cubic spline function given by

\[
b_{(\tau, \mathbf{s})}(t) = \beta_{1(\tau, \mathbf{s})}t + \beta_{2(\tau, \mathbf{s})}t^2 + \beta_{3(\tau, \mathbf{s})}t^3 + \sum_{j=1}^{K} \delta_{j(\tau, \mathbf{s})}(t - k_j)_{+},
\]

and \( \{k_j\}_{j=1}^{K} \) are prespecified set of knots. The knots are typically chosen as suitable quantiles of \( t \), as in spline smoothing. Then, we model the directional quantile for all possible directions of the subset of \( \mathcal{S}^{d-1} \). Therefore, for each direction \( \mathbf{s} \), the estimated directional quantile relates to the halfspace

\[
H(\mathbf{s}, Q_{(\tau, \mathbf{s})}(\mathbf{x}(t))) = \{ \mathbf{x} : s^T \mathbf{x} \geq Q_{(\tau, \mathbf{s})}(\mathbf{x}(t)) \},
\]

and the estimated directional quantile envelopes for the bivariate functional data at \( t \) are given by

\[
D_t(\tau) = \bigcap_{\mathbf{s} \in \mathcal{S}^{d-1}} H(\mathbf{s}, Q_{(\tau, \mathbf{s})}(\mathbf{x}(t)))).
\]

We call \( D_t(\tau) \) the marginal directional quantile envelopes, and define the marginal median envelope as the one corresponding to the highest \( \tau \) because the point corresponding to highest depth value is the deepest (Tukey, 1975). Then, the marginal median is defined as the average of the vertices of the median envelope, and the functional median is defined by connecting the marginal medians at all \( t \). Although the directional quantile envelopes can be estimated for any dimension \( d \geq 2 \), for our application and visualization purposes, we only consider the two-dimensional case, which can be easily visualized.

For the bivariate functional data observed at \( t_1, \ldots, t_m \), the parameters in the model based on the spline (3.1) are estimated by quantile regression at a given \( \tau \) for all \( \mathbf{s} \). Then, the \( \tau \)th directional quantile envelopes, which we call predicted functional directional quantile envelopes, can be constructed for any \( t \). For the bivariate case, we uniformly choose 100 directions along the unit circle. The parameters in the quantile regression model (3.1) are estimated using all the data for each direction,
but the predicted directional quantile envelopes are constructed for a given \( t \) using the estimated parameters. Then, the predicted directional quantile generates a halfspace for each direction. We take the intersection of these halfspaces, along 100 directions, and the polygon formed from the vertices of the intersecting lines forms the predicted directional quantile envelope.

**Extremal quantile envelopes**

In general, because of insufficient data in the tails of a distribution, the model for directional quantile envelopes for very small values of \( \tau \), e.g., 0.001 for 100 observations, will not be a good fit. Similar to Kong and Mizera (2012), for very small values of \( \tau \), we fit a generalized Pareto distribution (GPD) to the tails of the distribution of \( s^T X \), for each direction \( s \) and each \( t \), and obtain the estimate of the prescribed directional quantiles. For the projected distribution \( s^T X \), the GPD with shape parameter \( \xi \) and scale parameter \( \beta \) is given by

\[
\tau = F_{P_s}(s^T x) = \begin{cases} 
1 - (1 - \frac{\xi s^T x}{\beta})^{1/\xi}, & \xi \neq 0, \\
1 - \exp(-\frac{s^T x}{\beta}), & \xi = 0.
\end{cases}
\]

We consider 1000 directions along a unit circle. For each direction \( s \), we choose the 95th quantile of \( s^T x \) as the threshold and fit GPD to all exceedances. We estimate the parameters of the distribution using the method of maximum likelihood. After fitting the distribution, the directional extremal quantiles at \( t \) are estimated by

\[
q_{(s,t)}(\tau) = \begin{cases} 
\hat{\beta}_{(s,t)} \{1 - (1 - \tau)\hat{\xi}_{(s,t)}\}/\hat{\xi}_{(s,t)}, & \hat{\xi}_{(s,t)} \neq 0, \\
-\hat{\beta}_{(s,t)} \log(1 - \tau), & \hat{\xi}_{(s,t)} = 0,
\end{cases}
\]

where \( \hat{\xi}_{(s,t)} \) and \( \hat{\beta}_{(s,t)} \) are the estimated shape and scale parameters, respectively. For each \( t \), the extremal quantile envelopes are estimated as the intersection of the
halffspaces generated by the estimated directional extremal quantiles from the fitted GPD.

3.2.2 Outlier detection and visualization

We detect outliers using the extreme quantile envelopes described above and demonstrate it using the radiosonde wind data. We visualize the vertical profiles of the radiosonde wind across several atmospheric pressure levels using the estimated directional quantile envelopes, extreme quantile envelopes, and the estimated median. We devise a method of connecting the envelopes along pressure levels to characterize the changes. We also define two methods for functional outlier detection.

Detection of outliers

The radiosonde wind data has two components of wind, $u$-component and $v$-component, measured in meters per second (m/s), the components of wind towards east and north, respectively. These wind components are measured along eight standard atmospheric pressure levels: 70, 100, 200, 250, 300, 400, 500, and 700 hectopascals (hPa). The extreme quantile envelopes do not have the functional feature; however, we introduce it to detect bivariate outliers that are often of interest in radiosonde data, rather than the entire launch. The wind components were measured at each level directly from the atmosphere and, hence, the measurements are prone to outliers for many reasons, such as human error and data preprocessing errors. We estimate the extreme quantile envelopes for $\tau = 0.0002$ at each pressure level. The observations lying outside these extreme quantile envelopes are marked as bivariate outliers. The estimated extreme quantile envelopes for all eight pressure levels of radiosonde wind data at quantile level $\tau = 0.0002$ are shown in Figure 3.1. The bivariate outliers at each pressure level are marked by red dots. The outliers were in the range of 0.06% to 0.11% for all pressure levels.
Figure 3.1: Each of the eight panels shows a scatter plot of the $u$ and $v$ components of the radiosonde wind data and the estimated extreme quantile envelopes for $\tau = 0.0002$ at a given pressure level. The outliers are marked by red dots.

**Visualization of quantile envelopes**

We consider the two components of wind across pressure levels to be discretized bivariate functional data. Using the model and the estimation procedure defined in Section 3.2.1, we estimate the directional quantile envelopes of these bivariate functional data. For each $t$, we compute the average of the vertices of the directional quantile envelope at the highest $\tau$ for which the directional quantile is not empty and, then we connect the bivariate medians of each pressure level to obtain the functional median. Visualizing variability in bivariate functional data is not trivial, and there is no universal way of visualization of bivariate functional data. We propose a visualization tool for bivariate functional data to visualize the centrality of curves and variability in the curves by characterizing the marginal distributions by directional quantile envelopes and detecting outliers.

To illustrate our visualization tool for bivariate functional quantile envelopes, we
Figure 3.2: The solid contours are the estimated directional quantile envelopes for \( \tau = 0.05, 0.0125 \), and the dashed contours are the extreme quantile envelopes; the color variation represents the different pressure levels. The red dots are the outlying launches. The solid black line is the estimated functional median. (a) First time period 1962-1986, and (b) second time period 1987-2011.

divide the radiosonde wind data into two time periods, 1962-1986 and 1987-2011, and estimate the directional quantile envelopes at each pressure level. The estimated directional quantile envelopes for \( \tau = 0.05, 0.0125 \), the extreme quantile envelopes for \( \tau = 0.0002 \), the functional median, and the outlying launches are plotted in Figure 3.2. The directional quantile envelopes give an idea of the shape of the distribution, and we observe that the mid pressure levels have higher variability than the upper and lower pressure levels. From the median curve, it can be observed that the median wind is towards the southeast direction and the wind speed increases as the pressure level drop from 700 to 200, and then it decreases as the pressure level drops further, with the maximum speed attained at pressure level 200. The launches that had outliers at any of the pressure levels are marked as outlying launches. The first time
period, 1962-1986, contains 0.67% outliers, while the second time period, 1987-2011, contains 0.58% outliers. There are only slight differences in the distributions of two time periods, though they become more apparent in a two-dimensional plot where one pressure level is considered at a time.

**Connected envelopes**

To account for changes in the distribution along $t$, we introduce a method of connecting the directional quantile envelopes along $t$. Each directional quantile envelope is an intersection of the halfspaces corresponding to each direction. However, we do not have a point on the envelope that corresponds to each direction; rather, we have it just for the directions, whose halfspaces lie in the boundary of the envelope. We call such directions active directions. To obtain a point on the envelope corresponding to the directions that are not active, i.e., that lie outside the boundary of the envelope, we uniformly distribute the points between two active directions. Now, since we can obtain a point corresponding to each direction on the envelope, we connect the envelopes along each direction, and the connected envelopes characterize the changes along $t$ in all directions.

For radiosonde wind data, connected envelopes are useful to characterize the changes in the distribution of wind along pressure levels, which is often of interest. The connected envelopes for extreme quantiles also provide a way to flag certain potential outlying launches with respect to the observed distribution. Moreover, the proposed method of connected envelopes is different than constructing envelopes at each level and then connecting them. In the proposed methods, the contours at each level borrow information from other pressure levels. If we compute the contours separately at each pressure level, then the contours at each pressure level would be rougher, and it is also not clear on how to connect those separate contours.

Here, we demonstrate connected envelopes using the radiosonde wind data from
Figure 3.3: The three-dimensional plot represents the $u$ component on the x-axis, the $v$ component on the y-axis, and the pressure level on the z-axis. The grey dots represent the observed $u$ and $v$ components of wind speeds from the radiosonde data, and the black contours represent the estimated directional quantile envelopes for $\tau = 0.05$ during the time period 1987-2011. The different colors of the connected directional quantile envelopes for $\tau = 0.05$ represent different directions.

The second time period, 1987-2011. We obtain a point corresponding to each direction $s$, and connect the points by pressure level. We chose 100 directions to estimate the directional quantile envelopes. We estimate a point corresponding to each of the 100 directions on the envelope and connect the envelopes corresponding to each direction across the eight pressure levels. The connected directional quantile envelopes of the radiosonde wind data for $\tau = 0.05$ during the second time period are shown in Figure 3.3. At each pressure level, the grey dots represent the observed $u$ and $v$ components of the wind speed, and the black contours are the estimated directional
quantile envelopes. The different directions of the connected envelopes are represented by different colors. These connected lines show how the launches behave, for a given direction, at a quantile level of 0.05 or, equivalently, the launches with a Tukey depth of 0.05. If we increase the quantile level, i.e., increase the depth, then the contours will be smaller and closer to the center, and the connected lines will be closer to the center as well. Similarly, if we estimate the quantile envelopes for a lower quantile level, i.e., decrease the depth, the contours will be bigger, and the connected lines for each direction will represent launches with a smaller depth. We can also estimate the connected lines for the tails of the distribution at a very small quantile level by estimating the extreme quantile envelopes.

Functional outlier detection

Our proposed method provides quantile envelope estimation for bivariate functional data. Since the quantile based method is robust against outliers, it can be used to identify potential outliers. In Section 3.2.2, we detected the bivariate outliers using extreme quantile envelopes and marked the launches as outlying launches if they had outliers at any pressure level. Here, we present some other approaches of finding functional outliers. Functional outliers can be in the form of magnitude outliers, in which the curves that are distant from the majority of curves are marked as outliers, or shape outliers, where the shape of the curve is different from the shape of the majority of curves in the sample. We suggest two distance measures that might detect some magnitude and shape outliers. If outlier detection is the only purpose rather than visualization, there exist more sophisticated outlier detection procedures in the literature (Hubert et al., 2015; Rousseeuw and Hubert, 2018). To illustrate the two kinds of outliers, we simulate bivariate curves from a mean-zero bivariate Gaussian process with a bivariate Matérn covariance function (as explained in 3.3.1); then, we contaminate them to simulate a magnitude outlier and a shape outlier, as
shown in Figure 3.4. The curve in red is the shape outlier, while the magnitude outlier is shown in blue.

**Magnitude outliers**

To detect magnitude outliers, we need to find the curves that lie at an extreme distance from most of the other curves. We estimate the functional median of the sample using directional quantile envelopes and find the distance of all the curves from this estimated median curve. Specifically, a sample curve is denoted by $f(t)$, and the median curve is denoted by $f_M(t)$ for $t = t_1, \ldots, t_m$. The measure of distance used to find the magnitude outliers is given by

$$D_M(f) = \sum_{i=1}^{m} d(f(t_i), f_M(t_i)),$$
where $d$ is the Euclidean distance. We compute the distance measure for all the curves in the sample, and the outliers are marked using a boxplot. To illustrate this method, we simulate 1000 curves from a bivariate Gaussian distribution with mean zero and contaminate 1% of the curves with magnitude outliers. We estimate the functional median and compute $D_{M}$ for all the curves. We construct the boxplot for $D_{M}$; it is shown in Figure 3.5a. We observed that all the simulated magnitude outliers could be detected from the boxplot.

### Shape outliers

To detect shape outliers, we need to find the curves with a pattern different from the others. Following similar notations to those used for magnitude outliers, we define a distance measure, which is given by

$$D_{S}(\mathbf{f}) = \sum_{i=2}^{m} \frac{d(\mathbf{f}(t_i), \mathbf{f}(t_{i-1}))}{t_i - t_{i-1}}.$$
We contaminate 1% of the curves with shape outliers and compute $D_S$ for all the curves. Then, the outliers are marked using a boxplot. Figure 3.5b shows the boxplot for $D_S$; all the shape outliers are flagged as such.

### 3.2.3 Quantile envelope prediction

In this section, we propose two methods of predicting the directional quantile envelopes at any new $t^*$ within the support of the distribution. To construct a directional quantile envelope, there are two steps: first is to predict the directional quantile in a given direction; second is to take the intersection of the halfspace generated by the predicted directional quantile in all directions. We define two methods to predict the directional quantile in the first step. The first method uses a quantile regression model, while the second method focuses on predicting the directional quantile using univariate kriging.

#### Quantile regression model

We consider the discretized form of the bivariate functional data $(X(t_1), \ldots, X(t_m))$, where $X(t) = (X_1(t), X_2(t))$ for $t \in \mathcal{T}$ in $\mathbb{R}$. The first method of predicting directional quantile envelopes at a new $t^* \in \mathcal{T}$ is based on the quantile regression model (3.1). For a fixed $\tau \in (0, 1/2]$, the predicted conditional directional quantile in any direction $s$ can be modeled as

$$Q_{(\tau,s)}(X(t^*)) = \hat{\beta}_{(\tau,s)} + \hat{b}_{(\tau,s)}(t^*), \quad t^* \in \mathcal{T}.$$ 

The predicted conditional directional quantile in direction $s$, denoted by $Q_{(\tau,s)}(X(t^*))$, can be obtained by plugging all the estimated parameters into model (3.1). Once we predict the $\tau$th directional quantile in all directions, then we can construct the directional quantile envelope. The predicted directional quantile envelope at $t^*$ is given
by

\[ D_{t^*}(\tau) = \bigcap_{s \in S^{d-1}} H(s, Q_{(\tau,s)}(\mathbf{x}(t^*))). \]

Consequently, we can also predict the median at \( t^* \), by predicting the median envelope at \( t^* \), for the highest value of \( \tau \in (0, 1/2] \), such that the predicted envelope is not empty.

Quantile kriging

In previous studies, kriging models for fixed univariate quantiles have proven effective in building stochastic emulators (Plumlee and Tuo, 2014). Here, we propose a method to obtain bivariate functional quantiles by predicting the \( \tau \)th directional quantile in a direction \( s \) at \( t^* \) via kriging (Matheron, 1963; Cressie, 1993), conditional on the directional quantiles at \( t_1, \ldots, t_m \). If the \( \tau \)th directional quantile in direction \( s \) at \( t_i \), \( i = 1, 2, \ldots, m \), is denoted by \( Q_{(\tau,s)}(\mathbf{x}(t_i)) \), then the kriging predictor at \( t^* \) is given by

\[ Q_{(\tau,s)}(\mathbf{x}(t^*)) = \sum_{i=1}^{m} \lambda_i Q_{(\tau,s)}(\mathbf{x}(t_i)), \]

where \( \lambda_i, i = 1, 2, \ldots, m \) are the kriging weights. The kriging weights are chosen such that the predictor is unbiased, and the mean squared prediction error is minimized. For ordinary kriging, the estimated kriging weights \( \lambda = (\lambda_1, \ldots, \lambda_m) \) are given by

\[ \lambda = \Sigma^{-1} \left( c + 1 \frac{1 - 1^T \Sigma^{-1} c}{1^T \Sigma^{-1} 1} \right), \]

where \( \Sigma = \text{Cov}(Q, Q^T) \), \( c = \text{Cov}(Q, Q_{(\tau,s)}(\mathbf{x}(t^*))) \), and \( Q = (Q_{(\tau,s)}(\mathbf{x}(t_1)), \ldots, Q_{(\tau,s)}(\mathbf{x}(t_m))) \). To construct \( \Sigma \), we use the Matérn covariance function. The Matérn
correlation is given by

\[ M(h|\nu, \alpha) = \frac{2^{1-\nu}}{\Gamma(\nu)} \left( \frac{\|h\|}{\alpha} \right)^\nu K_{\nu}\left( \frac{\|h\|}{\alpha} \right), \]

(3.2)

where \( K_{\nu} \) is a modified Bessel function of the second kind, \( \nu \) is the smoothness parameter, and \( \alpha \) is the range parameter. The Matérn covariance function is given by \( C(h) = \sigma^2 M(h|\nu, \alpha) \), where \( \sigma^2 \) is the marginal variance. The covariance function depends only on the separation vector \( h \). The Matérn covariance model can be estimated by ordinary least squares, weighted least squares, or maximum likelihood estimation by assuming a Gaussian random field (Cressie, 1985, 1993; Sherman, 2011).

We use univariate kriging to predict the \( \tau \)th directional quantile in direction \( s \) at \( t^* \). Then, the predicted directional quantile envelope for a fixed \( \tau \in (0, 1/2] \) at \( t^* \) is given by

\[ D_{t^*}(\tau) = \bigcap_{s \in S^{d-1}} H(s, Q(\tau, s)(X(t^*))). \]

We can also predict the median using kriging by predicting the median envelope. Since the quantile kriging method involves the inversion of the covariance matrix, it will be slower than the quantile regression method for larger sample sizes.

### 3.3 Simulation study

To illustrate our proposed methods, we conduct two simulation studies in which we predict the directional quantile envelopes using the methods of quantile regression and quantile kriging. We predict the center of the distribution using the directional quantile envelope that corresponds to the highest depth, i.e., the median envelope, using different methods and then compare it with the true center of the distribution. In the first simulation, we simulate the curves from a bivariate Gaussian process
with mean zero, then contaminate the curves by introducing outliers. In the second simulation, we simulate curves from a model motivated by the radiosonde wind data.

### 3.3.1 Bivariate Gaussian random field

We simulate bivariate functional data $X(t) = (X_1(t), X_2(t))^T$ for $t = t_1, \ldots, t_m$ from a bivariate Gaussian random field with mean zero. We assume a stationary full bivariate Matérn covariance model (Gneiting et al., 2010). In a full bivariate Matérn model, the marginal covariance functions are given by

$$C_{11}(h) = \sigma_1^2 M(h|\nu_1, \alpha_1), \quad C_{22}(h) = \sigma_2^2 M(h|\nu_2, \alpha_2),$$

and the cross-covariance function is given by

$$C_{12}(h) = C_{21}(h) = \rho_{12} \sigma_1 \sigma_2 M(h|\nu_{12}, \alpha_{12}),$$

where $M()$ is of the form (3.2), $\sigma_1^2$ and $\sigma_2^2$ are the marginal variances, $\rho_{12}$ is the cross-correlation, and $\nu_{12}$ and $\alpha_{12}$ are the cross-covariance smoothness and range parameter, respectively. We simulate the bivariate curves and consider different outlier models. We set $\sigma_1 = 0.1$, $\sigma_2 = 0.1$, $\rho_{12} = 0.5$, $\nu_1 = 1$, $\nu_2 = 2$, and $\nu_{12} = 1$, let $\alpha_1 = \alpha_2 = \alpha_{12}$, and vary them from 0.1 to 0.5. The details of the four models used in the simulation are as follows:

1. Model 1: $X(t) = e(t)$, where $e(t) = (e_1(t), e_2(t))^T$ is the bivariate Gaussian process with mean zero and full bivariate Matérn covariance model.

2. Model 2: $X(t) = e(t) + d(t)$, $d(t) = (s_1 C, 0)$, where $C = 4$ indicates the contamination size. Here, $s_1 = 1$ with probability $p$; otherwise $s_1 = 0$. Hence, $s_1$ determines the contamination rate. We choose $p = 0.01$; therefore, the contamination rate is 1%. 

3. Model 3: Same as Model 2 except $p=0.05$, i.e., 5\% contamination rate.

4. Model 4: Same as Model 2 except $p=0.1$, i.e., 10\% contamination rate.

We generate 1000 curves from each of the four models and for each range parameter, 0.1, 0.3, and 0.5, at $m = 9$ points (70, 100, 200, 250, 300, 350, 400, 500, 700) scaled to $[0, 1]$. We predict the center of the distribution at a given $t$, leaving out the data for that particular $t$. We choose $t_0 = 0.4444$ and compare our proposed median estimation to the conditional mean, which is the best predictor for a Gaussian process (Krige, 1981).

We repeat the simulation 100 times and compare the predicted center with the true center of the distribution, which is $(0, 0)$. We note the mean square predicted error (MSPE) for all four models and different range parameter values. Figure 3.6 shows the bar plots of MSPE obtained using three methods, the two that we propose here using quantile regression and quantile kriging, and the method of conditional mean, for models with no outliers, 1\% outliers, 5\% outliers, and 10\% outliers, and for three different range parameters, 0.1, 0.3, and 0.5. In the case of no outliers, MSPE is comparable for the three methods, and all three methods perform well. However, in the case of no outliers, quantile kriging is better than the method of quantile regression. In this simulation setting, quantile kriging is better in prediction since we assume that the directional quantile along the pressure level follows a Gaussian process. When outliers are introduced, the method of the conditional mean is affected the most and performs the worst, while the two methods using quantile regression and quantile kriging are robust to outliers. As the contamination rate increases from 1\% to 10\%, this becomes more apparent, as MSPE for the method of conditional mean increases more quickly than for the other two methods. The method of quantile regression is least affected by the outliers, and, in the case of 10\% outliers, the method of quantile regression performs the best. The results are consistent across different range parameters. The computational costs of the two methods were comparable;
Figure 3.6: Results of simulations from a bivariate Gaussian process. The four panels represent the models with no outliers, 1% outliers, 5% outliers, and 10% outliers. Each panel shows bar plots of MSPE for the three methods, conditional mean (blue), quantile kriging (green), quantile regression (red), and for three cases of range parameters, 0.1, 0.3, and 0.5.

The cost of the quantile kriging method was slightly higher. The quantile regression method took 32.96 minutes, while quantile kriging took 37.89 minutes for running 100 simulations. The cost of kriging increases with the number of pressure levels. Therefore, the quantile regression method would be much faster than the kriging method with a high number of pressure levels.

3.3.2 Simulation motivated by radiosonde data

In this simulation design, we simulate curves from the linear models motivated by the radiosonde wind data. The bivariate functional data $X(t) = (X_1(t), X_2(t))^T$, for
\[ t = 70, 100, 200, 250, 300, 400, 500, 700, \text{ is generated from the linear model} \]

\[ X_i(t) = \beta_0 + \beta_1 D_i + b(t) + \epsilon_i, \quad i = 1, 2, \]

(3.3)

where \( b(t) \) is the cubic spline function given by

\[ b(t) = \beta_1 t + \beta_2 t^2 + \beta_3 t^3 + \sum_{j=1}^{K} \delta_j (t-k_j)^3 \]

and \( \{k_j\}_{j=1}^{K} \) are a prespecified set of knots. \( D_i \) is a dummy variable: \( D_i = 0 \) if \( i = 1 \) (\( u \)-component of wind) and \( D_i = 1 \) if \( i = 2 \) (\( v \)-component of wind). First, we fit this spline model (3.3) to the radiosonde wind data and obtain the estimated coefficients. Then, we generate data using the estimated coefficients, with \( \epsilon_1 \) and \( \epsilon_2 \) assumed to be independently simulated from a skew normal distribution with the shape parameter (denoted by \( \alpha \)) being, 0, 1, 2, and 3.

We simulate 1000 curves at eight pressure levels: 70, 100, 200, 250, 300, 400, 500, and 700. At pressure level 300, we leave out the data and predict the median using the three methods, quantile kriging, quantile regression, and the conditional mean. Then, we compare them with the truth. We repeat this simulation 100 times and compute MSPE. The MSPE results from the four models that correspond to the four shape parameters (\( \alpha = 0, 1, 2, \) and 3) are listed in Table 3.1 for the three methods. The quantile kriging method performs the best, followed by the method of quantile regression, and then conditional mean. In the case of normally distributed errors (i.e., in the case of \( \alpha = 0 \)), all three methods perform reasonably well and are comparable. However, as the skewness of the distribution increases (i.e., as \( \alpha \) increases), the method of conditional mean performs poorly, while the quantile kriging and quantile regression methods are still robust. We have done other simulations with more functional data points, and the results are similar. The simulation designs we show is to mimic the available radiosonde data.
Table 3.1: Mean square error from model motivated by real data.

<table>
<thead>
<tr>
<th>Method</th>
<th>α</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Quantile kriging</td>
<td></td>
<td>0.2453</td>
<td>0.2448</td>
<td>0.2455</td>
<td>0.2482</td>
</tr>
<tr>
<td>Quantile regression</td>
<td></td>
<td>0.3083</td>
<td>0.3157</td>
<td>0.2902</td>
<td>0.3144</td>
</tr>
<tr>
<td>Conditional mean</td>
<td></td>
<td>0.2896</td>
<td>1.1008</td>
<td>1.2979</td>
<td>1.3744</td>
</tr>
</tbody>
</table>

3.4 Applications to radiosonde winds

To demonstrate our methods, we use radiosonde wind data from a station in Denver, Colorado. The radiosonde wind data consists of two wind components, i.e., $u$-component and $v$-component, obtained by transforming polar coordinates of wind speed and direction to cartesian coordinates. We consider a set of 35,816 launches, which were recorded twice a day from March 3, 1962, to October 31, 2011, along eight standard atmospheric pressure levels, 70, 100, 200, 250, 300, 400, 500, and 700 hPa. The estimated directional quantile envelopes for the bivariate radiosonde wind data from two periods have been discussed in Section 3.2.2. The connected envelopes were estimated to characterize the changes in radiosonde wind data across pressure levels in Section 3.2.2. In this section, we apply the methods of predicting quantile envelopes that were discussed in Section 3.2.3 with the methods of quantile regression and quantile kriging.

3.4.1 Prediction of median

In this section, we predict the center of the bivariate distribution of radiosonde winds at each pressure level using the methods of quantile kriging, quantile regression, and conditional mean. First, we center the observations from each pressure level to eliminate the potential trend along the pressure level, by removing their respective bivariate Tukey median, which is a robust estimator of the center. Then, we perform kriging along the pressure level, assuming zero mean. We predict the median using
quantile kriging and quantile regression, and compare it with the estimated median of the observed data using the leave-one-out cross-validation approach. We also compute the conditional mean and compare it to the observed median. We leave one pressure level out at a time and compute the MSPE in the following way:

$$\text{MSPE} = \frac{1}{p} \sum_{i=1}^{p} d(m_i, \hat{m}_i),$$

where $d$ is the Euclidean distance, $m_i$ and $\hat{m}_i$ are the true and predicted medians of the $i^{th}$ pressure level, respectively, and $p$ is the number of pressure levels left out. We leave out each of the pressure levels 100, 200, 250, 300, and 400, one at a time. The MSPE results for both time periods, and the three methods are shown in Table 3.2.

For both time periods, we observe that the method of quantile kriging performs best. The reason is that the quantile kriging method takes advantage of the correlation among the vertical profiles of wind speed across pressure levels. For the first time period, the methods of quantile regression and conditional mean are comparable, with the conditional mean method performing slightly better, while for the second time period, the quantile regression method performs better than the conditional mean method. Therefore, the proposed method can predict the center of the distribution of radiosonde winds robustly for skewed and heavy-tailed data and provide summary statistics for the bivariate distribution of components of the wind.

Table 3.2: MSPE obtained by comparing the center of the radiosonde wind data distribution using leave-one-out cross-validation over two time periods.

<table>
<thead>
<tr>
<th>Method</th>
<th>MSPE$_1$</th>
<th>MSPE$_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Quantile kriging</td>
<td>0.2343</td>
<td>0.2224</td>
</tr>
<tr>
<td>Quantile regression</td>
<td>0.9318</td>
<td>0.7877</td>
</tr>
<tr>
<td>Conditional mean</td>
<td>0.9267</td>
<td>0.9431</td>
</tr>
</tbody>
</table>
3.4.2 Predicting envelope for a given $\tau$

Above, we demonstrated the application of quantile envelope prediction by predicting the median using methods based on quantile kriging, quantile regression, and the conditional mean. Here, we predict the envelopes for any $\tau \in [0, 1/2)$. We predict the distribution of the bivariate distribution of wind at unknown pressure levels, using the information recorded at the standard atmospheric pressure levels from the radiosonde wind data. We illustrate the prediction of quantile envelopes for a given pressure level and for a fixed $\tau$ by predicting the quantile envelope at each of the known pressure levels and then comparing it with the envelope estimated from the observations at the corresponding pressure level. We take the pressure levels 200 and 300 for $\tau = 0.05$ as an example. Figure 3.7 shows scatter plots between the $u$ and $v$ components of the radiosonde wind at the pressure levels 200 and 300. The quantile envelopes predicted by quantile kriging and quantile regression are shown in green and red, respectively, and the observed envelope in blue. The predictions of both methods are reasonably good, as they are very close to the observed envelope for both the pressure levels. These methods can be used to predict the bivariate distribution of wind in parts of the atmosphere where no radiosonde observations are available.

It is difficult to determine visually from Figure 3.7 which method performs better. In Section 3.2.2, we defined the idea of connected envelopes, where we obtain a point on an envelope corresponding to each direction. Using the connected envelopes, we compare the predicted quantile envelopes from both the methods numerically. We obtain MSPE as

$$\text{MSPE} = \frac{1}{100} \sum_{i=1}^{100} d(p_{s_i}, \hat{p}_{s_i}),$$

where $p_{s_i}$ and $\hat{p}_{s_i}$ are the points on an envelope corresponding to the direction $s_i$, $i = 1, 2, \ldots, 100$, and $d$ is the Euclidean distance. We estimate the quantile envelopes
Figure 3.7: The plots show the scatter between the $u$-component and $v$-component of the wind along with the estimated quantile envelope (blue) from the observed data and the predicted quantile envelopes from the method of kriging (green) and quantile regression (red). The left panel shows the plot for pressure level 200, and the right panel represents pressure level 300.

for the pressure levels 100, 200, 250, 300, and 400 for $\tau = 0.05$, and compare them to the observed envelope to obtain MSPE. Table 3.3 shows MSPE obtained using both methods. Overall, the quantile kriging method has a better performance. However, the quantile regression method performs better in some cases, such as the mid pressure levels. For instance, the quantile kriging method performs better for pressure level 200, while the quantile regression method is better for pressure level 300.

### 3.5 Discussion

In this chapter, we define a model for estimating and predicting the directional quantile envelopes of bivariate functional data, $X(t) = (X_1(t), X_2(t))^T$, where $t$ belongs to an interval $\mathcal{T}$ in $\mathbb{R}$, and extend the existing approaches for multivariate data. The functional aspect allows us to connect or predict those quantiles continuously
Table 3.3: MSPE obtained by comparing envelopes at each direction of the two methods kriging and quantile regression from the observed envelope.

<table>
<thead>
<tr>
<th>Pressure level</th>
<th>Quantile kriging</th>
<th>Quantile regression</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>2.6881</td>
<td>7.4713</td>
</tr>
<tr>
<td>200</td>
<td>2.8423</td>
<td>4.0829</td>
</tr>
<tr>
<td>250</td>
<td>2.5331</td>
<td>2.3622</td>
</tr>
<tr>
<td>300</td>
<td>3.4074</td>
<td>3.1900</td>
</tr>
<tr>
<td>400</td>
<td>4.0831</td>
<td>6.4693</td>
</tr>
</tbody>
</table>

along $t$. These directional quantile envelopes completely characterize the distribution marginally for each $t$. The proposed nonparametric approach is suitable for estimating bivariate distributions that are non-Gaussian and skewed. We use the proposed model to estimate the directional quantile envelopes for the $u$ and $v$ components of radiosonde wind data across several atmospheric pressure levels.

The developed visualization tool reflects the geometry of bivariate functional data with the help of the estimated directional quantile envelopes, functional median, and outlier detection. We visualize the geometry of the radiosonde wind data at eight standard atmospheric pressure levels by estimating the directional quantile envelopes at three quantile levels and estimating its functional median. To flag outliers and errors in the radiosonde measurements, we estimate the extreme quantile envelopes and detect the outliers at each pressure level. We define two methods of detecting two types of functional outliers, both magnitude and shape outliers.

We also define two methods of predicting directional quantile envelopes at a given $t$ for bivariate functional data. It provides a way of obtaining the bivariate conditional distributions of non-Gaussian processes. The first method uses quantile regression to model the $\tau^{th}$ directional quantile, and the second uses a kriging approach to predict the $\tau^{th}$ directional quantile at a new $t$. Using simulation studies, we show that both methods are robust against outliers and skewed distributions. Overall, the method of quantile kriging performed better, but there were some cases (e.g., Gaussian random
field with 10% outliers) when the quantile regression method worked better. Both of these proposed methods provide a robust prediction over the conditional mean.

The R code for the methods described in this chapter has been made publicly available on https://github.com/agarwag/Bivariate-Functional-Quantile-Envelopes. The readers can use the methods to visualize bivariate functional data and estimate even non-Gaussian, skewed, heavy-tailed distributions.
Chapter 4

Flexible Quantile Contours for Multivariate Functional Data: Beyond Convexity

4.1 Introduction

Functional data, where each point is a curve, surface, or anything else varying over a continuum, arise in all areas of science, medicine, and engineering. During the last two decades, the area of functional data analysis (FDA) has greatly attracted attention among researchers. However, most of the methods available in the literature are developed for univariate functional data, while the analysis of multivariate functional data has been largely overlooked. Multivariate functional data can be viewed as realizations of multivariate random functions; for example, children’s height and weight over age, patient’s CD4 cell count and viral load over the duration of therapy, and observed variables such as temperature, geopotential height, and PM$_{2.5}$ over time. Computing quantiles is one of the essential tasks for data analysis; however, the extension of quantiles to multivariate functional data is quite challenging. In this chapter, we propose a method to estimate the quantile contours of multivariate functional data.

PM$_{2.5}$ is an air pollutant that represents the concentration of fine particulate matter with a diameter of less than 2.5 µm. The exposure to high concentrations of PM$_{2.5}$ poses a considerable risk for human health and have known to cause respiratory and cardiovascular illness (World Health Organization, 2003; Cohen et al., 2005; Fuentes et al., 2006; Hu et al., 2008). Various meteorological variables, such as
temperature, wind, geopotential height, affect the PM$_{2.5}$ concentrations significantly (Li et al., 2015; Russell et al., 2017). However, most of these analyses were done in a univariate regression framework, and the effects were estimated at the mean level without reflecting any dynamic change in time. Consequently, a deeper understanding of the relationship between PM$_{2.5}$ and meteorological variables can help us formulate pollution control strategies and healthcare policies. In this chapter, we study the joint distribution of PM$_{2.5}$ and geopotential height at 850 hPa over time across the Northeastern United States. The bivariate distribution of PM$_{2.5}$ and geopotential height at a given time may have nonconvex features and is of particular interest as high-pressure systems can trap pollutants and cause high PM$_{2.5}$ concentrations. Previous literature has shown that several air pollution crises have occurred when maximum values of geopotential height were recorded above the city (Saunders and Waugh, 2015; He et al., 2018). We aim to develop an approach to estimate the joint distribution of PM$_{2.5}$ and geopotential height over time using quantiles.

For univariate data, the natural linear ordering makes it convenient and straightforward to define quantiles. However, the generalization to multivariate data is not trivial at all, let alone functional data. Most of the methods in the literature rely on ranking multivariate or functional data using depth functions (Hettmansperger et al., 1992; Chaudhuri, 1996b; Breckling et al., 2001; Chakraborty, 2001, 2003; Serfling, 2010). For example, Kong and Mizera (2012) use projection-based quantiles to estimate multivariate quantile contours, which coincide with Tukey depth contours (Tukey, 1975); see also Hallin et al. (2010); Paindaveine and Šiman (2012); Liu et al. (2013). For univariate functional data, Serfling and Wijesuriya (2017) use spatial depth to estimate quantile contours. Sun and Genton (2011) defined functional boxplots based on functional band depth (López-Pintado and Romo, 2009). Fraiman and Pateiro-López (2012) defined projection-based quantiles for high dimensional multivariate data and univariate functional data. For multivariate functional data, Kong
et al. (2015) modeled bivariate functional data using directional quantiles, and the estimated quantile contours for the marginal distribution coincides with the Tukey depth contours. However, all of these depth-based contours are convex by nature, which is not desirable for distributions with nonconvex support. These depth-based methods follow linear monotonicity axiom, which imposes the convexity of its contours (Liu, 1990; Zuo and Serfling, 2000). They also suffer from computational issues in higher dimensions, and even computing quantiles for three-dimensional data is very time-consuming and challenging (Zuo, 2019).

Wei (2008) proposed an approach based on stratified quantile regression to estimate bivariate quantile contours for longitudinal data. Their quantile contours can characterize nonconvex distributions; however, their method relies on sequential conditioning of response components, and the order adopted for that conditioning. Recently, Chernozhukov et al. (2017) introduced a concept of multivariate quantiles based on deterministic maps that transform a reference distribution to the distribution of interest. Their approach gives up the linear monotonicity axiom and can pick up nonconvex features of the distribution. Carlier et al. (2016) also proposed a conditional multivariate quantile function and a multivariate quantile regression model using an optimal transport approach. However, multivariate data methods cannot be used for functional data because of the dimensionality issues. To the best knowledge of the authors, there is no work in the current literature on the estimation of quantiles for multivariate functional data. Multivariate functional data are actively generated in various research areas, including meteorology, biology, biomedicine, engineering, and environmental science. The estimated quantile contours for multivariate functional data can be used to understand the joint distribution of the data better and for tasks such as clustering and outlier detection.

To model the multivariate functional data, we consider a cubic spline model in time and represent it as a linear combination of B-spline basis functions. The quan-
tile contours are constructed from the estimated multivariate quantile function for a particular time point, and functional quantile curves are constructed over time for each \( u \), where \( u \) is the multivariate quantile index. The estimated multivariate quantile function follows the property of monotonicity and uniqueness, and consistency has been established. The estimated quantile contour can characterize non-Gaussian and even nonconvex distributions, which the depth-based methods fail to do. The estimated multivariate quantile function is a continuous function of time, for a fixed quantile index, which is useful to capture the change in the distribution over time. Computationally, the proposed method is also efficient for higher dimensions, while existing methods only deal with bivariate functional data.

In the simulation studies, we apply the proposed methods to bivariate and trivariate functional data. We show that in the case of Gaussian marginals distribution, the predicted quantile contours coincide with the density contours, and for nonconvex marginal distribution, the contours can correctly pick up the shape of the distribution, while the depth-based methods failed to do so. The prediction was identified to be reasonably accurate numerically in predicting the center of the distribution. For the application, the bivariate distribution of PM\(_{2.5}\) and geopotential height is estimated using flexible quantile contours for the northeastern United States across six months. The quantile contours explain the nonconvexity in the marginal distribution, and the functional quantile curves capture the dynamic change in the distribution over time.

The rest of the chapter is organized as follows. In Section 4.2, we define the multivariate functional quantile model, discuss the estimation of multivariate functional quantiles and construction of quantile contours. We discuss the theoretical properties of the multivariate quantile function in Section 4.3. In Section 4.4, we conduct simulation studies on bivariate and trivariate functional data. In Section 4.5, we apply our methods to air pollution data, and, finally, Section 4.6 concludes the chapter with a discussion.
4.2 Methodology

4.2.1 Multivariate functional quantile model

Multivariate functional data are realizations of multivariate random processes on a given interval. We consider multivariate functional data, $\mathbf{Y}(t) = (Y_1(t), \ldots, Y_d(t))^\top \in \mathbb{R}^d$, where $t$ belongs to an interval $\mathcal{T}$ in $\mathbb{R}$, and $\mathbf{X} \in \mathbb{R}^p$ as the covariates. Let $\mathbf{U}$ be a latent vector that follows a reference distribution $F_U$, for instance uniform distribution on $(0, 1)^d$. The idea is to create a deterministic map from the random vector $\mathbf{U}$, with reference distribution $F_U$ to $\mathbf{Y}(t)$. We assume that the reference distribution $F_U$ has a probability density function $f_U$ with respect to the Lebesgue measure on $\mathbb{R}^d$ and has convex support on $\mathcal{U}$. Let $(\mathbf{Y}(t), \mathbf{X})$ and $\mathbf{U}$ be defined on the complete probability space $(\Omega, \mathcal{A}, P)$. Denote by $F_{(\mathbf{Y}(t), \mathbf{X})}$ the joint distribution of $(\mathbf{Y}(t), \mathbf{X})$, and $F_{\mathbf{Y}(t)|\mathbf{X}}$ the conditional distribution of $\mathbf{Y}(t)$ given $\mathbf{X}$. Let $\mathcal{U}\mathcal{X}$, and $\mathcal{Y}$ denote the support of $F_{\mathbf{U}|\mathbf{X}}$, and $F_{\mathbf{Y}(t)}$. The multivariate functional quantile model is the following linear regression model given by

$$
\mathbf{Y}(t) = Q_{\mathbf{Y}(t)|\mathbf{X}}(\mathbf{U}, \mathbf{X}) = \beta(\mathbf{U}, t)^\top \mathbf{X}, \quad \mathbf{U} \sim F_U, \quad \mathbb{E}(\mathbf{X} | \mathbf{U}) = \mathbb{E}(\mathbf{X}),
$$

(4.1)

where $\mathbf{u} \mapsto \beta(\mathbf{u}, t)^\top \mathbf{x}$ is a map from $\mathcal{U}$ to $\mathbb{R}^d$ such that it is a gradient of convex function:

$$
\beta(\mathbf{u}, t)^\top \mathbf{x} = \nabla_{\mathbf{u}} B(\mathbf{u}, t)^\top \mathbf{x}, \quad \text{for all } (\mathbf{u}, \mathbf{x}) \in \mathcal{U}\mathcal{X}, \text{ and } t \in \mathcal{T},
$$

(4.2)

where $\mathbf{u} \mapsto B(\mathbf{u}, t)^\top \mathbf{x}$ is a strictly convex map from $\mathcal{U}$ to $\mathbb{R}$. The constraint condition $\mathbb{E}(\mathbf{X} | \mathbf{U}) = \mathbb{E}(\mathbf{X})$ implies a mean independence condition between the latent vector $\mathbf{U}$ and covariates $\mathbf{X}$, which is a weaker condition than independence and follows naturally from the multivariate quantile regression model defined by Carlier et al. (2016). The parameter $\beta(\mathbf{u}, t)$ of the model is a continuous function of $t$, and is
indexed by the multivariate quantile index \( u \in \mathcal{U} \). For a time \( t \), \( \beta(u, t) \) is a \( d \times p \) matrix of regression coefficients, which reduces to a vector of quantile regression coefficients for \( d = 1 \). The map \( u \mapsto \beta(u, t) \top \mathbf{x} \) can be obtained as a gradient of the convex function \( \mathbf{B}(u, t) \top \mathbf{x} \). The proposed multivariate functional quantile model is a nonparametric, time-varying coefficient model where \( \mathbf{B}(u, t) \) are smooth functions of \( t \). The time varying coefficients can be well approximated by basis function expansions (Huang et al., 2002). Let \( \mathbf{H}(t) = (h_1(t), \ldots, h_K(t)) \top \) be a B-spline basis, where \( K \) is the number of B-spline functions, then \( \mathbf{B}(u, t) \) is approximated as \( \mathbf{B}(u, t) = \gamma(u) \mathbf{H}(t) \), where \( \gamma(u) \) is a \( p \times K \) matrix of coefficients. Then the equation (4.2) can be rewritten as

\[
\beta(u, t) \top \mathbf{x} = \nabla_u \{\gamma(u) \mathbf{H}(t)\} \top \mathbf{x}, \quad \text{for all } (u, \mathbf{x}) \in \mathcal{U} \mathcal{X}, \text{ and } t \in \mathcal{T}. \tag{4.3}
\]

The conditional multivariate quantile function, \( Q_{\mathbf{Y}(t)|\mathbf{X}}(u, \mathbf{x}) \), is a monotone map \( u \mapsto \beta(u, t) \top \mathbf{x} \) which is a gradient of a convex function \( \mathbf{B}(u, t) \), that transforms the reference distribution of \( U \) to the distribution of interest \( \mathbf{Y}(t) \) conditional on \( \mathbf{X} \). We approximate the coefficients \( \mathbf{B}(u, t) = (B_1(u, t), \ldots, B_p(u, t)) \top \) by B-splines expansion as they have compact support and are computationally efficient. We have \( B_l(u, t) = \sum_{i=1}^{K} \gamma_{li}(u) h_i(t) \) for \( l = 1, \ldots, p \), where \( h_i(t)(i = 1, \ldots, K) \) is a set of B-spline basis functions. We consider B-splines of cubic splines with a pre-specified set of knots. The number of knots are fixed and typically chosen as suitable quantiles of \( t \). The estimation of the model is not sensitive to the number of knots.

The conditional multivariate quantile function of \( \mathbf{Y}(t) \), denoted by \( Q_{\mathbf{Y}(t)|\mathbf{X}}(u, \mathbf{x}) \), is a continuous function of \( t \), for a fixed \( u \). In the univariate case, \( U \) is interpreted as a notion of rank, and the typical choice of the reference distribution is Uniform(0, 1). In the multivariate case, \( U \) can be interpreted as a multivariate rank of \( \mathbf{Y}(t) \), with a typical choice of Uniform(0, 1)\(^d\), since we create a deterministic map from \( U \) to
In this chapter, we consider only the intercept term in the model to obtain the multivariate quantiles of $Y(t)$. In principle, we can consider more covariates to study the multivariate quantiles of $Y(t)$ conditioned on the covariates. The conditional multivariate quantile function can be interpreted as the optimal transport from the reference distribution $F_U$ to $F_{Y(t)|X}$. There might be other maps $\tilde{Q} : \mathcal{U}X \mapsto \mathcal{Y}$, which transport $F_U$ to $F_{Y(t)|X}$, but $Q$ is the optimal transport such that it has the minimum Wasserstein distance $E\|Q_{Y(t)|X}(U, X) - U\|^2$, where $\|\|$ denotes the Euclidean norm of $\mathbb{R}^d$.

### 4.2.2 Estimation of conditional multivariate functional quantiles

To estimate the conditional multivariate quantile function, we estimate the regression coefficients of the multivariate functional quantile model (4.1) following the approach of Carlier et al. (2016) based on optimal transport. We minimize the Wasserstein distance between $V$ and $Y(t)$ considered over all random vectors $V$ defined on the probability space $(\Omega, \mathcal{F}, P)$ following reference distribution $F_U$, subject to mean independence constraint,

$$
\min_V \{E(\|V - Y(t)\|^2) : E(X | V) = E(X)\}.
$$

This problem is equivalent to the following covariance maximization problem

$$
\max_V \{E(V^T Y(t)) : E(X | V) = E(X)\}. \tag{4.4}
$$

For the estimation of conditional multivariate functional quantiles, we simply need to solve the convex optimization problem (4.4), which conveniently leads to a dual problem. The solution to the optimization problem (4.4) exist and is given by $V = U$, where $U = Q_{Y(t)|X}^{-1}(Y(t), X)$, which follows from Theorem 2.3 of Carlier et al. (2016).
The following paragraph gives the technical details of the implementation in practice.

Multivariate functional data is typically recorded at discrete points. Consider discretized multivariate functional data $Y(t)$, $t = t_1, \ldots, t_r$, with $m$ observations at each time point. Let $Y = \{Y(t_1)^T, \ldots, Y(t_r)^T\}$ be a $mr \times d$ matrix, where $Y(t_i)$ is $m \times d$ matrix with each row as a realization of $Y(t_i)$, $i = 1, \ldots, r$. Here, we consider only the intercept in our model, so $X$ is the $mr \times K$ B-spline matrix. We generate $n$ points for $U$ from the Uniform $(0, 1)^d$ reference distribution. Let $U$ be an $n \times d$ matrix with each row as a realization of $U$. The distributions $F_{YX}$ and $F_U$ can be approximated by discrete distributions $\nu$ and $\mu$. Let $\nu_j$ be the probability attached with $(y_j, x_j)$, $j = 1, \ldots, mr$ and $\nu = (\nu_1, \ldots, \nu_{mr})^T$. Let $\mu_i$ be the probability attached with $u_i$, $i = 1, \ldots, n$, and $\mu = (\mu_1, \ldots, \mu_n)^T$. Let $p_{ij}$ denote the probability attached with $(u_i, x_j, y_j)$, then the objective is to find the $n \times m$ optimal matrix $p = ((p_{ij}))$ such that it maximizes the objective function

$$
\sum_{ij} p_{ij} y_j^T u_i = \text{tr}(U^T p Y),
$$

with respect to constraints $p^T 1_n = \nu$ and $pX = \mu \nu^T X$. This constrained optimization problem is solved using generalized Lagrange function and the primal problem is given by

$$
\max_{\text{vec}(p) \geq 0} \text{vec}(U^T) \text{vec}(p) : \quad \left( I_{mr} \otimes 1_n^T \right) \text{vec}(p) = \nu, \left( X^T \otimes I_n \right) \text{vec}(p) = \text{vec}(\mu \nu^T X).
$$

(4.5)

Using generalized Lagrange function, the dual form can be derived as

$$
\min_{\psi, b} \psi^T \nu + \text{vec}(b)^T \text{vec}(\mu \nu^T X) : \quad \left( I_{mr} \otimes 1_n \right) \psi + \left( X \otimes I_n \right) \text{vec}(b) \geq \text{vec}(U^T), \quad (4.6)
$$

where $\psi$ is $mr \times 1$ vector and $b$ is an $n \times K$ matrix. The dual problem (4.6) was obtained from the primal problem (4.5) using Lagrange multipliers, as follows:
\[
\max_{\text{vec}(p) \geq 0} \text{vec}(U^{\top}Y) \quad \text{vec}(p) \quad \text{is equivalent to} \quad \min_{\text{vec}(p) \geq 0} \quad - \text{vec}(U^{\top}Y) \quad \text{vec}(p) \quad \text{with same constraints}. \quad \text{Consider the Lagrangian function,}
\]

\[
L(p, \psi, b, \lambda) = -\text{vec}(U^{\top}Y) \quad \text{vec}(p) + \text{vec}(\psi)^{\top} \left( (I_{mr} \otimes I_{n}^{\top}) \quad \text{vec}(p) - \nu \right) + \text{vec}(b)^{\top} \left( (X^{\top} \otimes I_{n}) \quad \text{vec}(p) - \text{vec}(\mu \nu^{\top} X) \right) - \lambda^{\top} \quad \text{vec}(p)
\]

Taking partial derivative of the Lagrangian function with respect to \(\text{vec}(p)\) and equating it to zero

\[
\frac{\partial L(p, \psi, b, \lambda)}{\partial \text{vec}(p)} = -\text{vec}(U^{\top}Y) \quad + \left( I_{mr} \otimes I_{n} \right) \psi + \left( X \otimes I_{n} \right) \text{vec}(b) - \lambda = 0
\]

Using dual feasibility: \(\lambda \geq 0\), we have

\[
\left( X \otimes I_{n} \right) \text{vec}(b) - \text{vec}(U^{\top}Y) \quad + \left( I_{mr} \otimes I_{n} \right) \psi \quad \geq \quad 0. \quad (4.7)
\]

Rearranging the Lagrangian function we get,

\[
L(p, \psi, b, \lambda) = -\psi^{\top} \nu - \text{vec}(b)^{\top} \text{vec}(\mu \nu^{\top} X) + \left( -\text{vec}(U^{\top}Y) \quad + \text{vec}(\psi)^{\top} \left( I_{mr} \otimes I_{n}^{\top} \right) + \text{vec}(b)^{\top} \left( X^{\top} \otimes I_{n} \right) \right) \text{vec}(p)
\]

The terms with \(\text{vec}(p)\) are neglected and the dual problem reduces to the minimization of \(-\psi^{\top} \nu - \text{vec}(b)^{\top} \text{vec}(\mu \nu^{\top} X)\) subject to constraint \(\lambda \geq 0\) which is equivalent to maximizing \(\psi^{\top} \nu + \text{vec}(b)^{\top} \text{vec}(\mu \nu^{\top} X)\) subject to constraint (4.7). Hence the dual problem is

\[
\min_{\psi, b} \psi^{\top} \nu + \text{vec}(b)^{\top} \text{vec}(\mu \nu^{\top} X) : \quad \left( I_{mr} \otimes I_{n} \right) \psi + \left( X \otimes I_{n} \right) \text{vec}(b) \quad \geq \text{vec}(U^{\top}Y),
\]
where \( \psi \) is \( mr \times 1 \) vector and \( b \) is a \( n \times K \) matrix. The matrices \( (I_{mr} \otimes 1_n) \psi \) and \( (X \otimes I_n) \text{vec}(b) \) can be very large matrices but they are sparse. This sparsity can be used for computational advantages. The above problems can be solved by linear programming software, such as Gurobi in R.

### 4.2.3 Quantile contour estimation and prediction

The multivariate quantile function can adequately capture the features of a wide range of distributions. In this section, we describe the estimation of quantile contours at time \( t \), and functional quantile curves over time from the predicted multivariate quantile function. The optimal solution of matrix \( b_{n \times K} \) from the dual problem (4.6) gives us the estimated regression coefficients \( b(u, t) = B(u, t) \) and, hence, gives the convex map \( u \mapsto B(u, t)^\top x \). The map \( u \mapsto \beta(u, t)^\top x \) is obtained by taking the gradient of convex function \( B(u, t)^\top x \). The multivariate quantile function in time, \( \hat{Q}_{Y(t)|X}(u, x) \), is then obtained as a convex map from the reference distribution. The components of the multivariate quantile function are given by

\[
\hat{Q}_{Y_1(t)|X}(u_1, \ldots, u_d) = \beta_1(u_1, \ldots, u_d, t)^\top x = \frac{\partial B}{\partial u_1}(u_1, \ldots, u_d, t)^\top x \\
\vdots \\
\hat{Q}_{Y_d(t)|X}(u_1, \ldots, u_d) = \beta_d(u_1, \ldots, u_d, t)^\top x = \frac{\partial B}{\partial u_d}(u_1, \ldots, u_d, t)^\top x.
\]

After obtaining the multivariate quantiles at time \( t \), we estimate the marginal quantile contours. To construct the quantile contours, we first map the reference distribution Uniform \( (0, 1)^d \) to a spherical uniform distribution with an optimal assignment algorithm. For \( d = 2 \), the spherical uniform distribution is generated randomly on circles with evenly spaced radii in \((0,1]\). An assignment problem is a matching problem that minimizes the total cost of the assignment; in our case, the cost is the Euclidean distance. Since the number of points in the reference distribution and
spherical uniform distribution is equal, it is a linear assignment problem. The assignment problem can be solved using algorithms such as the linear sum assignment algorithm (Burkard et al., 2009), Hungarian algorithm (Papadimitriou and Steiglitz, 1982), or an adaptation of simplex algorithm (Murty, 1983).

Once we obtain the map between the reference distribution and the spherical uniform distribution, we compute the norm of each point in the spherical reference distribution. Finally, we construct $\alpha$-hulls of $\hat{Q}_{Y(t)|X}(U \cap S(\tau))$, where $S(\tau) = \{x \in \mathbb{R}^d : \|x\| \leq \tau\}$ for $\tau \in (0, 1]$. The $\alpha$-hull is the $\tau$-quantile region, and the boundary of the $\alpha$-hull is the $\tau$-quantile contour. The algorithm of $\alpha$-hull is based on Delaunay triangulation and is explained in Edelsbrunner et al. (1983), and the boundary of $\alpha$-hulls are formed by the arc of open balls of radius $\alpha$. The proposed quantile contours are not restricted with convexity and work well for non-Gaussian, and even nonconvex distributions, without making any distributional assumptions. The functional quantile curves are also estimated using the multivariate quantile function to study the pattern over time. For a fixed $u$, the multivariate quantile function, $\hat{Q}_{Y(t)|X}(u, x)$, is a continuous function of $t$. The functional quantile curves are constructed by $\hat{Q}_{Y(t)|X}(u, x)$ over $t$, for each fixed $u$. The median functional quantile curve is also estimated to summarize the pattern over time, which is $\hat{Q}_{Y(t)|X}(u_m, x)$ over $t$, where $u_m$ is the median of the reference distribution. However, if we had neglected the functional nature of data and modeled the multivariate quantile function without the spline representation, then the estimated quantile contours would be rougher for each time point, and it would not be able to account for changes over time. Here, the proposed method borrows strength from other time points, and the functional quantile curves are well defined over time.
4.3 Theoretical properties

The conditional multivariate functional quantile has some desired properties that are analogous to the classical conditional quantile function:

Property 1. (Monotonicity) The conditional multivariate functional quantile is a gradient of a convex function. The gradient of a convex function is monotone (Minty, 1964). Therefore, the conditional multivariate quantile function follows monotonicity property with respect to $\mathbf{u} \in \mathcal{U}$:

$$\{Q_{Y(t)|X}(\mathbf{u}, x) - Q_{Y(t)|X}(\bar{\mathbf{u}}, x)\}^\top (\mathbf{u} - \bar{\mathbf{u}}) \geq 0,$$

for all $\mathbf{u}, \bar{\mathbf{u}} \in \mathcal{U}, t \in \mathcal{T}$.

Monotonicity is an evident feature of univariate quantile functions, and it is more complicated to define monotonicity properties in the multivariate case. The conditional multivariate functional quantile preserves the monotonicity property of the univariate quantile function. It implies that increasing a particular variable in $\mathcal{U}$ while keeping others fixed leads to an increase in the corresponding variable in $Y(t)$.

Property 2. (Uniqueness) The conditional multivariate quantile function $Q_{Y(t)|X}(\mathbf{u}, x)$ is a unique gradient of convex function, $\mathcal{F}_U$-almost everywhere. The gradient $\nabla_u B(\mathbf{u}, t)^\top x$ exists $\mathcal{F}_U$-almost everywhere, i.e., the set of points where it does not exist is negligible.

This property implies that the map $\mathbf{u} \mapsto Q_{Y(t)|X}(\mathbf{u}, x)$ that transforms the latent random vector $\mathbf{U}$ with distribution $\mathcal{F}_U$ into $Y(t)$ exists and is unique, such that $Y(t)$ conditional $X$ has distribution $F_{Y(t)|X}$. This map is unique in the sense that if there is any other map $\mathbf{u} \mapsto \tilde{Q}_{Y(t)|X}(\mathbf{u}, x)$ with the same properties of $Q$, then $\tilde{Q} = Q$.

The above properties immediately follow from the properties of the multivariate conditional quantile function defined by Carlier et al. (2016).

Theorem 1. (Uniform convergence) Let $\mathcal{U}$ denote the compact subset of $\mathbb{R}^d$ and the support of the reference distribution $\mathcal{F}_U$. Let $\hat{Q}_{Y(t)|X}(\mathbf{u})$ be the predicted conditional
multivariate quantile function on $\mathcal{U}$, then as $n \to \infty$

$$\sup_{u \in \mathcal{U}} \| \hat{Q}_{Y(t)|X}(u) - Q_{Y(t)|X}(u) \| \to 0.$$  

$\hat{Q}_{Y(t)|X}(u)$ uniformly converges to $Q_{Y(t)|X}(u)$ $F_U$-almost everywhere, i.e., the set where it does not converge is negligible.

**Proof.** Consider the sequence $\hat{Q}_n = \hat{Q}_{Y(t)|X}(u) = \nabla_{u_n} B(u_n, t)^\top x$, we want to prove that as $n \to \infty$, $\hat{Q}_n$ converges uniformly to $Q = \nabla_u B(u, t)^\top x$ for all $(u, x) \in \mathcal{U} \times T$.

**Lemma 1.** Let $D$ and $E$ be compact sets and $g : D \to E$ is continuous. Then for any convergent sequence $x_n \to x$ in $D$, the sequence of functions $g_n$ converges uniformly to $g$ if and only if $g_n(x_n) \to g(x)$.

Lemma 1 is referred to as the extended continuous mapping theorem, for the proof see Theorem B.3 of Bücher et al. (2014). Using this lemma, if $g_n = g$, then $g_n \to g$ trivially and hence, $g(x_n) \to g(x)$ for any convergent sequence $x_n \to x$.

Hence, for any convergent sequence $\{u_n\}$ such that $u_n \to u$, $u \in \mathcal{U}$, using lemma 1 we have

$$B(u_n, t)^\top x \to B(u, t)^\top x.$$  

**Lemma 2.** Consider a sequence $\{g_n\}$, such that $\lim_{n \to \infty} g_n(x_0)$ exists and is finite, $x_0 \in [a, b]$ and the sequence $\{\frac{\partial g_n}{\partial x}\}$ converges uniformly on $[a, b]$, then $g_n$ uniformly converges to $g$ and $\frac{\partial g_n}{\partial x}$ converges to $\frac{\partial g}{\partial x}$.

For the proof of lemma 2, see Theorem 7.17 of Rudin (1976). We denote $\nabla_{u_n} B(u_n, t)^\top x = \left( \frac{\partial}{\partial u_{n1}} B(u_n, t)^\top x, \ldots, \frac{\partial}{\partial u_{nN}} B(u_n, t)^\top x \right)$. Since $\nabla_{u_n} B(u_n, t)^\top x$ exists $F_U$-almost everywhere on $\mathcal{U}$, the set $\mathcal{U}$ where it does not exist is negligible. Therefore, the sequences $\frac{\partial}{\partial u_{1n}} B(u_n, t)^\top x, \ldots, \frac{\partial}{\partial u_{Nn}} B(u_n, t)^\top x$ converges $F_U$-almost everywhere on $\mathcal{U}$. Hence,
We have $d$ sequences that converge uniformly; we now want to prove that they converge jointly. Consider $d$ sequences $g_{1n}, \ldots, g_{dn}$ that uniformly converge to $g_1, \ldots, g_d$, respectively. From the definition of uniform convergence, we have the following:

For every $\epsilon_1 > 0$, there exists a natural number $N(\epsilon_1)$ such that $|g_{1n} - g_1| < \epsilon_1$, for all $n \geq N(\epsilon_1)$. Similarly, for every $\epsilon_k > 0$, there exists $N(\epsilon_k)$ such that $|g_{dn} - g_d| < \epsilon_d$, for all $n \geq N(\epsilon_d)$.

Consider

$$
\| (g_{1n}, \ldots, g_{dn}) - (g_1, \ldots, g_d) \| = \sqrt{(g_{1n} - g_1)^2 + \cdots + (g_{dn} - g_d)^2} 
$$

$$
\leq \sqrt{\epsilon_1^2 + \cdots + \epsilon_d^2} = \epsilon^*, \text{ for all } n \geq N = \max(N(\epsilon_1), \ldots, N(\epsilon_d)).
$$

For every $\epsilon^* > 0$, there exists a natural number $N = \max(N(\epsilon_1), \ldots, N(\epsilon_d))$ such that

$$
\| (g_{1n}, \ldots, g_{dn}) - (g_1, \ldots, g_d) \| < \epsilon^*, \text{ for all } n > N.
$$

Therefore,

$$(g_{1n}, \ldots, g_{dn}) \to (g_1, \ldots, g_d).$$

Using this property of uniform convergence, we show that

$$
\left( \frac{\partial}{\partial u_{1n}} B(u_{n}, t)^\top x, \ldots, \frac{\partial}{\partial u_{dn}} B(u_{n}, t)^\top x \right) \to \left( \frac{\partial}{\partial u_1} B(u, t)^\top x, \ldots, \frac{\partial}{\partial u_d} B(u, t)^\top x \right).
$$
Hence, we conclude that

$$\hat{Q}_{Y(t)|X}(u) \to Q_{Y(t)|X}(u).$$

This implies, as $n \to \infty$

$$\sup_{u \in U} \|\hat{Q}_{Y(t)|X}(u) - Q_{Y(t)|X}(u)\| \to 0,$$

i.e., $\hat{Q}_{Y(t)|X}(u)$ uniformly converges to $Q_{Y(t)|X}(u)$ for $u \in U$, $F_U$-almost everywhere.

\[\square\]

**Corollary 1.** Suppose $A$ is any non-empty compact subset of $U$, and denoting

$\hat{Q}(A) = \{\hat{Q}_{Y(t)|X}(u); u \in A\}$ and $Q(A) = \{Q_{Y(t)|X}(u); u \in A\}$, then as $n \to \infty$

$$\sup_{A \subseteq U} d_H(\hat{Q}(A), Q(A)) \to 0,$$

where $d_H$ is the Hausdorff distance.

**Proof.** The Hausdorff distance $d_H$ between two non-empty subsets $X$ and $Y$ of a Euclidean metric space is given by

$$d_H(X, Y) = \max\{\sup_{x \in X} \inf_{y \in Y} d(x, y), \sup_{y \in Y} \inf_{x \in X} d(x, y)\},$$

where $d$ is the euclidean distance. Let $A$ be a non empty subset of $U$, consider the
Hausdorff distance between $\hat{Q}(A)$ and $Q(A)$,

$$d_H(\hat{Q}(A), Q(A)) = \sup_{A \subseteq U} \left( \max \left\{ \sup_{u \in A} \inf_{u^* \in A} d(\hat{Q}(u), Q(u^*)), \sup_{u^* \in A} d(\hat{Q}(u^*), Q(u)) \right\} \right)$$

$$= \sup_{A \subseteq U} \left( \max \left\{ \sup_{u \in A} \inf_{u^* \in A} \| \hat{Q}(u) - Q(u^*) \|, \sup_{u^* \in A} \| \hat{Q}(u) - Q(u^*) \| \right\} \right)$$

$$\leq \sup_{A \subseteq U} \left( \max \left\{ \sup_{u \in A} \| \hat{Q}(u) - Q(u) \|, \sup_{u^* \in A} \| \hat{Q}(u^*) - Q(u^*) \| \right\} \right)$$

$$= \sup_{u \in U} \| \hat{Q}(u) - Q(u) \|$$

$$= \sup_{u \in U} \| \hat{Q}(u) - Q(u) \| \rightarrow 0, \quad \text{as } n \rightarrow \infty \quad \text{(Using Theorem 1)}.$$

This implies, as $n \rightarrow \infty$

$$\sup_{A \subseteq U} d_H(\hat{Q}(A), Q(A)) \rightarrow 0.$$

Hence, proved.

\[\square\]

Theorem 1 establishes the uniform consistency of the predicted multivariate quantile function $\hat{Q}_{Y \mid X}(u)$ to $Q_{Y \mid X}(u)$ $F_U$-almost everywhere. It implies that as the sample size of the reference distribution increases, the predicted conditional quantile function uniformly converges to the theoretical multivariate quantile function. The set $Q(A)$ with $A = (U \cap S(\tau))$ and $\hat{A} = (U \cap \hat{S}(\tau))$ is the $\tau$-quantile region and $\tau$-quantile contours, respectively, where $S(\tau) = \{x \in \mathbb{R}^d : \|x\| = \tau\}$. Therefore, Corollary 1 establishes the consistency of the predicted quantile region and contours to the theoretical quantile region and contours.
4.4 Simulation study

4.4.1 Bivariate functional data with convex margins

To illustrate our methods, we first conduct simulation studies by generating bivariate functional data with convex marginal distributions. Here, we simulate bivariate functional data \( Y(t) = (Y_1(t), Y_2(t))^\top \) with bivariate normal error distribution using the following model

\[
\begin{pmatrix}
Y_1(t) \\
Y_2(t)
\end{pmatrix} = \begin{pmatrix}
X_1^\top \theta_1(t) \\
X_2^\top \theta_2(t)
\end{pmatrix} + \begin{pmatrix}
\epsilon_1(t) \\
\epsilon_2(t)
\end{pmatrix}, \quad \text{where} \quad \begin{pmatrix}
\epsilon_1(t) \\
\epsilon_2(t)
\end{pmatrix} \sim N\left( \begin{pmatrix} 5 \\ 5 \end{pmatrix}, \begin{pmatrix} 0.1 & 0 \\ 0 & 0.1 \end{pmatrix} \right),
\]

and \( X_1^\top \theta_1(t) \) and \( X_2^\top \theta_2(t) \) represents cubic spline functions in \( t \). We represent the spline functions as a linear combinations of B-spline basis functions with the number of knots equal to 3. The two sets of coefficients are \( (1, 0.7, 0.9, 0.7, 0.8, 0.9) \) and \( (1, 0.8, 0.7, 0.8, 0.9, 0.7, 0.6) \), respectively. The data was generated for \( m = 300 \) and 10 equispaced points in \([0,1]\).

We then use the model (4.1) to estimate the conditional multivariate quantile function with reference distribution \( U(0,1)^2 \). We generate the points of the reference distributions from a \( 21 \times 21 \) uniformly spaced grid of points with \( n = 441 \). We then solve the linear programming problem (4.6) using Gurobi in R, and estimate the conditional multivariate quantile function.

The quantile contours are estimated for the marginal distributions of the simulated data, and Figure 4.1a shows the estimated quantile contours for \( t = 0.48 \), for three quantile levels, 0.25, 0.5, and 0.75. In the data generating model, the errors are simulated from a bivariate normal distribution, hence the marginal distribution is
known and given by

\[
\begin{pmatrix}
Y_1(t) \\
Y_2(t)
\end{pmatrix}
\sim
\mathcal{N}
\left[
\begin{pmatrix}
X_1^\top \theta_1(t) + 5 \\
X_2^\top \theta_2(t) + 5
\end{pmatrix},
\begin{pmatrix}
0.1 & 0 \\
0 & 0.1
\end{pmatrix}
\right].
\]

The density contours of the bivariate normal distribution are also plotted for \( t = 0.48 \) (Figure 4.1a) at the levels 0.25, 0.5, and 0.75. It can be observed that the estimated quantile contours coincide with the respective density contours of the bivariate normal distribution. The functional quantile curves are also estimated over time and are plotted in Figure 4.1b along with the true mean function of the simulated data. The estimated quantile curves pick up the pattern of the true mean function well. The median functional quantile curve predicts the center of the functional distribution, while the curves at all the multivariate quantiles predict the complete functional distribution.

To assess the accuracy of the prediction model, we compare the predicted median quantile curve with the true mean function by repeating the simulation and computing the mean square prediction error (MSPE). The above simulation study is repeated 100 times, and the MSPE is calculated in the following way

\[
\text{MSPE} = \frac{1}{100} \sum_{j=1}^{100} D_j(f(t), f_M(t)); \quad D_j(f(t), f_M(t)) = \frac{1}{p} \sum_{i=1}^{p} d(f(t_i), f_M(t_i)),
\]

where \( d \) is the euclidean distance, and \( f(t_i) \) and \( f_M(t_i), \ i = 1, \ldots, p \) are the true mean curves and predicted median curves, respectively. The obtained MSPE value is 0.1532 which shows that the prediction for the center of functional distribution is reasonably accurate. The predicted median quantile curves for all the simulations, plotted with the true mean function in Figure 4.1c, are quite close to the true center of the distribution. They correctly pick up the pattern of the mean function over time.

To deal with the uncertainty of the error distribution and estimation of contours, we
Figure 4.1: (a) Predicted contours for simulated data at $t = 0.48$ for bivariate functional data with convex margins are in solid black for quantile levels 0.25, 0.5, and 0.75, with the density contours of the true bivariate normal distribution overlayed with colored lines for quantiles level 0.25 (red), 0.5 (blue), 0.75 (green); and (b) predicted functional quantile curves for the 441 points of the reference distribution $U(0, 1)^2$ are shown in grey, and the true mean function is overlayed in red; (c) predicted median curves to estimate the center of the distribution for 100 simulations are in black, and the true center is in red.

average the contours over the 100 simulations. The averaged predicted contours for six quantile levels (0.1, 0.26, 0.42, 0.58, 0.74, 0.9) at $t = 0.48$ are plotted in Figure 4.2a. The averaged contours are obtained by constructing the $\alpha$-hull of the union of all sets of points of the predicted distribution from all the simulations.

We compare our method with the bivariate quantile varying coefficient model proposed by Kong et al. (2015). We estimate the directional quantile envelopes using Kong et al.’s method for the marginal distributions of the simulated data; the resulting quantile envelopes are plotted for $t = 0.48$, at six depth levels, in Figure 4.2b.
Figure 4.2: Predicted contours for bivariate functional data with convex margins at \( t = 0.48 \) using (a) multivariate functional quantile model (4.1) in the chapter for six quantile levels (0.1, 0.26, 0.42, 0.58, 0.74, 0.9) averaged over 100 simulations, and (b) Kong et al.’s method for six depth levels (0.01, 0.09, 0.17, 0.24, 0.32, 0.4).

The resulting quantile envelopes are convex and work well in the case of the convex marginal distribution.

4.4.2 Bivariate functional data with nonconvex margins

In this simulation study, we simulate bivariate functional data \( \mathbf{Y}(t) = (Y_1(t), Y_2(t))^\top \) with nonconvex error distribution using the following model

\[
\begin{pmatrix}
Y_1(t) \\
Y_2(t)
\end{pmatrix} = \begin{pmatrix}
\mathbf{X}_1^\top \mathbf{\theta}_1(t) \\
\mathbf{X}_2^\top \mathbf{\theta}_2(t)
\end{pmatrix} + \begin{pmatrix}
Z + R\cos\phi \\
Z^2 + R\sin\phi
\end{pmatrix},
\]

where \( Z \sim U[-0.5, 0.5], \phi \sim U[0, 2\pi], Z_0 \sim U[0, 0.5] \) are independent and \( R = 0.2Z_0(0.5 + (0.5 - |Z|)/2) \). In the data generating model, \( \mathbf{X}_1^\top \mathbf{\theta}_1(t) \) and \( \mathbf{X}_2^\top \mathbf{\theta}_2(t) \) represents cubic spline functions in \( t \) with fixed number of knots. The setting of the mean function is similar to the previous simulation model. Data are generated for \( m = 300 \) and 10 equispaced points in \([0,1]\).

Quantile contours are estimated for the nonconvex marginal distributions, and Figure 4.3a shows the estimated quantile contours for the marginal distribution, at
Figure 4.3: Predicted contours for bivariate functional data with nonconvex margins at \( t = 0.48 \) using (a) multivariate functional quantile model (4.1) for six quantile levels (0.1, 0.26, 0.42, 0.58, 0.74, 0.9), and (b) Kong et al.’s method for six depth levels (0.01, 0.07, 0.13, 0.18, 0.24, 0.3).

\( t = 0.48 \), for six quantile levels (0.1, 0.26, 0.42, 0.58, 0.74, 0.9). It can be seen that the estimated contours pick up the geometry of the distribution well and are flexible with respect to the nonconvex shape of the distribution. We estimate the directional quantile envelopes using Kong et al.’s method, and the same is plotted for \( t = 0.48 \) at six depth levels in Figure 4.3b. It can be observed that the quantile envelopes, which are essentially Tukey depth contours, cannot pick up the geometry of the distribution. This observation is common to most depth-based contours that follow the characteristics of convexity. For our method, the functional quantile curves are also estimated over time and plotted with the true mean function in Figure 4.4. The pattern of the estimated quantile curves is quite similar to the true mean function.

4.4.3 Trivariate functional data with nonconvex margins

One of the advantages of the proposed multivariate functional quantile model is that it is computationally efficient for higher dimensions. In this section, we simulate trivariate functional data \( \mathbf{Y}(t) = (Y_1(t), Y_2(t), Y_3(t))^\top \) with a nonconvex marginal distribution, using the following model
Figure 4.4: Predicted functional quantile curves for bivariate functional data with nonconvex margins for the 441 points of the reference distribution $U(0,1)^2$ are shown in grey and the true mean function is overlayed in red.

$$
\begin{pmatrix}
Y_1(t) \\
Y_2(t) \\
Y_3(t)
\end{pmatrix}
= \begin{pmatrix}
X_1^\top \theta_1(t) \\
X_2^\top \theta_2(t) \\
X_3^\top \theta_3(t)
\end{pmatrix}
+ \begin{pmatrix}
Z + R\cos\phi \\
Z^2 + R\sin\phi \\
Z + R
\end{pmatrix},
$$

with similar notations as Section 4.4.2. We simulate data for $m = 300$, for 10 equispaced points in $[0,1]$. We use model (4.1) to estimate the multivariate quantile function. We choose the reference distribution to be $U(0,1)^3$. The reference distribution is generated on a $9 \times 9 \times 9$ uniformly spaced grid in with $n = 729$, and mapped to the spherical uniform reference distribution that was generated on spheres with evenly spaced radii in $(0,1]$. The multivariate quantiles are estimated for the marginal distributions of the simulated functional data. The quantile shapes are plotted to visualize the geometry of the trivariate distribution, at $t = 0.48$, for six quantile level, $(0.1, 0.26, 0.42, 0.58, 0.74, 0.9)$ in Figure 4.5. The quantile shapes can correctly obtain the geometry of the distribution in the trivariate case as well. To evaluate the performance of our method, we compare the center of the trivariate functional distribution, which is known, with the predicted center, using the predicted median quantile curve, and compute the MSPE, using (4.8). The simulation is repeated 100 times, and the value we obtain for MSPE is 0.2641, which shows that the prediction
is reasonably accurate.

4.5 Applications to air pollution data

We now illustrate the flexibility of our proposed methodology with an application to air pollution data. We consider air quality data across the Northeastern United States region (Karl and Koss, 1984) defined by the National Climatic Data Center (New Hampshire, Vermont, New York, Massachusetts, Connecticut, Rhode Island, Pennsylvania, New Jersey, Delaware) for the years 2011 and 2012. In particular, we study the joint distribution of PM$_{2.5}$ and geopotential height at 850 hPa over different months. The data for PM$_{2.5}$ is obtained from the Environmental Protection Agency (EPA), which provides a daily average of values generated from a Community Multiscale Air Quality Modeling System (CMAQ, https://www.epa.gov/cmaq). The data for geopotential height is obtained from the North American Regional Reanalysis (NARR, https://www.esrl.noaa.gov/psd), which provides monthly average
values. We average the daily PM$_{2.5}$ concentrations with each month to comply with monthly geopotential height data. Early work on estimating multivariate marginal distribution relied heavily on parametric distributional assumptions or was restricted by the convexity of its contours, but recent research has emphasized the need for more flexible models to analyze such datasets.

Here, we apply the proposed methods to the bivariate distribution of the monthly averages of PM$_{2.5}$ and geopotential height at 850 hPa, across the various Northeastern states mentioned above, for two seasons, October 2011 to March 2012, and April 2012 to November 2012, and estimate the quantile contours for each month. The estimated quantile contours for October 2011 to March 2012 are shown in Figure 4.6a. The distribution of PM$_{2.5}$ and geopotential height at 850 hPa for most of the months is nonconvex, and the contours estimate the geometry of the distribution reasonably well. The estimated marginal distribution also explains the variability of the distribution with the contours and provides valuable summary statistics such as the center of the distribution by estimating the bivariate median. From Figure 4.6a, we can observe that in February 2012, the increase in the PM$_{2.5}$ concentration with the
Figure 4.7: (a) Predicted quantile contours for six quantile levels, (0.1, 0.26, 0.42, 0.58, 0.74, 0.9), and (b) predicted functional quantile curves over time for the distribution of PM$_{2.5}$ and geopotential height at 850 hPa from April 2012 to November 2012.

The increase in geopotential height is not linear, and the median of the distribution can be seen as the change point in geopotential height from where the PM$_{2.5}$ concentration gets much higher. The results for the second season, April 2012 to November 2012, are plotted in Figure 4.7. In the second season, many months have a bimodal distribution, which is clearly shown by the estimated contours, whereas most existing depth-based contours embodied with convexity cannot pick up such features.

Typically, univariate quantile curves are used to study the changes of a variable over time, but these methods aim at screening one variable at a time. Because PM$_{2.5}$ is significantly affected by geopotential height, the overall change should be assessed by studying bivariate quantile curves since univariate quantile curves fail to take correlations into account. The estimated bivariate functional quantile curves for the two seasons, October 2011 to March 2012, and April 2012 to November 2012, are shown in Figures 4.6b and 4.7b, respectively. The functional quantile curves show the changes in the bivariate distribution of PM$_{2.5}$ and geopotential height, over time, for the whole distribution. We denote the quantile curve corresponding to the median of the reference distribution as the median quantile curve. The median
curve summarizes the dynamic change in the distribution of PM$_{2.5}$ and geopotential height from October 2011 to March 2012. The median curves for the two seasons show different patterns of the bivariate distribution over time. The visualization of bivariate functional data is not trivial, and the estimated functional quantile curves visualize the centrality of curves and variability in the curves over the seasons, and the estimated contours characterize the marginal distributions for each month.

4.6 Discussion

This chapter proposes a flexible model to estimate quantiles for multivariate functional data and develop the estimation and prediction procedures. The estimated quantile contours can characterize non-Gaussian and even nonconvex multivariate distributions marginally without any parametric distributional assumptions. The estimated multivariate quantile function has desirable theoretical properties like monotonicity, uniqueness, and consistency. We also show that the method is computationally efficient for high dimensional data (more than 2). The proposed methodology is demonstrated on PM$_{2.5}$ and geopotential height data over time across the Northeastern United States. The estimated contours completely characterize the marginal distributions, and the bivariate quantile curves capture the dynamic change over time. The R code used for this research and links to the datasets used in the application has been made publicly available on https://github.com/agarwag/Flexible-quantile-contours-multivariate-functional-data.
Chapter 5

Copula-based Multiple Indicator Kriging for non-Gaussian Random Fields

5.1 Introduction

An important goal in geostatistics is to obtain predictions from the observed data at unsampled spatial locations. Gaussian random fields are extensively used in the analysis of spatial data due to their simplicity. The classical geostatistical tool, kriging, is the best linear unbiased predictor for Gaussian processes (Cressie, 1993). However, Gaussianity is a strong assumption and is rarely met in practice. Hence, kriging is not optimal for non-Gaussian data, e.g., skewed, heavy-tailed, or categorical data. The need to make accurate predictions from the observed data is common to many scientific disciplines. Spatial data is actively generated in various research areas including environmental science, earth science, ecological and hydrological sciences, and it often displays the presence of non-Gaussianity, for example, data from wind speed (Zhu and Genton, 2012), precipitation (Marchenko and Genton, 2010), heavy metal concentrations in soil (Lin et al., 2010), groundwater pollution (Arslan, 2012).

As a motivating application, we consider precipitation data from the European Climate Assessment & Dataset project (Klein Tank et al., 2002), publically available for research at https://www.ecad.eu. We study the average daily precipitation intensities in millimeters (mm) for November 2019 from 169 meteorological stations in Spain. The stations are sparsely located, and it is of great interest to predict the precipitation at new locations along with uncertainty and create probability exceedance
maps for high precipitation intensities, which can help reveal climate patterns. However, the marginal distribution of precipitation data is usually far from Gaussian distribution (Allcroft and Glasbey, 2003; Johns et al., 2003). Typically, precipitation data is positively skewed and has potential outliers. Hence, it is important to build flexible geostatistical models without distributional assumptions to make accurate predictions as well as explain the spatial variations in precipitation.

Recently, many approaches have been considered to model non-Gaussian geostatistical data by using, e.g., trans-Gaussian random fields (Cressie, 1993; De Oliveira et al., 1997), scale-mixing Gaussian random fields (Fonseca and Steel, 2011), multiple indicator kriging (Journel, 1983; Journel and Alabert, 1989), skew Gaussian processes (Zhang and El-Shaarawi, 2010; Rimstad and Omre, 2014). It is a standard approach in geostatistics to find a nonlinear transformation that enables to fit Gaussian processes; however, transformations do not guarantee normality and may change the features of the process. In particular, logarithm and square root transformations induce a link between the mean and covariance of data in the untransformed scale, resulting in nonstationarity (Wallin and Bolin, 2015). Of the previously mentioned approaches, multiple indicator kriging offers a nonlinear, nonparametric solution to spatial interpolation without any distributional assumptions on the marginal distribution. Indicator kriging has been previously used to study grade estimation in minerals (Vann and Guibal, 1998; Badel et al., 2011), soil or groundwater contamination (Juang and Lee, 2000; Van Meirvenne and Goovaerts, 2001; Goovaerts et al., 2005), spatial variation and mapping of precipitation (Haberlandt, 2007). The indicator approach involves taking a nonlinear indicator transform of the random field to create sets of indicator variables by thresholding the random field at several quantile levels. We formulate the spatial prediction problem as indicator cokriging. The proposed procedure can account for asymmetric spatial continuity of different quantiles of the distribution, since the correlation of low values may be significantly different
from the correlation of high values, which cannot be accommodated by Gaussian-type modeling techniques.

Typically, the spatial dependence structure of the underlying random field is modeled using variograms. However, variograms are strongly influenced by the marginal distribution of the random field, and the variogram estimates are adversely affected by outliers. Moreover, a variogram just describes the mean dependence between variables and not the dependence over the whole range of the distribution. Bárdossy (2006) proposed using copulas to model spatial variability, which circumvents the disadvantages of variograms. Copulas are standardized multivariate distribution functions with uniformly distributed margins, and can describe the dependence structure of multivariate distribution independently of their marginal distributions. Copulas generalize the concept of variograms for spatial modeling and can describe the dependence over the whole distribution. There are several applications of copulas in geostatistics, e.g., financial applications in pricing and credit risk analysis (Cherubini et al., 2004), estimation of groundwater quality parameters (Bárdossy and Li, 2008), modeling uncertainty of precipitation estimates (AghaKouchak et al., 2010), where the usual Gaussian dependence was not appropriate. Copulas are also used to describe the dependence between extremes (Salvadori and De Michele, 2013; Gräler, 2014). While existing methods for multiple indicator kriging use variograms to describe spatial dependence, we propose to use copulas to model the dependence structure of the indicator variables and use it for spatial interpolation. To model the spatial dependence of indicator variables using copulas, we derive the relationship between indicator covariances and copula functions and further develop estimation and prediction procedures.

One of the challenges in the spatial interpolation of non-Gaussian random fields is to predict the whole conditional distribution at unknown spatial locations. While point prediction just predicts a value, probabilistic prediction aims at predicting the
probability distribution of some uncertain quantity, which is more informative and valuable (Gneiting et al., 2007; Gneiting and Katzfuss, 2014). Another challenge of analyzing non-Gaussian data is the presence of potential extreme observations. If the model is not resistant to extreme observations, then it can create significant impacts on the estimation of parameters and prediction. Therefore, it is of great importance to build a flexible model that predicts the complete predictive distribution for non-Gaussian processes and is robust to outliers.

This chapter proposes a new method for spatial probabilistic prediction for non-Gaussian random fields using copulas. We propose a copula-based multiple indicator kriging (CMIK) model and a semiparametric maximum pseudo-likelihood estimator. The proposed model makes no distributional assumptions on the marginal distribution of the random field, therefore increasing the model’s flexibility. We describe the spatial dependence structure using copula functions by exploiting the relationship between the indicator covariances and copulas, which overcomes the disadvantages of variogram modeling. The proposed method provides a complete solution to the spatial prediction problem by predicting the local cumulative distribution function (CDF) at each unsampled location. We can compute the conditional quantiles from the conditional distribution function, and use the conditional median for point prediction and conditional tail quantiles to construct prediction intervals. We consider both Gaussian and non-Gaussian copulas to model the spatial dependence and evaluate our method’s predictive performance through simulation studies and applications to precipitation data and heavy metal concentrations in soil data. We show that the proposed methods are flexible to model non-Gaussian data and perform better than the existing variogram multiple indicator kriging approaches and those based on Gaussianity assumptions.

The rest of the chapter is organized as follows. In Section 5.2, we introduce the CMIK model framework and spatial modeling of indicator covariances using copulas.
We discuss the estimation of copulas and describe the dependence structure using copulas. We also present the spatial probabilistic prediction procedure by predicting the conditional CDF at unsampled locations. In Section 5.3, we simulate non-Gaussian random fields and evaluate the performance of the proposed method for point and probabilistic predictions. In Section 5.4, we consider two applications of our methodology, first, to the precipitation dataset from Spain, and second to the heavy metal dataset along the river Meuse. Finally, Section 5.5 contains discussions and conclusions.

5.2 Methodology

In this section, we propose flexible geostatistical models for spatial probabilistic prediction of non-Gaussian random fields. We formulate the spatial prediction problem as indicator cokriging and develop new estimation and prediction procedures. The existing methods of multiple indicator kriging use variograms to describe spatial dependence of the underlying random field. However, variograms are strongly influenced by the marginal distribution of the random field and are often susceptible to poor modeling. We propose a copula-based multiple indicator kriging model and describe the spatial dependence using copulas by exploiting the relationship between indicator covariances and copula functions. The proposed model predicts the conditional distribution function at each unsampled location, which allows us to quantify uncertainties of the estimates. The proposed method helps us to predict conditional quantiles at unknown locations, such as conditional median, which is more suitable for point prediction of skewed and heavy-tailed distributions, and conditional tail quantiles, which can be used to construct prediction intervals.
5.2.1 Copula-based multiple indicator kriging

Let $Z(s), s \in \mathbb{R}^d, d \geq 1$, be a random field observed at locations $s_1, \ldots, s_n$, and we want to predict $Z$ at an unknown location $s_0$. Here, our objective is to estimate the predictive distribution of $Z(s_0) \mid Z(s_1), \ldots, Z(s_n)$ for non-Gaussian random fields. We aim to predict the conditional CDF or conditional quantile function at a new location. The prediction of conditional quantiles provides a convenient way to construct point predictions, such as median, and use conditional tail quantiles to construct prediction intervals at a new location. In some applications, tail behaviors are of more interest; therefore, the prediction of conditional CDF or conditional quantile function is essential. The proposed method provides a probabilistic solution to the problem of spatial interpolation at unsampled locations, whereas existing methods focus on point prediction, e.g., Gaussian kriging, or rely on stronger distributional assumptions for probabilistic prediction.

We formulate the quantile spatial prediction problem as indicator cokriging. We first create $k$ sets of indicator variables by thresholding $Z(s_1), \ldots, Z(s_n)$ with each quantile $q \in (q_1, q_2, \ldots, q_k), q_1 < \ldots < q_k$,

$$Y_j(s_i) = I\{Z(s_i) \leq q_j\} = \begin{cases} 1, & Z(s_i) \leq q_j \\ 0, & Z(s_i) > q_j \end{cases}, \quad j = 1, \ldots, k.$$

Then, we treat them as multivariate binary spatial fields and apply the cokriging for spatial interpolation. The indicator approach offers a nonparametric, distribution-free solution to the problem of local estimation at unsampled locations. It provides an estimate of the local CDF, which allows uncertainty quantification. Theoretically, indicator cokriging is the best estimator of the predictive distribution function in the least-squares sense (Journel and Alabert, 1989). In indicator cokriging, each indicator variable is estimated using the values of all indicator variables and their associated
Let us denote the indicator covariance between variable \( Y_i \) at location \( s_u \) and variable \( Y_j \) at location \( s_v \) by \( C_{ij}(s_u, s_v) = \text{Cov}\{Y_i(s_u), Y_j(s_v)\} \). We propose to predict the vector \( Y(s_0) = \{Y_1(s_0), \ldots, Y_k(s_0)\}^T \) simultaneously by

\[
\hat{Y}_j(s_0) = \sum_{i=1}^{n} \sum_{j=1}^{k} \lambda_{ij} Y_j(s_i), j = 1, \ldots, k,
\]

where the weights \( \lambda_{ij} \)'s are evaluated by

\[
\begin{bmatrix}
\lambda_{11} \\
\vdots \\
\lambda_{n1} \\
\lambda_{12} \\
\vdots \\
\lambda_{nk}
\end{bmatrix} = \begin{bmatrix} C_{11} & \cdots & C_{1k} \\
\vdots & \ddots & \vdots \\
C_{k1} & \cdots & C_{kk} \end{bmatrix}^{-1} \begin{bmatrix}
C_{11}(s_0, s_1) \\
\vdots \\
C_{n1}(s_0, s_n) \\
C_{12}(s_0, s_1) \\
\vdots \\
C_{nk}(s_0, s_n)
\end{bmatrix}.
\]

The proposed predictor in (5.1) is the best linear unbiased predictor obtained by minimizing the least squares of the multivariate binary responses (Cressie, 1993). The \( n \times n \) cross-covariance matrix \( C_{ij} \) between variables \( Y_i \) and \( Y_j \) is given by

\[
C_{ij} = \begin{bmatrix}
C_{ij}(s_1, s_1) & \cdots & C_{ij}(s_1, s_n) \\
\vdots & \ddots & \vdots \\
C_{ij}(s_n, s_1) & \cdots & C_{ij}(s_n, s_n)
\end{bmatrix}.
\]

Indicator cokriging is the best estimator in the least-squares sense, while univariate indicator kriging (Journel, 1983) is a simplification of indicator cokriging to reduce the computational demands.

Typically, the spatial dependence of the indicator variables is modeled using variograms. The variogram is estimated for each indicator variable, and cross-variograms
(Journel and Huijbregts, 1978) are estimated for each pair of indicator variables by fitting parametric models to empirical variograms and cross-variograms, respectively. A single theoretical model is fitted to each indicator variogram and cross-variogram, and the linear model of coregionalization (Goovaerts, 1994b) is most widely used. However, variograms are strongly influenced by the marginal distribution of the random field and hence are susceptible to measurement anomalies. The predicted conditional CDF often suffers from order relation violations of a valid CDF due to poor modeling of variograms (Goovaerts, 1994b,a). Moreover, a variogram describes just the mean dependence and not the dependence over the whole range of distribution. To overcome these challenges, we propose to estimate the weights $\lambda_{ij}$’s (5.2) through copula to capture the multivariate spatial dependence. The covariance matrix $C_{ij}$ is unknown; we use copulas to estimate the cross-covariance matrices between the indicator variables. The covariance $C_{ij}(s_u, s_v) = \text{Cov}\{Y_i(s_u), Y_j(s_v)\}$ is related to the spatial copula describing the dependence of the random field $Z$ between locations $s_u$ and $s_v$, and it is established in the following subsection.

5.2.2 Spatial modeling using copulas

Copulas are standardized multivariate distributions with uniform marginals used to describe the dependence between random variables independently of their marginal distributions. We assume second-order stationarity, so the correlation between two locations only depends on the difference between the two locations. Let $Z = \{Z(s_1), \ldots, Z(s_n)\}^\top$ be a random vector with $n$-dimensional CDF $F_Z$. Using Sklar’s theorem, any multivariate distribution can be represented by a copula and univariate marginal distribution functions. Hence, there exists a unique copula $C_Z$, such that

$$F_Z(z_1, \ldots, z_n) = C_Z\{F_{Z(s_1)}(z_1), \ldots, F_{Z(s_n)}(z_n)\}. \quad (5.3)$$
First, we will discuss the estimation of unknown copula $C_Z$ from the sample $Z(s_1), \ldots, Z(s_n)$. Then, we estimate the auto and cross covariances of the indicator variables using the estimated copula.

**Estimation of copula**

Let $C_Z$ belongs to an absolutely continuous parametric family of copulas $\mathcal{C}$,

$$
\mathcal{C} = \{C_{\theta_c} : \theta_c \in \Theta\},
$$

where $\Theta$ is the parameter space of the copula parameters $\theta_c$. In spatial copula model, we have two types of parameters, i.e., parameters defining the spatial correlation structure and copula parameters. The natural choices of spatial correlation structure include spherical, exponential, Matérn correlation functions. We aim to estimate the parameters using the maximum likelihood approach. Differentiating the Sklar representation (5.3), the corresponding density $f_Z$ is given by

$$
f_Z(z_1, \ldots, z_n) = c_Z \left\{ F_{Z(s_1)}(z_1), \ldots, F_{Z(s_n)}(z_n) \right\} \prod_{i=1}^n f_{Z(s_i)}(z_i),
$$

where $c_Z$ is the copula density of $C_Z$, and $f_{Z(s_i)}(z_i)$ is the density of $F_{Z(s_i)}(z_i), i = 1, \ldots, n$. The parameters can be estimated using maximum likelihood estimator, so we obtain the log-likelihood function as

$$
\log L(\theta_1, \ldots, \theta_n, \theta_c) = \log f_Z(z_1, \ldots, z_n)
= \log c_Z \left\{ F_{Z(s_1)}(z_1 \mid \theta_1), \ldots, F_{Z(s_n)}(z_n \mid \theta_n) \mid \theta_c \right\}
+ \sum_{i=1}^n \log f_{Z(s_i)}(z_i \mid \theta_i). 
$$

(5.4)

The parameters $\theta_c$ can be estimated by maximizing the log-likelihood function (5.4). One may consider parametric models for the marginal CDFs, but the parametric
model approach has two limitations. First, the parametric assumptions can cause biases in the estimation of both marginal distribution and copulas. Second, the optimization is computationally difficult for high-dimensional parameter space, especially if \( n \) is large. To overcome these challenges, we propose to estimate the marginal distributions nonparametrically using the empirical CDF

\[
\hat{F}_{Z(s_i)}(z) = \frac{\sum_{j=1}^{n} I\{Z(s_j) \leq z\}}{n+1}, \quad i = 1, \ldots, n.
\]

The copula parameters can then be estimated by setting the estimated margins to (5.4) and maximizing the terms that depend on the parameter \( \theta_c \), i.e.,

\[
\hat{\theta}_c = \arg \max_{\theta_c} \log c_Z\{\hat{F}_{Z(s_1)}(z_1), \ldots, \hat{F}_{Z(s_n)}(z_n) \mid \theta_c\}.
\]

We refer to the proposed estimator \( \hat{\theta}_c \) as the semiparametric maximum pseudo-likelihood estimator. The method is semiparametric as the marginal is estimated nonparametrically, while the kriging coefficients are estimated through a working parametric copula model. In practice, the maximization is carried out using numerical optimization.

**Modeling spatial dependence using copulas**

Copula generalizes the concept of the variogram for spatial modeling (Bárdossy, 2006). Let us denote \( F_Z \) as the univariate distribution of the random field for each location \( s \) due to stationarity. Since all the multivariate distributions of the random fields can be defined using multivariate copulas, the relation between two locations with separation vector \( h \) is characterized by the bivariate distribution

\[
P\{Z(s) \leq z_1, Z(s + h) \leq z_2\} = C_h\{F_Z(z_1), F_Z(z_2)\}.
\]
Here, the dependence structure is described by the copula $C_h$, which is a function of separation vector $h$ (or distance $\|h\|$ in case of isotropy). Hence, copulas describe the dependence over the entire distribution and not just mean as the variogram does. We derive the auto and cross covariances of the indicator variables using spatial copula through

$$\text{Cov}\{Y_i(s_u), Y_j(s_v)\} = E[I\{Z(s_u) \leq q_i\}I\{Z(s_v) \leq q_j\}] - E[I\{Z(s_u) \leq q_i\}]E[I\{Z(s_v) \leq q_j\}]$$

$$= P\{Z(s_u) \leq q_i, Z(s_v) \leq q_j\} - P\{Z(s_u) \leq q_i\}P\{Z(s_v) \leq q_j\}$$

$$= F_Z(s_u, s_v)(q_i, q_j) - F_Z(q_i)F_Z(q_j). \quad (5.6)$$

Applying Sklar’s theorem to equation (5.6), we can establish a relationship between indicator covariances and copula functions in the following way

$$\text{Cov}\{Y_i(s_u), Y_j(s_v)\} = C_h\{F_Z(q_i), F_Z(q_j)\} - F_Z(q_i)F_Z(q_j). \quad (5.7)$$

Here, the marginal distributions are estimated using the empirical distribution function (5.5), and joint distributions are estimated using copulas $C_h\{\hat{F}_Z(q_i), \hat{F}_Z(q_j) | \hat{\theta}_c\}$. Although the estimation of covariances from copulas can be complicated, this approach has the benefits mentioned above, and covariances do not need to be calculated for indicator combinations separately; see equation (5.7).

Note that even if the copula is misspecified, the proposed kriging estimator in (5.1) is still unbiased. Although the optimality relies on the correct specification of copula, the users can choose from different copula families, e.g., elliptical copulas (Frahm et al., 2003; Rémillard et al., 2012), which includes Gaussian copula and $t$-copula, $\nu$-transformed multivariate normal copula (Bárdossy and Li, 2008), and Archimedean family of copulas (Genest and Rivest, 1993). And practically, we can use the diagnostic tools (defined in Section 5.2.3) to choose a copula that fits the data well. The advantage of using the elliptical copula family or the $\nu$-transformed multivariate
normal copula for estimation and modeling is that the correlation matrix explicitly appears in the analytical expression of elliptical distributions. The correlation matrix is parameterized by a spatial covariance model, and there is no need to estimate the marginal variance of the random field as it will be equal to 1. The reason is that variance is the property of the marginal distribution of the random field, and copula describes dependence without the information of marginal distribution.

5.2.3 Spatial probabilistic prediction

One of the primary goals in geostatistics is to make predictions at unknown spatial location \( s_0 \), from the observed data \( D_n \). There are two major approaches for spatial predictions: (i) point prediction (Cressie, 1993) and (ii) probabilistic prediction (Gneiting et al., 2007; Gneiting and Katzfuss, 2014). While point prediction just provides a value of \( Z(s_0) \), probabilistic prediction provides a comprehensive picture of the complete predictive distribution of \( Z(s_0) \mid D_n \), which is more informative and better at capturing the uncertainty in prediction. In this section, we derive the predictive distribution with the CMIK method, and the useful measures of central tendency and uncertainty.

Predictive distribution function estimation

Our objective is to estimate the local predictive distribution at each unsampled location. Plugging in the relationship between the auto and cross covariances using copula (5.7) to the cokriging estimator (5.1), we arrive at multiple indicator kriging based on spatial copula models. The vector \( Y(s_0) = \{Y_1(s_0), \ldots, Y_k(s_0)\}^\top \) is predicted simultaneously, and for each \( j = 1, \ldots, k \), \( \hat{Y}_j(s_0) = \sum_{i=1}^n \sum_{j=1}^k \hat{\lambda}_{ij} Y_j(s_i) \), where \( \hat{\lambda}_{ij} \) are the estimated indicator kriging weights. Since \( Y_j(s_0) = I\{Z(s_0) \leq q_j\} \), \( \hat{Y}_j(s_0) \) is an estimate of the conditional CDF of \( Z(s_0) \) at \( q_j \). This follows from the simple
relation:

\[ E[I\{Z(s_0) \leq q_j\} \mid D_n] = P\{Z(s_0) \leq q_j \mid D_n\} = F_{Z(s_0)|D_n}(q_j). \]

Hence, the result of multiple indicator kriging is a set of estimated values of the conditional CDF of \( Z(s_0) \) at \( q_j \), denoted by \( \hat{F}_{Z(s_0)|D_n}(q_j) \), for \( j = 1, \ldots, k \). Since the CDFs at different quantiles are estimated separately, it may not satisfy the order relations of a valid CDF (Journel, 1983). The source of this problem is often related to poorly modeled variograms. However, numerically, we did not observe any order relation violations with the proposed copula-based methods. Nevertheless, the order relation violations can be corrected using the upward and downward averaging algorithm (Deutsch and Journel, 1998). The final estimated CDFs then satisfy the relations

\[
\hat{F}_{Z(s_0)|D_n}(q_j) \in [0, 1] \quad \forall j, \\
\hat{F}_{Z(s_0)|D_n}(q_j) \leq \hat{F}_{Z(s_0)|D_n}(q_l), \quad \forall q_j < q_l.
\]

To obtain the complete distribution function, we need to interpolate between the estimated CDF values at the chosen quantiles and extrapolate beyond the first and last quantile. The upper (and lower) tails can be extrapolated to the midpoint between the uppermost (lowermost) cutoff and the maximum (minimum) value in the observed data. After filling the local distribution, the final estimator is defined as \( \hat{F}_{Z(s_0)|D_n}(q), \forall q \in S \), where \( S \) is the support of the distribution. The complete smooth local distribution can be obtained by fitting a linear, power, or hyperbolic model. E.g., a nonparametric kernel regression estimate can be used to smooth the CDF. The complete local CDF is practically obtained using the global empirical distribution function, where gaps and tails are fitted with the scaled global CDF. This preserves the shape of the global CDF in the intervals and scales it to the specified minimum
and maximum estimates. Hence, the complete predictive distribution is obtained using the estimated CDF values and the global empirical distribution function.

After obtaining the predictive distribution, it is possible to have any desired summary statistic. It allows us to compute central tendency measures such as mean or median, and measures of uncertainty such as variance and prediction confidence interval. We can also plot several maps, including the probability of exceeding or not exceeding a certain threshold. The \( p^{th} \) sample quantile is extracted from the predictive distribution:

\[
q_{s_0}(p) = \hat{F}^{-1}_{Z(s_0)|D_n}(p) = \inf\{x : \hat{F}_{Z(s_0)|D_n}(x) \geq p\}.
\]

Median is the robust central value of the distribution, and it is obtained as \( \hat{F}^{-1}_{Z(s_0)|D_n}(0.5) \).

With the predicted CDF, the prediction confidence intervals can be obtained directly. The \( 100(1-\alpha)\% \) prediction confidence interval is given by \( \{\hat{F}^{-1}_{Z(s_0)|D_n}(\alpha/2), \hat{F}^{-1}_{Z(s_0)|D_n}(1-\alpha/2)\} \). To calculate other summary statistics, the procedure is to obtain \( L \) quantiles (typically 200) from the nonparametric predictive distribution, \( q_{s_0}(p_l), p_l = l/(L+1), l = 1, \ldots, L \). The mean and variance at each location, denoted by \( m(s_0) \) and \( \sigma^2(s_0) \) respectively, can be computed using the extracted quantiles:

\[
m(s_0) \approx \frac{1}{L} \sum_{l=1}^{L} q_{s_0}(p_l); \quad \sigma^2(s_0) \approx \frac{1}{L} \sum_{l=1}^{L} \{q_{s_0}(p_l) - m(s_0)\}^2.
\]

The probability of exceeding a cutoff can be computed as the proportion of the extracted quantiles above that cutoff. The choice of thresholds is crucial in multiple indicator kriging approach as the predictive distribution is estimated at the chosen thresholds. It is constructive to choose more thresholds within the part of distribution with more interest, for instance, in the lower or upper tails of the distribution, and interesting inflection points such as median or exceedance cutoffs.
Model diagnostics

In this section, we introduce some diagnostic measures to evaluate the predictive performance of the proposed methods. The diagnostic tools can be used to compare the proposed model with existing methods such as Gaussian kriging, variogram-based multiple indicator kriging through cross-validation. In addition, these diagnostic measures can also be used in practice to help users make selections, e.g., to make a choice of copula and to choose a spatial correlation function. In the proposed method, probabilistic prediction takes the form of predictive cumulative probability distribution functions. Let us denote the predicted conditional distribution of $Z(s_0) \mid D_n$ as $F_{s_0}$, and the true conditional distribution as $G_{s_0}$. The performance of probabilistic prediction can be evaluated using two steps: (i) calibration and (ii) sharpness. Calibration refers to how close the true distribution is to the predicted distribution. Sharpness refers to the concentration of the predictive distribution around the true value. An important tool to assess the calibration is the probability integral transform (PIT) (Dawid, 1984; Diebold et al., 1998), the value that the predictive CDF attains at the observation, defined by

$$\text{PIT} = p_{s_0} = F_{s_0}(z(s_0)),$$

where $z(s_0)$ is the observed value from the true conditional distribution. If the prediction is ideal, i.e., $F_{s_0} = G_{s_0}$, then $p_{s_0}$ has a uniform distribution. In practice, PIT values can be calculated for different locations of the random field, and uniformity can be checked by plotting the PIT histogram. If the predictive distribution passes calibration, sharpness can be assessed by the average length and empirical coverage probabilities of the central 50% and 90% prediction intervals.

Numerical assessment of probabilistic predictions can be done through scoring rules which address calibration and sharpness simultaneously. In particular, the con-
tinuous rank probability score (CRPS, Gneiting et al. (2007)) is robust and defined directly in terms of the predictive CDF \( F_{s_0} \) as

\[
\text{CRPS}\{F_{s_0}, z(s_0)\} = \int_{-\infty}^{\infty} [F_{s_0}(x) - I\{z(s_0) \leq x\}]^2 dx,
\]

where \( I(\cdot) \) is an indicator function. It is a quadratic measure which represent the difference between the predictive CDF and the empirical CDF of the observation. The CRPS has desirable properties, and it can be computed using numerical quadrature rules (Unger, 1985).

### 5.3 Simulation study

In this section, we simulate non-Gaussian random fields to evaluate the performance of proposed multiple indicator kriging methods with spatial copulas. We simulate data from a transGaussian random field named Tukey \( g \)-and-\( h \) random field, which has very flexible marginal distributions (Xu and Genton, 2017). The random field is based on the Tukey’s \( g \)-and-\( h \) transformation (Tukey, 1977) given by

\[
\tau_{g,h}(x) = \begin{cases} 
  g^{-1}\{\exp(gx) - 1\}\exp(hx^2/2), & g \neq 0, h \geq 0, \\
  z\exp(hx^2/2), & g = 0, h \geq 0,
\end{cases}
\]

which is a strictly monotone function of \( x \in \mathbb{R} \). The standard Tukey \( g \)-and-\( h \) random field is obtained by taking the Tukey \( g \)-and-\( h \) transformation to a standard Gaussian random field and is defined by \( Z(s) = \tau_{g,h}\{X(s)\} \), where \( X(s) \) is a standard Gaussian random field with mean 0, variance 1, and correlation function \( \text{corr}\{X(s_1), X(s_2)\} = \rho_X(s_1, s_2) \). The parameter \( g \) governs the skewness of the random field \( Z(s) \), where \( g > 0 \) makes the random field positively skewed, and \( g < 0 \) makes it negatively skewed. The parameter \( h \) governs the random field’s tail behavior, where a larger value of \( h \) indicates a heavier tail.
In the simulation studies, we assume the latent Gaussian random field $X(s)$ has the exponential correlation function

$$\rho_X(s_1, s_2) = \exp(\|s_1 - s_2\|/\theta),$$

where $\|s_1 - s_2\|$ is the distance between the locations $s_1$ and $s_2$, and $\theta$ is the range parameter. We vary the strength of dependence by setting the range parameter to be 0.25, 0.5, and 0.75, where a larger value of $\theta$ means a higher autocorrelation. We consider multiple values for $g$ and $h$ in the TGH random fields for simulation. For $g = 0$ and $h > 0$, we can examine the effect of prediction when the data has heavy tails. For $h = 0$ and $g > 0$, the influence on the prediction when the data is asymmetric can be checked. We also consider a combination of skewness and elongation by setting $g > 0$, $h > 0$.

### 5.3.1 Evaluation of point prediction

In this simulation study, we simulate spatial data from Tukey $g$-and-$h$ random field with the exponential covariance function. The goal is to evaluate the performance of point prediction of the proposed copula-based multiple indicator kriging (CMIK) method and compare it with the variogram multiple indicator kriging (VMIK) and Gaussian kriging (GK) methods. We simulate 100 spatial observations at irregularly spaced locations over the region $[0, 1] \times [0, 1]$, not too close to each other. The irregular spatial locations are generated by the following procedure: First, $10 \times 10$ spatial locations are generated on the regular grid $[0, 1] \times [0, 1]$; then a random noise uniformly distributed on $[-0.4, 0.4]$ is added to each location. We randomly divide 80% data for training and leave 20% data for the validation. We repeat the simulation 500 times, and in each simulation, the same set of spatial locations is used for training and validation. Therefore, each location in the validation dataset has 500 predictions.
Table 5.1: MAD and standard errors (in parentheses) of GK, VMIK, and proposed CMIK\(_G\) and CMIK\(_t\) methods for Tukey \(g\)-and-\(h\) random fields with exponential covariance function and range parameters \(\theta = 0.25, 0.375, 0.5\) with multiple \(g\) and \(h\) values.

<table>
<thead>
<tr>
<th>((g,h))</th>
<th>(\theta)</th>
<th>GK (\text{MAD (SE)})</th>
<th>VMIK (\text{MAD (SE)})</th>
<th>CMIK(_G) (\text{MAD (SE)})</th>
<th>CMIK(_t) (\text{MAD (SE)})</th>
</tr>
</thead>
<tbody>
<tr>
<td>(g=0, h=0)</td>
<td>0.25</td>
<td>0.3727 (0.06)</td>
<td>0.3968 (0.06)</td>
<td>0.3810 (0.05)</td>
<td>0.3846 (0.05)</td>
</tr>
<tr>
<td></td>
<td>0.375</td>
<td>0.3074 (0.03)</td>
<td>0.3354 (0.05)</td>
<td>0.3221 (0.05)</td>
<td>0.3251 (0.05)</td>
</tr>
<tr>
<td></td>
<td>0.5</td>
<td>0.2695 (0.04)</td>
<td>0.2989 (0.04)</td>
<td>0.2932 (0.04)</td>
<td>0.2850 (0.04)</td>
</tr>
<tr>
<td>(g=0, h=0.3)</td>
<td>0.25</td>
<td>0.4927 (0.08)</td>
<td>0.4979 (0.06)</td>
<td>0.4838 (0.06)</td>
<td>0.4769 (0.07)</td>
</tr>
<tr>
<td></td>
<td>0.375</td>
<td>0.4417 (0.05)</td>
<td>0.4166 (0.04)</td>
<td>0.4129 (0.04)</td>
<td>0.4171 (0.04)</td>
</tr>
<tr>
<td></td>
<td>0.5</td>
<td>0.3788 (0.04)</td>
<td>0.3616 (0.04)</td>
<td>0.3577 (0.03)</td>
<td>0.3490 (0.04)</td>
</tr>
<tr>
<td>(g=0, h=0.5)</td>
<td>0.25</td>
<td>0.6630 (0.13)</td>
<td>0.5709 (0.1)</td>
<td>0.5515 (0.11)</td>
<td>0.5561 (0.10)</td>
</tr>
<tr>
<td></td>
<td>0.375</td>
<td>0.5299 (0.04)</td>
<td>0.4876 (0.05)</td>
<td>0.4570 (0.05)</td>
<td>0.4518 (0.05)</td>
</tr>
<tr>
<td></td>
<td>0.5</td>
<td>0.4693 (0.04)</td>
<td>0.4306 (0.06)</td>
<td>0.4083 (0.05)</td>
<td>0.4085 (0.04)</td>
</tr>
<tr>
<td>(g=0.5, h=0)</td>
<td>0.25</td>
<td>0.3924 (0.05)</td>
<td>0.3913 (0.06)</td>
<td>0.3846 (0.04)</td>
<td>0.3916 (0.05)</td>
</tr>
<tr>
<td></td>
<td>0.375</td>
<td>0.2993 (0.05)</td>
<td>0.2984 (0.04)</td>
<td>0.2935 (0.04)</td>
<td>0.2968 (0.04)</td>
</tr>
<tr>
<td></td>
<td>0.5</td>
<td>0.2783 (0.03)</td>
<td>0.2739 (0.03)</td>
<td>0.2784 (0.03)</td>
<td>0.2638 (0.03)</td>
</tr>
<tr>
<td>(g=0.75, h=0)</td>
<td>0.25</td>
<td>0.3872 (0.05)</td>
<td>0.3648 (0.05)</td>
<td>0.3580 (0.05)</td>
<td>0.3559 (0.05)</td>
</tr>
<tr>
<td></td>
<td>0.375</td>
<td>0.2980 (0.04)</td>
<td>0.2973 (0.04)</td>
<td>0.2881 (0.04)</td>
<td>0.2943 (0.04)</td>
</tr>
<tr>
<td></td>
<td>0.5</td>
<td>0.2794 (0.03)</td>
<td>0.2704 (0.03)</td>
<td>0.2756 (0.03)</td>
<td>0.2667 (0.04)</td>
</tr>
<tr>
<td>(g=0.5, h=0.3)</td>
<td>0.25</td>
<td>0.5543 (0.08)</td>
<td>0.5330 (0.08)</td>
<td>0.4930 (0.08)</td>
<td>0.5013 (0.08)</td>
</tr>
<tr>
<td></td>
<td>0.375</td>
<td>0.4245 (0.05)</td>
<td>0.4113 (0.06)</td>
<td>0.3878 (0.06)</td>
<td>0.3977 (0.05)</td>
</tr>
<tr>
<td></td>
<td>0.5</td>
<td>0.3580 (0.04)</td>
<td>0.3853 (0.05)</td>
<td>0.3648 (0.04)</td>
<td>0.3641 (0.04)</td>
</tr>
<tr>
<td>(g=0.3, h=0.4)</td>
<td>0.25</td>
<td>0.5883 (0.08)</td>
<td>0.5373 (0.06)</td>
<td>0.5263 (0.07)</td>
<td>0.5307 (0.06)</td>
</tr>
<tr>
<td></td>
<td>0.375</td>
<td>0.4897 (0.05)</td>
<td>0.4571 (0.05)</td>
<td>0.4459 (0.04)</td>
<td>0.4416 (0.04)</td>
</tr>
<tr>
<td></td>
<td>0.5</td>
<td>0.4610 (0.07)</td>
<td>0.3986 (0.07)</td>
<td>0.3913 (0.06)</td>
<td>0.3739 (0.06)</td>
</tr>
</tbody>
</table>

We consider two copula functions for CMIK: CMIK\(_G\) and CMIK\(_t\), which use the Gaussian copula and \(t\)-copula with unspecified degrees of freedom, respectively. We include the existing VMIK and GK methods for comparison of point prediction. For all four methods, we assume an exponential spatial correlation structure.

We choose 15 thresholds for the multiple indicator kriging methods at different quantiles of the global distribution such that interesting inflection points on the global CDF are covered. E.g., for a heavy-tailed distribution, we choose more thresholds in the tails of the distribution, instead of choosing equispaced quantile levels. We use the median as the predicted value for the multiple indicator kriging methods as the median gives a robust summary of the center of the distribution. We measure the
accuracy of point prediction at each spatial location using Median Absolute Deviation (MAD), which is given by

\[
\text{MAD}(\mathbf{s}_i) = \text{median}\{|z(\mathbf{s}_i)_b - \hat{z}(\mathbf{s}_i)_b|, \ b = 1, \ldots, 500\}, \ i = 1, \ldots, 20, \tag{5.8}
\]

where \(z(\mathbf{s}_i)_b\) and \(\hat{z}(\mathbf{s}_i)_b\) are the observed and predicted values at location \(\mathbf{s}_i\) for the \(b\)th simulation run, respectively. The MAD for each method is computed by averaging the MAD over all the locations in the validation set. The predictive performance of all four methods is summarized in Table 5.1. For the Gaussian random field (\(g = 0, h = 0\)), Gaussian kriging performs well, and multiple indicator kriging methods are comparable. The copula-based methods show improvement over the variogram method. As \(h\) increases, i.e., the marginal distribution of the random field has heavier tails, the proposed copula-based methods outperform the Gaussian kriging method, which is not flexible enough to accommodate the heavy tails and thus yields suboptimal prediction results. Similarly, as \(g\) increases, i.e., skewness increases, or if the marginal distribution is both skewed and heavy-tailed, the CMIK methods perform the best. For all the cases considered, the proposed CMIK methods consistently outperform the variogram method. When the strength of dependence is low (\(\theta = 0.25\)), CMIK_G performs better than CMIK_t for most of the cases. But as \(\theta\) increases, the performance of CMIK_t gets better than the performance of CMIK_G. This suggests that as the strength of dependence increases, \(t\)-copula can better model the spatial dependence.

### 5.3.2 Evaluation of probabilistic prediction

Here, the goal is to evaluate the performance of the probabilistic prediction of the proposed methods and compare it with the existing methods. We use the same simulation setting, as described in the previous subsection. We assess the statistical consistency between the predicted CDFs and the observed values using the PIT
histograms. The PIT values were computed at each prediction location for 500 simulation runs. We consider three types of non-Gaussian random fields: skewed \((g = 0.75, h = 0)\), heavy-tailed \((g = 0, h = 0.5)\), and both skewed and heavy-tailed \((g = 0.5, h = 0.3)\) with exponential covariance function with range equal to 0.25. The PIT histograms for the four methods and all three cases are plotted in Figure 5.1. For all three cases, the PIT histogram for the Gaussian kriging method is far from Uniform distribution. The PIT histograms for both CMIK\(_g\) and CMIK\(_t\) methods are closer to the uniform distribution and show improvement over the VMIK method, with the CMIK\(_g\) method performing the best. In the point prediction, the differences in the MAD were smaller between the Gaussian kriging and indicator kriging methods as compared to the differences between the PIT histograms, indicating that evaluating only the point prediction may be misleading to assess the predictive performance of the methods. With the assessment of PIT histograms, we get a more clear view of the improvement in prediction.

To study the sharpness of the four predictive distributions, we compute the average lengths and coverage probabilities of the 50% and 90% prediction intervals. The results are summarized in Table 5.2. It can be observed that prediction intervals of the GK method are too wide and hence overestimate the coverage probabilities. For 90% prediction intervals, the VMIK method leads to a low coverage, and the 50% prediction intervals are wider than those for the proposed CMIK methods. The CMIK\(_g\) and CMIK\(_t\) methods give the shortest prediction intervals on average with a coverage close to the nominal level. We also provide a numerical assessment of probabilistic prediction using the scoring rule CRPS. The accuracy of probabilistic prediction at each prediction location is measured by

\[
\text{CRPS}(s_i) = \text{mean}\{\text{CRPS}\{\hat{F}(s_i)_b, z(s_i)_b\}, b = 1, \ldots, 500, \ i = 1, \ldots, 20\}, \quad (5.9)
\]
where $\hat{F}(s_i)_b$ is the estimated predictive distribution at prediction location $s_i$ for the $b$th simulation run. The CRPS was also computed for the three cases (skewed, heavy-tailed, skewed, and heavy-tailed) and is reported in Table 5.3 for all the four methods. In all the three cases, the proposed copula-based methods perform better than the GK and the VMIK methods. The performance of both CMIK$_g$ and CMIK$_t$ are close, but the CMIK$_g$ performs the best for the given dependence structure. When the estimated degrees of freedom of $t$-copula is large, it approaches close to the performance of Gaussian copula. To allow more possibilities, we consider $t$-copula with lower degrees of freedom in such cases. Among the three kinds of marginal
Table 5.2: Coverage percentages (CP) and average lengths (Length) of prediction confidence intervals.

<table>
<thead>
<tr>
<th>$(g,h)$</th>
<th>Method</th>
<th>CP (50%)</th>
<th>Length</th>
<th>CP (90%)</th>
<th>Length</th>
</tr>
</thead>
<tbody>
<tr>
<td>$g=0.75$, $h=0$</td>
<td>GK</td>
<td>62.5%</td>
<td>1.19</td>
<td>90.1%</td>
<td>2.90</td>
</tr>
<tr>
<td></td>
<td>VMIK</td>
<td>50.3%</td>
<td>1.12</td>
<td>84.0%</td>
<td>2.62</td>
</tr>
<tr>
<td></td>
<td>CMIK$_g$</td>
<td>52.9%</td>
<td>1.11</td>
<td>87.1%</td>
<td>2.70</td>
</tr>
<tr>
<td></td>
<td>CMIK$_t$</td>
<td>54.1%</td>
<td>1.15</td>
<td>87.3%</td>
<td>2.81</td>
</tr>
<tr>
<td>$g=0$, $h=0.5$</td>
<td>GK</td>
<td>71.3%</td>
<td>3.29</td>
<td>92.5%</td>
<td>8.03</td>
</tr>
<tr>
<td></td>
<td>VMIK</td>
<td>50.9%</td>
<td>1.79</td>
<td>84.2%</td>
<td>4.75</td>
</tr>
<tr>
<td></td>
<td>CMIK$_g$</td>
<td>53.6%</td>
<td>1.73</td>
<td>88.1%</td>
<td>4.98</td>
</tr>
<tr>
<td></td>
<td>CMIK$_t$</td>
<td>54.5%</td>
<td>1.75</td>
<td>88.8%</td>
<td>5.05</td>
</tr>
<tr>
<td>$g=0.5$, $h=0.3$</td>
<td>GK</td>
<td>67.1%</td>
<td>2.06</td>
<td>91.9%</td>
<td>5.02</td>
</tr>
<tr>
<td></td>
<td>VMIK</td>
<td>52.2%</td>
<td>1.53</td>
<td>84.8%</td>
<td>3.90</td>
</tr>
<tr>
<td></td>
<td>CMIK$_g$</td>
<td>54.1%</td>
<td>1.50</td>
<td>88.0%</td>
<td>4.03</td>
</tr>
<tr>
<td></td>
<td>CMIK$_t$</td>
<td>55.3%</td>
<td>1.52</td>
<td>89.0%</td>
<td>4.15</td>
</tr>
</tbody>
</table>

Table 5.3: CRPS and standard errors (in parentheses) of GK, VMIK, and proposed CMIK$_G$ and CMIK$_t$ methods with three cases: skewed $(g = 0.75, h = 0)$, heavy-tailed $(g = 0, h = 0.5)$, and both skewed and heavy-tailed $(g = 0.5, h = 0.3)$.

<table>
<thead>
<tr>
<th>Marginal / Method</th>
<th>GK</th>
<th>VMIK</th>
<th>CMIK$_g$</th>
<th>CMIK$_t$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Skewed</td>
<td>0.4590 (0.03)</td>
<td>0.4497 (0.03)</td>
<td>0.4299 (0.03)</td>
<td>0.4350 (0.03)</td>
</tr>
<tr>
<td>Heavy-tailed</td>
<td>1.1489 (0.13)</td>
<td>0.9000 (0.10)</td>
<td>0.8594 (0.09)</td>
<td>0.8704 (0.09)</td>
</tr>
<tr>
<td>Skewed and heavy-tailed</td>
<td>0.7963 (0.05)</td>
<td>0.7231 (0.06)</td>
<td>0.6858 (0.05)</td>
<td>0.6945 (0.05)</td>
</tr>
</tbody>
</table>

distributions considered, the difference in the improvements of methods is more when the random field has heavy tails since the Gaussian kriging and variogram methods are susceptible to measurement anomalies.

5.4 Applications

5.4.1 Application to precipitation data

In this section, we apply the proposed methods of multiple indicator kriging using copulas to the precipitation dataset introduced in Section 5.1. The average daily precipitation (in millimeters) of Spain for November 2019 at 169 stations is plotted in Figure 5.2 using the R package ‘ggmap’ (Kahle and Wickham, 2013). The spatial locations appear quite sparse on the map, and a major human desire is to predict the
Figure 5.2: The daily average precipitation (mm) for November 2019 at 169 monitoring stations in Spain. (a) Map of the precipitation data; (b) histogram of the precipitation intensities.

observations at unknown locations and characterize the uncertainty. The histogram of
the precipitation is plotted in Figure 5.2, and it is evident that the marginal distribution is severely positively skewed. Therefore, we aim to provide accurate probabilistic predictions of the precipitation intensities and create maps at finer spatial resolutions with the observed data from the monitoring stations.

We consider the proposed multiple kriging method with Gaussian copula and \(t\)-copula: \(\text{CMIK}_g\) and \(\text{CMIK}_t\). We choose a total number of 17 thresholds to divide the global distribution adequately. Since the marginal distribution of the random field is positively skewed, we choose more thresholds at the upper quantiles of the distribution, such that it covers all the interesting inflection points on the global CDF. We include the method of Gaussian kriging on log-transformed data, denoted by \(\text{logGK}\). It aligns the distribution closer to normality, but the resulting estimates appear in the logarithmic scale, which makes the interpretation difficult, while back transformation introduces biases. We evaluate the predictive performance of our model and compare it with the \(\text{VMIK}\), \(\text{GK}\), and \(\text{logGK}\) methods using \(\text{PIT}\), \(\text{CRPS}\), and \(\text{MAD}\) values. We use the following resampling approach to compute these values. First,
we randomly divide 80% of the data to fit the proposed model. Second, we obtain
the predictive distribution function and compute PIT, CRPS values at the remaining
20% spatial locations. This procedure is repeated 500 times, and all the PIT and
CRPS values are recorded. We plot the PIT histograms to assess the calibration of
the predictive distributions for all the five methods in Figure 5.3. The PIT histogram
of CMIK$_t$ appears more uniform than all the other approaches. The PIT histogram
for the CMIK$_g$ method suggests that its prediction intervals are slightly wider. Both
the copula-based methods perform better than the variogram approach, which sug-
gests a departure from calibration. From the PIT histogram of Gaussian kriging, it
is evident that it performs poorly, and it is not suitable for probabilistic prediction
of this precipitation data. The log-transformed Gaussian kriging show improvements
Table 5.4: Numerical assessment of the predictive performance of GK, logGK, VMIK, and proposed CMIK$_G$ and CMIK$_t$ methods for the precipitation data.

<table>
<thead>
<tr>
<th>Measure</th>
<th>GK</th>
<th>logGK</th>
<th>VMIK</th>
<th>CMIK$_G$</th>
<th>CMIK$_t$</th>
</tr>
</thead>
<tbody>
<tr>
<td>MAD</td>
<td>4.06</td>
<td>4.22</td>
<td>4.04</td>
<td>3.71</td>
<td>3.68</td>
</tr>
<tr>
<td>CRPS</td>
<td>7.42</td>
<td>6.33</td>
<td>6.66</td>
<td>6.67</td>
<td>6.42</td>
</tr>
<tr>
<td>mCRPS</td>
<td>4.11</td>
<td>3.60</td>
<td>3.50</td>
<td>3.49</td>
<td>3.27</td>
</tr>
</tbody>
</table>

over Gaussian kriging; however, the PIT histogram is still not close to the uniform distribution, reflecting the biases in prediction obtained for non-Gaussian data.

We summarize the numerical assessments of the predictive performance of all the five methods in Table 5.4. The MAD and CRPS are defined by equations (5.8) and (5.9), respectively, and are computed using the resampling approach mentioned above. We also compute the mCRPS metric, which considers the median of the CRPS values instead of the mean, and it is more robust to outliers. The MAD measures the accuracy of point prediction, and the proposed copula-based methods perform better than the existing methods, with CMIK$_t$ performing the best. The mCRPS suggests that the CMIK$_t$ method performs the best. Both the CMIK methods have a better score than the variogram approach. The proposed CMIK methods have much better predictive performance than the Gaussian and log-Gaussian kriging methods. In practice, it can be difficult to find a unique transformation to make data normal, and it is better to use a flexible distribution-free approach. Considering all the diagnostic approaches, we argue that the proposed multiple indicator kriging with $t$-copula is the best method for spatial probabilistic prediction of this precipitation data.

Finally, we create high-resolution precipitation maps over Spain with the proposed CMIK$_t$ method using the observed data. In Figure 5.4a, we plot the precipitation map of predicted values, which shows high precipitation in the areas North of Spain and near Madrid. We plot the contour map of probability exceeding 50mm on average daily in Figure 5.4b. We can observe the spatial patterns over the region with different levels of probabilities exceeding the threshold.
Figure 5.4: Results from the proposed CMIK$_t$ method for the precipitation data. (a) Map of predicted values, (b) contour map for the probability of exceeding 50mm on average daily during November 2019 in Spain.

5.4.2 Application to Meuse data

We also illustrate the application of the proposed methodology on a classical geostatistical dataset of heavy metals in the topsoil of a floodplain along the river Meuse in the Netherlands (Burrough et al., 2015). The presence of heavy metal in high concentrations can have adverse effects on human health by ingestion, inhalation, or dermal absorption (Luo et al., 2012). Hence, it is vital to study the concentrations of these elements and their spatial distribution in the soil (Mariussen et al., 2017). In particular, we study the spatial distribution of zinc concentrations (in ppm) over the floodplain area. The zinc concentrations observed at 155 locations are plotted in Figure 5.5a, and the histogram of the concentrations is plotted in Figure 5.5b. The marginal distribution is far from Gaussian, and we use the proposed flexible method to model the data, and we aim to create finer resolution maps and identify the areas that are highly polluted with exceedance maps.

We applied the proposed method of multiple indicator kriging with Gaussian copula and $t$-copula and compared it with the commonly used variogram indicator krig-
Figure 5.5: The zinc concentrations (in ppm) at 155 locations in the topsoil of a floodplain along the river Meuse. (a) Map of zinc concentrations; (b) histogram of zinc concentrations.

The predictive performance of the methods was evaluated using MAD, CRPS, and $m$CRPS using the resampling approach mentioned above and are summarized in Table 5.5. We can see from the tabulated values that the copula-based methods have a better predictive performance than the variogram approach, log-Gaussian, Gaussian kriging. Although logGK has a good CRPS score, the back transformation of kriging estimates introduces biases in its prediction. Overall, the proposed method CMIK performs the best, and it is used for spatial prediction of this Meuse data. A fine-resolution map is predicted over the study area and is plotted in Figure 5.6a, and a contour map for the probability of exceeding zinc concentrations of 500 ppm is shown in Figure 5.6b. These plots suggest that the polluted sediments were carried by the Meuse river and deposited mainly along the banks of the river. The plots highlight the areas with high concentrations of zinc and where precautions should be taken. This application demonstrates the use of the proposed methodology for spatial risk assessment for a wide range of studies.
Table 5.5: Numerical assessment of the predictive performance of GK, logGK, VMIK, and proposed CMIK$_G$ and CMIK$_t$ methods for the Meuse data.

<table>
<thead>
<tr>
<th>Measure</th>
<th>GK</th>
<th>logGK</th>
<th>VMIK</th>
<th>CMIK$_G$</th>
<th>CMIK$_t$</th>
</tr>
</thead>
<tbody>
<tr>
<td>MAD</td>
<td>95.14</td>
<td>96.7</td>
<td>82.66</td>
<td>78.34</td>
<td>77.34</td>
</tr>
<tr>
<td>CRPS</td>
<td>121.28</td>
<td>106.83</td>
<td>116.53</td>
<td>109.98</td>
<td>108.93</td>
</tr>
<tr>
<td>mCRPS</td>
<td>73.64</td>
<td>70.15</td>
<td>79.38</td>
<td>72.49</td>
<td>70.42</td>
</tr>
</tbody>
</table>

Figure 5.6: Results from the proposed CMIK$_t$ method for the Meuse data. (a) Map of predicted values of zinc concentrations, (b) contour map for the probability of exceeding 500 ppm zinc concentrations.
5.5 Discussion

In this chapter, we build a flexible model for spatial probabilistic prediction for non-Gaussian spatial data, which can be applied to various applications. We introduce a copula-based multiple indicator kriging method and describe the spatial dependence of indicator variables using copulas. The proposed method is semiparametric, and makes no explicit assumption on the marginal distribution of the random field and is resistant to outliers. It provides a complete solution to the spatial prediction problem by estimating the conditional CDF at unknown spatial locations. The proposed copula-based methods circumvent the disadvantages of existing and commonly-used variogram indicator kriging approach and are more robust. The predictive performance of the proposed methodology is evaluated using extensive simulation studies and applications to precipitation and soil pollution data, all of which demonstrate the effectiveness of the proposed method. Another exciting feature of indicator kriging is that the predictive CDF can take very different and peculiar shapes from location to location in a study area. In contrast to Gaussian kriging, where the predictive CDF can only take the form of a normal curve determined by an estimate and estimation variance at a location.
Chapter 6

Concluding Remarks

6.1 Summary

The thesis research is focused on developing robust statistical methods for multivariate, functional, and multivariate functional data with an emphasis on environmental applications. It covers a vast area, from developing advanced quantile regression techniques for evaluating salinity tolerance in crops in bivariate data; to developing visualization tools for bivariate functional data by characterizing the distribution with directional quantile envelopes and detecting bivariate extremes with application to radiosonde wind data. It also includes estimation and prediction of quantiles for multivariate functional data and construction of flexible quantile contours, applied to air pollution data, and formulation of spatial prediction problem for non-Gaussian processes with application to precipitation data and soil pollution data. This thesis research can serve as an important tool for analyzing several aspects of non-Gaussian, heavy-tailed, or skewed data, which are common in real-world processes.

The proposed advanced quantile regression techniques analyze the traits that influence salt tolerance and barley’s yield in Chapter 2. I proposed to model the joint distribution of yield in the saline and non-saline conditions, given a plant trait, to build up a more detailed picture of the entire distribution. In this way, I could focus the analyses on those plant groups that displayed higher yields and greater tolerance and examine the main influencing factors. The univariate quantile analyses quantify yield under both conditions and identify the traits affecting salinity tolerance and are
more informative than mean regression. The bivariate quantile analyses allow linking plant traits to the salinity tolerance index directly by predicting the joint distribution of yield, and it also allows a nonlinear relationship between the yield and plant traits. The analysis revealed new dimensions of plant responses to salinity that could be relevant to salinity tolerance and could potentially inform future crop decisions, particularly in arid parts of the world.

I developed a set of quantile methods for bivariate data in Chapter 2. In recent times, multivariate functional data are frequently observed in many scientific fields, and the estimation of quantiles of these data is essential in data analysis. I extend the concept of directional quantile envelopes from bivariate data to bivariate functional data in Chapter 3. Motivated by radiosonde data, I study the joint distribution of wind speed and direction along several pressure levels, which is extensively used in climate models and weather predictions. I develop an effective visualization tool to demonstrate the features of this bivariate functional data that might not be apparent otherwise. The proposed nonparametric approach is suitable for estimating bivariate distributions that are non-Gaussian, heavy-tailed, and skewed. I also proposed two methods to predict directional quantile envelopes at any pressure level using quantile regression and quantile kriging. These methods allow us to predict the bivariate distribution of wind in parts of the atmosphere where no radiosonde observations are available.

However, directional quantile envelopes coincide with Tukey’s depth contours and are convex by definition. This feature is common to most existing depth concepts which have the convexity feature embodied in their contours. Hence, they are not suitable for nonconvex distributions, including multimodal ones. I overcome this challenge in Chapter 4 by defining a flexible quantile model for multivariate functional data that can adequately characterize non-Gaussian and even nonconvex multivariate distributions marginally. The proposed quantile model is a nonparametric, time-
varying coefficient model, which uses cubic splines for the estimation and prediction. The proposed method is computationally efficient for high dimensions, while existing methods focus only on bivariate functional data. I show that the estimated multivariate quantile function is monotonic and unique, and also establish the consistency of the estimator. The proposed methods are applied to study the joint distribution of particulate matter ($\text{PM}_{2.5}$) and geopotential height over time across the Northeastern United States. The estimated bivariate contours illustrate the nonconvex features, and the functional quantile curves capture the dynamic change across six months.

Finally, I propose a flexible model for spatial prediction of non-Gaussian processes in Chapter 5, which are common in real-world applications. The proposed copula-based multiple indicator method describes the spatial dependence using copulas by exploiting the relationship between indicator covariances and copula functions. The proposed method makes no distributional assumptions of the marginal distributions and thus provides more flexibility. It predicts the conditional quantile function or cumulative distribution function at unknown spatial locations, which is more informative than point prediction and better at capturing uncertainty. The proposed method outperforms the commonly used variogram approaches and those based on Gaussianity assumption in simulations as well as in applications to precipitation and soil pollution data.

6.2 Future research work

The work presented in this thesis can be extended in the following directions. In Chapter 2, I described a bivariate quantile regression model to study the effect of plant traits on the bivariate distribution of the yield in both saline and non-saline conditions. Although the model studies the effect of only a single covariate on the response at a time, it could be extended to study multiple covariates by including spline functions for multiple covariates in the model. It would be interesting to see if
multiple bivariate quantile regression can find some new observations.

There still exist some exciting open problems that require future research in Chapter 4. For example, I consider the intercept model for visualization in our simulation and application studies, but more covariates can be added in the model to study the dependence of the covariates on the response, using the multivariate quantiles. Moreover, in the cubic spline model for multivariate functional quantiles, I fix the number of knots. To choose the number of knots adaptively, we can use the cross-validation criterion by leaving out one curve at a time and use the estimated median curve as the prediction to compute the prediction errors. Then we can choose the number of knots corresponding to the minimum mean prediction error.

In Chapter 5, I use elliptical copulas to model the spatial dependence of the indicator variables, where the correlation matrix appears explicitly in their expression. However, other flexible copulas can also be explored for modeling spatial dependence, such as $\nu$-transformed multivariate normal copula, and the family of Archimedean copulas. Although I focus the analysis on continuous spatial data in our work, the proposed methodology can be modified to accommodate categorical spatial data also by considering spatial copula classification theory to model categorical data (Huang and Wang, 2018).

Detection of change points in functional data is often of interest, and functional data analysis applies to such kind of problems. Future research can focus on detecting a change point in the mean function. It will be interesting to detect a change point in the growth curves of barley plants to identify the time when the average plant yield of the saline and non-saline conditions start to differ. A linear spline model with a single knot can be used to detect change points with the maximum likelihood approach, the details of which can be explored.


Biography

Gaurav Agarwal is a Ph.D. candidate supervised by Prof. Ying Sun in the Statistics Program at King Abdullah University of Science and Technology. He received his Master in Statistics from the Indian Institute of Technology, Kanpur in 2016, and his Bachelor in Statistics from the Hindu College, University of Delhi in 2014. His research interests include robust statistics, quantile regression, functional data analysis, and spatial statistics.

Papers Submitted and Under Preparation