OUTLIERS DETECTION OF TEMPORAL GENE
EXPRESSIONS UNDER MULTIPLE CONDITIONS

A Thesis

Submitted to the Faculty of Graduate Studies and Research

In Partial Fulfillment of the Requirements

for the Degree of

Master of Science

In

Statistics

University of Regina

By

Yingjie Wei

Regina, Saskatchewan

December 2018

© Copyright 2018: Yingjie Wei
UNIVERSITY OF REGINA

FACULTY OF GRADUATE STUDIES AND RESEARCH

SUPERVISORY AND EXAMINING COMMITTEE

Yingjie Wei, candidate for the degree of Master of Science in Statistics, has presented a thesis titled, *Outliers Detection of Temporal Gene Expressions Under Multiple Conditions*, in an oral examination held on December 12, 2018. The following committee members have found the thesis acceptable in form and content, and that the candidate demonstrated satisfactory knowledge of the subject material.

External Examiner: Dr. Lisa Fan, Department of Computer Science

Supervisor: Dr. Dianliang Deng, Department of Mathematics & Statistics

Committee Member: Dr. Andrei Volodin, Department of Mathematics & Statistics

Chair of Defense: Dr. Zhanle Wang, Faculty of Engineering & Applied Science
Abstract

Temporal gene expression data have been studied and applied in biological, biomedical studies and early cancer detection. A set of temporal gene expression data in bacteria shows that the gene expression has different patterns under different biological conditions. The datasets are treated as functional data in this study, and the goal of my search is to detect outliers based on functional data theory. Then, we can identify the biological conditions that produce outliers and provide valuable information for biologists to find treatment to the bacteria.

The datasets with 21 genes in P. Aeruginosa expressed in 24 biological conditions have been studied in this thesis. The aim of my research is to find the influence of the biological conditions to each gene as we can visually find a few curves are significantly different from others, but they are supposed to have the same distribution as the rest curves. Therefore, we detect the conditions that produce outliers in each gene. We apply four functional depth notions, the Fraiman and Muniz depth, the h-modal depth, the random projection depth, and the random Tukey depth.
The simulation experiments to the outlier detection procedures are conducted accordingly and it performs well. Then, we apply the outlier detection procedures to the 21 genes datasets. In terms of the performance of each depth, the outlier detection results resemble the simulation results. We have identified the conditions that produce outliers of gene trajectories, and the results agree well with the classification of biological conditions result based on minimum Mahalanobis distance.
Acknowledgements

The deepest gratitude goes to my supervisor Dr. Dianliang Deng, whose knowledge, support, and encouragement played the most important role in my study and research.

I am grateful to the department of Mathematics and Statistics for offering me University Teaching Fellowship, teaching assistant opportunities, and providing me with scholarship and awards. Special thanks go to the Government of Saskatchewan and Faculty of Graduate Studies and Research for offering me the Saskatchewan Innovation and Excellence Graduate Scholarship. Thanks to all professors who delivered brilliant lectures. The same thanks go to administrative staff for their excellent support.

Special thanks will be given to Laboratory Instructor Sarah Carnochan Naqvi and Dr. Taehan Bae for providing me extra learning resources and opportunity.

Yingjie Wei, October 2018

Regina, Saskatchewan
Dedication

To my deceased father, my dearest mother and my family.
Contents

Abstract i

Acknowledgements iii

Dedication iv

Table of Contents v

List of Figures viii

List of Tables ix

1 Introduction 1

2 Preliminary Studies 8

3 Functional Data Theory 16
   3.1 Fundamental Knowledge of Hilbert Space 17
   3.2 Functional Depth Measures 21
3.2.1 The Fraiman and Muniz Depth (FMD) .......................... 23
3.2.2 The h-Modal Depth (MD) .................................... 25
3.2.3 The Random Projection Depth (RPD) ......................... 27
3.2.4 The Random Tukey Depth (RTD) ............................ 28
3.3 Trimmed Mean for Functional Data .............................. 29

4 Functional Outlier Detection Procedures .......................... 31

5 Simulation Results .................................................. 35

  5.1 Type I Error ..................................................... 36
  5.2 The Power Analysis of Outlier Detection Procedures ......... 37

6 Outlier Detection of Temporal Gene Datasets ..................... 42

  6.1 Main Results and Conclusions .................................. 43
  6.2 Questions and Discussions ...................................... 51

7 Summary and Future Research ..................................... 56

Appendix A R code example for outlier detection: gene E6 58

Appendix B R code example for type I error test: h-Modal Depth
when sample size \( n = 200 \) 64

Appendix C R code example for power test of contaminated sample:
h-Modal Depth when sample size $n = 200$ and $n_0 = 3$
### List of Figures

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.1</td>
<td>24 trajectories of gene A4</td>
<td>5</td>
</tr>
<tr>
<td>1.2</td>
<td>24 trajectories of gene A6</td>
<td>6</td>
</tr>
<tr>
<td>6.1</td>
<td>The 0.25-trimmed mean and the sample median of Gene A4</td>
<td>45</td>
</tr>
<tr>
<td>6.2</td>
<td>The 0.25-trimmed mean and the sample median of Gene A4</td>
<td>45</td>
</tr>
<tr>
<td>6.3</td>
<td>The 0.25-trimmed mean and the sample median of Gene A4</td>
<td>46</td>
</tr>
<tr>
<td>6.4</td>
<td>The 0.25-trimmed mean and the sample median of Gene A4</td>
<td>46</td>
</tr>
<tr>
<td>6.5</td>
<td>24 trajectories of gene E5</td>
<td>52</td>
</tr>
<tr>
<td>6.6</td>
<td>24 trajectories of gene D2</td>
<td>53</td>
</tr>
<tr>
<td>6.7</td>
<td>24 trajectories of gene F3</td>
<td>54</td>
</tr>
<tr>
<td>6.8</td>
<td>24 trajectories of gene G6</td>
<td>55</td>
</tr>
</tbody>
</table>
List of Tables

1.1 18 genes in P. aeruginosa expression ........................................... 7

5.1 Percentage of type I errors with outliers free datasets for sample size
n=100 and n=200 ................................................................. 37

5.2 Correct frequency (freq.) detection, false outliers rate of the applied
procedure for detecting functional outliers, and the frequency of no
outlier detection, where FMD, MD, RPD and RTD denote theFraiman
and Muniz depth, the h-mode depth, the random projection depth and
the Random Tukey depth respectively, and T stands for Trimming and
W stands for Weighting ............................................................ 41

6.1 Outliers detection results for 21 genes under 24 conditions .......... 47

6.2 Outlier detection Details of genes A4 ................................. 47

6.3 Outlier detection Details of genes A6 ................................. 48

6.4 Outlier detection Details of genes B4 ................................. 48
6.5 Outlier detection Details of genes B5 . . . . . . . . . . . . . . . . . . . 48
6.6 Outlier detection Details of genes C4 . . . . . . . . . . . . . . . . . . . 48
6.7 Outlier detection Details of genes D1 . . . . . . . . . . . . . . . . . . . 49
6.8 Outlier detection Details of genes E6 . . . . . . . . . . . . . . . . . . . 49
6.9 Outlier detection Details of genes F2 . . . . . . . . . . . . . . . . . . . 49
6.10 Outlier detection Details of genes G2 . . . . . . . . . . . . . . . . . . . 49
6.11 Outlier detection Details of genes G5 . . . . . . . . . . . . . . . . . . . 50
6.12 Outlier detection Details of genes H3 . . . . . . . . . . . . . . . . . . . 50
6.13 Outlier detection Details of genes H4 . . . . . . . . . . . . . . . . . . . 50
6.14 Outlier detection Details of genes S70 . . . . . . . . . . . . . . . . . . . 50
Chapter 1

Introduction

Outlier is the observation that stands out in contrast to other observations, as an extreme value, in some practical situations. It has been studied over a century by statisticians. The analysis of outliers is an important part of statistical analysis of data. Barnett and Lewis in [2] have introduced the history and outlier theory in basic principles, univariate data, multivariate and structured data, Bayesian approaches, time series and other special cases. It is natural to seek ways of interpreting or categorizing outliers, and identify the sources which have produced outliers. Therefore, three aims are associated with examining outliers: identification, rejection, and accommodation. Then we should be able to find methods for handling them - identifying them, rejecting them, or including them. This thesis aims to identify them and provide important information for scientific research.

Over the last two decades, functional data analysis (FDA) has established itself as
an important and dynamic area of statistics. The book [28] written by Ramsay and Silverman in 1997 has greatly contributed to popularize the FDA techniques among the users, offering a number of appealing case studies and practical methodologies. A further book [29] written by the same authors in 2002 is devoted to the applied aspects of FDA, with examples in growth analysis, meteorology, physiology, economics, and medicine, etc. The book [22] authored by Lajos Horváth and Piotr Kokoszka in 2012 focuses on the construction of test statistics and the relevant asymptotic theory, with emphasis on models for dependent functional data.

FDA is concerned with observations which are viewed as functions defined over a set $T$. Some functional data belong to the class of high dimensional data in the sense that every data object consists of a large number of scalar values, and the number of measurements per object may be larger than the sample size $N$. It offers effective new tools and has stimulated new methodological and theoretical developments. The field has become very broad, with many specialized directions of research.

In one dimension, order statistics and ranks have been applied to measure the centrality and outliers. $L$-estimates, which are defined as linear combinations of order statistics, are a well known class of robust location estimates. In particular, trimmed means, which are defined as the average of the most central $(1 - \alpha)n$ observations, $(0 \leq \alpha \leq 1)$ constitute a class of estimates that range from the sample mean to the sample median.
In more than one dimension, the concepts of order statistics and ranks are more involved and several definitions have been proposed: halfspace depth (or the Tukey depth), Tukey [34], and Brown [3]; Oja’s depth, Oja [27]; the simplicial depth, Liu [24], [25], Small1 [32], Gordaliza [20], Singh [31], Donoho and Gasko [12], Liu and Singh [26], Cuesta-Albertos, Gordaliza and Matrán [4], and Fraiman and Meloche [18]; the projection depth, Zuo [37] and Cuevas [9]; the random Tukey depth, Cuesta-Albertos [6]. All of them are based on different notions of depth. A data depth is a device introduced to measure the “centrality” of a multivariate data point within a given data cloud. Although these definitions are quite different for multivariate data, they are very similar when we look at them for univariate data.

The notion of depth was extended to functional data by Fraiman and Muniz [19]. The concept of depth of functional data offers a possible framework for identifying central and outlying observations: central observations which have maximal depth and potential outlier which have minimal depth. We will consider a set of curves, $x_1, \cdots, x_n$, generated from a stochastic process $X(\cdot)$ with sample paths in $C([a,b])$, which is the space of continuous functions defined on the interval $[a,b] \subseteq \mathbb{R}$.

Gene expression is the process by which information from a gene is used in the synthesis of a functional gene product. The process of gene expression is used by all known life, for example, it is utilized by viruses to generate the macromolecular machinery for life. Therefore, gene expressions have been studied and applied to
biological and biomedical studies as they contain ample information for research, such as, the development of technology of early stage cancer detection.

We are going to study the datasets containing 21 genes in P. Aeruginosa expressed in 24 conditions. Eighteen genes have been studied by Fang et al. [15]. Table 1.1 shows the detailed information of these 18 genes. Each gene was measured every half hour for 21 hours under each condition, therefore, 43 observations were obtained for each gene. The data are repeated measurements of the same subject densely over an ordered grid of points belonging to an interval of finite length 0.5 hour. Thus, for each gene, we observe a function, though the recording points are discrete, we may still consider the entire function as continuous based on the reality. It falls in the functional data analysis (FDA) research field. For example, Figure 1.1 and Figure 1.2 are 24 trajectories of gene A4 and A6 respectively under 24 different conditions within 21 hours. Each curve represents the expression of gene A4 and A6 respectively under each condition accordingly. Visually, the gene expressions demonstrate that a few trajectories are different from the others in shape. For example, the trajectory of gene A4 under condition C21 is apparently different with others, and it happens to the trajectory of gene A6 under condition C16. In functional data setting, a rigorous definition of outlier has not been given. In this thesis, we consider that a curve is an outlier if it has been generated by a stochastic process with a different distribution than the rest of curves, which are assumed to be identically distributed. Based on
this notion, in [16], M. Febrero et al. detected abnormal $NO_x$ levels using functional data by depth measures. Inspired by [16], we will detect outlier conditions in the datasets using four different functional depths, the Fraiman and Muniz depth, the h-modal depth, the random projection depth and the random Tukey depth. The first two depths have been applied in [16] to detect the outliers of $NO_x$ pollution. It is important to identify the biological conditions which are the sources of the outliers. Therefore, the biologists would be able to find ways to control the bacteria through the biological conditions.

Figure 1.1: 24 trajectories of gene A4
Figure 1.2: 24 trajectories of gene A6
<table>
<thead>
<tr>
<th>Code</th>
<th>Name</th>
<th>Protein</th>
<th>Ratio</th>
<th>Remarks</th>
</tr>
</thead>
<tbody>
<tr>
<td>A6</td>
<td>PA5283</td>
<td>Probable transcriptional regulator</td>
<td>99.68%</td>
<td>48% similar to putative transcriptional regulator (Bacillus subtilis)</td>
</tr>
<tr>
<td>B3</td>
<td>PA2975 (rluC)</td>
<td>Ribosomal large subunit pseudouridine synthase C</td>
<td>99.68%</td>
<td>Transcription, RNA processing &amp; degradation</td>
</tr>
<tr>
<td>B4</td>
<td>PA4991</td>
<td>Hypothetical protein</td>
<td>100%</td>
<td>Unknown</td>
</tr>
<tr>
<td>B5</td>
<td>PA5237</td>
<td>Conserved hypothetical protein</td>
<td>100%</td>
<td>87% similar to hypothetical yigC gene product of E.coli</td>
</tr>
<tr>
<td>C4</td>
<td>PA0287 (gpuP)</td>
<td>3-guanidinopropionate transport protein</td>
<td>100%</td>
<td>Transport of small molecules</td>
</tr>
<tr>
<td>D1</td>
<td>PA3115 (finV)</td>
<td>Motility protein FimV</td>
<td>100%</td>
<td>Membrane proteins; Motility &amp; Attachment</td>
</tr>
<tr>
<td>D2</td>
<td>PA3879 (narL)</td>
<td>Two-component response regulator NarL</td>
<td>99.67%</td>
<td>74% similar to E.coli NarL protein</td>
</tr>
<tr>
<td>D3</td>
<td>PA0894</td>
<td>Hypothetical protein</td>
<td>99.02%</td>
<td>Unknown</td>
</tr>
<tr>
<td>E5</td>
<td>PA1875</td>
<td>Probable outer membrane protein precursor</td>
<td>100%</td>
<td>41% similar to alkaline protease secretion protein AprF</td>
</tr>
<tr>
<td>E6</td>
<td>PA0573</td>
<td>Hypothetical protein</td>
<td>100%</td>
<td>Unknown</td>
</tr>
<tr>
<td>F2</td>
<td>PA3902</td>
<td>Hypothetical protein</td>
<td>100%</td>
<td>65% similar to putative amino acid abc transporter, ATP-binding protein (Helicobacter pylori J99)</td>
</tr>
<tr>
<td>F3</td>
<td>PA3212</td>
<td>Probable ATP-binding component of ABC transporter</td>
<td>100%</td>
<td>Energy metabolism</td>
</tr>
<tr>
<td>F5</td>
<td>PA2997 (nqrC)</td>
<td>Na+ translocating NADH:ubiquinone oxidoreductase subunit Nqr3</td>
<td>100%</td>
<td>Energy metabolism; Biosynthesis of co-factors, prosthetic groups &amp; carriers; Amino acid biosynthesis &amp; metabolism</td>
</tr>
<tr>
<td>G2</td>
<td>PA0649 (trpG)</td>
<td>Anthranilate synthase component II</td>
<td>100%</td>
<td>61% similar to putative enoyl-coA hydratase EchA3 (Mycobacterium tuberculosis)</td>
</tr>
<tr>
<td>G5</td>
<td>PA1748</td>
<td>Probable enoyl-CoA hydratase/isomerase</td>
<td>98.2%</td>
<td>54% similar to a region of putative regulatory protein (Streptomyces coelicolor)</td>
</tr>
<tr>
<td>G6</td>
<td>PA3771</td>
<td>Probable transcriptional regulator</td>
<td>99.22%</td>
<td>43% similar to hypothetical yeaK gene product of (E.coli)</td>
</tr>
<tr>
<td>H3</td>
<td>PA1841</td>
<td>Hypothetical protein</td>
<td>100%</td>
<td>As a control</td>
</tr>
</tbody>
</table>
Chapter 2

Preliminary Studies

Temporal gene expression data are of particular interest to researchers as it contains rich information in characterization of gene function. Many statistical techniques have been developed to interpret gene expression. For gene classification, various clustering procedures (Draghici et al. [13]), fold changes (Eisen et al. [14]), and ANOVA (Li et al. [23]), which are mostly related to the classical and parametric statistical methods, are widely used. However, when the number of observed values for each time point is less than the number of observed time points, the classical classification methods will not perform well and cannot be applied to analyze the gene expression data. Therefore, the methods of reduction of dimensionality for the vector of function of observed measurements were developed, for example, principal component analysis (PCA) (Anderson [1], Yeung and Ruzzo [35]), kernel principal component analysis (Kernel PCA) (Schölkopf and Müller [30]) and smooth spline
The datasets with 21 genes in P. Aeruginosa expressed in 24 conditions have been studied from different perspective using different statistical research methods. In order to analyze the datasets from a different point of view, we briefly review some of the main results, such as, the gene classification, the influence of conditions to the gene expressions, and the detection of threshold points for gene expressions under multiple biological conditions.

Fang et al. [15] proposed a non-linear regression model to characterize the relative change-rates of genes. Each individual expression trajectory is modeled as longitudinal data with changeable and covariance structure. For a given gene, let \( g_{j1}, g_{j2}, \ldots, g_{j(p_j+1)} \) be \( p_j + 1 \) observations under the \( j \)th condition, and \( t_j = (t_{j1}, t_{j2}, \ldots, t_{j(p_j+1)}) \) the corresponding set of times at which these measurements are made for \( j = 1, 2, \ldots, n \). The relative change-rate \( y_{jk} \) for \( g_{jk} \) at \( t_{jk} = (t_{j(k+1)} + t_{jk})/2 \), \( k = 1, 2, \ldots, p_j \) is

\[
y_{jk} = \frac{g_{j(k+1)} - g_{jk}}{t_{j(k+1)} - t_{jk}},
\]

where the observation interval \( |t_{j(k+1)} - t_{jk}| \) is small, so the relative change-rate \( y_{jk} \) is considered as the approximate derivative of \( g_{jk} \) at \( t_{jk} \).

Let \( Y_j(t) \) be the value of the \( j \)th curve at time \( t \) and write

\[
Y_j(t) = \mu(t) + \epsilon_j(t),
\]

where \( \mu(t) \) is the mean curve and \( \epsilon_j(t) \) is the random noise with zero mean and
variance $\sigma^2(t)$. Then a set of smoothing basis functions are used to represent the
derivative mean curve $\mu(t)$. Given a sequence of $K$ interior knots $0 < \tau_1 < \tau_2 < \cdots < 
\tau_K < \tau$, where $\tau$ is the end time of observations, the regression spline basis functions
of order $q$ are $1, t, t^2, \cdots, t^q, (t - \tau_1)^q, \cdots, (t - \tau_K)^q$. The vector of $r (= K + 1 + q)$
basis functions is denoted by

$$ B(t) = (1, t, t^2, \cdots, t^q, (t - \tau_1)^q, \cdots, (t - \tau_K)^q)^T. $$

Then $\mu(t)$ is smoothed by the regression spline, which is the linear combination of
the basis functions, and $Y_j(t)$ can be written as

$$ Y_j(t) = B(t)^T \beta + \epsilon_j(t), $$

where $\beta = (\beta_1, \cdots, \beta_r)^T$ is the basis coefficient vector. Hence the function $B(t)^T \beta$ is
a piecewise polynomial. Assumed that the correlation coefficient of two expressions
at times $t$ and $s$ is $\rho^{|t-s|}$, where $|\rho| < 1$, the variance and covariance functions will be:

$$ \sigma^2(t) = \sigma^2 \exp \{ \alpha \mu(t) \}, $$

and

$$ G(t, s) = \sigma(t)\sigma(s)\rho^{|t-s|}. $$

Therefore, the $y_j$ can be considered as realizations of mutually independent normal
random $Y_j$ at the times $t_j = (t_{j1}, t_{j2}, \cdots, t_{jp})$ with

$$ Y_j \sim MVN(\mu_j, V_j), $$

10
where $\mu_j = (\mu(t_{j1}), \ldots, \mu(t_{jp}))^T = B_j \beta$ with $B_j = (B(t_{j1}, \cdots, B(t_{jp}))^T$ and $V_j = V_j(\alpha, \rho)$ is the covariance matrix of $Y_j$ with the following form:

$$V_j = \begin{pmatrix}
\sigma^2(t_{j1}) & \sigma(t_{j1})\sigma(t_{j2})\rho|^{t_{j1}-t_{j2}|} & \cdots & \sigma(t_{j1})\sigma(t_{jp})\rho|^{t_{j1}-t_{jp}}| \\
\sigma(t_{j2})\sigma(t_{j1})\rho|^{t_{j2}-t_{j1}|} & \sigma^2(t_{j2}) & \cdots & \sigma(t_{j2})\sigma(t_{jp})\rho|^{t_{j2}-t_{jp}}| \\
\vdots & \vdots & \ddots & \vdots \\
\sigma(t_{jp})\sigma(t_{j1})\rho|^{t_{jp}-t_{j1}|} & \sigma(t_{jp})\sigma(t_{j2})\rho|^{t_{jp}-t_{j2}|} & \cdots & \sigma^2(t_{jp})
\end{pmatrix}$$

where $\sigma^2(t_{jk}) = \sigma^2 \exp\{\alpha \mu(t_{jk})\}, \ k = 1, 2, \cdots, p_j$.

Then the maximum likelihood estimates of parameters $\beta, \alpha, \sigma^2$ and $\rho$ are obtained. Furthermore, the expected information matrix and the covariance matrix of the estimators $(\hat{\beta}, \hat{\rho}, \hat{\sigma}^2, \hat{\alpha})$ are obtained. A chi-square test is proposed to test the equality of gene expression change-rates based on the parameter estimates. Furthermore, the Mahalanobis distance is used for the classification of 18 genes in 24 biological conditions. The 18 genes are divided into three clusters. The estimated relative change-rate function of D2 (two-component response regulator NarL) is significantly different from others. D2 alone is the first cluster. There is no significant difference between the relative change-rates of A6 and D3. A6, D3 and C4 constitute the second cluster. The rest of genes constitute the third cluster. Within the third cluster, Genes E6, F2, F5, G2 and H3 have similar relative change-rate. Gene F3, G5 and G6 have similar relative change-rate.
Deng et al. [10] proposed a non-linear regression model with log-normal distribution to characterize the variance functions of genes under the given conditions, then to investigate the influence of conditions to the gene expressions and to classify the conditions according to the variance analysis of data from temporal gene expression. Their interest is the dispersion function over time for all genes under a given condition and to compare the differences among all dispersion functions under all conditions. The variance function was estimated under each condition, then the classification was conducted by testing the similarity of two variance functions. An asymptotic test for the equality of variance functions was constructed using Wald statistics and Fisher information matrix. They drew the classification tree for 24 conditions based on minimum Mahalanobis distance. They obtained three clusters and some sub-clusters under the first cluster. The condition C21 alone was the third cluster, conditions C16 and C24 constituted the second cluster, and the rest conditions were classified as the first cluster including some sub clusters. Based on the chi-square test and 0.05 significance level, within the first cluster, they concluded that there is no significance difference among variances of conditions 1, 3, 6, 8, 9, 11, 12, 13, 14, 15, 19 and 22; conditions 4, 7, and 17 have similar variance.

Deng et al. [11] proposed three methods to detect the threshold points for the gene expressions. The algorithms constructed to detect the threshold time points were based on the classical Hotelling statistic, high dimensional test statistic and
empirical distribution based on statistic of sample derivatives for gene expressions. They concluded that B3 and G2 have similar threshold, D3, E6, F2, and G6 have similar threshold, F5, G2, G5 and H3 have similar threshold.

Researchers were trying to find the relationship among these genes or conditions in the previous studies, such as, the classification of the genes and the classification of the biological conditions. This thesis aims to find the influence of the biological conditions to each gene as we can visually find a few curves, which are supposed to have the same distribution as the rest curves, are significantly different from others for some genes. More specifically, we plan to detect the conditions that produce outliers in each gene. Although for each gene under one biological condition, the recording points are discrete, we may consider it as being continuously observed due to the reality. Then, each gene has 24 continuous trajectories corresponding to 24 biological conditions. Therefore, these trajectories can be treated as functional data.

Previous research methods are, more or less, based on the well known Mahalanobis distance. However, the Mahalanobis distance has some flaws since it is not defined if the inverse of the covariance matrix does not exist. Inspired by [9], [16], and [7], we will apply some functional depth notions in this study. In [9], Cuevas et al. considered five notions of data depth, the Fraiman and Muniz depth, the h-modal depth, the random projection depth, and the double random projection depth, then applied these depths to several functional classifiers (minimum, first quartile, median, mean,
third quartile and maximum) by simulations and real data sample. They concluded that the best performance among the depth measures corresponds to the h-modal depth and the best estimator is the trimmed mean (when combined with the h-modal depth) in all considered cases of models and distance measures. In [16], Febrero et al. proposed a method based on three functional depths, the Fraiman and Muniz depth, the h-modal depth and the double random projection depth, for the functional outlier detection. The performance of the proposed procedures were analyzed by simulation experiments, and it performed well with several cases over 90% of correct frequency detection. The h-modal depth and the double random projection depth appeared to have a better performance than the Fraiman and Muniz depth. The trimming bootstrap estimate of the cutoff made the procedure have a larger empirical power but at the cost of having a larger false outlier detection rate than the weighting bootstrap estimate. Then, the method was applied to a real dataset to identify the abnormal $NO_x$ level in a neighborhood of Barcelona. It identified two Fridays and the following Saturdays as outliers of $NO_x$ emission, and these days are the beginning of holidays in Spain. This validates the identification of the $NO_x$ curves on these days as outliers. In [7], Cuesta-Albertos et al. introduced the random Tukey depth, which is considered as a random approximation of the Tukey depth, can be extended to cover Hilbert valued data. The benefit of the random Tukey depth is that it is possible to obtain results similar to those obtained with more involved depths by taking only
a few one-dimensional projections and the computation cost is low. Hence, we will apply this depth to our dataset.

Therefore, The Fraiman and Muniz depth, the h-modal depth, the random projection depth, and the random Tukey depth will be applied in this gene expression data study.
Chapter 3

Functional Data Theory

Statistics is interested in obtaining information from observations \( x_1, x_2, \cdots, x_n \). The \( x_i \) can be scalars, vectors or other objects. Functional Data Analysis (FDA) is concerned with observations which are viewed as functions defined over a set \( T \). Some functional data belong to the class of high dimensional data in the sense that every data object consists of a large number of scalar values, and the number of measurements per object may be larger than the sample size \( n \). The data that presented in this thesis is of the form \( x_i(t), t \in [a, b] \), where \([a, b]\) is an interval on the line. Each observation is thus a curve, and such curves can arise in many ways. The value of a function at every point \( t \) may not be an outlier, but the curve itself may be a functional outlier. Generally speaking, once incorrectly recorded curves have been removed, a curve is an outlier if it comes from a populations with a different distribution in a functional space than the majority of the curves. An outlier may be
far away from the other curves, or may have a different shape. The concept of depth of functional data offers a possible framework for identifying central and outlying observations; those with maximal depth are central, and those with minimal depth are potential outliers.

3.1 Fundamental Knowledge of Hilbert Space

First, we introduce some fundamental concepts of the theory of operators in a Hilbert space, and then introduce the properties of random samples in $L^2$, the space of square integrable functions. The space $L^2$ is sufficient to handle the functional data in this thesis. This section follows [22] closely and references therein.

We denote $H$ as a separable Hilbert space with inner product $\langle \cdot, \cdot \rangle$ which generates the norm $\|\cdot\|$, and denote $\mathcal{L}$ as the space of bounded (continuous) linear operators on $H$ with the norm:

$$\|\Psi\|_\mathcal{L} = \sup \{ \|\Psi(x)\| : \|x\| \leq 1 \}.$$  

An operator $\Psi \in \mathcal{L}$ is said to be compact if there exist two orthonormal bases $\{\upsilon_j\}$, and $\{f_j\}$, and a real sequence $\{\lambda_j\}$ converging to zero, such that

$$\Psi(x) = \sum_{j=1}^{\infty} \lambda_j \langle x, \upsilon_j \rangle f_j, \quad x \in H. \tag{3.1}$$

The $\lambda_j$ may be assumed positive because one can replace $f_j$ by $-f_j$, if needed.

The existence of representation (3.1) is equivalent to the condition: $\Psi$ maps every
bounded set into a compact set. Another equivalent condition is the following: the convergence \( \langle y, x_n \rangle \to \langle y, x \rangle \) for every \( y \in H \) implies that \( \| \Psi(x_n) - \Psi(x) \| \to 0 \).

A compact operator admitting representation (3.1) is said to be a Hilbert-Schmidt operator if \( \sum_{j=1}^{\infty} \lambda_j^2 < \infty \). The space \( S \) of Hilbert-Schmidt operators is a separable Hilbert space with the scalar product

\[
\langle \Psi_1, \Psi_2 \rangle_S = \sum_{j=1}^{\infty} \langle \Psi_1(e_i), \Psi_2(e_i) \rangle,
\]

where \( \{e_i\} \) is an arbitrary orthonormal basis, and the value of (3.2) does not depend on it.

An operator \( \Psi \) is said to be symmetric if \( \langle \Psi(x), y \rangle = \langle x, \Psi(y) \rangle \), \( x, y \in H \), and positive-definite if \( \langle \Psi(x), x \rangle \geq 0 \), \( x \in H \). Sometimes, an operator with the above property is called positive semi-definite, and the term positive-definite is used when \( \langle \Psi(x), x \rangle > 0 \).

A symmetric positive-definite Hilbert-Schmidt operator \( \Psi \) admits the decomposition

\[
\Psi(x) = \sum_{j=1}^{\infty} \lambda_j \langle x, v_j \rangle v_j, \quad x \in H,
\]

with orthonormal \( v_j \) which are the eigenfunctions of \( \Psi \) corresponding to the eigenvalues \( \lambda_j \), i.e. \( \Psi(v_j) = \lambda_j v_j \).

The Space \( L^2 = L^2([0, 1]) \) is the set of measurable real-valued functions \( x \) defined on \([0,1]\) satisfying \( \int_0^1 x^2(t)dt < \infty \). The space \( L^2 \) is a separable Hilbert space with
the inner product
\[ \langle x, y \rangle = \int_0^1 x(t)y(t)dt. \]

If \( x, y \in L^2 \), the equality \( x = y \) means \( \int_0^1 [x(t) - y(t)]^2dt = 0 \).

An integral sign without the limits of integration is meant to denote the integral over the whole interval \([0, 1]\). The integral operators in \( L^2 \) are defined by
\[ \Psi(x)(t) = \int \psi(t, s)x(s)ds, x \in L^2, \]
where \( \psi(\cdot, \cdot) \) is the real kernel. These operators are Hilbert-Schmidt if and only if
\[ \iint \psi^2(t, s)dtds < \infty, \]
in which case
\[ \|\psi\|_2^2 = \iint \psi^2(t, s)dtds. \tag{3.4} \]

If \( \psi(s, t) = \psi(t, s) \) and \( \iint \psi(t, s)x(t)x(s)dtdt \geq 0 \), the integral operator \( \Psi \) is symmetric and positive-definite, and it follows from (3.3) that
\[ \psi(t, s) = \sum_{j=1}^{\infty} \lambda_j \nu_j(t)\nu_j(s), \quad x \in L^2([0, 1] \times [0, 1]). \tag{3.5} \]

If \( \psi \) is continuous, the above expansion holds for all \( s, t \in [0, 1] \), and the series converges uniformly.

We view a random curve \( X = \{X(t), t \in [0, 1]\} \) as a random element in \( L^2 \) equipped with the Borel \( \sigma \)-algebra. \( X \) is integrable, if \( E\|X\| = E[\int X^2(t)dt]^{\frac{1}{2}} < \infty \).

If \( X \) is integrable, there exists a unique function \( \mu \in L^2 \) such that \( E\langle y, X \rangle = \langle y, \mu \rangle \) for
any $y \in L^2$. It follows that $\mu(t) = E[X(t)]$ for almost all $t \in [0, 1]$. The expectation commutes with bounded operators, i.e. $E\Psi(X) = \Psi(EX)$, if $\Psi \in \mathcal{L}$ and $X$ is integrable.

If $X$ is square integrable, i.e.

$$
E\|X\|^2 = E \int X^2(t) dt < \infty,
$$

and $EX = 0$, the covariance operator of $X$ is defined by

$$
C(y) = E[\langle X, y \rangle X], \quad y \in L^2.
$$

Then

$$
C(y)(t) = \int c(t, s)y(s)ds, \quad \text{where} \quad c(t, s) = E[X(t)X(s)].
$$

It is easy to see that $c(t, s) = c(s, t)$, and

$$
\int \int c(t, s)y(t)y(s)dtds = \int \int E[X(t)X(s)]y(t)y(s)dtds
= E \left[ \left( \int X(t)y(t)dt \right)^2 \right] \geq 0.
$$

Therefore, $C$ is symmetric and positive-definite, so it has nonnegative eigenvalues.

However, it does not mean that every symmetric positive-definite operator in $L^2$ is a covariance operator. To explain, let $v_j$, $\lambda_j$, $j \geq 1$, be the eigenfunctions and the eigenvalues of the covariance operator $C$. The relation $C(v_j) = \lambda_jv_j$ implies that

$$
\lambda_j = \langle C(v_j), v_j \rangle = \langle E[\langle X, v_j \rangle X], v_j \rangle = E \left[ \langle X, v_j \rangle^2 \right].
$$
The eigenfunctions $\nu_j$ are orthogonal, and they can be normalized to have unit norm, hence $\{\nu_j\}$ forms a basis in $L^2$. By Parseval’s equality,

$$\sum_{j=1}^{\infty} \lambda_j = \sum_{j=1}^{\infty} E[(X, \nu_j)^2] = E\|X\|^2 < \infty.$$ 

It is easy to prove that $C \in \mathcal{L}(L^2)$ is a covariance operator if and only if it is symmetric positive-definite and its eigenvalues satisfy $\sum_{j=1}^{\infty} \lambda_j < \infty$.

### 3.2 Functional Depth Measures

Depth measures were originally introduced in multivariate data analysis for measuring the centrality of a point $x \in \mathbb{R}^m$, with respect to a data set generated from a probability distribution $F$ in $\mathbb{R}^m$. Therefore, depths provide a way to order in the Euclidean space from the centre to outward. Hence, points near the centre will have higher depth while points should have lower depth if they are far away from the centre.

The depth of a scalar data point can be defined in many ways, see Zuo and Serfling [36]. To illustrate, suppose $x_1, x_2, \cdots, x_n$ are scalar observations, and

$$F_n(x) = \frac{1}{n} \sum_{i=1}^{n} I\{x_i \leq x\},$$ \hspace{1cm} (3.9)

is their empirical distribution function. The Tukey depth or the Halfspace depth of the observation $x_i$ is defined as

$$TD_n(x_i) = HSD_n(x_i) = \min\{F_n(x_i), 1 - F_n(x_i)\}.$$ \hspace{1cm} (3.10)
If \( x_i \) is the median, then \( F_n(x_i) = \frac{1}{2} \), and \( HSD_n(x_i) = \frac{1}{2} \), which is the largest possible depth. If \( x_i \) is the largest point, then \( F_n(x_i) = 1 \), and \( HSD_n(x_i) = 0 \), which is the smallest depth.

The simplicial depth of \( x_i \) with respect to the sample \( x_1, x_2, \ldots, x_n \) is given by

\[
SD_n(x_i) = 2F_n(x_i)(1 - F_n(x_i)).
\]

If \( F_n(x_i) = 1 \), \( SD_n(x_i) = 0 \), then \( x_i \) has the least depth. If \( F_n(x_i) = 1/2 \), \( SD_n(x_i) = 1/2 \), then \( x_i \) has the largest depth.

Another way to define the measuring depth is

\[
D_n(x_i) = 1 - \left| \frac{1}{2} - F_n(x_i) \right|.
\] (3.11)

The largest possible depth will be 1 when \( x_i \) is the median, and the least possible depth is 1/2 when \( x_i \) is the largest point.

Suppose now that we have a sample of functions \( \{x_i(t), t \in [a,b], i = 1, 2, \ldots, n \} \). We define the empirical distribution function at point \( t \) by

\[
F_{n,t}(x) = \frac{1}{n} \sum_{i=1}^{n} I\{x_i(t) \leq x\},
\] (3.12)

where \( I(\cdot) \) is an indicator function. Then, we can define a functional depth by integrating one of the univariate depths. Now, we review four functional depths that will be used for outlier detection in this thesis.
3.2.1 The Fraiman and Muniz Depth (FMD)

Fraiman and Muniz [19] were the first to introduce a functional data depth. They defined a natural notion of depth for functional data. The idea is to measure “how long” a curve remains in the middle of a group of data curves, which are realizations of a stochastic process. Let \( x_1(t), x_2(t), \ldots, x_n(t) \) be independent and identically stochastic processes with continuous trajectories defined on an interval \([a, b]\). \( F_t \) stands for the marginal univariate distribution function of \( x_i(t) \). Let \( D_n \) be a depth defined on \( \mathbb{R} \). For each fixed \( t \in [a, b] \),

\[
D_n(x_i(t)) = Z_i(t)
\]

is defined as the univariate depth of \( x_i(t) \) at \( t \) with respect to \( x_1(t), x_2(t), \ldots, x_n(t) \). Therefore, at each single point we have ranked the values of \( x_1(t), x_2(t), \ldots, x_n(t) \) according to their depth \( Z_i(t), 1 \leq i \leq n \). However, we are more interested in the trend of a whole curve instead of a single point. Hence, we define

\[
I_i = \int_a^b Z_i(t) dt, \quad 1 \leq i \leq n,
\]

and rank the functions \( x_1(t), x_2(t), \ldots, x_n(t) \) according to the associated \( I_i \)’s values to obtain order statistics. Hence, the functional median will correspond to the \( x_i(t) \) for which \( I_i \) is the maximum. If we rank the observations \( x_1(t), x_2(t), \ldots, x_n(t) \) in decreasing order of values of \( I_i \) then we get order statistics \( x^{(1)}(t), x^{(2)}(t), \ldots, x^{(n)}(t) \), where \( x^{(1)}(t) \) is the functional median and \( x^{(n)}(t) \) corresponds to the \( x_i(t) \) for which
$I_i$ is the minimum.

Let $D_n$ be the depth associated with the univariate distribution $F_t$, and $x_i = x_i(t)$, $t \in [a,b]$ is continuous, $i = 1, 2, \cdots, n$.

Fraiman and Muniz functional depth (FMD) of a curve $x_i$ with respect to the set $x_1, x_2, \cdots, x_n$ is defined by

$$FMD_n(x_i) = \int_a^b D_n(x_i(t))dt. \quad (3.13)$$

Considering the univariate depth of the point $x_i(t)$ given by

$$D_n(x_i(t)) = 1 - \left| \frac{1}{2} - F_{n,t}(x_i(t)) \right|,$$

then, we have

$$FMD_n(x_i) = \int_a^b \left[ 1 - \left| \frac{1}{2} - F_{n,t}(x_i(t)) \right| \right] dt.$$ 

In most cases, these curves are observed at a discrete set of different time points $a \leq t_1 \leq t_2 \leq \cdots \leq t_m \leq b$, and a functional dataset of $n$ identically distributed functional curves, $x_1, x_2, \cdots, x_n$, observed at a grid of points, $t_1 \leq t_2 \leq \cdots \leq t_m$, is accordingly given by

$$\{x_i(t_j); i = 1, 2, \cdots, n, j = 1, 2, \cdots, m\}. \quad (3.14)$$

We can always apply transformation strategy to make them in the same grid of points if they are not.
The sample FMD of curves in equation (3.14) can be obtained by approximating the integral of equation (3.13). By using Reimann sums, we get

\[
FMD_n(x_i) = \sum_{j=2}^{m} \left[ 1 - \frac{1}{2} - F_{n,t_j}(x_i(t_j)) \right] \Delta_j, \quad i = 1, 2, \cdots, n, \quad (3.15)
\]

where \(\Delta_j = t_j - t_{j-1}\) and \(F_{n,t_j}(x_i(t_j))\) is the empirical cumulative distribution function of the points \(x_1(t_j), x_2(t_j), \cdots, x_n(t_j)\). Therefore, the approximation of sample FMD can be evaluated by

\[
FMD_n(x_i) = \sum_{j=2}^{m} \left[ 1 - \frac{1}{2} - \frac{1}{n} \sum_{k=1}^{n} I\{x_k(t_j) \leq x_i(t_j)\} \right] \Delta_j.
\]

Hence, the calculation of FMD of the curve \(x_i(t)\) is changed to calculate the weighted sum of the univariate depths of the datasets \(x_1(t_j), x_2(t_j), \cdots, x_n(t_j), j = 1, 2, \cdots, m\).

### 3.2.2 The h-Modal Depth (MD)

Cuevas et al. [8] defined a new functional depth based on mode. The basic idea is to select the trajectory most densely surrounded by other trajectories of the process.

Given a kernel function \(K : \mathbb{R} \to \mathbb{R}\) and a fixed tuning parameter \(h\), we define

\[
g(x; h) = E(K_h(\|x - X\|)) \quad (3.16)
\]

where \(X\) is the random element describing the population, \(\|\cdot\|\) is a suitable norm, and \(K_h(t)\) is a re-scaled kernel of type \(K_h(t) = \frac{1}{h} K(\frac{t}{h})\).

Given a random sample \(x_1, x_2, \cdots, x_n\) of \(X\), the sample version of \(g(x; h)\) is defined
in a natural way by replacing (3.16) by the empirical version

\[ \hat{g}(x; h) = \frac{1}{nh} \sum_{i=1}^{n} K(\frac{\|x - x_i\|}{h}) \]

\[ = \frac{1}{n} \sum_{i=1}^{n} K_h(\|x - x_i\|). \]

Therefore, the \( h \)-modal functional depth (MD) of curve \( x_i \) with respect to the set of curves \( x_1, x_2, \cdots, x_n \) is given by

\[ \text{MD}_n(x_i, h) = \frac{1}{n} \sum_{k=1}^{n} K_h(\|x_i - x_k\|), \quad (3.17) \]

where \( \| \cdot \| \) is the norm in the functional space, \( K_h : R^+ \rightarrow R^+ \) is the above mentioned kernel function, and \( h \) is a bandwidth.

In application, the norm, the kernel function \( K_h(\cdot) \), and the bandwidth \( h \) need to be chosen. Cuevas et al. [8] applied the \( L^2 \) and \( L^\infty \) norms, and they are defined as follows:

\[ \|x_i - x_k\|_2 = \left( \int_a^b (x_i(t) - x_k(t))^2 dt \right)^{\frac{1}{2}}, \]

and

\[ \|x_i - x_k\|_\infty = \sup_{t \in [a,b]} |x_i(t) - x_k(t)|, \]

and the Gaussian kernel

\[ K(t) = \frac{1}{2\pi} \exp\left(-\frac{t^2}{2}\right), \]

where \( h = 0.2 \max\{\|x_i - x_k\| : i, k = 1, 2, \cdots, n\} \), which means \( h \) is the 20th percentile of the empirical distribution of \( \{\|x_i - x_k\| : i, k = 1, 2, \cdots, n\} \). Hence, the curve that
attains the maximum value in equation (3.17) is defined as the functional mode. If
the curves are observed at discretized points, the functional norms are replaced by
the empirical forms
\[
\|x_i - x_k\|_2 = \left( \sum_{j=2}^{m} (x_i(t_j) - x_k(t_j))^2 \Delta_j \right)^{\frac{1}{2}},
\]
and
\[
\|x_i - x_k\|_\infty = \sup\{j=1,\ldots,m\} |x_i(t_j) - x_k(t_j)|.
\]
The authors explained that a wide range of values of $h$ can be applied, and the only
requirement is $h$ is not too small.

### 3.2.3 The Random Projection Depth (RPD)

Cuevas et al. [9] used the notion of the random projection depth based on the idea
of Custa-Albertos et al. [5] [6].

Given a sample $x_1, x_2, \cdots, x_n$ and a random direction $\nu$ (independent of $x_i$), and
project the data along this direction. Then the sample depth for a datum $x_i$ is defined
as the univariate Halfspace depth of the corresponding one-dimensional projection
(expressed in terms of order statistics so that the median is the deepest point). When
the sample is made of functional data, we will assume that the $x_i, i = 1, 2, \cdots, n$
belong to the Hilbert space $L^2[0, 1]$.

Given a functional data sample $x_1, x_2, \cdots, x_n$ and a random direction $\nu(t)$ (independent of $x_i$), and project the data along this direction. The projection of $x_i$ is given
by the standard inner product $P_{ν,i} = \langle ν, x_i \rangle = \int_0^1 ν(t)x_i(t)dt$. Hence we obtain a set of scalar dataset, $P_{ν,1}, P_{ν,2}, \cdots, P_{ν,n}$, their empirical distribution function is defined by

$$F_n(x) = \frac{1}{n} \sum_{i=1}^n I\{P_{ν,i} \leq x\}. \quad (3.18)$$

The Random Projection Depth of a functional datum $x_i$ will be the Halfspace depth of the projection of $x_i$ along a random direction $ν(t)$. Therefore, it can be defined by

$$\text{RPD}_n(x_i, ν) = \min\{F_n(P_{ν,i}), 1 - F_n(P_{ν,i})\}.$$  

In the finite-dimensional case the projection of $x = (x_1, x_2, \cdots, x_n)$ along the direction $ν = (ν_1, ν_2, \cdots, ν_n)$ is evaluated through the Euclidean inner product $\langle ν, x \rangle = x_1ν_1 + x_2ν_2 + \cdots + x_nν_n$.

The direction can be chosen according to a Gaussian distribution in the appropriate space, standardized to norm 1. This definition leads to a random measure of depth since it is based on the rank of the projections along a random direction.

### 3.2.4 The Random Tukey Depth (RTD)

The Tukey depth was introduced in [34] and can be defined as follows.

Let $P$ be a probability on $R^p$ and $μ \in R^p$. If $Π_ν$ denotes the projection on the one dimensional subspace generated by $ν$ and $P_μ$ the one-dimensional marginal of $P$ on the same subspace, then, the Tukey depth of $x$ with respect to $P$ is

$$\text{TD}(x, P) = \inf\{\text{TD}_1(Π_ν(x), P_μ): ν \in R^p\}, \quad (3.19)$$
where $TD_1(x, P) = \min\{P(-\infty, x], P[x, \infty)\}$ is the Tukey depth or the Halfspace depth of one dimension case which is introduced at the beginning of Section 3.2.

Some computational problems led the authors of [7] to introduce the random Tukey depth, which can be a random approximation of the Tukey depth. In [7], the following generalization to Hilbert spaces was proposed.

Let $H$ be a separable Hilbert space, $P$ be a probability distribution on $H$, $\nu$ be a Gaussian distribution with non-degenerated marginals on $H$ and $\nu_1, \cdots, \nu_k$ be i.i.d. random vectors with distribution $\nu$. The random Tukey depth of $x \in H$ with respect to $P$ based on $k$ random vectors chosen with $\nu$ is

$$RTD_{k, \nu}(x, P) = \min\{TD_1(\Pi_{\nu_i}(x), P_{\nu_i}), i = 1, 2, \cdots, k\}.$$ 

In this thesis, we will applied functional depths FMD, MD, RPD, and RTD in our study.

### 3.3 Trimmed Mean for Functional Data

In one dimension, order statistics and ranks are widely used for several applications, such as distribution free tests and some simple robust estimation procedures. In this case, they are defined through the natural order on the real line.

L-estimates, which are defined as linear combinations of order statistics, are a well-known class of robust estimates. Particularly, trimmed means, which are defined as the average of the most central $(1 - \alpha)n$ observations, where $0 \leq \alpha \leq 1$, constitute
a class of estimates that range from the sample mean to the sample median. In [9], Cuevas et al. concluded that the best estimator is the trimmed mean (when combined with the h-modal depth) in all considered cases of models and distance measures.

A functional version of the $\alpha$-trimmed mean is defined as the average of the $n - \lfloor n\alpha \rfloor$ deepest observations, where $\alpha$ satisfies $0 \leq \alpha \leq (n - 1)/n$, and $\lfloor t \rfloor$ denotes the largest integer that is less or equal to $t$.

Let $D_n(x_1), D_n(x_2), \ldots, D_n(x_n)$ be the functional depths of the set of trajectories $x_1, x_2, \ldots, x_n$ with any of the four functional depths introduced in Section 3.2. According to the multivariate case, the largest value of $D_n(x_i)$ the deepest is the curve $x_i$, among the curves $x_1, x_2, \ldots, x_n$. We get the ordered curves $x^{(1)}, x^{(2)}, \ldots, x^{(n)}$, such that $x^{(1)}$ is the deepest curve and $x^{(n)}$ is the curve with the least depth. Hence, we can define the location estimates for functional data based on the rank. For example, for the FMD depth, the trimmed mean will be

$$
\text{FMD}_\alpha(x_1(t), \ldots, x_n(t)) = \frac{1}{n - \lfloor n\alpha \rfloor} \sum_{i=1}^{n - \lfloor n\alpha \rfloor} x^{(i)}(t).
$$

(3.20)

We will let $\alpha = 0.25$, which is between 0.2 and 0.3, Fraiman and Muniz proposed in [19]. In [19], the authors compared trimmed mean estimates with the regular mean under basic and contaminated models and concluded that the functional trimmed mean estimates performed very well under all models – even under asymmetric contamination.
Chapter 4

Functional Outlier Detection Procedures

In order to detect outliers in functional datasets, we apply functional depths, FMD, MD, RPD and RTD. Based on the definition, depth and outliers are inverse notions, hence, trajectories corresponding to outliers in the dataset will have the least depths. Therefore, a way to identify functional outliers is to evaluate the depth of each functional curve and look for the curve with the smallest depth. We will apply the following functional outlier detection procedures, which were proposed by Febrero et al [16], to detect outliers in a given dataset consisting of functional curves $x_1, x_2, \cdots, x_n$:

1. Obtain the functional depths $D_n(x_1), D_n(x_2), \cdots, D_n(x_n)$, for one of the functional depths, FMD, MD, RPD and RTD.

2. Let $x_{i_1}, x_{i_2}, \cdots, x_{i_k}$ be the $k$ curves such that $D_n(x_{i_j}) \leq C$, $j = 1, 2, \cdots, k$, for a given cutoff $C$. Then, assume that $x_{i_1}, x_{i_2}, \cdots, x_{i_k}$ are outliers and delete them.
from the sample.

3. Repeat Step 1 with the new dataset which has deleted the outliers detected in
Step 2. Repeat it until no more outliers are detected.

Step 3 aims to avoid the outlier masking, which takes place when large outliers
mask the presence of other outliers. Therefore, if outliers are masked in one iter-
ation, they may be detected in a later iteration after removing the outliers have been
detected.

The critical element of this procedure is to determine the value of $C$ which should
be so small that only a small fraction, say 5%, of the curves are classified as outliers
when there are no outliers. It means that the percentage of correct observations are
misidentified as outliers is about 5%, i.e. $P_r(D_n(x_i) \leq C) = 0.05$, $i = 1, 2, \ldots, n$.

It is hard to evaluate $C$ since the distribution of a functional depth is unknown.
Therefore, two bootstrap methods are applied to obtain a robust estimate of the
percentile of the sample data. One is based on trimming the suspicious trajectories
of the sample to obtain smoothed bootstrap sets and estimate the percentile based
on it; another method is based on bootstrapping the curves of original dataset with
probability proportional to their depth. As a result, the depth with the largest value
will be sampled more frequently to avoid the presence of outliers in the bootstrap
samples. The procedure of these two methods will be discussed and applied.

The procedure of bootstrap method based on trimming is as follows:
1. Obtain the functional depths $D_n(x_1), D_n(x_2), \cdots, D_n(x_n)$ of FMD, MD, RPD and RTD.

2. Obtain $B$ standard bootstrap samples of size $n$ from the dataset of curves obtained after deleting $100\alpha\%$ less deepest curves. The bootstrap samples are denoted by $x_i^b$, for $i = 1, 2, \cdots, n$ and $b = 1, 2, \cdots, B$.

3. Obtain smoothed bootstrap samples $y_i^b = x_i^b + z_i^b$, where $z_i^b$ is such that $(z_i^b(t_1), z_i^b(t_2), \cdots, z_i^b(t_m))$ is normally distributed with mean 0 and covariance matrix $\gamma \Sigma_x$, where $\Sigma_x$ is the covariance matrix of $x(t_1), x(t_2), \cdots, x(t_m)$ and $\gamma$ is a bootstrap smoothing parameter. Let $y_i^b, i = 1, 2, \cdots, n$ and $b = 1, 2, \cdots, B$, be these samples.

4. For each bootstrap, set $b = 1, 2, \cdots, B$, obtain $C^b$ as the empirical 5th percentile of the distribution of the depths, $D_n(y_i^b), i = 1, 2, \cdots, n$.

5. Take $C$ as the median of the values of $C^b, b = 1, 2, \cdots, B$.

The procedure of bootstrap method based on weighting is very similar to the one based on trimming except step 2.

1. Obtain the functional depths $D_n(x_1), D_n(x_2), \cdots, D_n(x_n)$ of FMD, MD, RPD and RTD.

2. Obtain $B$ standard bootstrap samples from the curves in which each original curve is sampled with a probability proportional to its depth. The samples are denoted by $x_i^b$, for $i = 1, 2, \cdots, n$ and $b = 1, 2, \cdots, B$.

3. Obtain smoothed bootstrap samples $y_i^b = x_i^b + z_i^b$, where $z_i^b$ is such that
\((z^b_i(t_1), z^b_i(t_2), \cdots, z^b_i(t_m))\) is normally distributed with mean 0 and covariance matrix \(\gamma \Sigma_x\), where \(\Sigma_x\) is the covariance matrix of \(x(t_1), x(t_2), \cdots, x(t_m)\) and \(\gamma\) is a bootstrap smoothing parameter. Let \(y^b_i, i = 1, 2, \cdots, n\) and \(b = 1, 2, \cdots, B\) be these samples.

4. For each bootstrap, set \(b = 1, 2, \cdots, B\), obtain \(C^b\) as the empirical 5th percentile of the distribution of the depths, \(D_n(y^b_i), i = 1, 2, \cdots, n\).

5. Take \(C\) as the median of the values of \(C^b, b = 1, 2, \cdots, B\).
Chapter 5

Simulation Results

In this chapter, several aspects of the outlier detection procedures for functional data introduced in Chapter 4 will be explored by simulation experiments. Some R packages, such as fda.usc [17] and roahd [33], are applied in the simulation experiments and real data outlier detection. Functional data \(x_1, x_2, \ldots, x_n\) are obtained as realizations from a stochastic process \(X(\cdot)\), which have continuous trajectories on the observation period \([a, b] = [0, 1]\). The mechanism model for generating curves without outliers is a Gaussian process \(X(t)\) of the form:

\[
X(t) = E(t) + e(t)
\]

where \(E(t) = E(X(t)) = 50t^2(1-t)^{\frac{3}{2}}\) and \(e(t)\) is a Gaussian process with mean 0 and covariance matrix:

\[
E[e(t_i)e(t_j)] = 0.9 \exp \left( -\frac{|t_i - t_j|}{0.9} \right). \tag{5.21}
\]
5.1 Type I Error

A type I error is the incorrect rejection of a true null hypothesis. Usually a type I error leads one to conclude that a supposed effect or relationship exists when in fact it doesn’t. In this section, we will discuss the type I error of applied procedures in the outlier detection process.

We generate 100 datasets with curves from the outliers free model with sample size $n = 100$ and $n = 200$. These curves are observed at equidistant points $t_1 = 0, t_i = t_{i-1} + h, i = 2, \ldots, m$, $t_m = 1$, where $h = (t_m - t_1)/(m - 1)$. Let the number of grid points $m = 50$ in this simulation study. Based on trimming and weighting methods, the procedures discussed in Chapter 4 will be applied to the outlier free datasets respectively. We will use the bootstrap smoothing parameter $\gamma = 0.05$, bootstrap samples $B = 200$ and 5th percentile to determine the cutoff for bootstrap procedure.

The results of the type I errors corresponding to different depths based on trimming and weighting methods are shown in Table 5.1. The type I error is evaluated by the mean percentage of false outliers detected by the procedure in each dataset. Type I errors based on weighting method are much smaller than those based on trimming method. Type I errors based on weighting method are below 1% except for FMD, 1.01%, which is very close to 1%. Overall, all type I errors are much lower than the acceptable significance level 5%.
Table 5.1: Percentage of type I errors with outliers free datasets for sample size $n=100$ and $n=200$

<table>
<thead>
<tr>
<th>Sample Size</th>
<th>$n=100$</th>
<th>$n=200$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Method</td>
<td>Trimming</td>
<td>Weighting</td>
</tr>
<tr>
<td>FMD</td>
<td>2.63%</td>
<td>0.58%</td>
</tr>
<tr>
<td>MD</td>
<td>2.69%</td>
<td>0.20%</td>
</tr>
<tr>
<td>RPD</td>
<td>1.14%</td>
<td>0.28%</td>
</tr>
<tr>
<td>RTD</td>
<td>0.38%</td>
<td>0.20%</td>
</tr>
</tbody>
</table>

5.2 The Power Analysis of Outlier Detection Procedures

Now, we consider the datasets with contaminated curves and conduct simulation experiments in order to study the power of the outlier detection procedures. We will apply the alternative model for generating outlier curves, in which the mean function $E(t) = E(X(t)) = 50t^2(1-t)^{\frac{1}{2}}$ is replaced with $E(t) = 50t^{\frac{3}{2}}(1-t)^{\frac{3}{2}}$. We also generate 100 datasets with sample sizes $n=100$, $n=200$, and $n = n_0 + n_1$, where $n_0$ curves are generated from the contaminated model and $n_1$ curves are generated from the non-contaminated model. The frequency of correct outlier detection, the false outliers detected rate and the number of iterations that there are no outliers detected will be recorded during the simulation process.

Three cases are considered for sample sizes $n=100$ and $n=200$ respectively.

Case I: $n_0=1$ and $n_1=99$ when sample size $n=100$; $n_0=1$ and $n_1=199$ when sample size $n=200$.

Case II: $n_0=2$ and $n_1=98$ when sample size $n=100$; $n_0=2$ and $n_1=198$ when sample
size \( n = 200 \).

Case III: \( n_0 = 3 \) and \( n_1 = 97 \) when sample size \( n = 100 \); \( n_0 = 3 \) and \( n_1 = 197 \) when sample size \( n = 200 \).

The simulations will be conducted at equidistant points with 50 grid points as non-contaminated cases. We also use \( \gamma = 0.05 \), \( B = 200 \) bootstrap samples and 5th percentile to determine the cutoff for bootstrap procedure. The outlier detection procedures based on trimming and weighting will be applied to each dataset corresponding to the three cases.

The simulation results are shown in Table 5.2, row 4 to 7 show the frequency of correct outlier detection based on FMD, MD, RPD and RTD respectively. The frequency of correct outlier detection is the number of times that the true outliers are detected during the simulation process over the 100 datasets. Row 8 to 11 show the false outliers detection rate based on FMD, MD, RPD and RTD respectively. The false outliers detection rate is calculated as the number of false outliers detected divided by the product of sample size and the size of the dataset. Row 12 to 15 record the iteration procedure that there are no outliers detected either correct or incorrect.

We can derive the conclusion from Table 2 as follows:

First, the procedures of correct frequency detection based on trimming perform better than those based on weighting at a cost of higher false outliers detection rate. Some of them are large than 1%, but they all are below the acceptable significance.
Second, it appears that there are differences among the results from different depths. For example, in terms of trimming method, RTD has the worst performance. When sample size $n=100$, $n_0=1$, $n_0=2$ and $n_0=3$ respectively, the correct frequency detections are 68, 86 and 84, while the false outliers detection rates are small, 0.5%, 0.64% and 0.77% respectively, and the frequencies of no outliers being detected are 29, 13 and 16 accordingly. MD has the best performance of trimming method, when sample size $n=100$, the frequencies of correct detection are 100, 100 and 100, the false outliers detection rates are 2.02%, 1.02% and 1.35%, corresponding to $n_0=1$, $n_0=2$ and $n_0=3$; when sample size $n=200$, the frequencies of correct detection are 100, 100 and 100, the false outliers detection rates are 3.68%, 3.20% and 2.62%, corresponding to $n_0=1$, $n_0=2$ and $n_0=3$. The false outliers detection rates are all much less than 5%.

Third, the procedures of correct frequency detection perform better in larger sample size compared to smaller one with the same $n_0$. In terms of MD, when the sample size $n=100$, the frequencies of correct detection are 100, 77 and 81, the false outlier detecting rates are 0.24%, 0.03% and 0.1%, corresponding to $n_0=1$, $n_0=2$ and $n_0=3$, and when sample size $n=200$, the frequencies of correct detection are 100, 100 and 100, the false outliers detection rates are 0.47%, 0.46% and 0.43%, corresponding to $n_0=1$, $n_0=2$ and $n_0=3$, which are much lower than the acceptable significance level.
Overall, MD performs the best while RTD performs the worst among all the depths based on trimming and weighting methods; RPD performs better than FMD with only one exception, which RPD has 97 correct detection rate while FMD has 98; FMD has the largest false outliers detecting rate 4.42% based on trimming method, which is still less than 5%.

Some simulation results surprisingly agree with the conclusion in [9] and [16], such as, MD has the best performance among all the depth measures, the outlier detection procedure performed well with several cases over 90% of frequency detection; the trimming bootstrap estimate of the cutoff makes the procedure have a larger empirical power but at the cost of having a large false outlier detection rate than the weighting bootstrap estimate.
Table 5.2: Correct frequency (freq.) detection, false outliers rate of the applied procedure for detecting functional outliers, and the frequency of no outlier detection, where FMD, MD, RPD and RTD denote the Fraiman and Muniz depth, the h-mode depth, the random projection depth and the Random Tukey depth respectively, and T stands for Trimming and W stands for Weighting

<table>
<thead>
<tr>
<th>Number of Outliers</th>
<th>n₀=1</th>
<th>n₀=2</th>
<th>n₀=3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sample Size</td>
<td>n=100 n=200</td>
<td>n=100 n=200</td>
<td>n=100 n=200</td>
</tr>
<tr>
<td>Methods</td>
<td>T W T W</td>
<td>T W T W</td>
<td>T W T W</td>
</tr>
<tr>
<td>Freq. of Detection</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FMD</td>
<td>88 09 98 29</td>
<td>90 02 100 27</td>
<td>84 01 100 24</td>
</tr>
<tr>
<td>MD</td>
<td>100 100 100 100</td>
<td>100 77 100 100</td>
<td>100 81 100 100</td>
</tr>
<tr>
<td>RPD</td>
<td>96 57 97 84</td>
<td>99 55 100 93</td>
<td>97 26 100 85</td>
</tr>
<tr>
<td>RTD</td>
<td>68 50 84 63</td>
<td>86 62 90 76</td>
<td>84 31 89 74</td>
</tr>
<tr>
<td>False Rate%</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FMD</td>
<td>2.66 0.43 4.09 0.90</td>
<td>2.72 0.31 4.09 0.78</td>
<td>2.77 0.06 4.42 0.70</td>
</tr>
<tr>
<td>MD</td>
<td>2.02 0.24 3.68 0.47</td>
<td>1.02 0.03 3.20 0.46</td>
<td>1.35 0.10 2.62 0.43</td>
</tr>
<tr>
<td>RPD</td>
<td>1.86 0.23 2.61 0.60</td>
<td>2.23 0.11 3.12 0.55</td>
<td>2.39 0.04 3.08 0.42</td>
</tr>
<tr>
<td>RTD</td>
<td>0.50 0.06 1.00 0.34</td>
<td>0.64 0.08 0.98 0.36</td>
<td>0.77 0.02 0.89 0.31</td>
</tr>
<tr>
<td>Freq. of No Detection</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FMD</td>
<td>1 57 0 10</td>
<td>2 70 0 20</td>
<td>9 93 0 21</td>
</tr>
<tr>
<td>MD</td>
<td>0 0 0 0</td>
<td>0 23 0 0</td>
<td>0 17 0 0</td>
</tr>
<tr>
<td>RPD</td>
<td>4 42 3 14</td>
<td>1 44 0 7</td>
<td>3 73 0 15</td>
</tr>
<tr>
<td>RTD</td>
<td>29 49 16 31</td>
<td>13 38 9 22</td>
<td>16 69 11 24</td>
</tr>
</tbody>
</table>
In this chapter, we study the datasets with 21 genes in P. Aeruginosa expressed in 24 conditions introduced in Chapter 1. Each gene was measured every half hour for 21 hours under each condition, therefore, 43 observations were obtained for each gene. The data are repeated measurements of the same subject densely over an ordered grid of points belonging to an interval of finite length 0.5 hour. Thus, for each gene, we observe a function, though the recording points are discrete, we may still consider the entire function as continuous based on the reality. We will apply previous introduced procedures to identify outliers in the gene datasets in order to find out which conditions produce outliers or produce the most outliers. We conduct the outlier detection procedures based on four functional depth FMD, MD, RPD, and RTD to all 21 genes observation datasets.
In Section 6.1, we will show the main results from the outlier detection procedures and conclusions. In Section 6.2, we will list some problems that we found and the discussion of future studies. As before, we take $\gamma = 0.05$, $B = 200$, and use empirical 5th percentile to determine the cutoff of bootstrap procedure.

6.1 Main Results and Conclusions

Figure 6.1, Figure 6.2, Figure 6.3, and Figure 6.4 show the functional depth of gene A4, and the comparisons of the median and 0.25 – trimmed mean of FMD, MD, RPD and RTD respectively. It shows that the 0.25–trimmed mean agrees well with the median and sometimes they superpose. In [9], the authors concluded that the trimmed mean is the best estimator especially combined with $h$-modal depth. We applied the 0.25–trimmed mean to all of the depth calculations, and let the bootstrap smoothing parameter $\gamma = 0.05$.

Table 6.1 shows the outlier detection results, and Table 6.2 – Table 6.14 are the detail information of the outliers, including the cutoff value, the depth of the outlier, and the conditions of the outlier based on trimming and weighting methods respectively. In terms of the performance of each depth, the outlier detection results resemble the simulation results. The total frequency of outlier detection is 27, while 19 of the outliers are detected from trimming method, and seven of them are detected from $h$-modal depth based on trimming method. The trajectories of genes A4, B4,
B5, C4 and G2 are outliers under biological conditions C21; the trajectories of genes D1, E6, F2, G5 and $\sigma70$ are detected as outliers under biological condition C24; the trajectories of gene C4 under biological condition both C18 and C21 are detected as outliers; the trajectories of gene G5 under biological condition C19 and C24 are outliers; the trajectories of genes H3 and H4 are outliers under biological conditions C2 and C5 respectively. Therefore, biological conditions C21 and C24 need to be paid more attention for further study. The trajectory of A6 under condition C16 is detected as outlier four times: h-modal depth based on trimming, random projection depth based on trimming and weighting, and random Tukey depth based on weighting. This also needs further attention. In addition, it is worth of checking condition C2, C5, and C18.

In our study, conditions C21 and C24 are the most significant conditions that produce outliers of gene trajectories; conditions C16 and C19 produce outliers under four and two different depths respectively. To some extent, our results agree with the results in [10]. In [10], the classification chart based on minimum Mahalanobis distance shows there are three clusters among the 24 conditions. Condition C21 alone is classified as the third cluster, condition C16 and C24 stay at second cluster, and all the others constitute the first cluster including some sub-clusters.
Figure 6.1: The 0.25-trimmed mean and the sample median of Gene A4

Figure 6.2: The 0.25-trimmed mean and the sample median of Gene A4
Figure 6.3: The 0.25-trimmed mean and the sample median of Gene A4

Figure 6.4: The 0.25-trimmed mean and the sample median of Gene A4
Table 6.1: Outliers detection results for 21 genes under 24 conditions

<table>
<thead>
<tr>
<th>Depths</th>
<th>FMD</th>
<th>MD</th>
<th>RPD</th>
<th>RTD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Methods</td>
<td>T W</td>
<td>T W</td>
<td>T W</td>
<td>T W</td>
</tr>
<tr>
<td>A4</td>
<td></td>
<td>C21</td>
<td></td>
<td></td>
</tr>
<tr>
<td>A6</td>
<td>C16</td>
<td></td>
<td>C16</td>
<td>C16</td>
</tr>
<tr>
<td>B3</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>B4</td>
<td></td>
<td></td>
<td>C21</td>
<td></td>
</tr>
<tr>
<td>B5</td>
<td></td>
<td>C21</td>
<td></td>
<td>C21</td>
</tr>
<tr>
<td>C4</td>
<td>C18</td>
<td></td>
<td></td>
<td>C21</td>
</tr>
<tr>
<td>C5</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>D1</td>
<td>C24</td>
<td>C24</td>
<td></td>
<td></td>
</tr>
<tr>
<td>D2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>D3</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>E5</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>E6</td>
<td></td>
<td></td>
<td>C24</td>
<td>C24</td>
</tr>
<tr>
<td>F2</td>
<td>C24</td>
<td>C24</td>
<td></td>
<td></td>
</tr>
<tr>
<td>F3</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>F5</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>G2</td>
<td></td>
<td></td>
<td>C21</td>
<td>C21</td>
</tr>
<tr>
<td>G5</td>
<td>C24</td>
<td>C24</td>
<td>C19</td>
<td>C19</td>
</tr>
<tr>
<td>G6</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>H3</td>
<td></td>
<td></td>
<td>C2</td>
<td></td>
</tr>
<tr>
<td>H4</td>
<td></td>
<td></td>
<td></td>
<td>C5</td>
</tr>
<tr>
<td>S70</td>
<td>C24</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 6.2: Outlier detection Details of genes A4

<table>
<thead>
<tr>
<th>Method</th>
<th>Trimming</th>
<th>Weighting</th>
</tr>
</thead>
<tbody>
<tr>
<td>Depth</td>
<td>cutoff</td>
<td>Depth</td>
</tr>
<tr>
<td>FMD</td>
<td>0.44</td>
<td>0.40</td>
</tr>
</tbody>
</table>
Table 6.3: Outlier detection Details of genes A6

<table>
<thead>
<tr>
<th>Method</th>
<th>Trimming</th>
<th>Weighting</th>
</tr>
</thead>
<tbody>
<tr>
<td>Depth</td>
<td>cutoff</td>
<td>Depth</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FMD</td>
<td>0.64</td>
<td>0.40</td>
</tr>
<tr>
<td>MD</td>
<td>0.40</td>
<td></td>
</tr>
<tr>
<td>RPD</td>
<td>0.10</td>
<td>0.08</td>
</tr>
<tr>
<td>RTD</td>
<td>0.06</td>
<td>0.04</td>
</tr>
</tbody>
</table>

Table 6.4: Outlier detection Details of genes B4

<table>
<thead>
<tr>
<th>Method</th>
<th>Trimming</th>
<th>Weighting</th>
</tr>
</thead>
<tbody>
<tr>
<td>Depth</td>
<td>cutoff</td>
<td>Depth</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FMD</td>
<td>0.40</td>
<td></td>
</tr>
<tr>
<td>MD</td>
<td>0.40</td>
<td></td>
</tr>
<tr>
<td>RPD</td>
<td>0.05</td>
<td>0.04</td>
</tr>
<tr>
<td>RTD</td>
<td>0.05</td>
<td>0.04</td>
</tr>
</tbody>
</table>

Table 6.5: Outlier detection Details of genes B5

<table>
<thead>
<tr>
<th>Method</th>
<th>Trimming</th>
<th>Weighting</th>
</tr>
</thead>
<tbody>
<tr>
<td>Depth</td>
<td>cutoff</td>
<td>Depth</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FMD</td>
<td>0.40</td>
<td></td>
</tr>
<tr>
<td>MD</td>
<td>0.40</td>
<td></td>
</tr>
<tr>
<td>RPD</td>
<td>0.05</td>
<td>0.04</td>
</tr>
<tr>
<td>RTD</td>
<td>0.05</td>
<td>0.04</td>
</tr>
</tbody>
</table>

Table 6.6: Outlier detection Details of genes C4

<table>
<thead>
<tr>
<th>Method</th>
<th>Trimming</th>
<th>Weighting</th>
</tr>
</thead>
<tbody>
<tr>
<td>Depth</td>
<td>cutoff</td>
<td>Depth</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FMD</td>
<td>0.09</td>
<td>0.08</td>
</tr>
<tr>
<td>MD</td>
<td>0.42</td>
<td>0.41</td>
</tr>
<tr>
<td>RPD</td>
<td>0.05</td>
<td>0.04</td>
</tr>
<tr>
<td>RTD</td>
<td>0.05</td>
<td>0.04</td>
</tr>
</tbody>
</table>
Table 6.7: Outlier detection Details of genes D1

<table>
<thead>
<tr>
<th>Method</th>
<th>Trimming</th>
<th>Weighting</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Depth</td>
<td>Outlier</td>
</tr>
<tr>
<td></td>
<td>cutoff</td>
<td>Depth</td>
</tr>
<tr>
<td>FMD</td>
<td>0.11</td>
<td>0.08</td>
</tr>
<tr>
<td>MD</td>
<td></td>
<td>0.08</td>
</tr>
<tr>
<td>RPD</td>
<td></td>
<td>0.08</td>
</tr>
<tr>
<td>RTD</td>
<td></td>
<td>0.08</td>
</tr>
</tbody>
</table>

Table 6.8: Outlier detection Details of genes E6

<table>
<thead>
<tr>
<th>Method</th>
<th>Trimming</th>
<th>Weighting</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Depth</td>
<td>Outlier</td>
</tr>
<tr>
<td></td>
<td>cutoff</td>
<td>Depth</td>
</tr>
<tr>
<td>FMD</td>
<td>0.43</td>
<td>0.40</td>
</tr>
<tr>
<td>MD</td>
<td>0.06</td>
<td>0.05</td>
</tr>
<tr>
<td>RPD</td>
<td>0.05</td>
<td>0.04</td>
</tr>
<tr>
<td>RTD</td>
<td>0.05</td>
<td>0.04</td>
</tr>
</tbody>
</table>

Table 6.9: Outlier detection Details of genes F2

<table>
<thead>
<tr>
<th>Method</th>
<th>Trimming</th>
<th>Weighting</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Depth</td>
<td>Outlier</td>
</tr>
<tr>
<td></td>
<td>cutoff</td>
<td>Depth</td>
</tr>
<tr>
<td>FMD</td>
<td>0.03</td>
<td>0.01</td>
</tr>
<tr>
<td>MD</td>
<td>0.03</td>
<td>0.01</td>
</tr>
<tr>
<td>RPD</td>
<td>0.06</td>
<td>0.05</td>
</tr>
<tr>
<td>RTD</td>
<td>0.05</td>
<td>0.04</td>
</tr>
</tbody>
</table>

Table 6.10: Outlier detection Details of genes G2

<table>
<thead>
<tr>
<th>Method</th>
<th>Trimming</th>
<th>Weighting</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Depth</td>
<td>Outlier</td>
</tr>
<tr>
<td></td>
<td>cutoff</td>
<td>Depth</td>
</tr>
<tr>
<td>FMD</td>
<td>0.06</td>
<td>0.05</td>
</tr>
<tr>
<td>MD</td>
<td>0.04</td>
<td>0.04</td>
</tr>
<tr>
<td>RPD</td>
<td></td>
<td>0.04</td>
</tr>
<tr>
<td>RTD</td>
<td></td>
<td>0.04</td>
</tr>
</tbody>
</table>
Table 6.11: Outlier detection Details of genes G5

<table>
<thead>
<tr>
<th>Method</th>
<th>Trimming</th>
<th>Weighting</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Depth</td>
<td>Outlier</td>
</tr>
<tr>
<td>FMD</td>
<td>0.14</td>
<td>0.07</td>
</tr>
<tr>
<td>MD</td>
<td>0.44</td>
<td>0.40</td>
</tr>
<tr>
<td>RPD</td>
<td>0.07</td>
<td>0.07</td>
</tr>
<tr>
<td>RTD</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 6.12: Outlier detection Details of genes H3

<table>
<thead>
<tr>
<th>Method</th>
<th>Trimming</th>
<th>Weighting</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Depth</td>
<td>Outlier</td>
</tr>
<tr>
<td>FMD</td>
<td>0.08</td>
<td>0.07</td>
</tr>
<tr>
<td>MD</td>
<td></td>
<td></td>
</tr>
<tr>
<td>RPD</td>
<td>0.08</td>
<td>0.07</td>
</tr>
<tr>
<td>RTD</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 6.13: Outlier detection Details of genes H4

<table>
<thead>
<tr>
<th>Method</th>
<th>Trimming</th>
<th>Weighting</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Depth</td>
<td>Outlier</td>
</tr>
<tr>
<td>FMD</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MD</td>
<td></td>
<td></td>
</tr>
<tr>
<td>RPD</td>
<td>0.07</td>
<td>0.07</td>
</tr>
<tr>
<td>RTD</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 6.14: Outlier detection Details of genes S70

<table>
<thead>
<tr>
<th>Method</th>
<th>Trimming</th>
<th>Weighting</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Depth</td>
<td>Outlier</td>
</tr>
<tr>
<td>FMD</td>
<td>0.10</td>
<td>0.08</td>
</tr>
<tr>
<td>MD</td>
<td></td>
<td></td>
</tr>
<tr>
<td>RPD</td>
<td></td>
<td></td>
</tr>
<tr>
<td>RTD</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
6.2 Questions and Discussions

All outliers detected from the outlier detection procedures agree with the visualization result. However, there are no outliers detected from the outlier detection procedures for four genes, although some curves are significantly different from others visually. Figure 6.5, Figure 6.6, Figure 6.7 and Figure 6.8 are the curves of genes E5, D2, F3 and G6 in log scale respectively. For example, gene E5, about 50% of the curves increase at beginning and tend to be stable after a period, and these curves visually have different trends compared to the curve under control condition, while the rest curves have similar trends with the curve under control condition but with strong fluctuations. Therefore, we speculate that the outlier detection procedures may be invalid to the situation like E5 since these visually different elements have larger weight due to their amounts. This brings question to both statistics and the biology. In terms of statistical study, we need to develop some new tools to study this phenomenon and provide a better interpretation. For the biologists, they may need to take a second look with the biological conditions.
Figure 6.5: 24 trajectories of gene E5
Figure 6.6: 24 trajectories of gene D2
Figure 6.7: 24 trajectories of gene F3
Figure 6.8: 24 trajectories of gene G6
Chapter 7

Summary and Future Research

It is very important to identify outliers in a functional dataset. There are many reasons may produce outliers. Some outliers may come from the gross errors, such as experimental design, measurement, recording of information, and some outliers may exist by nature. Therefore, it is crucial to avoid or delete the outliers from gross errors, and it also has practical significance to locate the sources that produce the real outliers. The purpose of this thesis is to introduce a new method to study high-dimensional data from different perspectives. We apply some non-parametric outlier detection procedures to the high-dimensional temporal gene datasets. By conducting simulation experiments to different sample sizes and under different conditions, the outlier detection procedures are proved effective. We successfully detected some biological conditions that produce outliers to some genes, and these results agree with some previous study results which are derived from different statistical methods.
and perspectives. We also found the procedures did not work well for gene trajectories that have two distinct distributions visually. This may lead to a new research direction.

In the future research, we are going to focus on the following topics.

First, we may propose some new methods to study the functional dataset comes from two distinct distribution families almost evenly, which means neither the mean nor the median is a typical representative of these curves.

Second, we did not attempt to develop an asymptotic justification for the procedure described. Its performance is assessed by simulation experiments. Though such an approach is common in functional data, we will work towards to propose statistical procedures whose asymptotic validity can be established in our future study.
Appendix A

R code example for outlier detection:

gene E6
```r
library(RColorBrewer)
library(fda.usc)

## Warning: package 'fda.usc' was built under R version 3.5.1
## Loading required package: fda
## Warning: package 'fda' was built under R version 3.5.1
## Loading required package: splines
## Loading required package: Matrix
##
## Attaching package: 'fda'
## The following object is masked from 'package:graphics':
##
## matplot
##
## Warning: package 'MASS' was built under R version 3.5.1
## Loading required package: mgcv
## Loading required package: nlme
## This is mgcv 1.8-23. For overview type 'help("mgcv-package")'.
## Loading required package: rpart

E6 <- read.csv("I:/ResearchDataLogCSV/E6.csv", header=T)

col.names <- conditions

matplot(E6, type="l", pch=1:25)
```
TE6 <- t(E6)
FTE6 <- fdata(TE6, argvals=NULL, rangeval=NULL, names=NULL, fdata2d=FALSE)

depth.FM(FTE6, fdataori=FTE6, trim=0.25, scale=FALSE, dfunc="FM1", par.dfunc=list(scale=TRUE), draw=FALSE)

depth.mode(FTE6, fdataori=FTE6, trim=0.25, metric="metric.lp", h=NULL, scale=FALSE, draw=FALSE)

depth.RP(FTE6, fdataori=FTE6, trim=0.25, nproj=50, proj="vexponential", dfunc="TD1", par.dfunc=list(), scale=FALSE, draw=FALSE)

depth.RT(FTE6, fdataori=FTE6, trim = 0.25, nproj = 10, proj = 1, xeps = 1e-07, draw = FALSE)

outliers.depth.trim(): detecting outliers based on trimming method
outliers.depth.pond(): detecting outliers based on weighting method

outliers.depth.trim(FTE6, nb=200, smo=0.05, trim=0.25, quan=0.05, dfunc=depth.FM)
## $outliers
## character(0)
##
## $dep.out
## NULL
##
## $iteration
## numeric(0)
##
## $quantile
## 5%
## 0.0704186
##
## $Dep
## 1 2 3 4 5 6 7
## 0.5376744 0.1990698 0.6716279 0.4148837 0.5413953 0.6309677 0.7181395
## 8 9 10 11 12 13 14
## 0.2679070 0.1953488 0.6083721 0.3441860 0.5097674 0.2790698 0.7144186
## 15 16 17 18 19 20 21
## 0.7367442 0.6660465 0.7367442 0.5339535 0.5525581 0.3851163 0.3646512
## 22 23 24 25
## 0.3572093 0.5823256 0.1265116 0.8148837

outliers.depth.pond(FTE6, nb=200, smo=0.05, quan=0.05, dfunc=depth.FM)

## $outliers
## character(0)
##
## $dep.out
## NULL
##
## $iteration
## numeric(0)
##
## $quantile
## 5%
## 0.08372093
##
## $Dep
## 1 2 3 4 5 6 7
## 0.5376744 0.1990698 0.6716279 0.4148837 0.5413953 0.6309677 0.7181395
## 8 9 10 11 12 13 14
## 0.2679070 0.1953488 0.6083721 0.3441860 0.5097674 0.2790698 0.7144186
## 15 16 17 18 19 20 21
## 0.7367442 0.6660465 0.7367442 0.5339535 0.5525581 0.3851163 0.3646512
## 22 23 24 25
## 0.3572093 0.5823256 0.1265116 0.8148837

outliers.depth.trim(FTE6, nb=200, smo=0.05, quan=0.05, dfunc=depth.mode)

## $outliers
## character(0)
##
## $dep.out
## NULL
##
## $iteration
## numeric(0)
##
## $quantile
## 5%
## 0.08372093
##
## $Dep
## 1 2 3 4 5 6 7
## 0.5376744 0.1990698 0.6716279 0.4148837 0.5413953 0.6309677 0.7181395
## 8 9 10 11 12 13 14
## 0.2679070 0.1953488 0.6083721 0.3441860 0.5097674 0.2790698 0.7144186
## 15 16 17 18 19 20 21
## 0.7367442 0.6660465 0.7367442 0.5339535 0.5525581 0.3851163 0.3646512
## 22 23 24 25
## 0.3572093 0.5823256 0.1265116 0.8148837

outliers.depth.trim(FTE6, nb=200, smo=0.05, trim=0.25, quan=0.05, dfunc=depth.mode)

## $outliers
## [1] "C24"
## $dep.out
## [1] 0.3989423
##
## $iteration
## [1] 1
##
## $quantile
## 5%
## 0.5245457
##
## $Dep
## 1 2 3 4 5 6 7
## 2.4008470 0.6496839 2.9430730 0.8933397 2.3618842 2.7723702 3.7461218
## 8 9 10 11 12 13 14
## 1.3456888 2.5027392 3.5165780 2.9918116 1.4726572 2.5350945 3.1748282
## 15 16 17 18 19 20 21
## 3.9500035 3.8610601 3.7935308 3.4626883 2.4910748 2.9201509 0.7382501
## 22 23 24 25
## 0.9943610 2.4259651 0.3989423 3.8747116

outliers.depth.pond(FTE6, nb=200, smo=0.05, quan=0.05, dfunc=depth.mode)

## $outliers
## character(0)
##
## $dep.out
## NULL
##
## $iteration
## numeric(0)
##
## $quantile
## 5%
## 0.5245457
##
## $Dep
## 1 2 3 4 5 6 7
## 2.4008470 0.6496839 2.9430730 0.8933397 2.3618842 2.7723702 3.7461218
## 8 9 10 11 12 13 14
## 1.3456888 2.5027392 3.5165780 2.9918116 1.4726572 2.5350945 3.1748282
## 15 16 17 18 19 20 21
## 3.9500035 3.8610601 3.7935308 3.4626883 2.4910748 2.9201509 0.7382501
## 22 23 24 25
## 0.9943610 2.4259651 0.3989423 3.8747116

outliers.depth.trim(FTE6, nb=200, smo=0.05, quan=0.05, dfunc=depth.RP)

## $outliers
## [1] "C24" "C2"
##
## $dep.out
## [1] 0.0448 0.0600
## $iteration
## [1] 1 2
##
## $quantile
## 5%
## 0.07356
##
## $Dep
## 1 2 3 4 5 6 7 8 9 10
## 0.3152 0.1032 0.3736 0.3208 0.3768 0.1976 0.1184 0.2560
## 11 12 13 14 15 16 17 18 19 20
## 0.1560 0.2376 0.2080 0.3752 0.3944 0.3112 0.3920 0.2608 0.3216 0.2256
## 21 22 23 24 25
## 0.2008 0.1856 0.3208 0.0448 0.4288

outliers.depth.pond(FTE6, nb=200, smo=0.05, quan=0.05, dfunc=depth.RP)

## $outliers
## character(0)
##
## $dep.out
## NULL
##
## $iteration
## numeric(0)
##
## $quantile
## 5%
## 0.0456
##
## $Dep
## 1 2 3 4 5 6 7 8 9 10
## 0.3160 0.1216 0.3440 0.1856 0.2688 0.2936 0.3576 0.1840 0.1624 0.2832
## 11 12 13 14 15 16 17 18 19 20
## 0.2080 0.2208 0.1832 0.3600 0.3960 0.3352 0.3928 0.3080 0.2880 0.2440
## 21 22 23 24 25
## 0.2224 0.1832 0.2984 0.0520 0.4208

outliers.depth.trim(FTE6, nb=200, smo=0.05, trim=0.25, quan=0.05, dfunc=depth.RT)

## $outliers
## [1] "C24"
##
## $dep.out
## [1] 0.064
##
## $iteration
## [1] 1
##
## $quantile
## 5%
## 0.068
##

5
Appendix B

R code example for type I error test:

h-Modal Depth when sample size $n = 200$
library(roahd)
## Warning: package 'roahd' was built under R version 3.5.1
library(RColorBrewer)
library(fda.usc)
## Warning: package 'fda.usc' was built under R version 3.5.1
## Loading required package: fda
## Warning: package 'fda' was built under R version 3.5.1
## Loading required package: splines
## Loading required package: Matrix
## Attaching package: 'fda'
## The following object is masked from 'package:roahd':
##   fbplot
## The following object is masked from 'package:graphics':
##   matplot
## Loading required package: MASS
## Loading required package: mgcv
## Loading required package: nlme
## This is mgcv 1.8-23. For overview type 'help("mgcv-package")'.
## Loading required package: rpart
set.seed(1)
counter=0 # number of outliers
N = 200 # Sample size
P = 50 # Number of grid
n = 100 # Number of simulation times
grid = seq( 0, 1, length.out = P ) # t
Cov = exp_cov_function(grid, alpha = 0.9, beta = 0.9 )

for (i in 1:n) {
  Data = generate_gauss_fdata( N, centerline = 50+grid^2*(1-grid)^(1/2), Cov = Cov )
  FData <- fdata(Data, argvals=NULL, rangeval=NULL, names=NULL, fdata2d=FALSE)
  depth.mode(FData, fdataori=FData, trim=0.25,
MDoutliers <- outliers.depth.trim(FData, nb=200, smo=0.05, trim=0.1, quan=0.05, dfunc=depth.mode)

# print(MDoutliers)

ABC <- as.numeric(MDoutliers$outliers)
if (length(ABC)==0) ABC<-0
else counter <- counter + length(ABC)

}
counter

## [1] 793
Appendix C

R code example for power test of contaminated sample: h-Modal Depth

when sample size $n = 200$ and $n_0 = 3$
library(roahd)
## Warning: package 'roahd' was built under R version 3.5.1
library(RColorBrewer)
library(fda.usc)
## Warning: package 'fda.usc' was built under R version 3.5.1
## Loading required package: fda
## Warning: package 'fda' was built under R version 3.5.1
## Loading required package: splines
## Loading required package: Matrix
## Loading required package: MASS
## Loading required package: mgcv
## Loading required package: nlme
## This is mgcv 1.8-23. For overview type `help("mgcv-package")`.
## Loading required package: rpart
set.seed(1)
counter1=0 # number of correct outliers detection
counter2=0 # number of incorrect outliers detection alone
counter3=0 # number of no outliers detection
counter4=0 # number of false outliers coming with correct outliers detection
N1=197 # number of normal curves
N2=3 # number of contaminated curves
N = N1+N2
P = 50
n=100 # number of simulation times
grid = seq(0, 1, length.out = P)
Cov = exp_cov_function(grid, alpha = 0.9, beta = 0.9)
for (i in 1:n) {
  Data1 = generate_gauss_fdata(N1, centerline = 50*grid^2*(1-grid)^(1/2), Cov = Cov)
  Data2 = generate_gauss_fdata(N2, centerline = 50*grid^2*(1/2)*(1-grid)^(3/2), Cov = Cov)
Data=rbind(Data1, Data2)
FData <- fdata(Data, argvals=NULL, rangeval=NULL, names=NULL, fdata2d=FALSE)
depth.mode(FData, fdataori=FData, trim=0.25, metric=metric.lp, h=NULL, scale=FALSE, draw=FALSE)
MDoutliers<- outliers.depth.trim(FData, nb=200, smo=0.05, trim=0.1, quan=0.05, dfunc=depth.mode)
# print(MDoutliers)
ABC<- as.numeric(MDoutliers$outliers)
if (length(ABC)==0) {
  counter3<-counter3+1
} else if (any(ABC==N) & any(ABC==N-1) & (any(ABC==N-2))) {
  counter1<- counter1+1; counter4<-counter4+length(ABC)-3
} else if (any(ABC==N) & any(ABC==N-1))(any(ABC==N-1) & any(ABC==N-2))(any(ABC==N-1) & any(ABC==N-2)) {
  counter1<-counter1+1; counter4<-counter4+length(ABC)-2
} else if (any(ABC==N) & any(ABC==N-1) & any(ABC==N-2)) {
  counter1<-counter1+1; counter4<-counter4+length(ABC)-1
} else {
  counter2<-counter2+length(ABC)
}

counter1
## [1] 100
counter2
## [1] 0
counter3
## [1] 0
counter4
## [1] 524
Bibliography


