Robust Statistical Methods for the Functional Logistic Model

by

Melody B. Denhere

A dissertation submitted to the Graduate Faculty of
Auburn University
in partial fulfillment of the
requirements for the Degree of
Doctor of Philosophy

Auburn, Alabama
August 3, 2013

Keywords: functional data, outliers, diagnostics, logistic regression

Approved by

Nedret Billor, Chair, Associate Professor of Mathematics and Statistics
Asheber Abebe, Associate Professor of Mathematics and Statistics
Peng Zeng, Associate Professor of Mathematics and Statistics

Abstract

Over the last decade or so, a lot of interest has emerged in the field of functional data analysis. This interest spans from a broad spectrum of fields such as brain imaging studies, bio-metrics, genetics, e-commerce and computer science. Statistical tools, models and methods, whose strength is in recognizing this structural aspect of data are being discussed and developed; ranging from functional linear regression, functional ANOVA, functional principal component analysis and functional outlier detection. In this work, we discuss statistical methods for the functional logistic regression model; a model where the response is binary and the covariate(s) functional. Essentially, we consider ways that allow for the parameter estimator to be resistant to outliers, in addition to eliminating multicollinearity and high dimensional problems; issues which are inherent with functional data. The methods include robust techniques of estimating the parameter function for the model as well as diagnostic measures to assess the fit of the model. Two estimation approaches are discussed; the first one makes use of robust principal component estimation techniques and the second one uses a robust penalization approach. Results from a simulation study and a real world example are also presented to illustrate the performance of the proposed estimators.

Acknowledgments

The completion of this dissertation would not have been possible without the support and encouragement from the following individuals who have been instrumental in my graduate studies at Auburn University.

To my advisor, Dr. Nedret Billor, I am thankful for the many discussions and insights provided in this work. For stirring the interest in this field of research, for knowing how to keep me going and for always being available with immeasurable guidance, support and encouragement. I sincerely appreciate that. To the supervisory committee made up of Drs. Ash Abebe and Peng Zeng, whose knowledge and advice in this work were invaluable.

To all my fellow classmates, and in particular Guy-vanie Miakonkana, Achard Bindele, Jasdeep Pannu, Brice Nguelifack, Pallavi Sawant, Ruchika Sabharwal and Julian Allagan; the journey would not have been the same without you. To Rose and Dan Brauss, you were God-sent! No words could ever describe how much I appreciate you both. For always being there and believing in me, especially when I became discouraged. The road trips, the movie nights, the endless talks, the many many many wonderful memories - I am forever grateful.

To my amazing sister Chiedza, you kept me sane through it all. Thank you for indulging me and being everything a sister could ask for and more. To my brothers Godfrey, Ngonidzashe and Michael, I appreciate your support and prayers. To my mom, for nurturing my love for numbers at an early age and guiding me to the One who gave me wings to fly. And to my dad, for seeing so much more in me than I ever did and showing me that even the sky is not the limit!

This work is dedicated to the memory of my grandfather, the late B.M. Denhere, who encouraged us to pursue education and excel in all we do. I know you would be proud *Sekuru.* To God be the Glory, great things He has done! (Psalm 100:5)

Table of Contents

List of Figures

List of Tables

Chapter 1

Functional Data Analysis

1.1 Introduction

In the last decade, a substantial amount of attention has been drawn to the field of functional data analysis; resulting in the development and generalization of many statistical techniques to this type of data. Much work has been devoted to this field with publications on functional regression models forthcoming from James (22), Cardot and Sarda (4); Müller and StadtMüller (29); Escabias et al. (9); Ferraty and Vieu (13) to name just a few. Ramsay and Silverman (34) present functional data ideas and statistical methods in their book which gave impetus to the functional data analysis community. Interest has been from a broad spectrum of fields such as biometrics, genetics, e-commerce and computer science. Statistical tools, models and methods, whose strength is in recognizing this structural aspect of the data have been discussed; ranging from functional linear regression, functional ANOVA, functional principal component analysis and functional outlier detection. In fact, functional regression methods have resulted in a re-look at some of the ways used to analyze longitudinal data.

There are many steps that arise in the analysis of functional data that differentiate it from analysis approaches taken for non-functional data. An important aspect to understand when it comes to functional data is primarily in observing properties of functional data. The term 'functional' is used to describe the intrinsic structure of the data; one thinks of the observed data functions as whole entities and the analysis is based on these functions. Naturally, the functional data is observed and recorded as discrete observations. However, this data is viewed as having been generated by some function and so the analysis is made on the function rather than viewing the data as a sequence of individual observations. Ramsay and Silverman (34) discuss the details of the first steps in functional data analysis (FDA) which include data representation issues, data registration and plotting pairs of derivatives. In this chapter, we look at some of the major concepts that are integral in FDA. In particular, we explore issues on estimation of the observed function from the discrete time points.

1.2 Estimation of Smooth Functions

We have already established that whilst analysis in FDA is based on observed functions, in practice the observed data is recorded as discrete observations. Therefore, for n observations, discrete recordings of (t_{ij}, X_{ij}) pairs are made and X_{ij} is viewed as a snapshot of the function $X_i(t)$ at time t_j where $i = 1, ..., n$ represents the number of observations and $j = 1, ..., n_i$ represents the number of replicates for each observation. It's important to note that time is not the only continuum over which functional data can be recorded. Other continua such as spatial, position and frequency may be involved and the continuum time is being used loosely.

An initial step in FDA is the estimation of the function $X_i(t)$ based on the observed discrete points $X_i(t_j)$. We assume that the function $X_i(t)$ is smooth and therefore possesses one or more derivatives. Without the smoothness property, not much is gained by treating the data as functional rather than multivariate. This smoothness property ensures that any two adjacent data values $X_i(t_j)$ and $X_i(t_{j+1})$ are unlikely to be very different from each other and are linked together to a certain extent. To simplify notation for this chapter, we will denote the function to be estimated as $X(t)$ since the estimation of the functional observation from the discrete data is done independently for each i. There are several issues to consider in estimating the smooth function $X(t)$ from the discrete observed data. The actual recorded data might be far from being smooth due to the presence of measurement error; the data might be sparsely sampled or few in number making it difficult to adequately give a stable estimate of the function; the time points at which the data is recorded might be unevenly spaced. All these and other issues call for techniques that work around these specific problems with the sampled data.

Since smoothness is a property that is essential for the function $X(t)$, there is need to consider smoothing techniques in estimating this function from the raw discrete data in the presence of measurement error or noise. The observed discrete data is expressed as,

$$X_j = X(t_j) + \varepsilon_j, \;\; j = 1, ..., n_i, \quad\; i = 1, ..., n,$$

where the measurement error ε_j causes a roughness to the observed data. One might choose to filter out the noise or to alternatively leave the noise but ensure that the smoothness property is met in the analysis of the results. An approach taken in this dissertation work in estimating the smooth function, $X(t_j)$, is by way of some basis function.

By definition, a basis function system is a set of known functions, ϕ_k, that are mathematically independent from each other with the property that one can approximate arbitrarily well any function by taking a linear combination of a sufficiently large number of these functions. Therefore, the basis function approach represents a function $X(t)$ (in $L^2(T)$) in terms of K known basis functions ϕ_k as,

$$\begin{aligned} X(t) &= \sum_{k=1}^{K} c_k \phi_k(t), \\ &= \boldsymbol{c}' \boldsymbol{\phi}, \end{aligned} \tag{1.1}$$

where $\mathbf{c}$ and $\boldsymbol{\phi}$ are column vectors of length K. Two important issues arise from this basis representation of a function. One of the important issues is the choice of the basis function system. The determination of the appropriate basis function system is based on the characteristics that are inherent in the data. In general, Fourier basis functions are used to model periodic data and B-spline basis is used for non-periodic data. Other basis choices such as wavelets, trigonometric functions or even polynomial functions can also be used should that be appropriate (Ramsay and Silverman (34)).

A second important aspect in this representation is the determination of the dimension of the expansion, K. An interpolation is achieved when $K = n_i$. Therefore, K can be viewed as a smoothing parameter, since the degree to which the data, X_j is smoothed is determined by the number of basis functions. When an appropriate basis function system is selected for the observed data, fewer basis functions are required to make a good approximation of the function, i.e. the smaller K is. There are two opposing discussions when it comes to choosing the number, K, of basis functions. On the one hand, the larger the K, the better the fit. This is however, at the expense of fitting unnecessary noise that should ideally be ignored. On the other hand, the smaller the K, the smoother the function (Ramsay and Silverman (34)). The problem with this is that if K is made to be too small, then we risk missing out on some important aspects of the functions that is being estimated. Therefore, a trade-off between fit and smoothness is necessary. An assortment of algorithms exist for this purpose; including methods such as cross-validation and generalized cross-validation (Craven and Wahba (7)).

In order to estimate the coefficients of the linear expansion in (1.1), the ordinary least squares criterion can be used. This criterion is expressed as,

$$\begin{aligned} SMSSE(\mathbf{X}|\mathbf{c}) &= \sum_{j=1}^{n_i}\left[X_j - \sum_{k=1}^{K} c_k\phi_k(t_j)\right]^2 \\ &= (\mathbf{X} - \mathbf{\Phi c})^T(\mathbf{X} - \mathbf{\Phi c}). \end{aligned} \tag{1.2}$$

From this, the estimate, $\hat{\mathbf{c}}$, that minimizes SMSSE is ,

$$\hat{\mathbf{c}} = (\mathbf{\Phi}'\mathbf{\Phi})^{\mathbf{-1}}\mathbf{\Phi}'\mathbf{X}.$$

This approach is appropriate when we assume the standard model for error, i.e. the errors, ε_j are independently and identically distributed with mean zero and constant variance σ_ε. In the presence of non-stationary and/or autocorrelated errors, the weighted least squares fit would be more appropriate. Thus, the weighted least squares estimate is,

$$\hat{\mathbf{c}} = (\mathbf{\Phi}'\mathbf{W}\mathbf{\Phi})^{\mathbf{-1}}\mathbf{\Phi}'\mathbf{W}\mathbf{X},$$

where $\mathbf{W}$ is a symmetric positive definite matrix that incorporates the unequal weighting of the errors.

The next section looks at a more powerful option for approximating discrete data by a function. The shortcomings of using (weighted) least squares criterion is in that the degree of smoothness is established in a discontinuous manner and better results can be established with a roughness penalty.

1.3 Smoothing with Roughness Penalty

As discussed in the previous section, there are two competing objectives in function estimation: good fit vs. smooth fit. The competing objectives can be viewed in terms of the following statistical principle,

$$MSE = Bias^2 + Variance,$$

where MSE is the mean squared error; and bias and sampling variance are defined as,

$$\begin{aligned} Bias[\hat{X}(t)] &= X(t) - E[\hat{X}(t)], \\ Var[\hat{X}(t)] &= E[(\hat{X}(t) - E[\hat{X}(t)])^2]. \end{aligned}$$

Whilst an unbiased estimate is desirable, when the standard error model is not met, the high variance associated with the curve will result in a compromised MSE. Therefore, the MSE might be drastically reduced by sacrificing some bias so as to reduce some sampling variance. This is achieved by imposing some smoothness to the estimated curve. Therefore, the roughness penalty makes explicit what is sacrificed in bias to achieve an improved MSE.

The penalized residual sum of squares, defines the compromise of trading smoothness for fit as,

$$PENSSE(\mathbf{X} \mid \mathbf{c}) = (\mathbf{X} - \mathbf{\Phi c})^{\mathbf{T}}\mathbf{W}(\mathbf{X} - \mathbf{\Phi c}) + \lambda \times Pen(X(t)),$$

where $\mathbf{c}$ is the coefficient vector; $\mathbf{\Phi}$ is the appropriate basis vector; $\mathbf{W}$ is a symmetric positive definite matrix that incorporates the unequal weighting of the errors; λ is a smoothing parameter that measures the rate of exchange between the fit to the data; and $\mathrm{Pen}(X(t))$ is a measure of the function's roughness. The curvature of a function is usually used to quantify the idea of the roughness of a function. Therefore, a natural measure of a function's roughness is given as,

$$\begin{aligned} Pen(X(t)) &= \int_T \left[X^{''}(s)\right]^2 ds \\ &= \mathbf{c}^{'}\mathbf{P}\mathbf{c}, \end{aligned}$$

where

$$\mathbf{P} = \int_T \phi^{''}(s)\{\phi^{''}(s)\}^T ds,$$

is the roughness penalty matrix. Thus, the objective function being optimized becomes,

$$PENSSE(\mathbf{X} \mid \mathbf{c}) = (\mathbf{X} - \mathbf{\Phi c})^{\mathbf{T}}\mathbf{W}(\mathbf{X} - \mathbf{\Phi c}) + \lambda \times \mathbf{c}^{'}\mathbf{P}\mathbf{c},$$

resulting in an estimate of the coefficient vector being,

$$\hat{\mathbf{c}} = (\mathbf{\Phi}^{'}\mathbf{W}\mathbf{\Phi} + \lambda\mathbf{P})^{-1}\mathbf{\Phi}^{'}\mathbf{W}\mathbf{X}.$$

When $\lambda = 0$, this reduces to the weighted least squares estimate. It can also be noticed that as $\lambda \to \infty$, and more weight is given to the roughness term, the estimated function approaches the standard linear regression; and as $\lambda \to 0$, the estimated function approaches an interpolation of the data. The selection of λ can be achieved by cross-validation and generalized cross-validation methods which will be discussed in more detail in later chapters of this dissertation work.

1.4 Conclusion

In this chapter we introduced the concept of functional data and discussed some of the pertinent issues in the field of functional data analysis. The rest of this work is organized as follows. Chapter 2 gives some background to the functional logistic model and discusses some of the estimation techniques in literature as well as making a case for the contribution of this work. Chapter 3 explores the concept of robust principal component estimation which is one of the results of this work. The estimation technique is developed and its performance considered in a simulation study as well as with a real world example. Chapter 4 examines and proposes a robust penalized approach to the estimation of the parameter function in the functional logistic model, with simulation results and application to a real world example. Chapter 5, studies diagnostic issues for the functional logistic model and discusses some model validation methods, extending results from the standard logistic model. In the last chapter, there is a discussion of other aspects of the functional logistic model that may be explored as future work.

Chapter 2

Functional Logistic Regression

2.1 Introduction

Ramsay and Silverman (34) discussed functional linear regression models, a case in which the linear relationship between random variables and functions is explored. The different models that can arise from this set-up include,

- A functional response and a scalar independent variable(s). In this model, the observed values are of the form $\{X_{ij}, Y_i(t) : t \in T\}$ for $i = 1, ..., n$, $j = 1, ..., p$ and where T is the support of the functional predictors. The relationship between the functional response and scalar predictors can be formulated as,

$$Y_i(t) = \sum_{j=1}^{p} X_{ij}\beta_j(t) + \epsilon_i(t), \quad i = 1, ..., n.$$

- A scalar response and a functional independent variable(s). The observed values in this case are of the form $\{X_i(t), Y_i\}$ for $i = 1, ..., n$ and this relationship formulated as

$$Y_i = \int_T X_i(t)\beta(t)dt + \epsilon_i, \quad i = 1, ..., n.$$

- A functional response and a functional independent variable(s). This last form occurs when both the response and predictor are functional, i.e. $\{X_i(t), Y_i(t)\}$ for $i = 1, ..., n$. The model is defined as,

$$Y_i(t) = \int_T X_i(t)\beta(t)dt + \epsilon_i(t), \quad i = 1, ..., n.$$

Our work is focused on the functional logistic regression (FLR) model, where the response is binary and the predictor(s) functional. Different approaches have been developed in the estimation methods of the functional parameters of the functional logistic model. Escabias et al. (9) discuss two approaches of parameter estimation that employ principal component estimation. James (22) and Müller and StadtMüller (29) discuss the generalized functional linear model and consider estimation methods for its parameters. These estimation techniques, however, are not resistant to outliers and thus there is a need for some robust estimation methods.

We focus on functional logistic regression in particular as there are some interesting problems in FDA that could benefit from this model. In addition to modeling functional covariates with binary responses, functional logistic regression is also useful for classification methods. An increasing amount of research is focusing on functional logistic regression and its application to functional Magnetic Resonance Imaging (fMRI) data. Ratcliffe et al. (36), Reiss et al. (37), Escabias et al. (10), Leng and Müller (25) and Tian (41) are some examples of work that deal with applications of the functional logistic model to a variety of areas including fMRI data.

The presence of outliers in any data set is almost inevitable and their effect on the model unquestionable. One school of thought proposes developing outlier detection methods and using these to eliminate identified outliers when fitting the model. However, there are many drawbacks to this approach. Firstly, there might be sample curves that are just at the limit of outlyingness, and their removal from the data set might be too drastic an approach to make. Over and above this, the process might need to be repeated before the data set can be declared outlier-free. Secondly, there might be no consistency among different researchers in the approaches taken to eliminate outliers and therefore, there is an element of subjectivity. Lastly, making inferences from the model where data has been removed causes problems in that regard due to the dependence between the 'good' observations and the 'bad' observations (Victoria-Feser (44)). Therefore, robust estimation methods are more viable in that the effect of the outliers is minimized without removing them from the data set. Whilst there has been some work that addressed estimation of the parameter function for this model as cited above, to our knowledge there's no contribution to the field as far as robust estimation methods for this model is concerned. It is with this void in mind that this dissertation work is dedicated to contributing to the field of functional data analysis by proposing robust estimation methods for the functional logistic model.

2.2 The Model

We consider having independent functional covariates as $X_1(t), ..., X_n(t)$ where $t \in T$ and T is the support of the sample curves $X_i(t)$, $i = 1, 2, ..., n$; and a random sample of a binary response variable Y to be $Y_1, ..., Y_n$, that is $Y_i \in \{0, 1\}$, $i = 1, 2, ..., n$. Then the random variable Y is such that the observations,

$$Y \sim Bernoulli(\pi_i)$$

where π_i is the probability of a positive response given $X_i(t)$ which is given as,

$$\begin{aligned} \pi_i &= P(Y = 1 \mid \{X_i(t) : t \in T\}) \\ &= \frac{exp\{\beta_0 + \int_T X_i(t)\beta(t)dt\}}{1 + exp\{\beta_0 + \int_T X_i(t)\beta(t)dt\}}, \quad i = 1, ..., n. \end{aligned}$$

β_0 is a real parameter and $\beta(t)$ is a smooth function of t, of which both are unknown parameters. The logit transformation is,

$$\begin{aligned} l_i &= log\left\{\frac{\pi_i}{1 - \pi_i}\right\} \\ &= \beta_0 + \int_T X_i(t)\beta(t)dt, \quad i = 1, ..., n. \end{aligned} \tag{2.1}$$

This can be viewed in the more general sense as a generalized functional linear model with the link as the logit function. In this work, we consider the no-intercept model (i.e. $\beta_0 = 0$).

2.3 Estimation of Regression Parameters

Estimation of the parameters cannot be achieved by the usual method of maximum likelihood [Ramsay and Silverman (34)] resulting in a different approach for these functional data. In particular, we consider the functional covariates, $X_i(t)$, and functional parameter, $\beta(t)$, as belonging to a finite-dimension space generated by the same (not necessarily) basis $\{\phi_j(t)\}_{j \in \mathbb{N}}$. We assume $X_i(t) \in L^2(T)$ of squared integrable functions with the inner product,

$$\langle f, g \rangle_u = \int_T f(s)g(s)ds, \;\; \forall f, g \in L^2(T),$$

such that

$$X_i(t) = \sum_{j=1}^{K_X} c_{ij}\phi_j(t), \;\; i = 1, ..., n, \tag{2.2}$$

where $\{\phi_j(t)\}$ is an appropriate basis which is selected to reflect the characteristics of the data.

With this assumption, we are able to reconstruct the functional form of the functional covariates from the observed discrete points using two different approaches. On the one hand, if the functional covariate(s) is observed with some measurement error, then the i^{th} subject at the k^{th} replication is

$$X_{ik} = X_i(t_k) + \varepsilon_i(t_k), \quad k = 0, 1, .., n_i, \quad i = 1, ..., n,$$

where $\varepsilon(t)$ and $X(t)$ are independent. In this case where the functional covariate is observed with some noise, we use some least squares approximation approach to obtain the functional form of the covariates by approximating the basis coefficient $\{c_{ij}\}$ from the discrete observations. On the other hand, if the functional covariate(s) is observed without error, then the i^{th} subject at the k^{th} replication is

$$X_{ik} = X_i(t_k), \quad k = 0, ..., n_i, \quad i = 1, ..., n.$$

In this case, some interpolation method such as cubic spline interpolation can be used to get the functional form of the predictors.

We also define,

$$\beta(t) = \sum_{k=1}^{K_b} b_k \varphi_k(t), \tag{2.3}$$

where $\varphi_k(t), \ k = 1, ..., K_b$ is a known basis function.

Since estimates for the basis coefficient $\{c_{ij}\}$ can be found either by smoothing (as discussed in Chapter 1) or interpolation, using the basis expansion of the covariates and parameter function as defined in (2.2) and (2.3) in the regression model (2.1), the functional logistic model reduces to a standard multiple logistic one,

$$\begin{aligned} l_i &= \beta_0 + \int_T X_i(t)\beta(t)dt \\ &= \beta_0 + \sum_{j=1}^{K_X} \sum_{k=1}^{K_b} c_{ij}\psi_{jk} b_k, \quad i = 1, ..., n, \end{aligned}$$

where $\psi_{jk} = \int_T \phi_j(t)\varphi_k^{'}(t)dt$ for $j = 1, ..., K_X, k = 1, ..., K_b$; c_{ij} is the basis coefficient; β_0 is an unknown real parameter; b_k is the unknown basis coefficient used to estimate the parameter function $\beta(t)$. In matrix form, this can be written as,

$$\mathbf{L} = \beta_0 \mathbf{1} + \boldsymbol{C\psi b}, \tag{2.4}$$

where $\mathbf{L}= (l_1, ..., l_n)^{'}$, $\mathbf{C}= \{c_{ij}\}_{n\times K_X}$, $\boldsymbol{\psi} = \{\psi_{jk}\}_{K_X\times K_b}$, $\mathbf{1}= (1, ..., 1)^{'}$ and $\boldsymbol{b} = (b_1, ..., b_{K_b})^{'}$.

Three different approaches of estimation can be discussed for this model (2.4). Firstly, there is the maximum likelihood estimation (MLE) which results in unstable and inaccurate estimators due to the higher collinear nature of the covariates. To deal with this shortcoming of the MLE, the second approach of principal component estimation as discussed by Escabias et al. (9) can be considered. Alternatively, a third approach is the penalized maximum likelihood estimation approach which is also explored in this work; and an empirical comparison of the three approaches is given by way of a simulation study.

2.3.1 Maximum Likelihood Estimation

The FLR model in (2.4) is now in the same form as the standard logistic regression model and, therefore, the maximum likelihood parameter estimates can be found. The responses, Y_i, are assumed to follow a Bernoulli(π_i) distribution, such that the log-likelihood function can be written as,

$$L(\pi; Y) = \sum_{i=1}^{n} \left[Y_i log \left(\frac{\pi_i}{1-\pi_i} \right) + log(1-\pi_i) \right]. \tag{2.5}$$

The derivative of this log-likelihood function w.r.t. π_i is

$$\frac{\partial L}{\partial \pi_i} = \sum_{i=1}^{n} \frac{Y_i - \pi_i}{\pi_i(1-\pi_i)}.$$

Using the fact that,

$$log \left(\frac{\pi_i}{1-\pi_i} \right) = l_i = \beta_0 + \mathbf{C}_i^{'} \boldsymbol{\psi} \mathbf{b},$$

together with the chain rule, the derivative of the log-likelihood w.r.t. β_0 is,

$$\begin{aligned}\frac{\partial L}{\partial \beta_0} &= \sum_{i=1}^{n} \frac{\partial L}{\partial \pi_i}\frac{\partial \pi_i}{\partial l}\frac{\partial l}{\partial \beta_0} \\ &= \sum_{i=1}^{n} \frac{Y_i - \pi_i}{\pi_i(1-\pi_i)} \times \pi_i(1-\pi_i) \times 1 \\ &= \sum_{i=1}^{n} (Y_i - \pi_i).\end{aligned}$$

Similarly, the derivative of the log-likelihood w.r.t. $\mathbf{b}$ is,

$$\begin{aligned}\frac{\partial L}{\partial \mathbf{b}} &= \sum_{i=1}^{n} \frac{\partial L}{\partial \pi_i}\frac{\partial \pi_i}{\partial l}\frac{\partial l}{\partial \mathbf{b}} \\ &= \sum_{i=1}^{n} \frac{Y_i - \pi_i}{\pi_i(1-\pi_i)} \times \pi_i(1-\pi_i) \times \mathbf{C}_i^{'}\boldsymbol{\psi} \\ &= \sum_{i=1}^{n} (Y_i - \pi_i)\mathbf{C}_i^{'}\boldsymbol{\psi}.\end{aligned}$$

Therefore the likelihood equations for model (2.4), are:

$$\sum_{i=1}^{n} [Y_i - \pi_i] = 0. \tag{2.6}$$

$$\sum_{i=1}^{n} \Big[(Y_i - \pi_i)\mathbf{C}_i^{'}\boldsymbol{\psi}\Big] = 0. \tag{2.7}$$

Since these equations are non-linear in the parameters of interest, β_0 and $\mathbf{b}$, an iterative method is used to solve these equations and get estimates of the parameters. One such method is the Newton-Raphson method which is used to solve these non-linear likelihood equations,

$$\mathbf{X}^{'}(\mathbf{Y} - \boldsymbol{\pi}) = \mathbf{0},$$

where from (2.4), $\mathbf{X}$=($\mathbf{1}|$ $\mathbf{C}\boldsymbol{\psi}$), $\mathbf{Y}$= $(Y_1, ..., Y_n)^{'}$ and $\boldsymbol{\pi} = (\pi_1, ..., \pi_n)^{'}$. The approximated parameter function estimate for our functional logistic model will then be,

$$\begin{aligned}\hat{\beta}(t) &= \sum_{k=1}^{K_b} \hat{b}_k \varphi_k(t) \qquad (2.8)\\ &= \hat{\boldsymbol{b}}' \boldsymbol{\varphi},\end{aligned}$$

where $\boldsymbol{\varphi}$ is a known basis function and $\hat{\boldsymbol{b}}$ is the maximum likelihood estimate of the reduced functional logistic model (2.4).

The estimation of the parameter function obtained using the maximum likelihood approach is not very accurate in the presence of highly correlated data. In fact, the design matrix, $\boldsymbol{C\psi}$ in (2.4),has high correlation among its columns. Thus, there is a need to eliminate multicollinearity in order to obtain a more reliable estimation of the parameter function, $\beta(t)$ in our functional model. One such approach is discussed by Escabias et al. (9) and uses the idea of principal component estimation. Another well-known adapted approach which we explore in this chapter is ridge estimation.

2.3.2 Principal Component Estimation

The reduced functional logistic model in (2.4) is considered and in order to deal with the multicollinearity problem highlighted, the principal components (PCs) of the design matrix $\boldsymbol{C\psi}$ are utilized instead. We let $\mathbf{Z} = \{\xi_{ij}\}_{n \times K_b}$ be the matrix of PCs of the design matrix, such that

$$\mathbf{Z} = \boldsymbol{C\psi V},$$

where $\mathbf{V}$ is a $K_b \times K_b$ matrix whose columns are the eigenvectors associated with the eigenvalues of the covariance matrix of $\boldsymbol{C\psi}$. Then the model becomes,

$$\mathbf{L} = \beta_0\mathbf{1} + \mathbf{Z}\boldsymbol{\gamma}, \tag{2.9}$$

where $\boldsymbol{\gamma} = \mathbf{V}^{'}\mathbf{b}$. As with the MLE, the $\beta(t)$ parameter function is then estimated as,

$$\hat{\beta}(t) = \hat{\boldsymbol{b}}'\boldsymbol{\varphi},$$

where $\hat{\boldsymbol{b}} = \boldsymbol{V}\hat{\boldsymbol{\gamma}}$. A selected number of PCs are included in the model instead, so that (2.9) becomes,

$$\mathbf{L}^{(s)} = \beta_0\mathbf{1}^{(s)} + \mathbf{Z}^{(s)}\boldsymbol{\gamma}^{(s)}, \tag{2.10}$$

where s denotes the number of selected PCs retained in the model. The number of selected PCs can be determined either by the cross-validation method or the generalized cross-validation method. These and other methods for selecting the number of PCs are discussed in greater detail in Chapter 3.

2.3.3 Ridge Estimation

A second approach that can be used to deal with the highly correlated covariates in model (2.4), would be in the form of penalized maximum likelihood estimation. Le Cessie and van Houwelingen (23), extended the ridge regression theory used in standard linear regression to logistic regression. We adopt this approach to define a ridge estimator for the reduced functional logistic regression model. In earlier work, Schaefer (39) also discussed the ridge logit estimator to handle collinear data. In that work, the parameter estimator is given as,

$$\hat{\beta}^{\nu} = (\boldsymbol{X}'\boldsymbol{V}\boldsymbol{X} + \nu\mathbf{I})^{-1}\boldsymbol{X}'\boldsymbol{V}\boldsymbol{X}\hat{\beta}_{MLE},$$

where $\hat{\beta}_{MLE}$ denotes the maximum likelihood estimator of the parameter; $\mathbf{X}$ is the design matrix; $\mathbf{V}$ is the diagonal matrix of MLEs of success probabilities; $\mathbf{I}$ is the identity matrix and ν is the ridge parameter.

We consider the reduced functional logistic regression model (2.4) obtained by considering the functional observations, $X_i(t)$, and the parameter function, $\beta(t)$, as belonging to the finite-dimensional space generated by the bases $\{\phi_j(t)\}_{j=1}^{K_X}$ and $\{\varphi_k(t)\}_{k=1}^{K_b}$, respectively. Under this model, the log-likelihood function as stated earlier is,

$$L(\pi; Y) = \sum_{i=1}^{n} \left[Y_i log \left(\frac{\pi_i}{1-\pi_i} \right) + log(1-\pi_i) \right].$$

Using the approach discussed by Le Cessie and van Houwelingen (23), we consider the log-likelihood function above with a penalty on the norm of $\mathbf{b}$, the unknown parameter in model (2.4) as,

$$L_{\nu}(\pi; Y) = L(\pi; Y) - \nu\|\mathbf{b}\|^2, \tag{2.11}$$

where $\|\mathbf{b}\| = \left(\sum_{k} b_k^2 \right)^{\frac{1}{2}}$. Thus, instead of maximization of the log-likelihood function(2.5), the maximization is on $L_{\nu}(\pi; Y)$ with a penalty on the norm of $\mathbf{b}$. In this case, ν controls the shrinkage of the norm of the parameter $\mathbf{b}$. When $\nu = 0$, the solution will be the ordinary MLE. As the ridge parameter, ν, approaches ∞, the b_k's tend to 0. As with the concept of ridge regression, the idea is to introduce some bias so as to get a biased estimator but whose variance is smaller than that of the unbiased estimator (MLE in this case). The choice of ν would be such that $MSE(\hat{\beta}^{\nu}) < MSE(\hat{\beta}_{MLE})$.

Therefore the likelihood equations for model (2.11), are:

$$\sum_{i=1}^{n} [Y_i - \pi_i] = 0.$$

$$\sum_{i=1}^{n} \Big[(Y_i - \pi_i)\mathbf{C}_i^{'}\boldsymbol{\psi}\Big] - 2\nu\mathbf{b} = 0. \tag{2.12}$$

These likelihood equations, which are non-linear in the parameters, are solved by some iterative method to obtain the ridge estimator, $\hat{\mathbf{b}}^{\nu}$. Much discussion on the choice of the ridge parameter, ν, has been made and one approach is based on minimizing an estimate of the prediction error of the model. Le Cessie and van Houwelingen (23) consider different measures that quantify the prediction error. Such measures include,

(i) Classification Error (CE),

$$CE = \begin{cases} 1 & : Y_{new} = 1 \ and \ \hat{\pi}_{new} < 0.5 \\ \frac{1}{2} & : \hat{\pi}_{new} = 0.5 \\ 0 & : otherwise \end{cases}$$

(ii) Squared Error,

$$SE = (Y_{new} - \hat{\pi}_{new})^2.$$

(iii) Minus Log-Likelihood Error

$$ML = -\{Y_{new} log(\hat{\pi}_{new}) + (1 - Y_{new}) log(1 - \hat{\pi}_{new})\},$$

where Y_{new} denotes the response for a new observation and $\hat{\pi}_{new}$ denotes the probability that the response is positive for a new covariate $X_{new}(t)$. Other measures include use of the cross-validation method and this criterion is the one we used in the simulation study presented in the next section. This is defined for the mean squared error as,

$$MSE_{CV} = \frac{1}{n}\sum_{i=1}^{n}\{Y_i - \hat{\pi}_{(-i)}\}^2,$$

where $\hat{\pi}_{(-i)}$ is the estimated probability of a positive response, when the model is fit without the i^{th} observation.

We compare the performance of each of these estimation methods discussed in this section by way of a simulation study.

2.4 Simulation Study

The following steps were carried out in this simulation study in order to investigate how the different estimation techniques discussed in the previous section compare in the estimation of the parameters of the functional logistic regression model.

Step 1: Generate the simulated data.

(a) Generate n sample curves for the functional logistic model.

We generate 50 sample functional observations of a known stochastic process $X(\cdot)$ considered over the interval $[0, 10]$ which has 21 equally spaced time points. We define this stochastic process as $X(t) = Z(t)+t/4+E$ where $Z(t)$ is a zero mean Gaussian stochastic process with covariance function $C(s,t) = (\frac{1}{2}) \times (\frac{1}{2})^{80|s-t|}$ and E is a Bernoulli random variable with $p = 0.1$. This model is illustrated in Figure 2.1. To obtain the functional form of these sample curves, we consider them to belong to the finite-dimensional space generated by a cubic B-spline basis defined on equally spaced nodes on the $[0, 10]$ interval. We used the generalized cross-validation (GCV) method to determine the number of basis functions (K_X) to use as well as the smoothing parameter (λ) in the estimation of the functional form of the covariate, $X(t)$. The criterion is expressed as,

$$GCV(\lambda) = \left(\frac{n}{n - df(\lambda)}\right)\left(\frac{SSE}{n - df(\lambda)}\right)$$

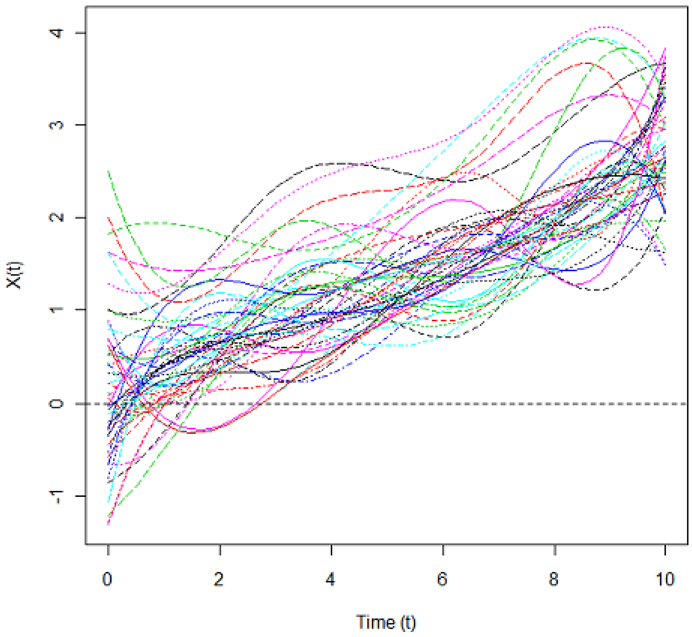


Figure 2.1: 50 sample curves generated from the stochastic process $X(t) = Z(t) + t/4 + E$

where $df(\lambda)$ =trace $\{\mathbf{S}_{\phi,\lambda}\}$ is the equivalent degrees of measure and $\mathbf{S}_{\phi,\lambda}$ is the projection operator such that $\hat{\mathbf{X}} = \mathbf{S}_{\phi,\lambda}\mathbf{X}$; SSE is the residual sum of squares; and λ is the smoothing parameter that measures the rate of exchange between the fit to the data. The optimal number of basis was achieved at $K_X = 10$ with $\lambda = 2048$, which is what was used for all the replications.

(b) Select a parameter function.

The natural cubic spline interpolation of the parameter function $sin(t+\pi/4)$ is selected. The basis coefficients, $\mathbf{b} = (b_1, ..., b_{K_b})'$ of the parameter function are known and used to assess estimation techniques presented in this work.

(c) Generate the values of the response variable.

The probabilities are,

$$\pi_i = \frac{exp\{l_i\}}{1 + exp\{l_i\}},$$

where l_i is as defined in (2.1), β_0 is fixed at 0 and $i = 1, .., n$. The n values of the response are obtained by simulating observations of a Bernoulli distribution with probabilities π_i.

The reduced functional logistic model in (2.4) was fit with the coefficient matrix as the covariate. The Hosmer-Lemeshow goodness of fit test was also carried out [Hosmer and Lemeshow (20)], which confirmed validity of the model for the generated data.

Step 2: Obtain the approximated estimations for the parameter function.

(a) Use maximum likelihood estimation with the coefficient matrix as the covariate of the logistic model.

(b) Perform PCA on the covariates in (2.4) and fit the logistic model with the retained s PCs as the covariates.

(c) Use penalized maximum likelihood estimation with the coefficient matrix as the covariate.

Small values of MSEB indicate better estimation when the three different estimation techniques are compared.

All simulations were implemented using R [R Development Core Team (33)]. The packages *fda*[Ramsay et al. (35)], *rrcov*[Todorov (42)] and *penalized* [Goeman (17)] were particularly useful in obtaining the functional form of the data; performing principal component estimation and penalized maximum likelihood estimation, respectively. The cross-validated likelihood criterion was used to determine the optimal shrinkage parameter for the penalized ML estimation.

The median estimates of $\beta(t)$ for 250 simulations are shown in Figures (2.2) and (2.3), illustrating how the estimates compared with the true estimates of the function parameter. As is evident in the graphs, the ML estimation method does not give accurate estimates since no consideration is taken for the multicollinearity issue that is inherent in the design matrix.

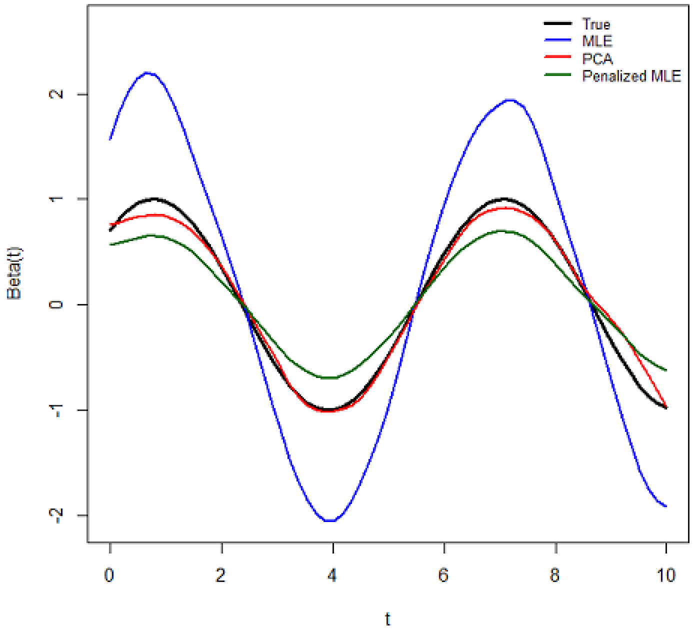
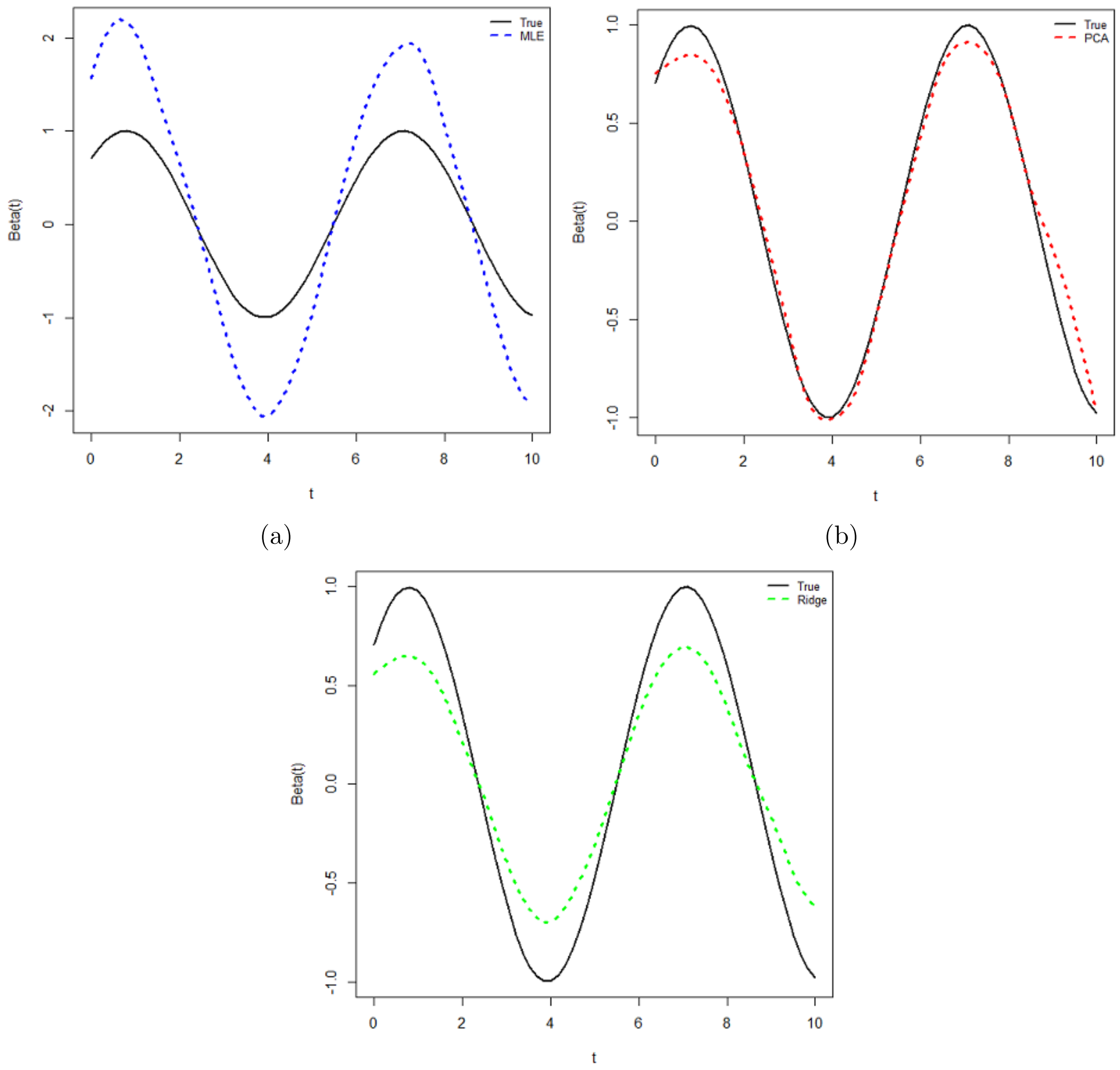


Figure 2.2: Comparison of the median $\beta(t)$ estimates for the three different techniques against the true parameter function $sin(t + \pi/4)$.

Both the PC-based and penalized ML estimation approaches give more accurate results, with the estimates of the parameter function being much closer to the true $\beta(t)$ function. This is especially evident when the mean $\beta(t)$ estimates (results not shown) are used. The median MSEB for each of the techniques is 2.1311 for MLE; 0.1762 for PC estimation; and 0.2934 for the ridge estimation.

2.5 Conclusion

In this chapter, we have introduced the FLR model and reviewed some of the estimation methods discussed in literature. It has been shown that due to the formulation of the design matrix of the FLR model developed in this work, the estimation of the parameter function, $\beta(t)$, is greatly affected by the presence of multicollinearity. Two approaches discussed to minimize the collinear covariates were those of principal component estimation [Escabias et al. (9)] and of penalized maximum likelihood estimation. These alternative approaches provide more reliable estimates of $\beta(t)$. These results are consistent with literature [Barker



(a)

(b)

(c)

Figure 2.3: Median $\beta(t)$ estimates (a) Maximum Likelihood Estimation (b) Principal Component Estimation and (c) Ridge Estimation

and Brown (2), Le Cessie and van Houwelingen (23), Escabias et al. (9) and Vágó and Kemenéy (43)] and we conclude that estimation of the function parameter $\beta(t)$ can be established with better accuracy using principal component estimation or penalized maximum likelihood estimation techniques. However, in the presence of outlying sample curves, there is a need to consider robust estimation methods. In chapters 3 and 4, we turn our focus to this by proposing robust methods for the functional logistic regression model.

Chapter 3

Robust Principal Component Functional Logistic Regression

3.1 Introduction

Inherent with most functional data, and the estimation techniques discussed in this work, is the fact that we are often dealing with highly correlated data, which poses problems in parameter estimation resulting in estimations that might be far from accurate and unreliable. The presence of functional observations that deviate from the overall pattern of the data, creates additional inefficiencies in many procedures. In this work, we propose a robust estimation approach for functional data with a binary response and functional covariates that addresses these inadequacies. To our knowledge, there is no work that has been done in this area of robust estimation for the functional logistic regression model. There are some robust methods proposed for the functional linear regression model and for functional principal component analysis. Boente and Fraiman (3), Gervini (15), Gervini (16), Bali et al. (1), Sawant et al. (38) and Lee et al. (24) are examples of such work. This work is different from these others in that it looks at robust methods for the functional logistic regression model, for which no known robust approach has been proposed. However, functional outliers are inevitable and therefore it is important that robust techniques for this model be developed.

The first approach taken in this chapter to develop robust estimators is that of using robust principal component estimation. The functional logistic model is reduced to a standard multiple logistic model by approximating the functional covariates as a linear combination of an appropriate basis [Ramsay and Silverman (34)]. This estimation approach of the functional data results in collinear and (potentially) high dimensional data and therefore, an initial focus is the elimination of such problems. Some of the proposed approaches in literature that eliminate these issues are by way of principal component estimation or ridge estimation as discussed in Chapter 2. However, in the presence of atypical curves, these estimations are unstable and inaccurate. Febrero et al. (11) define a functional outlier as that curve that has been generated by a stochastic process with a different distribution than the rest of the curves, which are assumed to be identically distributed. This definition appears to be all-encompassing as it refers to those curves that could be far away from most of the curves; curves that have a different pattern from the rest or even those curves that are atypical in some sub-interval of the period of interest. With real word data, outliers are inevitable and this makes it necessary to develop robust estimators, of which we consider robust estimation techniques using principal component estimation.

3.2 Estimation of $X(t)$ and $\beta(t)$

We consider the functional covariates, $X_i(t)$, and functional parameter, $\beta(t)$, as belonging to a finite-dimensional space generated by (not necessarily) the same basis $\{\phi_j(t)\}_{j=1}^{K_X}$ and $\{\varphi_k(t)\}_{k=1}^{K_b}$, respectively. We consider $X_i(t) \in L^2(T)$ with the usual inner product, such that

$$X_i(t) \approx \sum_{j=1}^{K_X} c_{ij}\phi_j(t), \tag{3.1}$$

where $\phi_j(t), \;\; j = 1, ..., K_X$, is an appropriate basis, selected to reflect the characteristics and main features of the data. As discussed in Chapter 1, the truncation lag, K_X, is a parameter that is selected based on the features and characteristics of the data. This determines the dimension of expansion; the larger K_X is, the better the fit to the data is. However, this can result in problems of over-fitting and therefore capturing noise or variation in the data that might be ignored. Smaller K_X on the other hand, whilst being desirable should not be too small such that there is a risk of overlooking important features of the data.

The selection of the basis system is an important aspect in functional data in that the features that might be evident in the observed data should be adequately met by the selection of an appropriate basis system. A good basis system selection potentially results in a smaller K_X which means less computational time in estimating the functional covariate and ensuring that the basis coefficients, c_{ij}, serve as interesting descriptors of the given data. In general, the Fourier basis functions are used to model periodic data and the B-spline basis is used for non-periodic data. However, this is by no means the status-quo and great consideration needs to be made in this regard. Thus, with the assumption that $X_i(t)$ belongs to the Hilbert space, we are able to reconstruct the functional form of the functional covariates from the observed discrete points using two different approaches - either by smoothing or interpolating.

We also define the functional parameter,

$$\beta(t) = \sum_{k=1}^{K_b} b_k \varphi_k(t), \tag{3.2}$$

where $\varphi_k(t),\ \ k = 1, ..., K_b$ is a known basis function; and K_b is selected in such a way that $K_X \geq K_b$. Since the estimate for the basis coefficient $\{c_{ij}\}$ can be found either by smoothing or interpolation, using the basis expansion of the covariates and parameter function as defined in (3.1) and (3.2) in the regression model (2.1), the functional model reduces to a standard multiple one,

$$\begin{aligned} l_i &= \beta_0 + \int_T X_i(t)\beta(t)dt \\ &= \beta_0 + \sum_{j=1}^{K_X} \sum_{k=1}^{K_b} c_{ij}\psi_{jk} b_k,\ \ i = 1, ..., n, \end{aligned}$$

where $\psi_{jk} = \int_T \phi_j(t)\varphi_k^{'}(t)dt\,; j = 1, ..., K_X, k = 1, ..., K_b$; c_{ij} is the basis coefficient; β_0 is an unknown real parameter; b_k is the unknown basis coefficient used to estimate the parameter function $\beta(t)$. In matrix form, this can be written as,

$$\mathbf{L} = \beta_0 \mathbf{1} + \boldsymbol{C\psi b}, \tag{3.3}$$

where $\mathbf{L}= (l_1, ..., l_n)^{'}$, $\mathbf{C}= \{c_{ij}\}_{n\times K_X}$, $\boldsymbol{\psi} = \{\psi_{jk}\}_{K_X\times K_b}$, $\mathbf{1}= (1, ..., 1)^{'}$ and $\boldsymbol{b} = (b_1, ..., b_{K_b})^{'}$.

We develop the principal component estimation technique discussed in Escabias et al. (9) further to cater for cases where there are functional outliers in the data. Due to the fact that principal component estimation makes use of the eigen decomposition of the covariance matrix of the design matrix, $\boldsymbol{C\psi}$, the presence of outliers will greatly influence the PCs. This sensitivity to outliers results in the first few PCs being attracted towards the outliers, and therefore this approach might miss the main modes of variability of the rest of the observations.

Our proposed approach uses robust PC estimation techniques that eliminate multicollinearity and reduces the effect of functional outliers, resulting in a more accurate estimator in the presence of outliers. We use robust PCA methods on the covariate matrix to obtain robust PCs which are used as the design matrix in the standard multiple logistic model. Robust Principal Component Analysis, ROBPCA, by Hubert et al. (21) is one such approach which uses projection pursuit and robust covariance estimation based on the Minimum Covariance Determinant (MCD) method. The MCD method is based on seeking an h-subset of observations whose classical covariance matrix has the smallest determinant.

The three basic steps in ROBPCA are given in the following algorithm:

Input: Data matrix $\mathbf{A} = \mathbf{C}\boldsymbol{\psi}_{n\times K_b}$ where n is the number of observations and K_b represents the initial number of variables.

Output: Robust PC scores $\mathbf{Z}_{n\times k_0}$ where $k_0 < K_b$ is the number of eigenvectors retained.

1. A singular value decomposition (SVD) of the data is performed so as to project the observations on the space spanned by themselves. This step is especially useful when $K_b \geq n$ as it yields a huge dimension reduction.

2. A measure of outlyingness is defined for each data point. This is achieved by projecting all the points onto many univariate directions, $\mathbf{v}$, and then determining the

standardized distance of each projected point to the center of the data. The $h(< n)$ least outlying data points are determined and retained, where outlyingness is defined as,

$$Out(\mathbf{a}_i) = max_v \frac{\mid \mathbf{a_i^{'}v} - \hat{\mu}_r \mid}{\hat{\sigma}_r}, \qquad i = 1, ..., n,$$

where $\hat{\mu}_r$ and $\hat{\sigma}_r$ are the univariate MCD based location and scale estimates for the projected data points, $\mathbf{a_i^{'}v}$, respectively. The h data points are projected on the subspace spanned by the first k_0 eigenvectors of the sample covariance matrix of the h-subset.

3. The covariance matrix of the mean-centered data matrix, $\mathbf{A}^*_{n\times k_0}$, obtained in the second step using the MCD estimator is robustly estimated and PCA is applied on to the data matrix.

We consider the functional logistic regression model as defined before in (2.1) where the functional covariate is defined as (3.1), resulting in the standard multiple logistic regression (3.3). We let $\mathbf{Z}_{(r)} = \{\xi_{ij}\}_{n\times k_0}$ be the matrix of robust PCs of the design matrix, s.t.

$$\mathbf{Z}_{(r)} = \mathbf{A}^*\mathbf{V}_{(r)},$$

where $\mathbf{A}^*$ is the design matrix as described in Step 3 of the ROBPCA algorithm; $\mathbf{V}_{(r)}$ is a $K_b \times k_0$ matrix whose columns are the eigenvectors associated with the eigenvalues of the robust covariance estimation based on the MCD of $\mathbf{A}^*_{n\times k_0}$, the mean-centered data matrix. The logit transformation of the reduced functional logistic model becomes,

$$\mathbf{L}^{(s)}_{(r)} = \beta^{(s)}_0\mathbf{1} + \mathbf{Z}^{(s)}_{(r)}\gamma^{(s)}, \tag{3.4}$$

where $\gamma = \mathbf{V}^{'}_{(r)}\mathbf{b}$ and s is the number of retained PCs in the model.

The design matrix of this model is now void of collinear columns, and also because of the robust approach taken to compute the PCs, the effect of outlying curves is minimized.

Therefore, the estimate of the functional parameter is given by,

$$\hat{\beta}(t) = \hat{\mathbf{b}}'\boldsymbol{\varphi},$$

where $\hat{\mathbf{b}} = \mathbf{V}_{(r)}\hat{\gamma}$.

3.3 Model Selection Issues

There are different criteria used in deciding which PCs should be included in the model. The natural order is to include PCs based on the explained variability. There are other criterion available as discussed by Hocking (19) and by Müller and StadtMüller (29) which take into consideration the predictive ability of the PCs. In the simulation study carried out in the next section, the simplistic natural order of explained variability is used to decide which PCs to include in the model.

Another issue in model selection is the decision concerning the number of PCs to include in the model. Some of the measures that can be used include the integrated mean squared error of the $\beta(t)$ parameter function (IMSEB). This is defined as,

$$IMSEB^{(s)} = \frac{1}{T}\int_T (\beta(t) - \hat{\beta}^{(s)}(t))^2 dt,$$

where $\hat{\beta}^{(s)}(t)$ is the estimated parameter function for the logistic model with s PCs.

Another available measure is the mean squared error of $\beta(t)$ parameters (MSEB), which is defined as,

$$MSEB^{(s)} = \frac{1}{K_b + 1}\left((\beta_0 - \hat{\beta}_0^{(s)})^2 + (\beta(t) - \hat{\beta}(t)^{(s)})^2\right),$$

where K_b is the number of basis functions, $\hat{\beta}_0^{(s)}$ is the estimated intercept in the model with s PCs and $\hat{\beta}(t)^{(s)}$ is the estimated parameter function in the functional logistic model with s PCs. The optimal model is selected as that model whose IMSEB or MSEB is the smallest. In the simulation study, we use the MSEB to determine s, the number of PCs to include in the model. The PCs are then added based on the explained variability of each PC, starting with the one with the largest explained variability until the optimal number of PCs is attained.

However, in the case of real data, both these methods cannot be obtained and therefore more practical approaches are required. Escabias et al. (9) suggest the use of the estimated variance of the estimated parameters which is defined as,

$$Var(\hat{\beta}(t)^{(s)}) = \mathbf{V}_{(r)}^{(s)}(\mathbf{Z}_{(r)}^{'(s)}\mathbf{W}_{(r)}^{(s)}\mathbf{Z}_{(r)}^{(s)})^{-1}\mathbf{V}_{(r)}^{'(s)},$$

where $\mathbf{W}_{(r)}^{(s)} = diag[\hat{\pi}^{(s)}(1-\hat{\pi}^{(s)})]$, $\mathbf{Z}_{(r)}^{(s)}$ is the matrix of s robust PC scores from the design matrix and $\mathbf{V}_{(r)}^{(s)}$ is the matrix whose columns are the eigenvectors associated with the eigenvalues of the robust covariance estimation based on the MCD of the design matrix. Escabias et al. (9) suggested selecting the optimum number of PCs by plotting the estimated variance against the number of PCs, s, and selecting s just before a significant increase in the estimated variance. In the case where there are potentially multiple cases like this, the smallest s is then selected. The percent of variance explained (PVE) is another alternative approach to determine the number of PCs to retain in the model. In this case, PCs with eigenvalues greater than 1.0 can be retained. In both these instances, the idea is to ensure that much of the variability in the model is retained.

Another method is the cross validation (CV) method. Cross validation involves partitioning the data into two sets; the first set known as the training set is used to determine a predictive model whilst the second set, known as the test set is used to validate the predictive model. The leave-one out cross validation method leaves out one observation and fits the model with the remaining $n-1$ observations. Prediction is then made for the left-out observation using this model, and this procedure is repeated for all the observations. For the logistic model, this is defined as,

$$CV^{(s)} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{\pi}_{i,-i}^{(s)})^2,$$

where $\hat{\pi}_{i,-i}$ indicates the predicted response with observation i missing from the predictive model. An optimal number of PCs are selected as those with a minimum CV. The Information Criterion (IC) method is another alternative. The IC can be viewed as a compromise between the goodness of fit and the complexity of the model. The Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) are some of the widely used IC in which case to select the optimal number of PCs, the IC is computed for varying s values and the optimal s would then be chosen at the minimum.

3.4 Numerical Examples

In this section, we study the performance of our proposed estimation approach by way of a simulation study, as well as applying the methodology to the Canadian Weather data. We show the improved accuracy in the parameter function estimation in the case where outliers are present.

3.4.1 Simulation Study

The following steps were carried out in this simulation study in order to investigate how the proposed estimation technique performs and compares to other existing approaches in the estimation of the parameter function of the functional logistic regression model. In the first step to generate the data, the following were done,

(a) Generate $n = 50$ sample curves for the functional logistic model.

We generate 50 sample functional observations of a known stochastic process $X(\cdot)$ considered over the interval $[0, 10]$ which has 21 equally spaced time slots. We define this

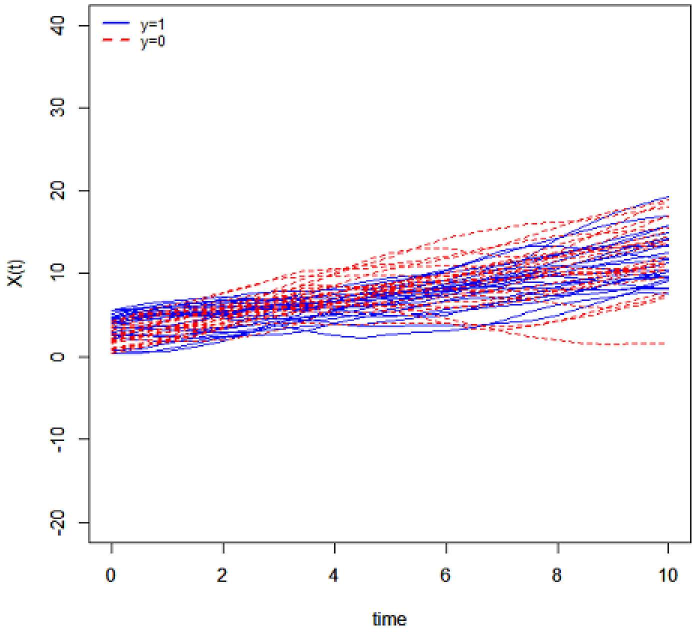


Figure 3.1: 50 sample curves generated from the overlapping stochastic process $X_i(t) = a_{i1} + a_{i2}t + W_i(t)$ and differentiated for each class of the response

process as,

$$X_i(t) = a_{i1} + a_{i2}t + W_i(t)$$

$$W_i(t) = \sum_{r=1}^{10} b_{i1} sin(\frac{2\pi}{10}rt) + b_{i2} cos(\frac{2\pi}{10}rt)$$

where $a_{i1} \sim U[1,4]$ or $a_{i1} \sim U[2,4]$, $a_{i2} \sim N[1,0.2]$ or $a_{i2} \sim N[1,0.6]$, $b_{i1}, b_{i2} \sim N[0,1/r^2]$. This model is illustrated in Figure 3.1. Since the response is binary, the sample curves are differentiated for the two classes. To obtain the functional form of these sample curves, we consider them to belong to the finite-dimensional space generated by some basis, where $\phi(t)$ is a cubic B-spline basis defined on equally spaced nodes on the $[0, 10]$ interval.

We used the generalized cross-validation (GCV) method to determine the number of basis functions to use as well as to determine the smoothing parameter in estimating the functional covariate, $X(t)$. The criterion is expressed as,

$$GCV(\lambda) = \left(\frac{n}{n - df(\lambda)}\right)\left(\frac{SSE}{n - df(\lambda)}\right)$$

where $df(\lambda) = \text{trace}\ \{\mathbf{S}_{\phi,\lambda}\}$ is the equivalent degrees of measure and $\mathbf{S}_{\phi,\lambda}$ is the projection operator such that $\hat{\mathbf{X}} = \mathbf{S}_{\phi,\lambda}\mathbf{X}$; SSE is the residual sum of squares; and λ is the smoothing parameter that measures the rate of exchange between the fit to the data. The minimization of GCV with respect to λ is achieved by doing a grid search of some values of λ and selecting that λ with the minimum GCV value.

(b) The natural cubic spline interpolation of the parameter function $sin(t + \pi/4)$ is selected as the true $\beta(t)$ function. The basis coefficients, $\mathbf{b} = (b_1, ..., b_{K_b})'$ of the parameter function are known and used to assess the estimation techniques presented in this work.

(c) The probabilities of a positive response ($Y_i = 1$) given $X_i(t)$ are,

$$\pi_i = \frac{exp\{l_i\}}{1 + exp\{l_i\}}, \quad i = 1, .., n$$

where l_i is as defined before in (3.3); and β_0 is fixed at 0. The values of the response, Y_i, are obtained by simulating observations of a Bernoulli distribution with probabilities π_i.

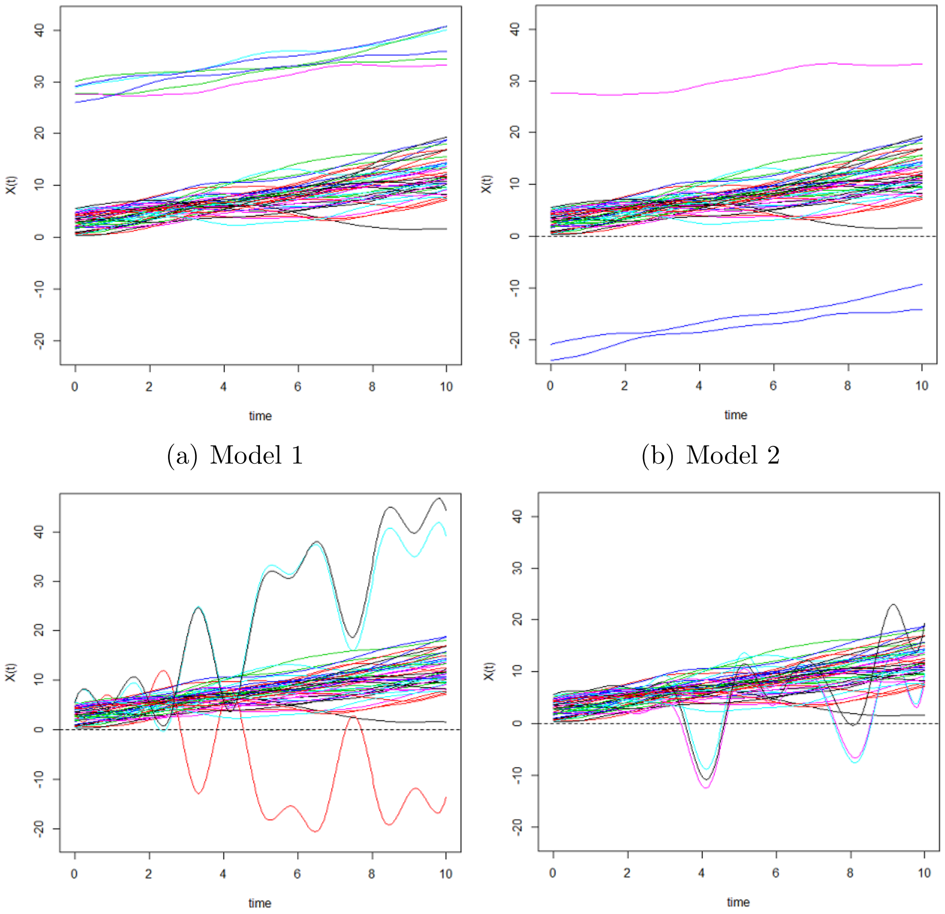
The second step involves contamination of the simulated data. We adopted the contamination process as discussed by Fraiman and Muniz (14):

(i) **Model 0: No Contamination**

$X(t) = a_1 + a_2t + W(t)$ is the generated data as discussed before.

(ii) **Model 1: Asymmetric Contamination**

$Z(t) = X(t) + cM$ where c is 1 with probability q and 0 with probability $1 - q$ and $q = \{0\%, 5\%, 10\%, 15\%, 20\%\}$ is the contamination level; M is the contamination constant size taking a value of 25 and $X(t)$ is as defined in Model 0.



(a) Model 1

(b) Model 2

(c) Model 3

(d) Model 4

Figure 3.2: Sampling curves for Model 1 - 4 at $q = 5\%$

(iii) **Model 2: Symmetric Contamination**

$Z(t) = X(t) + c\sigma M$ where $X(t)$, c and M are as defined before and σ is a sequence of random variables independent of c that takes the values 1 and -1 with probability 0.5

(iv) **Model 3: Partial Contamination**

$Z(t) = X(t) + c\sigma M$ if $t > T$ and $Z(t) = X(t)$ if $t < T$, where T is a random number generated from a uniform distribution on $[0, 10]$

(v) **Model 4: Peak Contamination**

$Z(t) = X(t) + c\sigma M$ if $T \leq t \leq T + l$ and $Z(t) = X(t)$ if $t \notin [T, T + l]$ where $l = 2$ and T is a random number from a uniform distribution on $[0, 10 - l]$.

Table 3.1: Median MSEB (standard error) for the estimation of the functional parameter based on the optimum model for Model 1 and Model 2

	Asymmetric		Symmetric	
Cont. (%)	CPCA	RPCA	CPCA	RPCA
0	0.1903 (0.0216)	0.1935 (0.0217)	0.1903 (0.0216)	0.1935 (0.0217)
5	0.1970 (0.0209)	0.1731 (0.0203)	0.1995 (0.10193)	0.1615 (0.0194)
10	0.1802 (0.0186)	0.1780 (0.0175)	0.1844 (0.0176)	0.1728 (0.0177)
15	0.1795 (0.0172)	0.1862 (0.0171)	0.1798 (0.0167)	0.1768 (0.0169)
20	0.1854 (0.0167)	0.1824 (0.0168)	0.1815 (0.0160)	0.1769 (0.0165)

Table 3.2: Median MSEB (standard error) for the estimation of the functional parameter based on the optimum model for Model 3 and Model 4

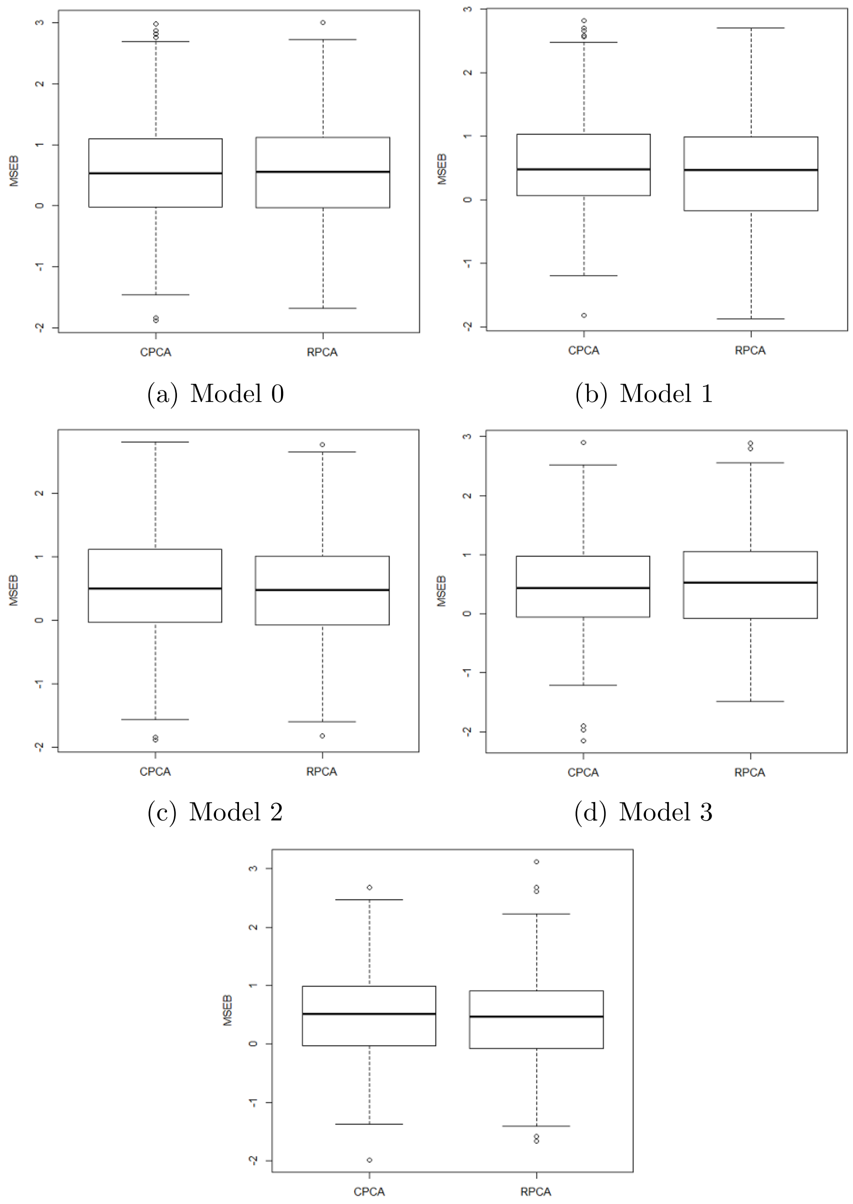
	Partial		Peak	
Cont. (%)	CPCA	RPCA	CPCA	RPCA
0	0.1790 (0.0216)	0.1839 (0.0217)	0.1790 (0.0216)	0.1839 (0.0217)
5	0.2929 (0.0204)	0.2152 (0.0197)	0.2866 (0.0209)	0.2263 (0.0196)
10	0.2833 (0.0207)	0.2416 (0.0192)	0.3059 (0.0200)	0.2554 (0.0194)
15	0.2806 (0.0206)	0.2504 (0.0199)	0.2879 (0.0203)	0.2594 (0.0202)
20	0.2779 (0.0205)	0.2787 (0.0209)	0.2878 (0.0201)	0.3070 (0.0196)

The effects of these different types of contamination are shown in Figure 3.2. The reduced logistic model in (3.3) was fit with the coefficient matrix, $\boldsymbol{C\psi}$, as the covariate. The Hosmer-Lemeshow (20) goodness of fit test was also carried out, which indicated the model to be valid for the generated data. In the final step, we obtained the approximated estimations for the parameter function using the robust approach proposed (RPCA) in Section 3.2.

The approximated estimates are compared with the maximum likelihood estimate (MLE) as well as the classical principal component estimate (CPCA) as discussed by Escabias et al. (9). The MSEB is used to determine how many PCs to retain in the model, with small MBSE values indicating better estimation when these three different estimation techniques are compared. All simulations were implemented using R (33). The packages *fda* (35) and *rrcov* (42) were particularly useful in obtaining the functional form of the data and performing principal component estimation, respectively.

The simulations were replicated 200 times and Tables 3.1 and 3.2 summarize the effect of outliers on the median MSEB of the models. The median was used to reduce the influence of the extreme observations. In most of the models, and especially so for Models 3 and 4, the proposed robust approach yields better results at the varying contamination levels. It should be noted that though higher contamination levels (i.e. $q = 15\%, 20\%$) were also attempted with similar consequences, high contamination levels for the logistic regression model make it difficult to distinguish between the contaminated and the simulated data. This would explain the breakdown of the estimations at these higher contamination levels. The estimates for the ML approach were unstable and inefficient as expected, and the median MSEB values are excluded from the tables.

Figure 3.3 shows the distribution of $\hat{\beta}_0$ for the principal component estimation methods for the 5 different models at $q = 5\%$. The true parameter value is 0 and there seems to be no major difference in the distributions of this scalar parameter. Figure 3.4 , on the other hand, shows the median estimates for the parameter function for the two PC estimation techniques, CPCA and RPCA. In this instance, 4 PCs were retained in the regression models and the median $\beta(t)$ estimates compared for the robust and classical PCA approaches. 4 PCs were retained as this was the typical (median) number of PCs retained for the optimal model for both methods and for the different contamination models. The effect of contamination is evident in that when there are no functional outliers (Model 0), the $\beta(t)$ estimates are no different as shown in Figure 3.4(a). However, the estimates deteriorate for the classical PCA estimation techniques, even at $q = 5\%$ contamination. This is especially so with the partial and peak contamination models. In this overlay comparison of the robust method (RPCA) vs. the non-robust approach (CPCA), it can be seen that the robust estimation



(a) Model 0

(b) Model 1

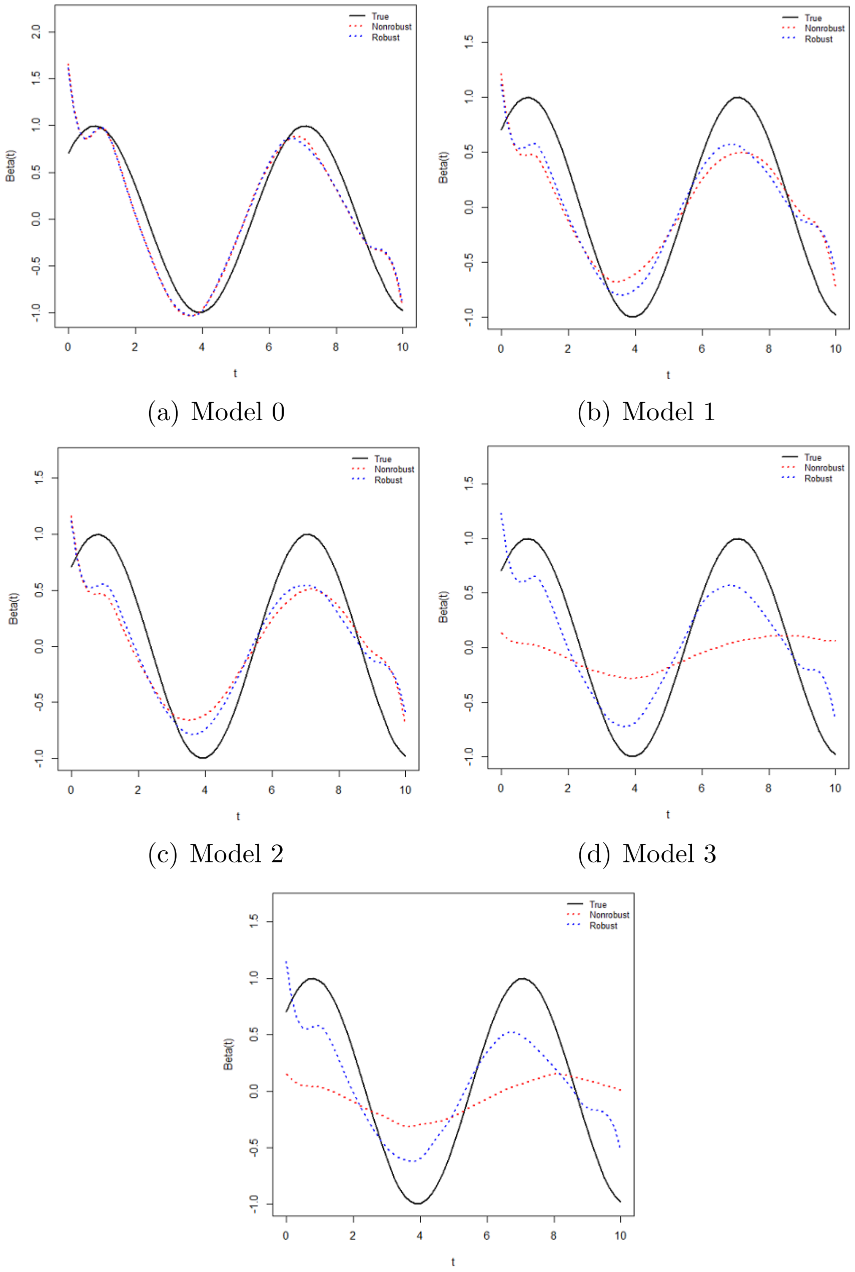
(c) Model 2

(d) Model 3

(e) Model 4

Figure 3.3: Side-by-side boxplots on the effect of outlier curves on maximum likelihood estimate (MLE),classical PCA (CPCA) and robust PCA (RPCA) estimates of β_0 at $q = 5\%$ contamination

remains closer to the true simulated curve when compared with the non-robust approach of PC estimation. Therefore, we obtain better parameter estimation results by making use of the proposed method when there are outliers present in the data.



(a) Model 0

(b) Model 1

(c) Model 2

(d) Model 3

(e) Model 4

Figure 3.4: Overlay comparison of the effect of outlier curves on classical PCR and robust PCR when the first 4 PCs are included in the model at $q = 5\%$ contamination.

3.4.2 Canadian Weather Data Set

In their paper on modeling environmental data by functional principal component estimation, Escabias et al. (10) used the Canadian weather data from Ramsay and Silverman (34) to predict the risk of drought ($Y = 1$ for a station where there is no drought risk;

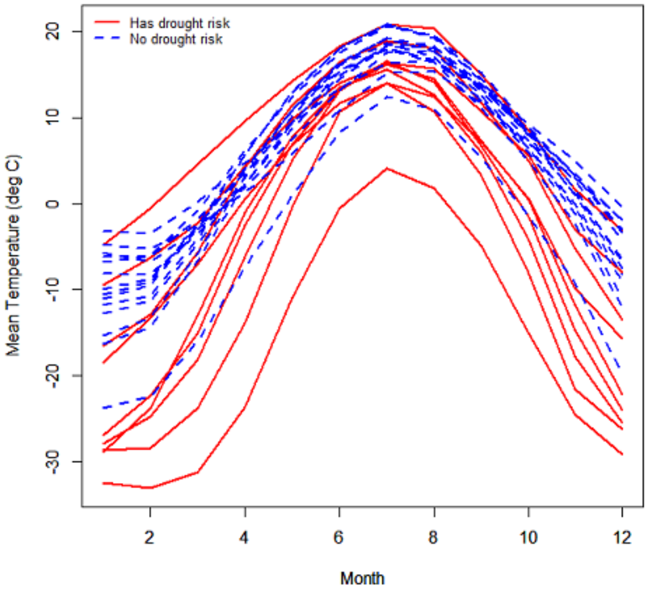
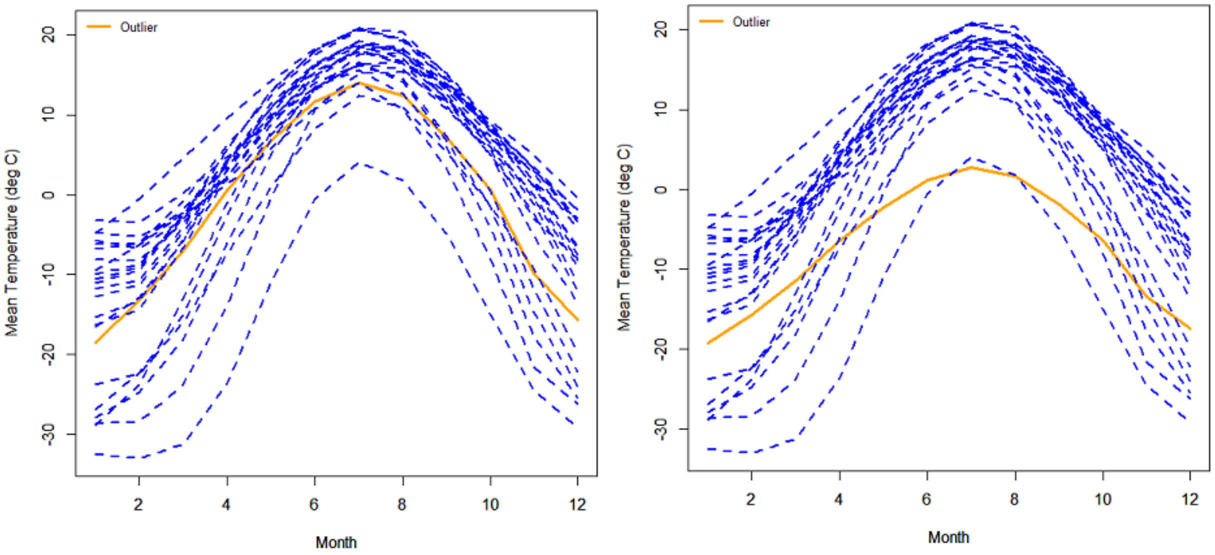


Figure 3.5: The mean monthly temperatures for 23 Canadian weather stations used to predict the risk of drought

$Y = 0$ for a station where there is drought risk) based on the monthly average temperatures recorded over 12 months. The annual precipitations for each area were used to determine whether an area had risk of drought or not. An area is said to have drought risk if the precipitation along a year in that area is lower than the 25th percentile of the total annual precipitation in the entire country.

There are 23 samples representing the weather stations, each with 12 mean monthly temperatures recorded. Figure 3.5 shows the sample curves for the 23 weather stations with an indication of the drought risk for each station. From this dataset, $n_1 = 9$ stations have drought risk and the rest, $n_2 = 14$, do not have drought risk based on the precipitation records. Due to the sinusoidal nature of the sample curves, the Fourier basis was used in the approximation of the temperature function for each of the weather stations. The generalized cross-validation method was used to determine the order of expansion for the functional covariate of which 11 basis functions are used with a smoothing parameter, $\lambda \approx 0.0009$.

We introduce an outlying sample curve by reducing all the temperatures for the Churchill station by 10 degrees Celsius and slightly changing the pattern of the temperature curve by



(a) Before

(b) After

Figure 3.6: Churchill weather station's temperatures are altered

stretching it by a factor of 0.675. The weather station was randomly selected. Figure 3.6 illustrates the effect of that shift and stretch on the original data. The AIC was used in order to determine how many PCs to retain in the principal component-based methods. Table (3.3) gives the details of the retained PCs for the classical principal component analysis (CPCA) approach as well as our proposed robust principal component analysis (RPCA). In both cases, the first three PCs were retained and both models have an AIC of 8. There is little difference in the percent of variation explained (PVE) with the inclusion of these PCs in the logistic model, all of the models having over 99% of variance explained.

Table 3.3: The model details for each estimation technique

	Original Sample			Contaminated Sample		
	Retained PCs	AIC	PVE	Retained PCs	AIC	PVE
CPCA	1,2,3	8	99.62%	1,2,3	8	99.63%
RPCA	1,2,3	8	99.70%	1,2,3	8	99.72%

Due to multicollinearity issues, the maximum likelihood approach is not ideal in estimating the parameter function, $\beta(t)$, of the functional logistic regression model for predicting the risk of drought. Figure 3.7 shows the parameter estimate using maximum likelihood

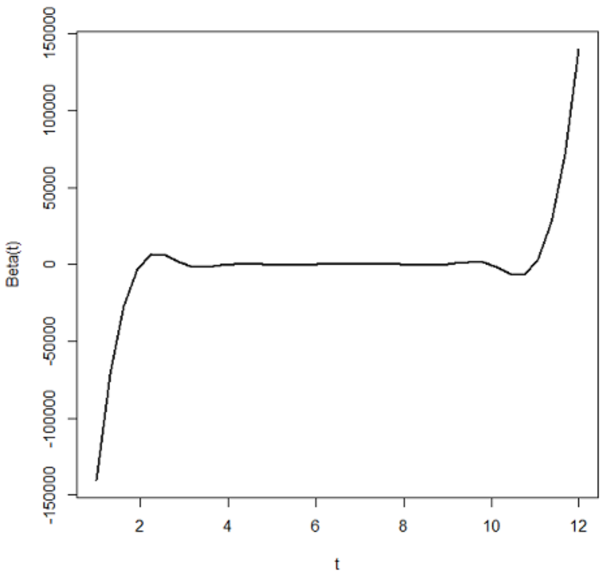


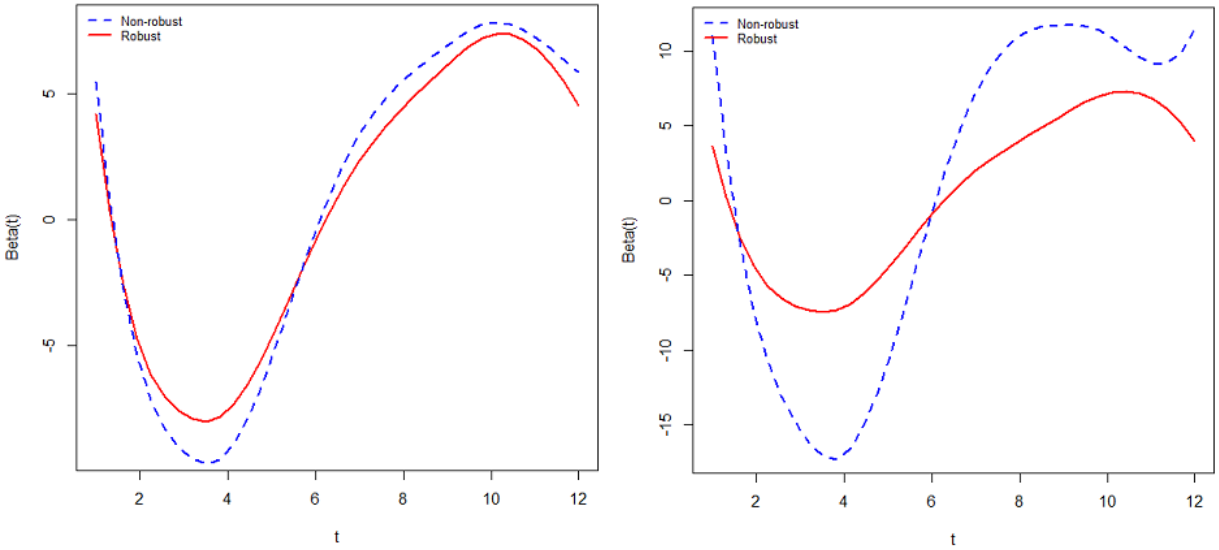
Figure 3.7: Parameter estimation without using principal component estimation

estimation. Figure 3.8 shows the function parameter estimates of $\beta(t)$ using the two different approaches of principal component estimation. In the absence of any outlying curves, the non-robust and robust principal component estimations for the parameter function are almost similar. The effect of contamination on the estimation of $\beta(t)$ is quite noticeable for the non-robust approach. The parameter estimate is distinctly different, and therefore one would have a different interpretation when calculating the odds of drought for certain seasons or time intervals. On the other hand, the presence of this outlying sample curve has minimal effect on the proposed robust approach. The parameter function estimate remains quite unchanged, as can be seen in Figure (3.9).

Table 3.4: Goodness of fit measures

	Original Sample			Contaminated Sample		
Method	AUC	Z	p-value	AUC	Z	p-value
MLE	0.6476	-0.8869	0.3751	0.6672	-0.8152	0.4149
Classical PCA	0.6508	-0.6428	0.5204	0.6481	-1.0026	0.3161
Robust PCA	0.6587	-0.5862	0.5773	**0.7125**	-0.5259	0.5989

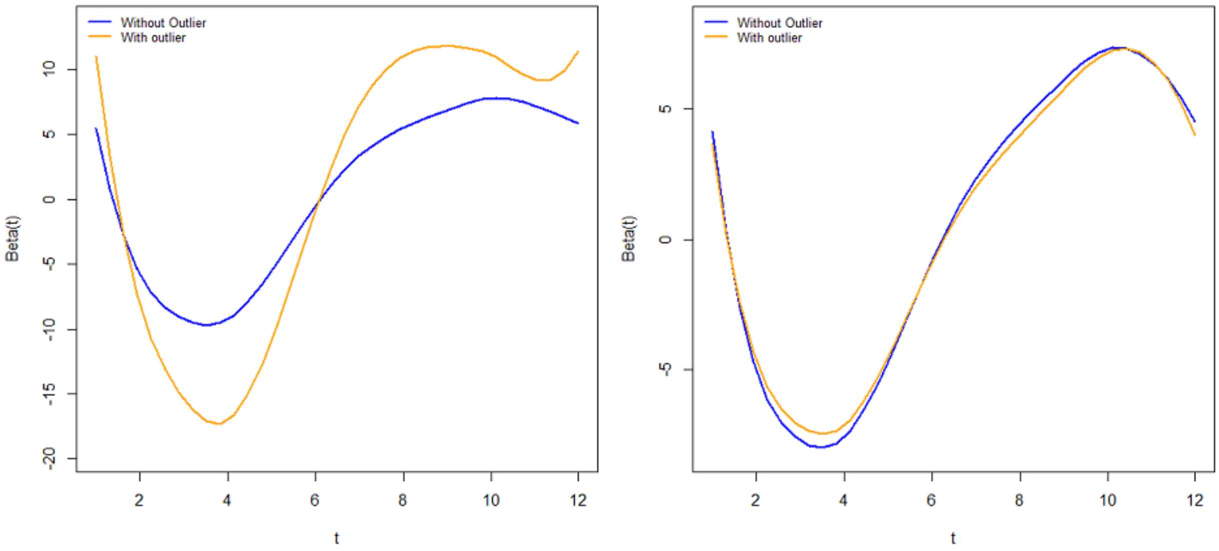
Goodness of fit measures were also conducted for the three different approaches as summarized in Table (3.4). The measures provided are the area under the ROC curve



(a) Before

(b) After

Figure 3.8: Parameter estimation when Churchill weather station's temperatures are shifted and its effect on the principal component estimation methods



(a) Classical PCA

(b) Robust PCA

Figure 3.9: CPCA vs RPCA when Churchill weather station's temperatures are shifted

(AUC) as well as the goodness of fit statistic (Z) and its p-value. For the goodness-of-fit test, the Le Cessie-van Houwelingen normal test statistic (Le Cessie and van Houwelingen

(23)) for the unweighted sum of squared errors is used. This is defined as,

$$\hat{T}_r = \sum_{i=1}^{n} \frac{\hat{r}_{si}^2}{var(\hat{r}_{si}^2)},$$

where $\hat{r}_{si} = \sum_{j=1}^{n} w_{ij} \left\{ \frac{y_j - \hat{\pi}_j}{\sqrt{\hat{\pi}_j(1-\hat{\pi}_j)}} \right\}$ and the w_{ij}'s are weights. All three approaches provide good fits (p-value > 0.05). The robust PCA approach has the highest area under the ROC curves, especially in the presence of the outlying sample curve. These goodness-of-fit measures indicate that the model that uses the robust approach has AUC of 0.7125 whilst that which uses the non-robust approach has one of 0.6755. This is considered as excellent discrimination by Hosmer and Lemeshow (20), an indication that the model based on the robust PCs better predicts the risk of drought.

3.5 Conclusion

The objective in this chapter was to suggest a robust estimation technique for the functional logistic regression model. The estimation of the functional parameter in this model cannot be achieved by the regular method of maximum likelihood. Therefore, we approximate the functional observations and define the parameter function in a finite-dimensional space generated by a known basis. This reduces the functional model to a multiple model, albeit with highly collinear covariates. The presence of multicollinearity and outliers causes the maximum likelihood estimate from this multiple logistic model to be unstable and therefore, unreliable.

Robust estimation methods for the functional logistic model are therefore an important tool in estimation of the parameter function derived in this manner. In this work, we have proposed an approach that makes use of robust principal component estimation, and essentially reduces dimensionality and improves the estimation of the parameter function in the presence of multicollinearity and outliers. From the simulation study, we have shown that in the presence of outliers, this approach results in better estimations for the parameter function, and subsequently better interpretation of the model. We also illustrated the improved performance of the proposed method, by analyzing a real data set.

Chapter 4

Robust Penalized Functional Logistic Regression

4.1 Introduction

In Chapter 3, we proposed an estimation technique that makes use of robust principal component estimation. This method is shown to be a more reliable approach in estimating the functional parameter in the functional logistic regression model, especially since it became necessary to look at an approach that eliminated multicollinearity as the resulting design matrix in the reduced standard logistic model was ridden with collinear columns. Even though the use of a robust principal component estimation technique resulted in better estimations for the model in the presence of outliers, the influence of outliers, however, is still noticeable. This can be attributed to the fact that this approach still makes use of the maximum likelihood estimation approach.

It has been shown that for the standard logistic model, the MLE is sensitive to outliers [Hosmer and Lemeshow (20), Croux et al. (8), Carroll and Pederson (5), Pregibon (32)]. The robustness properties for the MLE were analyzed in terms of the influence function. The influence function (IF) of an estimator is an asymptotic approximation to the behavior of that estimator when the sample contains a small fraction of identical outliers. An estimator with an unbounded IF is non-robust since an infinitesimal model deviation causes the bias to be arbitrarily large. The IF of the MLE for the standard logistic model is,

$$IF(y, \mathbf{x}, \boldsymbol{\beta}) = \mathbf{M}^{-1}(y - \pi(\boldsymbol{\beta}^{'}\mathbf{x}))\mathbf{x},$$

where $\mathbf{M}$ is the Fisher information matrix, $\mathbf{M} = E[\pi^{'}(\boldsymbol{\beta}\mathbf{x})\mathbf{x}\mathbf{x}^{'}]$. Therefore, the MLE is unbounded in x and bounded in y. In fact, the kind of outliers whose influence is large for the logistic model are such that either $\| \mathbf{x}_i \| \rightarrow \infty$, $y_i = 1$ and $\boldsymbol{\beta}'\boldsymbol{x}_i$ is bounded away from ∞ or $\| \mathbf{x}_i \| \rightarrow \infty$, $y_i = 0$ and $\boldsymbol{\beta}'\boldsymbol{x}_i$ is bounded away from $-\infty$. In other words, extreme values in the design space lead to a biased MLE. Moreover, it has also been shown that mis-classification errors, i.e. errors in the response, lead to a biased MLE [Pregibon (32), Copas (6)]. Therefore, there is a need to develop robust estimation methods that deal with these issues.

There have been many approaches discussed in literature that look at ways to turn the MLE into an estimate with bounded influence. Such methods involve techniques of down-weighting high-leverage observations. In this chapter, we look at one such alternative approach that is a Mallows-type estimate. We adopt the same estimation procedures as before that reduce the functional logistic model to a standard logistic model by approximating the functional covariate and the functional parameter as a linear combination of some appropriate basis. A Huber-type loss function is then used to down-weight the high-leverage observations and there is also a roughness penalty term to ensure that the estimated parameter function, $\hat{\beta}(t)$, is smooth.

4.2 Estimation of $X(t)$ and $\beta(t)$

We consider the functional covariates, $X_i(t)$, and functional parameter, $\beta(t)$, as belonging to a finite-dimensional space generated by (not necessarily) the same basis. We consider $X_i(t) \in \mathbb{H}$ (i.e. Hilbert space), such that,

$$X_i(t) \approx \sum_{j=1}^{K_X} c_{ij}\psi_j(t), \tag{4.1}$$

where $\{\psi_j(t)\}_{j=1}^{K_X}$, is an appropriate basis that is selected to reflect the characteristics of the data.

We assume $\beta(t) \in \mathbb{H}$ and that the parameter is defined as a linear combination of a cubic B-spline basis $\{\phi_k(t)\}_{k=1}^{K_b}$, such that,

$$\beta(t) = \sum_{k=1}^{K_b} b_k \phi_k(t),$$

where the order of expansion, K_b, is estimated either by cross-validation or generalized cross-validation; and so that $K_b \leq K_X$. This results in the functional logistic model being reduced to a standard logistic model,

$$\begin{aligned} \pi_i &= P(Y_i = 1 \mid X_i(t) \in L^2(T)) \\ &= \frac{exp\left(\beta_0 + \sum_{j=1}^{K_X}\sum_{k=1}^{K_b} c_{ij} J_{jk} b_k\right)}{1 + exp\left(\beta_0 + \sum_{j=1}^{K_X}\sum_{k=1}^{K_b} c_{ij} J_{jk} b_k\right)}, \quad i = 1, ..., n, \end{aligned} \tag{4.2}$$

whose logit transformation is given by,

$$\begin{aligned} l_i &= \beta_0 + \sum_{j=1}^{K_X}\sum_{k=1}^{K_b} c_{ij} J_{jk} b_k, \quad i = 1, ..., n, \\ \mathbf{L} &= \beta_0 \mathbf{1} + \boldsymbol{CJb}, \end{aligned} \tag{4.3}$$

where $\mathbf{L}= (l_1, ..., l_n)'$; $\mathbf{C}= \{c_{ij}\}_{n \times K_X}$; $\mathbf{J} = \{J_{jk}\}_{K_X \times K_b}$ with $J_{jk} = \int_T \psi_j(t)\phi_k(t)dt$; $\mathbf{1}= (1, ..., 1)'$; and $\boldsymbol{b} = (b_1, ..., b_{K_b})'$. In this reduced form, the maximum likelihood estimate can be derived and the unknown parameters established using an iterative method such as Newton-Raphson. The next section explains a robust penalization method that minimizes the effect of outliers on the estimate, resulting in a smooth robust estimate for the functional logistic model.

4.3 Robust Penalized Maximum Likelihood Estimation

Cardot and Sarda (4), Goldsmith et al. (18) and Ogden and Reiss (30) have considered penalization in the estimation of the functional logistic model or generalized functional linear model. The introduction of the penalty term in the maximization of the log-likelihood function is to ensure that the smoothness property of the parameter function is met. This regularization approach in the estimation of smooth functional estimates is common, especially with estimating the functional form of the predictor, X(t), as discussed in Chapter 1 and also in functional principal component analysis [Ramsay and Silverman (34)].

The general form of the penalized likelihood is,

$$L_\lambda(\pi; Y) = L(\pi; Y) - \lambda \times Pen(\beta(t)),$$

where $L(\pi; Y)$ is the log-likelihood function; λ is the smoothing parameter and Pen$(\beta(t))$ is the penalty term that ensures that the parameter function is smooth. Cardot and Sarda (4) considered a roughness penalty of the form $\| \boldsymbol{\varphi}_k^{(m)'} \mathbf{b} \|^2$ where $\boldsymbol{\varphi}_k$ denotes the vector of all the B-splines and $\boldsymbol{\varphi}_k^{(m)}$ the vector of derivatives of order m of all the B-splines for some integer m. The GCV criterion is then used to determine the value of the smoothing parameter, λ.

Goldsmith et al. (18), on the other hand, proposed an automated way of deducing the smoothing parameter in the penalty term. They considered the parameter function, $\beta(t)$, as the truncated power series spline basis expansion,

$$\beta(t) = b_0 + b_1 t + \sum_{k=3}^{K_b} b_k (t - K_k)_+$$

where $\{K_k\}_{k=3}^{K_b}$ are knots. They also considered the reduced functional model as a mixed effects model with $K_b - 2$ random effects, $\{b_k\}_{k=3}^{K_b} \sim N(0, \sigma_b^2 \mathbf{I})$. Therefore, the penalty term is given as, $\frac{1}{\sigma_b^2}\mathbf{b}^{'}\mathbf{D}\mathbf{b}$, where $\mathbf{D}$ is the penalty matrix and $\frac{1}{\sigma_b^2}$ is the smoothing parameter.

Thus, the smoothing parameter is estimated as a variance component in the mixed effects model.

We build on these ideas and propose a robust penalization method. It is important to note that the penalized likelihood methods discussed above and such similar methods in literature are sensitive to outlying observations. From (4.2), we denote the parameter vector as $\boldsymbol{\gamma} = (\beta_0, \mathbf{b}^{'})^{'}$, and the covariate vector as $\mathbf{a}_i = (1, \{\mathbf{CJ}\}_i^{'})^{'}$ to simplify the notation. The proposed estimator [RPMLE] is defined as,

$$\hat{\boldsymbol{\gamma}} = \arg \max_{\gamma} \sum_{i=1}^{n} w_i \{y_i log(\pi_i) + (1 - y_i) log(1 - \pi_i)\} - \frac{\lambda}{2} \boldsymbol{\gamma}^{'} \mathbf{D}_0 \boldsymbol{\gamma}, \tag{4.4}$$

where π_i is short for $\pi(\mathbf{a}_i^{'}\boldsymbol{\gamma}; y_i)$; λ is a smoothing parameter; w_i are weights that might depend on $\mathbf{a}_i$, y_i or both; and $\mathbf{D}_0$ is the penalty matrix augmented by a leading column and row of $K_b + 1$ zeros. When $w_i = 1, \forall i$ and $\lambda = 0$, then (4.4) yields the maximum likelihood estimator. We consider a Mallows class estimator (Mallows (26)), that addresses the problem of outliers in the design space for the reduced functional logistic model. The main idea with a Mallows-type estimator is to down-weight observations which have high leverage. In our proposed robust estimator in (4.4), this is achieved by the following weighting scheme as proposed by Stefanski (40),

$$\begin{aligned} w_i &= W(h_n(\mathbf{x}_i)), \\ h_n(\mathbf{x}) &= \Big[(\mathbf{x} - \hat{\boldsymbol{\mu}}_n)^{'} \hat{\boldsymbol{\Sigma}}_n^{-1} (\mathbf{x} - \hat{\boldsymbol{\mu}}_n), \Big]^{\frac{1}{2}} \end{aligned}$$

with $\hat{\boldsymbol{\mu}}_n$ and $\hat{\boldsymbol{\Sigma}}_n$ being a robust location vector and a robust dispersion matrix estimate of the design matrix, respectively. W is a non-increasing function such that $W(u)u$ is bounded. We use the weight function corresponding to Huber's Ψ's, which is defined as,

$$W_k(u) = min\Big\{1, \frac{k}{\mid u \mid}\Big\},$$

where $k = 1.345$ is a tuning constant. It is noted that while the Mallows-class estimator can be less efficient than the usual logistic MLE, its bias can be smaller than that of the MLE. Carroll and Pederson (5) were able to illustrate that in the case where the design has extreme design points, selective down-weighting can lead to less biased estimates and this decrease in bias together with the robustness property, may be worth the lower efficiency.

The estimation of the smoothing parameter, λ, can be achieved by cross-validation or by the GCV method. The smoothness of the parameter function, $\beta(t)$, is based on the idea of the curvature of a function. This curvature is quantified as,

$$\begin{aligned}\int_T \{\beta^{''}(t)\}^2 dt &= \int_T \sum_{k=1}^{K_b} b_k \phi_k^{''}(t)\phi_k^{''}(t) b_k dt \\ &= \sum_{k=1}^{K_b} b_k D(t) b_k \qquad (4.5) \\ &= \mathbf{b}^{'}\mathbf{D}\mathbf{b} \qquad (4.6)\end{aligned}$$

where $D(t) = \int_T \phi_k^{''}(t)\phi_k^{''}(t)dt$ is the smoother. The approximation of the smoother is possible by means of a numerical quadrature scheme.

4.3.1 Details of the Estimation Technique

Iterative algorithms are usually used for the computation of M-estimators. Since the likelihood equations for (4.4) are non-linear in the parameter of interest, $\boldsymbol{\gamma}$, an iterative method has to be used to solve these equations and get estimates of the parameters. A modified scoring algorithm or Newton-Raphson method may be used to solve the non-linear likelihood equations.

To make use of the Newton-Raphson method, the gradient and hessian matrices are required. We consider the likelihood function in (4.4) and consider that,

$$\begin{aligned} log(\pi_i) &= log\left\{\frac{exp(\mathbf{a}_i^{'}\boldsymbol{\gamma})}{1+exp(\mathbf{a}_i^{'}\boldsymbol{\gamma})}\right\}, \\ &= \boldsymbol{a_i^{'}}\boldsymbol{\gamma} - log(1+exp(\boldsymbol{a_i^{'}}\boldsymbol{\gamma})), \end{aligned}$$

and similarly,

$$\begin{aligned} log(1-\pi_i) &= log\left\{\frac{1}{1+exp(\boldsymbol{a_i^{'}}\boldsymbol{\gamma})}\right\} \\ &= -log(1+exp(\boldsymbol{a_i^{'}}\boldsymbol{\gamma})) \end{aligned}$$

The likelihood function in (4.4) can then be re-written as,

$$\begin{aligned} L(\boldsymbol{\gamma}; \boldsymbol{y}) &= \sum_{i=1}^{n} w_i\{y_i log(\pi_i) + (1-y_i)log(1-\pi_i)\} - \frac{\lambda}{2}\boldsymbol{\gamma}^{'}\mathbf{D}_0\boldsymbol{\gamma} \\ &= \sum_{i=1}^{n} w_i\{y_i \boldsymbol{a_i^{'}}\boldsymbol{\gamma} - log(1+exp(\boldsymbol{a_i^{'}}\boldsymbol{\gamma})\} - \frac{\lambda}{2}\boldsymbol{\gamma}^{'}\mathbf{D}_0\boldsymbol{\gamma}. \end{aligned}$$

The first derivative (the gradient matrix) then becomes,

$$\begin{aligned} L^{'}(\boldsymbol{\gamma}; \boldsymbol{y}) &= \sum_{i=1}^{n} w_i\left\{y_i\mathbf{a}_i - \frac{\mathbf{a}_i exp(\boldsymbol{a_i^{'}}\boldsymbol{\gamma})}{1+exp(\boldsymbol{a_i^{'}}\boldsymbol{\gamma})}\right\} - \lambda\boldsymbol{\gamma}^{'}\mathbf{D}_0 \\ &= \sum_{i=1}^{n} w_i\mathbf{a}_i\{y_i - \pi_i\} - \lambda\boldsymbol{\gamma}^{'}\mathbf{D}_0 \\ &= \mathbf{A}^{'}\mathbf{W}\mathbf{r} - \lambda\boldsymbol{\gamma}^{'}\mathbf{D}_0, \end{aligned}$$

where $\mathbf{r} = \mathbf{y} - n\boldsymbol{\pi}$ and $\mathbf{W} = \text{diag}(w_i)$. The second derivative (the hessian matrix) becomes,

$$\begin{aligned}
L^{''}(\boldsymbol{\gamma};\boldsymbol{y}) &= \sum_{i=1}^{n} -w_i \left\{ \frac{(\mathbf{a}_i^{'})^2 exp(\boldsymbol{a_i^{'}}\boldsymbol{\gamma})(1 + exp(\boldsymbol{a_i^{'}}\boldsymbol{\gamma}) - (\mathbf{a}_i^{'})^2(exp(\boldsymbol{a_i^{'}}\boldsymbol{\gamma}))^2}{[1 + exp(\boldsymbol{a_i^{'}}\boldsymbol{\gamma})]^2} \right\} - \lambda\mathbf{D}_0 \\
&= \sum_{i=1}^{n} -w_i\{(\mathbf{a}_i^{'})^2\pi_i - (\mathbf{a}_i^{'})^2(\pi_i)^2\} - \lambda\mathbf{D}_0 \\
&= \sum_{i=1}^{n} -w_i\{\mathbf{a}_i\pi_i(1-\pi_i)\mathbf{a}_i^{'}\} - \lambda\mathbf{D}_0 \\
&= -\mathbf{A}^{'}\mathbf{V}^{\frac{1}{2}}\mathbf{W}\mathbf{V}^{\frac{1}{2}}\mathbf{A} - \lambda\mathbf{D}_0,
\end{aligned}$$

where $\mathbf{V}$ =diag$(n\pi_i(1-\pi_i))$. The derivation of the hessian matrix can be computationally expensive and an approximation based on the gradient may be used instead. The Newton-Raphson method leads to the iterative scheme,

$$\boldsymbol{\gamma}^{k+1} = \boldsymbol{\gamma}^{k} + (\mathbf{A}^{'}\mathbf{V}^{\frac{1}{2}}\mathbf{W}\mathbf{V}^{\frac{1}{2}}\mathbf{A} + \lambda\mathbf{D}_0)^{-1}(\mathbf{A}^{'}\mathbf{W}\mathbf{r} - \lambda\boldsymbol{\gamma}^{'}\mathbf{D}_0), \tag{4.7}$$

in which case $\mathbf{V}$ and $\mathbf{r}$ are evaluated at $\boldsymbol{\gamma}^k$. When all the weights are unity, i.e. $w_i = 1 \;\; \forall i$ and $\lambda = 0$, then the above reiterative scheme simplifies to that of of the usual maximum likelihood iterative procedure for $\hat{\boldsymbol{\gamma}}$. A grid search is done to obtain the optimal smoothing parameter, λ, by way of the cross-validation method. This selection criteria is discussed in the following section.

4.3.2 Tuning Parameter Selection

The influence of the penalty is based on the magnitude of the smoothing parameter, which is non-negative. When $\lambda = 0$ then we have un-penalized estimates and thus, have unbiased estimates where smoothness is not emphasized; whereas large values of λ indicate that more weight is given on the smoothness of the estimate. Increasing the value of λ makes the $\beta(t)$ smoother and an optimum is reached by doing a grid search of λ. The smoothing parameter, λ, is searched by varying it systematically and then monitoring the prediction error using techniques such as cross-validation. The factor $\frac{1}{2}$ is used so as to get rid of a factor 2 that occurs when the penalty is differentiated.

The leave-one out cross validation method leaves out one observation and fits the models with the remaining $n-1$ observations. Prediction is then made for the left-out observation using this predictive model, and this procedure is repeated for all the observations. For the logistic model, this is defined as,

$$CV = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{\pi}_{i,-i})^2,$$

where $\hat{\pi}_{i,-i}$ corresponds to the predicted probability of a positive response given the predictor with observation i missing from the predictive model. The CV is computed for a variety of λ and an optimal smoothing parameter with a minimum CV is then selected. Another option for exponential family models is minimization of the information criterion (IC). The IC can be viewed as a compromise between the goodness of fit and the complexity of the model. The Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) are some of the widely used ICs. To select the optimal smoothing parameter, a logarithmic grid search on the non-negative penalty parameter λ is performed and IC computed. The optimal λ is chosen at the minimum.

4.4 Numerical Examples

To display the robustness properties of the proposed estimator, we conduct a simulation study and compare the robust approach discussed in this chapter with the existing maximum likelihood estimate (MLE),

$$\hat{\boldsymbol{\gamma}} = \arg\max_{\boldsymbol{\gamma}} \sum_{i=1}^{n}\{y_i log(\pi_i) + (1-y_i)log(1-\pi_i)\} \tag{4.8}$$

as well as the penalized maximum likelihood estimate (PMLE),

$$\hat{\boldsymbol{\gamma}} = \arg\max_{\boldsymbol{\gamma}} \sum_{i=1}^{n} \{y_i log(\pi_i) + (1 - y_i) log(1 - \pi_i)\} - \frac{\lambda}{2} \boldsymbol{\gamma}' \boldsymbol{D_0} \boldsymbol{\gamma}. \tag{4.9}$$

An application of the estimation method to a real data set is also explored and the results discussed in this section.

4.4.1 Simulation Study

We use the same data generated in Chapter 3 for this simulation study. That is, we generate 50 sample functional observations of a known stochastic process $X(\cdot)$ considered over the interval $[0, 10]$ which has 21 equally spaced time slots where the process is defined as,

$$\begin{aligned} X_i(t) &= a_{i1} + a_{i2} t + W_i(t) \\ W_i(t) &= \sum_{r=1}^{10} b_{i1} sin(\frac{2\pi}{10} rt) + b_{i2} cos(\frac{2\pi}{10} rt) \end{aligned}$$

where $a_{i1} \sim U[1, 4]$ or $a_{i1} \sim U[2, 4]$, $a_{i2} \sim N[1, 0.2]$ or $a_{i2} \sim N[1, 0.6]$, $b_{i1}, b_{i2} \sim N[0, 1/r^2]$.

The functional predictor is estimated using a cubic B-spline basis expansion where the order of expansion, K_X, is determined by way of the GCV method. The true $\beta(t)$ function is taken to be $sin(t + \pi/4)$ and the natural cubic spline of this used to get an approximation of the function. Outlying sample curves are introduced for the stochastic process $X(\cdot)$ using the model of Fraiman and Muniz (14) as discussed in Chapter 3. These are asymmetric contamination (Model 1), symmetric contamination (Model 2), partial contamination (Model 3) and peak contamination (Model 4).

The smoothing parameter is determined by the cross-validation method. The Huber-type weights makes use of the Mahalanobis distance, where the robust location and dispersion

Table 4.1: Mean MSEB (standard error) for the estimation of the functional parameter, $\beta(t)$, for Models 2, 3 and 4

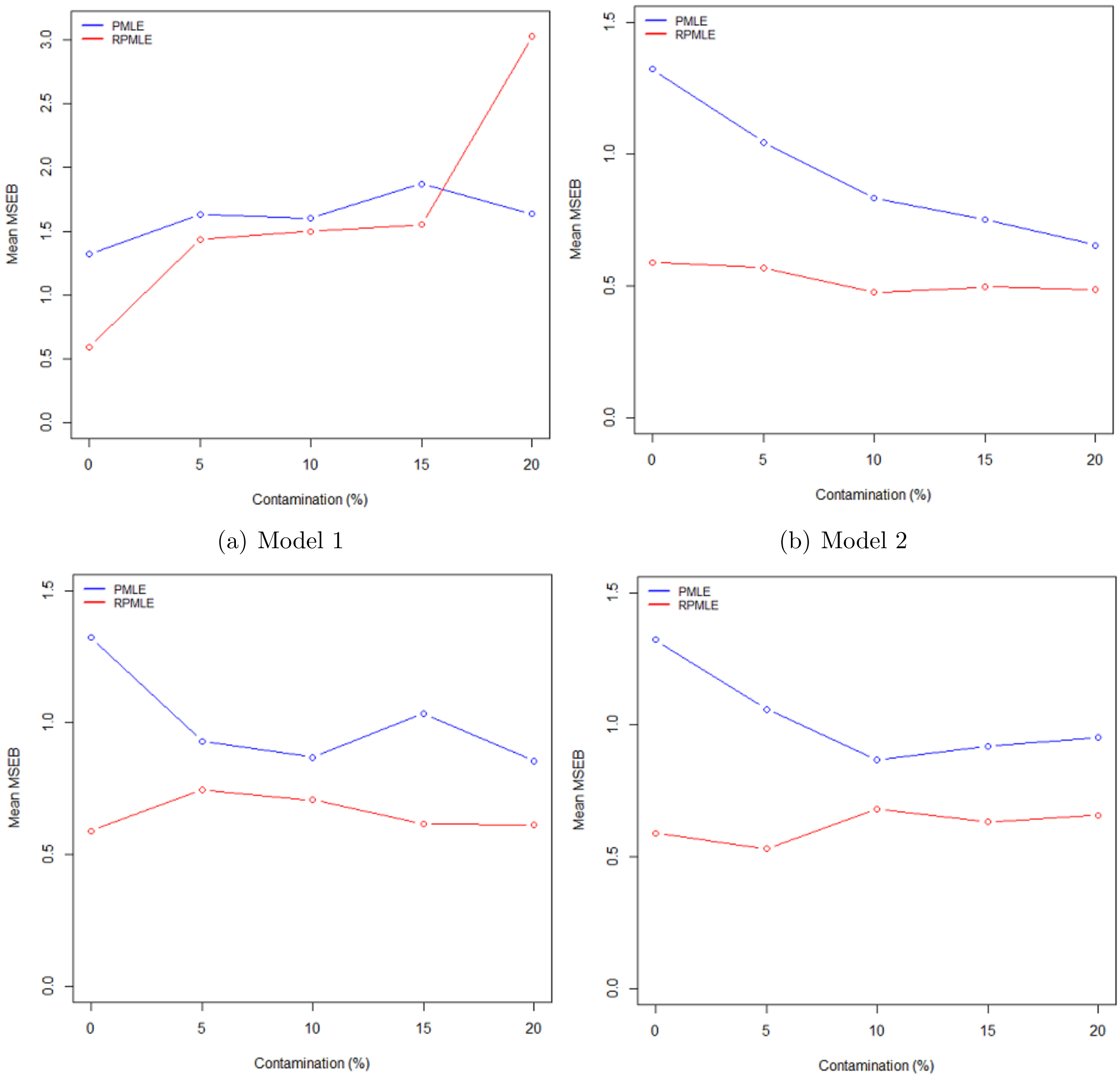
	Model 2			Model 3			Model 4		
Cont. (%)	MLE	PMLE	RPMLE	MLE	PMLE	RPMLE	MLE	PMLE	RPMLE
0	26.3043	1.3211	0.5894	26.3043	1.3211	0.5894	26.3043	1.3211	0.5894
	(9.5266)	(0.5115)	(0.1081)	(9.5266)	(0.5115)	(0.1081)	(9.5266)	(0.5115)	(0.1081)
5	20.5941	1.0453	0.5681	16.3381	0.9292	0.7444	20.5692	1.0591	0.5308
	(12.7574)	(0.3335)	(0.1072)	(6.2826)	(0.2919)	(0.3015)	(6.9599)	(0.3239)	(0.0883)
10	8.1512	0.8327	0.4756	12.4196	0.8676	0.7055	18.2714	0.8676	0.6798
	(3.6822)	(0.2573)	(0.0386)	(4.9123)	(0.2793)	(0.2218)	(4.7792)	(0.2182)	(0.2067)
15	4.3307	0.7522	0.4958	11.5515	1.0341	0.6168	17.0797	0.9183	0.6318
	(2.9919)	(0.2844)	(0.0469)	(4.7697)	(0.3466)	(0.1429)	(4.4364)	(0.2361)	(0.1510)
20	3.7119	0.6544	0.4858	7.8925	0.8528	0.6115	17.5679	0.9504	0.6578
	(4.4867)	(0.2118)	(0.0214)	(3.3326)	(0.2776)	(0.1223)	(5.1051)	(0.2697)	(0.1408)

estimates are both MCD-based. Table 4.1 summarizes the average of the MSEB measures for the three estimation methods discussed in this chapter, i.e. the maximum likelihood estimate (MLE), the penalized maximum likelihood estimate (PMLE) and our proposed robust penalized approach (RPMLE). There is improved estimation with the inclusion of the smoothing term, as expected and discussed by Cardot and Sarda (4) and Goldsmith et al. (18). In the presence of outliers, however, our robust approach is a markedly better estimator.

A comparison of our robust penalized estimation approach with the non-robust penalized approach is shown in Figure (4.1). It is evident from this that, in the presence of outliers, down-weighting these leverage points results in a more reliable estimate of the parameters (with notable exception of high contamination levels for Model 1). Essentially, the effect of the outlying curves is minimized in the estimation of the functional parameter, $\beta(t)$, resulting in a more reliable functional logistic model for the observed data.

4.4.2 Poblenou NO_X Levels Data Set

This dataset was used by Febrero et al. (11) and is a collection of NO_X emissions in Poblenou, Barcelona (Spain) over a period of 115 days with recordings starting on 23 February and ending on 26 June, in 2005. Over that period of time, hourly measurements of the NO_X are recorded and therefore we split the whole sample of hourly measures in a



(a) Model 1

(b) Model 2

(c) Model 3

(d) Model 4

Figure 4.1: Comparison of the penalized method and robust penalized method at differing contamination levels

dataset of functional trajectories of 24 h observations (each curve represents the evolution of the levels in 1 day). There were twelve curves that contained missing data that were excluded in the study.

We consider the functional logistic regression model where the functional predictor is the functional trajectories of NO_X levels over a 24 hour period with a binary response indicating whether the day is a working day ($Y = 1$) or a non-working day ($Y = 0$). This data set is

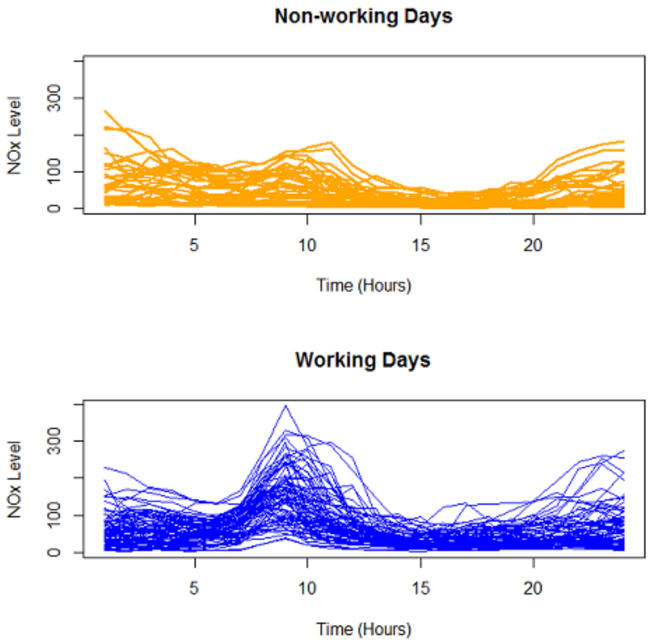


Figure 4.2: NO_X emmission levels for non-working (top) and working (bottom) days

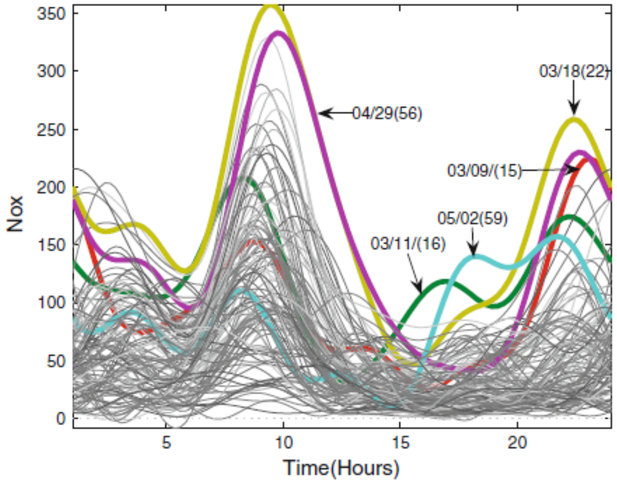


Figure 4.3: Functional form of the hourly trajectories with the outliers detected in the NO_X dataset

known to have outliers and a robust functional principal component method by Sawant et al. (38) identified the outlying curves as shown in Figure 4.3. These outliers were identified for the following five days 03/09, 03/11, 03/18, 04/29 and 05/02, all of which were working days.

The functional trajectories are estimated using a truncated Karhunen-Loeve decomposition. We let $\sum_{k=1}^{\infty} \lambda_k \psi_k(s) \psi_k(t)$ be the spectral decomposition of the covariance function, C(s,t), where $\lambda_1 \geq \lambda_2 \geq \cdots$ are the non-increasing eigenvalues associated with the orthonormal eigen-functions $\psi_1(t), \psi_2(t), \cdots$. We suppose $X_i(t) \in L^2(T)$ and centered, then

$$X_i(t) = \sum_{j=1}^{K_X} c_{ij}\psi_j(t),$$

where $\psi_j(t)$ are the first K_X eigenfunctions of the smooth covariance function $C(s,t) = cov[X_i(s), X_i(t)]$ and $c_{ij} = \int_T X_i(t)\psi_j(t)dt$. The truncation lag, K_X, is determined using a 99% cut-off of the percent variance explained (PVE). This is defined as,

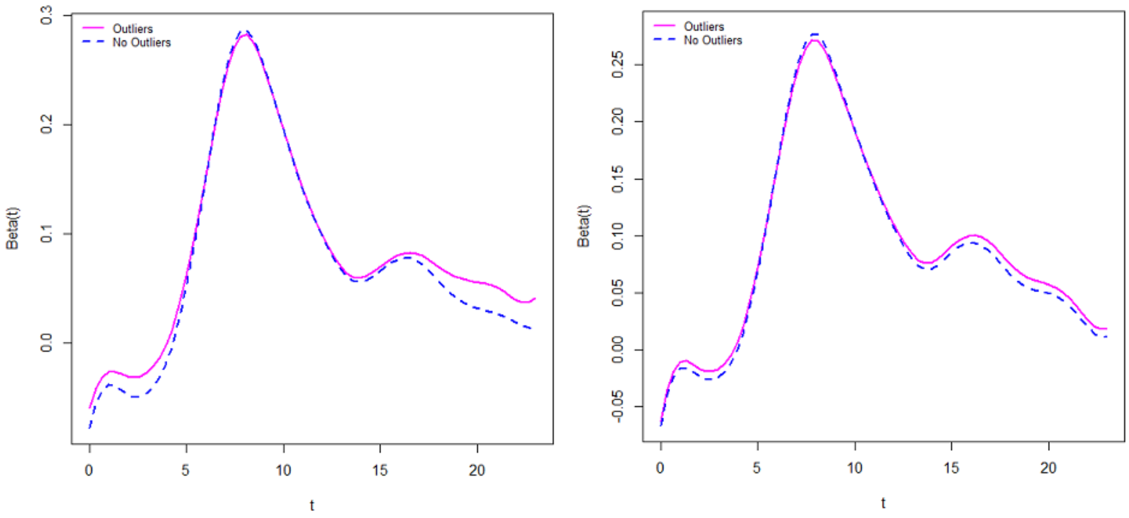
$$PVE(K_X) = \frac{\sum\limits_{p=1}^{K_X} \lambda_p}{\sum\limits_{p=1}^{24} \lambda_p},$$

where K_X represents the minimum number of principal components needed to explain 99% of the total variation in the model.

The estimated parameter function for the three methods discussed in this chapter are shown in Figure (4.4). The pattern of the parameter is similar for all three methods with slight differences for the late afternoon to early morning time interval (i.e. 18:00 hrs to 24:00 hrs and also between 01:00 hrs and 03:00 hrs). This difference is more pronounced for the MLE and PMLE in the presence of outliers, which inevitably gives a different model interpretation for those sub-intervals. Due to the down-weighting of high leverage observations, the effect of outlying curves is minimized in our robust approach. It is interesting to note that the affected sub-interval for the estimators is the same sub-interval where the functional trajectories' behave differently from the rest of the sample curves.

4.5 Conclusion

In this chapter we proposed a robust estimation approach that downweighs high leverage points. This was achieved by reducing the functional logistic model by way of basis expansion.



(a) PMLE

(b) RPMLE

Figure 4.4: A comparison of the estimated $\beta(t)$ with and without the 5 outliers for the non-robust and robust penalized methods

The influence function for the maximum likelihood estimator for this reduced logistic model is known to be unbounded in the x-direction. Therefore, functional observations whose pattern differ from the rest of the observations, even in some sub-interval, greatly influence the estimator. It was with this in mind that weights based on the design space were used together with some regularization to impose the smoothness property of the estimated parameter.

The proposed robust penalized method is a Mallows-type estimator that uses Huber-type weights to down-weight outliers in the x-direction by using the robust Mahalanobis distance of the covariate. The penalty term makes use of the curvature of the parameter function and ensures smoothness of the estimator. The Monte Carlo study showed the increased efficiency in estimation by the robust penalized estimator. The same conclusion was arrived at for a real world data example.

Chapter 5

Diagnostic Methods for the Functional Logistic Model

5.1 Introduction

Often after fitting a model, one needs to assess the fit of the model before using it to make any inference. For this to be done, it is imperative that there's an understanding of the model as well as the fitting procedure used. The idea is on assessing how good the fit is, i.e. how well does the fit model describe the outcome. In this chapter, we adapt diagnostics methods for the standard logistic model and extend these ideas to the functional logistic model. We contribute to the diagnostics measures of this model.

Febrero-Bande et al. (12) proposed two statistics that measure the influence of each curve on the functional slope of the functional linear model, with a scalar response and functional predictor(s), when the principal components method was used in the estimation of the model. Müller and Chiou (28) developed diagnostic measures for functional regression models where the response was functional and the predictor was either multivariate vectors or random functions by proposing residual processes. Malloy et al. (27) proposed a method that extended the Box-Tidwell score test and involved construction of residual plots to detect non-linearity of the functional predictor(s) in the functional generalized linear model which has a scalar response and functional predictors. Our proposed diagnostic measures differ from all this work in two primary aspects. Firstly, our measures are for the functional logistic model which has a binary response and functional predictor(s). Secondly, we focus on measures of goodness-of fit as well as measures to identify ill-fitting observations for the model and/or observations that have a dominant influence in the fit of the model.

Pregibon (32) made a significant contribution on diagnostic measures for the standard logistic model, by extending ideas from linear regression. We explore ways that these and other diagnostic methods can be adapted for data were the predictor is functional and the response binary. We base our measures on the principal component based method discussed in chapters 2 and 3. This methodology can be extended to other estimation methods and we illustrate this by assessing the fit of the numerical examples analyzed in chapters 3 and 4.

5.2 Measures of Goodness-of-Fit

We consider the functional logistic model where the functional predictor, $X(t) \in L^2(T)$, defined on the closed interval $T = [a, b] \subset \mathbb{R}$. We assume that the unknown parameter function, $\beta(t) \in L^2(T)$, and that both functions are represented as a linear combination of a known basis, such that the probability of a positive response given the functional predictor is given as in (2.4). As discussed in chapter 2, the maximum likelihood estimate of this model is unstable and inefficient due to multicollinearity issues with the design matrix $\mathbf{X} = (\mathbf{1}|\mathbf{C}\psi)$. Therefore, we consider the principal component based approach, where we let $\mathbf{Z} = \{\xi_{ij}\}_{n\times K_X}$ be the matrix of PCs of the design matrix, such that $\mathbf{Z} = \boldsymbol{C\psi V}$, where $\mathbf{V}$ is a $K_X \times K_X$ matrix whose columns are the eigenvectors associated with the eigenvalues of the covariance matrix of $\boldsymbol{C\psi}$. Then the logit model (2.4) becomes,

$$\mathbf{L}^{(s)} = \beta_0 \mathbf{1}^{(s)} + \mathbf{Z}^{(s)} \boldsymbol{\gamma}^{(s)}, \tag{5.1}$$

in matrix form, where $\boldsymbol{\gamma} = \mathbf{V}^{'}\mathbf{b}$ and s denotes the number of principal components retained in the model. The maximum likelihood estimate is then,

$$\hat{\boldsymbol{\gamma}} = \arg\max_{\boldsymbol{\gamma}} \sum_{i=1}^{n} y_i log[\pi(\mathbf{z_i}\boldsymbol{\gamma}; y_i)] + (1 - y_i) log[1 - \pi(\mathbf{z_i}\boldsymbol{\gamma}; y_i)]. \tag{5.2}$$

Goodness-of-fit summary measures are useful in giving an indication of the fit of the model. In order to asses the fit of the model, covariate patterns need to be considered. These are distinct groupings of the covariates which we will denote by J. Therefore, should some subjects have the same values for the covariate, then $J < n$. The number of observations with the same covariate pattern is denoted by m_j for $j = 1, ..., J$ such that $\sum m_j = n$. The number of positive responses among the m_j observations having the same covariate pattern is also denoted by y_j. And so the sum of y_j would be the total number of observations with the same covariate pattern having a positive response. Therefore, the maximum likelihood estimate in (5.2) can be re-written as,

$$\hat{\boldsymbol{\gamma}} = \arg\max_{\boldsymbol{\gamma}} \sum_{j=1}^{J} y_j log[\pi(\mathbf{z_j}\boldsymbol{\gamma}; y_j)] + (m_j - y_j) log[1 - \pi(\mathbf{z_j}\boldsymbol{\gamma}; y_j)]. \qquad (5.3)$$

We discuss two measures that can be used to measure the difference between the observed and fitted values. We denote the fitted values as,

$$\hat{y_j} = m_j\hat{\pi}_j = m_j \frac{exp\left\{\hat{\beta}_0 + \sum\limits_{k=1}^{s} z_{jk}\hat{\gamma}_k\right\}}{1 + exp\left\{\hat{\beta}_0 + \sum\limits_{k=1}^{s} z_{jk}\hat{\gamma}_k\right\}} \quad j = 1, ..., J.$$

The Pearson residual is defined as,

$$r(y_j, \hat{\pi}_j) = \frac{y_j - m_j\hat{\pi}_j}{\sqrt{m_j\hat{\pi}_j(1 - \hat{\pi}_j)}}, \quad j = 1, ..., J,$$

For the case that $y_j = 0$, this reduces to,

$$r(y_j, \hat{\pi}_j) = -\sqrt{m_j}\sqrt{\frac{\hat{\pi}_j}{1 - \hat{\pi}_j}},$$

whereas in the case that $y_j = m_j$, we have

$$r(y_j, \hat{\pi}_j) = \sqrt{m_j}\sqrt{\frac{1 - \hat{\pi}_j}{\hat{\pi}_j}}$$

This has the Pearson chi-square statistic that is based on it,

$$X^2 = \sum_{j=1}^{J} r(y_j, \hat{\pi}_j)^2. \tag{5.4}$$

The deviance residual is defined as,

$$d(y_j, \hat{\pi}_j) = \pm \left\{ 2 \left[y_j log \left(\frac{y_j}{m_j \hat{\pi}_j} \right) + (m_j - y_j) log \left(\frac{m_j - y_j}{m_j(1 - \hat{\pi}_j)} \right) \right] \right\}^{\frac{1}{2}}, \ j = 1, ..., J,$$

where the sign is determined by the sign of $(y_j - m_j\hat{\pi}_j)$. In the case of no positive response, i.e. $y_j = 0$, then

$$\begin{aligned} d(y_j, \hat{\pi}_j) &= -\left\{ 2 \times m_j log \left(\frac{1}{1 - \hat{\pi}_j} \right) \right\}^{\frac{1}{2}} \\ &= -\{2m_j \mid log(1 - \hat{\pi}_j) \mid\}^{\frac{1}{2}}, \end{aligned}$$

whereas, when all observations with that covariate pattern have a positive response, i.e. $y_j = m_j$,

$$\begin{aligned} d(y_j, \hat{\pi}_j) &= \left\{ 2 \times m_j log \left(\frac{1}{\hat{\pi}_j} \right) \right\}^{\frac{1}{2}} \\ &= \{2m_j \mid log(\hat{\pi}_j) \mid\}^{\frac{1}{2}}. \end{aligned}$$

The deviance is then the summary statistic based on the deviance residual and is given as,

$$D = \sum_{j=1}^{J} d(y_j, \hat{\pi}_j)^2. \tag{5.5}$$

Both these summary statistics are distributed as χ^2 with degrees of freedom $J - (s+1)$, under the assumption that the fitted model is correct. The statistics measure the goodness of the fit of the model; X^2 measures the relative deviation between the observed and fitted values, whilst D measures the disagreement between maxima of the observed and the fitted log likelihood functions. Large values would be an indication that the observations are poorly accounted for by the model.

Graphical plots of these measures of goodness of fit can be used to allow for easier detection of those observations that are ill-fitting to the model. A plot of the Pearson residuals against standard normal quantiles is one such approach where deviations from a linear pattern is an indication of the lack of goodness of fit. Other plots that can be utilized are index plots of the residuals discussed against the observations or more commonly the discussed residuals against the fitted values, i.e. r_j vs. $\hat{\pi}_j$ or d_j vs. $\hat{\pi}_j$. Ill-fitting observations would have large values for the residuals.

5.3 Regression Diagnostics

The diagnostic measures discussed in this section will assist in identifying which of the observations are not well-explained by the model, or in pointing out those observations having dominance in some aspect of the fit and quantifying their effect on the fit. The work by Pregibon (32) was key in establishing the theoretical work that extended diagnostics from the linear regression model to logistic regression. We seek to discuss these ideas for the functional logistic model.

The analog of the projection matrix for the reduced functional logistic model (5.1) is given by,

$$\begin{aligned}\mathbf{P} &= \mathbf{I} - \mathbf{H} \\ &= \mathbf{I} - \mathbf{U}^{\frac{1}{2}}\mathbf{Z}(\mathbf{Z}'\mathbf{U}\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{U}^{\frac{1}{2}},\end{aligned}$$

where $\mathbf{U}$ is an $J \times J$ diagonal matrix with the general element $u_j = m_j\hat{\pi}_j(1-\hat{\pi}_j)$ and $\mathbf{H}$ is the hat matrix. A diagonal element from this hat matrix, h_j, is therefore of the form,

$$\begin{aligned} h_j &= m_j\hat{\pi}_j(1-\hat{\pi}_j)\mathbf{z}_j^{'}(\mathbf{Z}^{'}\mathbf{U}\mathbf{Z})^{-1}\mathbf{z}_j^{'} \\ &= u_j \times q_j \end{aligned}$$

where $q_j = \mathbf{z}_j^{'}(\mathbf{Z}^{'}\mathbf{U}\mathbf{Z})^{-1}\mathbf{z}_j^{'}$. Pregibon (32) showed that, just as is the case with the linear model, $\mathbf{P}$ is symmetric, idempotent and spans the residual space. Therefore, small values of p_{jj} detect extreme points in the design space. Diagnostic plots that would be useful in identifying outlying and influential curves include the Pearson residual (r_j), the deviance residual (d_j) as discussed in the last section as well as the diagonal elements of the projection matrix (p_{jj}).

Finally, we discuss diagnostic measures that look at the effect of infinitesimal model perturbations on the model fit and therefore, quantify the effect of each observation (or subset of observations) on the model fit. The diagnostic to measure the sensitivity of the estimated parameter estimate in the reduced functional logistic model to infinitesimal perturbations of all observations with the same covariate pattern is given by,

$$\begin{aligned} \triangle\hat{\boldsymbol{\gamma}}_j &= \frac{r_j^2 h_j}{(1-h_j)^2} \\ &= \frac{r_{sj}^2 h_j}{1-h_j}, \end{aligned}$$

where r_{sj} is the studentized Pearson residual which is defined as,

$$r_{sj} = \frac{r_j}{\sqrt{1-h_j}},$$

where r_j is the Pearson residual. This measure, $\Delta\hat{\boldsymbol{\gamma}}_j$, is analogous to Cook's distance for the linear model. Peña (31) proposed a measure that differs from that of Cook's measure for the linear regression model. Unlike Cook's Distance where the influence of an observation was determined by deleting the observation and measuring the influence of that on the predictors; Peña's measure looks at the influence of an observation based on how that observation is being influenced by the rest of the data. In the case of the reduced functional logistic model, the influence of the j^{th} covariate pattern is measured as,

$$P_j = \frac{\left(\hat{y}_j - \hat{y}_{j(-j)}\right)^2}{m_j \hat{\pi}_j (1 - \hat{\pi}_j)},$$

where $\hat{y}_{j(-j)}$ is the predicted response when the j^{th} covariate pattern is removed from the sample.

If those observations with the j^{th} covariate pattern are not well fit by the model, infinitesimal perturbations cause changes in the deviance and chi-square statistics discussed in the previous section, which most often can be isolated to the residual components. To measure the effect of each covariate pattern on the model fit, the rate of change of the deviance statistic and chi-square statistic can be approximated by,

$$\Delta D_j = d_j^2 + \frac{r_j^2 h_j}{1 - h_j},$$

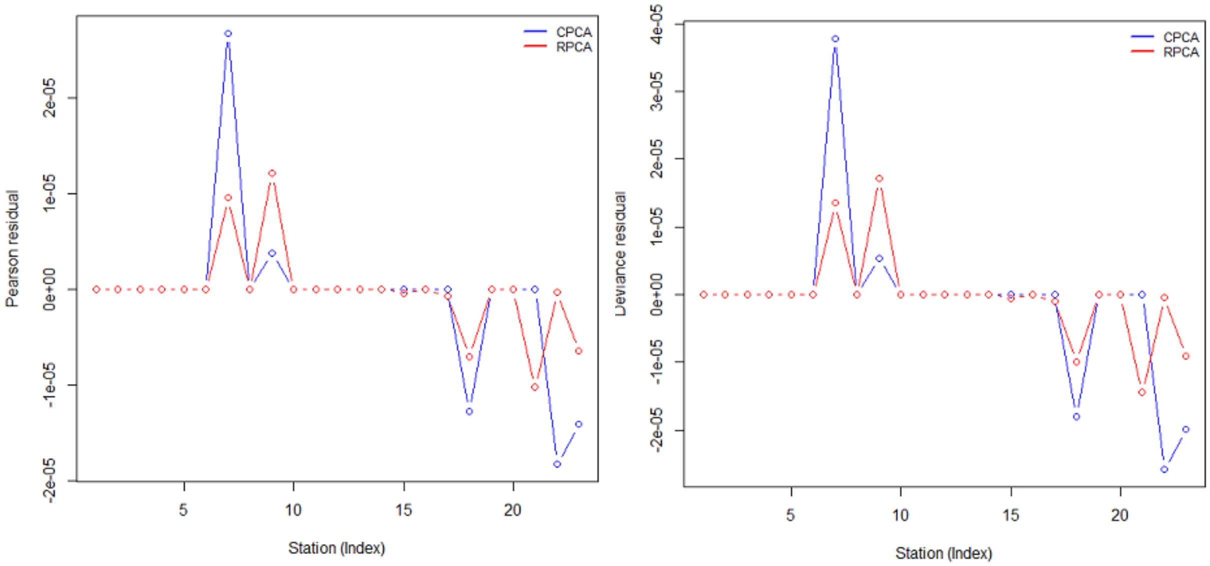
and

$$\Delta X_j^2 = \frac{r_j^2}{1 - h_j},$$

respectively. Those covariate patterns that are poorly fit by the model will be identified by large values for ΔD_j and ΔX_j^2. Similarly, those covariate patterns having the greatest influence on the values of the estimated parameter are identified by large values for $\Delta \hat{\gamma}_j$.

5.4 Numerical Examples

The diagnostic measures discussed in this chapter are applied to data sets that were previously used in earlier chapters. The Canadian Weather data set with the outlier is fitted



(a) r_i vs. index

(b) d_i vs. index

Figure 5.1: Index plots for the Canadian Weather data comparing residuals of the principal-component based approaches

using the (robust) principal component based approach. The Poblenou NO_X emission data which has been identified to have have 5 outlying curves is fitted using a (robust) penalized maximum likelihood approach. In both instances, we compute the diagnostic measures and goodness-of-fit measures for the robust and the non-robust approaches discussed in this dissertation work and by utilizing the two different estimation methods - the principal component based estimation; and the penalized maximum likelihood based estimation.

5.4.1 Canadian Weather Diagnostics

This data set has 23 samples representing the weather stations, each with 12 mean monthly temperatures recorded. Of these sample curves, $n_1 = 9$ stations have drought risk and the rest, $n_2 = 14$, do not have drought risk. The functional logistic model from this data set was used to predict risk of drought for these 23 locations based on the functional temperature predictor. An outlier was introduced in this data set by shifting and stretching a random curve.

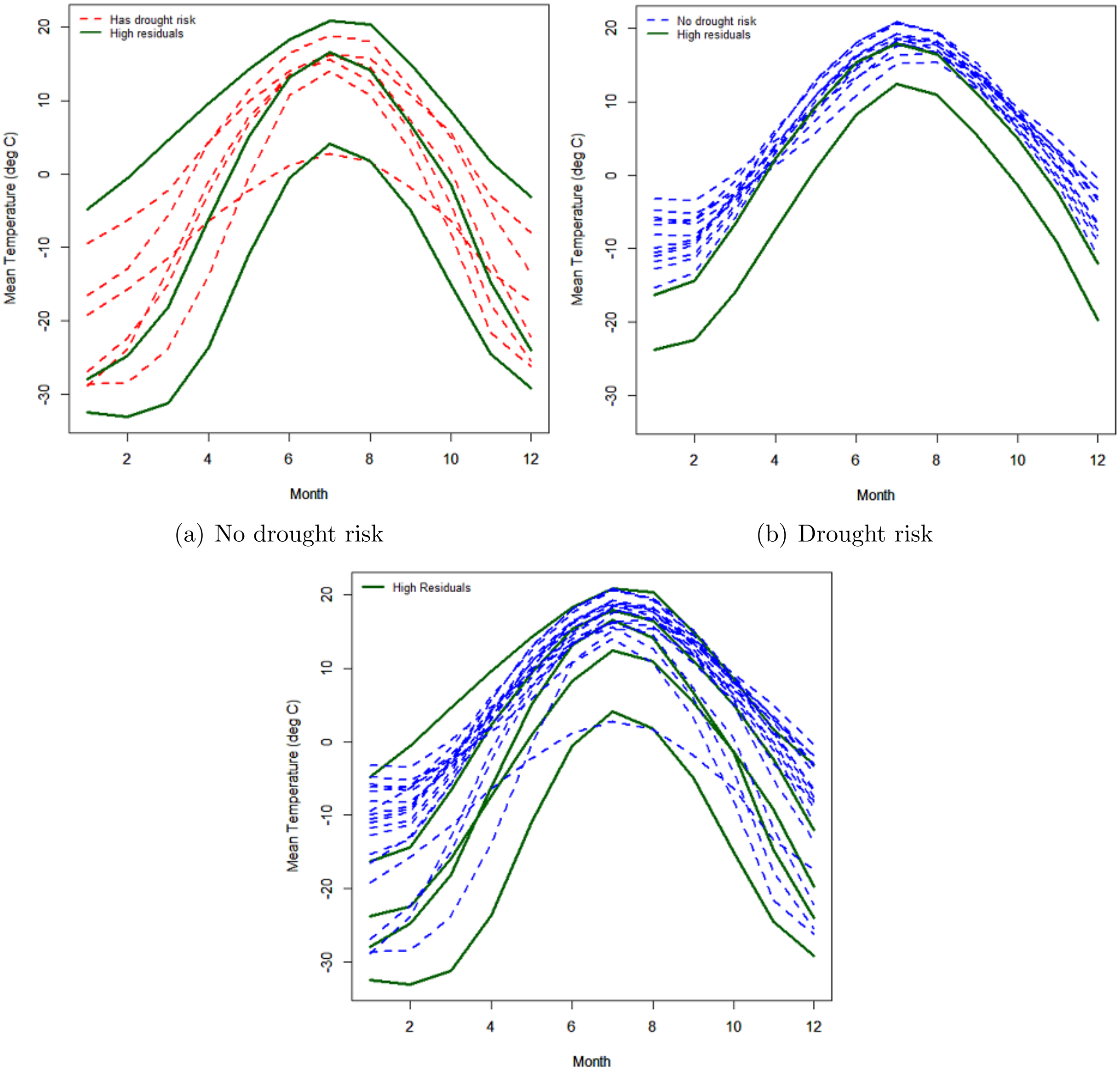
The classical principal component-based estimation, as explained in this chapter, as well as our proposed robust principal component-based estimation technique discussed in Chapter 3 are used to fit this data and their diagnostic measures compared. There were no covariate patterns distinguished and so $J = n$. Figure (5.1) displays the index plots for the two types of residuals discussed, i.e. Pearson residual and deviance residual. The residuals in both instances are relatively small, indicating that both models are a good fit of the data. On average, the robust PCA fit model has smaller residuals than the classical PCA fit functional logistic model. The summary statistics that measure the appropriateness of the fitted models were $D = 2.8414 \times 10^{-9}$ and $X^2 = 1.4207 \times 10^{-9}$ for the fit that used classical PCA approach, whilst the fit that used our proposed robust PCA method had $D = 8.6886 \times 10^{-10}$ and $X^2 = 4.344 \times 10^{-10}$ on 19 degrees of freedom in both cases.

The weather stations that were identified as having the higher residuals by both fits are shown in Figure (5.2). The identified stations are Scheffervll, Bagottville, Kamloops, Yellowknife and Resolute - the first two are classified as having no risk of drought whilst the last 3 are identified as having risk of drought. However, it's important to note that these residuals are still relatively small and so the appropriateness of the fit is met.

The diagnostic measures $\Delta\hat{\gamma}_j$, ΔX_j^2 and ΔD_j were evaluated and a summary of these statistics plotted against h_{jj} as shown in Figure (5.3). The values of these statistics are small and therefore, we can conclude that there are no observations that are poorly fit or having a great influence on the values of the estimated parameter. The temperature function for Scheffervll (#7) has noticeably higher values for the non-robust model fit.

5.4.2 Poblenou NO_X Emission Diagnostics

The Poblenou NO_X example is considered as a functional logistic regression model where the functional predictor is the functional trajectories of NO_X levels over a 24 hour period with a binary response indicating whether the day is a working day ($Y = 1$) or a non-working day ($Y = 0$).



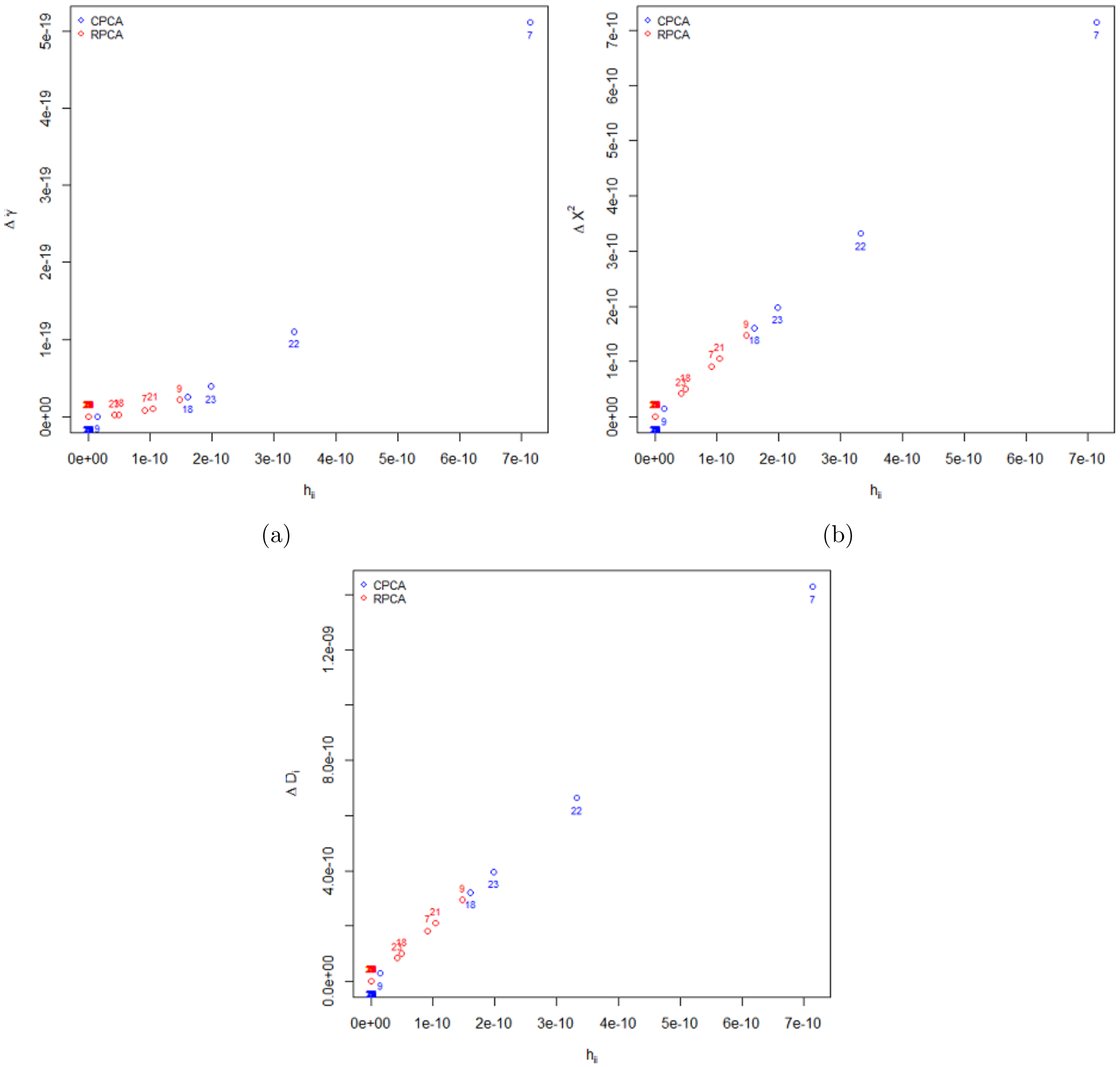
(a) No drought risk

(b) Drought risk

(c) All

Figure 5.2: Weather stations with the larger residuals shown as peaks in Figure (5.1)

The approach taken to fit this model was that of (robust) penalized maximum likelihood as discussed in Chapter 4. With this example we demonstrate that the diagnostic methods discussed in this chapter can be extended to other estimation approaches. Figure (5.4) displays the index plots of the deviance residuals (d_j) against the days for which the NO_X were recorded. There are several days with large values for the deviance residuals. Three



(a)

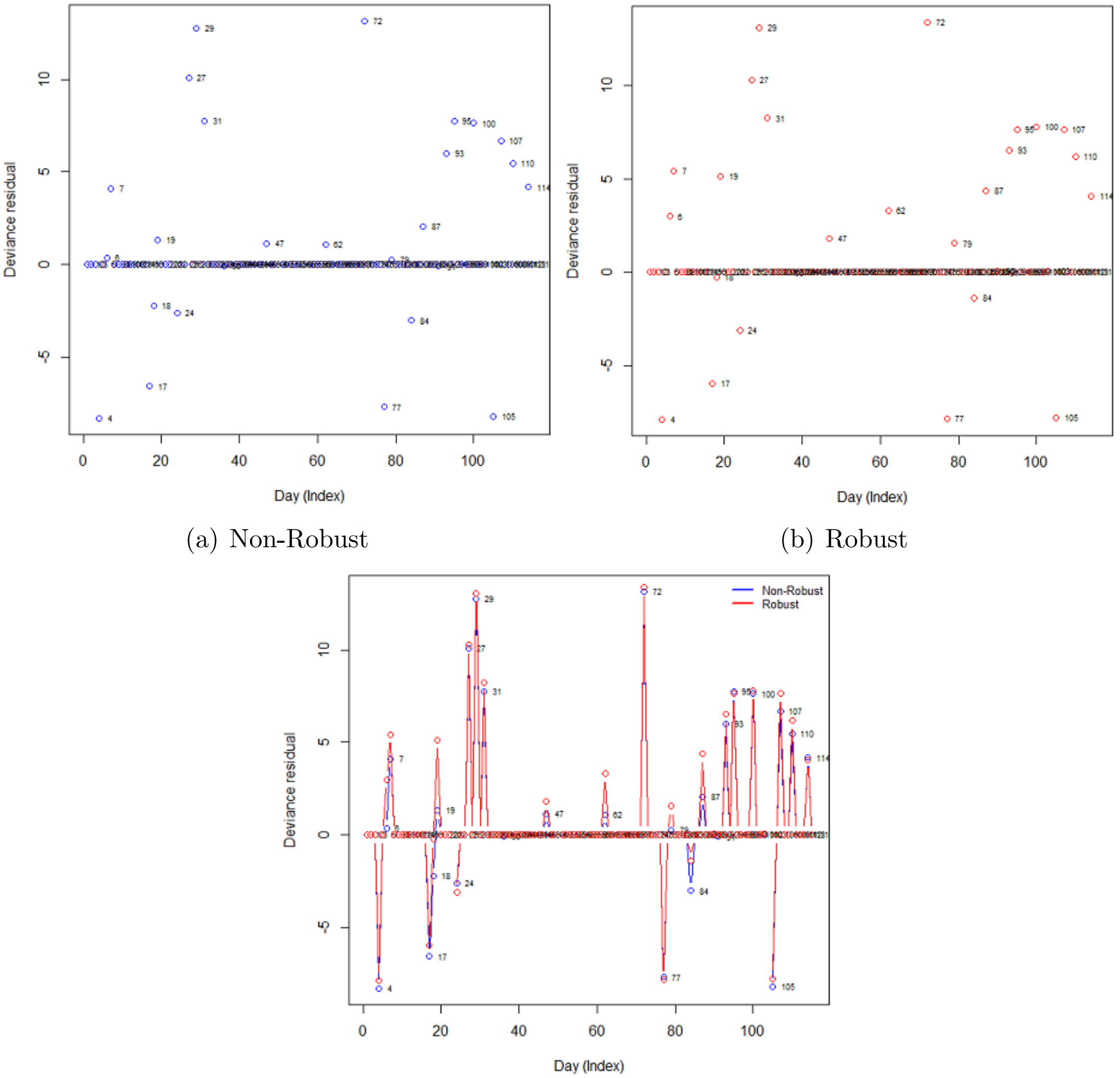
(b)

(c)

Figure 5.3: Diagnostic plots that show the influence of each observation on the different diagnostic measures versus h_{ii}

noticeable days are 03/23(#27), 03/25(#29) and 05/16(#72) with deviance residuals above 10.

The summary statistics for the goodness of fit for the non-robust penalized approach and the robust penalized approach were extremely large in both cases (p-value $<< 0.01$). Therefore, both models are poor fits of the data and drawing inferences on the odds of a working day based on these hourly NO_X emission levels would be inappropriate.



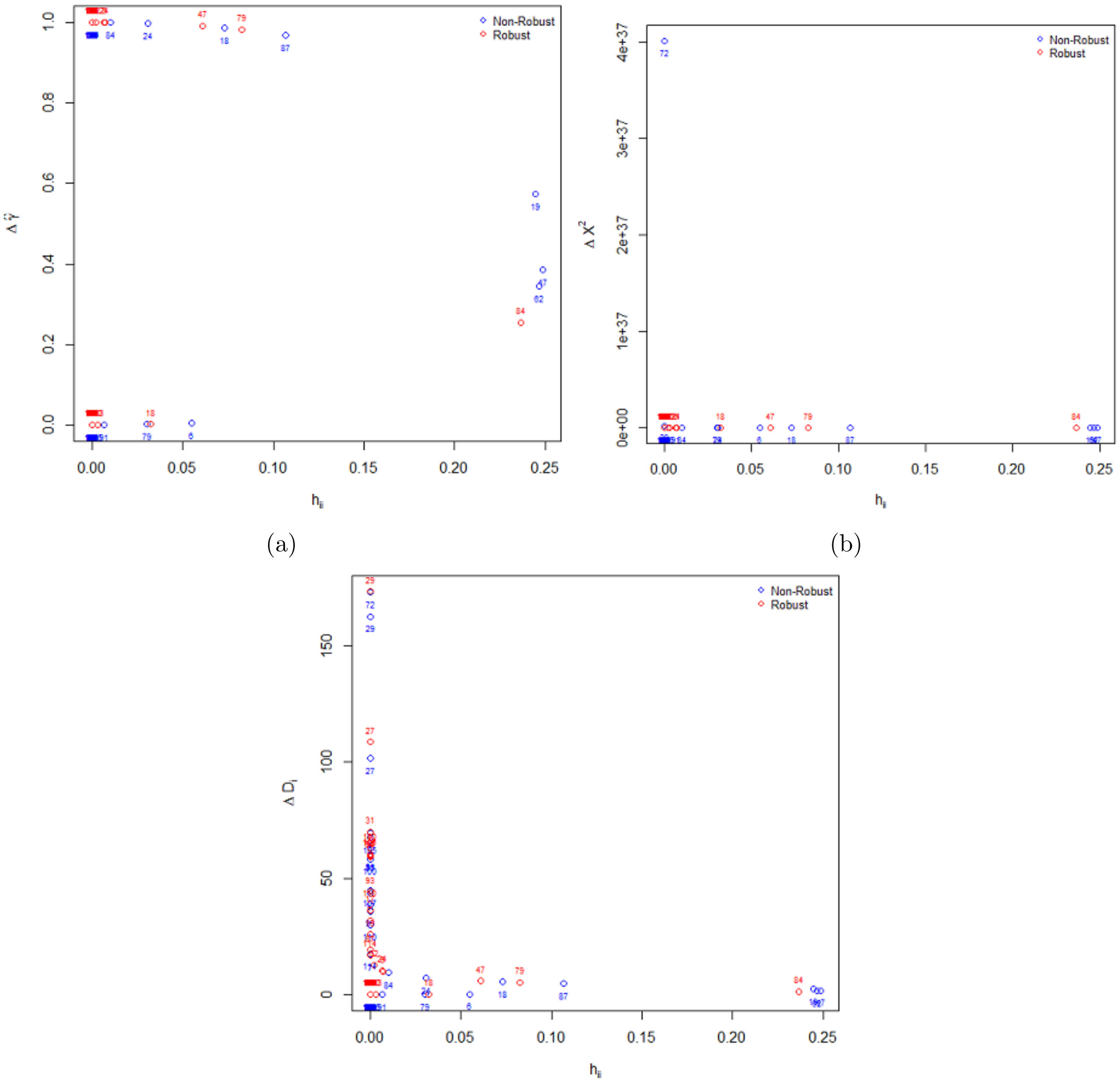
(a) Non-Robust

(b) Robust

(c) Overlay

Figure 5.4: Index plot of the deviance residuals for the NO_X penalized maximum likelihood fitted models

The diagnostic plots in Figure (5.5) also indicate the observations with the largest influence on the parameter estimate as well as the diagnostic summaries. These observations are consistent with those in the index plot.



(a)

(b)

(c)

Figure 5.5: Diagnostic plots that show the influence of each observation on the different diagnostic measures versus h_{ii} for the NO_X model

Chapter 6

Conclusion

The aim of this dissertation work was to propose robust statistical methods for the functional logistic model, which has functional predictors and a binary response. Whilst an increasing amount of work has been directed in functional data analysis, we are not aware of any work specifically looking at robust methods for the functional logistic model.

Firstly, we proposed a robust estimation technique that was principal component based. Both the functional predictor and functional parameter were estimated using basis expansion, where the order of expansion was determined by the generalized cross-validation method. The reduced functional logistic model had multicollinearity issues that were eliminated by robustly obtaining principal components that were used as the covariates in the model. The estimated functional parameter was shown to be more efficient in a simulation study as well as with an application to a real dataset.

Secondly, we proposed a Mallows-type estimator. This other robust estimation approach made use of a penalization technique as well as down-weighting high leverage points. The penalty term was introduced to ensure that the estimator retained the smoothness property and it was based on the curvature properties of the $\hat{\beta}(t)$ function. Huber-type weights that were based on the robust Mahalanobis distance of the covariate were used to minimize the effect of those observations that were outliers in the design space. The proposed estimator was shown to be better and more efficient in a simulation study and on real functional data with outlying curves.

Lastly, we explored goodness-of-fit and diagnostic measures for the functional logistic model. These measures were generalizations of widely-used and known diagnostics measures of the standard logistic model. We showed that some adaptions are necessary when dealing with the functional logistic model and illustrated their usefulness and performance by analyzing the fit of the real-world examples discussed in prior chapters.

6.1 Future Work

The robust estimation approaches proposed in Chapters 3 and 4 only considered outliers in the design space. However, for the functional logistic model, outliers can be viewed as either extreme sample curves in the design space or as mis-classification errors in the response. The robust methods proposed, whilst resistant to outlying curves in the design space, are not resistant to mis-classification errors. These errors occur when observations in the $y = 0$ class are recorded as being in $y = 1$ class, and vice versa. To simultaneously addresses the problem of mis-classification as well as dealing with extreme observations in the design space, consideration of a weight function of the form $w_i = w_{y_i} \cdot w_{x_i}$ where w_{y_i} denotes weights for mis-classification problems and w_{x_i} denotes weights to down-weigh extreme data points in the design space can be made for the robust penalized approach proposed in Chapter 4. Croux et al. (8) showed that the breakdown point for the MLE is $\leq 2(p-1)/n$ where p indicates the number of predictors in the standard logistic model. The basis for the w_{x_i} is therefore made from the hat matrix $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-\mathbf{1}}\mathbf{X}'$ where $\mathbf{X}$ is the $n \times (p+1)$ design matrix for our reduced functional model. In particular, one can consider the Huber-type weights

$$\begin{aligned} w_{x_i} &= min\left\{1, \frac{2K_b n}{h_{ii}}\right\}, \\ w_{y_i} &= min\left\{1, c\left|\frac{y_i - \pi_i}{[\pi_i(1-\pi_i)]^{\frac{1}{2}}}\right|^{-1}\right\} \end{aligned}$$

where c is a tuning constant that controls the degree of robustness.

The estimation methods proposed in this dissertation work have taken an approach that reduces the functional predictors into multivariate predictors by way of basis expansion. Therefore, in both cases robust multivariate techniques were used. Zhang and Chen (45) showed that the smoothing step in these approaches may result in some bias. A fully functional approach would be ideal, especially in the case of the principal component approach. Escabias et al. (9)'s work contrasted a functional PCA approach with one that used multivariate PCA techniques on a reduced functional logistic regression. Gervini (15) proposed a fully functional approach for spherical PCA. Lee et al. (24) also recently proposed an M-estimation type functional principal component analysis method. These methods result in robust principal functions that can be used as functional predictors in the functional logistic model.

The nature of the functional logistic model, allows it to be used for classification purposes where the data has two classes and the observations are functional. Robust functional classification methods would be applicable in diverse fields where there are two or more classes. The functional logistic model is part of the generalized functional linear model family. Therefore, a broader focus on robust estimation methods for this family could be made. Finally, the attraction to work with functional data stemmed from brain imaging studies and the application of statistical methods and tools with functional Magnetic Resonance Imaging (fMRI) data. An increasing amount of study and focus has emerged with brain images and other neuro-imaging studies. In future work, a look at the application of these functional statistical methods in dealing with this particular data will be a main focus.

Bibliography

[1] J. L. Bali, G. Boente, D. E. Tyler, and J.-L. Wang. Robust functional principal components: A projection-pursuit approach. *The Annals of Statistics*, 39:2852 – 2882, 2011.

[2] L. Barker and C. Brown. Logistic regression when binary predictor variables are highly correlated. *Statistics in Medicine*, 20:1431 – 1442, 2001.

[3] G. Boente and R. Fraiman. Discussion of robust principal components for functional data by locantore et al. *Test*, 8:28 – 35, 1999.

[4] H. Cardot and P. Sarda. Estimation in generalized linear models for functional data via penalized likelihood. *Journal of Multivariate Analysis*, 92(1):24 – 41, 2005.

[5] R.J. Carroll and S. Pederson. On robustness in the logistic regression model. *Journal of the Royal Statistical Society (B)*, 55:693 – 706, 1993.

[6] J.B. Copas. Binary regression models for contaminated data. *Journal of Royal Statistical Society (B)*, 50:225 – 265, 1988.

[7] P. Craven and G. Wahba. 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.

[8] C. Croux, G. Dhaene, and D. Hoorelbeke. The breakdown behaviour of the maximum likelihood estimator in the logistic regression model. *Statistics and Probability Letters*, 60:377 – 386, 2002.

[9] M. Escabias, A.M. Aguilera, and M.J. Valderrama. Principal component estimation of functional logistic regression: Discussion of two different approaches. *Journal of Nonparametric Statistics*, 16(3 – 4):365 – 384, 2004.

[10] M. Escabias, A.M. Aguilera, and M.J. Valderrama. Modeling environment data by functional principal component logistic regression. *Environmetrics*, 16:95 – 107, 2005.

[11] M. Febrero, P. Galeano, and W. González-Manteiga. Outlier detection in functional data by depth measures, with application to identify abnormal NO_x levels. *Environmetrics*, 19:331 – 345, 2007.

[12] M. Febrero-Bande, P. Galeano, and W. González-Manteiga. Influence in the functional linear model with scalar response. *Journal of Multivariate Analysis*, 101(2):327 – 339, 2010.

[13] F. Ferraty and P. Vieu. *Nonparametric Functional Data Analysis: Theory and Practice*. Springer, 2006.

[14] R. Fraiman and G. Muniz. Trimmed means for functional data. *Test*, 10:419 – 440, 2001.

[15] D. Gervini. Robust functional estimation using the median and spherical principal components. *Biometrika*, 95:587 – 600, 2008.

[16] D. Gervini. Detecting and handling outlying trajectories in irregularly sampled functional datasets. *The Annals of Applied Statistics*, 3:1758 – 1775, 2009.

[17] J. J. Goeman. L1 penalized estimation in the cox proportional hazards model. *Biometrical Journal*, 52(1):70–84, 2010.

[18] J. Goldsmith, J. Bobb, C.M. Crainiceanu, B. Caffo, and D. Reich. Penalized functional regression. *Journal of Computational and Graphical Statistics*, 20:830 – 851, 2011.

[19] R.R. Hocking. The analysis and selection of variables in linear regression. *Biometrics*, 32:1 – 49, 1976.

[20] D.W. Hosmer and S. Lemeshow. *Applied Logistic Regression.* Wiley, second edition, 2000.

[21] M. Hubert, P.J. Rousseeuw, and K.V. Branden. Robpca: a new approach to robust principal component analysis. *Technometrics*, 47(1):64 – 79, 2005.

[22] G.M. James. Generalized linear models with functional predictors. *Journal of the Royal Statistical Society, Series B*, 64(3):411 – 432, 2002.

[23] S. Le Cessie and J.C. van Houwelingen. Ridge estimators in logistic regression. *Journal of the Royal Statistical Society*, 41(1):191 – 201, 1992.

[24] S. Lee, H. Shin, and N. Billor. M-type smoothing spline estimation for principal functions. *Computational Statistics and Data Analysis*, 2013.

[25] X. Leng and H.G. Müller. Classification using functional data analysis for temporal gene expression data. *Bioinformatics*, 22:68 – 76, 2006.

[26] C.L. Mallows. On some topics in robustness. *Technical Memorandum*, 1975.

[27] E.J. Malloy, E.J. Bedrick, and T. Goldsmith. Diagnostics for the scale of functional predictors in generalized linear models. *Technometrics*, 49(4):480 – 489, 2007.

[28] H-G. Müller and J-M. Chiou. Diagnostics for functional regression via residual processes. *Computational Statistics and Data Analysis*, 51:4849 – 4863, 2007.

[29] H.G. Müller and U. StadtMüller. Generalized functional linear models. *The Annals of Statistics*, 33(2):774 – 805, 2005.

[30] R.T. Ogden and P.T. Reiss. Functional generalized linear models with images as predictors. *Biometrics*, 66:61 – 69, 2010.

[31] D. Peña. A new statistic for influence in linear regression. *Technometrics*, 47(1):1 – 12, 2005.

[32] D. Pregibon. Logistic regression diagnostics. *The Annals of Statistics*, 9(4):705 – 724, 1981.

[33] R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2011. URL `http://www.R-project.org`. ISBN 3-900051-07-0.

[34] J.O. Ramsay and B.W. Silverman. *Functional Data Analysis*. Springer, second edition, 2005.

[35] J.O. Ramsay, H. Wickham, S. Graves, and G. Hooker. *Functional Data Analysis*, 2012. URL `http://www.cran.r-project.org/web/packages/fda`.

[36] S.J. Ratcliffe, G.Z. Heller, and L.R. Leader. Functional data analysis with application to periodically simulated foetal heart rate data ii: Functional logistic regression. *Statistics in Medicine*, 21:1115 – 1127, 2002.

[37] P.T. Reiss, R.T. Ogden, J.J. Mann, and R.V. Parsey. Functional logistic regression with pet imaging data: A voxel-level clinical diagnostic tool. *Journal of Cerebral Blood Flow and Metabolism*, 25(S635), 2005.

[38] P. Sawant, N. Billor, and H. Shin. Functional outlier detection with robust functional principal component analysis. *Computational Statistics*, 27(1):83 – 102, 2012.

[39] R.L. Schaefer. Alternate estimators in logistic regression when the data are collinear. *Journal of Statistical Computation and Simulation*, 25(1 – 2):75 – 91, 1986.

[40] L.A. Stefanski. Infuence and measurement error in logistic regression. *PhD Thesis*, 1983.

[41] T.S. Tian. Functional data analysis in brain imaging studies. *Frontiers in Psychology*, 1(35), 2010.

[42] V. Todorov. *Robust Location and Scatter Estimation and Robust Multivariate Analysis with High Breakdown Point*, 2012. URL `http://www.cran.r-project.org/web/packages/rrcov`.

[43] E. Vágó and S. Kemenéy. Logistic ridge regression for clinical data analysis (a case study). *Applied Ecology and Environmental Research*, 4(2):171 – 179, 2006.

[44] M. Victoria-Feser. Robust logistic regression for binomial responses. 2000. URL `http://dx.doi.org/10.2139/ssrn.1763301`.

[45] J.T. Zhang and J. Chen. Statistical inferences for functional data. *The Annals of Statistics*, 35:1052 – 1079, 2007.