ROBUST METHODS FOR FUNCTIONAL DATA ANALYSIS

Except where reference is made to the work of others, the work described in this thesis is my own or was done in collaboration with my advisory committee. This thesis does not include proprietary or classified information.

Pallavi Sawant

Certificate of Approval:

Asheber Abebe
Associate Professor
Mathematics and Statistics

Nedret Billor, Chair
Associate Professor
Mathematics and Statistics

Hyejin Shin
Assistant Professor
Mathematics and Statistics

George T. Flowers
Dean
Graduate School

ROBUST METHODS FOR FUNCTIONAL DATA ANALYSIS

Pallavi Sawant

A Thesis

Submitted to

the Graduate Faculty of

Auburn University

in Partial Fulfillment of the

Requirements for the

Degree of

Master of Science

Auburn, Alabama
August 10, 2009

Robust Methods for Functional Data Analysis

Pallavi Sawant

Permission is granted to Auburn University to make copies of this thesis at its discretion, upon the request of individuals or institutions and at their expense. The author reserves all publication rights.

Signature of Author

Date of Graduation

VITA

Pallavi Rajesh Sawant was born on the 9th of October, 1972 in Thane, Maharashtra, India. She graduated with Bachelor of Science in Statistics from University of Mumbai, Mumbai, Maharashtra, India in April 1993. She continued her Master's program in University of Mumbai and was awarded Master of Science in Statistics in 1995. Following her graduation she taught computer programming courses at computer institutes in Mumbai, India. She joined Auburn University in fall 2006 for the Masters program in the Department of Mathematics and Statistics.

THESIS ABSTRACT

ROBUST METHODS FOR FUNCTIONAL DATA ANALYSIS

Pallavi Sawant

Master of Science, August 10, 2009
(M.S., Mumbai University, 1995)
(B.S., Mumbai University, 1993)

91 Typed Pages

Directed by Nedret Billor

Functional data consist of observed functions or curves at a finite subset of an interval. Each functional observation is a realization from a stochastic process. This thesis aims to develop a suitable statistical methodologies for functional data analysis in the presence of outliers.

Statistical methodologies assume that functional data are homogeneous but in reality they contain functional outliers. Exploratory methods in functional data analysis are outlier sensitive. In this thesis we explore the effect of outliers in functional principal component analysis and propose a tool for identifying functional outliers by using robust functional principal components in a functional data. This is done by means of robust multivariate principal component analysis. Diagnostic plots based on functional principal component analysis are also found to be useful for identification and classification of functional outliers. Extensive simulation study is conducted to evaluate the performance of the proposed procedures and also real dataset is employed to illustrate the goodness of the method.

In addition, regression diagnostics for a functional regression model where regressors are functional data such as curves and the response is a scalar are discussed. We proposed a robust principal component based method for the estimation of the functional parameter in this type of functional regression model. Further we introduce robust diagnostic measures for identifying influential observations. A real dataset is also used to illustrate the usefulness of the proposed robust measures for detecting influential observations.

ACKNOWLEDGMENTS

I express my sincere gratitude to my advisor, Dr. Nedret Billor, who suggested this thesis and generously gave, continued guidance and support during the course of this work. I highly appreciate her insights and valuable suggestions for this work, without which I could not have completed this thesis. My special thanks to Dr. Hyejin Shin, who served on my committee, for the many helpful discussions and advice. I would also like to acknowledge my committee member Dr. Asheber Abebe for his valuable support.

Many thanks to the faculty, staff and fellow students at the Department of Mathematics and Statistics for making my study at Auburn such a wonderful experience. I would like to express my special gratitude to my parents for their never-ending love and support throughout my life. I feel lucky to have Rajesh as my life partner, whose companionship, and ever lasting support and commitment during all these years of studying abroad have been a source of strength and encouragement. Finally it's the grace and blessings of the God that has helped me fulfill my aspirations.

Style manual or journal used Journal of Approximation Theory (together with the style known as "aums"). Bibliograpy follows van Leunen's *A Handbook for Scholars.*

Computer software used The document preparation package TEX (specifically LATEX) together with the departmental style-file `aums.sty`.

TABLE OF CONTENTS

LIST OF FIGURES x

LIST OF TABLES xii

1 INTRODUCTION 1

2 ROBUST FUNCTIONAL PRINCIPAL COMPONENT ANALYSIS AND OUTLIER DETECTION 10
- 2.1 Introduction 10
- 2.2 Classical and Robust Principal Component Analysis 11
  - 2.2.1 Classical Principal Component Analysis 12
  - 2.2.2 Robust Principal Component Analysis 12
- 2.3 Classical and Robust Functional Principal Component Analysis and Outlier Detection 18
  - 2.3.1 Classical Functional Principal Component Analysis 18
  - 2.3.2 Robust Functional Principal Component Analysis and Outlier Detection 21

3 NUMERICAL EXAMPLES 29
- 3.1 Dataset: NOx Data 29
- 3.2 Simulation 43

4 ROBUST FUNCTIONAL PRINCIPAL COMPONENT REGRESSION 54
- 4.1 Introduction 54
- 4.2 Estimation of Functional Parameter β 55
- 4.3 Functional Regression Diagnostic measures 59
- 4.4 Numerical Example 66

5 CONCLUSION 76

BIBLIOGRAPHY 77

LIST OF FIGURES

1.1 Example of (a)Functional Observation and (b)Functional dataset . . . 2

2.1 (a)Different types of outliers when a 3 dimensional dataset is projected on robust 2 dimensional PCA-subspace. (b)Different types of outliers in Orthogonal-Score distance plot (Hubert *et al.* [22]). 28

3.1 Sample curves of NOx data by using (a) B-spline Basis and (b) Fourier Basis 30

3.2 Generalized Cross-Validation by using (a) B-spline Basis and (b) Fourier Basis for NOx data 31

3.3 Orthogonal-score plot for whole sample by using B-spline basis computed with (a)CPCA, (b)ROBPCA, (c)BACONPCA 32

3.4 Orthogonal-score plot for whole sample by using Fourier basis computed with (a)CPCA, (b)ROBPCA, (c)BACONPCA 33

3.5 Working Days of NOx data by using (a) B-spline Basis and (b) Fourier Basis 34

3.6 Orthogonal-score plot for working days by using B-spline basis computed with (a)CPCA, (b)ROBPCA, (c)BACONPCA 35

3.7 Orthogonal-score plot for working days by using Fourier computed with (a)CPCA, (b)ROBPCA, (c)BACONPCA 36

3.8 Non-working Days of NOx data by using (a) B-spline Basis and (b) Fourier Basis 37

3.9 Orthogonal-score plot for Non-working Days by using B-spline basis computed with (a)CPCA, (b)ROBPCA, (c)BACONPCA 38

3.10 Orthogonal-score plot for Non-working Days by using Fourier basis computed with (a)CPCA, (b)ROBPCA, (c)BACONPCA 39

3.11 Outliers detected by proposed method for (a)Whole sample, (b)Working days, (c)Non-working days 40

3.12 Curves generated from model 1 (without contamination), model 2 (asymmetric contamination), model 3 (symmetric contamination), model 4 (partial contamination) and model 5 (peak contamination) with n=50, p=100, M=10 and q=0.1. 51

3.13 Boxplots of norm when there is no contamination (0%) and symmetric contamination (5%, 10% and 15%) for high dimensional cases for CPCA(C) ROBPCA(R) and BACONPCA(B). 52

3.14 Boxplots of norm when there is no contamination (0%) and symmetric contamination (5%, 10% and 15%) for low dimensional cases for CPCA(C) ROBPCA(R) and BACONPCA(B). 53

4.1 Potential-Residual (P-R) plot. 64

4.2 (a)Sample curves (X-data) of the Canadian data; (b)Boxplot of the log-precipitation (y-data) data. 70

4.3 Diagnostic plot of the Candian dataset based on (a)Classical Sore plot; (b)Robust Sore plot. 70

4.4 (a)Estimated beta function by Robust and classical FPCA; (b)First eigenfunction and (c)Second eigenfunction by Robust and Classical FPCR. 71

4.5 Fitted values against the residuals (a)Classical; (b)Robust. 72

4.6 (a)Leverages; (b)Cook's Influence Measures; (c)Hadi's Influence Measures (Classical(o) and Robust(+)). 73

4.7 P-R plot (a)Classical(o); (b)Robust(+); (c)Classical(o) and Robust(+). 74

4.8 Sample curves of the Canadian data. 75

LIST OF TABLES

3.1 Outliers detected by three PCA methods. (.) denotes case number for three datasets 42

3.2 Outliers detected by Febrero *et al.* [13] for the NOx data 42

3.3 Simulation results of the Mean Proportion of Explained Variability when there is no contamination (0%) and with symmetric contamination (5%, 10% and 15%) for low and high dimensional cases. 47

3.4 Simulation results of the Mean Proportion of Explained Variability when there is no contamination (0%) and with asymmetric contamination (5%, 10% and 15%) for low and high dimensional cases. . . . 48

3.5 Simulation results of the Mean Proportion of Explained Variability when there is no contamination (0%) and with partial contamination (5%, 10% and 15%) for low and high dimensional cases. 49

3.6 Simulation results of the Mean Proportion of Explained Variability when there is no contamination (0%) and with peak contamination (5%, 10% and 15%) for low and high dimensional cases. 50

4.1 Names of the Candian weather stations 68

4.2 Classical and Robust influence measures for the Candian weather stations 69

CHAPTER 1

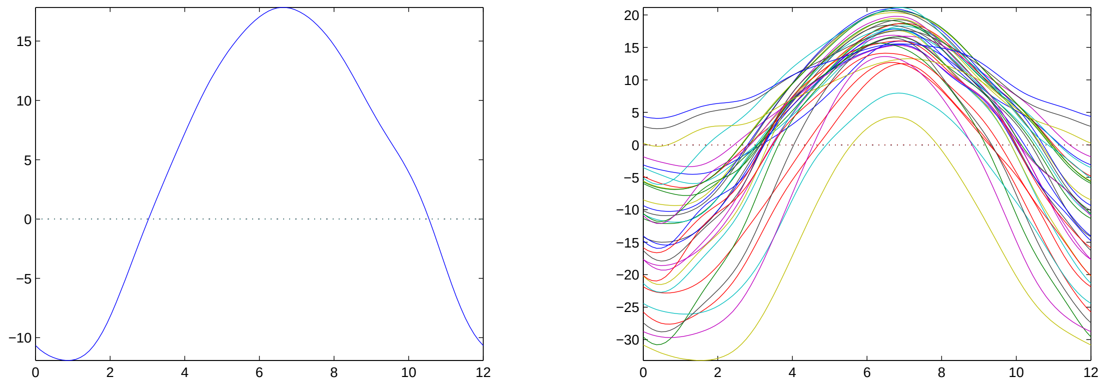
INTRODUCTION

Functional Data Analysis (FDA), comparatively new area in the statistical modeling, has become more popular in recent years. FDA is an assemblage of different methods in statistical analysis for analyzing curves or functional data. The complexity of data generated and large size of databases mandate use of new tools for analysis such as FDA [32, 26]. Functional data analysis helps to extract additional information from densely sampled observations over a time or space. In standard statistical methodology the focus is on the set of data vectors whereas, in FDA focus is on the type of data structure such as curves, shapes, images, or set of functional observations.

In FDA, each observed curve is thought of as a single observation rather than a collection of individual observations. A curve can be regarded as an infinite-dimensional vector, whose dimensions may not be countable (Figure 1.1(a),(b)).

In a traditional statistical methodology, the usual data types are univariate and multivariate. A univariate dataset contains numbers as its observations; while a multivariate dataset contains vectors as its observations. A number is one-dimensional while a vector is multi-dimensional. Multivariate Data Analysis (MDA) is an extension of Univariate Data Analysis and FDA is an extension of multivariate analysis, where the random vectors are of infinite dimensions.

In number of situations functional data can be treated as multivariate data. However, treating data directly as multivariate data may pose difficulty, such as when



(a)　(b)

Figure 1.1: Example of (a)Functional Observation and (b)Functional dataset

design points are not equal in subjects. So, direct multivariate treatment may not be possible in this case. This calls for the development of functional data analysis.

When each functional observation is sampled at a same set of design points, the functional data we get may look like multivariate data. But functional data is in general different from the multivariate data in the following aspects: 1. For a functional observation, the observed data is sampled from an underlying smooth function, whereas in a multivariate dataset for an observed vector there is no such structure. 2. The dimension of a functional observation is so large that it is regarded as a continuous function. This can be seen in Figure 1.1(a). This dimension is often larger than the sample size. 3. The time points can be different from one data point to another. All of these different aspects necessitate development of functional data analysis [39].

There are three advantages in treating data in functional forms. First, by representing data in functional form with small number of parameters reduces its size considerably. Second, since FDA deals with continuous functions; information between observed points is not lost. For finite sets of observations, FDA first estimates functions from the observed data, and then discretizes the function at any suitable choice of time points for further analysis. The free choice of analyzed points is attractive when observational points are different in each subject. Thirdly, it is very useful to have particularly interesting features in some time interval to have more closely spaced points.

In functional data framework, the random variables are defined on the functional space. To model the population of these random functions we think of a functional data observation as a realization of a stochastic process, $X(t)$, $t \in T$, where T is a bounded interval in $\Re$.

Some mathematical concepts used for FDA are explained here. In FDA, we work with a functional Hilbert space L_2 (e.g., inner product space) which is determined by an inner product $\langle x, y \rangle$ [25]. In a finite dimension with $x = (x_1, \ldots, x_n)$ and $y = (y_1, \ldots, y_n)$, the Euclidean inner product is defined in the following way:

$$\langle x, y \rangle_{\Re^n} = \Sigma x_i y_i.$$

In a functional space where $x = x(t)$ and $y = y(t)$ are functions, the L_2 inner product is defined as:

$$\langle x, y \rangle_{L_2} = \int x(t) y(t) dt,$$

where $x, y \in L_2$. For the convenience we drop the subscripts $\langle x, y \rangle_{L_2}$ and just use $\langle x, y \rangle$.

The L_2 norm is the most common type of norm, which is related to the inner product. The norm on an inner product space is defined by:

$$\| x \|^2 = \langle x, x \rangle = \int x^2(t), dt$$

and a distance between x and y is given by:

$$d(x, y) = \parallel x - y \parallel = (\langle x - y, x - y \rangle)^{1/2}.$$

Consider the functions as processes in continuous time defined over an interval, say $T \in [t_{min},\, t_{max}]$. The i^{th} replication of functional observation is denoted as $x_i(t) \in L_2[T]$, $i = 1, \ldots, n$. In practice, it is impossible to observe the functional values in continuous time. We usually obtain the data only on a finite and discrete grid $t = \{t_1, t_2, \ldots, t_p\} \in T$ in the following manner:

$$y_i = x_i(t) + \epsilon_i, \qquad 1 \leq i \leq n,$$

where ϵ_i is a random error or noise with zero mean and variance function $\sigma_i^2(t)$. For simplicity, we assume that all processes are observed at the same time points, which are equally spaced on T and is denoted by $t = \{t_1, t_2, \ldots, t_p\}$, but in reality $t_j \in T$ can be different, i.e., t_{ij} depending on i where $1 \leq j \leq p_i$.

Functional Basis Expansion

For any data analysis in the FDA framework first step is functional data smoothing. It is done to convert raw discrete data points into smooth functions i.e., to convert data to functional form. Smoothing method is used to minimize noise in raw data for calculations and analysis. There are different types of smoothers that can be applied to functional data. In this thesis we use smoothing based on basis-function method. By the use of basis function discrete data is represented as a smooth function this is also known as functional data smoothing. In the basis expansion method; the function x_i can be represented as a linear combination of first k known basis functions

ϕ_K, $K = 1, \ldots, k$, where k is large enough, $k < p$. In this approach, a functional observation x_i is expressed as:

$$x_i(t) = \sum_{K=1}^{k} c_{iK}\phi_K(t),$$

where ϕ is vector-valued function having components $\phi_1, \ldots, \phi_k$. The C is $n \times k$ coefficient matrix of the expansion, where $C = [c_{iK}]$, $1 \leq i \leq n$, $1 \leq K \leq k$. The simultaneous expansion of all n curves can be expressed in matrix notation as:

$$x = C\phi,$$

where x is a vector-valued function with x_i, $1 \leq i \leq n$, as its components. This approach is preferred since it makes good approximation of the data with a relatively small number of parameters. This may be considered to be a dimension reduction operation.

There are many basis functions possible. Fourier and B-spline bases are most frequently used bases functions which are summarized below:

Fourier basis system

This is the best known basis expression for the periodic data which is based on Fourier series. The Fourier basis expansion is given by:

$$\hat{x}(t) = c_0 + c_1 sin(wt) + c_2 cos(wt) + c_3 sin(2wt) + c_4 cos(2wt) + \ldots,.$$

Here,

$$\phi_j(t) = 1, \quad j = 0$$
$$\phi_j(t) = sin(rwt), \quad j = 2r - 1$$

$$\phi_j(t) = cos(rwt), \quad j = 2r.$$

The period and the length of the interval $|T| = 2\pi/w$ is determined by the frequency w. The Fourier basis is said to be orthogonal if the values of t_j are equally spaced on T and the period is equal to the length of T. Due to orthogonal property the cross-product matrix $\phi'\phi$ is diagonal, and can be made equal to the identity by dividing the basis function by suitable constants $n^{1/2}$ for $j = 0$ and $(n/2)^{1/2}$ for all j. This basis is well known partially due to the Fast Fourier Transformation (FFT) Algorithm, which makes it possible to compute the coefficients speedily and efficiently.

B-Spline basis system

B-Spline basis is a well-known functional basis for non-periodic data. The interval T on which basis is defined is divided into L subintervals separated by values τ_l $l = 1, \ldots, L-1$ called breakpoints or knots. A spline function is determined by two quantities: the order of the B-spline and the number of knots. A spline is piecewise polynomial function of order m over each interval, which is smoothly connected at breakpoints. The notation $B_K(t, \tau)$ of the B-spline function defined by the breakpoint sequence τ indicate the value at t and K refers to the number of the largest knot at or to the immediate left of value t.

The notation $\phi_K(t)$ is B-spline of order m and a knot sequence τ given by:

$$\phi_K(t) = B_K(t, \tau),\ K = 1, \ldots, m + L - 1.$$

The advantage of this basis function is that they are flexible and fast.

Since no basis is universally good, choosing one is a complex issue. However, there are guidelines for specific situations as each candidate function for the basis has the unique characteristics; for example if the data are periodic then a Fourier basis is used and for non-periodic data or data that have a lot of local features B-spline works better. The selection of the basis function $\phi_K(t)$ is done by observing the data. Selecting proper order of expression k (the number of basis functions) is important question in the basis expansion. There are many ways to decide the number of basis functions like Cross-Validation (CV), Generalized Cross Validation (GCV) or other similar criteria. In this thesis we use GCV developed by Craven and Wahba [9], which is described in Chapter 2.

Since $X(t)$ is a random function the mean, variance, covariance and correlation of $X(t)$ are defined as:

$$\begin{gathered}
VarX(t) = E[X(t) - EX(t)]^2, \qquad t \in T \\
CovX(s,t) = E[X(s) - EX(s)][X(t) - EX(t)], \qquad s,t \in T \\
CorrX(s,t) = \tfrac{CovX(s,t)}{\sqrt{[VarX(s)VarX(t)]}}, \qquad s,t \in T
\end{gathered}$$

where $VarX(s)$, $VarX(t) > 0$.

The functional sample descriptive statistics, where we have an n-dimensional subspace for the sample space, $x_i(t)$, $i = 1, \dots, n$ are also be defined as:

$$\begin{gathered}
\overline{x}(t) = \tfrac{1}{n}[\sum_{i=1}^{n} x_i(t)], \qquad t \in T \\
\widehat{Varx}(t) = \tfrac{1}{n-1}[\sum_{i=1}^{n} [x_i(t) - \overline{x}(t)]^2], \; t \in T \\
\widehat{Covx}(s,t) = \tfrac{1}{n-1}[\sum_{i=1}^{n} [x_i(s) - \overline{x}(s)][x_i(t) - \overline{x}(t)]], \qquad s,t \in T \\
\widehat{Corrx}(s,t) = \tfrac{\widehat{Covx}(s,t)}{\sqrt{[\widehat{Varx}(s)\widehat{Varx}(t)]}}, \qquad s,t \in T.
\end{gathered}$$

Majority of statistical techniques used in traditional and functional data analysis assume that the dataset is free of outliers. However, outliers occur very frequently in real data. Outlier is defined as a data point appearing to be inconsistent with the rest of the data. Possible sources of outliers are errors in recording and measurement, incorrect distribution assumption, unknown data structure, or novel phenomenon [23]. In any statistical data analysis, investigation of outliers is important. Since traditional statistical methods are sensitive to outliers, presence of outliers in a dataset make estimators and statistical conclusions unreliable. In addition presence of outliers severely affects modeling and prediction.

High dimensional data occurrence is natural in some practical applications such as studies involving image analysis and microarray datasets in genomic studies. In such applications dimension p is greater than n, sample size. The FDA analysis and high-dimensionality are closely related as the functional data generally come in a discretized manner so that a function x_i in the sample is in fact given by $(x_i(t_1), \ldots, x_i(t_p))$. High dimensionality problem has two distinct features: first the dimension p depends on the descretization order. This is not given in advance and can be arbitrarily increased. Second, the data from the discretized functions likely to be highly correlated thus creating difficulty in estimation of the covariance matrices.

Functional Principal Component Analysis (FPCA) is a useful statistical technique for understanding the structure of data. They are effective dimension reduction tools for functional data. FPCA aims to explain the covariance structure of data by means of small number of functional components. These functional components are linear combinations of the original variables. This gives better interpretation of the different sources of variation. Thus effectiveness of FPCA in data reduction is useful in analysis of high dimensional data. In the presence of outliers, dimension reduction via FPCA would yield untrustworthy results since FPCA is known to be sensitive to outliers.

The main contribution of this work is the construction of the method to detect outliers in functional data through robust FPCA. We have used library developed by Ramsay and Silverman [31] to construct our code for functional data analysis. This thesis is organized as follows. Chapter 2 reviews functional principal component analysis and explore sensitivity of FPCA to outliers. We also propose a functional outlier detection procedure based on robust multivariate techniques. We have described accompanying diagnostic plots that can be used to detect and classify possible outliers. Chapter 3 consists of numerical examples, real dataset and simulation study for the proposed procedure. Chapter 4 reviews the functional linear model with scalar response. In this chapter we propose a robust technique of parameter function estimation based on the robust functional principal components method. We also introduce robust diagnostic measures for identifying influential observations. Finally, in Chapter 5, we give the conclusions of this work.

CHAPTER 2

ROBUST FUNCTIONAL PRINCIPAL COMPONENT ANALYSIS AND OUTLIER DETECTION

2.1 Introduction

In various areas such as chemometrics, biometrics, engineering, genetics, and e-commerce the data come from the observation of continuous phenomenons of time or space known as functional data. Due to advancement of new techniques it is now possible to record large number of variables simultaneously. The nature of this data in many applications is high dimensional where the number of variables (p) is greater than the number of observations (n) ($n \ll p$). The focus of researchers is on analysis of such data due to the emergence of statistical problems while applying various statistical tools for data analysis. Functional principal component analysis (FPCA) is a useful tool to reduce the dimension of the functional data.

In functional data, the first step is to represent the data in a lower dimensional space in order to have better interpretation. This is done by performing FPCA to capture the main modes of variability of the data by means of small number of components which are linear combinations of original variables that allow for better interpretation of various sources of variation.

Sensitivity of the variance and the covariance matrix to irregular observations make it vulnerable to outliers and may not capture the variation of the regular observations. Therefore data reduction based on FPCA becomes unreliable. This necessiates the need of the robust FPCA method.

Lacontore *et al.*[27], has proposed a robust functional principal component analysis based on spherical (SPHER) and elliptical (ELL) PCA [27]. However, these methods have two drawbacks. First drawback is that SPHER and ELL only estimate the principal components and not their eigenvalues. Second drawback is that SPHER and ELL PCA are influenced by outliers [22]. Febrero *et al.* [13, 14] also proposed two methods for outlier detection that are based on the idea of functional depths and distance measures.

The main contribution of our work is to construct a robust PCA method to achieve dimension reduction of data and to develop tools for detection of functional outliers.

The outline of this chapter is as follows. In Section 2.2, a brief description of classical principal component analysis (CPCA) method and robust PCA methods are given. The main concepts of PCA which are used in FPCA are discussed in Section 2.3. In this section, the outlier detection procedure via robust FPCA is also described.

2.2 Classical and Robust Principal Component Analysis

Principal Component Analysis (PCA) is used for understanding the structure of a multivariate dataset. PCA is a useful tool for data reduction, which is achieved by identifying main modes of variability of a given dataset. Unfortunately, if the data contains outliers then data reduction based on classical principal component analysis becomes unreliable. The goal of the robust PCA methods is to obtain principal components that are not affected by outliers.

2.2.1 Classical Principal Component Analysis

In multivariate data the central notion is to find weight vectors $\gamma_j \in \Re^p$ for which a linear combination of centered variable values

$$Z_i = \sum_{j=1}^{p} \gamma_j x_{ij} = \gamma' x_i \qquad i = 1, \ldots, n \tag{2.1}$$

that have maximal variance subject to constraints $\gamma'_m \gamma_r = \mathcal{I}(m = r)$ for $m < r$, where $\gamma = [\gamma_1, \ldots, \gamma_p]'$ and $x_i = [x_{i1}, \ldots, x_{ip}]'$. The solution is obtained by the means of spectral decomposition of the variance-covariance matrix [24, 31].

2.2.2 Robust Principal Component Analysis

In this Section three robust methods for PCA for multivariate data are reviewed. These are 1. Elliptical Principal Component Analysis, 2. Robust Principal Component Analysis (ROBPCA), 3. Blocked Adaptive Computationally Efficient Outlier Nominators Principal Component Analysis (BACONPCA).

Elliptical Principal Component Analysis

In order to find the robust principal directions, Spherical (SPHER) Principal Component Analysis and Elliptical (ELL) Principal Component Analysis, developed by Lacontore *et al.*[27], are used. In this method: the first step is to find the robust estimate of the center of the population. This is done by considering the spatial median which is given by:

$$\hat{\theta} = \arg\max_{\theta} \sum_{i=1}^{n} \parallel x_i - \theta \parallel_2,$$

where $\| \cdot \|_2$ denotes the Euclidean norm on $\Re^p$ and $x_1, \ldots, x_n \in \Re^p$. To overcome the problem of high dimensionality SPHER PCA, which is a robust version of PCA, is performed by projecting the data on the unit sphere in $\Re^p$. For coordinates measured in different scales, ELL PCA is adopted by first scaling the data through robust scale estimates. Prescaling is done by the Median Absolute Deviations (MAD). By doing this the vertical axis is compressed and the bulk of the data look like a sphere. Then the data is projected on this scale. After the projected data is rescaled, which lie on an ellipse, classical PCA is performed.

The SPHER and ELL PCA methods facilitate fast algorithm for performing robust PCA. SPHER and ELL only estimate the principal components and do not calculate estimates of the eigenvalues, so computing score distances is not possible. Hubert *et al.* [22] showed that SPHER and ELL PCA are influenced by outliers when the data are high-dimensional or when there is a large percentage of contamination in the data. In such instances these methods convert the bad leverage points into good leverage points and orthogonal outliers.

ROBPCA and BACONPCA

Let the original data be $n \times p$ matrix $X = X_{n,p}$, where n denotes the number of observations and p denotes the number of variables. ROBPCA [20, 21, 22] is based on the minimum covariance determinant (MCD) estimator [35, 36] of multivariate location vector and scatter matrix. BACONPCA [3] is based on the estimator of the location vector and scatter matrix obtained from the basic subset which is found algorithmically by utilizing BACON approach [2]. We will describe these two methods in details for low and high dimensional cases.

ROBPCA for low dimensional data ($n > p$)

Minimum covariance determinant (MCD) estimator [35, 36] of multivariate location and covariance matrix are popular for this case because of its high resistance to outliers. It also provides fast algorithm (FAST-MCD) [37] for computation.

In this algorithm: the initial MCD estimators are defined, based on h observations ($h < n$), as the mean $\hat{\mu}_0$ and the covariance matrix $\hat{\Sigma}_0$. The covariance matrix of these h observations has the lowest determinant and h should be at least $[(n+p+1)/2]$. MCD estimator can resist $n-h$ outliers and with this choice the MCD estimator has a breakdown value of $(n-h+1)/n$. The value of h is taken approximately between $0.5n$ and $0.75n$. The value $h \approx 0.5n$ is taken when there is 50% contamination and if there is 25% contamination then the value of $h \approx 0.75n$. When there are smaller number of outliers the value of h is increased for a more precise estimates.

Reweighting is then done to increase the finite sample efficiency. Each data point receives a weight 1 if its robust distance $RD(x_i) = \sqrt{(x_i - \hat{\mu}_0)'\hat{\Sigma}_0^{-1}(x_i - \hat{\mu}_0)}$ is $\leq \sqrt{\chi^2_{p,0.975}}$ and weight 0 otherwise. For the observations with weight one the reweighted MCD estimator is then defined as the classical mean $\hat{\mu}_M$ and covariance matrix $\hat{\Sigma}_M$. The robust loadings are the first k_1 eigenvectors of the MCD estimator of $\hat{\Sigma}_M$ sorted in descending order of the eigenvalues [20, 21, 22].

ROBPCA for high dimensional data ($p > n$)

For data with high-dimension ($p > n$), the MCD estimator can not be used because the covariance matrix of $h < p$ observations is always singular and can not be minimized. In this case ROBPCA method suggested by Hubert *et al.* [20, 21, 22] is used on the $X_{n,p}$ data. ROBPCA method is a combination of both projection pursuit technique (PP)[19] and MCD covariance estimation in lower dimensions. PP is used first to reduce dimension. The MCD method is then applied to this low dimensional subspace to estimate the center and the scatter of the data.

Initial data preprocessing is done by applying singular value decomposition of $X_{n,p}$. This results in huge dimension reduction as the data are represented using at most $n - 1 = rank(\tilde{X}_{n,p})$ variables without loss of information.

By applying PP, the high dimensional data points are projected on many univariate directions. Then the robust center $\hat{\mu}_r$ and scale $\hat{\sigma}_r$ (based on univariate MCD method) of these projected data points on every direction are computed. For each projected data point, the Stahel-Donoho's outlyingness measure:

$$Outl(z_j) = max_v \frac{\mid z_i' v - \hat{\mu}_r \mid}{\hat{\sigma}_r}, \quad i = 1, \ldots, n.$$

is used to form h subset, that has smallest outlyingness. Optimal $k_0 \ll p$ principal components are then selected from the covariance matrix of the final h subset. The data are then projected onto this k_0 dimensional subspace. Next the reweighted MCD estimator is used to compute the center and the scatter of the data points in this low-dimensional subspace. The dominant k_1 eigenvectors of this covariance matrix are the k_1 robust principal components, and the MCD location and covariance matrix estimates serve as robust estimates for the location vector μ and covariance matrix σ.

BACONPCA in low dimensions ($n > p$)

The Blocked Adaptive Computationally Efficient Outlier Nominators (BACON) algorithm developed by Billor *et al.* [2] is used for this robust procedure. BACON is a cost effective, fast computational method with high breakdown point based on measuring robust distances from a basic subset, which is free of outliers. The initial basic subset is derived algorithmically in two ways by Mahalanobis distances or by Euclidean distances.

For BACONPCA in low dimensions, initial dimension reduction of the mean centered data matrix X_c is done by singular value decomposition (SVD).

$$X_c = (X - 1_n\hat{\mu}') = UD\Gamma',$$

where $\hat{\mu}$ is the classical mean vector, D is a $p \times p$ diagonal matrix of the eigenvalues of the $X_c'X_c$ and $U'U = I_p = \Gamma'\Gamma$, Γ is the matrix of the eigenvectors corresponding to the eigenvalues of $X_c'X_c$. I_p is the $p \times p$ identity matrix. The next step is to obtain the score matrix, $Z = X_c\Gamma$. The robust mean, $\hat{\mu}_B$ and the robust variance-covariance matrix, $\hat{\Sigma}_B$, are computed from clean observations obtained from BACON algorithm. From the BACON based covariance matrix, $\hat{\Sigma}_B$, number of robust PCs are determined as k_1. Γ_1 is the matrix of the eigenvectors corresponding to the nonzero eigenvalues of $\hat{\Sigma}_B$. Finally, robust score matrix, $Z_1 = (Z - 1_n\hat{\mu}_B')\Gamma_1$, is obtained [3].

BACONPCA in high dimensions ($p > n$)

In this case BACONPCA method, suggested by Billor *et al.* [3], is used on the centered X_c data. The mean centered data matrix X_c are preprocessed by singular value decomposition (SVD) based on the eigenvalues and the eigenvectors of X_cX_c' instead of $X_c'X_c$. Since decomposition of X_cX_c' is much faster than that of $X_c'X_c$. Then the score matrix $Z = X_c\Gamma$ is obtained, where Γ is the matrix of the eigenvectors corresponding to the eigenvalues of X_cX_c'.

Since BACON or MCD methods are useful only when $n > p$, these methods cannot be used, where $n < p$, to determine clean observations of Z because of singularity of the covariance matrix. Stahel-Donoho's outlyingness measure is useful to determine a clean set of h observations of Z (Hubert *et al.* [22]). The high dimensional data points, z_i, are projected onto many univariate directions v. For every direction v, robust center μ_r and robust standard deviation, $\hat{\sigma}_r$ (based on univariate BACON method) are obtained for the projected observations, $z_i'v$ $(i = 1, \ldots, n)$. Outlyingness measure based on these robust center and scale values can be defined as:

$$Outl(z_j) = max_v \frac{\mid z_i'v - \hat{\mu}_r \mid}{\hat{\sigma}_r}, \quad i = 1, \ldots, n.$$

This measure will detect the points which are outlying on the projection vector. Therefore, this will result into h clean set of observations (h=0.75n). For h observations the mean, $\hat{\mu}_1$, and the scatter matrix, $\hat{\Sigma}_1$, of the Z matrix are obtained. The spectral decomposition of $\hat{\Sigma}_1$, gives:

$$\hat{\Sigma}_1 = \Gamma_1 \Lambda_1 \Gamma_1',$$

where Γ_1 is the matrix of the eigenvectors corresponding to the eigenvalues of $\hat{\Sigma}_1$, $\Lambda = diag(\lambda_1, \ldots, \lambda_p)$ is the diagonal matrix of the eigenvalues of $\hat{\Sigma}_1$. Then we determine the retaining number of principal components $k_0 < p$ by using some techniques, like a scree plot. The data are then projected onto the subspace spanned by the first k_0 eigenvectors of the covariance matrix $\hat{\Sigma}_1$, that is

$$Z_2 = (Z_{n \times p} - 1_n \hat{\mu}_1') \Gamma_{p \times k_0}$$

where $\Gamma_{p\times k_0}$ is a matrix of the first k_0 eigenvectors Γ_1. Next, BACON algorithm is applied to find the mean vector, $\hat{\mu}_B$, and scatter matrix, $\hat{\Sigma}_B$, of the matrix, Z_2. Based on the robust covariance matrix, $\hat{\Sigma}_B$, the robust PCs are obtained as:

$$Z_3 = (Z_2 - 1_n\hat{\mu}'_B)\Gamma^*_{p\times k_1},$$

where $\Gamma^*_{p\times k_1}$ is the matrix of eigenvectors corresponding to the first k_1 eigenvalues, that are determined by a selection criteria (e.g., a scree plot), of the robust BACON based covariance matrix $\hat{\Sigma}_B$ [3].

2.3 Classical and Robust Functional Principal Component Analysis and Outlier Detection

When the dataset is in the form of a curve, the procedure for PCA can be generalized for functional principal component analysis (FPCA) to obtain main modes of variability for the curves.

2.3.1 Classical Functional Principal Component Analysis

When the dataset is in the form of a curve, the procedure for classical PCA can be generalized for Functional Principal Component Analysis (FPCA) to obtain main modes of variability for the curves. Instead of variable values x_{ij}, used in PCA, functional values $x_i(t)$ are used in FPCA, so that the discrete index j in the multivariate context is replaced by continuous index t. Unlike multivariate PCA, components in functional PCs are functions rather than vectors. So summations over j are replaced by integrations over t.

Let $\{x(t), t \in T\}$ be a stochastic process where T is some index set which is a bounded interval on $\Re$. The principal component scores corresponding to weight γ is generalized to an integral form,

$$Z_i = \int \gamma_j(t) x_i(t) dt. \tag{2.2}$$

The weight function $\gamma_j(t)$ is obtained by solving

$$\max_{\langle \gamma_\mathbf{j}, \, \gamma_\mathbf{m} \rangle = \mathcal{I}(\mathbf{j}=\mathbf{m}), \, \mathbf{j} \leq \mathbf{m}} N^{-1} \sum \left(\int \gamma_j x_i \right)^2 \tag{2.3}$$

or equivalent to solving the functional eigenequation

$$\int \psi(s,t) \gamma(t) dt = \lambda \gamma(s), \qquad \gamma \in L_2, \tag{2.4}$$

where ψ is the covariance function of the $x(t)$. The sequence of eigenfunctions γ_i, $i = 1, 2, \ldots$, sorted with respect to the corresponding eigenvalues $\lambda_1 \geq \lambda_2 \geq \ldots$ solves the FPCA problem (2.3). The eigenequation is the same general equation as in PCA, except here γ is now an eigenfunction rather than an eigenvector. There is a major difference between the multivariate and functional eigenanalysis. In multivariate case the eigenvalue-eigenfunction pairs are p (number of variables) whereas, in functional case they are infinite (number of functional values). In practice, the unknown covariance function ψ needs to be estimated by the sample values $x_i(t)$, $1 \leq i \leq n$, where for each i, $x_i(t)$ is observed on a discrete set of points $t = \{t_1, \ldots, t_p\}$ for finite p.

FPCA problem can be represented in terms of basis function approach. In which, first k bases functions in a basis $\{\phi_1, \ldots, \phi_k\}$ are used, where k is large enough, so that these functions will be able to describe most of the features of the data. The bases are selected based on the nature of the data; for example if the data are smooth and periodic then a Fourier basis might be ideal and for data that have a lot of local features then B-splines might work better. Approximate each x_i by:

$$\hat{x}_i(t) = \sum_{K=1}^{k} c_{iK}\phi_K(t). \tag{2.5}$$

We can express all n curves simultaneously by defining the vector-valued function x to have components $x_1, x_2, \ldots, x_n$ and the vector valued function ϕ to have components $\phi_1, \ldots, \phi_k$ as:

$$x = C\phi, \tag{2.6}$$

where the coefficient matrix C is $n \times k$. In matrix terms the variance-covariance function is:

$$\psi(s,t) = n^{-1}\phi(s)'C'C\phi(t). \tag{2.7}$$

Define W as a symmetric matrix of order k

$$W = \int \phi\phi'. \tag{2.8}$$

Suppose that the weight function γ has the expansion

$$\gamma(s) = \sum b_K\phi_K(s) \tag{2.9}$$

and in matrix notation, $\gamma(s) = \phi(s)'b$. Using equations (2.6-2.9) the left hand side of eigen equation (2.4) becomes

$$\int \psi(s,t)\gamma(t)dt = \int n^{-1}\phi(s)'C'C\phi(t)\phi(t)'bdt$$

$$= \phi(s)'n^{-1}C'CW'b.$$

The eigenequation can be written as:

$$\phi(s)' n^{-1} C' C W b = \lambda \phi(s)' b. \tag{2.10}$$

As this equation holds true for all s, it can be written in matrix form in following manner:

$$n^{-1} C' C W b = \lambda b. \tag{2.11}$$

As $\| \gamma \| = 1$ implies $b' W b = 1$ and similarly, two functions γ_1 and γ_2 will be orthogonal if and only if the corresponding vectors of coefficients satisfy $b_1' W b_2 = 0$. We define $u = W^{1/2} b$ to get the required principal components by solving equivalent symmetric eigenvalue problem

$$n^{-1} W^{1/2} C' C W^{1/2} u = \lambda u \tag{2.12}$$

and compute $b = W^{-1/2} u$ for each eigenvector. If the basis is orthonormal then $W = I$. The functional PCA problem reduces to the standard multivariate PCA of the coefficient array C.

In this section, we examined FPCA as a dimension reduction tool. Although, FPCA solves dimensionality problem, it fails to deal with data containing outliers. In next section a new robust method, robust FPCA method, is given to overcome this problem.

2.3.2 Robust Functional Principal Component Analysis and Outlier Detection

In this section, an outlier detection method for functional data via robust FPCA is given. The robust FPCA method is to obtain functional principal components that are less influenced by outliers. Outlier detection method proposed by Febrero *et al.*[13] is discussed and then the construction of the robust FPCA method is described.

Outlier Detection using Likelihood Ratio Test

Outlier detection procedure in functional data using Likelihood Ratio Test (LRT) is developed by Febrero *et al.*[13], which is based on distance measure. Let the functional sample be $x_1, \ldots, x_n$ and the statistic is given as:

$$O_\alpha(x_i) = \| \frac{x_i - \hat{\mu}_{TM,\alpha}}{\hat{\sigma}_{TSD,\alpha}} \|,$$

$$\Lambda = \max_{1 \leq i \leq n} O_\alpha(x_i),$$

where $\| \cdot \|$ is a norm in the functional space ($\| \cdot \|_1$, $\| \cdot \|_2$ or $\| \cdot \|_\infty$), $\hat{\mu}_{TM,\alpha}$ is the α-trimmed mean and $\hat{\sigma}_{TSD,\alpha}$ is the α-trimmed standard deviation. Hence, $O_\alpha(x_i)$ is the distance between x_i and $\hat{\mu}_{TM,\alpha}$ relative to $\hat{\sigma}_{TSD,\alpha}$. The presence of outliers is determined by comparing the test statistic (Λ) with some threshold and an iterative procedure.

Description of the Proposed Method

Consider the functions as processes in continuous time defined over an interval, say $T \in [t_{min},\ t_{max}]$. The i^{th} replication of functional observation is denoted as $x_i(t) \in L_2[T]$, $i = 1, \ldots, n$. In practice, it is impossible to observe the functional values in continuous time. We usually obtain data only on a finite and discrete grid

$t = \{t_1, t_2, \ldots, t_p\} \in T$ in the following manner:

$$y_i = x_i(t) + \epsilon_i, \qquad 1 \leq i \leq n,$$

where ϵ_i is a random error or noise with zero mean and constant variance function $\sigma_i^2(t)$. For simplicity, we assume that all processes are observed at the same time points, which are equally spaced on T and is denoted by $t = \{t_1, t_2, \ldots, t_p\}$.

The function x_i can be represented as a linear combination of the first k basis functions ϕ_K, $K = 1, \ldots, k$, where k is large enough, $k < p$ using basis expansion method given in Chapter 1 as:

$$x_i(t) = \sum_{K=1}^{k} c_{iK}\phi_K(t),$$

where ϕ is vector-valued function having components $\phi_1, \ldots, \phi_k$. The C is n x k coefficient matrix of the expansion, where $C = [c_{iK}]$, $1 \leq i \leq n$, $1 \leq K \leq k$. The simultaneous expansion of all n curves can be expressed in matrix notation as:

$$x = C\phi,$$

where x is a vector-valued function with x_i, $1 \leq i \leq n$, as its components.

We assume that basis function is orthonormal. To select optimal number of basis functions, k, GCV developed by Craven and Wahba [9], is used which is described in the following section.

Coefficient Estimation: On partially observed functions the coefficients c_{iK} are computed by using the least squares approach, for $i = 1, \ldots, n$ and $K = 1, \ldots, k$,

$$\sum_{i=1}^{n}[y_i(t) - \sum_{K=1}^{k} c_{iK}\phi_K(t)]^2$$

$$= (y - C\phi)'(y - C\phi)$$

$$= \parallel y - C\phi \parallel^2 .$$

Since we deal with basis function that is orthonormal the functional PCA problem reduces to the standard multivariate PCA of the coefficient array C (see Section 2.3.1). Instead of dealing with FPCA we apply classical PCA on C. Applying PCA on C would provide the equivalent information about the structure of the covariance function of functional data $x(t)$. Outliers in C will be equivalent to the outliers in functional data $x(t)$. Therefore, the diagnostic plots developed to detect outliers by using multivariate PCA method can also be used to detect functional outliers. Diagnostic plot, Orthogonal-score plot [22], which is a by-product the robust PCA method is used for identification and classification of outliers. By using PCA method we obtain robust scores Z in the following manner:

$$Z = C \times V,$$

where Z is $n \times k_1$ matrix, C is $n \times k$ matrix of the coefficients, V is $k \times k_1$ robust eigenvectors and $k_1 \leq k$. The selection criteria to choose the components k_1 is based on the eigenvalues. The predetermined threshold value is 90%. The optimal number of components k_1 is the minimal value for which the cumulative percentage of total variance is greater than or equal to 90%.

Robust coefficients are obtained by transforming the data back to $\Re^k$ as:

$$\hat{C} = Z \times V'.$$

Finally we obtain functional data which are robust in following way:

$$\hat{x} = \hat{C} \times \phi.$$

Selecting Number of Basis

Selecting optimal number of bases, k, is important because if k is too large it may introduce small variation with large bias and if k is too small then we may miss some aspects of smooth function x that we want to estimate. This will also introduce less bias with large variance. To choose the appropriate number of basis functions a popular measure in the smoothing methods known as generalized cross validation (GCV) developed by Craven and Wahba [9] is used. This criterion is defined as:

$$k = \arg\min_{\mathbf{j}}(GCV(j)),$$

where

$$GCV(j) = \frac{n \times SSE}{(n-j)^2}, \quad j = 3, \ldots, p-1,$$

$$SSE = \sum_{i=1}^{n}(x_i - \hat{x}_i)^2, \quad \hat{x}_i = \sum_{K=1}^{k} c_{iK}\phi_K.$$

There is another technique cross-validation (CV) based on minimizing mean squared error (MSE). Minimizing CV can lead to under-smoothing the data by introducing large variation. However, GCV has advantage over CV technique as it has less tendency to undersmooth the data.

The choice of number of bases relies on $\hat{x}_i$. The coefficients $C = [c_{iK}]$ are computed by using least squares method and then $\hat{x}_i$ are estimated. Since, least squares method is sensitive to outliers, the choice of number of bases is also affected by outliers. Robust version of this criteria for selecting number of bases can be obtained by estimating the coefficients robustly.

Diagnostic Plot for Detection of Outliers

The diagnostic plot developed to detect outliers by using PCA method for multivariate data can be used to detect functional outliers. Outliers in C will be equivalent to the outliers in $x(t)$ functional curves. Orthogonal score plot proposed by Hubert *et al.* [22] is used to distinguish between regular observations and the three types of outliers. This diagnostic plot is a scatter plot of the orthogonal distance Od_i versus the robust score distance Sd_i. The score distance is defined as:

$$Sd_i = \sqrt{\sum_{j=1}^{k_1} z_{ij}^2/\lambda_j}, \quad i = 1, \ldots, n,$$

where z_{ij} are the scores and λ_j are the eigenvalues. The orthogonal distance which measures the distance between an observation x_i and its projection in the k_1-dimensional PCA-subspace, Od_i, is given by:

$$Od_i = \| x_i - \hat{\mu} - \Gamma_{p,k_1} z_i' \|, \quad i = 1, \ldots, n,$$

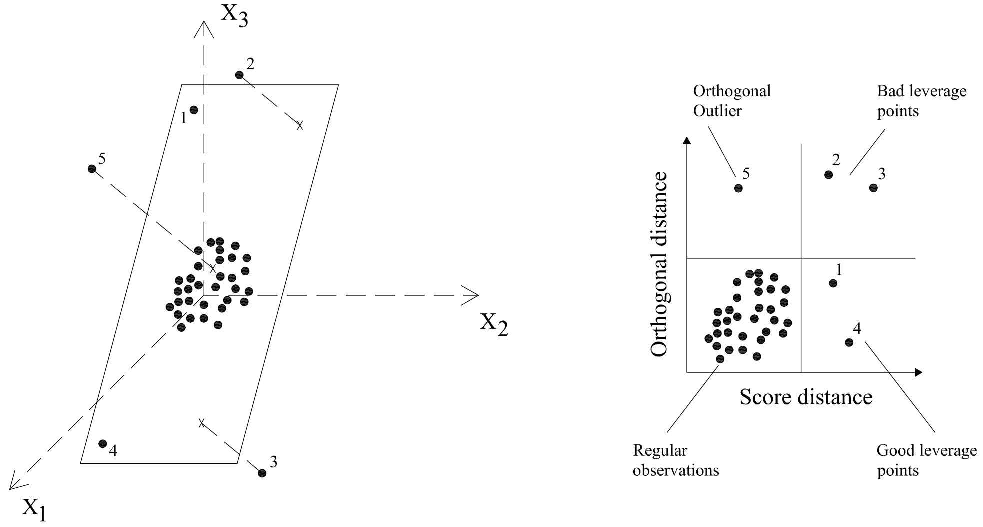
where Γ_{p,k_1} is the $p \times k_1$ matrix of eigenvectors and z_i' is the i^{th} row of the score matrix.

If Sd_i is large and Od_i is small, then the i^{th} observation is far away from the homogeneous observations and close to the PCA space (e.g., Observations 1 and 4 in Figure 2.1(a),(b)). If Od_i is large and Sd_i is small, then the i^{th} observation is far away from the PCA space orthogonally (e.g., Observation 5 in Figure 2.1(a),(b)). If Sd_i

and Od_i are both large, then the i^{th} observation is far away from the homogeneous observations and the PCA space (e.g., Observations 2 and 3 in Figure 2.1(a),(b)).

Therefore observations can be classified in 1) Homogeneous observations (close to PCA and not far away from the remaining data), 2) Orthogonal outliers, 3)Good Leverage points, and 4) Bad Leverage points (See Figure 2.1).

Two cutoff lines are used to classify the observations. The cutoff value horizontal line is $\sqrt{\chi^2_{k,0.975}}$ when $k > 1$ and $\pm\sqrt{\chi^2_{k,0.975}}$ when $k = 1$. For the cutoff value of orthogonal distances the Wilson-Hilferty approximation for a χ^2 distribution is used, where the $Od^{2/3} \sim N(\mu, \sigma^2)$. Estimates of $\hat{\mu}$ and $\hat{\sigma}^2$ are obtained by using univariate MCD. The cutoff value of the vertical line equals $(\hat{\mu} + \hat{\sigma} z_{.975})^2$ where $z_{.975}$ is 97.5% quantile of the normal distribution.



(a) (b)

Figure 2.1: (a)Different types of outliers when a 3 dimensional dataset is projected on robust 2 dimensional PCA-subspace. (b)Different types of outliers in Orthogonal-Score distance plot (Hubert *et al.* [22]).

CHAPTER 3

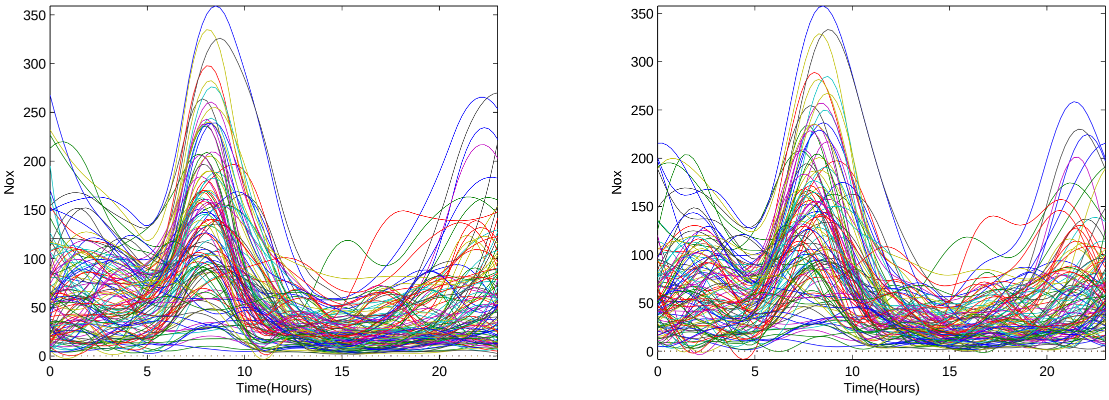
NUMERICAL EXAMPLES

In this chapter a real data and simulation study are given to demonstrate optimality of the proposed method for outlier detection in FDA.

3.1 Dataset: NOx Data

The aim of our analysis is to illustrate the performance of the robust FPCA on the NOx data, which was used by Febrero *et al.* [13, 14]. The NOx emission levels data collected by a control station near a power plant in Barcelona in year 2005 is analyzed by using techniques for functional data. The dataset consists of NOx levels ($\mu g/m^3$) measured every hour for the period February 23, 2005 to June 26, 2005. Only NOx levels for 115 days are available due to missing observations problem for several consecutive hours of some days. The dataset of NOx emission levels are displayed in Figure 3.1.

The whole NOx data is divided into working days ($n_1 = 76$ curves) and non working days (n_2 =39 curves). Non working days are weekends and holidays during the given data period. For using the functional data analysis it is essential to first convert the discrete data to functional form (i.e., continuous function) by using basis function. We will utilize these three datasets 1) to explore the effect of different basis expansion (Fourier and B-spline) on outlier detection. 2) to compare the proposed method for detection of outliers with the Febrero's results.



(a) (b)

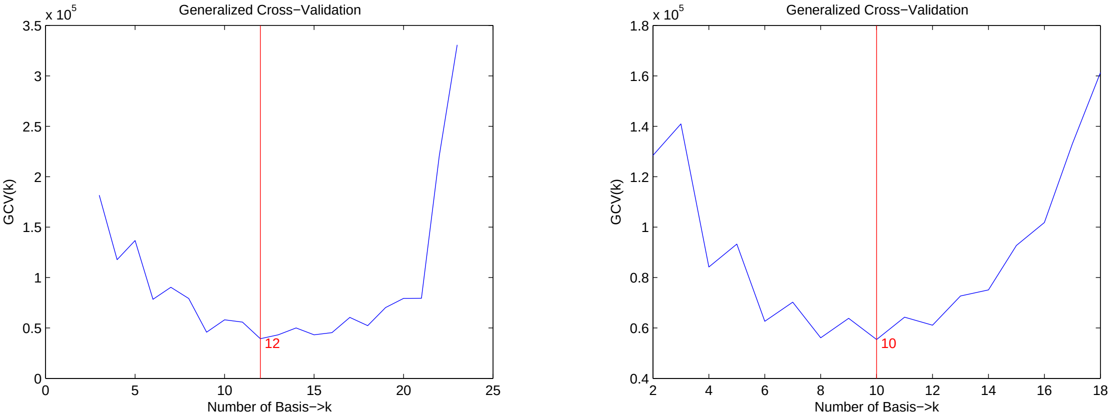
Figure 3.1: Sample curves of NOx data by using (a) B-spline Basis and (b) Fourier Basis

From the Figure 3.1 we can say that NOx levels increase in the morning and reach peak value around 8:00 am, then decrease until 2:00 pm, and again increase in the evening.

The Figures 3.5 and 3.8 exhibit the sample curves for working days and non working days, respectively. In Figures 3.1, 3.5 and 3.8, the group of curves shows presence of a few trajectories that are in some way different from the rest.

NOx (Whole sample):

Initial dimension of the dataset is 115 × 24. GCV method is used to determine optimal number of bases for B-spline and Fourier bases, and the resulting k based on GCV are 12 for B-spline basis (Figure 3.2(a)) and 10 for Fourier basis (Figure 3.2(b)). Whole sample data curves by using 12 B-spline bases and 11 Fourier bases are shown in Figure 3.1(a) and (b), respectively. Since there is a high correlation among the variables of coefficient matrix we apply CPCA, ROBPCA and BACONPCA on coefficient matrix for dimension reduction and outlier detection. For B-spline and Fourier



(a) (b)

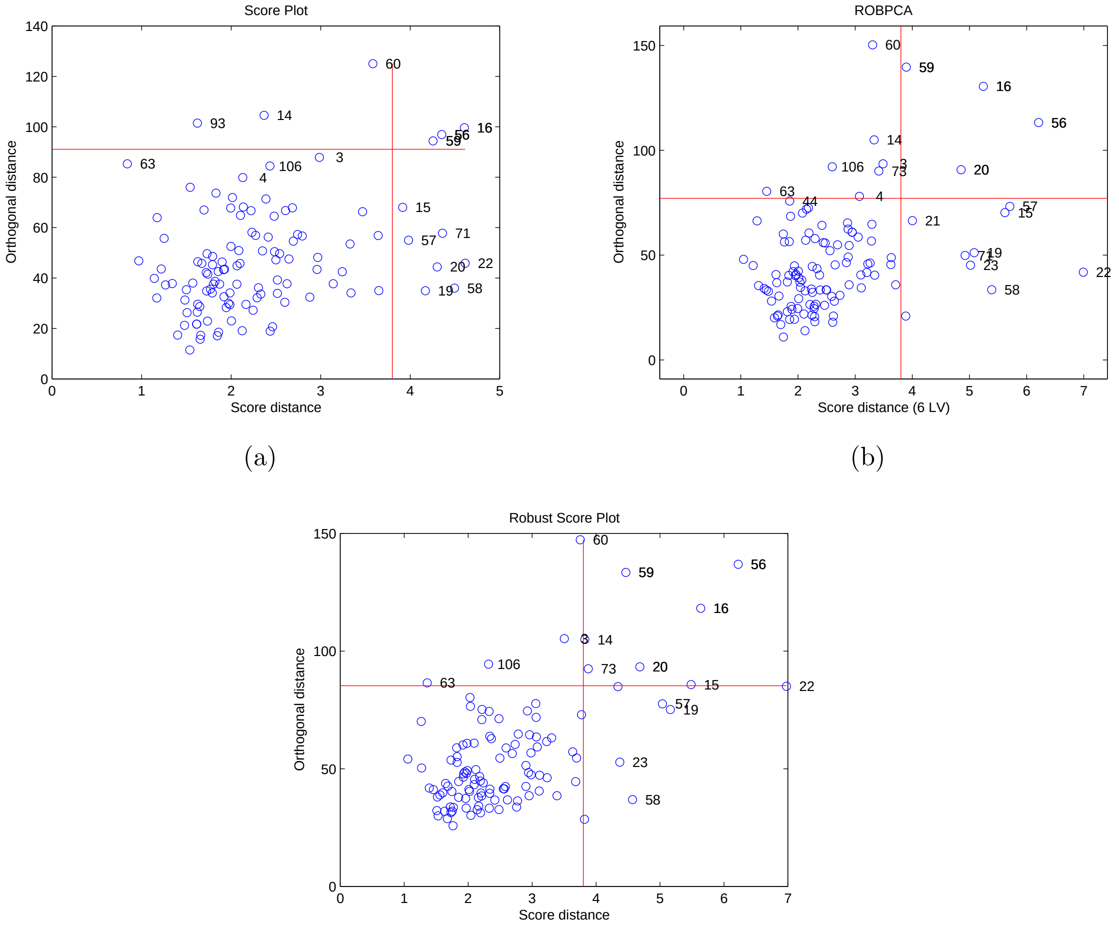
Figure 3.2: Generalized Cross-Validation by using (a) B-spline Basis and (b) Fourier Basis for NOx data

bases, six and five principal components were retained, respectively each for CPCA, ROBPCA and BACONPCA, yielding a classical and robust explanation percentage more than 90%.

The resulting diagnostic plots for the three PCA methods by using B-spline and Fourier bases are given in Figures 3.3 and 3.4, respectively. All bad leverage points, detected by these diagnostic plots (orthogonal-score plots) formed by using the three PCA methods based on both bases are listed in Table 3.1.

NOx (Working Days):

Initial dimension of the dataset is 76 × 24. Optimal number of bases based on GCV is obtained as k=12 for B-spline, k=10 for Fourier basis. Data curves for working days by using 12 B-spline bases and 11 Fourier bases are given in Figure 3.5(a) and (b), respectively. Due to high correlation among the variables in C, we apply CPCA, ROBPCA and BACONPCA on coefficient matrix. For B-spline and Fourier bases, seven and five principal components were retained, respectively each for CPCA,

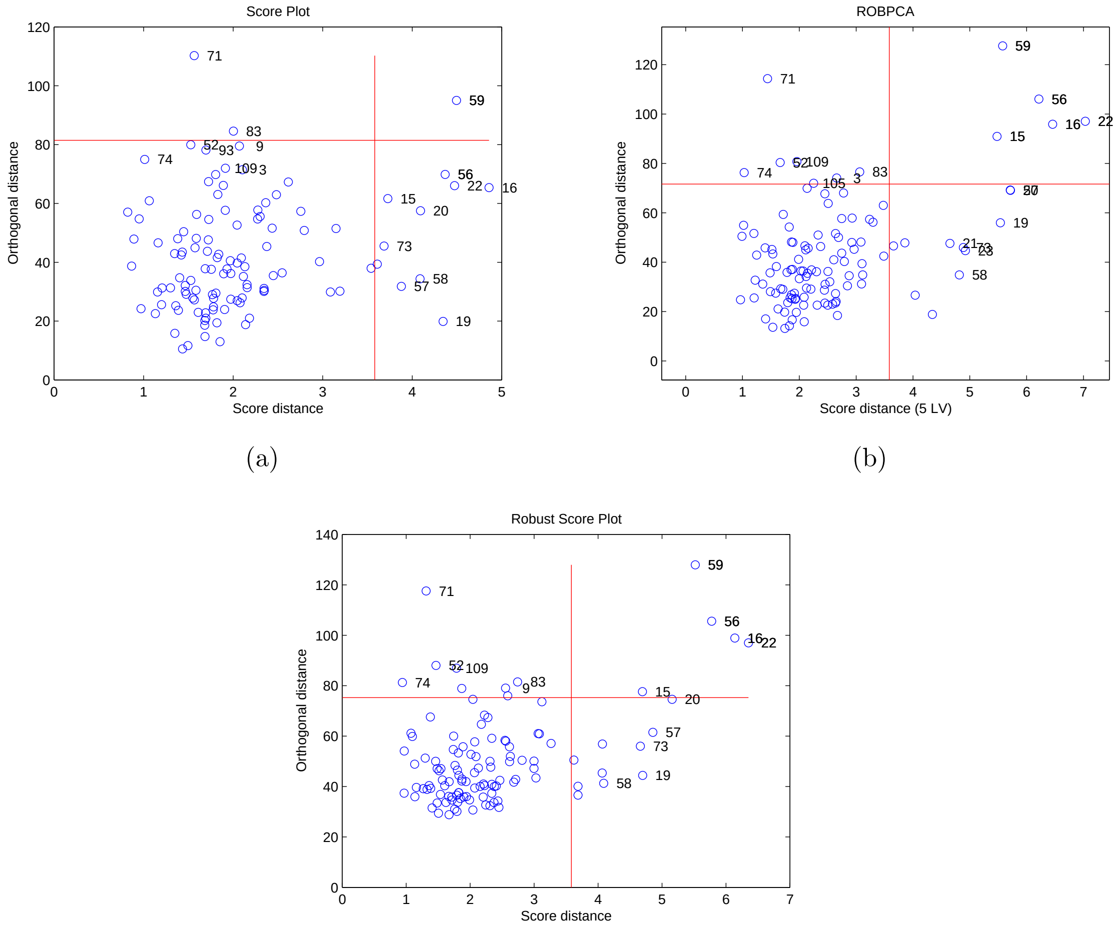


(a)

(b)

(c)

Figure 3.3: Orthogonal-score plot for whole sample by using B-spline basis computed with (a)CPCA, (b)ROBPCA, (c)BACONPCA

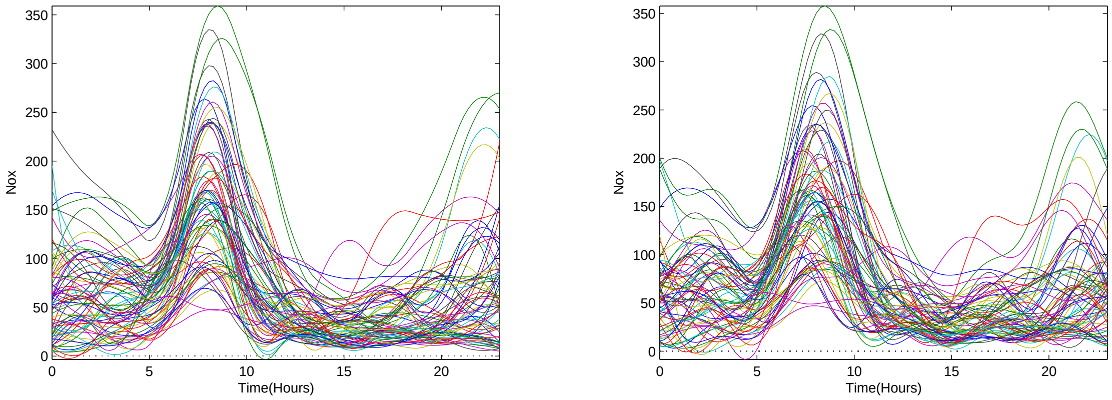


(a)

(b)

(c)

Figure 3.4: Orthogonal-score plot for whole sample by using Fourier basis computed with (a)CPCA, (b)ROBPCA, (c)BACONPCA



(a) (b)

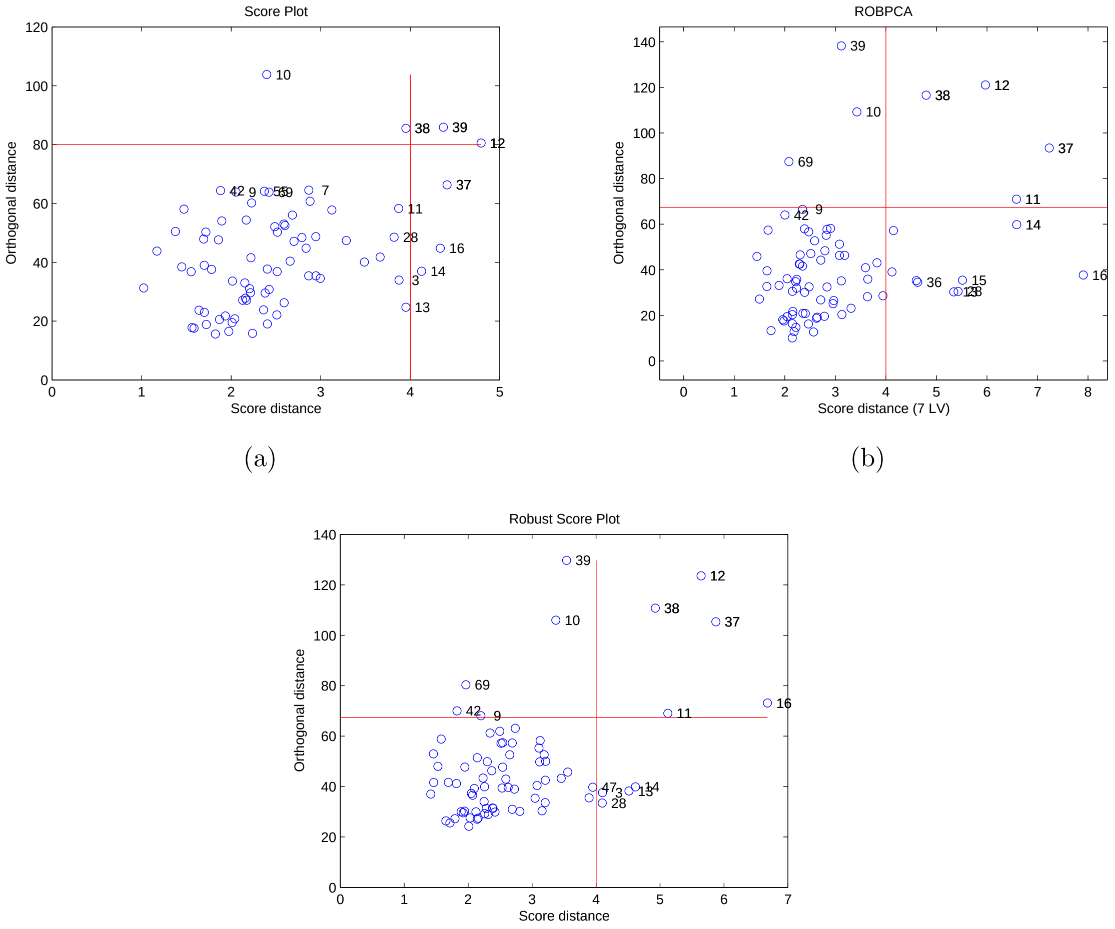
Figure 3.5: Working Days of NOx data by using (a) B-spline Basis and (b) Fourier Basis

ROBPCA and BACONPCA, yielding a classical and robust explanation percentage more than 90%.

The resulting diagnostic plots for the three PCA methods by using these bases are displayed in Figures 3.6 and 3.7, respectively. All bad leverage points detected by these diagnostic plots (orthogonal-score plots) formed by using the three PCA methods based on both bases are listed in Table 3.1.

NOx (Non-Working Days):

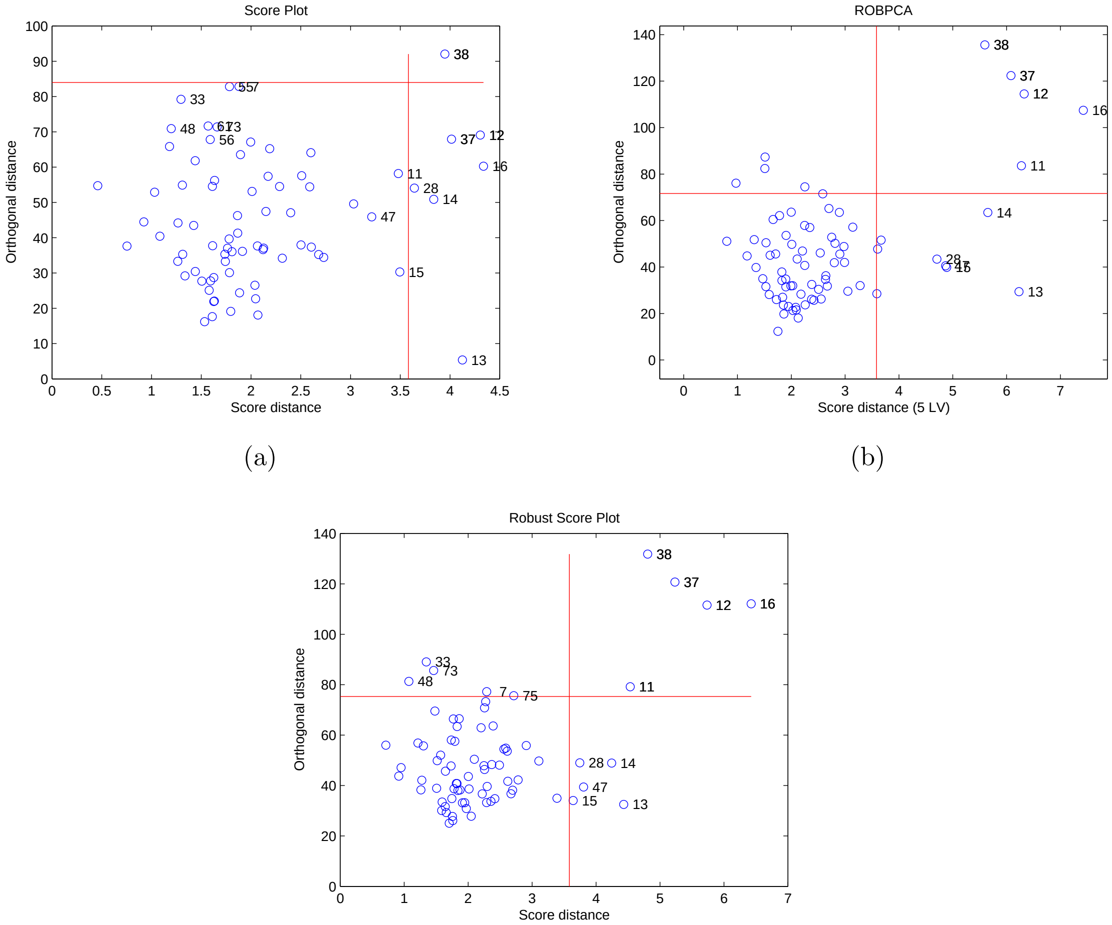
Initial dimension of the dataset is 39×24. By GCV method the optimal numbers of bases for B-spline and Fourier bases are k=12 for B-spline bases and k=6 for Fourier basis. Data curves for Non-working days by using 12 B-spline bases and 7 Fourier bases are given in Figure 3.8(a) and (b), respectively. We again apply CPCA, ROBPCA and BACONPCA on coefficient matrix. For B-spline and Fourier basis, five and three principal components were retained, respectively each for CPCA, ROBPCA



(a) (b)

(c)

Figure 3.6: Orthogonal-score plot for working days by using B-spline basis computed with (a)CPCA, (b)ROBPCA, (c)BACONPCA

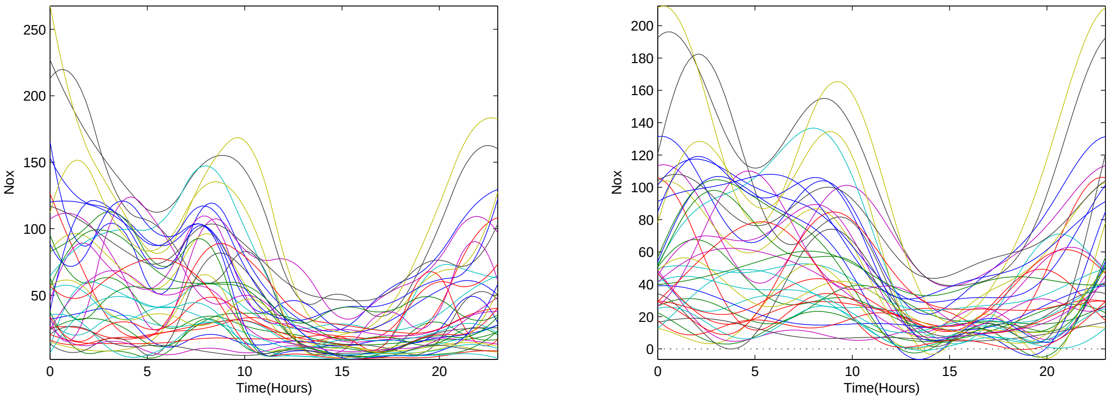


(a)

(b)

(c)

Figure 3.7: Orthogonal-score plot for working days by using Fourier computed with (a)CPCA, (b)ROBPCA, (c)BACONPCA



(a)

(b)

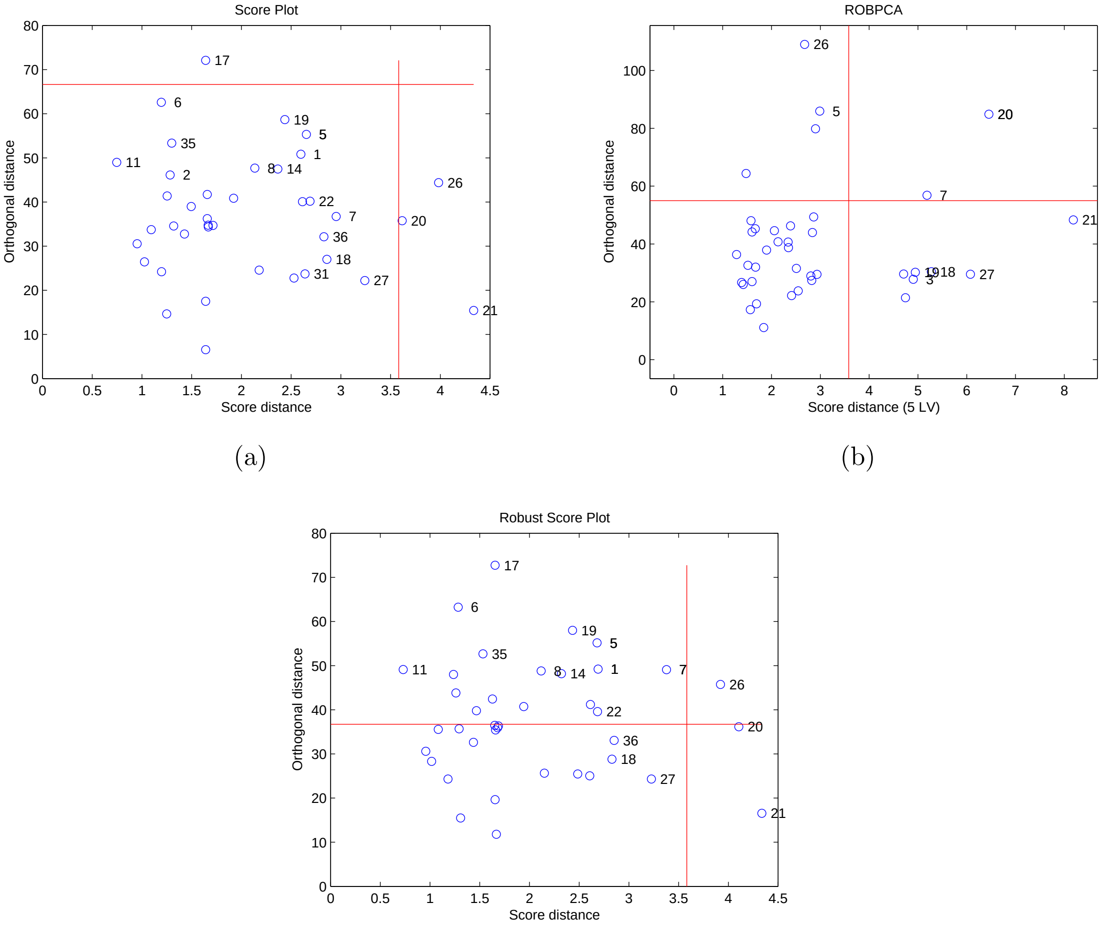
Figure 3.8: Non-working Days of NOx data by using (a) B-spline Basis and (b) Fourier Basis

and BACONPCA, yielding a classical and robust explanation percentage more than 90%.

The resulting diagnostic plots for the three PCA methods by using B-spline and Fourier bases which are given in Figures 3.9 and 3.10, respectively, illustrate all types of outliers in this dataset. All bad leverage points detected by these diagnostic plots (orthogonal-score plots) formed by using the three PCA methods based on both bases are listed in Table 3.1.

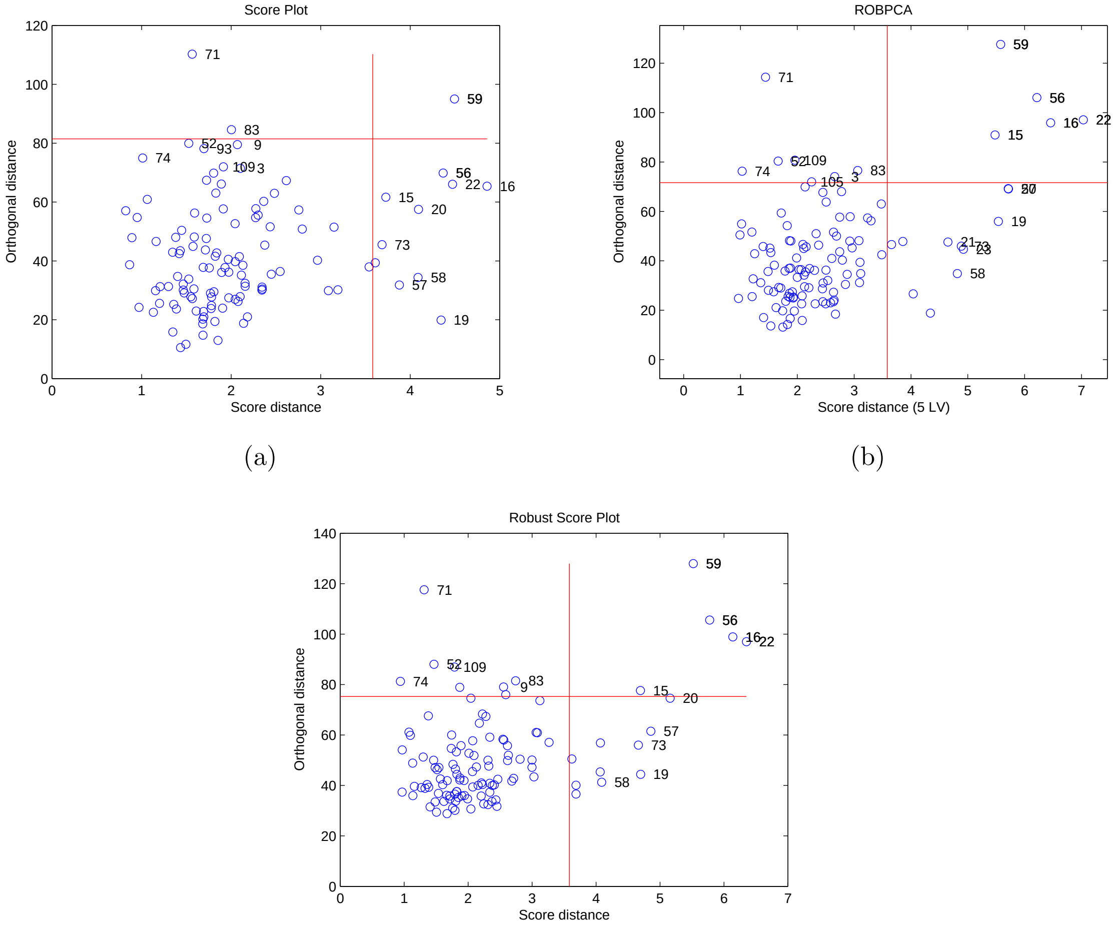
Figure 3.11 shows the outliers identified by the proposed method for the three datasets. All bad leverage points detected by three PCA methods by using two different bases are summarized in Table 3.1. Outliers detected by Febrero *et al.* [13] for the same dataset are also given in Table 3.2 for comparison purposes.

From Table 3.1 we conclude that both ROBPCA and BACONPCA detected similar outliers in whole sample by using Fourier basis. These results match with the results obtained by Febrero *et al.* [13] (refer Table 3.2). However, we have detected



(a) (b) (c)

Figure 3.9: Orthogonal-score plot for Non-working Days by using B-spline basis computed with (a)CPCA, (b)ROBPCA, (c)BACONPCA

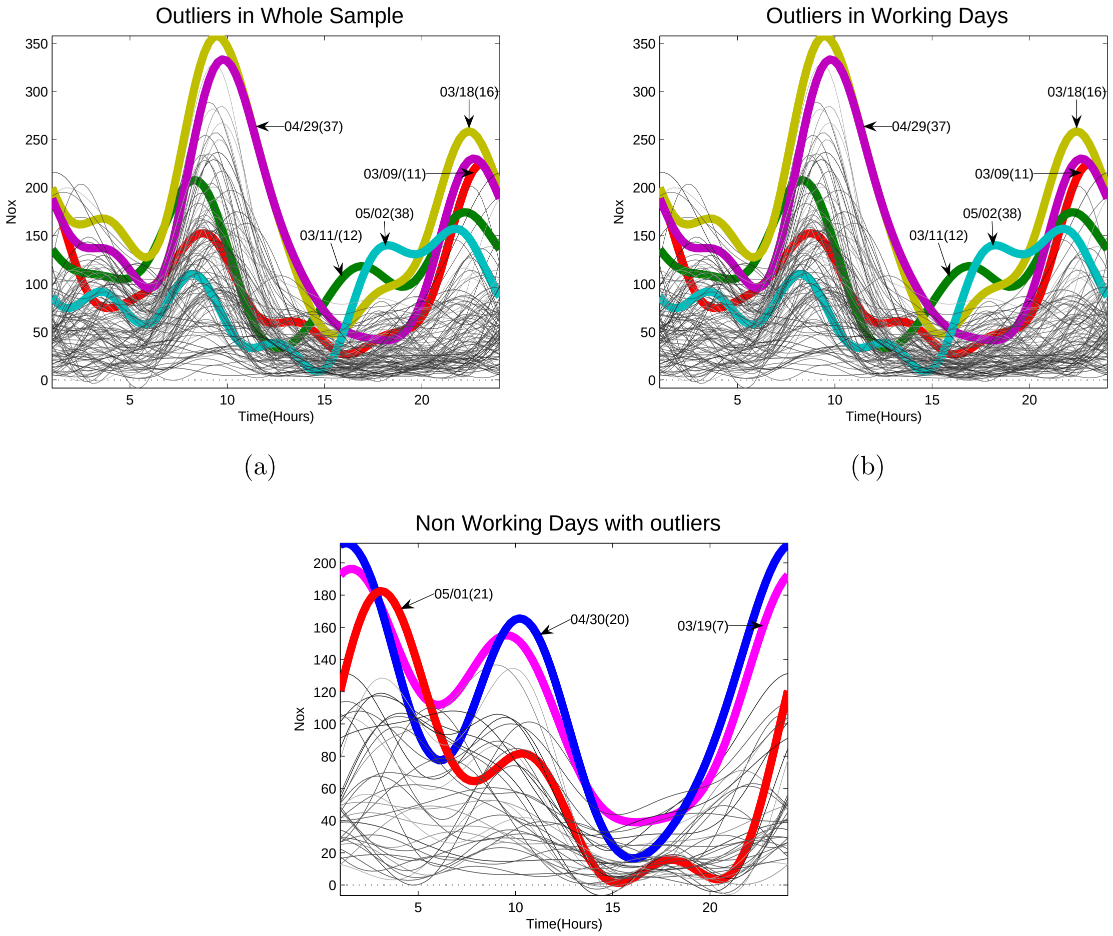


(a)

(b)

(c)

Figure 3.10: Orthogonal-score plot for Non-working Days by using Fourier basis computed with (a)CPCA, (b)ROBPCA, (c)BACONPCA



(a)

(b)

(c)

Figure 3.11: Outliers detected by proposed method for (a)Whole sample, (b)Working days, (c)Non-working days

one additional outlier (03/09) by using both ROBPCA and BACONPCA methods. The orthogonal score plot based on CPCA detects only one bad leverage point.

For whole samples by using B-spline basis four similar outliers are detected by both ROBPCA and BACONPCA methods. CPCA detected three outliers. But outliers detected by these three PCA methods using B-spline basis do not conform with outliers detected by Febrero *et al.* [13].

For working days dataset we obtained five similar outliers for both ROBPCA and BACONPCA by using Fourier basis. Outliers thus detected by us match with outliers detected by Febrero *et al.* [13] for working days dataset using Fourier basis, except that we have detected two additional outliers (05/02 and 03/09) (refer Table 3.1). CPCA method using Fourier basis detected only one outlier.

Under B-spline basis four outliers are detected by ROBPCA method and five outliers are detected by BACONPCA method for working dataset. Results obtained by both PCA methods do not match with results obtained by Febrero *et al.* [13]. Here also CPCA detected only one outlier and converted bad leverage points detected by ROBPCA and BACONPCA into good leverage points.

In non-working days dataset under Fourier basis ROBPCA detected two outliers and BACONPCA detected three outliers. These results match with the findings of Febrero *et al.* [13] except that BACONPCA detected one additional outlier (05/01). Under Fourier basis CPCA failed to detect any outlier.

Similarly for non-working dataset using B-spline basis the CPCA method has not detected any outlier. Outliers detected by ROBPCA and BACONPCA using B-spline basis are different from those obtained by Febrero *et al.* [13].

After detecting outliers, we checked for sources for abnormal values of these curves. We expected to provide information about the abnormally large NOx emissions on these particular days. We found that Friday, March 11 is the beginning

Table 3.1: Outliers detected by three PCA methods. (.) denotes case number for three datasets

	Fourier Basis			B-spline Basis		
Dataset	CPCA	ROBPCA	BACONPCA	CPCA	ROBPCA	BACONPCA
Whole sample	05/02(59)	03/18(22) 04/29(56) 03/11(16) 05/02(59) 03/09(15)	03/18(22) 04/29(56) 03/11(16) 05/02(59) 03/09(15)	04/29(56) 03/11(16) 05/02(59)	04/29(56) 03/11(16) 05/02(59) 03/16(20)	04/29(56) 03/11(16) 05/02(59) 03/16(20)
Working days	05/02(38)	03/18(16) 04/29(37) 03/11(12) 05/02(38) 03/09(11)	03/18(16) 04/29(37) 03/11(12) 05/02(38) 03/09(11)	05/03(39)	04/29(37) 03/11(12) 05/02(38) 03/09(11)	04/29(37) 03/11(12) 05/02(38) 03/09(11) 03/18(16)
Non working days	—	04/30(20) 03/19(07)	04/30(20) 03/19(07) 05/01(21)	—	04/30(20) 03/19(07)	05/15(26)

Table 3.2: Outliers detected by Febrero *et al.* [13] for the NOx data

Dataset	$\|\cdot\|_1$	$\|\cdot\|_2$	$\|\cdot\|_\infty$
Whole sample	03/18 04/29 03/11	03/18 04/29 03/11 05/02	03/18 04/29 03/11 05/02
Working days	03/18 04/29 03/11	03/18 04/29 03/11	03/18 04/29 03/11
Non working days	04/30 03/19	04/30 03/19	04/30 03/19

of a weekend. The Friday, March 18 and Saturday, March 19 are the beginning of the Eastern vacation in Spain in the year 2005. Also Friday, April 29, Saturday, April 30, Sunday, May 1, and Monday, May 2 correspond to long weekend. There is sudden increase in traffic on these small vacation periods. So we conclude that abnormal observations on specific days can be attributed to increase in traffic due to small vacation periods. We have also detected outlier on Wednesday, March 9. It is observed that high NOx emissions are recorded on March 9 after 8:00 pm. Since the observation on March 10^{th} is missing and thus not included in analysis, we could not pinpoint the reason behind this abnormal observation on March 9^{th}.

3.2 Simulation

The simulation study is conducted to compare the performance of ROBPCA and BACONPCA with the classical PCA (CPCA) on coefficient matrix. The simulation setting given by Fraiman and Muniz [12, 28], with few changes, is used here. For simulation we consider functional data $x_1, \ldots, x_n$ obtained as realizations from a stochastic process $X(\cdot)$. This functional data has continuous paths on the observation period $[t_{min}, t_{max}] = [0, 1]$. Curves are generated from different models. Model 1 was generated without contamination and several other models were generated with different types of contaminations.

Model 1 (no contamination): $X_i(t) = g(t) + e_i(t), 1 \leq i \leq n$, where model error term $e_i(t)$ is a stochastic Gaussian process with zero mean and covariance function $\vartheta(s,t) = (1/2)(1/2)^{(0.9)|t-s|}$ and g(t)=4t, with $t \in [0, 1]$.

Model 2 (asymmetric contamination): $Y_i(t) = X_i(t) + c_i M$, $1 \leq i \leq n$, where c_i is 1 with probability q and 0 with probability $1 - q$; M is the contamination size constant.

Model 3 (symmetric contamination): $Y_i(t) = X_i(t) + c_i \sigma_i M$, $1 \leq i \leq n$, where c_i and M are defined as in model 2 and σ_i is a sequence of random variables independent of c_i taking values 1 and -1 with probability 1/2.

Model 4 (partially contaminated): $Y_i(t) = X_i(t) + c_i \sigma_i M$, if $t \geq Ti$, $1 \leq i \leq n$, and $Y_i(t) = X_i(t)$, if $t < T_i$, where T_i is a random number generated from a uniform distribution on $[0, 1]$.

Model 5 (Peak contamination): $Y_i(t) = X_i(t) + c_i \sigma_i M$, if $T_i \leq t \leq T_i + \ell$, $1 \leq i \leq n$, and $Y_i(t) = X_i(t)$, if $t \notin [T_i, T_i + \ell]$, where $\ell = 2/30$ and T_i is a random number from a uniform distribution in $[0, 1 - \ell]$. Figure 3.12 exhibits curves simulated from these five models.

For each model, we generated 100 replications, with two settings each for low and high dimensional data. For low dimensional data we consider 1) $n = 100,\ p = 12,\ k = 8$ and 2) $n = 50,\ p = 5,\ k = 4$ settings. For high dimensional data we also consider two settings with 1) $n = 50,\ p = 100,\ k = 51$ and 2) $n = 50,\ p = 500,\ k = 151$. For the model 1 contamination percent is $q = 0$ and contamination constant is $M = 0$. For each contaminated model (2, 3, 4 and 5) we considered several levels of contamination: $q = 5,\ 10,\ 15$ percentage and contamination constants $M = 10$ and 25. Classical PCA and robust methods ROBPCA and BACONPCA are used on the simulated functional data based on the five models.

Two quantitative measures of the goodness of the methods are considered. The first one is Mean proportion of variability (MPV) :

$$MPV = 1/N \sum_{m=1}^{N} \frac{\hat{\lambda}_1^m + \hat{\lambda}_2^m + \ldots + \hat{\lambda}_k^m}{\lambda_1^m + \lambda_2^m + \ldots + \lambda_k^m + \ldots + \lambda_p^m}$$

where $N = 100$ denotes the number of iterations and λ_j^m is an j^{th} eigenvalue at m^{th} replication obtained from the covariance matrix of coefficient matrix of uncontaminated model. $\hat{\lambda}_j^m$ is the estimated value of λ_j^m at the m^{th} replication. $\hat{\lambda}_j^m$ is obtained by using classical or robust multivariate techniques on coefficient matrix of contaminated or uncontaminated model. 90% of variability is explained by the first three components for each setting. For the mean proportion of explained variability the optimal values are 0.9 for low and high dimensional data.

The second quantitative measure is the Norm of the difference between $\hat{\lambda}_1^m$ and λ_1^m which is given as $||\hat{\lambda}_1^m - \lambda_1^m||$, where λ_1^m is largest eigenvalue obtained from the covariance matrix of coefficient matrix of uncontaminated model. $\hat{\lambda}_1^m$ is the estimated value of λ_1^m at m^{th} replication. $\hat{\lambda}_1^m$ is largest eigenvalue obtained by using classical or robust multivariate techniques on coefficient matrix of contaminated or uncontaminated model. The optimal value is zero or near zero.

Model 1 is compared with models 2, 3, 4 and 5. The simulation results of mean proportion of variability for four comparisons are given in Tables 3.3 - 3.6. It is clear that CPCA provides the best mean proportion of explained variability when there is no contamination in the data, which is expected. For the uncontaminated data robust methods also yield comparable results. However, when contamination is introduced to the data (models 2-5) the eigenvalues obtained with CPCA are overestimated. Since estimated percentages of MPV are larger than 100%. In ROBPCA and BACONPCA we obtain MPV of 90% for low dimensional data without and with contamination. For high dimensional data the mean percentage of explained variability is similarly 90% for without and with contamination. The main reason behind this is the optimal direction obtained by ROBPCA and BACONPCA are robust to outliers. CPCA clearly fails and provides the worst possible result because mean proportion of variability is above 100%. It is clear from Tables 3.3 - 3.5 that the MPV for ROBPCA at 15% contamination level is above 100% in most of the cases except for model 5. BACONPCA gives better results than ROBPCA at 15% for both low and high dimensional case. From these results we can deduce that BACONPCA and ROBPCA outperform the CPCA.

Simulation results for the norm with $N = 100$ iterations and different contamination levels for comparison of model 1 vs models 2 are summarized in Figures 3.13 - 3.14. For this comparison, we used two high and two low dimensional settings with the two values of M (10 and 25). The ideal value of norm must be very small or near zero. We conclude that the norm is near zero when there is no contamination for all methods. This is an indication of ROBPCA and BACONPCA being also effective methods for uncontaminated data. The norm for CPCA tends to increase as contamination level increases. For contaminated data, norms corresponding to ROBPCA and BACONPCA method yield minimum value which is near zero for high and low dimensional settings. The comparisons of model 1 vs model 3-5 for low and high dimensional settings yielded very similar results observed in Figures 3.13-3.14, therefore they are not repeated here.

We conducted the simulation given above for $M = 5$, but the results were not up to the mark. This is because in this case the outliers are not yet very well separated from the regular data group. As soon as the contamination lies somewhat further, robust methods are capable to distinguish the outliers. Therefore the results of the simulation for this case is not reported in this thesis since we aimed at distinguishing outliers from the regular points.

Table 3.3: Simulation results of the Mean Proportion of Explained Variability when there is no contamination (0%) and with symmetric contamination (5%, 10% and 15%) for low and high dimensional cases.

High dimension:n=50, p=100, k=51						
Contamination	M=10			M=25		
	CPCA	ROBPCA	BACONPCA	CPCA	ROBPCA	BACONPCA
0%	0.943	0.790	0.912	0.944	0.801	0.914
5%	13.213	0.833	0.918	71.297	0.838	0.915
10%	21.955	0.874	0.910	121.020	0.866	0.917
15%	32.354	2.472	0.918	195.299	12.720	0.927

High dimension:n=50, p=500, k=151						
Contamination	M=10			M=25		
	CPCA	ROBPCA	BACONPCA	CPCA	ROBPCA	BACONPCA
0%	0.949	0.793	0.907	0.948	0.785	0.909
5%	11.749	0.821	0.917	73.250	0.804	0.912
10%	21.131	0.842	0.908	126.076	0.840	0.910
15%	31.715	1.512	0.919	201.703	14.303	0.928

Low dimension:n=100, p=12, k=8						
Contamination	M=10			M=25		
	CPCA	ROBPCA	BACONPCA	CPCA	ROBPCA	BACONPCA
0%	0.914	0.833	0.892	0.917	0.830	0.898
5%	10.441	0.856	0.901	55.600	0.852	0.897
10%	18.859	0.860	0.889	114.111	0.870	0.903
15%	30.120	1.707	0.899	169.042	0.885	0.896

Low dimension:n=50, p=5, k=4						
Contamination	M=10			M=25		
	CPCA	ROBPCA	BACONPCA	CPCA	ROBPCA	BACONPCA
0%	0.955	0.801	0.913	0.956	0.812	0.918
5%	9.736	0.837	0.923	56.574	0.830	0.913
10%	18.992	0.866	0.910	107.322	0.0.875	0.922
15%	25.839	1.360	0.921	161.370	3.781	0.920

Table 3.4: Simulation results of the Mean Proportion of Explained Variability when there is no contamination (0%) and with asymmetric contamination (5%, 10% and 15%) for low and high dimensional cases.

High dimension:n=50, p=100, k=51						
Contamination	M=10			M=25		
	CPCA	ROBPCA	BACONPCA	CPCA	ROBPCA	BACONPCA
0%	0.943	0.815	0.921	0.944	0.801	0.914
5%	10.254	0.837	0.920	70.863	0.839	0.914
10%	20.233	0.862	0.919	115.722	0.869	0.914
15%	26.345	0.895	0.919	165.056	9.011	0.921

High dimension:n=50, p=500, k=151						
Contamination	M=10			M=25		
	CPCA	ROBPCA	BACONPCA	CPCA	ROBPCA	BACONPCA
0%	0.948	0.785	0.909	0.948	0.785	0.909
5%	12.728	0.804	0.909	74.551	0.806	0.905
10%	20.080	0.849	0.910	120.222	0.847	0.911
15%	28.722	2.110	0.916	173.636	9.202	0.917

Low dimension:n=100, p=12, k=8						
Contamination	M=10			M=25		
	CPCA	ROBPCA	BACONPCA	CPCA	ROBPCA	BACONPCA
0%	0.914	0.833	0.892	0.915	0.834	0.896
5%	9.856	0.846	0.888	57.283	0.847	0.897
10%	17.285	0.870	0.896	106.619	0.868	0.897
15%	24.770	0.887	0.914	152.289	3.153	0.901

Low dimension:n=50, p=5, k=4						
Contamination	M=10			M=25		
	CPCA	ROBPCA	BACONPCA	CPCA	ROBPCA	BACONPCA
0%	0.953	0.800	0.899	0.954	0.824	0.923
5%	10.930	0.826	0.908	51.805	0.844	0.922
10%	16.986	0.853	0.907	105.488	0.873	0.920
15%	24.150	1.923	0.920	135.844	0.903	0.921

Table 3.5: Simulation results of the Mean Proportion of Explained Variability when there is no contamination (0%) and with partial contamination (5%, 10% and 15%) for low and high dimensional cases.

High dimension:n=50, p=100, k=51						
Contamination	M=10			M=25		
	CPCA	ROBPCA	BACONPCA	CPCA	ROBPCA	BACONPCA
0%	0.944	0.793	0.913	0.942	0.781	0.911
5%	5.641	0.831	0.925	34.047	0.819	0.917
10%	10.920	0.865	0.910	60.144	0.843	0.905
15%	15.836	1.280	0.908	99.545	1.487	0.918

High dimension:n=50, p=500, k=151						
Contamination	M=10			M=25		
	CPCA	ROBPCA	BACONPCA	CPCA	ROBPCA	BACONPCA
0%	0.948	0.785	0.909	0.949	0.822	0.911
5%	6.514	0.820	0.909	33.517	0.850	0.929
10%	11.222	0.861	0.920	69.062	0.872	0.922
15%	15.815	1.095	0.919	98.250	1.209	0.927

Low dimension:n=100, p=12, k=8						
Contamination	M=10			M=25		
	CPCA	ROBPCA	BACONPCA	CPCA	ROBPCA	BACONPCA
0%	0.915	0.805	0.882	0.915	0.829	0.892
5%	5.792	0.830	0.885	30.162	0.842	0.895
10%	11.325	0.856	0.885	62.695	0.860	0.894
15%	14.734	0.855	0.873	90.655	1.119	0.897

Low dimension:n=50, p=5, k=4						
Contamination	M=10			M=25		
	CPCA	ROBPCA	BACONPCA	CPCA	ROBPCA	BACONPCA
0%	0.955	0.819	0.910	0.956	0.816	0.917
5%	6.745	0.831	0.903	35.184	0.841	0.912
10%	11.983	0.856	0.918	65.041	0.870	0.929
15%	18.445	1.577	0.931	100.715	4.946	0.920

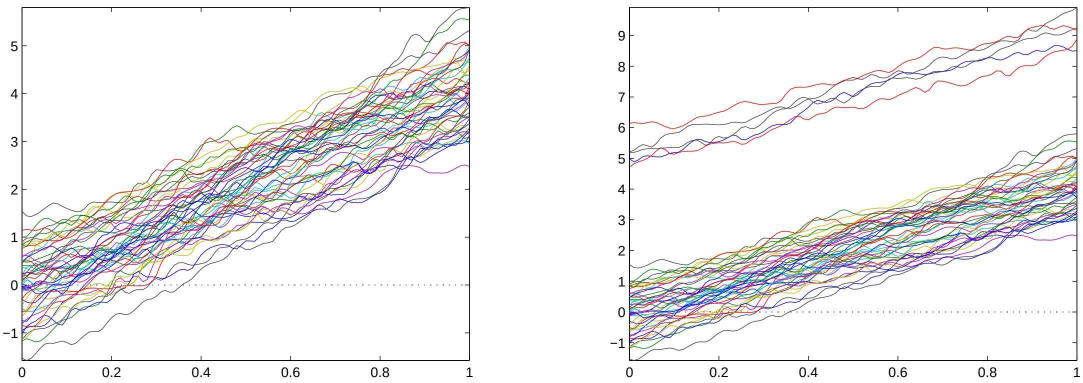
Table 3.6: Simulation results of the Mean Proportion of Explained Variability when there is no contamination (0%) and with peak contamination (5%, 10% and 15%) for low and high dimensional cases.

High dimension:n=50, p=100, k=51						
Contamination	M=10			M=25		
	CPCA	ROBPCA	BACONPCA	CPCA	ROBPCA	BACONPCA
0%	0.945	0.802	0.922	0.944	0.812	0.916
5%	1.505	0.832	0.913	5.285	0.851	0.923
10%	1.792	0.882	0.904	7.427	0.871	0.917
15%	2.070	0.921	0.873	8.989	0.912	0.898

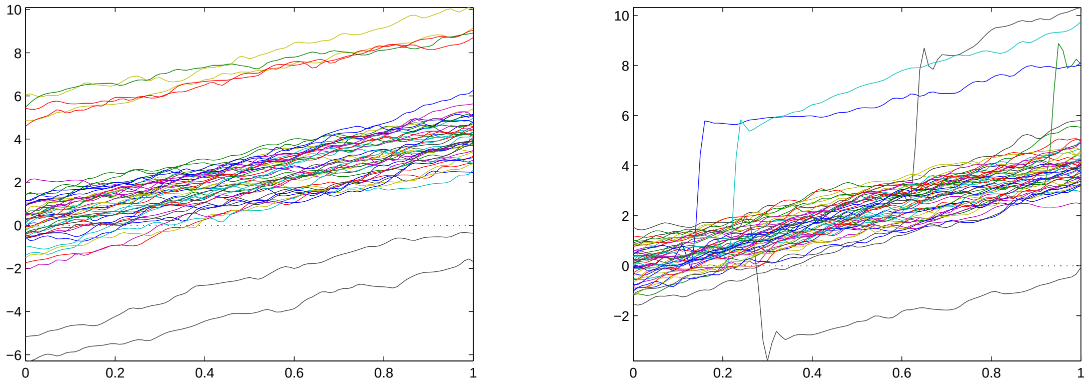
High dimension:n=50, p=500, k=151						
Contamination	M=10			M=25		
	CPCA	ROBPCA	BACONPCA	CPCA	ROBPCA	BACONPCA
0%	0.949	0.822	0.911	0.948	0.785	0.909
5%	1.462	0.855	0.925	5.123	0.824	0.910
10%	1.752	0.878	0.902	7.006	0.860	0.900
15%	1.998	0.933	0.901	8.347	0.896	0.896

Low dimension:n=100, p=12, k=8						
Contamination	M=10			M=25		
	CPCA	ROBPCA	BACONPCA	CPCA	ROBPCA	BACONPCA
0%	0.916	0.841	0.893	0.916	0.823	0.895
5%	1.703	0.852	0.893	6.506	0.834	0.896
10%	2.419	0.850	0.887	10.273	0.851	0.893
15%	2.917	0.862	0.889	15.164	0.871	0.899

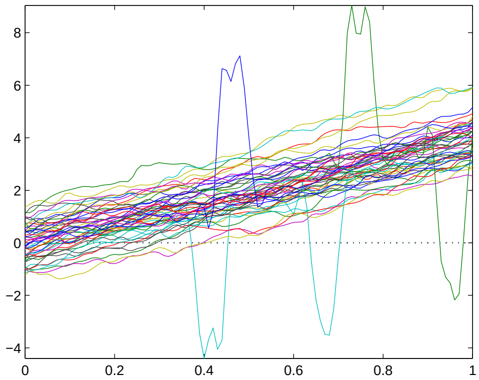
Low dimension:n=50, p=5, k=4						
Contamination	M=10			M=25		
	CPCA	ROBPCA	BACONPCA	CPCA	ROBPCA	BACONPCA
0%	0.954	0.799	0.901	0.954	0.814	0.898
5%	1.591	0.808	0.904	5.652	0.818	0.901
10%	2.456	0.811	0.903	9.499	0.828	0.897
15%	3.369	0.820	0.908	15.931	0.831	0.896



(a) Model 1 (b) Model 2



(c) Model 3 (d) Model 4



(e) Model 5

Figure 3.12: Curves generated from model 1 (without contamination), model 2 (asymmetric contamination), model 3 (symmetric contamination), model 4 (partial contamination) and model 5 (peak contamination) with n=50, p=100, M=10 and q=0.1.

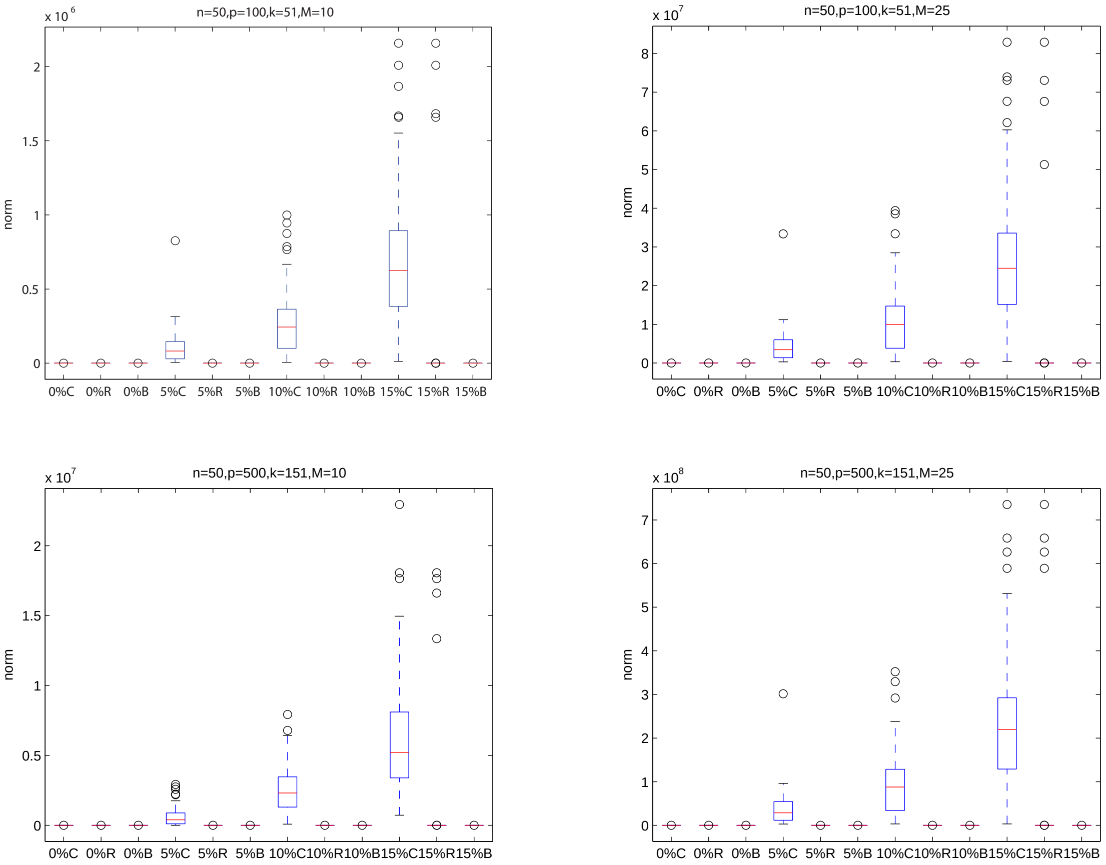


Figure 3.13: Boxplots of norm when there is no contamination (0%) and symmetric contamination (5%, 10% and 15%) for high dimensional cases for CPCA(C) ROBPCA(R) and BACONPCA(B).

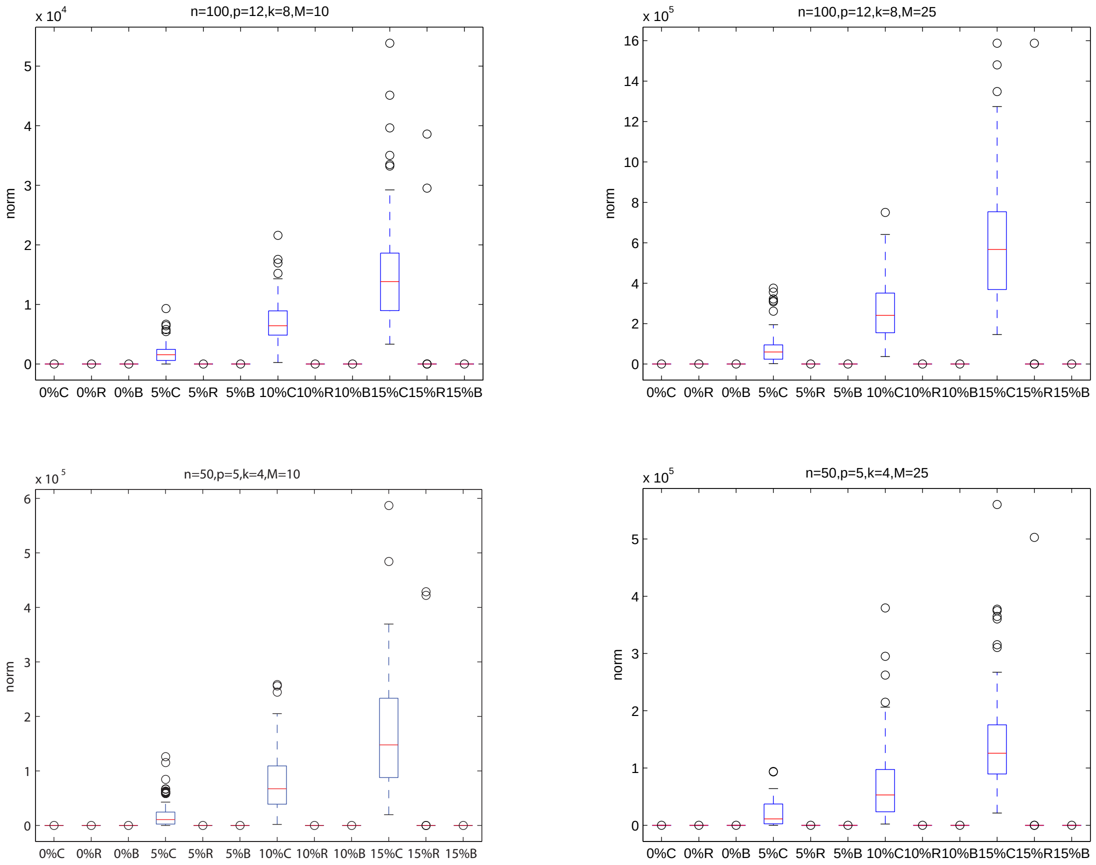


Figure 3.14: Boxplots of norm when there is no contamination (0%) and symmetric contamination (5%, 10% and 15%) for low dimensional cases for CPCA(C) ROBPCA(R) and BACONPCA(B).

CHAPTER 4

ROBUST FUNCTIONAL PRINCIPAL COMPONENT REGRESSION

4.1 Introduction

Recently researchers have put more attention to functional linear models in which the regressors and/or the response are of a functional nature and proposed several methods for estimating the functional parameter [10, 11, 15, 16, 30]. Functional regressors are infinite in nature. Problem with infinite dimensionality of the regressor is that, it results into infinitely many sets of solutions or suffers from multicollinearity. Therefore the first step in this regression setup is to reduce dimension by using FPCA and then is to regress the response onto these components obtained from FPCA.

The presence of outliers or influential observations has a serious effect on the estimation and prediction of the functional linear model. In the presence of outliers, the decomposition of the classical covariance matrix is unreliable. In such situation, both the FPCA stage and the regression stage called Functional Principal Component Regression (FPCR) yield unreliable results. Influential observations in a given dataset can have a strong impact on analysis. If these outlying or influential observations are removed from the data then this may substantially affect the statistical inference. The functional versions of the diagnostic measures based on Cook's distance [8] is introduced by Chiou and M*ü*ller [7] and Shen and Xu [38] for the models where the regressors are real or curves and the responses are functional.

Recently, Febrero *et al.* [16] reviewed estimation based on the classical functional principal components method and then analyzed influence in the functional linear model with scalar response. They have proposed three measures of influence by generalizing the measures proposed for the standard regression model by Cook [8] and Pẽna [29]. In this chapter we propose a robust functional principal component regression method which consists of two parts. First we apply a robust FPCA method on the regressors, and then regress the response variables on the scores, which are discussed in Section 4.2. In Section 4.3 we propose robustified influence regression diagnostic measures to detect which observations have strong influence. In Section 4.4 the practical use of these measures is illustrated by means of a real dataset.

4.2 Estimation of Functional Parameter β

The functional linear model with a scalar response is a regression model with the regressor which is a random curve and the response which is real random variable defined on the same probability space. We assume that (X, y) is a pair of random variables where $X = (X(t))$, $X \in L_2(T)$, $t \in T = [t_{min}, t_{max}] \subset \Re$ and y is a real random variable. For easy computation we assume that both X and y are centered i.e., $E[X(t)] = 0$, and $E[y] = 0$. Assuming $E(\parallel X \parallel^2) < \infty$, the dependence between the scalar response y and the functional random variable X is written as:

$$y = \langle X, \beta \rangle + \epsilon = \int_T X(t)\beta(t)dt + \epsilon, \tag{4.1}$$

where $\langle ., . \rangle$, denotes the $L_2(T)$ inner product, β is a square integrable function defined on T and errors, ϵ, is a real random variable with $E[\epsilon] = 0$, $E[X(t)\epsilon] = 0$ and finite variance equal to σ^2.

Suppose that a random sample of pairs (X_i, y_i), $i = 1, \ldots, n$, is observed where X_i and y_i $(i = 1, \ldots, n)$ are realizations of the functional X and y, respectively.

The estimate of β can be obtained by finding such β that minimizes residual sum of squares

$$RSS(\beta) = \sum_{i=1}^{n} (y_i - \langle X_i, \beta \rangle)^2. \tag{4.2}$$

Such β is a functional parameter, that has high dimensionality problem. Thus minimization of RSS can be accomplished by using PC approach. The sample covariance operator of X denoted by ψ_X allows a spectral decomposition into orthonormal eigenfunctions $\gamma_1, \gamma_2, \ldots$ [33]. By Mercer's Theorem, an orthogonal expansion for ψ_X in L_2 is given by:

$$\psi_X(s,t) = \sum_{K=1}^{\infty} \lambda_K \gamma_K(s) \gamma_K(t), \tag{4.3}$$

with ordered nonnegative eigenvalues $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_n \geq 0 = \lambda_{n+1} \geq \ldots$ The sequence of eigenvalue-eigenvector pairs satisfies the eigenequation given by $\varphi_X \gamma_K = \lambda_K \gamma_K$, for $K \geq 1$ and $\langle \gamma_K, \gamma_l \rangle = 1$ if $K = l$ and $\langle \gamma_K, \gamma_l \rangle = 0$ otherwise.

By using Karhunen-Loève expansion [1] the functional variables X_i and the functional parameter β can be written in terms of the eigenfunctions γ_K in following manner:

$$X_i = \sum_{K=1}^{\infty} \xi_{iK} \gamma_K, \tag{4.4}$$

$$\beta = \sum_{K=1}^{\infty} \beta_K \gamma_K, \tag{4.5}$$

where $\xi_{iK} = \langle X_i, \gamma_K \rangle$, such that $\xi_{iK} = 0$, for $K > n$ and $\beta_K = \langle \beta, \gamma_K \rangle$, respectively, for $i = 1, \ldots, n$ and $K = 1, 2, \ldots$. Eigenfunctions form an orthonormal basis of the functional space $L_2(T)$ [30].

By using the new definitions of X_i and β the residual sum of squares in equation (4.2) can be written as:

$$RSS(\beta) = \sum_{i=1}^{n}(y_i - \sum_{K=1}^{n} \xi_{iK}\beta_K)^2. \tag{4.6}$$

The dimension of β in equation (4.6) is reduced from ∞ to n. But minimizing this equation will give us a perfect fit of the response variable. To avoid this problem alternate method proposed by Cardot *et al.* [4] to estimate β is used. In this method $\beta_K = 0$, for $K \geq k_n + 1$, where $0 < k_n < n$ and $\lambda_{k_n} > 0$. The coefficients β_K, for $K = 1, \ldots, k_n$, are obtained by minimizing the residual sum of squares given by:

$$RSS(\beta_{1:k_n}) = \sum_{i=1}^{n}(y_i - \sum_{K=1}^{k_n} \xi_{iK}\beta_K)^2 = \parallel Y - \xi_{(1:k_n)}\beta_{(1:k_n)} \parallel^2, \tag{4.7}$$

where $Y = (y_1, \ldots, y_n)'$ is the $n \times 1$ vector, $\beta_{(1:k_n)} = (\beta_1, \ldots, \beta_{k_n})'$ is the $k_n \times 1$ vector and $\xi_{(1:k_n)}$ is $n \times k_n$ score matrix whose K^{th} column is the vector $\xi_{.K} = (\xi_{1K}, \ldots, \xi_{nK})'$, the K^{th} principal component score, which satisfies $Var(\xi_{.K}) = \lambda_K$ and $cov(\widehat{\xi_{.K}, \xi_{.m}}) = 0$ for $K \neq m$. The least-squares estimate of $\beta_{(1:k_n)}$ is then given by:

$$\hat{\beta}_{(1:k_n)} = (\xi'_{1:k_n}\xi_{1:k_n})^{-1}\xi'_{1:k_n}Y,$$

where $\xi'_{1:k_n}\xi_{1:k_n}$ is a $k_n \times k_n$ diagonal matrix whose $(K, K)^{th}$ element is $\xi'_{.K}\xi_{.K} = n\lambda_K$ and $\xi'_{.K}\xi_{.m} = 0$, for $K \neq m$. And $\xi'_{1:k_n}Y$ is a $k_n \times 1$ vector whose K^{th} element is $n \times cov(\widehat{\xi_{.K}, Y})$,

$$\hat{\beta}_{(1:k_n)} = (\frac{\xi'_{.1}Y}{n\lambda_1}, \ldots, \frac{\xi'_{.k_n}Y}{n\lambda_{k_n}}). \tag{4.8}$$

This expression defines the least-squares estimate of the slope β, denoted by $\hat{\beta}_{(k_n)}$, as follows:

$$\hat{\beta}_{(k_n)} = \sum_{K=1}^{k_n} \hat{\beta}_K \gamma_K = \sum_{K=1}^{k_n} \frac{\xi'_{.K} Y}{n\lambda_K} \gamma_K = \sum_{K=1}^{k_n} \frac{\widehat{cov(\xi_{.K}, Y)}}{\lambda_K} \gamma_K. \tag{4.9}$$

$\hat{\beta}_{(k_n)}$ is an estimator of functional regression parameter obtained by using classical functional principal component regression (CFPCR). Cardot *et al.*[4] showed that, under several conditions, $\langle X, \hat{\beta}_{(k_n)} \rangle$ converges in probability and almost surely to $\langle X, \beta \rangle$. The principal components become more rough with increase in value of K. k_n acts as a smoothing parameter which has to be determined. There are many ways to determine the value of k_n. The selection criteria to choose k_n is based on the eigenvalues. The predetermined threshold value is 90%. The optimal number of components k_n is then the minimal value for which the cumulative percentage of total variance is greater than or equal to 90%.

If dataset contains outliers then spectral decomposition of covariance in (4.3) is unreliable. Both the expressions in equations (4.4) and (4.5) are not reliable. To obtain the robust estimate of the $\hat{\beta}_{(k_n)}$ we employ robust method of principal components (e.g., BACONPCA Section 2.2.2) on X. Robust values of scores $\xi_{iK}^{(r)}$ $i = 1, \ldots, n$ and $K = 1, \ldots$ and robust eigenvector $\gamma_K^{(r)}$ $K = 1, \ldots$ are obtained. Using robust values both the expression in equation (4.4) and (4.5) can be written as:

$$X_i^{(r)} = \sum_{K=1}^{\infty} \xi_{iK}^{(r)} \gamma_K^{(r)}, \tag{4.10}$$

$$\beta^{(r)} = \sum_{K=1}^{\infty} \beta_K^{(r)} \gamma_K^{(r)}. \tag{4.11}$$

Both expansions allow to write the residual sum of squares in equation (4.7) as:

$$RSS(\beta^{(r)}_{1:k_n}) = \sum_{i=1}^{n}(y_i - \sum_{K=1}^{k_n} \xi^{(r)}_{iK}\beta^{(r)}_{K})^2. \tag{4.12}$$

By minimizing the residual sum of squares the robust estimate of the slope $\beta_{(k_n)}$, denoted by $\hat{\beta}^{(r)}_{(k_n)}$ is obtained as:

$$\hat{\beta}^{(r)}_{(k_n)} = \sum_{K=1}^{k_n} \hat{\beta_K}^{(r)}\gamma^{(r)}_K = \sum_{K=1}^{k_n} \frac{\xi'^{(r)}_{.K}Y}{n\lambda^{(r)}_K}\gamma^{(r)}_K. \tag{4.13}$$

$\hat{\beta}^{(r)}_{(k_n)}$ is an estimator of functional regression parameter obtained by using robust functional principal component regression (RFPCR).

4.3 Functional Regression Diagnostic measures

A pair of i^{th} observation of the form (X_i, y_i) is called influential whose deletion would lead to a noticeable change in the regression parameter estimates. To detect the presence of influential observations or outliers, diagnostic measures defined for ordinary linear regression model can be extended to functional linear model with scalar response [16]. We will start defining robustified versions of residuals, fitted values, leverages and then some other diagnostic measures such as Cook's D [8], Hadi's potential-residual measure [17] for functional data.

Similar to the standard regression model the fitted values and residuals are useful in defining the influence measures of single observation for functional linear model with scalar response.

Fitted values and Leverages

Let the fitted values of the response variable be denoted by $\hat{y}_i$ and can be obtained from equations (4.1) and (4.9) in the following manner:

$$\hat{y}_i = \langle X_i, \hat{\beta}_{(k_n)} \rangle = \sum_{K=1}^{k_n} \xi_{iK} \hat{\beta}_K = \sum_{K=1}^{k_n} \xi_{iK} \frac{\xi'_{.K} Y}{n\lambda_K}, \tag{4.14}$$

for $i = 1, \ldots, n$, which allows to define the $n \times 1$ vector of fitted values $\hat{Y} = (\hat{y}_1, \ldots, \hat{y}_n)'$ The matrix form of above equation can be written as follows: $\hat{Y} = H_{(k_n)} Y$, where $H_{(k_n)}$ is the $n \times n$ hat matrix, given by:

$$H_{(k_n)} = \xi_{1:k_n} (\xi'_{1:k_n} \xi_{1:k_n})^{-1} \xi'_{1:k_n},$$

which can be written as:

$$H_{(k_n)} = \zeta_{1:k_n} \zeta'_{1:k_n},$$

where $\zeta_{1:k_n}$ is the $n \times k_n$ matrix whose K^{th} column is the vector

$$\zeta_{.K} = \frac{\xi_{.K}}{\sqrt{n\lambda_K}},$$

$$H_{(k_n)} = \frac{1}{n} \left(\frac{\xi_{.1} \xi'_{.1}}{\lambda_1} + \ldots + \frac{\xi_{.k_n} \xi'_{.k_n}}{\lambda_{k_n}} \right). \tag{4.15}$$

The diagonal elements of $H_{(k_n)}$ are leverage values denoted by $H_{(k_n),ii}$, given by:

$$H_{(k_n),ii} = \frac{1}{n} \left(\frac{\xi_{i1}^2}{\lambda_1} + \ldots + \frac{\xi_{ik_n}^2}{\lambda_{k_n}} \right), \tag{4.16}$$

where $0 \leq H_{(kn),ii} \leq 1$ and $Trace H_{(k_n)} = k_n$. The leverage can be used as a quick way to measure influential observation in prediction. Smaller the values of hat diagonal for the pair (X_i, y_i), the better is predicted y_i. We can say that if the leverage value $H_{(k_n),ii}$ of any observation (X_i, y_i) exceeds $2 \times k_n/n$, then that observation might be an influential observation. But, an observation with high leverage value may not necessarily be influential. Observations with high leverage values are the outliers in $X - space$ but the converse is not necessarily true.

Residuals (Ordinary and other)

The residuals are defined as $e = Y - \hat{Y} = (I_n - H_{(k_n)})Y$, where I_n is the $n \times n$ identity matrix. Using equation (4.1) the relationship between ϵ and e can be established in following manner:

$$e = (I_n - H_{(k_n)})Y = \xi_{(k_n+1:n)}\beta_{(k_n+1:n)} + (I_n - H_{(k_n)})\epsilon, \tag{4.17}$$

where matrix $\xi_{(k_n+1:n)}$ is the $n \times (n - k_n)$ whose columns are the vectors $\xi_{.K}$, for $K = k_n + 1, \ldots, n$ and $\beta_{(k_n+1:n)} = (\beta_{k_n+1}, \ldots, \beta_n)$.

As ϵ has zero mean and covariance $\sigma^2 I_n$, then the vector of residuals e conditional on $X_1, \ldots, X_n$ has mean $\xi_{(k_n+1:n)}\beta_{(k_n+1:n)}$ and covariance $\sigma^2(I_n - H_{(k_n)})$. If n is large then the term $\xi_{(k_n+1:n)}\beta_{(k_n+1:n)}$ can be omitted [5, 18]. Since $Trace(I_n - H_{(k_n)}) = n - k_n$

$$E[e'e|X_1, \ldots, X_n] = n(\frac{\beta_{k_n+1}^2}{\lambda_{k_n+1}} + \ldots + \frac{\beta_n^2}{\lambda_n}) + (n - k_n)\sigma^2. \tag{4.18}$$

Therefore, the error variance σ^2 may be estimated by the functional residual variance estimate, $\hat{s}^2$, given by:

$$\hat{s}^2 = \frac{e'e}{n - k_n}. \tag{4.19}$$

Internally studentized residual

The internally studentized residual is given as:

$$r_i^2 = \frac{e_i^2}{\hat{s}^2(1 - H_{(k_n),ii})}. \tag{4.20}$$

Cook's D Measure

The Cook's distance is the standardized difference in estimating β with and without the observation (X_i, y_i)

$$CP_i = \frac{(\hat{y} - \hat{y}_{(-i,k_n)})'(\hat{y} - \hat{y}_{(-i,k_n)})}{k_n \hat{s}^2}, \tag{4.21}$$

where $\hat{y}_{(-i,k_n)}$ denotes the prediction of the response vector y excluding the i^{th} observation (X_i, y_i) in the estimation. A high value of CP_i indicates that the i^{th} observation has influence on estimated responses in the sense that deleting it from the dataset will alter the prediction value. To compute CP_i in equation (4.21) requires n standard linear regressions with an observation deleted. Cook's distance can be written as a measure which is a function of quantities related to the full dataset

$$CP_i = \frac{e_i^2}{k_n \hat{s}^2(1 - H_{(k_n),ii})} \frac{H_{(k_n),ii}}{(1 - H_{(k_n),ii})} = r_i^2 \frac{H_{(k_n),ii}}{k_n(1 - H_{(k_n),ii})}, \tag{4.22}$$

where r_i^2 is the i^{th} internally studentized residual. Observations with large residual values (r_i^2) are called outliers.

Hadi's Potential-Residual Measure

This measure combines two measures $H_{(k_n),ii}$, which provides information about high-leverage points and r_i, which contains information about outliers. High leverage and outlier observations may influence the regression results and conclusions based on them. High leverage points are likely to have small residuals, so detecting residuals alone is not sufficient to find influential observations. The Cook's measure is a multiplicative function of the residual and the leverage value. The drawback of multiplicative function is that when one of the two components is small or near zero, it suppresses the other component. If one of the component is too small then the multiplicative measure is also small. The observations with large leverage value are likely to have small residuals and this will result in small values for multiplicative measures. This may lead to incorrect conclusions that these observations are not influential. But an additive measure which is a function of the residual and leverage value, by contrast, is large if either or both components are large. Instead of multiplicative measure we use additive measure of influence suggested by Hadi [17, 6] and is defined as:

$$HM_i^2 = \frac{k_n}{(1 - H_{(k_n),ii})} \frac{d_i^2}{(1 - d_i^2)} + \frac{H_{(k_n),ii}}{(1 - H_{(k_n),ii})}, \tag{4.23}$$

$i = 1, \ldots, n$, where $d_i^2 = e_i^2/e'e$ is the square of the i^{th} normalized residual and $\sum d_i^2 = 1$. The first component is a function of the i^{th} normalized residual. The second component is the ratio of the $Var(\hat{y}_i)$ and $Var(e_i)$ and this quantity is known as the *"potential"*. HM_i^2 is monotonically increasing in both the leverage value and the squared residual. Therefore, a large value of H_i^2 may be due to a large value of e_i, a large value of $H_{(k_n),ii}$ or both. The additive measures can identify outliers in the X-space, in the y-space, or in both. The cut-off value for HM_i^2 suggested by Hadi [17] is $mean(HM_i^2) + c\sqrt{Var(HM_i^2)}$. Since mean and variance are nonrobust; median and median absolute deviations are used to estimate the cut-off value, respectively.

The Potential-Residual (P-R) plot can be used to distinguish between regular observations, outliers and leverage points (Figure 4.1). This diagnostic plot is a

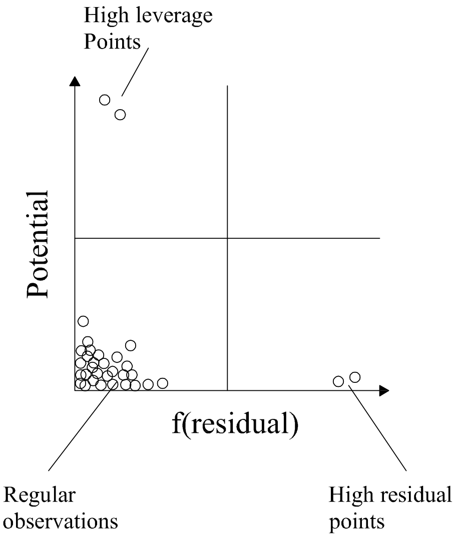


Figure 4.1: Potential-Residual (P-R) plot.

scatter plot of $\frac{H_{(k_n),ii}}{(1-H_{(k_n),ii})}$ versus $\frac{k_n}{(1-H_{(k_n),ii})}\frac{e_i^2}{(1-d_i^2)}$. Since the first component is known as potential and the second component is a function of the residual, this plot is referred as the Potential-Residaul (P-R) plot. In this plot the high leverage points are located in the upper left area and observations with high prediction error are located in the lower right area [17].

If dataset contains outliers then all diagnostic measures defined earlier will be sensitive to outliers. To obtain the robust estimates of these diagnostic measures we apply robust method of principal components (e.g., Section 2.2.2) on X. Robust eigenvector $\gamma_K^{(r)}$, $K = 1, \ldots$; robust eigenvalue $\lambda_K^{(r)}$, $K = 1, \ldots$; and robust values of scores $\xi_{iK}^{(r)}$, $i = 1, \ldots, n$ and $K = 1, \ldots$ are obtained. Using robust values both the expression in equations (4.14) and (4.16) can be redefined.

Robust fitted values are given as:

$$\hat{y_i}^{(r)} = \langle X_i, \hat{\beta}^{(r)}_{(k_n)} \rangle = \sum_{K=1}^{k_n} \xi_{iK}^{(r)} \frac{\xi_{.K}^{(r)'} Y}{n \lambda_K^{(r)}}. \tag{4.24}$$

Robust Leverages are defined as:

$$H^{(r)}_{(k_n),ii} = \frac{1}{n} (\frac{\xi_{i1}^{(r)^2}}{\lambda_1^{(r)}} + \ldots + \frac{\xi_{ik_n}^{(r)^2}}{\lambda_{k_n}^{(r)}}). \tag{4.25}$$

The robust version of Cook's measure can also be defined in the following manner:

$$CP_i^{(r)} = \frac{e_i^{(r)^2}}{k_n \hat{s}^{(r)2} (1 - H^{(r)}_{(k_n),ii})} \frac{H^{(r)}_{(k_n),ii}}{(1 - H^{(r)}_{(k_n),ii})}, \tag{4.26}$$

where $e_i^{(r)} = y_i - \hat{y_i}^{(r)}$, $i = 1, \ldots, n$, is the i^{th} residual.

The robust version of Hadi's measure is given as follows:

$$HM_i^{2(r)} = \frac{k_n}{(1 - H^{(r)}_{(k_n),ii})} \frac{d_i^{2(r)}}{(1 - d_i^{2(r)})} + \frac{H^{(r)}_{k_n,ii}}{(1 - H^{(r)}_{(k_n),ii})}, \tag{4.27}$$

where $e_i^{(r)} = y_i - \hat{y_i}^{(r)}$, $i = 1, \ldots, n$, is the i^{th} robust residual and $d_i^{2(r)} = e_i^{2(r)} / e^{(r)'} e^{(r)}$ is the square of the i^{th} normalized robust residual.

Section 4.4 compares the classical and robustified versions of diagnostic measures by utilizing a real dataset.

4.4 Numerical Example

In this section we demonstrate the practical use of proposed influence measures. For this we have considered a dataset analyzed by Ramsay and Silverman [32]. The data is from thirty-five Canadian weather stations which are listed in Table 4.1. In this dataset the regressor set of curves is the mean daily temperatures (in degree Celsius) and the response is the logarithm (base 10) of total annual precipitation (in mm). The temperature is assumed to be periodic due to its cyclical behavior during years. Therefore, we used the Fourier basis functions. The Figure 4.2(a) shows the curves dataset using 65 Fourier functions, while a boxplot of the log-precipitation data is shown in the Figure 4.2(b). The curves dataset is assumed to be observed in the interval $[t_{min}, t_{max}] = [0, 1]$. The time unit is one year which is discretized in 365 days. For this analysis both regressor and response variable are centered by substracting their means.

To select the cutoff value k_n we estimate the coefficients by using the least squares method and then we compute the total variance. The optimal number of components $k_n = 2$ as the cumulative percentage of total variance for two components is 96%. Orthogonal-score plots for two components are constructed using CPCA and BACON-PCA which are given in Figure 4.3. The orthogonal-score plots for CPCA indicated no bad leverage points while BACONPCA revealed three bad leverage points (21, 11, and 10).

The estimated beta function by classical FPCA and robust FPCA and the two eigenfunctions by CFPCR and RFPCR are depicted in Figure 4.4. The estimated beta function and first eigenfunction for both methods show differences; on the other hand second eigenfunction do not differ too much. The estimated beta function and first eigenfunction clearly indicates that the classical methods are not robust to outliers and there is a need for robust method to estimate the parameter function [20].

Table 4.2 contains values for diagnostic measures. Residual plots by classical (Figure 4.5(a)) and robust methods (Figure 4.5(b)) show that observation 12 (Kamloops) is an outlier as it has high residual value as compared to other stations. From both figures we can deduce that residual values are much larger for robust method than those of classical method.

The leverages which are the diagonal elements of the hat matrix indicate which stations might be influential a priori. We can clearly see that leverage value for observations 21 (Resolute), 11 (Iqaluit) and 10 (Inuvik) computed by robust method are larger than those of the classical method (Figure 4.6(a)).

The influential observations detected by Cook's measure are 21 (Resolute) and 11 (Iquluit) by using the robust method and only 21 (Resolute) by using the classical method (Figure 4.6(b)). Comparing influential observations for both methods the observations 21 and 11 have much higher $CP_i^{(r)}$ value than CP_i.

The cut-off value for HM_i^2 is 0.43 and for $HM_i^{2(r)}$ is 0.46. The influential observations detected by Hadi's influence measure are 21 (Resolute) and 11 (Iquluit) by the robust method and 21 (Resolute) by the classical method (Figure 4.6(c)). Comparing influential observations for both methods the observations 21 and 11 have much higher $HM_i^{(r)}$ value than HM_i.

By comparing Hadi's measure with Cook's measure for classical method we can conclude that Hadi's measure is superior in terms of distinguishing influential observations from regular observations. Observation 11 (Figure 4.6(c)) identified by Hadi's measure is clearly distinguished from regular observations when compared with observation 11 (Figure 4.6(b)) by Cook's measure. Therefore we can say that the additive

Table 4.1: Names of the Candian weather stations

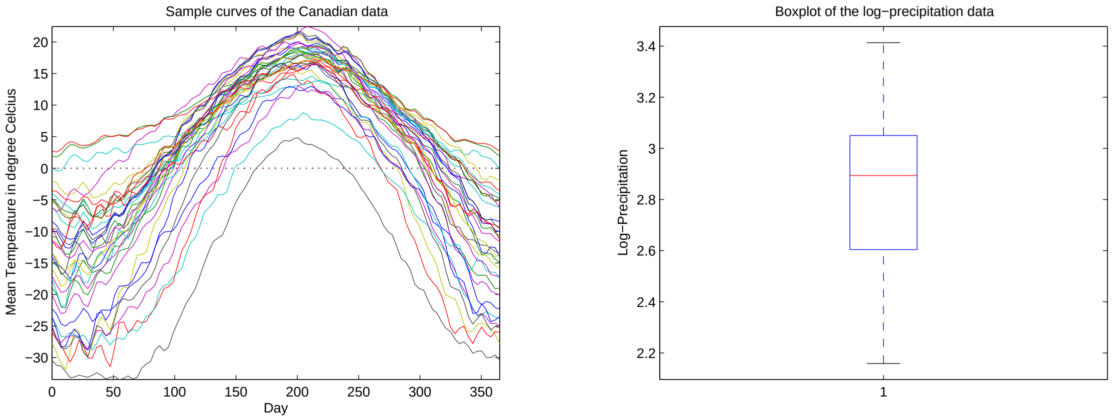
1)Arvida	2)Bagottvi	3)Calgary	4)Charlott	5)Churchil	6)Dawson
7)Edmonton	8)Frederic	9)Halifax	10)Inuvik	11)Iqaluit	12)Kamloops
13)London	14)Montreal	15)Ottawa	16)Princeal	17)Princege	18)Princeru
19)Quebec	20)Regina	21)Resolute	22)Scheffer	23)Sherbroo	24)Stjohns
25)Sydney	26)Thepas	27)Thunderb	28)Toronto	29)Uraniumc	30)Vancouvr
31)Victoria	32)Whitehor	33)Winnipeg	34)Yarmouth	35)Yellowkn	

measure used by Hadi's measure, is much more efficient than multiplicative measure used by Cook's measure.

The P-R plot shown in Figure 4.7(a) indicates that five observations with high values of HM_i^2 can be classified as follows: observation 21 is a high leverage point, observation 12 is an outlier and observation 11, 18 and 31 are combinations of both. The P-R plot by robust method indicates that three observations which have large $HM_i^{2(r)}$ values can be classified as follows: observation 21 is an high leverage point, observation 12 is an outlier and observation 11 is a combination of both (Figure 4.7(b)). Here P-R plot helps to identify the outliers, leverages and regular observations. By comparing the classical and robust P-R plots (Figure 4.7(c)) we find that observations 21, 12 and 11 yielded much larger values for robust method than the ones obtained from the classical method. Observations identified as bad leverages in orthogonal score plot (Figure 4.3) and observations identified as outliers and leverages in P-R plot by using both methods (Figure 4.7) are highlighted in Figure 4.8.

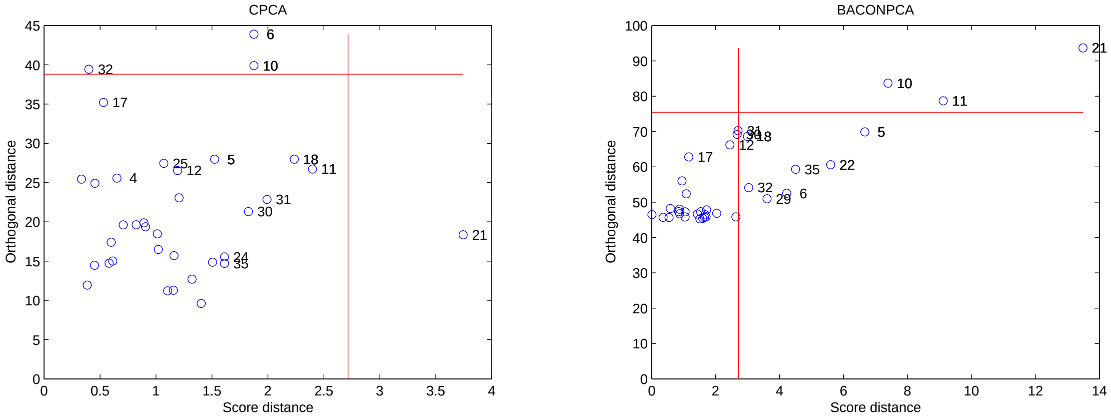
Table 4.2: Classical and Robust influence measures for the Candian weather stations

n	e_i	$e_i^{(r)}$	$H_{(k_n),ii}$	$H^{(r)}_{(k_n),ii}$	CP_i	$CP_i^{(r)}$	HM_i^2	$HM_i^{2(r)}$
1	0.16	0.11	0.01	0.01	0.01	0.00	0.02	0.01
2	0.19	0.16	0.01	0.01	0.01	0.01	0.01	0.01
3	-0.30	-0.33	0.01	0.00	0.01	0.01	0.02	0.01
4	0.13	0.09	0.01	0.01	0.00	0.00	0.01	0.01
5	0.06	0.23	0.07	0.15	0.00	0.24	0.07	0.19
6	0.03	0.08	0.10	0.12	0.00	0.02	0.12	0.13
7	-0.14	-0.16	0.00	0.00	0.00	0.00	0.00	0.00
8	0.15	0.08	0.01	0.01	0.00	0.00	0.01	0.01
9	0.18	0.13	0.02	0.02	0.01	0.01	0.02	0.02
10	**-0.03**	**0.16**	**0.10**	**0.20**	**0.00**	**0.16**	**0.12**	**0.24**
11	**0.05**	**0.32**	**0.17**	**0.34**	**0.01**	**1.62**	**0.20**	**0.53**
12	**-0.57**	**-0.71**	0.04	0.05	**0.21**	**0.63**	**0.22**	**0.28**
13	0.02	-0.08	0.02	0.03	0.00	0.00	0.02	0.03
14	0.09	-0.01	0.03	0.04	0.00	0.00	0.03	0.04
15	0.09	-0.01	0.03	0.04	0.00	0.00	0.03	0.04
16	-0.08	-0.10	0.04	0.04	0.00	0.01	0.04	0.04
17	-0.13	-0.16	0.01	0.01	0.00	0.00	0.01	0.01
18	0.26	0.25	0.15	0.14	0.18	0.25	0.18	0.17
19	0.24	0.18	0.01	0.01	0.01	0.01	0.02	0.01
20	-0.18	-0.25	0.04	0.04	0.02	0.05	0.04	0.04
21	**-0.26**	**0.15**	**0.41**	**0.77**	**1.03**	**7.38**	**0.71**	**3.42**
22	0.27	0.41	0.04	0.10	0.05	0.45	0.05	0.14
23	0.18	0.13	0.01	0.00	0.00	0.00	0.01	0.00
24	0.14	0.16	0.08	0.08	0.02	0.05	0.08	0.09
25	0.17	0.15	0.03	0.03	0.01	0.02	0.04	0.03
26	-0.01	-0.02	0.04	0.04	0.00	0.00	0.04	0.04
27	0.04	0.02	0.00	0.00	0.00	0.00	0.00	0.00
28	-0.07	-0.17	0.02	0.03	0.00	0.02	0.02	0.03
29	0.01	0.06	0.07	0.08	0.00	0.01	0.07	0.09
30	-0.11	-0.21	0.10	0.09	0.02	0.10	0.11	0.10
31	-0.26	-0.33	0.12	0.11	0.15	0.31	0.14	0.13
32	-0.30	-0.26	0.00	0.02	0.01	0.02	0.02	0.02
33	-0.01	-0.08	0.06	0.06	0.00	0.01	0.06	0.07
34	0.05	0.00	0.05	0.05	0.00	0.00	0.05	0.05
35	-0.07	0.00	0.08	0.10	0.01	0.00	0.08	0.11



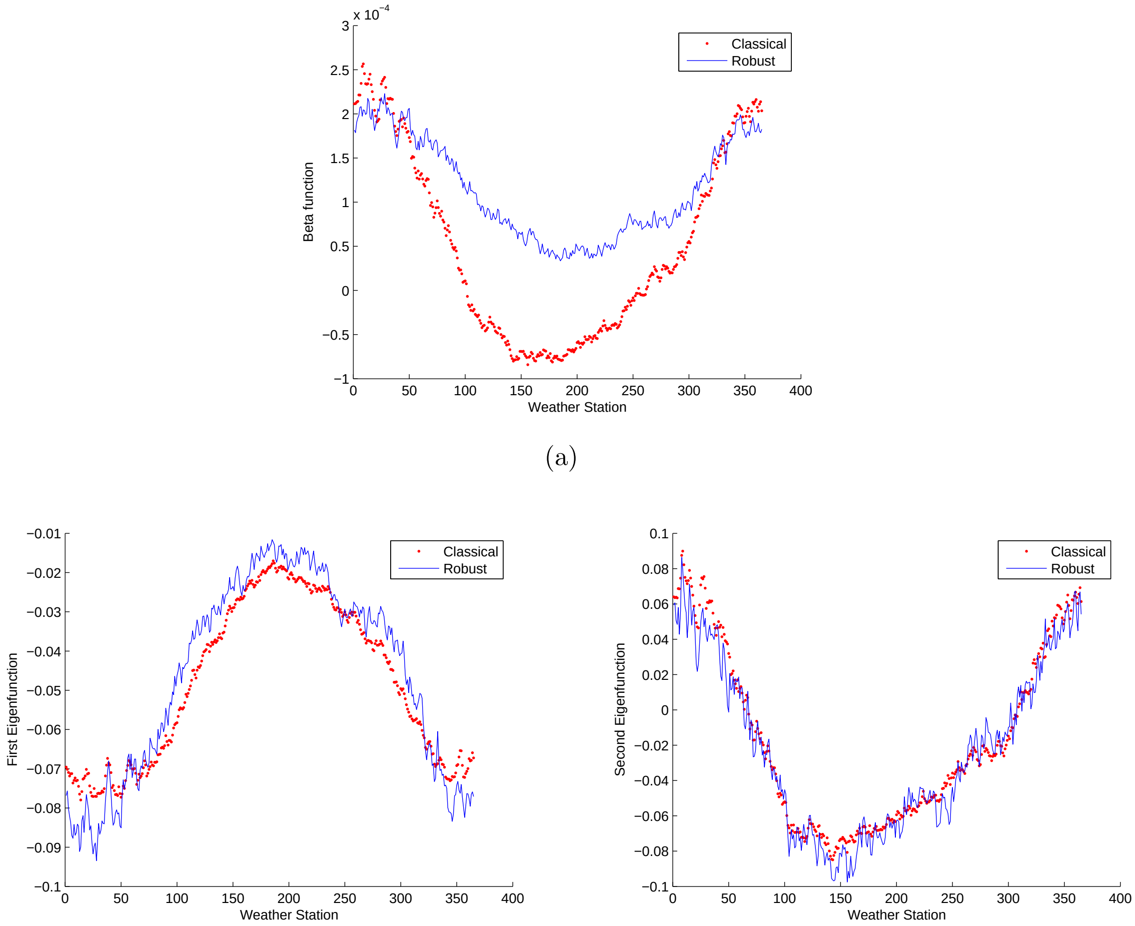
(a) (b)

Figure 4.2: (a)Sample curves (X-data) of the Canadian data; (b)Boxplot of the log-precipitation (y-data) data.



(a) (b)

Figure 4.3: Diagnostic plot of the Candian dataset based on (a)Classical Sore plot; (b)Robust Sore plot.

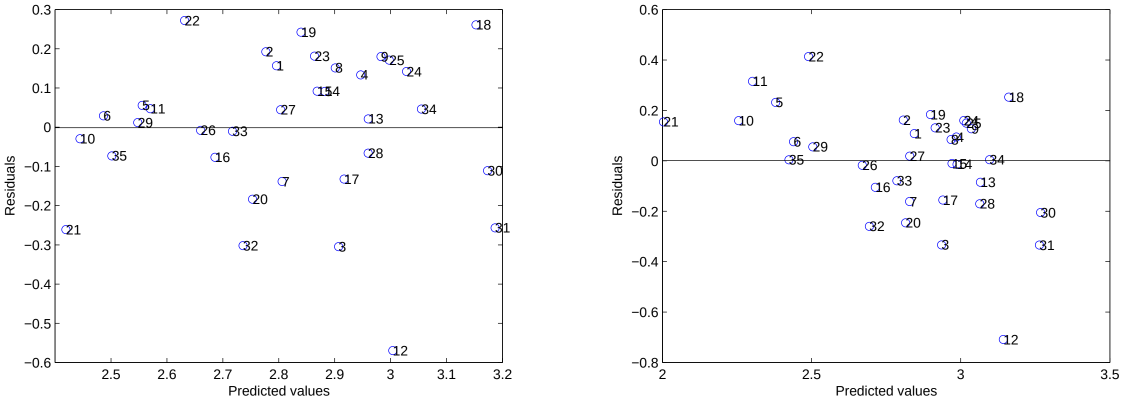


(a)

(b)

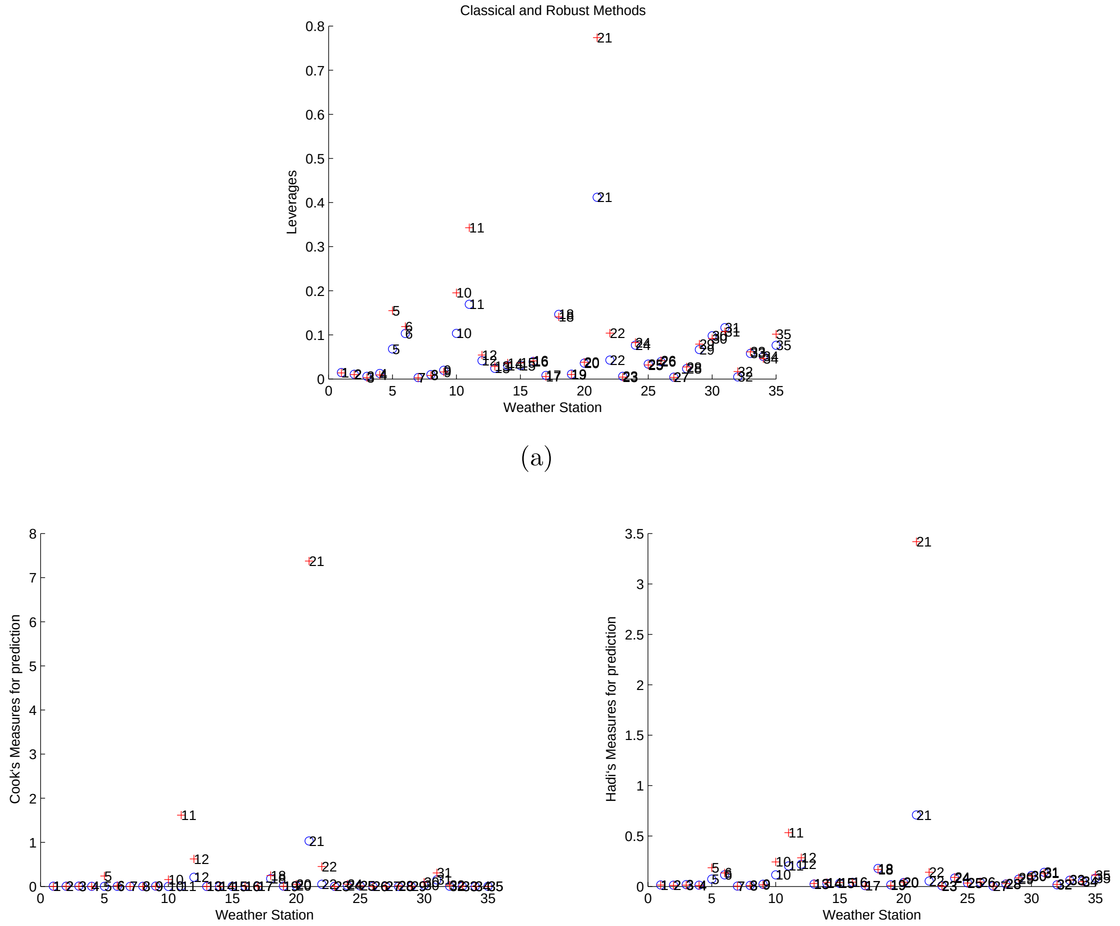
(c)

Figure 4.4: (a)Estimated beta function by Robust and classical FPCA; (b)First eigenfunction and (c)Second eigenfunction by Robust and Classical FPCR.



(a) (b)

Figure 4.5: Fitted values against the residuals (a)Classical; (b)Robust.

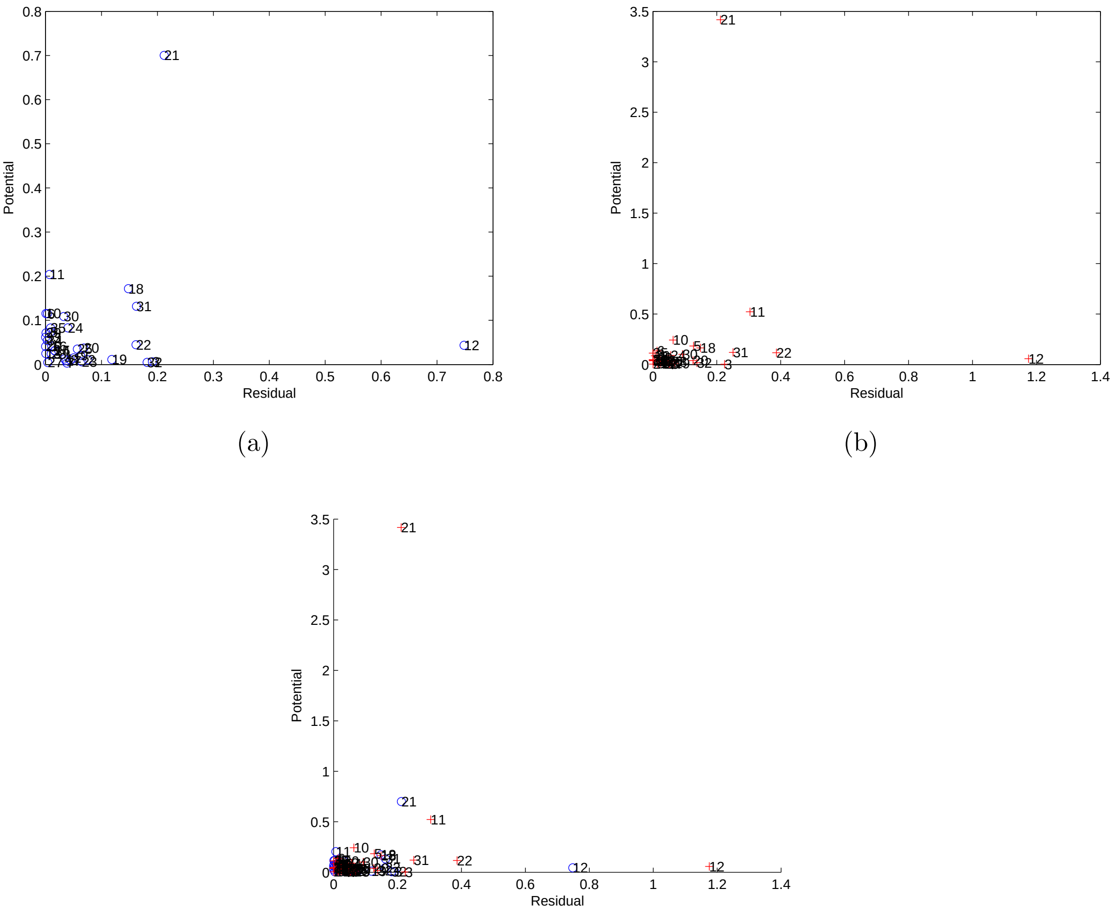


(a)

(b)

(c)

Figure 4.6: (a)Leverages; (b)Cook's Influence Measures; (c)Hadi's Influence Measures (Classical(o) and Robust(+)).



(a)

(b)

(c)

Figure 4.7: P-R plot (a)Classical(o); (b)Robust(+); (c)Classical(o) and Robust(+).

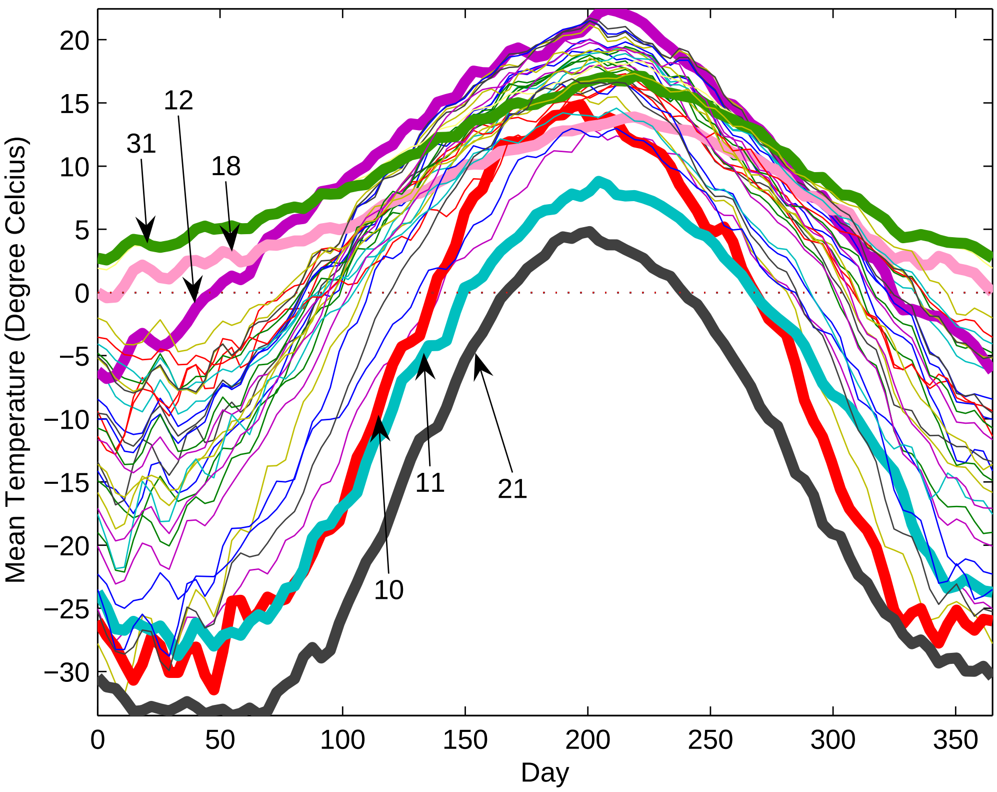


Figure 4.8: Sample curves of the Canadian data.

Chapter 5

Conclusion

Robust PCA based Functional Data Analysis has been developed for dimension reduction purpose. This method can also be used to detect functional outliers and classify these outliers in three classes.

An extensive simulation study was conducted and a real dataset was used to asses the performance of the robust FPCA. From this simulation study based on different contamination configurations (symmetric, asymmetric, partial and peak), we concluded that robust PCA based Functional Data Analysis yields better results than CPCA based Functional Data Analysis.

Besides, we have developed robust FPCR method and constructed robustified influence regression diagnostic measures. The practical use of robust FPCR and diagnostic measures are illustrated by means of a real data example. In view of this example, we concluded that the proposed robust FPCR method performs much better than classical FPCR and robustified influence measures appears to be more useful diagnostic tools for detecting heterogeneity in functional regression model with scalar response.

As future work, the different methods to estimate beta function can be explored. Simulation study can be conducted to assess the performance of the robust FPCR in presence of outliers, for a variety of scenarios. It is stated that the optimal number of bases is an important issue in FDA. Since, least squares method is sensitive to outliers, the choice of number of bases is affected by outliers. Robust version of this criteria for selecting number of bases can be determined.

BIBLIOGRAPHY

[1] Ash, R. B. and Gardner, M. F., *Topics in stochastic processes.* New York: Academic Press, 1975.

[2] Billor, Nedret, Hadi A. S., and Paul, F. Velleman P. F., "BACON: blocked adaptive computationally efficient outlier nominators", *Computatuional Statistics and Data Analysis*, 34, 279-298, 2000.

[3] Billor, N., Kiral, G. and Turkmen, A., "Outlier Detection Using Principal Components", *Twelfth International Conference on Statistics, Combinatorics, Mathematics and Applications*, Auburn, 2005. (Unpublished Manuscript).

[4] Cardot, H., Ferraty F. and Sarda, P., "Functional linear model", *Statistics and Probability Letters*, 45, 11-22, 1999.

[5] Cardot, H., Ferraty F. and Sarda, P., "Spline estimators for the functional linear model", *Statistica Simica*, 13, 571-591, 2003.

[6] Chatterjee, S., and Hadi, A. S., *Sensitivity Analysis in Linear Regression.* New York: Wiley, 1988.

[7] Chiou, J. M. and Müller, H. G., "Diagnostic for functional regression via residual processes", *Computational Statistics and Data Analysis*, 51, 4849-4863, 2006.

[8] Cook, R. D., "Detection of influential observations in linear regression", *Technometrics*, 19, 15-18, 1977.

[9] Craven, P. and Wahba G., "Smoothing noisy data with spline functions: estimating the correct degree of smoothing by the method of generalized cross-validation", *Numerische Mathematik*, 31, 377-403, 1979.

[10] Cuevas, A., Febrero, M., Fraiman R., "Linear Functional regression: the case of fixed design and functional response", *Canad. J. Statist*, 30, 285-300, 2002.

[11] Farawayy, J. J., "Regression analysis for a functional response", *Technometrics*, 39, 254-261, 1977.

[12] Fraiman, R. and Muniz, G., "Trimmed means for functional data", *Test*, 10, 419-440, 2001.

[13] Febrero, M., Galeano, P. and Gonzales-Mantegia, W., "A functional analysis of NOx levels: location and scale estimation and outlier detection", *Computational Statistics*, 22, 411-427, 2007.

[14] Febrero, M., Galeano, P. and Gonzales-Mantegia, W., "Outlier detection in functional data by depth measures, with application to identify abnormal NOx levels", *Environmetrics*, 19, 331-345, 2008.

[15] Febrero, M., Galeano, P. and Gonzales-Mantegia, W., "Influence in the Functional Linear Model with Scalar Response", In *Functional and Operatorial Statistics*, edited by Sophie Dabo-Niang and Frdric Ferraty. Physica-Verlag Heidelberg, 1 edition, 165-171, 2008.

[16] Febrero, M., Galeano, P. and Gonzales-Mantegia, W., "Measures of influence for the functional linear model with scalar response", *Journal of Multivariate Analysis*, In Press, Corrected Proof, Available online 25 December 2008.

[17] Hadi, A. S., "A new measure of overall potential influence in linear regression", *Computatuional Statistics and Data Analysis*, 14, 1-27, 1992.

[18] Hall, P. and Hosseini-Nasab, M., "On properties of functional principal components analysis", *Journal of the Royal Statistical Society*, Series B, 68 109-126, 2006.

[19] Huber, P. J. "Projection pursuit", *Annals of Statistics,* 13, no. 2, 435-525, 1985.

[20] Hubert, M., and Verboven, S., "A robust PCR method for high-dimensional regressors", *Journal of Chemometrics*, 17, 438-452, 2003.

[21] Hubert, M., and Engelen, S., "Robust PCA and classification in biosciences", *Bioinformatics*, 20, 1728-1736, 2004.

[22] Hubert, M., Rousseeuw, P. J. and Branden, K. V., "ROBPCA: A new approach to Robust Principal Component analysis", *Technometrics*, 47, no. 1, 64-79, 2005.

[23] Iglewicz, B. and Hoaglin D. C., "How to Detect and Handle Outliers", Milwaukee: American Society for Quality Control, WI, 1993.

[24] Joliffe, I. T. *Principal Component Analysis.* New York: Springer-Verlag, 1986.

[25] Kolmogorov, A. N., and Fomin, S. V., *Introductory real analysis* [English translation], Englewood Cliffs, NJ: Prentice-Hall, 1968.

[26] Levitin, D. J., Nuzzo, R. L., Vines, B. W. and Ramsay, J. O., "Introduction to Functional Data Analysis", *Canadian Psychology,* 48, no. 3, 135-155, 2007.

[27] Locantore, N., Marron, J. S., Simpson, D. G., Tripoli, N., Zhang, J. T. and Cohen, K. L., “Robust principal component analysis for functional data”, *TEST*, 8, no. 1, 1-73, 1999.

[28] Lopez-Pintado, S., and Romo J., “Depth based inference for functional data”, *Computational Statistics and Data Analysis*, 51, no. 10, 4957-4968, 2007.

[29] Pẽna, D. “A new statistic for influence in linear regression”, *Technometrics*, 47, 1-12, 2005.

[30] Ramsay, J. O. and Dalzell, C. J., “Some tools for functional data analysis”, *Journal of Royal Statistical Society*, 53, no. 3, 539-572, 1991.

[31] Ramsay, J. O. and Silverman, B. W., ”Functional Data Analysis Software”, MATLAB edition.
Online at http://www.psych.mcgill.ca/faculty/ramsay/software.html, 2001.

[32] Ramsay, J. O. and Silverman, B. W., *Functional Data Analysis.* Second Edition. New York: Springer-Verlag, 2005.

[33] Rice, J. A. and Silverman, B. W., “Estimating the mean and covariance structure nonparametrically when the data are curves”, *Journal of the Royal Statistical Society, Series B*, 53, 233-243, 1991.

[34] Roman, D. R. “A tool for Functional Data Analysis and Experimentation”, M.S. thesis, University of Puerto Rico, 2005.

[35] Rousseeuw, P. J. “Least median of squares regression”, *Journal of the American Statistical Association*, 79, 871-880, 1984.

[36] Rousseeuw, P. J. “Multivariate estimation with high break-down point”, In *Mathematical Statistics and Applications*, edited by Grossmann, W. *et al.*. Reidel, Dordrecht, Vol. B., 283-297, 1985.

[37] Rousseeuw, P. J. and Van Driessen, K., “A fast algorithm for the minimum covariance determinant estimator”, *Technometrics*, 41, 212-223, 1999.

[38] Shen, Q. and Xu, H., “Diagnostic for linear models with functional resposes”, *Technometrics*, 49, 26-33, 2007.

[39] Zhang, Jin-Ting. “Smoothed Functional Data Analysis”, PhD dissertation, The University of North Carolina at Chapel Hill, 1999.