Outliers Detection Based on Buantiles and Depth Functions

Fidence Munyamahoro

A Thesis
in the Department
of
Mathematics and Statistics

Presented in Partial Fulfillment of the Requirements
for the Degree of
Master of Science (Mathematics) at
Concordia University
Montréal, Québec, Canada

January 2024

CONCORDIA UNIVERSITY
School of Graduate Studies

This is to certify that the thesis prepared

By: Fidence Munyamahoro

Entitled: Outliers Detection Based On Quantiles And Depth Functions

and submitted in partial fulfillment of the requirements for the degree of

Master of Science (Mathematics)

complies with the regulations of the University and meets the accepted standards with respect to originality and quality.

Signed by the final examining committee:

Chair

Examiner

Examiner
Dr. Yogen Chaubey

Thesis Supervisor(s)
Dr. Melina Mailhot

Thesis Supervisor(s)

Approved by
Dr. Lea Popovic, Graduate Program Director

Dr. Pascale Sicotte, Dean of Faculty

Abstract

Outliers Detection Based on Quantiles and Depth Functions

Fidence Munyamahoro

Outlier detection plays a crucial role in data analysis and is employed in various domains such as finance, healthcare, and anomaly detection. This thesis presents a novel approach for detecting outliers using quantiles and depth functions, and we apply it to an air quality dataset. Quantiles provide a statistical measure of the distribution of data, while depth functions assess the centrality of observations relative to the entire dataset. Combining these two techniques, we propose a robust and effective method to identify outliers in multidimensional datasets. Our approach is particularly useful in scenarios where traditional outlier detection methods may be inadequate or fail to capture the complex patterns present in the data. By considering multiple quantiles, we can identify outliers that deviate from different aspects of the data distribution. Additionally, we incorporate depth functions, which measure the centrality of observations within a dataset, to further refine our outlier detection process. To evaluate the effectiveness of our approach, we apply it to a real-world air quality dataset. The data is about the New York Air Quality Measurements of 1973 for five months from May to September recorded daily. It contains 153 observations of 6 variables. By applying our method, we can identify outliers representing unusual air quality patterns, potentially indicating anomalies or errors in the data collection process. Our experimental findings support the proposed approach and effectively detect outliers in the air quality dataset. Compared to traditional outlier detection techniques, our method achieves higher accuracy and provides more detailed in- sights into the nature of the outliers. Furthermore, we show that the identified outliers can be valuable in understanding the factors contributing to air pollution and in improving the quality ofair quality monitoring systems. The findings of this research contribute to the advancement of outlier detection methodologies and provide valuable insights for practitioners in identifying and handling outliers in real-world applications.

Keywords: Outliers detection, Quantiles, Geometric quantiles and Depth function.

Acknowledgement

I would like to take this opportunity to express my deepest gratitude to Dr. Mélina Mailhot for her unwavering support, guidance, and invaluable insights throughout the research process. Her expertise, patience, and dedication have been instrumental in shaping this work. I extend my appreciation to Dr. Galia Dafni for her valuable support. I am indebted to all Department of Mathematics and Statistics staff members who offered beneficial facilities and insights, shared their experiences and provided a supportive academic environment. Lastly, I want to thank all those whose names may not be mentioned here but who, in various ways, contributed to my academic and personal growth. Thank you to everyone who played a part in this significant chapter of my academic life.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

In data analysis, outliers in a dataset may cause problems for the statisticians, pushing them to develop methods for identifying them. Outlier detection aims to identify abnormal patterns (outliers) from data sets. Outliers are considered irregularities, anomalies, faults, deviations, and exceptions in many applications. The source of outliers could be false sampling (samples from other than the target population or sampling unusual elements), data recording or (and) entry errors. However, according to the definition of outliers, a dataset may have many or no such cases. Both terms outlier and anomaly are used for identifying atypical observations. Nevertheless, there is a substantial difference between these two groups in their application methods. The term "outlier detection" is by statisticians, while "anomaly detection" is traditionally used by the machine learning community.

Based on the source of outliers, their treatment should be different. For example, unusual members from the target population would remain for data analysis. Those caused by data collection or entry errors will likely be removed from the sample before the data analysis. Identifying inconsistent observations from standard data is of great interest in many applications. The typical research problem underlying these applications is outlier detection, fault detection, anomaly detection, or novelty detection.

There are three major categories of outliers in statistics and data science: The first is Global outliers (or point anomalies) when the data point's value is far outside the entirety of the data set. The second is Contextual (or conditional) outliers, when the data point value significantly deviates from the rest of the observations in the same context. The third is Collective outliers, when a subset of data points in a data set deviate significantly from the entire data set.

In Statistics, detecting outliers in multivariate data is essential because that kind of data can distort any statistical analysis or procedure. In many scientific fields like quality control, finance, medicine, chemistry, and image analysis, the task of detecting multivariate outliers is valuable action. This study focuses on outliers detection based on quantiles and depth functions. In multivariate data sets, quantile contours are potent tools for detecting outliers. Recently, [34] used quantiles for anomaly detection of multivariate sensor data, and [14] brought a novel outlier de-

tection method for multivariate data. Outlier detection using quantile has also been studied in the last decade; for example, [8] introduced a detecting outliers method using geometric quantile approaches by plotting quantile contours.

1.1 The notion of quantiles

Quantiles are statistical measures used to divide a dataset into equal portions, typically into quarters (called quartiles), tenths (deciles), or hundredths (percentiles). They are useful for understanding the distribution of a dataset and identifying key points within it. More formally, for a probability distribution or a dataset, a quantile is a value that divides the distribution or dataset into segments with equal probabilities or equal frequencies. They are powerful tools for understanding data distribution and have a wide range of applications across various fields, from statistics and finance to healthcare and machine learning.

The most common quantiles are: 1. Median (Q_2): The 50^{th} percentile divides the data into two equal halves. Half of the data values are greater than the median, and half are smaller. 2. Quartiles (Q_1 and Q_3): These are the 25^{th} and 75^{th} percentiles dividing the data into four equal parts, respectively. The first quartile (Q_1) represents the 25% of data below it, and the third quartile (Q_3) represents the 25% of data above it. 3. Percentiles: General quantiles that divide the data into 100 equal parts. For instance, the 90^{th} percentile represents the value below which 90% of the data falls. Quantiles preserve the order of the data. For instance, Q_2 (median) will always lie between Q_1 and Q_3 in a sorted dataset. They are less affected by extreme values (outliers) compared to the mean. This makes them helpful in analyzing datasets with outliers. In a dataset with distinct values, quantiles are unique. However, there might be different ways to calculate quantiles in datasets with repeated values.

Quantiles are essential tools in data analysis and have various applications: In finance and insurance, quantiles are used to calculate Value at Risk (VaR) - a measure of potential losses in a portfolio or an investment. In manufacturing, quantiles can be used to monitor the variation in production processes, identifying potential defects or deviations. In medical studies, quantiles can be used to understand the distribution of various health indicators, such as blood pressure or cholesterol levels. Boxplots (box-and-whisker plots) use quartiles to visually represent the data distribution, showing the median, quartiles, and potential outliers. In chapter 4 of this thesis, we used box-and-whisker plots to detect outliers from air quality data. Quantiles are used in machine learning for techniques like quantile regression, which estimates conditional quantiles, providing a more comprehensive view of the relationship between variables. They also use a Q-Q plot to compare the distribution of the data set. The Q-Q plot is created by plotting the dataset's quantiles against the theoretical distribution's quantiles. Q-Q plot (Quantile-Quantile Plot): A Q-Q plot is

a graphical technique used to compare the distribution of a dataset to a theoretical distribution, typically the normal distribution. The main idea is to visualize whether the data points follow a particular theoretical distribution. In a Q-Q plot, the x-axis represents the assumed distribution's theoretical quantiles (expected values), and the y-axis represents the observed quantiles from the analyzed dataset. If the data closely follows the theoretical distribution, the Q-Q plot points will fall approximately along a straight line. Deviations from a straight line indicate departures from the assumed distribution. The drawback of the Q-Q plot is its limitation with variables; it is a popular tool of only univariate analysis. Geometric quantile can overcome this limitation; [11] used this notion for multivariate data. Overall, the Q-Q plot is a useful visual tool for comparing the distribution of your data to a theoretical distribution and detecting departures from it. It is widely used in statistical analysis to assess the appropriateness of various assumptions underlying data modelling and to identify potential data outliers or discrepancies.

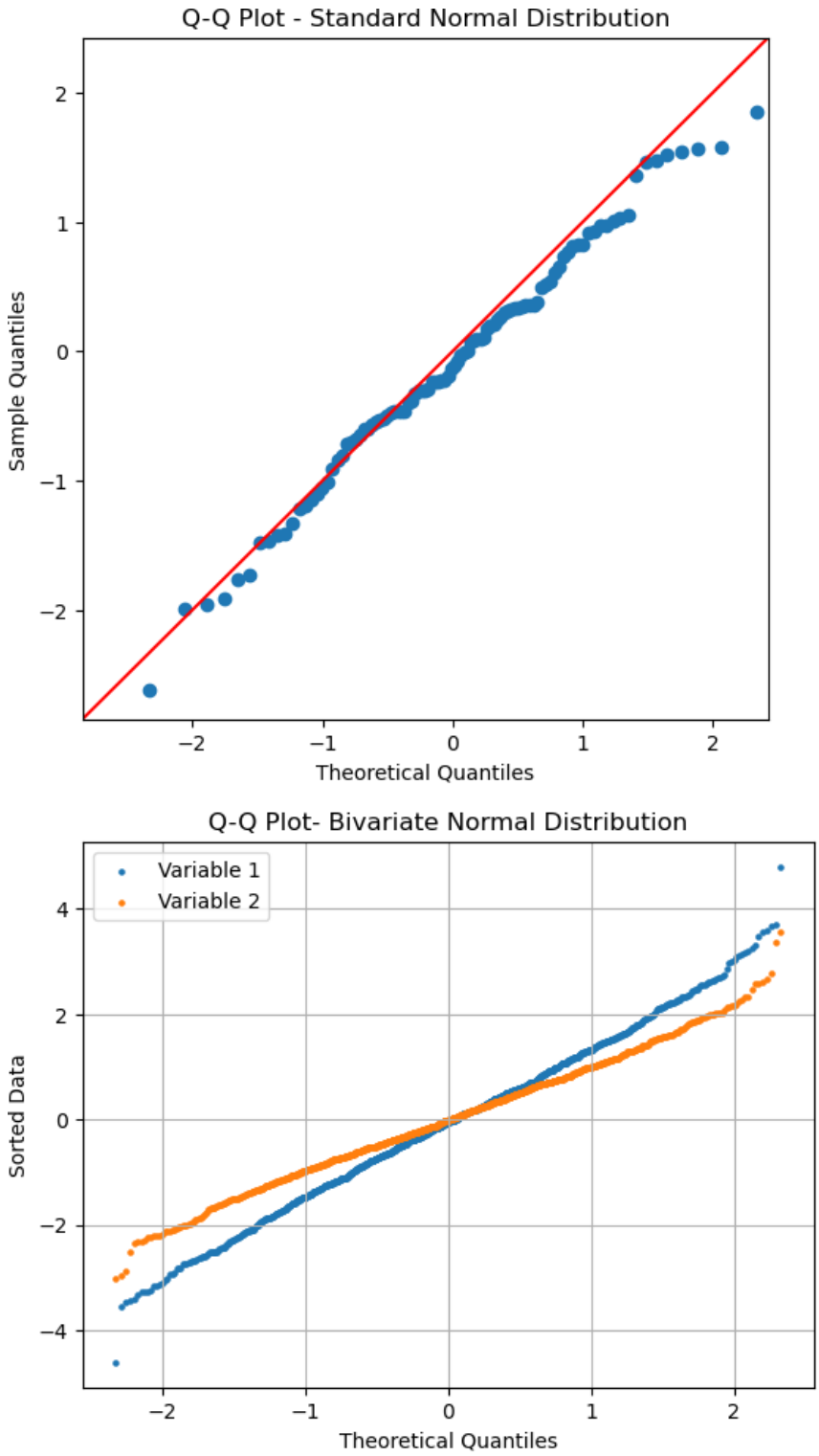

Figure 1.1: Quantile-Quantile(Q-Q) plot

At the top of the Figure 1.1, a Q-Q plot from 250 observations simulated from a standard Normal distribution, and in the bottom, the Q-Q plot from 1000 observations from a bivariate Normal distribution $N(0, \Sigma)$ with $\Sigma = \begin{pmatrix} 2 & 0.9 \\ 0.9 & 1 \end{pmatrix}$. If the data perfectly follows the theoretical distribution, the points in the Q-Q plot will fall along a straight line, usually at a 45-degree angle. Deviations from this line suggest differences between the observed data and the theoretical distribution. In Figure 1.1, the top image shows a Normal distribution plotted using simulated 250 observations, resulting in a reasonably straight line which means the data fit with the Normal distribution. In contrast, the bottom image gives information about the bivariate Normal distribution. The Q-Q plots provide the idea of selecting the distribution that can be used in data analysis.

1.1.1 Univariate quantiles

Univariate quantiles are values that partition the distribution of a univariate dataset into intervals with specified proportions of the data falling below them. In simpler terms, they represent pointsin a dataset below which a certain percentage of the data falls. They provide a way to measure individual data points' relative position or rank within a distribution. The quantiles allow us to understand the spread and distribution of data and are commonly used in statistics and data anal- ysis. Consider a real random variable X having distribution function F and $0 < \tau < 1$, then a τ^{th} quantile of X is any number x such that $Pr(X \leq x) \geq \tau$ and $Pr(X \geq x) \geq 1 - \tau$. It can be written as $Q(\tau, X) = \inf\{x \in \mathsf{R} : F(x) \geq \tau\}$. This expression emphasizes that given the value of τ, we needto find some x, which results in $F(x)$ returning a value not less than τ.

However, we could find many values of x that meet the condition, and if this happens, we take the smallest x of those values.If random variable X is continuous, then $Q(\tau, X)$ satisfies $F_X(Q(\tau, X)) = Pr(X \leq Q(\tau, X)) = \tau$. Let $X_{1:n}, X_{2:n}, X_{3:n}, \cdots, X_{n:n}$ be the order statistics from therandom sample $X_1, X_2, X_3, \cdots, X_n$, then the sample estimator of ordinary quantiles function can be expressed as follows:

$$\hat{Q}_n(\tau) = \hat{F}_n^{-1}(\tau) = \inf\{x \in \mathsf{R} : \hat{F}_n(x) \geq \tau\} = X_{\lfloor n\tau \rfloor + 1:n} \tag{1.1.1}$$

where $\hat{F}_n(x)$ is the empirical cumulative distribution of random sample X, and $\lfloor \cdot \rfloor$ denotes the greatest integer part. In Table 1.1, we find the values of $\hat{Q}_n(\tau)$ from the binary dataset of size six with $\tau = 0.25, 0.50, 0.75$.

Level τ	Data1 {0, 0, 0, 0, 0, 1}	Data2 {0, 0, 0, 0, 1, 1}	Data3 {0, 0, 0, 1, 1, 1}	Data4 {0, 0, 1, 1, 1, 1}	Data5 {0, 1, 1, 1, 1, 1}
0.25	0	0	0	0	1
0.50	0	0	1	1	1
0.75	0	1	1	1	1

Table 1.1: Values of $\hat{Q}_n(\tau)$ from binary dataset of size six.

When the values of the product $n\tau$ are non-integer values, this classical estimate could be inappropriate. A good way to handle this problem is to perform the interpolation between two order statistics with indices closest to the value of $n\tau$.

Given the dataset, the quantiles plot shows the percent of fraction of points below the given value. If F is invertible, we write it as

$$Q(\tau, X) = F_X^{-1}(\tau).$$

However, another equivalent definition of the τ^{th} quantile is expressed as the solution to an optimization problem. Let $E(X)$ be the expectation of X, it is known (from derivation) that $E(X) = \text{argmin}_{\tau \in \mathbb{R}} E[(X - \tau)^2]$ and the median is a solution of $\text{med}(X) = \text{argmin}_{\tau \in \mathbb{R}} E[(X - \tau)]$. We also know that the median is the 0.5-quantile; we can show that the τ^{th} quantile of X can be given by $Q(\tau, X) = \text{argmin}_{\tau \in \mathbb{R}} E[\rho_\tau (X - \tau)]$, where ρ_τ is called a check function. The check function ρ_τ is also called the convex loss function, and it is a generalization of the absolute value for $0 < \tau < 1$.

1.1.2 Some beneficial properties of quantiles

Quantiles are statistical measures that divide a dataset into equal-sized groups or segments. The translation and scale ivariance properties of quantiles refer to how they behave under translations (shifts) and rescaling (changes in scale) of the data [2]

- **Scale invariance:** For every $\tau \in (0, 1)$, any arbitrary positive constant $a \in \mathbb{R}$, and for the given random variable X, $Q(\tau, X)$ is scale-invariant.

$$\begin{aligned} Q(\tau, aX) &= \inf\{v : P(aX \le v) \ge \tau\}, \qquad \text{for } v = au \\ &= \inf\{au : P(aX \le au) \ge \tau\} \\ &= a \inf\{u : P(X \le u) \ge \tau\} \\ &= aQ(\tau, X) \end{aligned}$$

Quantiles are scale-invariant, meaning multiplying all data points by a constant factor will not

alter the quantiles. In other words, if you scale the entire dataset (multiply each data point by a constant), the quantiles will remain unchanged. This property ensures that quantiles are not affected by changes in the spread or variability of the data. For example, if we have a dataset and calculate the 25^{th}, 50^{th}, and 75^{th} percentiles and then multiply each data point by a constant factor, the new dataset's 25^{th}, 50^{th}, and 75^{th} percentiles will remain at the same relative positions in the shifted dataset.

- **Translation invariance:** For every $\tau \in (0, 1)$, any arbitrary positive constant $b \in \mathbb{R}$, and for the given random variable X, $Q(\tau, X)$ is translation invariant.

$$\begin{aligned} Q(\tau, X+b) &= \inf\{v : P(X+b \le v) \ge \tau\} \qquad \text{for } v = u+b \\ &= \inf\{u+b : P(X+b \le u+b) \ge \tau\} \\ &= \inf\{u : P(X+b \le u+b) \ge \tau\} + b \\ &= \inf\{u : P(X \le u) \ge \tau\} + b \\ &= Q(\tau, X) + b \end{aligned}$$

Translation invariance: Quantiles are translation invariant, meaning that adding a constant value to all data points in a dataset will not change the quantiles. In other words, if you shift the entire dataset by a fixed amount, the quantiles will remain the same. This property is helpful because it ensures that quantiles are not affected by changes in the location (center) of the data. For example, if we have a dataset and calculate the 25^{th}, 50^{th}, and 75^{th} percentiles (quantiles) of the original dataset, and then add a constant value to each data point, the 25^{th}, 50^{th}, and 75^{th} percentiles of the new dataset will be shifted by the same constant value and the quantiles would still be at the same relative positions.

These properties make quantiles valuable for summarizing and comparing datasets because they focus on the relative distribution of data points rather than their absolute values, making them robust to changes in location and scale.

1.1.3 Conditional Quantiles

Conditional quantile is a statistical concept used to estimate and analyze the distribution of a random variable Y given certain conditions or values of other variables X. In other words, they provide information about how the quantiles of Y change with different values of X. Conditional quantiles are particularly useful in understanding how one variable is influenced by or associated with another variable. In 1978, [23] introduced the notion of conditional quantiles in parametric framework, and since then, they have extended their investigation to other frameworks. Estimating conditional quantiles is useful in various applications, such as risk management, econometrics,

finance, and environmental sciences. It allows us to examine how a particular response variable (Y) varies across different predictor variable levels (X) and how the relationship between the two variables changes across the distribution.

When we talk about conditional quantiles, we consider quantiles of one variable under the condition or constraint of another variable. In other words, we're interested in how the distribution of one variable (let's call it Y) changes or varies depending on the value of another variable (let'scall it X). For instance, in environmental science, you should study how the concentration ofa pollutant (Y) varies with temperature (X). You could calculate conditional quantiles of Y for different values or ranges of X. To calculate conditional quantiles, you would first partition your dataset into different groups or categories based on the values of X. Then, within each group, you calculate the quantiles of Y. This helps you understand how the quantiles of Y change as X varies. Figure 1.2 shows Pareto and Normal distributions' conditional mean and conditional quantiles.We simulated 200 observations for both Pareto and Normal distributions; at the upper left corner arethe scatter plots of the Normal distribution with conditional mean; the upper right is a conditionalquantile of the Normal distribution. Then, on the bottom left is a scatter plot of asymmetric Pareto with conditional mean, and on the bottom right is a conditional quantile of asymmetric Pareto. From the Normal distribution, we observe that the distance among responses increases directly to the value of x. Furthermore, in the upper right of Figure 1.2, we observe that the true conditionalmean and the actual conditional median tend to coincide. In contrast, in the Pareto plot (bottom left), we observe a significant difference between the true mean and estimated conditional median due to the distribution's asymmetry.

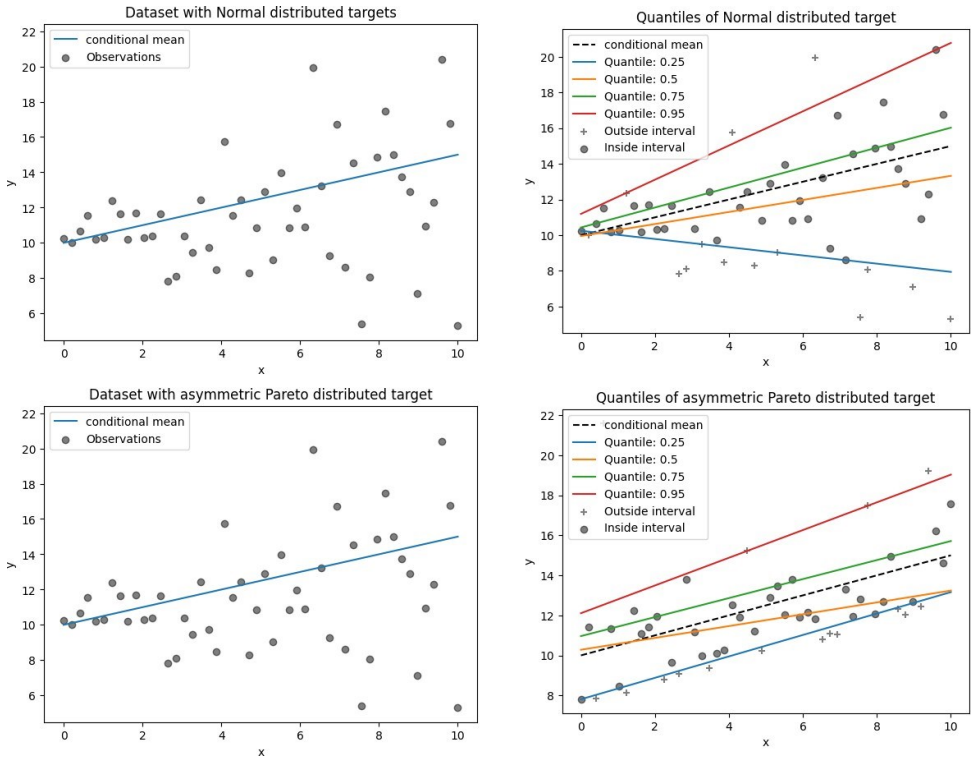

Figure 1.2: Conditional quantile of normal vs Pareto distribution.

The conditional quantiles have many applications in areas concerned with the effects of the covariates on the outcome. For example, an insurance company concerned about knowledge of the claim size that will exceed the financial coverage of a policyholder (say threshold) will use covariates variables to calculate the conditional mean and variance of their variable of interest. We define conditional quantiles analogously to standard quantiles. Let us take a random variable Y distribution depending on the covariate X. The conditional quantile function of Y given X is the inverse of the corresponding conditional distribution function, i.e.,

$$Q_\tau = Q_Y(\tau|X) = F_Y^{-1}(\tau|X = x) = \inf\{F_Y(y|X = x) \geq \tau\}, \tag{1.1.2}$$

where $F_Y(y|X) = P(Y \leq y|X)$, and $0 < \tau < 1$. τ represents the desired probability level, such as$\tau = 0.25$ for the first quartile, $\tau = 0.5$ for the median, and $\tau = 0.75$ for the third quartile.The estimation of $Q_\tau\ (X)$ can be achieved by replacing the conditional expectation with a suitable estimator and then by solving the minimization problem. We are more concerned with estimating $Q_\tau\,(X)$ when $\tau \uparrow 1$. The conditional function of Y given X fully captures the relationship between$Y$ and covariates X. Conditional quantiles are key to investigating how covariates X affect the re- sponse variable Y. When $E(Y|x) < \infty$, the quantile function defined in equation 1.1.2 can minimize

expected asymmetric loss in a conditional way:

$$Q_\tau(x) = \min_q E[\rho_\tau(Y - q)|X = x], \tag{1.1.3}$$

where the check function ρ_τ is defined as $\rho_\tau(t) = \tau t 1_{\{t \geq 0\}} - (1 - \tau) t 1_{\{t \leq 0\}}$ [23]. Conditional quantiles help us understand how the distribution of one variable is influenced or conditioned byanother variable. They are valuable in various fields for gaining insights into relationships and making data-driven decisions.

1.1.4 Geometric Quantiles

Geometric quantile is a concept in statistics and probability theory used to divide a dataset or probability distribution into equal-sized subsets based on the geometric mean. They are particularly useful when dealing with data that is positively skewed or follows a multiplicative pattern. In a multivariate set-up, geometric quantiles are viewed as an extension of univariate quantiles and are helpful in data analysis of complex surveys. For example, geometric quantiles are essential in detecting outliers employing quantile contour plots in a multivariate data set. This application attracted many researchers, such as [8], who introduced geometric quantile approaches for plot- ting quantile contours and [24] defined geometric quantile as an extension of multivariate quantiles based on the geometry of the multivariate data set and norm minimization. De Gooijer [21] also ap- plied the multivariate quantiles method for financial time series analysis. In multivariate data sets, quantile contour plots are potent tools for detecting outliers, and the region enclosed by quantilecontours is considered a multivariate analog of box plots.

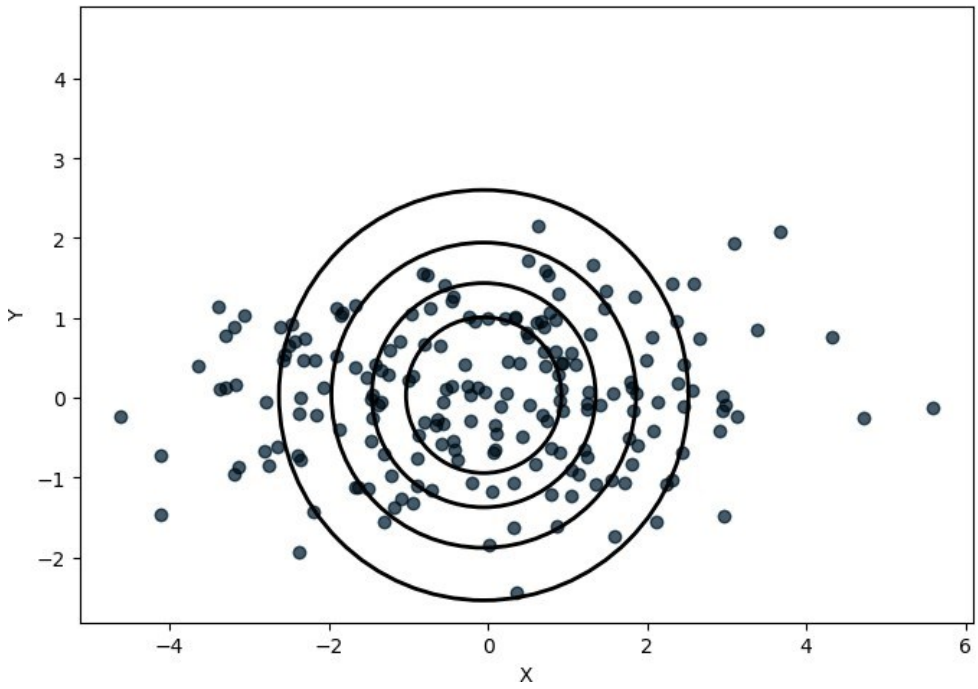

Figure 1.3: Geometric quantiles contour plot of bivariate Normal.
Geometric quantiles contour plot for $\tau = 0.20, 0.40, 0.60$ and $\tau = 0.80$ of 200 points simulated from $N(0, \Sigma)$ distribution with $\Sigma = \begin{pmatrix} 3 & 0.4 \\ 0.4 & 1 \end{pmatrix}$.

In this visual representation (Figure 1.3), by employing a quantile threshold of 0.80, observations lying outside the contour line associated with the $\tau = 0.80$ quantile are designated as outliers. This technique leverages the distribution of the data points, with the contour line delineating the boundary beyond which observations significantly deviate from the expected pattern. Consequently, points positioned beyond this contour line are deemed potentially anomalous or aberrant instances within the dataset, warranting further investigation or consideration in subsequent analyses. This concept of identifying outlier observations through the utilization of a quantile contour plot was discussed in details by [8].

However, plotting geometric quantile contours requires many steps and a recursive Newton–Raphson method, which is complicated and time-consuming when dealing with many observations. In many applications, Multivariate quantile contours are helpful and have been studied by many researchers; for example, [35] introduced quantile functions for multivariate analysis. In addition, different technics of multivariate quantiles were suggested by [5, 8, 24, 32, 35]. However, although quantile contour plots are potent tools for detecting outliers in multivariate data sets, there is no easy solution without solid assumptions on selected distributions.

Geometric quantiles have many valuable properties, which lead to much usage in real applica-

tions. Those properties can be classified into two parties:

(i) Central properties of geometric quantiles are referred extensively to as the median. It is challenging to deal with large samples of high-dimensional data; nowadays, most data are such. A median is defined as the most central quantile in the concept of a multidimensional quantile. Many methods in literature (for example, least square) are sensitive to outliers which is the drawback in outlier detection. As an indicator of centrality, the median can overcome this weakness in high-dimensional data where outlying observations should be brutal to detect. The geometric median also called the spatial median, is given by:

$$\text{med}(X) = \underset{q\in\mathsf{R}}{\text{argmin}} E[\|(X-q)\| - \|X\|],$$

where X is a random variable taking values in R.

(ii) Extreme geometric quantiles $Q(u)$ when $\|u\|$ close to one: In this thesis, we focus on outliers detection based on geometric quantile introduced by [11]. The author of [11] introduced geometric quantiles based on the euclidean distance, which also gives the idea of extreme observations. The insurer wants to know the claim size to avoid bankruptcy. Most insurance data are multivariate; the quantile contour plot will help them locate extreme observations (outliers). In this study, we will use a quantile contour plot to analyze the behaviour of extreme observations.

For fixed probability $\tau \in (0,1)$ and for $t \in \mathsf{R}$, let us define the loss function used in [23] as follows:

$$\varphi(1-2\tau, t) = \|t\| + (1-2\tau)t. \tag{1.1.4}$$

where $\varphi(.,.)$ is the multivariate loss function, with $\|\cdot\|$ being the usual euclidean norm. The function $\varphi(.,.)$ is well-defined as for any given input $(1-2\tau, t)$, the output $\|t\| + (1-2\tau)t$ is unique. This can be shown by demonstrating that the expression on the right-hand side of the equation yields a unique value for any given τ and t. The norm of t is always non-negative, and scalar multiplication is straightforward; thus, the sum of these two terms results in a unique valuefor any τ and t.

Let $u = 1-2\tau$ be one-to-one function that maps the open unit interval $(0,1)$ onto the open interval $(-1,1)$; according to [11] for a univariate Y, if $E(Y) < \infty$ then

$$Q(u) = \underset{q\in\mathsf{R}}{\text{argmin}} E[\varphi(u, Y-q)]. \tag{1.1.5}$$

As $\varphi(.,.)$ is a well-defined function, the difference $(\varphi(u, Y-\theta) - \varphi(u, Y))$ is also well-defined and its expected value should exist provided that the expected value of each individual term exists. Hence $E[\varphi(u, Y-\theta) - \varphi(u, Y)]$ is well-defined.

Using the idea of [22], if these two functions in $E[\varphi(u, Y-\theta) - \varphi(u, Y)]$ admit the minimum, it will be the same value for both functions . This leads us to the new definition of geometric quantiles as follows:

$$Q(u) = \underset{q \in \mathbb{R}}{\operatorname{argmin}} E[\varphi(u, Y - q) - \varphi(u, Y)]. \tag{1.1.6}$$

We observe u because $\|u\|$ close to one corresponds to an extreme quantile, whereas $\|u\|$, close to zero, corresponds to a central quantile. We can show that in the univariate setting ($d = 1$), the multivariate geometric quantile function 1.1.7 reduces to the univariate quantile function $F_Y^{-1}(u)$. First, let us rewrite the multivariate geometric quantile function $Q(u)$ in terms of a scalar variable q, representing the quantile level. We minimize the expected value of φ with respect to q:

$$Q(u) = \underset{q \in \mathbb{R}}{\operatorname{argmin}} E[\varphi(u, Y - q) - \varphi(u, Y)]. \tag{1.1.7}$$

Now, let's consider $d = 1$, then u is a scalar. So, u can be treated as a constant. Let us denote $\varphi(u, Y) = \varphi_u(y)$, the expression 1.1.7 can be written as follows:

$$Q(u) = \underset{q \in \mathbb{R}}{\operatorname{argmin}} \int [\varphi(u, Y - q) - \varphi(u, Y)] dF(y)$$

$$Q(u) = \underset{q \in \mathbb{R}}{\operatorname{argmin}} \int [\varphi_u(y - q) - \varphi_u(y)] dF(y)$$

$$Q(u) = \underset{q \in \mathbb{R}}{\operatorname{argmin}} \int \varphi_u(y - q) dF(y) - \int \varphi_u(y) dF(y)$$

$$Q(u) = F_Y^{-1}(u)$$

Hence, when $d = 1$, $Q(u)$ can be reduced to the univariate quantile in the univariate setting.

In their article,[18] have found some harmful properties of geometric quantiles, such as: The magnitude of extreme geometric quantiles tends to infinity, and asymptotic geometric quantiles are those with $\|u\|$ close to 1. These properties explain that even if a random variable has compact support, the norm of extreme geometric quantiles diverges to infinity. Their results from theorem 2 conclude that when the variance is the smallest, the norm of an extreme geometric quantile is the largest. The shapes of iso-density surfaces and extreme geometric quantile contours are orthogonal for elliptically contoured distributions.

1.1.5 Geometric Conditional Quantiles

Let us consider a k-dimensional covariate X and a d-dimensional response variable Y to extend the geometric quantile to the conditional geometric quantile. For any vector $u = (u_1, u_2, \cdots, u_d)^T$ belonging to the open unit ball $B^d = \{u | u \in \mathbb{R}^d, \|u\| < 1\}$ and for any $t \in \mathbb{R}^d$, let $\varphi(u, t) = \|u\| + \langle u, t \rangle$ be the multivariate loss function, with $\|\cdot\|$ being the usual euclidean norm and $\langle ., . \rangle$ the associated scalar product. Then, we express the u^{th} geometric conditional quantile of Y given $X = x$ as follows:

$$Q(u|x) = \underset{q \in \mathbb{R}}{\operatorname{argmin}} E[\varphi(u, Y - q) - \varphi(u, Y) | X = x]. \tag{1.1.8}$$

The vector u plays a significant role in the estimation of a quantile $Q(u|x)$ because it determines the order of the quantile with the fact that for $\|u\|$ close to 1 corresponds to an extreme quantile, whereas $\|u\|$ which is closed to zero corresponds to a central quantile. In our estimation, we will focus on extreme quantiles, that is, $|u|$ close to 1. For any non-zero vector $v \in \mathbb{R}^d$, let define $S(v) = \frac{v}{\|v\|}$ and if Y is an absolutely continuous random variable; then the conditional geometric quantile can be seen as the solution y of the following equation [10]:

$$E[(S(y - Y) | X = x] = u. \tag{1.1.9}$$

Conditional quantile analysis helps model spatial dependence structure and construct confidence intervals. However, there are few potential applications of conditional spatial quantile in literature, maybe because the classical extension of conditional quantile estimation for dependent random variables to spatial quantile regression is not trivial.

1.2 The notion of depth function

Statistical depth functions are mathematical constructs used in multivariate statistics to measure the centrality or outlyingness of a data point relative to a data set. In multivariate data analysis, statistical depth functions have been proven to be a beneficial nonparametric method. A depth function, noted by $D(x, F)$, is a nonnegative real-valued function that measures the centrality or closeness of a point $x \in \mathbb{R}^d$ with respect to a distribution function F, given the ordering of points $x \in \mathbb{R}^d$ and the distribution function F.

In the context of a depth function in $\mathbb{R}^d$, the ordering of points typically refers to the arrange- ment of points in terms of their "depth" or how far they are from a reference point or hyperplane.The ordering of points would arrange the points from the deepest to the shallowest. In other words, points with higher depths are considered to be more central or less "outlying" than those with lower

depths. The statistical data depth function determines how centrally a point should be located in a data cloud.

Depth functions can measure the outlying-ness or extremeness of a data point with respect to a given data set; hence they can detect extreme observations relative to the rest of the observations, called outliers. For $d > 1$, the depth function is considered a method of rank and median general- izing to multivariate data. The element of the dataset with the largest depth is called the median of that set, and a point in R^d with the largest depth is called the center of the dataset. As they summa- rize the location of points of the dataset as a single point, the center of the dataset and median arecalled location estimators. The empirical depth function $D(x, F_n)$ can be used in ordering datasets with empirical distribution F_n.

1.2.1 Halfspace depth (or Tukey depth)

The halfspace depth was named after the prominent statistician John Tukey [36] with a vital role in ordering multivariate data. Tukey depth is particularly useful for identifying outliers and assessing the centrality of data points.

The Tukey depth of a point $x \in \mathsf{R}^d$ with respect to a probability distribution F is defined as the most negligible probability mass of any closed halfspace containing x, and it is given by:

$$D_H(x, F) = \inf\{F(H)|H \text{ closed halfspace}, x \in H\}. \tag{1.2.1}$$

It is straightforward to say that lower halfspace depth is associated with greater outlyingness. Tukey depth is flexible and famous because of its possession of many desirable properties: upper semicontinuous, quasiconcave, monotonicity relative to the deepest point and affine invariant. To compute the Tukey depth of a point in a dataset, we first choose a point in the dataset for which you want to compute the Tukey depth. This point is often referred to as the "center" or "reference" point. Then, we calculate the depth of the reference point. The Tukey depth of the reference point is determined by calculating the fraction of data points in the dataset that are farther from the reference point than it is. In other words, it's the proportion of points in the dataset that are more "outlying" than the reference point. Mathematically, if you have a dataset of n data points and you are interested in the depth of a reference point x, you would calculate it using equation 1.2.1 as follows:

$$D_{Hn}(x, F_n) = \frac{1}{n}\sum_{i=1}^{n} \min\{\#\{x_i : x_i \in H\}|x \in H\}. \tag{1.2.2}$$

We can also use the following expression involving the distance between data points and the center

point:

$$D_{Hn}(x, F_n) = \frac{1}{1 + \frac{1}{n}\sum_{i=1}^{n} d(x_i, c)} \tag{1.2.3}$$

where: x_i is the i-th data point, c the center point, $d(x_i, c)$ the Euclidean distance between the data point x_i and the center point c, and n is the total number of data points. The Euclidean distance between point x_i and the center point c in n-dimensional space can be calculated using the following formula:

$$d(x_i, c) = \sqrt{\sum_{j=1}^{n} (x_{ij} - c_j)^2}$$

with $x_i = (x_{i1}, x_{i2}, x_{i3}, \ldots, x_{in})$ the coordinate vector of the data point x_i in n-dimensions and $c_j = (c_1, c_2, c_3, \ldots, c_n)$ the coordinate vector of the center point c. This formula computes the square root of the sum of the squared differences between corresponding coordinates of the two points.

We can see that both equations 1.2.2 and 1.2.3 capture the notion of the "average proximity" of the points in H, whether it's in terms of the number of points in H or their average distance fromthe centroid c.

The equation 1.2.2, computes the average minimum count of elements in H among all the samples x_i. It is essentially measuring the "density" of points in H where the term $\#\{x_i : x_i \in H\}$ represents the number of points x_i in H. This term is essentially counting how many points are there in the set H. Let us denote: $m_i = \min\{\#\{x_i : x_i \in H\} | x \in H\}$. This represents the minimum count of points in any half-space defined by x_i among all points in the sample. Then, the equation **1.2.2** becomes: $D_{Hn}(x, F_n) = \frac{1}{n}\sum_{i=1}^{n} m_i$.

Now, we notice that m_i is essentially counting the number of points in the half-space defined by x_i that x belongs to. This is similar to the concept of distance when considering the center of the data. So, we can rewrite m_i in terms of the distance as:

$m_i = \frac{1}{1+d(x_i, c)}$.

By substituting this into equation 1.2.2, we get: $D_{Hn}(x, F_n) = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{1+d(x_i, c)}$ and this expression is exactly the same as equation 1.2.3. Thus, we have shown that quation 1.2.2 and Equation **1.2.3** are equivalent.

As these two measures increase or decrease in the same way with respect to changes in x and F_n, they are indeed equivalent.

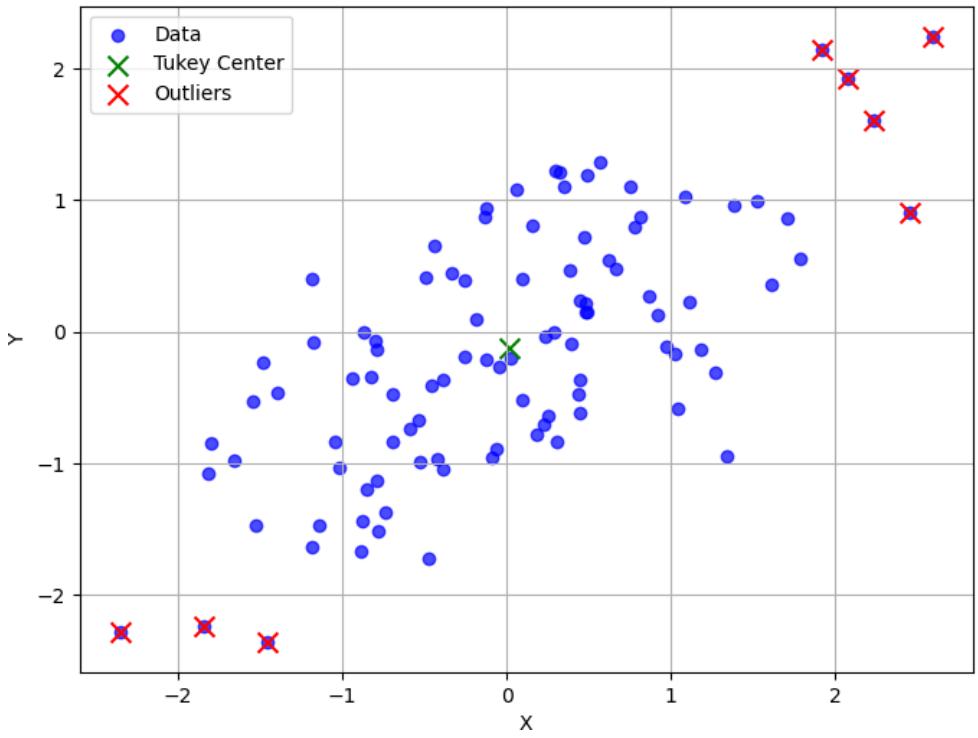

Figure 1.4: Tukey outliers detection.

To create the Figure 1.4 that demonstrates Tukey depth outlier detection for a bivariate normal distribution, we generate a bivariate normal distribution with mean zero and $\Sigma = \begin{pmatrix} 1 & 0.7 \\ 0.7 & 1 \end{pmatrix}$, calculate the Tukey depth for each data point, identify outliers based on a threshold, and then plot the data points, highlight the outliers in red and indicating the Tukey center in green. Higher Tukey depth values indicate that a point is more central within the dataset, while lower values suggest that a point is more of an outlier. Tukey depth is a valuable tool in robust statistics and data analysis because it's resistant to the influence of outliers. It provides a way to assess centrality based on the distribution of data points rather than relying solely on traditional measures like the mean or median, which can be heavily affected by outliers.

1.2.2 Mahalanobis depth

Mahalanobis depth is based on the transform of Mahalanobis distance proposed by Mahalanobis [26]. Mahalanobis distance is a measure used in statistics and machine learning to quantify the distance between two points in a multidimensional space, while taking into account the correlation structure of the data. The formula for Mahalanobis distance between two points X_i and X_j in a dataset with mean vector μ and covariance matrix $\sum$ is given by:

$$MD_{ij} = \sqrt{(X_i - X_j)^T \sum^{-1} (X_i - X_j)}$$

Mahalanobis depth is a statistical measure used to determine the distance of a data point from the center of a dataset, taking into account the covariance structure of the data. It is an extension of the concept of multivariate distance and is particularly useful when dealing with datasets that have correlated variables. The Mahalanobis depth of a data point measures how far away it is from the center of the data, considering the correlation between variables and the spread of the data along each axis. Given the ordering of points $X = x_1, \cdots, x_n \in \mathsf{R}^n$ and distribution function F, the Mahalanobis depth is given by:

$$D_M(x, F) = \frac{1}{1 + (x - \mu)^T \sum^{-1} (x - \mu)}, \tag{1.2.4}$$

where μ is the mean of distribution F, $\sum^{-1}$ is the inverse of the covariance matrix. The empirical Mahalanobis depth for a point x in a dataset $x_1, \cdots, x_n$ with sample mean $\bar{x}$ is given by:

$$D_{Mn}(x, F_n) = \frac{1}{1 + (x - \bar{x})^T S^{-1} (x - \bar{x})}$$

where S is the sample covariance matrix of the dataset. This empirical depth measures how deep the point X_i is within the distribution F . In the denominator, the second term is called the Mahalanobis distance. Mahalanobis distance is a distance metric that computes the distance between the point and distribution. It uses a covariance matrix of variables to find the distance between the center and data points, which is effective method in multivariate analysis. This means that Mahalanobis depth detects outliers based on the distribution pattern of data points, contrary to the euclidean distance.

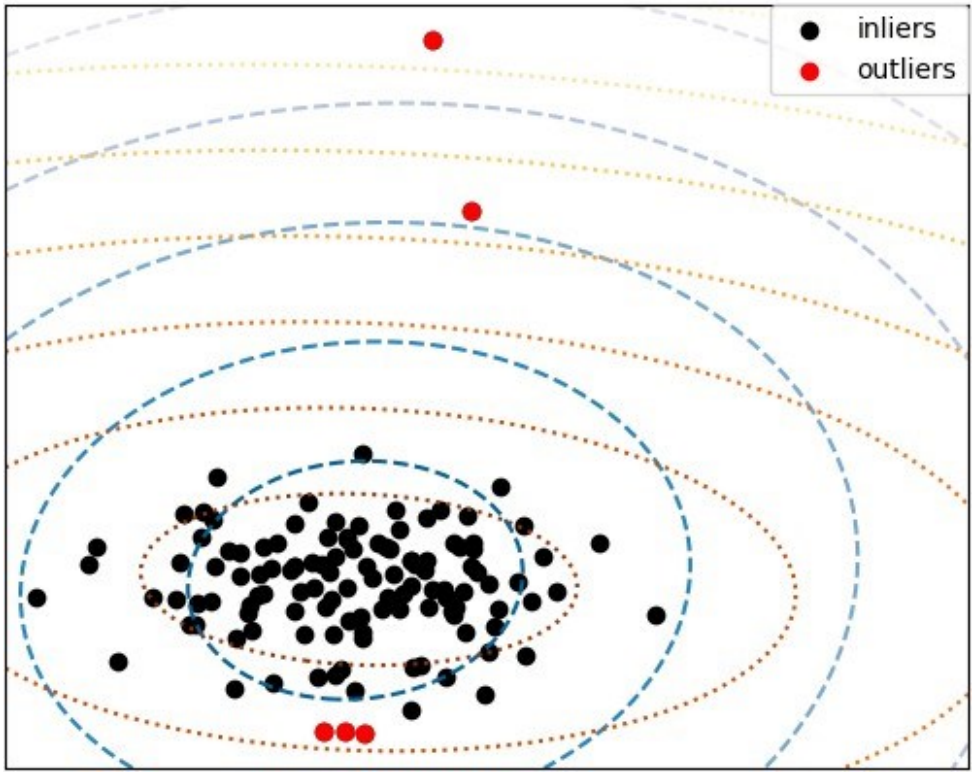

Figure 1.5: Mahalanobis outlier detection.

In the multi-dimensional space of variables, two points may look to have the same distance from the mean, yet one of them is far away from the data cloud (thus, it is an outlier), which is marked as red in the figure 1.5. In this space, the euclidean metric can fail to find the correct distance as it tries to get a straight-line distance.

1.2.3 Spatial depth function.

Spatial depth is a measure that quantifies how deep a point lies within a dataset. It measures how outlying or central a point is concerning other points in the dataset. The notion of spatial depth introduced by Brown [7]was used in many researches. For example The author of [35] used this introduction and studied multivariate spatial depth based on the geometry of the data. He considered a random variable Y with distribution F and the spatial depth of point $x \in \mathsf{R}^d$ related to F is given by:

$$D_s(x, F) = 1 - \left\| \int S(x - y) dF(y) \right\|$$

and the sample functional spatial depth of point $x \in \mathbb{R}^d$ related to F is given by:

$D_{s,n}(x, F_n) = 1 - {}_n\frac{1}{\sum_y} S(x-y) \|$

where $\|.\|$ is the Euclidean norm in $\mathbb{R}^d$ and $S(\cdot)$ is a spatial sign function given by:

$$S(x) = \begin{cases} \frac{x}{x}, & \text{when } x \neq 0 \\ 0, & \text{when } x = 0 \end{cases}.$$

This formula helps to find maximal depth points like spatial median points, central points and outlying points. the spatial depth is discussed in chapters 3, and 4.

The remaining components of this thesis are presented in the following manner: Chapter 2 presents the literature review summarizing outlier detection's research background using quantiles and depth functions. Chapter 3 details the method of analysis and computation of quantiles and depth functions for detecting outliers. We explain the procedures, methods and assumptions used in computations and the identification of outliers using quantiles and depth functions. Chapter 4 represents the application of quantiles and depth function to detect outliers using air quality data. Finally, Chapter 5 brings a conclusion that summarizes the results found and their contributions to the area of study, and it also provides suggestions for future research.

Chapter 2

Literature Review

In multivariate data, detecting outliers is essential because that kind of data can give misleading results or false accounts to the statistical analyst. In many scientific fields like quality control, finance, medicine, chemistry, and image analysis, detecting outliers is helpful. Detecting outliers based on quantiles and depth functions is a statistical technique used to identify observations that significantly deviate from the majority of the data. Talking about outliers detection, the quantile contour plots is a method used to identify data points that deviate significantly from the overall pattern of the data. This approach involves visualizing the data using quantile contour plots, which depict the contours of specified quantiles of the data distribution.

In [8], the authors introduced equivariant quantile contour plots for studying the geometry of multivariate data clouds and detecting outliers. This geometry of multivariate data clouds refers to the spatial arrangement of data points in a high-dimensional space. Quantile contours offer a flexible approach to outlier detection because they can accommodate different types of distributions and provide a visual representation of the data's spread and central tendency.

Detecting outliers based on multivariate quantile contours is another technique used in statistics and machine learning to identify observations that deviate significantly from the bulk of the data in multiple dimensions. Quantile contours provide a robust way to characterize the distribution of multivariate data, allowing for the identification of extreme observations. Multivariate quantile contours are helpful and have been studied by many researchers; for example, [35] introduced quantile functions for multivariate analysis, [15] used nested sequence of sets to define multivariate quantiles. Furthermore, different techniques of multivariate quantiles were suggested by: [5, 24, 30, 32]. However, quantile contour plots are potent tools for detecting outliers in multivariate data sets; there is no easy solution without solid assumptions on selected distributions. In this thesis, we will use the geometric quantiles introduced by [11] to detect outliers. The authors presented geometric quantiles based on the euclidean distance. It attracted much research because of its generalization of univariate quantiles and the uniqueness and existence of the quantile when a

random variable of interest is not concentrated on a single straight line.

The notion of geometric quantiles for outlier detection was introduced by [11], and he established some valuable properties. In his study, he stated that when the distribution of a random variable is not concentrated on a single straight line in R^d, the geometric quantiles with index vector u in unit open ball happen to be rotational equivariant. Another vital property stated by [11] said that geometric quantiles are equivariant under any homogeneous scale transformation of the coordinates of the multivariate observations. Finally, another property stated by [24] says that if two random variables yield the same quantile function, they have the same distribution function. That makes geometric quantiles more attractive in many applications, including multivariate analysis.

In economics and other financial sciences, conditional quantiles (or quantile regression) play an essential role in analyzing the effect of a set of covariates on the outcome of conditional distribution (see [23]). Conditional quantiles have attracted many researchers like [6] and [13] studied portfolio returns-based attribution using quantile regression. Also, [29] used conditional quantile to analyze the cross-country assessment of systemic risk in the European stock market. In environmental modelling, [1] used conditional quantiles to analyze factors contributing to extremely low infant birth weights, and [28] used conditional quantiles for regional flood frequency analysis. In the multivariate sense, there must be considered both the distance of an observation from the centroid of the data and the shape of the data. The Mahalanobis distance [26] is a well-known measure which takes it into account.

2.1 Outlier detection based on censored quantile regression

Because of its flexibility in modelling the effect of covariates, censored quantile regression has attracted considerable attention in survival analysis. For example, [25] tackled the censored quantile regression problem that facilitates inference and asymptotic studies. They used censored quantile regression for survival data subject to conditionally independent censoring. When dependent variables are censored in population studies, the censored quantile regression model introduced by [31] can consistently estimate conditional quantiles. In his model, [31] noticed that if we observe censoring values C_i,for all $i = 1, \cdots, n$ and for the given linear latent variable model, $L_i = X_i^T \beta(\tau) + u_i$ where u_i (sometimes called a random error) is an independent and identically distributed with distribution function F and $\beta(\tau)$ for some $\tau \in (0, 1)$ is a d-dimensional quantile coefficient vector. If we also observe $Y_i = \min\{C_i, L_i\}$ then the conditional quantile functions given by : $Q_{Y_i|X_i}(\tau|X_i) = \inf\{x : F(x|X_i) \geq \tau\} = X^T \beta_i(\tau)$ can be consistently estimated by

$$\hat{\beta} = \underset{t \in \mathrm{R}^d}{\operatorname{argmin}} \sum_{i=1}^{n} \rho_\tau(Y_i - \min\{C_i, X_i^T t\}),$$

where $\rho_\tau(u) = u(\tau - 1_{\{u<0\}})$ is the check function.

These censored quantile regressions have advantages compared to a standard censored regression problem. They are an alternative to conditional means and notice similarities with conditional quantiles. As seen by [31], they allow consistent estimation in the censored regression model.

Quantile regression plays a vital role in numerous scientific disciplines, and it comes up with a natural way to capture the effects of covariates at different response distribution tails. However, due sparsity of data estimation from quantile regression, analysis from quantile regression for heavy-tailed distributions is often unstable at the tails without any assumption on distributional function. The popular and easy way of understating the tail of the distribution is considering the extreme quantiles.

Recently, [16] proposed three outlier detection algorithms based on censored quantile regression: residual-based, scoring, and boxplot algorithms. The residual-based outlier detection algorithm is based on fitting a censored quantile regression model, calculating the residuals and computing covariates.

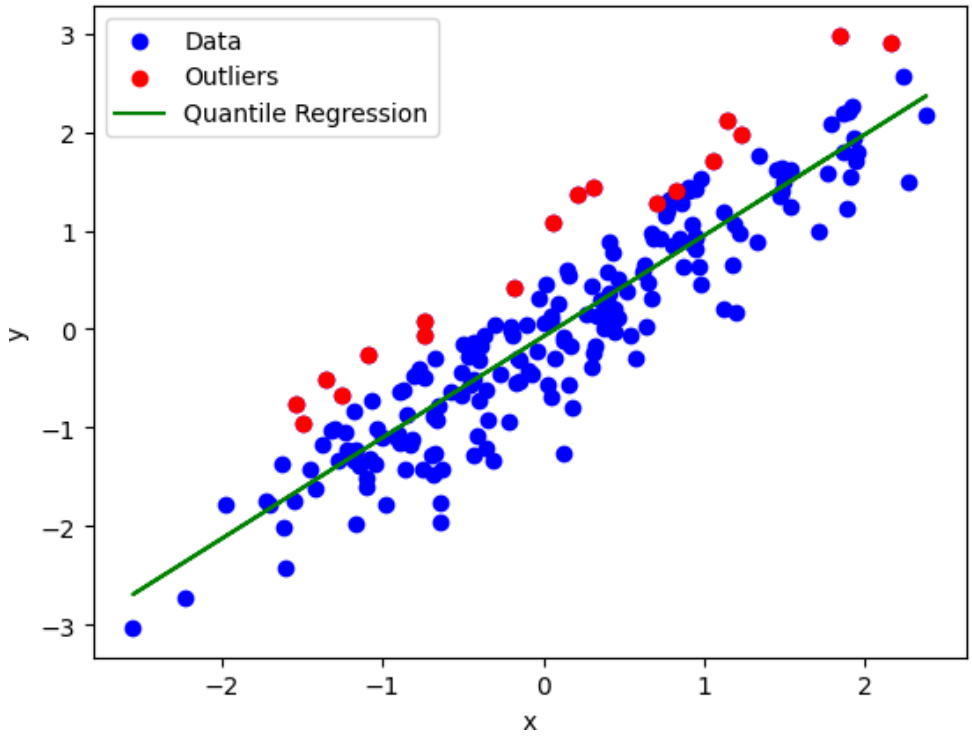

Figure 2.1: Outlier detection based on Censored quantile regression model.

In Figure 2.1, we first generate 200 observations from Normal distribution with two covariates (X) and a response variable (Y). Then, we fit a censored quantile regression model using the sm.QuantReg class. The sm.QuantReg refers to the quantile regression class provided by the

Statsmodels library in Python. It allows us to compare quantile regression results with those of ordinary least squares (OLS) regression to assess whether the relationships between variables differ across quantiles. It's a statistical technique that extends traditional linear regression by estimating the conditional quantiles of the dependent variable. In other words, instead of predicting the mean (expected value) of the response variable, quantile regression allows you to model various quantiles, such as the median, lower quantiles, or upper quantiles. We set the desired quantile for outlier detection (in this case, the 90^{th} percentile) and fit the model using the fit method. Af- ter fitting the model, we calculate the predicted quantile values using the prediction method. Wethen calculate the residuals by subtracting the predicted quantiles from the observed response val- ues. Additionally, we calculate the censored values by taking the maximum between the observed response values and the predicted quantiles. Finally, we identify outliers based on the residuals (using a simple rule of being more than two standard deviations away from the mean) or based on the censored values. The indices of the identified outliers are printed as the output in red (Figure 2.1). The observation is an outlier if its residual is larger than the estimated variance-covariance value. A boxplot outlier detection algorithm is based on a simple outlier detection approach based on a boxplot introduced by [36] and has been used widely for uncensored data. Finally, the scoring outlier detection algorithm is based on fitting a censored quantile regression model and calculating the outlying score. This algorithm provides the outlying degree that shows the magnitude of deviation from the probability distribution given the covariates.

2.2 Extreme quantiles

Extreme quantiles, also known as tail quantiles or high quantiles, refer to the quantiles of a probability distribution that correspond to the tails of the distribution. They play a significant role in forecasting rare events. We may have to infer beyond the observations to estimate high quantiles of an unknown probability distribution. It is essential to predict extremely high quantiles in numerous disciplines accurately. The author in [20] studied the estimation of conditional quantiles nonparametrically with emphasis on the range beyond the data. They proposed a flexible nonparametric two-stage procedure for estimating extreme quantiles in a regression setting. There is much interest in many financial modelling areas in using extreme quantiles. Recent contributions in the literature in the case of the sparsity of covariates, like [13] studied conditional extremes and near extremes, and [12] brought nonparametric extreme regression quantiles. Also, [17] examined the extreme geometric quantiles in a multivariate regular variation framework, and they provided an equivalent of magnitude and the direction of an extreme geometric quantile. The author presented some Intriguing properties of extreme geometric quantiles. One consequence of their results is that the magnitude of extreme geometric quantiles of a random vector with a finite covariance matrix

grows at a fixed rate. Furthermore, they found that with underlying distribution possessing a finite covariance matrix, the extreme geometric quantile could be accurately estimated using a standard empirical estimator of the covariance matrix regardless of its extremity.

2.3 Multivariate quantiles and outliers detection

Detecting data outliers is one of the main tasks in statistical analysis. The search for outliers in multivariate data is usually based on the location and spread of the data, not only the distance of an observation from the centroid of the data but also the shape of the data.

Univariate quantiles are very popular due to their applications and simplicity in descriptive statistics. However, extending the notion of quantiles to find suitable quantile analogs for multivariate data poses a big problem. Many researchers presented several approaches. In 1985 Eddy [15] studied different methods for ordering multivariate data; [33] stated that "multivariate quan- tiles are an attractive alternative to quantiles based on an estimate of the multivariate density asthey are simple to compute and do not suffer from the well-known 'curse of dimensionality' prob-lem inherent in most nonparametric density estimation procedures". Multivariate M-quantiles have applications in outlier detection, as they are probability-based ordering techniques for multidimensional data. Also, [8] developed introduced equivariant quantile contour plots for studying the geometry of multivariate data clouds and detecting outliers.

Multivariate quantile contours are useful and have been studied by many researchers; for example, [35] introduced quantile functions for multivariate analysis, [15] used nested sequence of sets to define multivariate quantiles. Different techniques of multivariate quantiles were suggested by [5, 24, 30, 32]. However, quantile contour plots are powerful tools for detecting outliers in multivariate data sets; there is no easy solution without strong assumptions on selected distributions. This thesis focuses on the extreme conditional case of geometric quantile introduced by [11]. Thiswork [11]) also introduced geometric quantile based on the euclidean distance; it attracted many researchers because of its generalization of univariate quantiles and its uniqueness and existence of the quantile when a random variable of interest is not concentrated on a single straight line. For more illustration, let us simulate an example of a bivariate data set of 200 observations following a bivariate standard normal distribution.

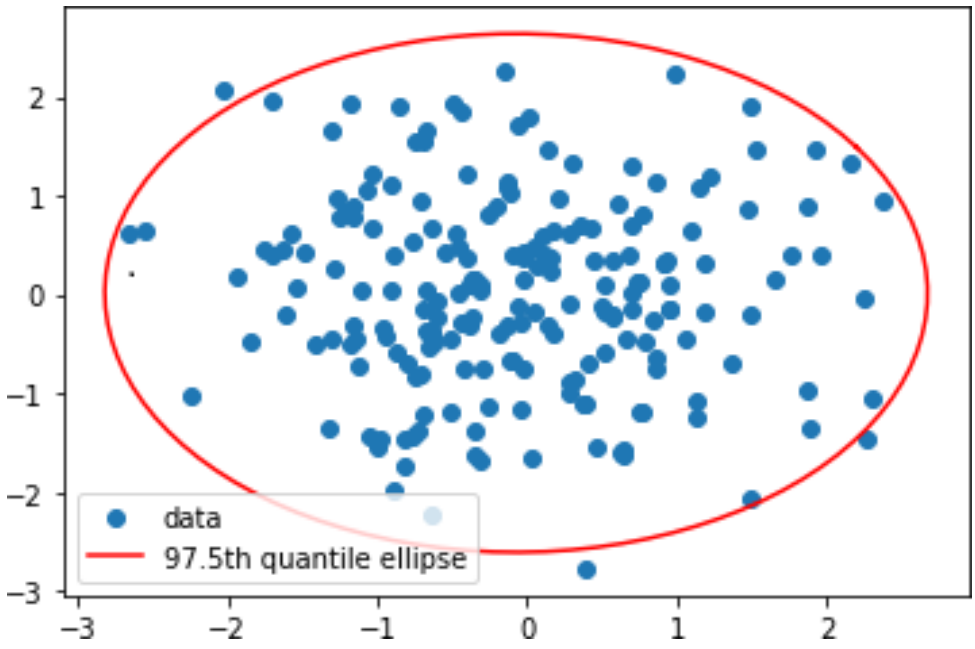

Figure 2.2: Outliers (outside of ellipse) identified by $\chi^2_{2;0.95}$.

From the Figure 2.2, the points outside of the ellipse are considered an outlier. The figure presents the observations (points in blue) and 0.975^{th} quantile ellipse generated with the help of chi-square with two degrees of freedom (solid red line). The line $\chi^2_{2;0.95}$ identifies only three outliers in 200 observations outside the ellipse.

2.4 Outlier detection based on Mahalanobis distance

To detect outliers in the multivariate sense, there must be consideration of both the shape of the data and the distance of an observation from the centroid of the data. The well-known measure, which considers it, is Mahalanobis distance [26]. It was stated by [19] that the distribution of the squared Mahalanobis distance is known to be chi-squared with k degrees of freedom. Then, selecting the threshold as the 0.975 quantiles of the χ^2_k embraced as a method for identifying the outliers. We apply outlier detection based on the Mahalanobis distance method to the Airquality dataset. The data is about the New York Air Quality Measurements of 1973 for five months from May to September recorded daily. It contains 153 observations of 6 variables. For more details about dataset, see https://r-data.pmagunia.com/iframe/airquality.html. The summary of the data is found in Table 2.1, where:

- Ozone: Mean ozone in parts per billion from 1300 to 1500 hours at Roosevelt Island.

- Solar.R: At Central Park, solar radiation in Langleys in the frequency band 4000–7700 Angstroms from 0800 to 1200 hours.
- Wind: Average wind speed in miles per hour at 700 and 1000 hours at LaGuardia Airport.
- Temp: Maximum daily temperature in degrees Fahrenheit at La Guardia Airport.

	Ozone	**Solar.R**	**Wind**	**Temp**	**Month**	**Day**
Minimum	1	7	1.7	56	5	1
1^{st} Quartile	18	115.8	7.4	72	6	8
Median	31.5	205	9.7	79	7	16
Mean	42.13	185.9	9.958	77.88	NA	NA
3^{rd} Quartile	63.25	258.8	11.5	85	8	23
maximum	168	334	20.7	97	9	31

Table 2.1: Air quality data summary statistics showing minimum value, first quartile, median, mean, third quartile and maximum value.

Applying the Mahalanobis distance method for outlier detection is a valid approach to identifying unusual data points in the airquality dataset or any other dataset. The Mahalanobis distance quantifies the dissimilarity between an observation and a dataset by considering the covariance between variables. Its utility extends to outlier detection in multivariate datasets. To apply the Mahalanobis distance method to the airquality dataset, we will need to follow these general steps: Handle miss- ing values, scale the variables if needed, and ensure the dataset is ready for analysis. 2. Calculate the Mahalanobis distance of each data point from the mean or centroid of the dataset using the covariance matrix. Decide on a threshold value or a critical region that defines the boundary beyond which data points are considered outliers. This threshold could be based on statistical methods like the chi-square distribution or other domain-specific knowledge. 3. Compare the Mahalanobis distances of each data point to the threshold. Data points with Mahalanobis distances greater than the threshold are considered outliers.

Cutoff	Ozone	Temp
30	115	79
62	135	84
117	168	81

Table 2.2: Identification of outliers with Mahalanobis distance method.
Ozone in parts per billion and temperature in degrees Fahrenheit. The term "cutoff" typically refers to a threshold or boundary value used to determine whether a particular depth measurement is significant or not.

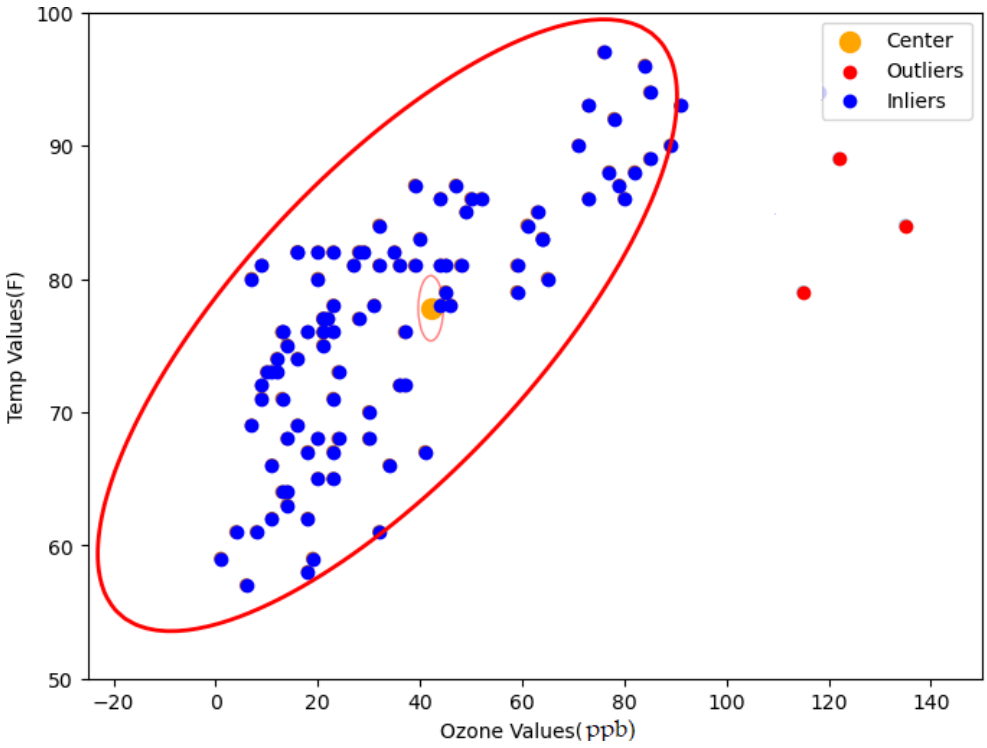

Figure 2.3: Scatter plot of ozone and temperature variables.
Ozone in parts per billion and temperature in degrees Fahrenheit.

Considering two variables: ozone and temperature, the outlier detection based on Mahalanobis distance is shown in the Figure 2.3. From the figure, the yellow point in the ellipse shows the center point. The Ozone — Temp variables observations are blue points in the ellipse, while red points outside the ellipse are considered outliers. The Mahalanobis Distance method calculates the distance between each point and center. Then we find distances and use the Chi-Square value as a cut-off to find outliers. In statistical depth functions, the cutoff refers to a predetermined value used to determine whether a point is considered "deep" or "outlying" within a dataset.

Our identification of the outliers in the multivariate data found 30, 62 and 117 observations (first column in Table 2.2) as the same as the points outside of the ellipse in the Figure 2.3. Considering the other two variables: solar radiation and wind, the outlier detection based on Mahalanobis distance identifies 9, 48 and 53 observations in the first column of Table 2.3.

Cutoff	Solar.R	Wind
9	19	20.1
48	284	20.7
53	59	1.7

Table 2.3: Identification of outliers with Mahalanobis distance method of solar radiation and wind. Solar.R in the frequency band and wind speed in miles per hour.

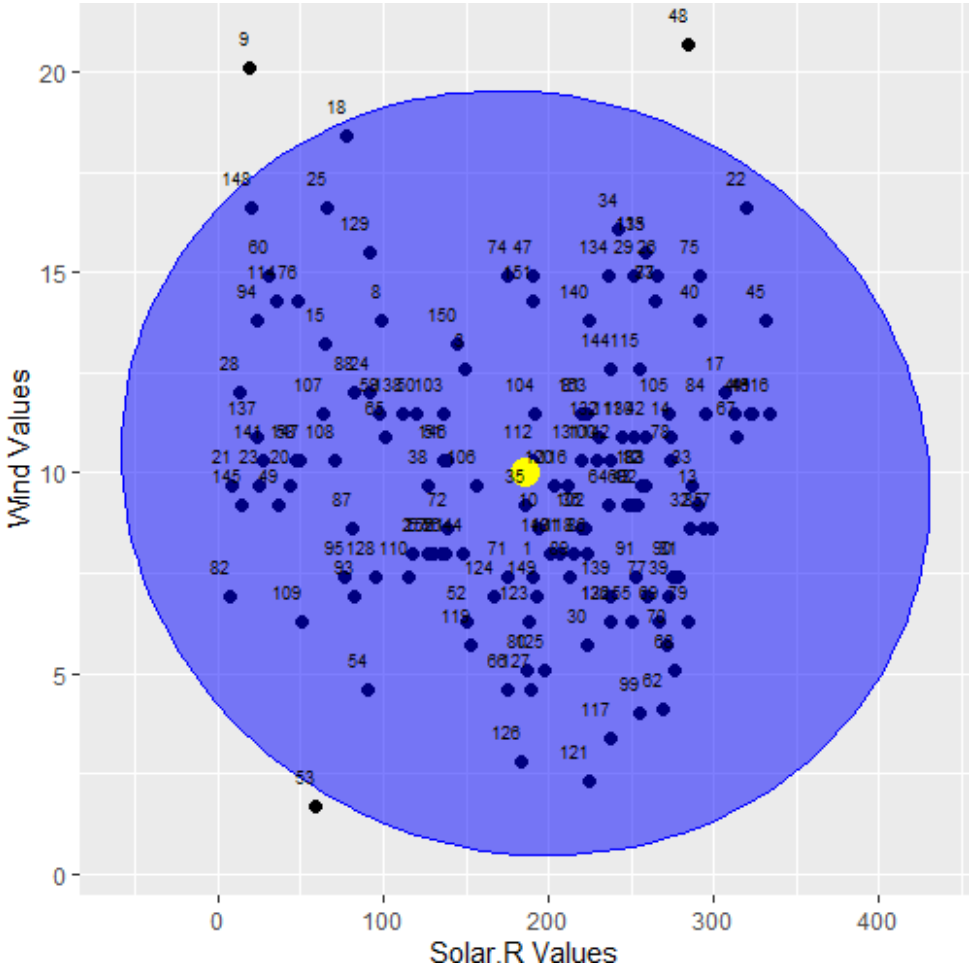

Figure 2.4: Scatter plot of solar radiation and wind variables.
Solar radiation in the frequency bands and wind speed in miles per hour.

From the Figure 2.4, the outliers points are 9,48 and 53, the same points found in the first column in the Table 2.3. In measuring the distance between a point and a distribution, Mahalanobis distance is considered an extremely useful multivariate distance metric with excellent applications in multivariate outlier detection and classification on highly imbalanced datasets.

Chapter 3

Methodology

In a multivariate set-up, geometric quantiles are an extension of the concept of univariate quantiles. Geometric quantiles are a type of quantiles that partition a dataset into intervals such that each interval contains approximately the same number of data points when measured on a logarithmic scale. In a multivariate data set, geometric quantiles play an essential role in detecting outliers through quantile contour plots.

Geometric quantiles find applications in various fields, including finance, environmental science, and engineering, where analyzing data with exponential behavior is common. Many researchers have used quantiles to analyze multivariate datasets like [21] applied multivariate quantiles for performing the financial time series analysis. In a multidimensional setting, a significant challenge is the lack of a natural basis for ordering multivariate observations.

Recently, Chakraborty [8] developed useful equivariant quantile contour plots to study the geometry of multivariate data clouds and detect outliers. By plotting quantile contours, he suggested that any point outside the plot should be suspected of being an outlier. Those quantiles are also helpful for constructing multivariate Q-Q plots to check how well the proposed multivariate dis- tribution fits the data. In a multivariate data cloud, it is reasonable to study the low points and thehigh points because the observations may have low values in some directions and high values in other directions. Those intrinsic geometric features of multivariate observations can be capturedby multivariate quantile analysis such as geometric quantile contour plots.

In his study, [11] stated that when the distribution of a random variable is not concentrated on a single straight line in R^d then the geometric quantiles with index vector u in a unit open balls are rotationally equivariant. Another substantial property stated by [11] said that geometric quantiles are equivariant for any homogeneous scale transformation of the coordinates under multivariate observations. Also, [24] stated that if two random variables yield the same quantile function, then they have the same distribution function. These mentioned properties make geometric quantiles more attractive in many applications, including multivariate analysis.

The quantile analysis method has attracted many researchers because of its straightforward interpretation. In [23], the authors constructed the quantile regression analysis as a generalization of median regression. Quantile regression illustrates the conditional quantile functions of response variables in terms of a set of covariates. For example, we can use this method to predict the expected i^{th} percentile of the exam scores versus the number of hours studied. Figure 3.1 shows the result from this example by considering 200 students at some universities. From the figure, we can see the estimated regression equation is

$$90^{th}\text{percentile of exam score} = 63.2752 + 2.5050 * (\text{number of hours})$$

```
================================================================================
Dep. Variable:                  score   Pseudo R-squared:   0.5308
Model:                       QuantReg   Bandwidth:          2.306
Method:                 Least Squares   Sparsity:           10.82
                                        No. Observations:   200
                                        Df Residuals:       198
                                        Df Model:           1
================================================================================
            coef      std err      t      P>|t|
--------------------------------------------------------------------------------
Intercept  63.2752     0.455    139.060     0.000
hours      2.5050      0.071    35.324      0.000


================================================================================
```

Figure 3.1: Prediction of 90^{th} percentile of exam scores based on the number of hours studied.

For example, the 90^{th} percentile score for all students who study 10 hours is expected to be 88.3252. Quantile regression analysis is also a standard tool for modeling the relationship between a response and covariates variables. In health services and health economics, many researchers commonly use conditional quantile regression in the quantile regression framework. For example, [27] used conditional quantile regression to model the impact of price demand for alcohol, and [4] used conditional quantile regression to assess gender difference in thrombolytic therapy timeliness. Conditional quantile regression estimates the coefficients as a conditional quantile function; this means that the estimated coefficients in quantile regression quantify the expected change in the distribution of the response variable as the covariate variable increases.

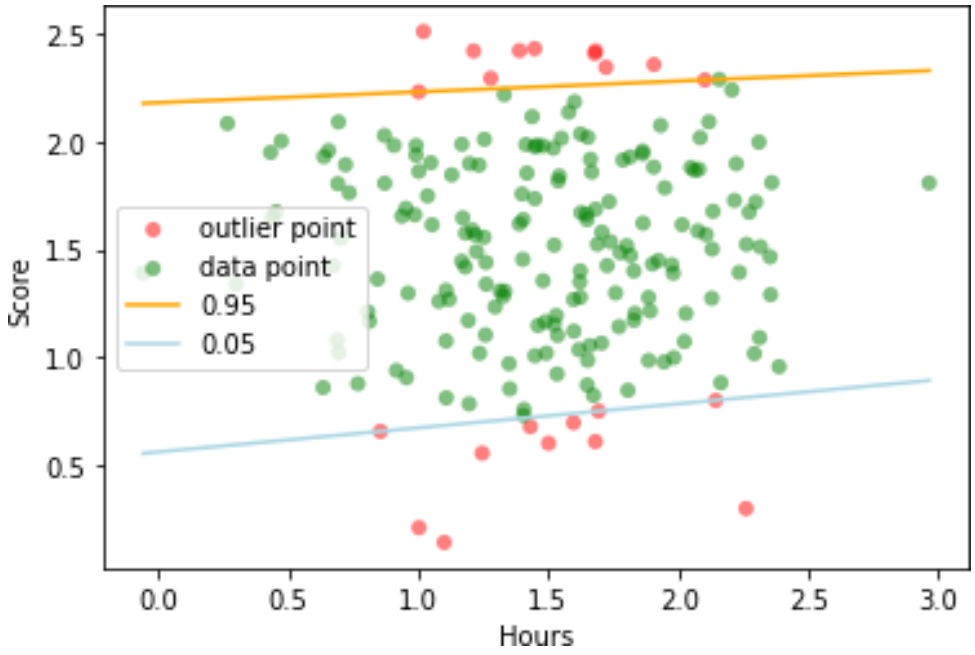

Figure 3.2: Outlier labelling of student scores based on quantile regression method. The hours are in 10^{th} while scores are multiples of 40.

According to the figure 3.2, for better forecasting of student performance, it is wise to treat the red points (outlier points) in two ways either removing them before analysis or managing their impact in data analysis. Those scores are not proportional to the number of hours used by the students; hence they will harm the prediction.

3.1 Definition of geometric quantiles (Chaudhuri 1996)

Let us consider a random vector X in R^d with underlying distribution function F, if $u \in \mathsf{R}^d$ is an arbitrary vector and if a geometric u–th quantile of X exists, then, it is a solution to the following optimization problem:

$$Q_F(u) = \underset{q \in \mathsf{R}^d}{\operatorname{argmin}} E[\|X - q\| - \|X\|] - \langle u, q \rangle, \tag{3.1.1}$$

where $\|\cdot\|$ and $\langle\cdot\rangle$ are Euclidean norm and Euclidean scalar product, respectively. It is notewor- thy that they possess both magnitude and direction [11]. He pointed out that when the distributionof X is not concentrated on a single straight line, the quantile $Q_F(u)$ indexed by vector u belongsto the unit open ball B^d, which is unique. In this definition, the vector u corresponds to a direction, which explains the geometric aspect. Those geometric quantiles have many valuable properties:

1. Under any orthogonal transformation, geometric quantiles are equivariant [11] .
2. For every vector $u \in \mathsf{R}^d$ in the unit open ball B^d,and whenever the distribution of random variable X is not concentrated on a single straight line in R^d , there exists a unique geometric quantile [11].
3. With index vector u, the norm of the geometric quantile $Q_F(u)$ diverges to infinity as $\|u\| \uparrow 1$ [17].
4. The geometric quantiles function characterizes the associated distribution. This property means that if two random variables yield the same quantile function, then they have the same distribution [24].

Identifying extreme values is very crucial in insurance, economics and other financial statistical studies. In the univariate case, this action can be done with the help of a boxplot. A quantile contour plot can be an alternative to a boxplot for outliers detection in a multivariate framework. We can refer to the geometric quantile contour plot suggested by [8] to detect outliers. Any point outside regions enclosed by a quantile contour plot would be suspected to be an outlier. Quantile contour plots are potent tools for detecting outliers in multivariate data sets [8] .

3.2 Statistical analysis of geometric quantile

In multivariate data sets, geometric quantiles play a vital role in detecting outliers employing quan- tile contour plots. Many statistical literatures have proposed the notion of geometric or spatial quantiles. We can refer to [11], who studied a geometric notion of quantiles for multivariate data. In his study, he introduced geometric quantiles based on optimization algorithms. His definition attracted many researchers. In this thesis, we refer to his definition and build a geometric quantile contour plot to detect the outliers. In recent literature, [9] also used geometric quantiles to detect outliers. We can also refer to some researchers who performed the geometric quantile based on estimation like [3], who estimated conditional geometric median and [37], who estimated conditional geometric median in infinite-dimensional covariate.

In this thesis, we focus on the geometric quantile introduced by [11] and use geometric quantile contour plots to detect outliers. Let us recall the equation 3.1.1 from the above definition and redefining it as a continuous function in R as follows:

$$\psi(u, q) = E[\|X - q\| - \|X\|] - \langle u, q \rangle,$$

where $\psi : \mathsf{R}^d \times \mathsf{R}^d \to \mathsf{R}$ is a continuous and derivative function in R. For simplicity, assume that $\frac{x}{\|x\|} = 0$ if $x = 0$. If $q(u)$ is a solution of equation 3.1.1 with u an arbitrary vector in R^d, then

$q(u)$ is a root of $\frac{\partial \psi(u,q)}{\partial q}$. That is $q(u)$ is zero of

$$u^{[} \quad E \quad \frac{X - q(u)}{\|X - q(u)\|} \quad = 0 \tag{3.2.1}$$

The vector u plays a significant role in estimating the quantile $Q_F(u)$. Calculating the norm of u helps find the order of spatial quantiles. If $\|u\|$ is close to one, then $Q_F(u)$ is an extreme quantile, and when $\|u\|$ is close to zero, then $Q_F(u)$ is central quantile [11]. Let us consider an exampleof a contrived bivariate data set of size $n = 20$, $\{X_1, X_2, \cdots, X_n\} \in \mathbb{R}^2$. Contrived bivariate datarefers to a set of paired observations deliberately created for the purpose of illustrating a particular concept. The contrived nature of the dataset implies that it is intentionally created rather than naturally occurring.

$X_i = Q_F(u_i)$	u_i	$\|u_i\|$
(0,1)	(0.0195, 0.0487)	0.0524
(0,-1)	(0.0195, -0.0487)	0.0524
(1,0)	(0.1782, -0.0882)	0.1989
(0,1.5)	(0.0295, 0.1785)	0.1809
(0,3)	(0.0401, 0.3320)	0.3344
(-3,0)	(-0.2628, -0.0521)	0.2679
(1.5,0)	(0.3115, -0.0784)	0.3212
(0,5)	(0.0369, 0.4699)	0.4713
(-5,0)	(-0.3934, -0.0293)	0.3945
(-10,0)	(-0.5236, -0.0057)	0.5236
(10,0)	(0.5236, -0.0057)	0.5236
(0,-12)	(0.0210, -0.5209)	0.5213
(0,15)	(0.017316, 0.5793)	0.5796
(-15,0)	(-0.5850, 0.00064)	0.5850
(-16,0)	(-0.5865, 0.0012)	0.5865
(0,-17)	(0.0155, -0.5831)	0.5833
(0,18)	(0.0147, 0.5847)	0.5848
(-20,0)	(-0.5906, 0.0024)	0.5906
(0,20)	(0.0133, 0.5872)	0.5873
(0,-20)	(0.0133, -0.5872)	0.5873

Table 3.1: Spatial quantile estimation for contrived bivariate data set.

In the table 3.1, the middle half of the data should consist of ten points out of twenty. The observations X_3, X_4, X_5, X_6, X_7, X_8 and X_9 are in the region $\{Q(u) : \|u\| \leq 0.5\}$, which does not represent half of the observations, and according to [11], there is no extreme quantile in our example since there is no $\{Q(u_i) : \|u_i\| \uparrow 1\}$.

The important message from the above table and equation 3.2.1 is that we can give a geometric quantile interpretation to each observation X_i a spatial quantile $Q_F(ui)$ with index vector u_i indicates that X_i has F as a distribution function. Also, we can measure the outlyingness of any point x_i by the corresponding magnitude $\|u_i\|$ quantitatively.

3.3 Simulation and outliers detection based on quantiles

Outlier detection is a common technique used in data analysis to identify observations that deviate significantly from most data. It is handy for detecting datasets' anomalies, errors, or unusual patterns. In the case of simulated observations, outlier detection can help identify any unusual or unexpected patterns that may have occurred during the simulation process. In this thesis, we focus on outliers detection based on quantiles methods defined in chapter one, and we compare the usefulness of those methods. Outlier detection based on quantiles is a common method for identifying outliers in a dataset. The idea is to define thresholds based on the quantiles of the data distribution and consider observations that fall outside these thresholds as potential outliers.

3.3.1 Outliers detection based on univariate quantiles

Univariate quantiles can be used for outlier detection in a dataset. The general idea is to identify data points that fall significantly outside the expected range defined by the quantiles. Outliers significantly deviate from the rest of the data, and quantiles provide a way to divide the data into equal-sized portions. In chapter one, we defined univariate quantiles as follows: $Q_X(\tau) = \inf\{x \in \mathbb{R} : F(x) \geq \tau\}$. This expression emphasizes that given the value of τ, we need to find some x, which results in $F(x)$ returning a value not less than τ. Here is a general approach for outlier detection based on univariate quantiles: We choose a quantile level by determining the desired quantile level that will be used to identify outliers. The most commonly used quantiles are the lower quartile ($Q1$), the median ($Q2$), and the upper quartile ($Q3$). These correspond to the 25^{th}, 50^{th}, and 75^{th} percentiles, respectively. Then we calculate the quantiles by sorting the data in ascending order and calculating the chosen quantiles. Then we determine the interquartile range (IQR): The interquartile range is calculated as the disparity between the upper quartile ($Q3$)and the lower quartile ($Q1$), i.e., $IQR = Q3 - Q1$. After we define outlier thresholds by multiplying the IQR by a factor (usually 1.5 or 3) to determine the lower and upper thresholds for outliers, these thresholds

are used to identify values that fall below $Q1 - (IQR * factor)$ or above $Q3 + (IQR * factor)$. Lastly, we Identify outliers by Comparing each data point with the defined thresholds. Any valuebelow the lower or above the upper threshold is considered an outlier. Outliers detection based on univariate quantiles uses the boxplots technique in exploratory data analysis. Boxplots depict the data distribution within a dataset, including its median, quartiles, and potential outliers.

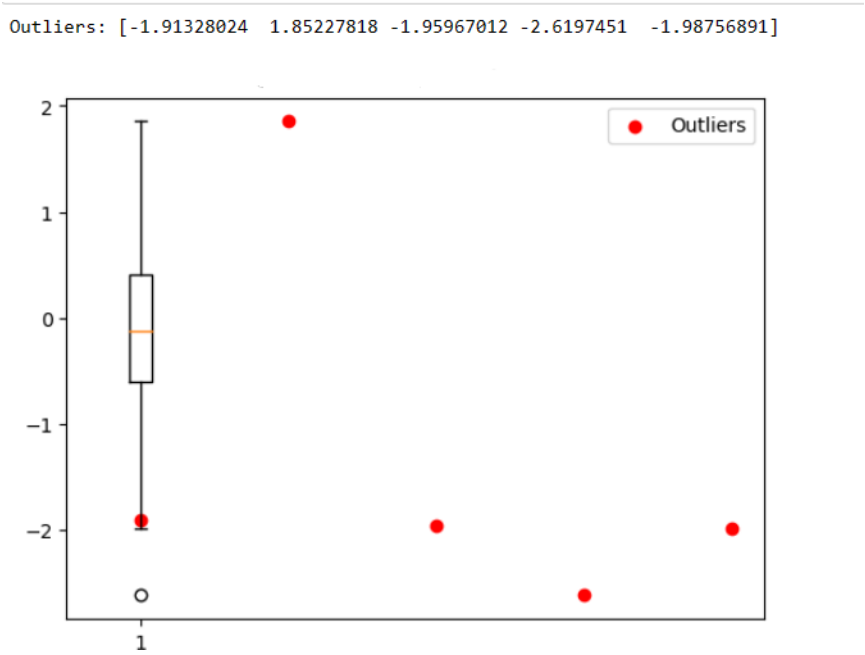

Figure 3.3: Outlier detection based on univariate quantiles using boxplots.

Figure 3.5 is the result of simulated 100 observations with the help of standard normal distribution; the box represents the middle 50% of the data, with the median indicated by a horizontal line inside the box. The boxplot identified five points that fall outside the box. These points are considered potential outliers. Typically, values beyond 1.5 times the IQR (i.e., $Q3 + 1.5 * IQR$ or$Q1 - 1.5 * IQR$) are considered outliers. It is worth noting that while boxplots are a helpful tool for identifying outliers, they may not capture all types of outliers, especially those present in skewed distributions or those occurring in multivariate datasets. It is important to note that this approach assumes the data follows a univariate distribution and that outliers can be identified based on a single variable. For multivariate data, we use geometric quantile analysis through quantile contour plots for multivariate outlier detection or anomaly detection.

3.3.2 Outliers detection based on conditional quantiles

Outliers detection based on conditional quantiles is a statistical approach used to identify extreme observations that deviate significantly from the expected values within specific conditional con-

texts. This method focuses on assessing the distribution of data points within different subsets or conditions of the dataset rather than analyzing the overall distribution. We define the conditional context by Identifying a variable or set of variables that determine the conditions under which outliers will be assessed. Then we Split the dataset into subsets based on the defined conditional context. Each subset represents a specific condition or context under which outliers will be evaluated. For each subset, we calculate the conditional quantiles. Quantiles provide information about the distribution of the data and help determine the threshold beyond which observations are considered outliers. Then we identify outliers by comparing individual data points within each subset to the corresponding conditional quantiles. If a data point falls below the lower quantile or above the upper quantile, it is considered an outlier within that specific condition. The implementation of outlier detection based on conditional quantiles can vary depending on the statistical techniques used and the dataset's characteristics. Additionally, determining appropriate thresholds for outliers might require experimentation or domain knowledge. This approach allows for detecting outliers within specific conditions, which can be helpful when analyzing datasets with varying contexts or when outliers may differ across subsets.

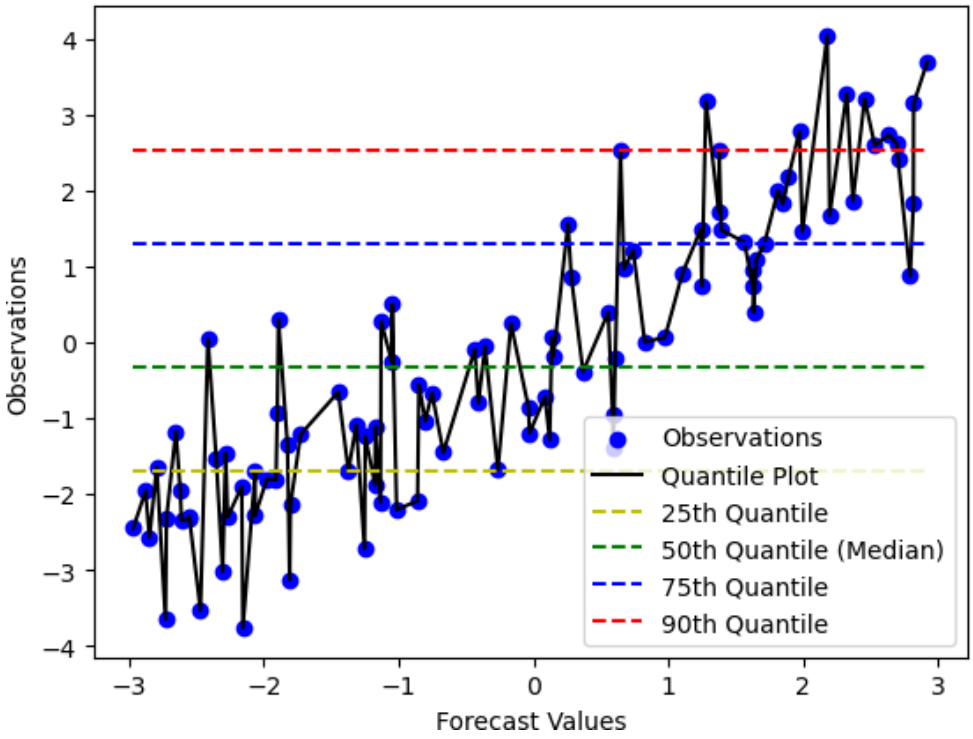

Figure 3.4: Outlier detection based on conditional quantiles.

We generated a sample of 100 observations from the standard normal distribution to create a conditional quantile plot from a standard normal distribution. We then sort the data based on the

forecast values to ensure the correct ordering for the quantile plot. Next, we calculate the desired quantiles of the sorted observations. The resulting plot (figure 3.4) shows how the quantiles of the conditional distributions change as the forecast values vary. By setting the 90^{th} quantile as a threshold, we found ten outliers representing ten percent of the observations.

3.3.3 Outliers detection based on geometric quantile contour plot

A quantile contour plot is a type of graphical representation that visualizes data distribution in a two-dimensional space. It is commonly used in statistical analysis and data visualization to show the relationship between two variables and identify data patterns or trends. The data is first divided into a grid of cells to create a quantile contour plot, each representing a range of values for the two variables being analyzed. The number of cells can vary depending on the size of the dataset and the desired level of detail in the plot. Next, the quantiles of the data are calculated for each cell. Quantiles are a way of dividing a dataset into equal parts, such as quartiles (dividing the data into four equal parts) or deciles (dividing the data into ten equal parts). The quantiles are then used to create contour lines on the plot. Each contour line represents a different quantile, with all lines closer to the center of the plot representing higher quantiles and the lines further away representing lower quantiles. Finally, the data points are plotted on top of the quantile contour plot. The points are colour-coded or shaded based on which quantile they fall into, and this allows the viewer to see where most of the data falls and identify any outliers or unusual patterns. Quantile contour plots are particularly useful for visualizing non-linear data or complex relationships between variables. They are also helpful in the comparison of multiple datasets or in analyzing changes in data over time. Quantile contour plots provide a powerful tool for data visualization and analysis, allowing researchers and analysts to identify patterns, trends, and outliers in their data with greater clarity and precision. In this part, we investigate the behaviour and location of outliers through geometric quantile contour plots using simulated data. We use two-dimensional cases (i.e., supposing that d = 2) to interpret and realize these contour plots easily. In a statistical study, the identification of outliers in sample data is a considerable and helpful action.

For an illustration of the geometric quantile contour plots, we simulate 100 observations according to standard normal distribution (i.e. the multinormal distribution $N_2(0, I_2)$ (Figure 3.5) and $N(0, \Sigma)$ normal distribution with $\Sigma = \begin{pmatrix} 2 & 0.9 \\ 0.9 & 1 \end{pmatrix}$ (Figure 3.6).

In both cases, geometric quantile contour plots help you visualize the spread and shape of the data distribution. They are useful for identifying patterns and understanding the relationships between variables, especially when dealing with multivariate data. The contour lines represent quantiles of the data, allowing you to see how data points are distributed in different plot regions.

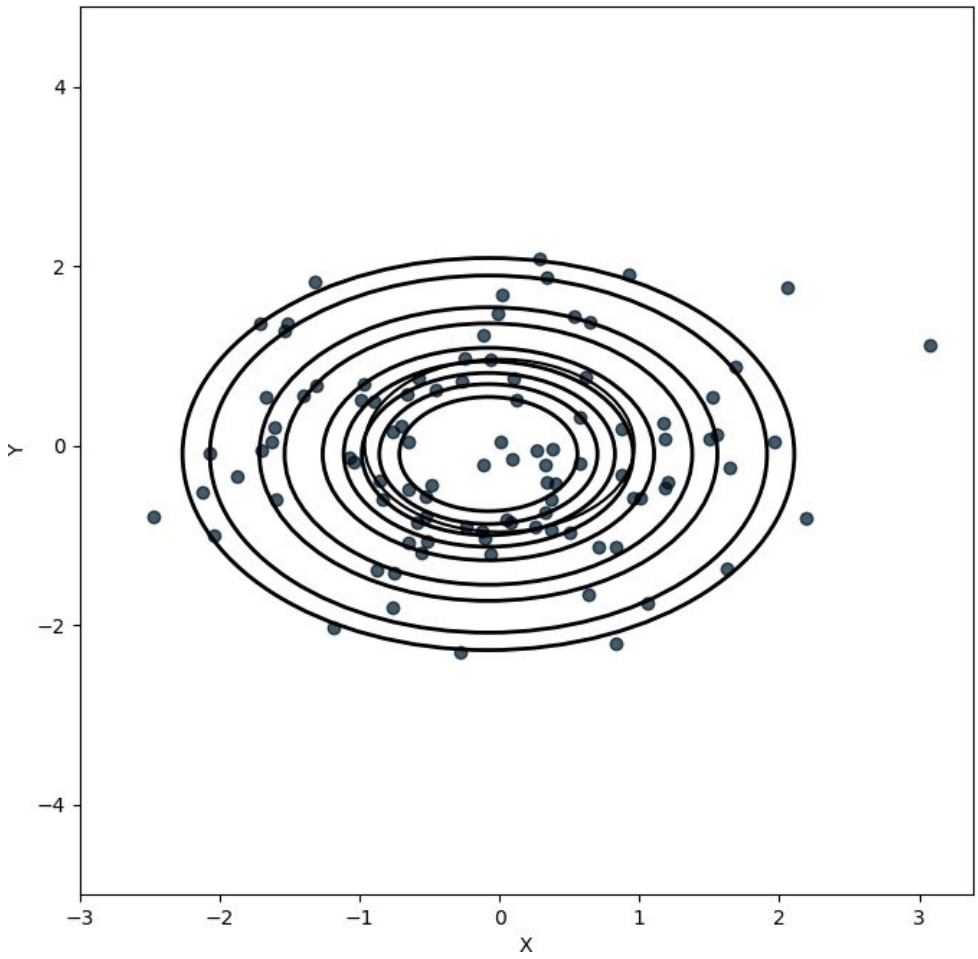

Figure 3.5: Quantile contour plot of 100 observation with standard Normal distribution. Quantile contour plot from 10% to 90% for observations given by a bivariate standard normal distribution.

From the Figure 3.5, ten observations are located outside of the contour line corresponding with $Q(u) = 0.90$. Those observations are considered as outliers. We can say that with $Q(u) = 0.90$, ten percent of observations are labelled as outliers.

In the Figure 3.6, we simulated 100 observations from a multivariate normal distribution $N(0, \Sigma)$ with $\Sigma = \begin{pmatrix} 2 & 0.9 \\ 0.9 & 1 \end{pmatrix}$.

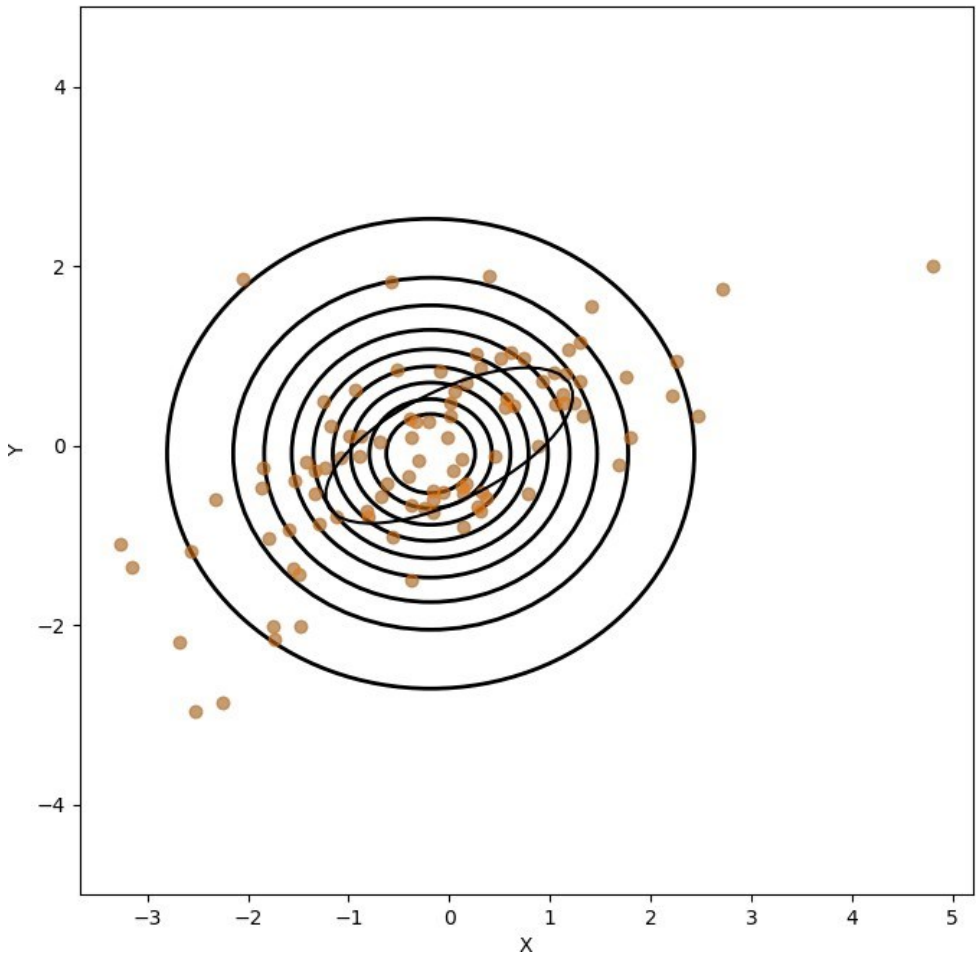

Figure 3.6: Quantile contour plot of 100 observation with Normal distribution. Quantile contour plot from 10% to 90% , from a $N(0, I_2)$ normal distribution whith $\Sigma = \begin{pmatrix} 2 & 0.9 \\ 0.9 & 1 \end{pmatrix}$.

From the Figure 3.6, only nine observations are located outside of the contour line, correspond- ing with $Q(u) = 0.90$. Those observations are labelled as outliers.

In a geometric quantile contour plot, contour lines represent constant quantiles of the data distribution. These lines connect points on the plot that have the same quantile value. Specifically, each contour line encloses areas of the plot where the data points have the same quantile value, allowing for visualization of how the data is distributed across different quantiles. The contour lines in such plots aid in understanding the shape and spread of the data distribution, especially in relation to quantiles.

In Figure 3.6, the contour lines represent quantiles in 10% intervals. This implies that approx- imately 10% of the points generated from the estimated Normal distribution are below the lowest

contour, 20% are below the next contour, and so on. The highest contour has about 90% of the points below it. The points outside of the contour line of 90% are considered as outliers. Quantile contour plots provide a visual representation of the distribution of data. By displaying contour lines that connect points of equal quantile, they reveal patterns and structures within the data that may not be apparent from summary statistics alone. These plots help identify clusters, outliers, and regions of high or low density, aiding in understanding the shape and characteristics of the distribution. Quantile contour plots allow us to assess data variability across different quantiles. By observing the spacing and shape of the contour lines, we can infer how the spread of values changes across the range of quantiles. This information is precious in fields such as finance, where understanding the variability of asset returns or risk measures at different percentiles is crucial. Quantile contour plots effectively identify extreme values or outliers in the data. Outliers can be detected as isolated points or regions with significantly different contours from the rest of the dataset. Identifying these extreme observations is important in many applications, such as anomaly detection or quality control, where detecting unusual values is critical. In summary, quantile contour plots play a vital role in understanding a dataset's distribution, variability, and relationships. They are powerful data exploration, visualization, and communication tools, providing valuable insights that can drive further analysis and decision-making.

3.4 Simulation and outliers detection based on depth functions.

Outlier detection based on depth functions is a statistical method used to identify observations that deviate significantly from the majority of the data. Depth functions assign a value to each data point, indicating its relative centrality within the dataset. Outliers are then identified as points with lower depth values. Once the depth values are calculated for each data point, a threshold is set to determine which points are outliers.in this thesis, we set the 90th percentile of the halfspace depths as a cutoff. The specific threshold depends on the application and can be determined empirically or based on the statistical properties of the depth values. We can note that depth-based methodsare robust to outliers and can handle high-dimensional data effectively. They are instrumental in situations where traditional distance-based methods may fail due to the presence of outliers or skewed distributions.

3.4.1 Outliers detection based on Halfspace depth (or Tukey depth).

Outlier detection based on halfspace depth (also known as Tukey depth) is a statistical technique used to identify outliers in a dataset. Halfspace depth is a measure of centrality or outlyingness of a point in a dataset with respect to a set of hyperplanes. It quantifies how far a point is from

the center of the dataset. Halfspace depth is computed by determining the fraction of hyperplanes that separate a point from the other points in the dataset. Here is a step-by-step procedure for outlier detection using Tukey depth: We first calculate Tukey Depth for Each Data Point, then Sort the dataset in ascending order to calculate the median of the dataset (M). then Divide the dataset into two halves: values less than or equal to the median (group L) and values greater thanor equal to the median (group H). After that, we calculate the median of each group (ML for group L, MH for group H). then calculate the Tukey depth (TD) for each data point (x_i) using the formula: $TD(x_i) = 0.5 * (1 + (R(x_i) - ML)/(MH - ML))$ where $R(x_i)$ is the rank of the data point x_i in the sorted dataset. Halfspace depth can be computed using the following formula (1.2.1): $HSD(x) = min\{F(x, H) : H$ is a closed halfspace containing $x\}$.

To Identify Outliers, we compare each data point's Tukey depth (TD) to the critical value (CV) determined by the Tukey depth. Finding the critical value (CV) for outliers detection based on Tukey depth involves determining the threshold values that define the Tukey depth. The critical value is represented by the constant "k" A common choice for "k" is 1.5, but it can be adjusted depending on the specific characteristics of your dataset and the desired sensitivity to outliers. Generally, a larger "k" value will result in fewer outliers being detected as it widens the range for potential outliers. Conversely, a smaller "k" value will lead to more outliers being identified as the range for potential outliers becomes narrower. Remember that outlier detection is a crucial step in data analysis. However, it is essential to interpret and handle outliers carefully based on the data's context and the analysis's goals. Not all outliers are necessarily errors or anomalies; they could carry valuable information or indicate exciting patterns in the data.

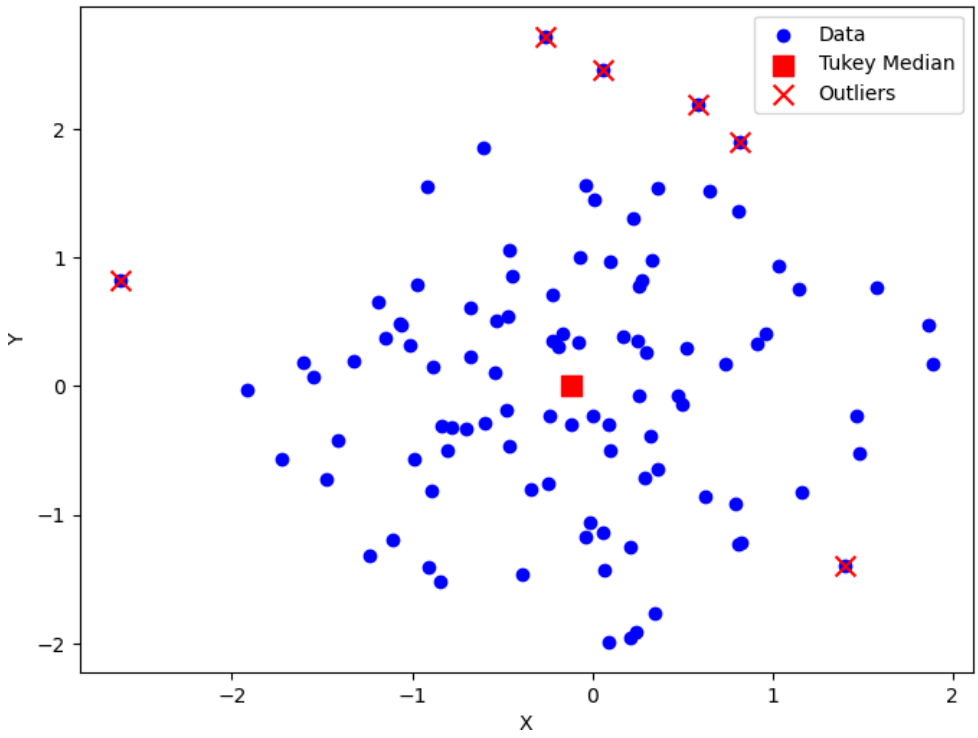

Figure 3.7: Outlier detection based on Tukey depth.

With the help of standard normal distribution, we simulated 100 observations and identified outliers using the Tukey depth method (see Figure 3.7). From Figure 3.7, we flag six data points corresponding with the halfspace depth exceeding the threshold as outliers.

In Tukey's depth method for identifying outliers, the distance of a point from the center (median) is not the only factor considered. Tukey's depth is based on the concept of regions of depth, and points outside these regions are flagged as potential outliers. Even if a point has a small distance from the center (median), it might still be considered an outlier if it falls outside the specified fences. This is because Tukey's depth takes into account not only the relative position of a point but also the distribution of the data. If a point is significantly distant from the bulk of the data, it is flagged as a potential outlier.

In summary, outliers in Tukey's depth method are determined by both the point's depth relative to other points and its position with respect to the defined fences based on quartiles and a user-defined constant. Small distances from the center alone may not be sufficient to exclude a point from being considered an outlier if it falls outside the specified boundaries.

3.4.2 Outliers detection based on Mahalanobis depth.

Mahalanobis depth is a measure that can be used for outlier detection in multivariate datasets. It assesses the depth of a point within a distribution, considering the correlation and covariance structure of the variables. Data should be adequately preprocessed and standardized to performa Mahalanobis depth analysis. The variables should have zero mean and unit variance, as Mahalanobis distance is sensitive to scale. For each data point, calculate its Mahalanobis distance using equation 1.2.4. Note that the Mahalanobis depth method assumes that the data follow a multivariate Gaussian distribution. Therefore, if the data violates this assumption, the results may not be accurate. Additionally, outliers can be influenced by the size of the dataset, dimensionality, and the presence of influential observations.

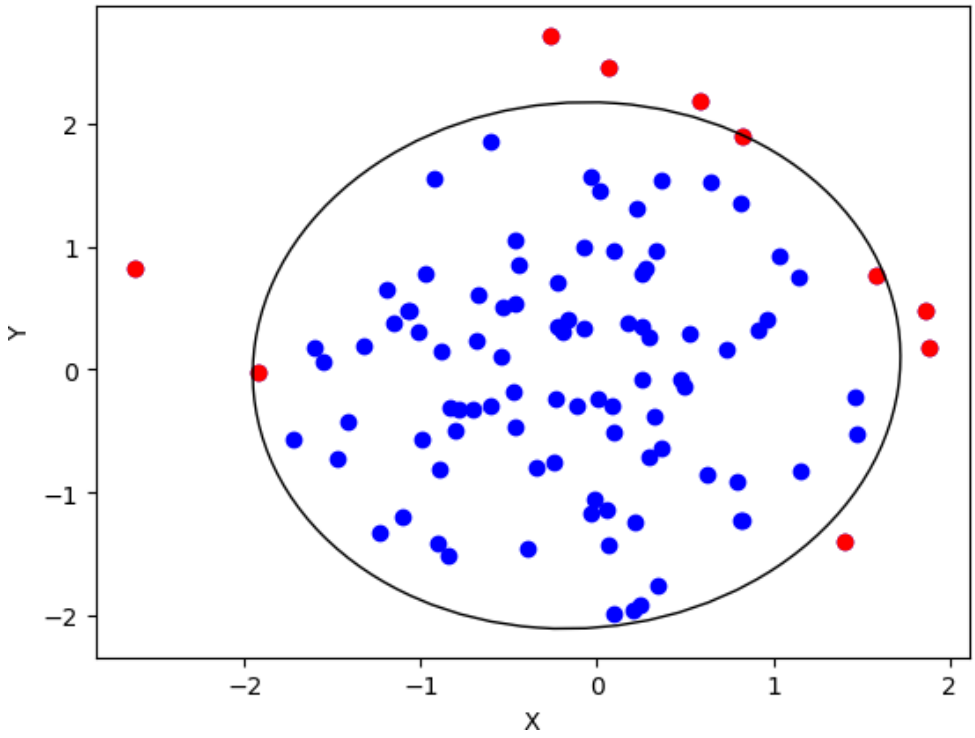

Figure 3.8: Outlier detection based on Mahalanobis depth.

In Figure 3.8, We generate 100 observations from a standard normal distribution, calculate the Mahalanobis distance for each observation in the dataset, then calculate the Mahalanobis depth for each data point. The Mahalanobis depth represents the depth of a point within the data distribution, indicating how far it is from the boundaries of the distribution. In this experiment, we used a chi-square distribution with two degrees of freedom. Then we determine a threshold value to classify points as outliers. The threshold represents the minimum Mahalanobis depth required fora point to be considered an outlier. Choosing an appropriate threshold depends on the specific

application and desired level of sensitivity to outliers. Here we used quantiles of the Mahalanobis depth distribution with the 90^{th} percentile as the threshold because we used the same level in other methods used here and needed to compare them. Lastly, we compare the Mahalanobis depth of each data point with the chosen threshold value based on the significance level $alpha = 0.10$. Ten red points are observations with Mahalanobis distance greater than the threshold and are outliers.

3.4.3 Spatial depth and spatial outlyingness.

A quantile contour plot based on spatial depth and spatial outlyingness is a graphical tool to identify spatial outliers in a dataset. It is based on the concepts of spatial depth and spatial outlyingness, which are used to measure the centrality and outlyingness of a point in a spatial dataset. Spatial depth is a measure of how central a point is in a dataset based on its location relative to other points. Points with a high spatial depth are located in the center of a dataset, while points with a low spatial depth are located on the edges or in the outliers of a dataset. Spatial outlyingness, however, measures how different a point is from its neighbouring points regarding its attribute values. Points with a high spatial outlyingness are significantly different from their neighbours, indicating that they may be spatial outliers. To create a quantile contour plot based on spatial depth, we would first calculate the spatial depth for each point in the dataset. This can be done using various algorithms, such as the Half Space Depth (HSD) and the Local Outlier Factor (LOF). Once we have calculated the spatial depth for each point, we can plot them on a two-dimensional graph. The plot is divided into a grid of cells, each containing a specific range of spatial depth and spatial outlyingness values. The contour lines on the plot represent the quantiles of the dataset. The contour lines closer to the center of the plot represent the higher quantiles, while the contour lines further away from the center represent the lower quantiles. The points in the dataset are then plotted on top of the quantile contour plot. Points that fall outside the contour lines are considered spatial outliers and may require further investigation.

The idea of spatial depth was introduced by Brown [7]; he introduced the concept of spatial median and investigated the location estimation for spatial data in two dimensions. Also, [35] studied multivariate spatial depth based on the geometry of the data. He considered a random variable Y with distribution F and the spatial depth of point $x \in \mathsf{R}^d$ related to F is given by:

$$D_s(x, F) = 1 - \int S(x - y)dF(y) \quad . \| \tag{3.4.1}$$

When we have a sample of curves say (Y_i), $i = 1, \cdots, n$, the spatial depth can be calculated with the

following formula

$$D_{s,n}(x, F_n) = 1 - \left\| \frac{1}{n} \sum_{y} S(x - y) \right\|, \tag{3.4.2}$$

Where $S(\cdot)$ is a spatial sign function given by:

$$S(x) = \begin{cases} \frac{x}{\|x\|}, & \text{when } x \neq 0 \\ 0, & \text{when } x = 0 \end{cases}.$$

Recalling the measure of outlyingness $\|Q_F^{-1}(x)\|$ with respect to a distribution function F on R^d and spatial quantile function Q_F associated with F, the spatial depth is defined as :

$$D_s(x, F) = 1 - \|Q_F^{-1}(x)\| \tag{3.4.3}$$

From the relation 3.4.1, if x has a low spatial depth value, its associated spatial quantile of vector u has a high norm, and vice versa. From the above discussion, this depth function is affine equivariance. The spatial median is the point of maximal depth while outlyingness increases with respect to decreasing depth. In the one-dimensional case (that is $d = 1$) and for $-1 < u < 1$, the quantile function is simply computed as follows $Q_F(u) = F^{-1}(1/2 + u/2)$ and median med $= Q_F(0) = F^{-1}(1/2)$ which results $Q_F^{-1}(x) = 2F(x) - 1$ and then in R, $\|2F(x) - 1\|$ is a measure of the outlyingness of x with respect to the distribution F. From the equation 3.4.1 combined with the above results, the spatial depth is defined as :

$$D_s(x, F) = 1 - \|2F(x) - 1\| = 2 \min\{F(x), 1 - F(x)\}$$

Considering the case when F is the uniform distribution on the unit square, the spatial depth functions used for this distribution function are easy and straightforward to obtain. We consider a contrived bivariate data set of size $n = 20$; the spatial depth can be computed with the help of the equation 3.4.2 or following equation. Both equations give the same results:

$$D_s((x, y), F) = 1 - \sqrt{a^2(x, y) + a^2(y, x)},$$

With $a(x, y) = \int_0^1 \left(\sqrt{(1 - x)^2 + (t - y)^2} - \sqrt{x^2 + (t - y)^2} \right) dt$. The spatial median is a point of maximal depth, and decreasing depth corresponds to increasing outlyingness [35].

From the data set of Table 3.2, the spatial median is (0, 0), while the (0, 20) and (0, -20) are the

most outlying, and (0, 1) and (0, -1) the most central points.

x	$D_{s,n}(x, F)$
(0,1)	0.830595
(0,-1)	0.830595
(1,0)	0.630929
(0,1.5)	0.770295
(0,3)	0.638976
(-3,0)	0.66401
(1.5,0)	0.581474
(0,5)	0.503291
(-5,0)	0.679659
(-10,0)	0.502024
(10,0)	0.502024
(0,-12)	0.326771
(0,15)	0.308963
(-15,0)	0.353535
(-16,0)	0.246347
(0,-17)	0.201388
(0,18)	0.197523
(-20,0)	0.122997
(0,20)	0.091734
(0,-20)	0.091734

Table 3.2: Spatial depth for contrived bivariate dataset.
This contrived bivariate dataset was explained in Table 3.1.

In summary, a quantile contour plot based on spatial depth is a helpful tool for identifying spatial outliers in a dataset. It allows you to visualize the centrality and outlyingness of each point in the dataset and provides a way to quantify the degree of outlyingness for each point.

Chapter 4

Applications

Quantiles and depth functions are commonly used in outlier detection to identify and analyze datasets' potential outliers. Quantiles divide a dataset into equal-sized subsets, providing informa- tion about the distribution of the data. Outliers can be identified by comparing data points to the quantiles. For example, the median is the 50^{th} percentile, dividing the data into two equal halves. If a data point significantly deviates from the median, it can be considered a potential outlier. Also, quartiles divide the data into four equal parts. The interquartile range (IQR) is determined by the lower quartile, which represents the 25^{th}percentile, and the upper quartile, which corresponds to the 75^{th} percentile. Data points outside the range of 1.5 times the IQR can be flagged as outliers. Percentiles other than quartiles, such as deciles or percentiles, can also be used to identify outliers based on the desired granularity.

Depth functions also play a considerable role in outliers detection. Depth functions measure the centrality of a data point within a dataset by assigning a depth value. Points with lower depth values are more likely to be outliers. In this thesis, we applied Half-space depth, Mahalanobis depth and spatial depth to air quality data to detect outliers. It is worth mentioning that these methods provide a starting point for identifying potential outliers, and further analysis or domain knowledge is often required to confirm and interpret outliers appropriately.

In this section, we analyze outliers detection based on quantiles and depth functions, compute geometric quantiles contour plots using the air-quality dataset consisting of 153 observations of 6 variables of measurements of 1973 for five months from May to September recorded daily. For more information about dataset, check https://r-data.pmagunia.com/iframe/airquality.html.

4.1 Application of univariate quantiles to air quality data.

Univariate quantiles can be used to detect outliers in a dataset by considering every single variable. Univariate quantiles provide a way to measure the position of individual data points within

a dataset's distribution. Using univariate quantiles for outlier detection is a simple and widely applicable method. However, we must rely on something other than this method for more complex scenarios or when considering multiple variables simultaneously.

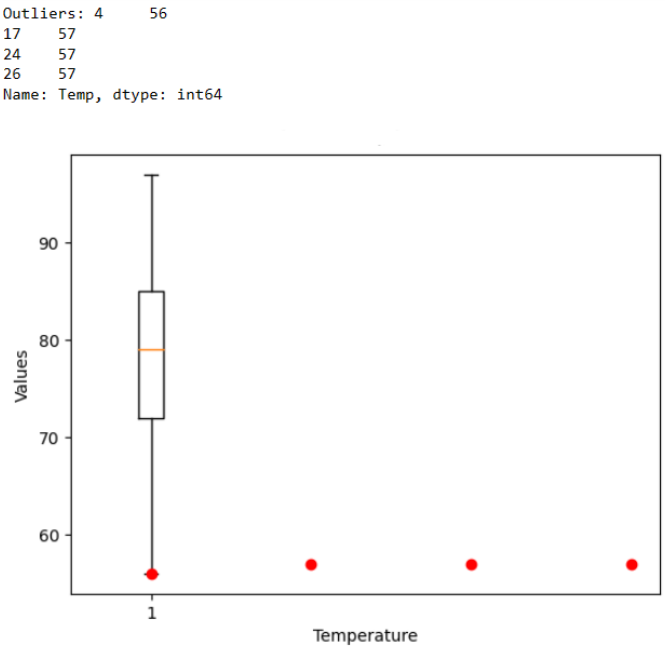

Figure 4.1: Outlier detection of temperature values based on univariate quantiles using boxplots. Temperature in degrees Fahrenheit.

By performing the same procedure as in Chapter 3 to the temperature values from the air quality dataset, four outliers are identified from 153 observations (Figure 4.1). Applying univariate quantiles to air quality data for outlier detection has some drawbacks: Univariate quantiles only consider a single variable at a time, ignoring potential relationships and dependencies between multiple variables. Various factors, such as temperature, solar radiation, wind speed, and pollutant concentrations, influence air quality. By focusing on a single variable, crucial contextual information may be overlooked. Univariate quantiles are not effective in detecting outliers that co-occur across multiple variables. Multivariate outliers, which may indicate abnormal patterns or interac- tions between different air quality parameters, could be missed using this approach. To overcome this issue, we use geometric quantile contour plots for multivariate outlier detection algorithms.

Air quality data often exhibit temporal and spatial variations. Univariate quantiles do not account for spatial context, so they may not be sensitive to outliers occurring only at specific locations. An outlier detection method incorporating spatial analysis would be more suitable in such cases. In this thesis, we used spatial depth analysis, and this method can provide a more comprehensive and robust analysis of outliers in air quality data.

4.2 Application of geometric quantile contour plot to air-quality data.

Applying geometric quantiles to air quality data can provide valuable insights and facilitate a better understanding of the distribution and variability of pollutant levels in the atmosphere. Geometric quantile contour plots are valuable tools for visualizing and analyzing multivariate data. They pro- vide insights into the distributional characteristics of a dataset and can be applied in various fields. Here are some applications of geometric quantile contour plots: 1. Risk Analysis: Geometric quantile contour plots can be used in risk analysis to visualize and understand the risk profiles ofdifferent variables or portfolios. Plotting the data's quantiles on contour lines makes it easier to identify regions of high or low risk and examine the relationship between different variables. (Forexample, [12]) 2. Environmental Monitoring: Geometric quantile contour plots can be employed to study environmental variables, such as pollutant concentrations or air quality measurements. These plots help identify areas with high or low concentrations, assess the spatial distribution of pollutants, and monitor changes over time. For example, [9]) considered the Kola Ecogeochemistry project data dealing with the detection of pollution rates around an industrial zone from measures of barium (Ba) and calcium (Ca) found in different plants. 3. Financial Data Analysis: In finance, geometric quantile contour plots can assist in analyzing multivariate financial data. They can be used to visualize the joint distribution of asset returns or portfolio performance, identify risk di- versification opportunities or assess tail risk. 4. Quality Control: Geometric quantile contour plots are helpful in quality control to analyze multivariate data and identify regions that deviate from desired specifications. Plotting the quantiles on contour lines makes it easier to detect outliers, as- sess the distributional characteristics of the data, and make decisions based on quality standards. 5. Machine Learning: Geometric quantile contour plots can aid in understanding and evaluating the performance of machine learning models. They can be used to visualize the relationship between model inputs and outputs, analyze prediction uncertainties, and assess the model's robustness and sensitivity to different variables. 6. Spatial Analysis: Geometric quantile contour plots can be applied in spatial analysis to study geographical data, such as temperature, rainfall, or population density. They help visualize the spatial patterns and variations in the data, identify hotspots or cold spots, and support decision-making in urban planning, resource management, or epidemiology.

In this thesis, geometric quantile contour plots will help analyze the air quality dataset. These are just a few examples of how geometric quantile contour plots can be applied. The plots provide valuable insights into the distributional properties of multivariate data and offer a visual represen- tation of relationships and patterns that might not be immediately apparent from the raw data.

Applying an air quality dataset to geometric contour quantiles can help analyze and visualize the distribution of air pollution levels across different geographical areas. Here are a few reasons

why it can be a perfect idea: Geometric quantiles contour plot provides a spatial representation of data by dividing the study area into regions with similar values. It allows for clearly visualizing how air quality varies across different locations and helps identify outlier regions. Applying geometric quantile contour plots to an air-quality dataset allows for aggregating the data points within each region and calculating summary statistics, such as the mean, median, or percentiles. It helps in understanding the overall air pollution levels in specific areas. Geometric quantiles contour plots enable easy comparison between different regions. By overlaying the contours of multiple areas, we can visually compare air quality levels and identify patterns or disparities. It can be handy for identifying pollution hotspots or areas with better air quality. To the decision-makers, analyzing air-quality data using geometric quantiles contour plots can provide valuable insights for decision-making processes. It can help policymakers, urban planners, and environmental agencies identify areas requiring targeted interventions for improving air quality and prioritize resources accordingly. Lastly, geometric quantiles contour plots can be visually appealing and easily understandable by the general public. By presenting air quality information through contour maps, it becomes easier to communicate the extent and severity of pollution in different areas, raising public awareness about environmental issues.

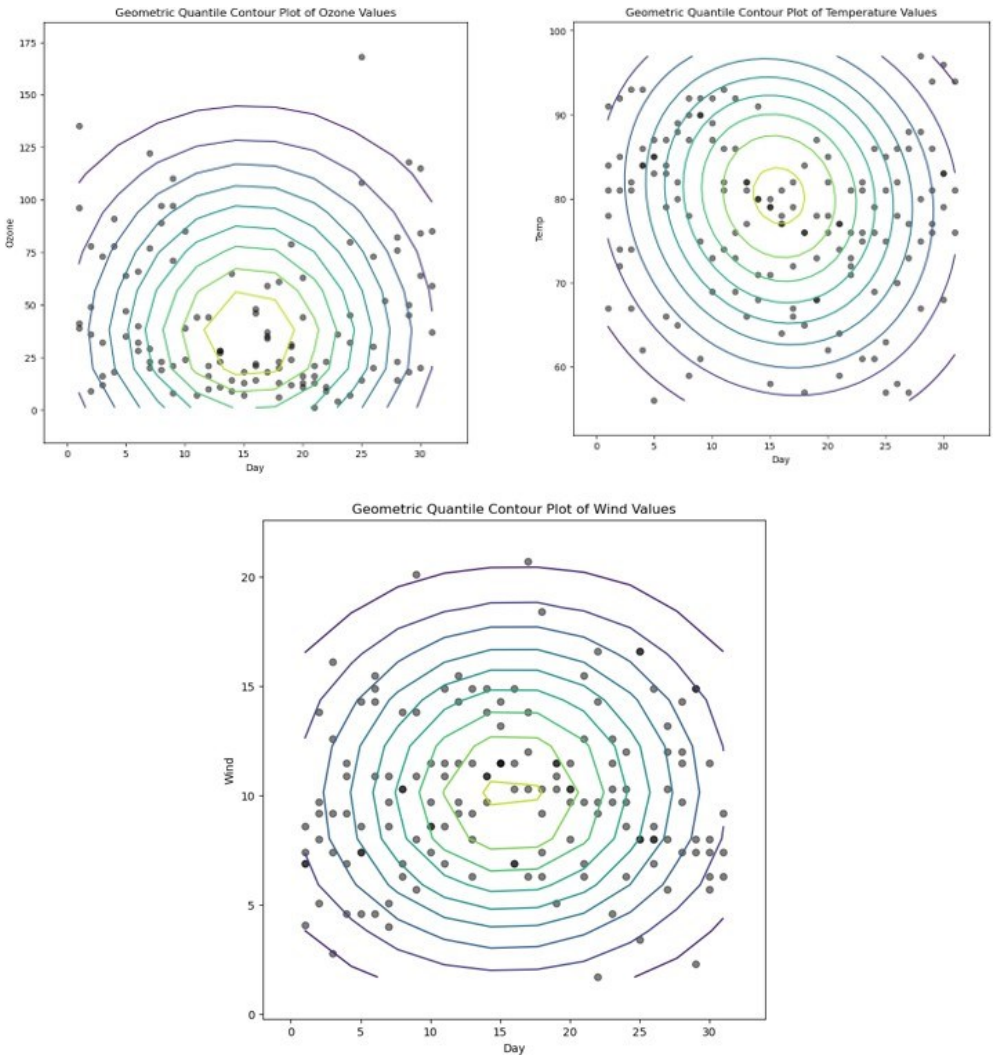

Figure 4.2: Geometric quantiles contour plot of wind, ozone and temperature vs day. Wind speed in miles per hour, ozone in parts per billion and temperature in degrees Fahrenheit.

Figure 4.2 shows a quantile contour plot from 10% to 90% for wind values (at the bottom), ozone (at the top left corner), and temperature (at the top right corner) recorded daily over a four-month, and Figure 4.3 represents a quantile contour plot from 10% to 90% for wind values and temperature values recorded monthly over a four-month period.

Quantile contour plot from 10% to 90% for wind values (at the bottom), ozone (at the top left corner), and temperature (at the top right corner) recorded daily over four months. The plot representing wind was generated based on the calculation of quantiles at 10%, 20%, 30%, ..., 90% for each day by grouping the wind values by the day they were recorded. This step allows calculating the quantiles for each day separately. Once the quantiles for each day are calculated, we create contour plots. Contour plots typically involve two variables, such as wind and day values, with the wind speed values quantiles on the y-axis and recorded days on the x-axis, and use contour lines

to represent the quantile levels. From the bottom figure (representing wind values), only three observations are located outside of the contour line corresponding with $Q(u) = 0.90$. The first was recorded on the 9^{th} day in May with a wind speed value of 20.1 miles per hour; the second was recorded on the 1^{s} day in June with a wind speed value of 20.7 miles per hour; the third was recorded on the 22^{nd} day in June with wind speed value of 1.7 miles per hour. Those observations are labelled as outliers. From the top left corner figure (representing ozone values), four observations are located outside of the contour line corresponding with $Q(u) = 0.90$. Those four detected outliers are 168, 135,118 and 115 parts per billion (ppb) measured on 25^{th} August, 1^{st} January, 29^{th} August and 30^{th} May, respectively. Lastly, four temperature records were marked as outliers (figure at the top right corner). These measurements recorded on 5^{th} May, 29^{th} August, 30^{th} August, and 31^{st} August are 56, 94,96 and 94 degrees Fahrenheit. Geometric quantiles are less sensitive to extreme values or outliers than traditional quantile methods such as median or quartiles. This robustness is particularly relevant in air quality data, as pollutant concentrations can sometimes exhibit extreme values due to sporadic events or measurement errors. By considering geometric quantiles, we can obtain a more reliable representation of the central tendency and spread of the data distribution.

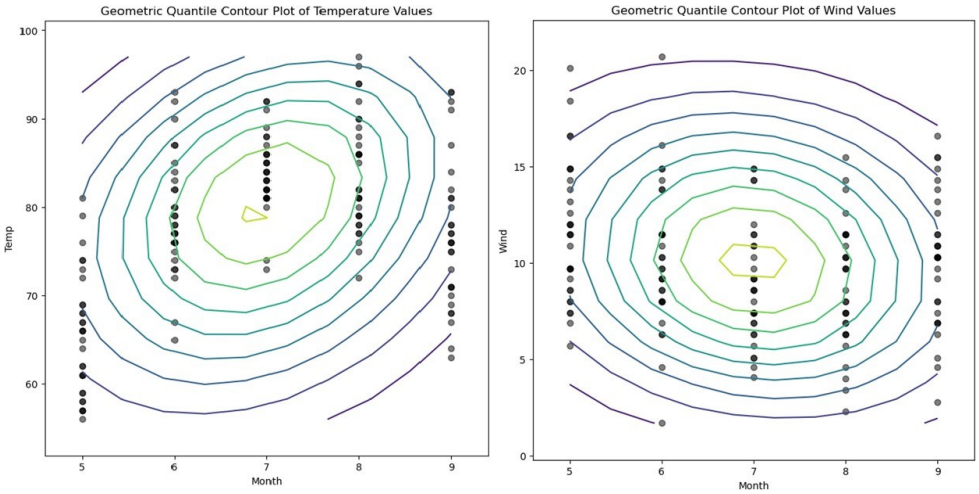

Figure 4.3: Geometric quantiles contour plot of wind, and temperature vs month. Wind speed in miles per hour and temperature in degrees Fahrenheit.

Air pollution has significant implications for human health. Geometric quantiles can be used to estimate the exposure levels experienced by individuals at different quantile levels. This information is crucial for assessing the health risks associated with different pollutant levels and establishing guidelines for minimizing exposure to harmful pollutants.

Figure 4.3 (left) plots monthly temperature values; three outliers were identified in August, and one was identified in May. Those measurements are precisely the same found in Figure 4.2, also,

Figure 4.3 (right) represents wind values recorded monthly; three outliers were recorded, one in May and 2 in June. These outliers are the same as in the Figure 4.2.

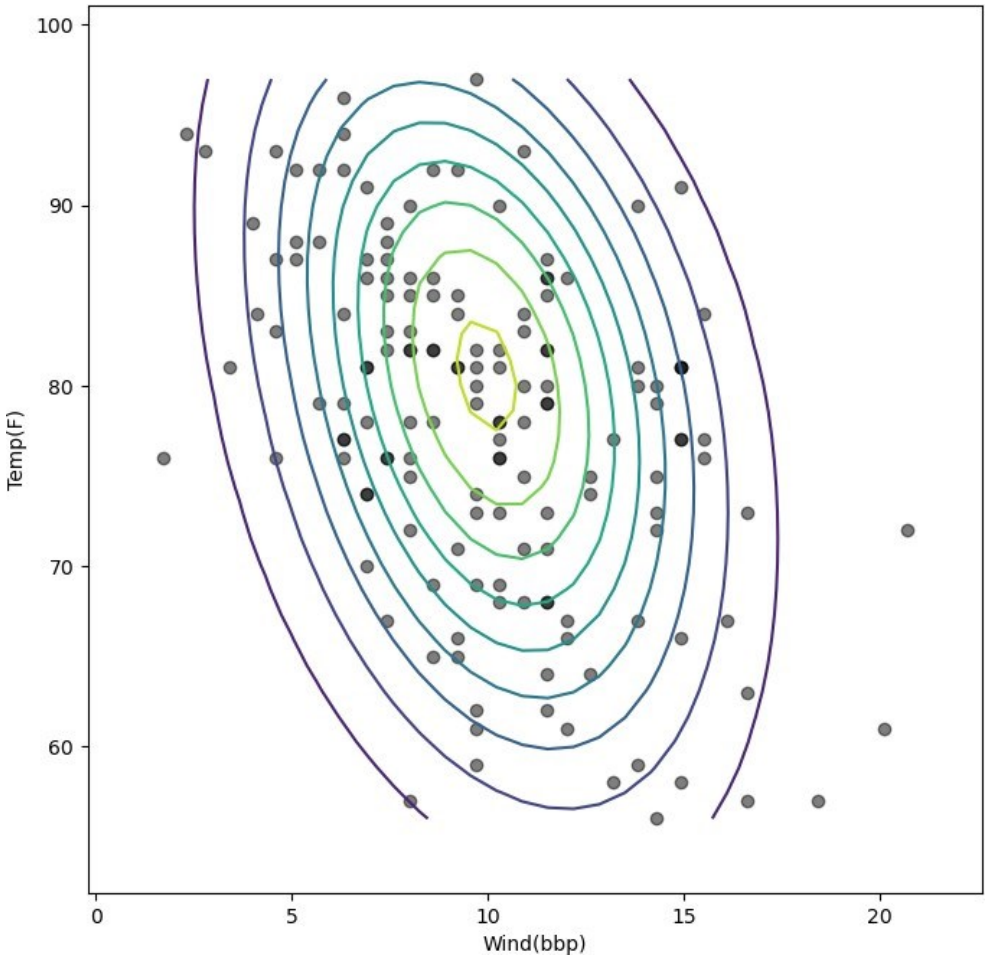

Figure 4.4: Geometric quantiles contour plot of temperature and wind. Wind speed in miles per hour and temperature in degrees Fahrenheit.

The geometric quantiles contour plot of temperature versus wind offers several advantages in visualizing and analyzing the relationship between these two variables. The contour plot presents a comprehensive overview of the joint distribution of temperature and wind. By plotting multiple quantiles (10^{th}, $\cdots$, 90^{th}), the contour plot reveals how temperature and wind interact across different levels or percentiles. This visualization enables a better understanding of the relationship between the two variables throughout the range of data. In Figure 4.4, six outliers were identified based on the 90^{th} quantiles.

Overall, applying air quality datasets to geometric contour quantiles enhances the analysis and visualization of spatial data, enabling a better understanding of air pollution patterns and support- ing informed decision-making for environmental management.

Geometric quantiles allow us to describe pollutant levels at different quantile levels, such as the 90th percentile or the 95th percentile. These quantiles are often used as thresholds or benchmarks for assessing air quality and determining compliance with regulatory standards. By analyzing geometric quantiles (in Figure 4.2 and Figure 4.3), we identified high-pollution episodes, evaluated the frequency and duration of pollutant exceedances, and assessed the overall severity of air pollution. Geometric quantiles provide a concise summary of the data distribution and can support decision-making processes related to air quality management. Policymakers can utilize information derived from geometric quantiles to set appropriate air quality standards, design pollution control strategies, and allocate resources effectively to areas with the highest pollution levels.

Applying geometric quantiles to air quality data enables a more comprehensive analysis of pollutant concentrations, enhances our understanding of spatial and temporal patterns, supports health risk assessment, and informs decision-making processes for improving air quality.

4.3 Application of Halfspace depth (or Tukey depth) to air-quality data

Halfspace depth, also known as Tukey depth, quantifies the centrality of a point within a dataset. It is advantageous in multivariate analysis to determine the depth of a point relative to a given dataset.

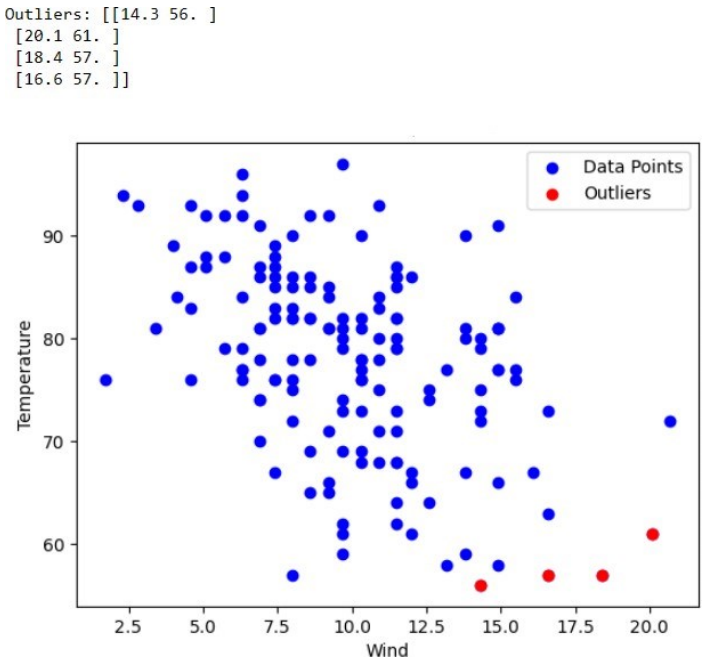

Figure 4.5: Outlier detection of temperature and wind based on Tukey depth. Wind speed in miles per hour and temperature in degrees Fahrenheit.

In the context of outlier detection using Halfspace depth (also known as Tukey depth), the goal is to identify data points that are extreme or unusual compared to the rest of the data based on their combinations of two or more variables, typically wind and temperature. When you talk about identifying "outlier combinations of wind and temperatures using Halfspace depth," it means you want to find data points where the combination of wind and temperature values is extreme or unusual compared to the rest of the data when considering all possible directions or combinations of these two variables.

Our air quality data typically consists of multiple variables, such as pollutant concentrations (e.g., ozone), meteorological parameters (e.g., temperature, humidity, wind speed), and geographical coordinates. These variables collectively define the air quality conditions at different monitoring locations or over time. This method calculates the air quality dataset's halfspace depth for each point. This involves computing the depth value for each point based on its position relative to the other points in the dataset. Higher depth values indicate points that are more central within the dataset, while lower values represent more extreme or peripheral points, which can be labelled as outliers.

Using Halfspace depth (Tukey depth) for outlier detection in combinations of wind and temperature helps you find data points where the joint behaviour of these variables is different from what you would typically expect, making them potential candidates for further analysis or anomaly detection in applications such as environmental monitoring, climate research, or quality control.

Ozone	Solar.R	Wind	Temp
Confidence level=90%($\theta = 0.10$)			
8	19	20.1	61
6	78	18.4	57
1	8	9.7	59
11	320	16.6	73
37	284	20.7	72
135	269	4.1	84
16	7	6.9	74
122	255	4	89
168	238	3.4	81
118	225	2.3	94
73	183	2.8	93
Confidence level=95%($\theta = 0.05$)			
8	19	20.1	61
6	78	18.4	57
11	320	16.6	73
37	284	20.7	72
168	238	3.4	81
118	225	2.3	94
Confidence level=99%($\theta = 0.01$)			
1	8	9.7	59
4	25	9.7	61

Table 4.1: Multivariate outliers detection based on Halfspace depth.
Ozone in parts per billion, solar.R in the frequency band, wind speed in miles per hour and temperature in degrees Fahrenheit. The term "cutoff" typically refers to a threshold or boundary value used to determine whether a particular depth measurement is significant or not.

Applying halfspace depth to air quality data helps us identify four outliers from the pattern of temperature and wind values (Figure 4.5). Those outliers are also identified in Figure 4.4 by geometric quantiles contour plot methods. Halfspace depth can be extended to multivariate (more than two variables) scenarios, and here we consider a dataset with multiple variables (dimensions). In this context, the halfspace depth of a point is defined as the minimum number of points in any halfspace that contains the point. The idea behind this approach is that outliers tend to have low

halfspace depth because they lie far away from the bulk of the data points.

Table 4.1 shows outliers based on halfspace depth of multivariate variables of air quality by considering four columns: "Ozone," "Solar.R," "Wind," and "Temp" with cutoff values of 1%,5% and 10%. For example, the "cutoff" value is set here at the 5^{th} percentile of the halfspace depths, meaning we are considering the lowest 5% points with the lowest halfspace depth as outliers. It is important to note that outlier detection is a nuanced task, and the choice of method and the critical value depends on various factors, such as the nature of the data, the underlying assumptions, and the context of the analysis.

The application of halfspace depth to air quality data provides valuable insights for outlier detection, robust measures of central tendency, data visualization, anomaly detection, and early warning systems. It helps to understand the distributional characteristics of the data and improve the analysis and decision-making processes related to air quality management and monitoring.

4.4 Application of Mahalanobis depth to air-quality data

Mahalanobis depth is a statistical measure used to determine the distance of a data point from the center of a multivariate distribution. It considers the interdependencies among variables through their covariance structure of the data, making it a valuable tool for outlier detection and anomaly analysis. Mahalanobis depth can help detect abnormal patterns or anomalies in air-quality data. By calculating the Mahalanobis depth for each data point, we can identify outliers that deviate significantly from the typical multivariate distribution of the air quality parameters. These outliers could indicate potential sensor malfunctions, extreme pollution events, or unusual atmospheric conditions. Data points with a low Mahalanobis depth value, indicating they are close to the center of the distribution, can be considered representative and reliable observations. Conversely, data points with a high Mahalanobis depth, indicating they are far from the center of the distribution, might raise concerns about measurement errors, data corruption, or other issues that require further investigation.

By applying Mahalanobis depth to air quality data by considering two variables (temperature and wind), using Chi-squared distribution with a confidence level of 90%, we identify six outliers that deviate significantly from the expected behaviour (Figure 4.6). These outliers could represent abnormal air quality conditions or measurement errors, allowing for targeted investigation and potential corrective actions.

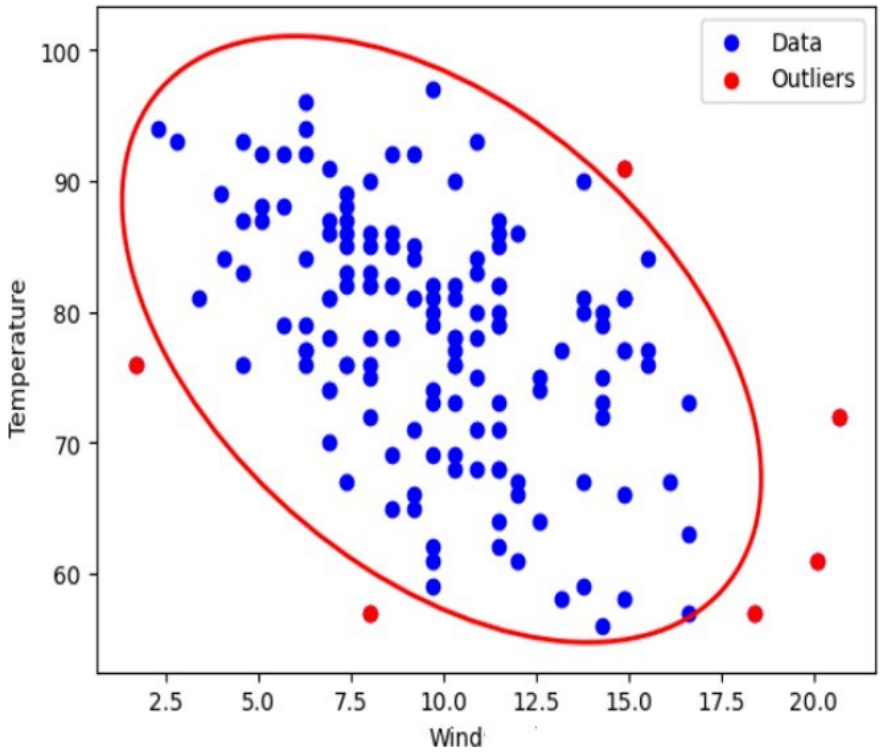

Figure 4.6: Outlier detection of temperature and wind based on Mahalanobis depth. Wind speed in miles per hour and temperature in degrees Fahrenheit.

When using the Mahalanobis depth for outlier detection, we are looking for data points that have unusually large Mahalanobis distances. This would mean wind and temperature combinations significantly differ from the average combinations observed in our air quality dataset while considering the correlation (covariance) between these two variables. Applying the Mahalanobis depth method to the wind and temperature variables lets you calculate Mahalanobis distances for each data point. Data points with Mahalanobis distances above a certain threshold are considered outliers. These outlier combinations of wind and temperature are those that deviate significantly from the typical patterns observed in your dataset, considering the joint behaviour of both vari- ables.

Detecting outliers using Mahalanobis depth can be done for both multivariate variables (more than two variables) and bivariate variables, but the approach is slightly different in each case. Mahalanobis depth measures how far a data point is from the center of a multivariate dataset, taking into account the covariance structure of the data. It is particularly useful for detecting outliers in datasets with multiple correlated variables.

In summary, Mahalanobis depth can be used for both multivariate and bivariate variables to detect outliers. For multivariate data, the depth considers the entire covariance structure, while for bivariate data, it simplifies to a measure based on the covariance between the two variables.

We generated Table 4.2 to show outliers detected by Mahalanobis depth of multivariate variables of air quality by considering four columns: "Ozone," "Solar.R," "Wind," and "Temp" with cutoff values of 1%,5% and 10% and figure 4.6 shows outliers detected by Mahalanobis depth of bivariate variables of air quality by considering two columns: "Wind," and "Temp".

Using the Mahalanobis depth method to detect outlier combinations of wind and temperature means you're looking for unusual pairs or combinations of wind speed and temperature values in your dataset that stand out when considering the covariance structure between these two variables.

Cutoff	Ozone	Solar.R	Wind	Temp
Confidence level=90%($\theta = 0.10$)				
3	18	313	11.5	62
8	8	19	20.1	61
15	14	334	11.5	64
17	6	78	18.4	57
20	1	8	9.7	59
29	115	223	5.7	79
47	37	284	20.7	72
61	135	269	4.1	84
116	168	238	3.4	81
128	32	92	15.5	84
147	14	20	16.6	63
Confidence level=95%($\theta = 0.05$)				
8	8	19	20.1	61
17	6	78	18.4	57
29	115	223	5.7	79
47	37	284	20.7	72
61	135	269	4.1	84
116	168	238	3.4	81
Confidence level=99%($\theta = 0.01$)				
47	37	284	20.7	72
116	168	238	3.4	81

Table 4.2: Multivariate outliers detection based on Mahalanobis depth.
Ozone in parts per billion, solar.R in the frequency band, wind speed in miles per hour and temperature in degrees Fahrenheit. The cutoff serves as a threshold for identifying outliers.

Applying Mahalanobis depth to air quality data offers a practical statistical approach for identifying outliers, detecting anomalies, optimizing monitoring networks, implementing early warning systems, and ensuring data quality. It can assist in improving air quality management strategies, facilitating timely interventions, and enhancing our understanding of complex multivariate air quality datasets. Applying Mahalanobis depth to air quality data enhances data analysis, quality control, decision-making, and the development of effective strategies for air quality management. By incorporating this statistical tool, it becomes possible to derive valuable insights and improve our understanding of complex air pollution dynamics.

4.5 Application of Spatial depth to air quality data

Spatial depth is a non-parametric approach, meaning it does not make assumptions about the underlying data distribution. This makes it applicable to a wide range of data types and ensures its effectiveness in various domains and datasets. Spatial depth is a statistical measure that provides information about the centrality and outlyingness of data points in a multidimensional space. It can be applied to various data types, including air quality data, to gain insights into pollution levels' spatial distribution and patterns. Here are some potential applications of spatial depth to air quality data:

a. Identifying pollution hotspots: Spatial depth can help identify areas with consistently high or low levels of air pollution. By calculating the spatial depth of pollution measurements at different locations, we can pinpoint regions that deviate significantly from the central tendency. These hotspots may indicate areas with poor air quality that require immediate attention and targeted interventions.

b. Assessing spatial variations: Air pollution can exhibit significant spatial variations due to various factors such as local emissions sources, topography, and meteorological conditions. Spatial depth analysis can help characterize the extent and magnitude of these spatial variations. By calculating spatial depth at multiple locations, we can determine whether pollution levels are consistent across the region or if there are distinct clusters or gradients of pollution.

c. Outlier detection: Outliers in air quality data can indicate unusual pollution events or measurement errors. Spatial depth can be used to identify such outliers by assessing the relative outlyingness of individual data points. It can help detect anomalous pollution episodes requiring investigation or data validation.

d. Spatial interpolation: Spatial depth measures can be incorporated into interpolation techniques to generate more accurate estimates of air pollution levels in locations where measurements are unavailable. By considering the spatial depth of neighbouring measurements, the interpolation algorithm can give higher weight to data points more representative of the central tendency,

improving the reliability of interpolated values.

e. Environmental monitoring network optimization: Spatial depth analysis can guide the optimization of air quality monitoring networks. By identifying areas with high spatial depth values, indicating a significant departure from the central tendency, we can determine the most critical locations for placing monitoring stations. It helps ensure that monitoring efforts focus on areas where pollution levels vary significantly from surrounding regions.

Spatial depth (also called L1-depth) is a depth function based on distance exploiting the idea of spatial quantiles introduced by [11] and [24], formulated by [38] and [35].

Day	Wind	Temp	$D_{s,n}(x, F)$
1	7.4	67	0.4268
2	8	72	0.3462
3	12.6	74	0.2856
4	11.5	62	0.7152
5	14.3	56	0.1621
6	14.9	66	0.5027
7	8.6	65	0.5716
8	13.8	59	0.4443
9	20.1	61	0.2036
10	8.6	69	0.4735
11	6.9	74	0.2374
12	9.7	69	0.5507
13	9.2	66	0.6817
14	10.9	68	0.6821
15	13.2	58	0.3362
16	11.5	64	0.8851
17	12	66	0.8444
18	18.4	57	0.1340
19	11.5	68	0.6533
20	9.7	62	0.5775
21	9.7	59	0.3314
22	16.6	73	0.2377
23	9.7	61	0.4908
24	12	61	0.6072
25	16.6	57	0.1915
26	14.9	58	0.3135
27	8	57	0.1697
28	12	67	0.7423
29	14.9	81	0.0852
30	5.7	79	0.0783
31	7.4	76	0.1776

Table 4.3: Spatial depth of wind and temperature in month of May. Wind speed in miles per hour and temperature in degrees Fahrenheit.

Table 4.3 shows the calculation of the spatial depth of wind and temperature recorded in May using the equation 3.4.2. From Table 4.3, (11.5, 64) and (12, 66) are the most central points with spatial depth values of 0.8851 and 0.8444, respectively, while the (5.7, 79) and (14.9, 81) are the most outlying with the spatial depth value 0.0783 and 0.0852 respectively. In spatial depth analysis, the terms "most central points" and "most outliers" refer to specific characteristics of a spatial dataset. The most central points are the data points closest to the dataset's spatial distribution center. The concept of centrality is determined based on a chosen measure of central tendency, such as the mean, median, or spatial median. These points are considered representative of the central tendency of the dataset and are often used to characterize the data's overall spatial pattern or tendency. The most outliers, on the other hand, are data points that are located far away from the center or the main cluster of the dataset. Outliers are data points that deviate significantly from the general pattern or distribution of the data. Various factors, such as measurement errors, sampling bias, or the presence of rare or unusual phenomena, can cause them. Outliers are often of interest because they may indicate important spatial patterns or anomalies in the dataset. In spatial depth analysis, identifying the most central points and most outliers can provide insights into the overall structure and characteristics of the spatial data, allowing for a better understanding of the underlying patterns and potential anomalies in the dataset.

Applying spatial depth to air quality data can enhance our understanding of pollution levels' spatial patterns, variability, and outliers, facilitating better decision-making in environmental management and public health interventions.

	Ozone	Solar.R	Wind	Temp	Month	Day
4	NaN	NaN	14.3	56	5	5
17	6.0	78.0	18.4	57	5	18
24	NaN	66.0	16.6	57	5	25
26	NaN	NaN	8.0	57	5	27

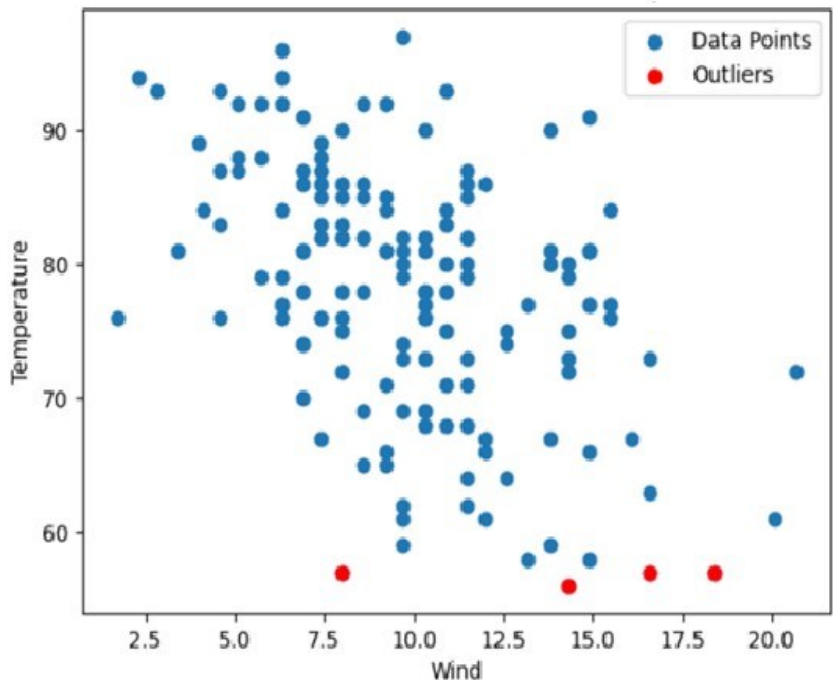

Figure 4.7: Spatial depth of pair points of wind and temperature.
Wind speed in miles per hour and temperature in degrees Fahrenheit.

Using the spatial depth outliers detection method for wind and temperature involves assessing the centrality of data points in the joint distribution of these variables and identifying combinations that deviate significantly from the expected patterns, potentially indicating unique or abnormal conditions.

The resulting plot (Figure 4.7) shows the spatial depth function values for each data point of the column of wind and temperature. The identified outliers are marked with red points on the plot. The Python code calculates the spatial depth function by computing the euclidean distances between each pair of data points' coordinates. Then, it calculates the spatial depths' median and median absolute deviation (MAD) to define a threshold for outlier detection. Data points with spatial depth values above this threshold are considered outliers. Pair points (14.3,56) ;(18.4,57) ;(16.6,57), and (8,57) are identified as outliers. Those points were also identified in Table 4.3 with spatial depth values of 0.1621, 0.1340, 0.1915 and 0.1697, respectively.

Spatial depth functions take into account the spatial distribution and multivariate nature of the data, allowing for the detection of outliers even in the presence of skewed or heavy-tailed

distributions. Air quality data often exhibit spatial autocorrelation, meaning neighbouring locations tend to have similar measurements. By incorporating spatial information, a spatial depth function can capture this context and identify outliers that deviate significantly from their spatial neighbours. It is advantageous in detecting localized pollution events or anomalous air quality readings.

Chapter 5

Concluding Remarks

Applying quantiles and depth functions to outlier detection is essential for several reasons. These techniques provide robust and effective ways to identify and handle outliers in a dataset. Here are some key reasons why they are important: 1. Robustness: Outliers can significantly affect statistical measures such as the mean and standard deviation, leading to biased results. Quantiles and depth functions provide robust estimators less sensitive to outliers, allowing for more accurate and reliable analyses. 2. Non-parametric nature: Quantiles and depth functions are non-parametric methods, meaning they do not rely on specific assumptions about the underlying distribution of the data. This makes them more flexible and applicable to various data types without requiring assumptions about their shape. 3. Distribution-free: Outliers can occur in any distribution, and traditional techniques based on assumptions about normality or other specific distributions may fail to detect them. Quantiles and depth functions provide distribution-free methods to identify outliers regardless of the underlying distribution. 4. Multivariate outlier detection: Quantiles and depth functions can be extended to handle multivariate data, where outliers are defined by their deviation in multiple dimensions. These techniques consider the overall structure of the data and provide a comprehensive approach to detect outliers in high-dimensional datasets. 5. Interpretabil- ity: Quantiles and depth functions clearly interpret outlier detection results. For example, quantiles provide thresholds based on percentiles, allowing you to define outliers as values exceeding a certain threshold. Depth functions provide a measure of centrality, where points with lower depths are considered outliers. This interpretability enhances the understanding and decision-making process in outlier detection. 6. Application in various fields: Outlier detection using quantiles and depth functions has applications in numerous fields, including finance, healthcare, fraud detection, anomaly detection in industrial processes, and more. These techniques help identify unusual and potentially important observations that can significantly impact analysis and decision-making. Ta- ble 5.1 highlights some advantages and limitations of outliers detection based on quantiles and depth functions.

Comparison of methods of outliers detections	
Advantages	**Limitations**
Geometric quantile contour plot	
1.Capturing multivariate outliers: Geometric quantile contour plot takes into account the joint distribution of variables. 2. It provides a visual representation of the data, making it easier to interpret and understand the outliers in the context of the entire dataset. 3. Geometric quantile contour plot is less affected by univariate outliers since it considers the overall distribution of the variables.	1.Computational complexity: Constructing the contour plot and determining the outliers can be computationally expensive, especially for large datasets. 2. Subjectivity in contour selection: Selecting appropriate contour boundaries can be subjective and may require some expertise or trial-and-error.
Univariate quantiles	
1.Simplicity: Univariate quantile-based methods are relatively simple to implement and interpret, as they focus on the distribution of each variable separately. 2. Computational efficiency: Compared to the multivariate approach, univariate methods are generally computationally more efficient, making them suitable for large datasets.	1. Limited capturing of multivariate outliers: Univariate methods may miss outliers that are only detectable when considering the relationships between variables. 2. Univariate methods can be influenced by extreme values in a single variable. 3.Lack of context: Univariate methods do not consider the joint behavior of variables, which may overlook outliers that are not apparent in individual variable distributions.
Mahalanobis depth	
1. Mahalanobis depth incorporates information about the covariance structure, offering a quantitative measure of outlyingness. 2. Provides a quantitative measure of outlyingness, allowing for ranking and prioritization of outliers	1. Assumes a multivariate normal distribution, which may not hold in all datasets. 2. Sensitive to outliers themselves, as they can impact the estimation of the covariance
Halfspace depth	
1. Halfspace depth is known for its robustness in high-dimensional datasets, making it useful for outlier detection in complex data structures. 2. No subjective threshold: Unlike GQCP, halfspace depth does not require setting a specific threshold, making it less subjective and more consistent.	1. Computational complexity: Computing the halfspace depth for each data point can be computationally expensive, especially in large datasets or datasets with a high number of dimensions. 2.Sensitivity to data distribution: Halfspace depth may be sensitive to the underlying distribution of the data, and its performance may vary depending on the shape and structure of the dataset.
Spatial depth	
1. Robust against certain types of outliers. 2. Less sensitive to the dimensionality of the data.	1. Assumes that outliers are distant from the center of the data distribution, which may not always hold. 2. Limited effectiveness in cases where outliers are concentrated in specific regions.

Table 5.1: Advantages and limitations of using quantiles and depth functions to outliers detections.

In summary, applying quantiles and depth functions to outlier detection is vital due to their robustness, non-parametric nature, distribution-free approach, ability to handle multivariate data, interpretability, and broad applicability across different domains. By comparing all mentioned methods in the Table 5.1, geometric quantiles provide a robust and interpretable approach for identifying outliers in various data analysis scenarios. Their application can enhance the understanding of data distributions, aid in anomaly detection, and inform decision-making processes. Applying geometric quantile contour plots to air quality data offers a powerful tool for understanding air pollutants' distribution, spatial patterns, and temporal changes. It assists in identifying pollution hotspots, evaluating compliance with regulations, assessing exposure levels, monitoring trends, and making informed decisions to improve air quality and safeguard public health. However, the interpretation and treatment of outliers should always be done with caution, considering the spe- cific context and objectives of the analysis. Using spatial depth for outliers detection can also be highly beneficial in various data analysis tasks. Spatial depth refers to a concept that measures the centrality or outlyingness of a point within a dataset in a multivariate space. It provides a robustmeasure of how typical or atypical a data point is in relation to other points.

Method used	**Variable(s)**	**Confidence levels**	**Outliers found**	**Reference figure and (or)table**
Univariate quantile	Temperature	Q1-1.5*IQR And Q3+1.5*IQR	56,57,57,57	Figure 4.1
Geometric quantile	Temperature	90^{th} percentile	56,94,96	Figure 4.2
Geometric quantile	Temperature and wind	90^{th} percentile	(1.7,76);(16.6,57); (18.4,57);(20.1,61); (20.7,72);(2.3,94)	Figure 4.4
Halfspace depth	Temperature and wind	cutoff value of 10% (here we used a 90^{th} percentile)	(14.3,56);(20.1,61); (18.4,57);(16.6,57)	Figure 4.5
Mahalanobis depth	Temperature and wind	Threshold: based on the chi-squared distribution with a 90% confidence level (alpha = 0.10).	(1.7,76);(8,57); (18.4,57);(20.1,61); (20.7,72);(14.9,91)	Figure 4.6
Spatial depth	Temperature and wind	Threshold: Taking alpha= 10% (90% confidence level)	(8,57);(14.3,56); (16.6,57);(18.4,57)	Figure 4.7 and Table 4.3

Table 5.2: Results summary from used outliers detection methods .

In this research, we use spatial depth to analyze the air-quality data. Outliers in an air quality dataset can indicate exceptional events or measurement errors. By applying a spatial depth function, which measures the centrality or outlyingness of a data point concerning the entire dataset, it becomes possible to identify and flag potential outliers. This is crucial for ensuring the accuracy and reliability of the dataset. In Table 4.3, we highlighted some most central and most outlying points of wind and temperature values. Table 5.2 compares five different outliers detection methods based on the outputs found. They almost detected a similar outlier point but slightly different based on confidence levels and limitations. In summary, Geometric quantile contour plot-based outlier detection is advantageous when capturing multivariate outliers and understanding the relationships between variables is crucial. Univariate quantile-based methods are more straightforward and computationally efficient, but they may overlook certain outliers and are more vulnerable to the influence of extreme values in individual variables.

From tables 5.1 and 5.2, geometric quantiles are less affected by extreme values and can provide a more accurate representation of the underlying data distribution. In this thesis, we used geometric quantile contour plots to analyze the air-quality data. By representing different quantiles with contour lines, it becomes easier to identify areas with high or low pollutant concentrations(see Figures 4.2,4.3 and 4.4). This visual representation helps identify pollution hotspots, regions with potential health risks, or areas requiring specific attention for mitigation measures. Geometric quantiles can also be extended to multivariate datasets, allowing the identification of outliers based on their distance from the geometric center in a higher-dimensional space. This enables the detection of outliers that exhibit unusual behaviour across multiple variables simultaneously. Air quality standards and regulations often define acceptable limits for pollutants. Geometric quantile contour plots can help evaluate compliance with these standards by indicating areas where pollutant concentrations exceed the acceptable thresholds. This information is crucial for environmental agencies, policymakers, and researchers to identify areas that require immediate action to improve air quality.

Applying depth functions to outlier detection is also crucial because depth functions provide a robust and effective way to measure the outlyingness of data points in a dataset. Depth functions assign a score or a depth value to each data point based on its relative centrality within the dataset. Depth functions consider the entire dataset's context when assessing the outlyingness of a data point. They can handle multivariate data, where multiple attributes or features represent each data point. By considering the overall distribution and relationship among the attributes, depth functions can provide a more comprehensive measure of outlyingness than univariate methods.

Spatial depth was proven to be a proper depth function that detects outliers by considering the distribution of data points as a whole. It is less sensitive to individual extreme values or outliers compared to traditional measures such as distance-based methods (e.g., euclidean distance). This

robustness allows for more reliable identification of outliers, even in the presence of noisy or contaminated data. Spatial depth takes into account the multivariate nature of the data. It considers the combined relationships among variables, allowing for a comprehensive assessment of outlyingness. Spatial depth provides a geometric interpretation of outlyingness. It measures how deep a data point lies within the data cloud or distribution. Outliers can also arise due to instrument malfunctions or calibration issues. Applying a spatial depth function to air quality data makes it possible to identify potential instrument anomalies that might be affecting the measurements. This information can guide validation processes, help calibrate instruments, and ensure data accuracy and reliability. Spatial depth enables the ranking of outliers based on their outlyingness scores. This ranking provides a helpful tool for prioritizing and focusing on the most extreme or influen-tial outliers. By identifying the most significant outliers, resources can be efficiently allocated to address or investigate these data points. Based on scores in Table 3.2, the pairs (0, 20) and (0, -20) are the most outlying points, and from Table 4.3, the pairs (5.7, 79) and (14.9, 81) are the most outlying points. Spatial depth has been successfully applied in diverse fields, such as finance, environmental monitoring, medical diagnostics, and image analysis. Its effectiveness in different domains highlights its versatility and practicality for outlier detection tasks.

Applying a spatial depth function to an air quality dataset for outlier detection improves data quality, enhances spatial contextualization, supports decision-making, and assists in data validation and calibration. These benefits contribute to a more accurate understanding of air quality and facilitate effective environmental management strategies.

Bibliography

[1] Jason Abrevaya. The effects of demographics and maternal behavior on the distribution of birth outcomes. *Empirical Economics*, 26:247–257, 2001.

[2] Alan Agresti and Barbara Finlay. Statistical methods for the social sciences. *Pearson*, 4th ed, 2009.

[3] Cadre Ali, Gannoun / Benoît. Asymptotic normality of consistent estimate of theconditional l1-median. *Annales de l'ISUP*, pages 13–33, 2000.

[4] Peter C. Austin, Jack V. Tu, Paul A. Daly, and David A. Alter. The use of quantile regression in health care research: a case study examining gender differences in the timeliness of thrombolytic therapy. *Statistics in Medicine*, 24(5):791–816, 2005.

[5] Vic Barnett. The ordering of multivariate data. *Journal of the Royal Statistical Society. Series A (General)*, 139(3):318–355, 1976.

[6] Gilbert W. Bassett and Hsiu-Lang Chen. Portfolio style: Return-based attribution using quan- tile regression. *Empirical Economics*, 26:293–305, 2001.

[7] Basil Montgomery Brown. Statistical uses of the spatial median. *Royal Statistical Society*, 45(1):25–30, 1983.

[8] Biman Chakraborty. On affine equivariant multivariate quantiles. *Annals of the Institute of Statistical Mathematics*, 53(2):380–403, 2001.

[9] Goga Chaouch, Mohamed /Camelia. Design-based estimation for geometric quantiles with application to outlier detection. *Computational Statistics and Data Analysis (CSDA)*, 54(10): 2214–2229, October 2010.

[10] Mohamed Chaouch, Ali Gannoun, and Jérôme Saracco. Conditional spatial quantile: Characterization andnonparametric estimation. *GREThA*, 2008.

[11] Probal Chaudhuri. On a geometric notion of quantiles for multivariate data. *Journal of the American Statistical Association*, 91(434):862–872, 1996.

[12] Victor Chernozhukov. Nonparametric extreme regression quantiles. Working paper, Stanford Univ. Presented at Princeton Econometrics Seminar, December 1998.

[13] Victor Chernozhukov. Conditional extremes and near-extremes. MIT Dept. of Economics Working Paper No. 01-21, June 2001.

[14] Yahya Almardeny Noureddine BoujnahFrances Cleary. A novel outlier detection method for multivariate data. *IEEE Transactions on Knowledge and Data Engineering*, 39(9):4052– 4062, 2022.

[15] William F. Eddy. Ordering of multivariate data. computer science and statistics. *The Interface, ed. L. Billard, Amesterdam: North-Holland*, pages 25–30, 1985.

[16] Soo-Heang Eo, Seung-Mo Hong, and HyungJun Cho. Identification of outlying observations with quantile regression for censored data. *arXiv:1404.7710v1 [stat.CO]*, 2014.

[17] Stéphane Girard and Gilles Stupfler. Extreme geometric quantiles in a multivariate regular variation framework. *Extremes : Statistical Theory and Applications in Science, Engineering and Economics*, 18(4):629–663, 2015.

[18] Stéphane Girard and Gilles Stupfler. Intriguing properties of extreme geometricquantiles. *REVSTAT – Statistical Journal*, 15(1):107–139, 2017.

[19] Ramanathan Gnanadesikan and Jon R. Kettenring. Robust estimates, residuals, and outlier detection with multiresponse data. *Biometrics*, page 81–124, 1972.

[20] Yuri Goegebeur Jan Beirlant, Tertius De Wet. Nonparametric estimation of extreme conditional quantiles. *Journal of Statistical Computation and Simulation*, 74(8):567–580, 2004.

[21] Dawit Zerom Jan DeGooijer, Ali Gannoun. A multivariate quantile predictor. *Communications in Statistics: Theory and Methods*, 25(1):133–147, 2006.

[22] Johannes Henricus Bernardus Kemperman. The median of a finite measure on a banach space. *Statistical Data Analysis Based on the Li Norm and Related Methods*, 1987.

[23] Roger Koenker and Gilbert Bassett. Regression quantiles. *Econometrica*, 46(1):33–50, 1978.

[24] V. I. Koltchinskii. M-estimation, convexity and quantiles. *Annals of Statistics*, 25(2):435– 477, 1997.

[25] Yijian Huang Limin Peng. Survival analysis with quantile regression models. *Journal of the American Statistical Association*, 103(482):637–649, 2008.

[26] Prasanta Chandra Mahalanobis. On the generalized distance in statistics. *Proceedings of the National Institute of Science (India)*, 2(1):49–55, 1936.

[27] Willard G. Manning, Linda Blumberg, and Lawrence H. Moulton. The demand for alcohol: The differential response to price. *Journal of Health Economics*, 14(2):123–148, 1995.

[28] GR Pandey and V-T-V Nguyen. A comparative study of regression based methods in regional flood frequency analysis. *Journal of Hydrology*, 225(1):92–101, 1999.

[29] Lea Petrella, Alessandro Gustavo Laporta, and Luca Merlo. Cross-country assessment of systemic risk in the european stock market: Evidence from a CoVaR analysis. *Springer*, 146 (1):169–186, 2019.

[30] Robin L. Plackett. Comment on ordering of multivariate data by v. barnett. *Journal of the Royal statistical Society, Ser.A*, 139:303–325, 1976.

[31] James L. Powell. Least absolute deviations estimation for the censored regression model. *Journal of Econometrics*, 25:303–325, 1984.

[32] Rolf-Dieter Reiss. *Approximate Distributions of Order StatisticsWith Applications to Non-parametric Statistics*. New York : Springer-Verlag, 1989.

[33] David W Scott. *Multivariate Density Estimation. Theory, Practice, and Visualization*. Wiley Series in Probability and Statistics, 1992. URL book.

[34] Seunghyoung RyuJiyeon YimJunghoon SeoYonggyun YuHogeon Seo. Quantile autoencoder with abnormality accumulation for anomaly detection of multivariate sensor data. *IEEE Access*, 10:70428–70439, 2022.

[35] Robert Serfling. Quantile functions for multivariate analysis: approaches and applications. *Statistica Neerlandica*, 56(2):214–232, 2002.

[36] John Wilder Tukey. Mathematics and picturing data. *In Proceedings of the International Congress on Mathematics (R. D. James, ed.)*, 2:523–531, 1975.

[37] Cheng / Jan G. De Gooijer Yebin. On the uth geometric conditional quantile. *The Journal of Statistical Planning and Inference*, 137(6):1914–1930, June 2007.

[38] Yijun Zuo and Robert Serfling. General notions of statistical depth function. *The Annals of Statistics*, 28(2):461–482, 2000.