Robust Group Variable Selection Methods for Multiple Functional Regression Model

by

Jasdeep Pannu

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 1, 2015

Keywords: Functional Regression Model; LAD-LASSO, Outliers, Variable selection, Robust

Approved by

Nedret Billor, Chair, Professor of Mathematics and Statistics
Asheber Abebe, Associate Professor of Mathematics and Statistics
Peng Zeng, Associate Professor of Mathematics and Statistics
Guanqun Cao, Assistant Professor of Mathematics and Statistics
George T. Flowers, Dean, Graduate School

Abstract

With the advancements in science and ever changing technology to collect data, functional data have become common these days, especially in various fields such as neuroscience, chemometrics, e-commerce and computer science. Thus, in last two decades a vast amount of new statistical methodologies to analyze such data, so-called, functional data analysis, have been developed. Much research has been done in various areas of functional data analysis like functional linear regression, functional logistic regression, functional ANOVA, functional principal component analysis and functional outlier detection.

Just as in ordinary multiple regression analysis, variable selection is an important problem in the functional regression framework. The area of functional variable selection is seldom discussed in functional data analysis. The classical existing functional variable selection methods are all based on minimizing the penalized residual sum of squares, which is non-robust in nature, in the presence of outliers. In this work, we study robust variable selection methods for functional regression model with a scalar response and functional predictors in the presence of outliers.

Essentially, we consider ways that minimize the effect of outliers on the parameter estimator and selector. Since multiple parameters exist for a functional predictor group variable selection methods are used for selecting functional predictors that select grouped variables rather than individual variables. We consider the problem of selecting functional predictors using the L1 regularization in a functional linear regression model with a scalar response and functional predictors in the presence of outliers.

Four estimation approaches are discussed: functional *LAD- group LASSO*, functional *Weighted LAD- group LASSO*, functional *LAD- Adaptive group LASSO* and functional *Weighted LAD- Adaptive group LASSO*. We present an extensive simulation study and a real world example to illustrate the performance of the proposed estimators.

Acknowledgments

I take this opportunity to express my sincere thanks to my PhD adviser, Dr. Nedret Billor. Her motivation, patience, flexibility, genuine caring and concern, and faith in me during the dissertation process enabled me to attend to life while also earning my Ph.D. She provided valuable insight and direction to the end of this work. I cannot thank her enough. I am also very grateful to the remaining members of my dissertation committee, Drs. Asheber Abebe, Peng Zeng, Guanqun Cao and Alejandro Lazarte for their indispensable advice in this work.

My gratitude extends to my friends in Auburn, Achard Bindele, Sandy Kaur and Maryam Fatima, who kept things light and me smiling. Special thanks go to Melody Denhere, with whom I have a lifelong best friend and colleague. I also thank Sandeep Sandhu, who I call "sister from another mother". She has played the part of a friend, confidant and conscience. Not to forget my precious nephews and nieces, Sahib, Harleen, Aikam, Shaan, Manav, Aryan and Sada, who gave me a reprieve from work and school. Especially, Harleen and Aikam who gave me a reason to not give up when I totally gave up at one point in my life. Thank you, babies. I love you very much. Then I thank my parents who taught me about hard work, self-respect, persistence and how to be independent. Mom, especially, was a great role model of resilience, strength and character. Love you mom and dad. I also thank my brother Inder, who offered me everything that a big brother could offer. I also appreciate my sisters, Sona and Rupinder for believing in me. Last, but certainly not least, I must acknowledge God, with tremendous and deep thanks, for everything that I have in my life.

Finally, this work is dedicated to the loving memory of my grandfathers, the late S. Hari Singh and the late S. Harnam Singh, who always taught us to become better humans first, before becoming anything else. Hopefully I have made you proud *Daadu* and *Nannu*.

Table of Contents

List of Figures

List of Tables

Chapter 1

Functional Data Analysis

1.1 Introduction

Functional data analysis has become increasingly frequent and important in diverse fields of sciences, engineering, and humanities, in the last two decades. Imperative data pertaining to these fields is functional in nature, for instance, genomics data, $fMRI$ data, DTI data, weather data.

Functional data analysis has basically the same goals as those of any other branch of statistics, as pointed out in Ramsay and Silverman [26] . These include:

- Representing the data in ways that helps further analysis.
- Displaying the data in a way that highlights various characteristics.
- Studying significant sources of pattern and variation among the data.
- Explaining variation in a dependent variable using independent variable information.
- Comparing several sets of data with respect to certain types of variation, where different sets of data can contain different sets of replicates of the same functions, or different functions for a common set of replicates.

Functional data analysis includes techniques like functional logistic regression, functional linear regression, functional ANOVA, functional principal component analysis, functional outlier detection, functional linear discriminant analysis, functional variable selection, to list a few. Several authors such as Ferraty and Vieu [9], Ramsay and Silverman [26], Bali et. al [2], Cardot and Sarda [4], Gertheiss et al [11] have explored several areas of functional data

analysis.

Functional data are usually observed and recorded discretely as n pairs (t_j, y_j), and y_j is a snapshot of the function at time t_j. This can be expressed in notation as

$$y_j = x(t_j) + \epsilon_j \tag{1.1}$$

where the error term ϵ_j contributes to a roughness to the raw data.

Time is so often the continuum over which functional data are recorded, but certainly other continua, such as spatial position, frequency, weight, etc. may be involved. The term "functional" in reference to observed data refers to the intrinsic structure of the data rather than to their explicit form. The basic principle of functional data analysis is to think of observed data functions as single entities, rather than merely as a sequence of individual observations. Functional data analysis uses the fact that functions defined on a specific domain form an inner product vector space and can be treated algebraically like vectors. This helps in executing the counterparts of ordinary multivariate statistical methods in functional space rather than in the space spanned by vectors of individual observations.

For example, $fMRI$ data shown in Figure 1.1 [32], are considered functional for the same reason. Data delivered by the functional magnetic resonance imaging ($fMRI$) scans can be considered continuous functions of time sampled at the inter-scan interval and can be treated as functional data. Another example of functional data is annual Canadian weather dataset [26] which includes mean monthly temperatures at 35 Canadian weather stations, as shown in Figure 1.2. This dataset can be treated as functional data because the data (temperatures) are collected over 12 months (time). In addition, the underlying function x is assumed to be smooth, so that a pair of adjacent data values, y_j and y_{j+1} are necessarily linked together to some extent and unlikely to be too different from each other. If this smoothness property did not apply, there would be nothing much to be gained by treating the data as functional rather than just multivariate.

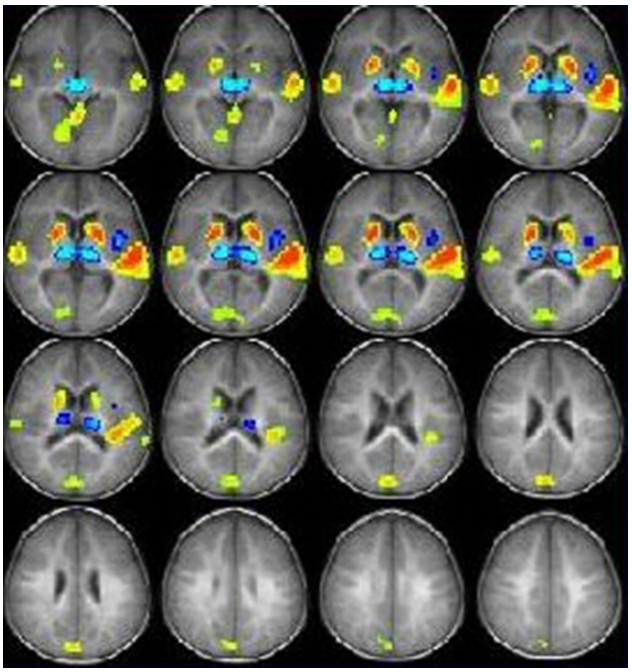


Figure 1.1: $fMRI$ Images

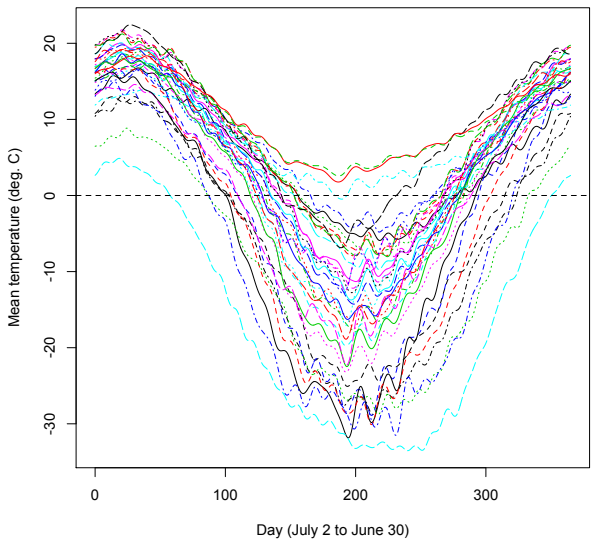


Figure 1.2: Annual Canadian Weather Data

By smooth, it usually means that function x possesses first or higher order derivatives. Usually the discrete data y_j, $j = 1, \ldots, n$ is used to estimate the function x and at the same time a certain number of its derivatives, as described next.

Representing functions using basis functions:

Assuming that a functional datum for replication i arrives as a set of discrete measured values, $y_{i1}, \ldots, y_{in}$, the first task is to convert these values to a function x_i with values $x_i(t)$ computable for any desired argument value t. If the discrete values are assumed to be error-less, then the process is interpolation, but if they have some observational error that needs removing, then the conversion from discrete data to functions may involve smoothing.

The use of linear combinations of basis functions is a well adapted computational device to represent functions. The curve x can be estimated using K known basis functions ϕ_k, as below:

$$x(t) = \Sigma_{k=1}^{K} a_k \phi_k(t) \tag{1.2}$$

which, in matrix notation becomes

$$x(t) = \mathbf{a}^T \phi \tag{1.3}$$

where, ϕ is the functional vector whose elements are the basis functions $\phi_k(t)$ and $\mathbf{a}$ is the vector of length K of the coefficients a_k.

Basis expansion method, in effect is a way to represent the infinite dimensional world of functions within the finite-dimensional framework of vectors like $\mathbf{a}$. K, therefore becomes the new dimension. There are several ways to choose K, some of which are described in Ramsay and Silverman [26]. However, It would be a mistake to infer that this simply reduces functional data analysis to multivariate data analysis, as the choice of basis system plays an important role here. Technically, basis functions should have features that match

to the functions being estimated. To summarize, most commonly used basis system these days are Fourier basis for periodic data, B-spline basis for non-periodic data and Wavelet bases where derivatives are not required, for functional data analyses.

Next we provide the following tools and summary statistics for functional data:

- The size of a function x is measured by the norm of the function x, $\| x \|$. A basic type of norm is L^2 norm given by

$$\| x \|^2 = < x, x > = \int x^2$$

  where $< x, x >$ is the inner product of x.

- The size of the second derivative of a function x, that is, $|D^2x(t)|$ or $[D^2x(t)]^2$ is used to measure its curvature at argument t.

- The functional mean:

$$\bar{x}(t) = N^{-1}\Sigma_{i=1}^{N} x_i(t) \tag{1.4}$$

  It is the average of the functions point-wise across N records.

- The variance function:

$$var_X(t) = (N-1)^{-1}\Sigma_{i=1}^{N}[x_i(t) - \bar{x}(t)]^2 \tag{1.5}$$

  and the standard deviation function is the square root of the variance function.

- The covariance function:

$$cov_X(t_1, t_2) = (N-1)^{-1}\Sigma_{i=1}^{N}[(x_i(t_1) - \bar{x}(t_1))(x_i(t_2) - \bar{x}(t_2))] \tag{1.6}$$

The covariance function summarizes the dependence of records across different t_1 and t_2. Basically, functional summary statistics are extensions of ordinary univariate summary statistics to functional data.

1.2 Functional Regression Model

Functional data are different from ordinary data because there is assumed to be an underlying curve describing the data. Functional Data can be considered a set of data consisting of a sequence (X_i, Y_i) for $i = 1, 2, \ldots, N$. Different set-ups of X_i and Y_i give rise to the following regression models as described in Ramsay and Silverman [26]:

1. A model with both functional response and functional predictor(s): In this case, the observed data are in the form of $(X_{ij}(t), Y_i(t) : t \; \varepsilon \; \mathcal{T})$, where $i = 1, \ldots, N$, $j = 1, \ldots, p$ and $\mathcal{T}$ is the support of the functional response and functional predictors which need not be same for all the predictors. Here both $X_j(t)$ and $Y_i(t)$ are real functions defined on some interval of $\Re$. The model is defined as,

$$Y_i(t) = \alpha(t) + \Sigma_{j=1}^{p} \int_{\mathcal{T}} X_{ij}(t)\beta_j(t)dt + \epsilon_i(t), \quad i = 1, \ldots, N; \; j = 1, \ldots, p \tag{1.7}$$

   where $Y_i(t)$ is the functional response, $\alpha(t)$ is the intercept function, $\beta(t)$ is the coefficient function and $\epsilon_i(t)$ is the residual function.

2. A model with a functional response and scalar predictor(s): The observed data in this case are in the form of $(X_{ij}, Y_i(t) : t \; \varepsilon \; \mathcal{T})$, where $i = 1, \ldots, N$, $j = 1, \ldots, p$ and $\mathcal{T}$ is the support of the functional response. The model is as below,

$$Y_i(t) = \alpha(t) + \Sigma_{j=1}^{p} X_{ij}\beta_j(t) + \epsilon_i(t), \quad i = 1, \ldots, N; \; j = 1, \ldots, p \tag{1.8}$$

   where $Y_i(t)$ is the functional response, X_{ij} is the scalar predictor, $\alpha(t)$ is the intercept function, $\beta(t)$ is the coefficient function and $\epsilon_i(t)$ is the residual function.

3. A model with a scalar response and functional predictor(s): In this case, the observed data are in the form of $(X_{ij}(t), Y_i : t \; \varepsilon \; \mathcal{T})$, where $i = 1, \ldots, N$, $j = 1, \ldots, p$ and $\mathcal{T}$ is the support of the functional predictor(s). The model becomes,

$$Y_i = \alpha + \Sigma_{j=1}^{p} \int_{\mathcal{T}} X_{ij}(t)\beta_j(t)dt + \epsilon_i, \quad i = 1, \ldots, N; \; j = 1, \ldots, p \tag{1.9}$$

where Y_i is the scalar response, α is the intercept, $\beta(t)$ is the coefficient function and ϵ_i is a sequence of iid centered random variables uncorrelated with X_i.

A functional model can be thought of as a continuous version of Multivariate Linear Regression. The link between predictors and responses is analyzed through the above relations. In this dissertation, we focus on the model described in equation (1.9). We develop robust variable selection methods for this model scenario. There has been an evolving literature devoted to understanding the performance of estimation of functional predictors. Escabias et al. [7], Denhere and Billor [5], Boente and Fraiman [3], Gervini [12], Bali et al. [2], Sawant et al. [29], Goldsmith et al. [15] and Ogden and Reiss [25] proposed some robust parameter estimation techniques in functional logistic regression model, functional principal component analysis and generalized functional linear models, respectively.

Just as in ordinary data analysis, variable selection is also an important aspect of functional data analysis. Functional data suffer from high dimensionality and multicollinearity among functional predictors. This could lead us to wrong model selection and hence wrong scientific conclusions. Collinearity also gives rise to issues of over fitting and model misidentification. So it is very important to perform variable selection on functional covariates. With sparsity, variable selection effectively identifies the subset of significant predictors, which improves the estimation accuracy and therefore, enhances the model interpretability. However, in the presence of outliers, that are curves deviating from the remaining of functional data, the

effective and correct selection of significant functional covariates become even more challenging.

Not much work has been done in the area of variable selection for functional predictors in functional regression models. Gertheiss et al. [11], Matsui and Konishi [22], Lian [18], Zhu and Cox [38] and Zhaoa et al. [37] proposed some variable selection techniques for functional predictors via *L1* and *L2* regularizations, for instance, using various roughness penalties like *gLASSO*, Wavelet based- *LASSO*, *gSCAD* for the generalized functional linear models. However, these methods do not work well in the presence of outliers. Since these variable selection techniques are all based on the estimation of the coefficient functions in which the estimates are obtained by minimizing the penalized residual sum of squares, which is known to be non-robust in nature. Thus, there is a need for a robust variable selection method which is resistant to outliers. Lilly and Billor [19] have proposed group $LAD - LASSO$ for multiple regression model, but to our knowledge, there is no work that has been done in the area of robust variable selection of the functional linear model.

1.3 Summary and Discussion

In this chapter we provide an insight on functional data , functional data analysis and different functional regression models. We also describe the research problem we considered for this dissertation. To summarize, in this dissertation we propose new methodologies that minimize the effect of outliers in the estimation and selection of the functional covariates in functional linear models. In this research work, we consider the problem of selecting functional predictors using the $L1$ regularization in a functional linear regression model with a scalar response and functional predictors in the presence of outliers. The first step that we take is to reformulate the functional linear model as a multiple linear one by approximating the functional covariates as a linear combination of an appropriate basis as discussed in Ramsay and Silverman [26]. Then we apply robust methods like functional *LAD-group LASSO*, *LAD-Adaptive group LASSO*, *Weighted LAD-group LASSO* and *Weighted LAD-*

Adaptive group LASSO procedures for selection of grouped variables where each functional predictor is assumed to have grouped parameters.

Chapter 2

Variable Selection Methods

In this chapter, we review existing variable selection methods, based on different penalty functions, for both ordinary data and functional data. Section 2.1 presents some variable selection methods for ordinary Multiple Regression Model and Section 2.2 discusses all the classical existing variable selection methods for functional data to our knowledge.

2.1 Variable Selection Techniques for Multiple Regression Model

In this section, we review variable selection methods using penalized estimation for ordinary data . Consider the following standard multiple regression model for an ordinary data (x_{ij}, y_i) with p predictors.

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \tag{2.1}$$

where $\mathbf{y}$ is an $n \times 1$ vector of response y_i, $\mathbf{X}$ is $n \times p$ matrix of predictors X_{ij}, $\boldsymbol{\beta}$ is an $p \times 1$ vector of coefficients β_j and $\boldsymbol{\epsilon}$ is $n \times 1$ vector of errors ϵ_i for $i= 1, \ldots, n$; $j = 1, \ldots, p$. Moreover, the ϵ_i are assumed to be statistically independent, each with mean 0 and (unknown) standard deviation σ.

In general, the estimates of the unknown coefficients β_j can be considered as minimizers of

$$\tfrac{1}{2} \parallel \mathbf{y} - \mathbf{X}\boldsymbol{\beta} \parallel^2 + \lambda^* \Sigma_{j=1}^{p} |\beta_j|^q \tag{2.2}$$

where $\lambda^* = n\lambda/q$ and $0 \leq q \leq 2$.

This is a general case of PLS (Penalized Least Squares) estimation with L_q penalty, where λ is the shrinkage parameter. Different values of q and λ give the following special cases of

PLS Regression:

1) **The *OLS* (ordinary least squares) regression ($\lambda = 0$)**

If λ is 0, then we get ordinary least squares regression, which does not use any penalization.

2) **The Ridge regression ($q = 2$)**

PLS becomes Ridge regression if the value of q is 2. Ridge regression is an L_2 regularization method developed by Hoerl and Kennard [16] in 1970.

The estimates of the regression coefficients by this method $\hat{\boldsymbol{\beta}}$ are

$$\hat{\boldsymbol{\beta}} = \underset{\beta}{\operatorname{argmin}} \ (\| \mathbf{y} - \mathbf{X}\boldsymbol{\beta} \|^2 + \lambda \Sigma_{j=1}^{p} \beta_j^2) \tag{2.3}$$

Here λ is the shrinkage parameter that controls the size of the coefficients and the amount of regularization. Also as $\lambda \downarrow 0$, we obtain the least squares solutions and as $\lambda \uparrow \infty$, we get $\hat{\beta}_{\lambda=\infty}^{ridge} = 0$, that is, intercept-only model.

3) **The *LASSO* ($q = 1$)**

If the value of q is 1, then *PLS* is called *LASSO*. *LASSO* is a variable selection method which stands for Least Absolute Shrinkage and Selection Operator. Since ridge regression fails to provide a parsimonious model with few parameters, Tibshirani [30] introduced an improved method *LASSO* in 1996. This method is an $L1$ regularization technique that simultaneously performs model selection and parameter estimation by shrinking certain coefficients to exactly 0, excluding those predictors from the model. The other, non-zero, coefficients represent variables that are relevant to the model.

The estimates of the regression coefficients by the *LASSO* method $\hat{\boldsymbol{\beta}}$ are

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}} \ (\tfrac{1}{2} \| \mathbf{y} - \mathbf{X}\boldsymbol{\beta} \|^2 + \lambda \Sigma_{j=1}^{p} |\beta_j|). \tag{2.4}$$

In general a good penalty function should result in an estimator with the following three desired properties.

1. *Unbiasedness*: The resulting estimator is nearly unbiased when the true unknown parameter is large to avoid unnecessary modeling bias.

2. *Sparsity*: The resulting estimator is a thresholding rule, which automatically sets small estimated coefficients to zero to reduce model complexity.

3. *Continuity*: The resulting estimator is continuous to avoid instability in model prediction.

The L_q penalty functions described above do not simultaneously satisfy the mathematical conditions for unbiasedness, sparsity, and continuity. But the following two penalty functions called $SCAD$ (developed by Fan and Li [8]) and Adaptive $LASSO$ (developed by Zou [39]), give estimators that satisfy these three desired properties. Next we provide the details of these penalties.

4) $SCAD$

Since the L_q penalty functions described previously do not posses the three desired properties mentioned above, Fan and Li [8] in 2001, propose a continuous differentiable penalty function called $SCAD$ (Smoothly Clipped Absolute Deviation). $SCAD$ penalty function is symmetric and have singularities at the origin to produce sparse solutions. Furthermore, $SCAD$ is also bounded by a constant to reduce bias and satisfies certain conditions to yield continuous solutions. Fan and Li [8] also show that $SCAD$ has oracle property that is, it works as well as if the correct sub-model were known in advance.

The estimates of the regression coefficients by the $SCAD$ method $\hat{\boldsymbol{\beta}}$ are minimizers of

$$\ell(\beta) + n\Sigma_{j=1}^{p} p_\lambda(|\beta_j|)$$

where $\ell(\beta)$ is a loss function of β and $p_\lambda(|\beta_j|)$ is the $SCAD$ penalty.

The first derivative of $p_\lambda(|\beta_j|)$ is given by

$$p'_\lambda(|\beta|) = \lambda\{I(|\beta| \leq \lambda) + \frac{(a\lambda - |\beta|)_+}{(a-1)\lambda} I(|\beta| > \lambda)\} \tag{2.5}$$

for some $a > 2$ and $\theta > 0$. In practice, the best pair (λ, a) can be searched for over the two-dimensional grids using some criteria, such as cross-validation and generalized cross-validation.

5) **Adaptive** *LASSO*

Since there are scenarios in which the *LASSO* selection is not consistent, therefore to overcome this problem, Zou [39] propose a new version of the *LASSO*, the Adaptive *LASSO* in 2006. Adaptive *LASSO* uses data dependent adaptive weights for penalizing different coefficients in the $L1$ penalty. The *LASSO* penalizes all the coefficients using same λ, but the Adaptive *LASSO* enforces different weights on different coefficients.

The estimates of regression coefficients by the Adaptive *LASSO* method $\hat{\boldsymbol{\beta}}$ are

$$\hat{\boldsymbol{\beta}} = \underset{\beta}{\operatorname{argmin}} \, (\tfrac{1}{2} \parallel \mathbf{y} - \mathbf{X}\boldsymbol{\beta} \parallel^2 + \lambda \Sigma_{j=1}^{p} \nu_j |\beta_j|) \tag{2.6}$$

where ν_j are data adaptive weights. For a given $\gamma > 0$, the weight ν_j can be defined as $\nu_j = \frac{1}{|\ddot{\beta}_j|^\gamma}$, where $\ddot{\beta}_j$ is an initial estimate of β_j. For example, $\hat{\beta}(ols)$ can be used as an initial estimate of β_j. Zou [39] also show that the Adaptive *LASSO* possesses the oracle property, which means it performs as well as if the true underlying model were given in advance.

Furthermore, in many statistical modeling problems, known grouping structures of the variables arise naturally. Several methods have been proposed for variable selection that respect grouping structure in variables. Below we discuss the group versions of some methods provided above.

6) **Group** *LASSO*

Group *LASSO* (*gLASSO*) is a natural extension of the *LASSO* and selects variables in a grouped manner. It is proposed by Yuan and Lin [36] in 2006.

For this consider independent observations $(y_i, x_i), \quad i = 1, \ldots, n$ where, $x_i = (x'_{i1}, \ldots, x'_{iM})'$ and x_{im} represents a group of predictors. Then the linear regression model with M group of predictors is defined as

$$Y = \alpha + \Sigma_{m=1}^{M} x_m \beta_m + \epsilon \tag{2.7}$$

where $\alpha \ \epsilon \ \Re$ is the intercept, each β_m is a vector whose components are the regression coefficients for the mth group of predictors and $Y_{n\times 1}$ is the vector of responses.

The coefficients are defined as

$$\hat{\beta} = \underset{\beta}{\text{argmin}} \ (||Y - X\beta||_2^2 + \lambda \Sigma_{m=1}^{M} ||\beta_m||_2) \tag{2.8}$$

where λ is a tuning parameter and $\beta = (\alpha, \beta'_1, \ldots, \beta'_M)'$. The penalty is a mixture of L_1 and L_2 regularization methods, the *LASSO* and the Ridge regression penalties.

7) **Adaptive group** *LASSO*

Adaptive group *LASSO* (*agLASSO*) is the group version of the Adaptive *LASSO*. It is proposed by Wang and Leng [34] to overcome the limitations of *gLASSO*, like estimation inefficiency and selection inconsistency.

Consider the same model defined in (2.7), then the Adaptive group *LASSO* based estimator minimizes the following objective function:

$$Q(\boldsymbol{\beta}) = \underset{\beta}{\text{argmin}} \ (||\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}||_2^2 + \Sigma_{m=1}^{M} \lambda_m ||\beta_m||_2) \tag{2.9}$$

As can be seen, the key difference between the *agLASSO* and *gLASSO* is that the *agLASSO* allows for different tuning parameters for different factors. Such a flexibility in turn produces

different amounts of shrinkage for different factors. Intuitively, if a relatively larger amount of shrinkage is applied to the zero coefficients and a relatively smaller amount is used for the nonzero coefficients, an estimator with a better efficiency can be obtained.

2.2 Variable Selection Techniques for Multiple Functional Regression Model

In this section, we review variable selection methods for functional data existing in literature. These methods include variable selection techniques for functional predictors via L_1 and L_2 regularizations, for instance, using roughness penalties like Wavelet based-$LASSO$, Group $SCAD$ ($gSCAD$) and Group $LASSO$ ($gLASSO$). Variable selection of the functional predictors based on non- group structured methods fails since multiple parameters exist for a functional predictor. Therefore, group based methods are used for selecting functional predictors since they selects grouped variables rather than individual variables.

Consider a functional regression modeling setup where the response Y_i is scalar for the ith subject and $X_1, X_2, \ldots, X_p$ are the squared integrable random curves, $X_j : \mathcal{T}_I \subset \Re \to \Re$ and $X_{i1}, X_{i2}, \ldots, X_{ip}$ denote their independent realizations, respectively.

For the sake of simplicity, each X_{ij} is considered to be observed without measurement error at a grid of time points $\{t_{j1}, t_{j2}, \ldots, t_{jN_j}\}$. Then a functional linear model with the scalar response and p functional predictors can be defined as :

$$Y_i = \alpha + \Sigma_{j=1}^{p} \int_{\mathcal{T}_I} X_{ij}(t)\beta_j(t)dt + \epsilon_i, \qquad i = 1, \ldots, N. \tag{2.10}$$

The main object of interest in this model is to estimate the regression coefficient functions which are assumed to be smooth and squared integrable. The random error terms ϵ_i are assumed to be independent normally distributed with mean 0 and variance σ^2. α is a scalar parameter and $\beta_j(t)$ is a parameter function for $j = i, \ldots, p$.

To overcome infinite dimensionality issue which is inherent with functional data, a multivariate generalization can be applied to this functional form of the model using basis expansion.

The curves $X_{ij}(t)$ can be discretized on a finite grid and expressed as linear combination of basis functions:

$$X_{ij}(t) = \Sigma_{b=1}^{K} a_{ijb}\phi_{jb}(t) \tag{2.11}$$

where $\phi_{jb}(t)$ are the known basis functions and a_{ijb} are the corresponding coefficients.
The coefficient functions β_j can also be expressed as linear combination of some known basis functions as:

$$\beta_j(t) = \Sigma_{b=1}^{K} c_{jb}\phi_{jb}(t) \tag{2.12}$$

where $\phi_{jb}(t)$ are the known basis functions (need not be the same as used for $X_{ij}(t)$) and c_{jb} are the unknown corresponding coefficients. These basis coefficients become the predictors in the transformed space, that need to be estimated.
After these modification to $X_{ij}(t)$ and $\beta_j(t)$ the model in (2.10) can be written in matrix form as:

$$\boldsymbol{Y} = \boldsymbol{Z}\boldsymbol{\beta} + \boldsymbol{\epsilon}. \tag{2.13}$$

We summarize the existing functional variable selection methods in the following:

1) **Wavelet based -** $LASSO$

Wavelet based-$LASSO$ method is proposed by Zhao et al [37]. Wavelets are the basis functions that can be used to efficiently approximate other functions with relatively few nonzero wavelet coefficients. The construction of a wavelet family starts with two related and suitably chosen orthonormal basic functions: the scaling function ϕ and the mother wavelet ψ. Both the $\beta_j(t)$ function and the functional predictors $X_{ij}(t)$ in model (2.10) are expressed in terms of their wavelet components.

The wavelet coefficients become the predictors in the transformed space. The $LASSO$ estimates for the coefficients of the regression model (2.13) is obtained by

$$\hat{\beta}_\lambda = \underset{\beta}{\operatorname{argmin}} \frac{1}{2}(\mathbf{Y} - \mathbf{Z}\boldsymbol{\beta})'(\mathbf{Y} - \mathbf{Z}\boldsymbol{\beta}) + \lambda \Sigma_{j=1}^{p} |\beta_j|, \tag{2.14}$$

where λ is a nonnegative tuning parameter, chosen by K-fold cross validation (CV).

2) **Group** $SCAD$

Group $SCAD$ ($gSCAD$) penalty for functional data has been proposed in two different scenarios by Matsui and Konishi [22] and Lian [18]. We discuss both approaches below:

(i) $gSCAD$ **By Matsui and Konishi [22]**

In this method, Gaussian basis functions expansion is used for both $\beta_j(t)$ function and the functional predictor $X_{ij}(t)$ in model (2.10) and then a penalized log-likelihood function is maximized

$$\ell_\lambda(\theta) = \ell(\theta) - n\Sigma_{m=1}^{M} p_\lambda(||b_m^*||_2) \tag{2.15}$$

where $l(\theta) = \Sigma_{a=1}^{n} logf(y_\alpha|x_\alpha;\theta)$ is a log-likelihood function, $p_\lambda(.)$ is a $SCAD$ penalty function, b_m^* are parameter vectors and $\parallel b_m^* \parallel_2 = \sqrt{({b_m^*}^T G_m b_m^*)}$, with the $p_m \times p_m$ positive semi- definite matrix G_m. The first derivative of the $SCAD$ penalty $p(.)$ is given by (2.5). As mentioned in Section 2.1, the $SCAD$ penalty possesses nice properties like Sparsity, Continuity, unbiasedness and oracle property.

(ii) $gSCAD$ **By Lian [18]**

In this method functional principle component analysis ($FPCA$)-based estimation is combined with $gSCAD$. As suggested by Mercer's theorem and Karhunen-Loève theorem, the basis can be taken to be the eigenfunctions of the covariance operator K, where if $k(s,t) =$

$Cov(X(s), X(t))$ then K is given by:

$$KX(t) = \int X(t)k(s,t)ds. \tag{2.16}$$

The coefficients a_{ijb} in (2.11) in this case are called the functional principal components scores of the functional data. Eigen basis functions can be estimated using various techniques. Some of the methods are described in Ramsay and Silverman [26].

To select functional predictors simultaneously, the following criterion function is minimized

$$\ell(\theta) + n\Sigma_{j=1}^{p} p_\lambda(||\beta_j||_2)$$

where $\ell(\theta)$ is least squares loss function, $||.||_2$ is the L^2 norm and $p_\lambda(.)$ is the $SCAD$ penalty as defined in (2.5).

Furthermore, Wang et al. [35] proposed using same group $SCAD$ penalty for estimating varying-coefficient models with scalar predictors and a functional response.

3) **Functional group** $LASSO$

Functional group $LASSO$ is proposed by Gerthesis et al. [11] in 2013. In this method the curves $X_{ij}(t)$ are discretized as Riemann Integration as below:

$$\int X_{ij}(t)\beta_j(t)dt \approx \Sigma_m X_{ij}(t_m)\beta_j(t_m). \tag{2.17}$$

The coefficient functions β_j are defined using B- spline basis functions as described in (2.12). The following objective function is minimized in this method:

$$\ell(\beta) + P_{\lambda,\varphi}(\beta_j) \tag{2.18}$$

where $\ell(\beta)$ is a least square loss function and $P_{\lambda,\varphi}(\beta_j)$ is the penalty function defined by Meier et al. [23].

Specifically,

$$P_{\lambda,\varphi}(\beta_j) = \lambda(||\beta_j||_2^2 + \varphi||\beta_j^{''}||_2^2)^{1/2} \tag{2.19}$$

where $||.||_2^2 = \int(.)^2 dt$ is the L^2 norm and $\beta_j^{''}$ is the second derivative of β_j.

Here, λ is the parameter that controls sparseness and φ is the smoothing parameter that controls smoothness of the coefficients. As the sparsity parameter λ increases, the estimated coefficient functions β(t)'s are shrunk and at some value, set to zero. As the smoothing parameter φ increases, the departure from linearity is penalized stronger and thus the estimated curves become closer to a linear function. Smaller values for φ result in very wiggly and difficult to interpret estimated coefficient functions. For optimal estimates (in terms of accuracy and interpretability), an adequate (λ, φ) combination has to be chosen. λ and φ are selected via K-fold cross-validation. The most commonly used values of K are 5 and 10.

4) **Functional Adaptive group** $LASSO$

Functional Adaptive group $LASSO$ is also proposed by Gerthesis et al. [11].

The penalty function $P_{\lambda,\varphi}(\beta_j)$ in equation (2.23) is modified as below for this method:

$$P_{\lambda,\varphi}(\beta_j) = \lambda(\kappa_j||\beta_j||_2^2 + \varphi\nu_j||\beta_j^{''}||_2^2)^{1/2}. \tag{2.20}$$

where the weights κ_j and ν_j are chosen in a data-adaptive way. The choice of weights is meant to reflect some subjectivity about the true parameter functions and to allow for different shrinkage and smoothness for the different covariates. One possibility for choosing the weights is to use initial parameters estimates, based on smoothing solely, but without using sparseness-assumptions. Adaptive estimation has been shown to reduce the number of false positives considerably in penalty-based variable selection.

2.3 Summary and Discussion

In this chapter we provided a literature review of some of the variable selection methods for ordinary multiple regression model and all variable selection methods for functional regression model to our knowledge. We notice that, since multiple parameters exist for a functional predictor, so the methods which select grouped variables rather than individual variables, should be used for functional regression models. We also notice that the variable selection methods in the functional linear regression literature are merely special cases of the penalized least squares regression and, as a result, suffer from the presence of outliers. Some outlier resistant methods such as Maronna and Yohai [21] have been proposed to address the robustness in estimation of the functional parameters. However, to our knowledge, no robust method has been proposed for simultaneous parameter estimation and selection for functional regression models. This sets up the bases for our work proposed in the next chapters, where we introduce four robust variable selection methods for functional regression model. These methods are called functional *LAD- group LASSO*, functional *LAD- Adaptive group LASSO*, functional *WLAD- group LASSO* and functional *WLAD- Adaptive group LASSO*. Functional *LAD- group LASSO* and functional *LAD- Adaptive group LASSO* are discussed in Chapter 3 and functional *WLAD- group LASSO*, functional *WLAD- Adaptive group LASSO* are discussed in Chapter 4.

Chapter 3

Robust Group Variable Selection Methods for Multiple Functional Model in the Presence of Outliers in the Response Variable

3.1 Introduction

Variable selection is an important problem in functional regression analysis, just as in ordinary regression analysis. Usually, a large number of predictors are introduced at the initial stage of the regression model to mitigate possible modeling biases. However, including unnecessary predictors can vitiate the estimation and prediction efficiency of the resulting procedure. On the other hand, omitting an important explanatory variable may produce biased parameter estimates and prediction results.

Functional variable selection problem when multiple functional observations exist is fairly new research problem in functional regression model, therefore only a few studies have been published on this statistical problem. Because functional regression coefficients in a multiple functional regression model are far more complicated than scalar regression coefficients in classical multiple linear regression, selection of functional predictors for predicting the responses, even if p, the number of functional predictors, is small, requires the development of new variable selection methodologies or the extension of the existing ones to the multiple functional regression model.

Further, since multiple parameters exist for a functional predictor, so a group structure based techniques, which select grouped variables rather than individual variables, should be used for functional models. Therefore, since variable selection is an important task in functional regression analysis, a number of methods, including the functional group $LASSO$ [11] and functional group $SCAD$ [18], [22] have been proposed.

The variable selection methods in the functional linear regression literature are special cases

of the penalized least squares regression and, as a result, the presence of outliers has a serious effect on the resulting estimators. Some outlier resistant loss functions such as biweight Maronna and Yohai [21] have been proposed to address the robustness in estimation of the functional parameters. However, to our knowledge, no robust method has been proposed that carries out robust parameter estimation and variable selection simultaneously for functional regression models.

In this chapter we propose to use a combination of the well known robust loss function least absolute deviation (LAD) and penalty function group $LASSO$, where the functional parameters are estimated and selected through the minimization of the sum of the absolute value of the errors and penalizing the parameter functions. This method is called functional *LAD-group LASSO* (*LAD-gLASSO*). However, in this method same amount of penalty is applied to all the parameters. In order to reflect some subjectivity about the true parameter functions and to allow for different shrinkage and smoothness for the different functional predictors, in this chapter we also propose an alternative penalty function based on adaptive weights for the penalized estimation criterion. This method is called functional *LAD-Adaptive group LASSO* (*LAD - agLASSO*).

The LAD regression method, which is a special case of the M-estimation method, is particularly well-suited to the heavy-tailed error distributions. However, it is well-known that the LAD based method is only resistant to the outlier in the response variable, but not resistant to the outliers in the explanatory variables (leverage points). Wang and Leng [33] also point out that combining the LAD and the $LASSO$ methods can only produce estimators that are only resistant to the outliers in the response variable.

To deal with the outliers in the functional explanatory variables we propose a weighted version of the functional *LAD- gLASSO* method. This method is called functional *Weighted LAD-gLASSO* (*WLAD- gLASSO*). This method is not only resistant to outliers in the response variable but also minimizes the effect of the leverage points by introducing weights

which are only dependent on the explanatory variables. This method is discussed in chapter 4. In chapter 4, we also provide an adaptive version of the functional *WLAD-gLASSO* in which adaptive *LASSO* penalty criterion is used to assign different weights to different coefficients to penalize them differently. This method is called functional *Weighted LAD-Adaptive group LASSO* (*WLAD- agLASSO*).

3.2 Methodology

In order to estimate the parameter functions based on multivariate variable selection idea we follow two steps. The first step is to formulate the given functional model in a usual multiple regression model form to overcome infinite dimensionality issue which is inherent with functional data. The second step is to apply a robust variable selection method based on robust version of group *LASSO* that would select the influential functional predictors in predicting the response.

In this section we will first give a description for a functional regression model with a scalar response and functional predictors and present a method to reformulate this model as an ordinary multiple regression model.

As mentioned in chapter 1, functional data are usually sampled discretely over a continuum, usually time and we assume that there is an underlying curve describing data. In the usual functional regression modeling setup, we assume that the response Y_i is scalar for the ith subject and X_1, X_2, ..., X_p are the squared integrable random curves, $X_j : \mathcal{T}_I \subset \Re \to \Re$ and X_{i1}, X_{i2}, ..., X_{ip} denote their independent realizations, respectively.

We also assume that the mean function of the underlying trajectories, X_j is equal to zero. For the sake of simplicity, each X_{ij} is considered to be observed without measurement error at a dense grid of time points $\{t_{j1}, t_{j2}, \ldots, t_{jN_j}\}$. Then a functional linear model with the scalar response and p-functional predictors can be defined as :

$$Y_i = \alpha + \Sigma_{j=1}^{p} \int_{\mathcal{T}_I} X_{ij}(t)\beta_j(t)dt + \epsilon_i, \quad i = 1, \ldots, N. \tag{3.1}$$

Our main intent in this model is to estimate the regression coefficient functions, which are assumed to be smooth and squared integrable. The random error terms ϵ_i are assumed to be independent normally distributed with mean 0 and variance σ^2. α is a scalar parameter and $\beta_j(t)$ is a parameter function for $j = 1,\ldots,p$.

To overcome infinite dimensionality problem, we use basis approximation method. This requires the use of pre-set basis functions expansion for approximation of the parameter functions, $\beta_j(t)$ as well as for approximation of the functional predictors, $X_{ij}(.)$. The choices of basis functions are associated with characteristics of the parameter functions and functional predictors and they do not have to be the same basis functions. Then the integral in (3.1) can be approximated by Riemann sum as

$$\int X_{ij}(t)\beta_j(t)dt \approx \Sigma_m X_{ij}(t_m)\beta_j(t_m) \tag{3.2}$$

where,

$$\beta_j(t) = \Sigma_{b=1}^{l} c_{jb}\phi_{jb}(t) \tag{3.3}$$

Here $\boldsymbol{\Phi_j}(t) = (\Phi_{j1}(t),\ldots,\Phi_{jl}(t))$ is a finite basis and c_{jb} are the corresponding basis coefficients for $b = 1,\ldots,l$.

Using (3.2) and (3.3), the integral on the right side of the model equation in (3.1) approximates to the following:

$$\int X_{ij}(t)\beta_j(t)dt \approx \Sigma_b\{\delta_j\Sigma_m X_{ij}(t_{jm})\phi_j(t_{jm})\}c_{jb} = \Sigma_b\Phi_{ijb}c_{jb} = \boldsymbol{\Phi_{ij}^T c_j} \tag{3.4}$$

where $\delta_j = t_{jm} - t_{j,m-1}$, $\mathbf{c_j} = (c_{j1},\ldots,c_{jl})^T$, $\mathbf{\Phi_{ij}} = (\Phi_{ij1},\ldots,\Phi_{ijl})^T$ and $\Phi_{ijb} = \delta_j\Sigma_m X_{ij}(t_{jm})\phi_{jb}(t_{jm})$. for $i = 1,\ldots,N$ and $j = 1,\ldots,p$.

The new model in the usual multiple regression form is then written as

$$Y_i = \alpha + \Sigma_{j=1}^{p}\mathbf{\Phi_{ij}^T c_j} + \epsilon_i, \quad i = 1,\ldots,N \tag{3.5}$$

where $\mathbf{\Phi_{ij}}$ are known and α and $\mathbf{c_j}$'s are the unknown regression coefficients that need to be estimated. In general, due to preset grouping structure of the parameters $\mathbf{c_j}$'s, methods using ordinary penalty function cannot be applied directly to functional data. In particular, group variable selection methods are employed for functional predictors. One of the main assumptions in functional regression model as in classical multiple regression model is that data should be homogeneous, that is free of outliers. However this is almost never true in real life. Therefore, it is desirable to develop statistical methods that is robust to such curves that behave differently from the remaining curves in a functional data. In the following sections we propose two robust functional variable selection methods, which are based on shrinkage estimation, in the presence of outliers in the response variable. These are: functional *LAD group LASSO* (functional *LAD-gLASSO*) and functional *LAD Adaptive group LASSO* (functional *LAD- agLASSO*).

3.2.1 Functional LAD- group LASSO

First we discuss functional *LAD-gLASSO* and then functional *LAD- agLASSO* in the next section 3.2.2. For the simultaneous estimation of the parameter functions and sparseness of the solution, Gertheiss et al. [11] proposed a sparsity-smoothness penalty technique, which is based on the *group LASSO* penalty function, given by

$$\Sigma_{i=1}^{n}(Y_i - \alpha - \Sigma_{j=1}^{p}\mathbf{\Phi_{ij}}^T\mathbf{c_j})^2 + P_{\lambda,\varphi}(\beta_j) \tag{3.6}$$

where $P_{\lambda,\varphi}(\beta_j)$ is the *group LASSO* penalty function defined by Meier et al. [23].

However, as discussed in Section 3.1 this method is based on minimization of least squares and thus suffers from the presence of outliers, therefore necessitating a different type of approach to handle this issue. Let us write the objective function in general as

$$\Sigma_{i=1}^{n}\rho(Y_i - \alpha - \Sigma_{j=1}^{p}\mathbf{\Phi_{ij}}^T\mathbf{c_j}) + P_{\lambda,\varphi}(\beta_j) \tag{3.7}$$

where ρ is a loss function, which is robust in nature to take into account the effect of outliers. There are several robust loss functions such as biweight function. For our research we choose ρ to be an absolute value function which gives us a new criterion called functional *LAD-group LASSO*. According to this criterion, α and $\mathbf{c_j}$ can be estimated by minimizing the following:

$$\Sigma_{i=1}^{n}|Y_i - \alpha - \Sigma_{j=1}^{p}\mathbf{\Phi_{ij}}^T\mathbf{c_j}| + P_{\lambda,\varphi}(\beta_j) \tag{3.8}$$

where $P_{\lambda,\varphi}(\beta_j)$ is the penalty function as introduced by Meier et al. [23] and used by Gerthesis et al. [11] for functional variable selection. Specifically,

$$P_{\lambda,\varphi}(\beta_j) = \lambda(||\beta_j||_2^2 + \varphi||\beta_j^{''}||_2^2)^{1/2} \tag{3.9}$$

where $||.||_2^2 = \int(.)^2dt$ is the L^2 norm and $\beta_j^{''}$ is the second derivative of β_j.

Here, λ is the parameter that controls sparseness and φ is the smoothing parameter that controls smoothness of the regression coefficient functions. As the sparsity parameter λ increases, the estimated coefficient functions β(t)'s are shrunk and at some value, set to zero. As the smoothing parameter φ increases, the departure from linearity is penalized stronger and thus the estimated curves become closer to a linear function. Smaller values for φ result in very wiggly and difficult to interpret estimated coefficient functions. For optimal estimates (in terms of accuracy and interpretability), an adequate (λ, φ) combination has to be chosen. λ and φ are selected via K-fold cross-validation, in which the prediction error of the model is minimized, which is discussed in details in Section 3.3.3. Then we redefine the penalty function $P_{\lambda,\varphi}(\beta_j)$ in (3.9), as proposed by Gerthesis et al. [11].

$$P_{\lambda,\varphi}(\beta_j) = \lambda(\mathbf{c_j}^T(C_{\varphi,j})\mathbf{c_j})^{1/2} \tag{3.10}$$

where $C_{\varphi,j} = \Psi_j + \varphi\Omega_j$ is a l x l symmetric and positive definite matrix, $\Psi_j$is a l x l matrix whose $(b,k)th$ element is $\int \phi_{jb}(t)\phi_{jk}(t)dt$ and Ω_j is a l x l matrix whose $(b,k)th$ element is $\int \phi''_{jb}(t)\phi''_{jk}(t)dt$ for $b,k = 1,\ldots,l$.

Further $C_{\varphi,j}$ can be decomposed using Cholesky decomposition as following:

$$C_{\varphi,j} = L_{\varphi,j}L_{\varphi,j}^T \tag{3.11}$$

where $L_{\varphi,j}$ is non-singular lower triangular matrix. Now using (3.10) and (3.11), equation (3.8) reduces to the following:

$$\Sigma_{i=1}^n |Y_i - \alpha - \tilde{\Phi}_{ij}^T \tilde{\mathbf{c}}_\mathbf{j}| + \lambda \Sigma_{j=1}^p ||\tilde{\mathbf{c}}_\mathbf{j}|| \tag{3.12}$$

where $\tilde{\mathbf{c}}_\mathbf{j} = L_{\varphi,j}^T \mathbf{c}_\mathbf{j}$ and $\tilde{\Phi}_{ij} = L_{\varphi,j}^{-1}\Phi_{ij}$.

Now $\hat{\alpha}$ and $\hat{\mathbf{c}}_\mathbf{j}$'s are the minimizers of (3.12) and the coefficient function $\beta(t)$ is estimated by $\hat{\beta}_j(t) = \Sigma_{b=1}^l \phi_{jb}(t)\hat{c_{jb}}$ for $j = 1,\ldots,p$.

3.2.2 Functional LAD- Adaptive group LASSO

Because the group *LASSO* applies same amount of shrinkage to all of the regression coefficients, this method is not consistent in terms of model selection (Fan and Li [8]). Efficiency can also suffer due to the one shrinkage parameter (Zou [39]). As a result, an adaptive tuning parameter is introduced, which assigns a different tuning parameter for each group, allowing the shrinkage to vary from group to group. The same problems arise for *LAD-gLASSO* as well.

In this section we consider a different penalty function to allow for different shrinkage and smoothness for the different covariates. The penalty is supposed to be adaptive in nature. The functional *gLASSO* penalty function discussed in section 3.2.1 imposes same penalty on all the coefficient functions, whereas the adaptive penalty reflects some subjectivity about the true parameter functions. We call this method as functional *LAD- Adaptive group LASSO*

(*LAD - agLASSO*).

Reconsider the equation (3.7):

$$\Sigma_{i=1}^{n}|Y_i - \alpha - \Sigma_{j=1}^{p}\mathbf{\Phi_{ij}}^T\mathbf{c_j}| + P_{\lambda,\varphi}(\beta_j) \tag{3.13}$$

here, $P_{\lambda,\varphi}(\beta_j)$ is the Adaptive $LASSO$ penalty function as introduced by Zou [39] and used by Gerthesis et al. [11] for functional variable selection. Specifically,

$$P_{\lambda,\varphi}(\beta_j) = \lambda(\kappa_j||\beta_j||_2^2 + \nu_j\varphi||\beta_j^{''}||_2^2)^{1/2} \tag{3.14}$$

where $||.||_2^2 = \int(.)^2dt$ is the L^2 norm, $\beta_j^{''}$ is the second derivative of β_j, κ_j and ν_j are the data adaptive weights.

The choice of weights κ_j and ν_j is meant to reflect some subjectivity about the true parameter functions and to allow for different shrinkage and smoothness for the different covariates. One possibility for choosing the weights is to use initial parameters estimates, based on smoothing solely, but without using sparseness-assumptions.

Consider a generalized functional linear model with multiple functional covariates, and let $\ddot{\beta_j}^{'}$s, be the initial estimates of the coefficient functions $\beta_j^{'}$s, using for example, quantile regression implemented in the R package *quantreg*. Then, the adaptive weights can be defined as $\kappa_j = 1/ \parallel \ddot{\beta}_j \parallel$ and $\nu_j = 1/ \parallel \ddot{\beta}_j^{''} \parallel$. Further we consider the following three versions of Adaptive $LASSO$:

Adapt 1: In Adapt 1, weights κ_j are considered 1, which means only smoothness is concern. That is, only weights ν_j are used in the penalty function. The following function is minimized in this method:

$$\Sigma_{i=1}^{n}|Y_i - \alpha - \Sigma_{j=1}^{p}\mathbf{\Phi_{ij}}^T\mathbf{c_j}| + \lambda(||\beta_j||_2^2 + \nu_j\varphi||\beta_j^{''}||_2^2)^{1/2}. \tag{3.15}$$

Adapt 2: In Adapt 2, weights ν_j are considered 1, which means only shrinkage is concern. That is, only weights κ_j are used in the penalty function. The following function is minimized in this method:

$$\Sigma_{i=1}^{n}|Y_i - \alpha - \Sigma_{j=1}^{p}\mathbf{\Phi_{ij}}^T\mathbf{c_j}| + \lambda(\kappa_j||\beta_j||_2^2 + \varphi||\beta_j^{''}||_2^2)^{1/2}. \tag{3.16}$$

Adapt 3: In Adapt 3, both weights κ_j and ν_j are used in the penalty function, which means both smoothness and shrinkage are concern for different covariates. The following function is minimized in this method:

$$\Sigma_{i=1}^{n}|Y_i - \alpha - \Sigma_{j=1}^{p}\mathbf{\Phi_{ij}}^T\mathbf{c_j}| + \lambda(\kappa_j||\beta_j||_2^2 + \nu_j\varphi||\beta_j^{''}||_2^2)^{1/2}. \tag{3.17}$$

3.2.3 Choosing the tuning parameters

In this section we discuss how the tuning parameters λ and ϕ in (3.8) and (3.13) are selected. We consider K-fold cross-validation to select λ and ϕ. Explicitly, the original sample is randomly split into K smaller sets (roughly equal-sized). For each of the K folds, a model is trained using *(K -1)* of the folds as training data. The resulting model is validated on the remaining part of the data using the prediction error of the model given by sum of squared errors $\Sigma_i(Y_i - \hat{Y}_i)^2$. The K estimates of the prediction error are averaged and the values of the tuning parameters that minimize the overall prediction error are selected by the criterion. Most commonly used values of K in the literature are 5 and 10.

Next, in order to show the goodness of the proposed methods we perform a numerical study in which a Toy example is considered and then a simulation study is conducted. Furthermore, we show a real data application of the proposed methods.

3.3 Numerical Study

In this section we provide the numerical performance of the methods proposed in this chapter. We consider the following three models for numerical study:

- **Model (0)**: No outliers in the scalar response Y and the functional predictors X(t).
- **Model (1)**: Presence of outliers in the scalar response Y only.
- **Model (2)**: Presence of outliers both in the scalar response Y and the functional predictors X(t).

We take following steps to carry out the numerical study:

A. Generating data:

Generating Functional Predictors $X_j(t)$: The functional predictors $X(t)$ are generated similarly as in Tutz & Gerthesis [31] from

$$X_{ij}(t) = [\sigma(t)]^{-1}\Sigma_{r=1}^{5}(a_{ijr}sin(\pi t(5 - a_{ijr})/150) - m_{ijr}) \tag{3.18}$$

where i is the number of curves, j is the number of different predictors, $a_{ijr} \sim U(0,5)$, $m_{ijr} \sim U(0,2\pi)$ and $\sigma(t)$ is defined so that $\text{var}[X_{ij}(t)] = 0.01$.

Generating Y: Response Y is generated from:

$$Y_i = \alpha + \int_{\mathcal{T}} \beta_j(t)X_{ij}(t)dt + \epsilon_i \tag{3.19}$$

where i is the number of curves, j is the number of predictors and $\epsilon_i \sim N(0, 4)$.

B. Contamination of data

Contamination of Y:

In order to create outliers in response Y, the errors ϵ are generated from the standard normal distribution, the t-distribution with 2 degrees of freedom, and the t-distribution with 7 degrees of freedom. Several contamination levels (0%, 15%, 25% and 40%) are considered. However, we only present the results for 0% and 15% contamination levels. In addition, we provide information based on other attempted contamination levels and comment on empirical breakdown point based on toy example and simulation study.

Contamination of functional predictors $X_j(t)$:

We consider contaminating $X_j(t)$ at 15% level to produce functional outliers. We also considered other contamination levels, 0%, 25% and 40%, but only the result based on 0% and 15% will be given. The contamination process is carried out as described by Fraiman & Muniz [10]. The following three types of outlier curves are considered:

- **Case (1)**: Asymmetric contamination $Z_{ij}(t) = X_{ij}(t) + cM$ where c is 1 with probability q and 0 with probability 1- q and $q = \{0\%; 15\%\}$; M is the contamination constant size equal to 10 and $X_{ij}(t)$ is as defined in (3.18).

- **Case (2)**: Symmetric contamination $Z_{ij}(t) = X_{ij}(t) + c\sigma M$ where $X_{ij}(t)$, 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.

- **Case (3)**: Partial contamination $Z_{ij}(t) = X_{ij}(t) + c\sigma M$ if $t > T$ and $Z_{ij}(t) = X_{ij}(t)$ if $t < T$, where $T \sim U[0, 10]$.

Numerical Study for functional *LAD-gLASSO*

First we present numerical study for functional *LAD-gLASSO*, in which Toy example and simulation study are presented.

3.3.1 Toy Example

For Toy example we consider two functional covariates $X_1(t)$ and $X_2(t)$ which are observed at 50 equidistant points in (0, 50). 50 replications of each predictor are generated. The data are generated as described in (3.18) and (3.19). We set up the model where the response Y is related to the $X_1(t)$ and not to $X_2(t)$. The true model is :

$$Y_i = \alpha + \int_0^{50} \beta_1(t) X_{i1}(t) dt + \epsilon_i \tag{3.20}$$

where $i = 1, \ldots, 50$ and $\epsilon_i \sim N(0, 4)$.

The parameter function $\beta_1(t)$ corresponding to $X_1(t)$ has a sine-wave function shape as shown in Figure 3.1. We consider contaminating both $X_{i1}(t)$ and $X_{i2}(t)$ at 15% level. The effects of these different types of contamination on $X_{i1}(t)$ at 15 % level are shown in Figure 3.2. We apply functional *LAD-gLASSO* to all three model settings: Model (0), Model (1) and Model (2). The results are as following.

Model (0): No outliers in the scalar response Y and the functional predictors $X(t)$.
First we apply our proposed method functional *LAD-gLASSO* to Model (0) and compare it with classical functional *gLASSO*. Model (0) has neither outliers in scalar response Y nor in the functional predictors $X_1(t)$ and $X_2(t)$. The response Y is dependent only on the first predictor $X_1(t)$. Figure 3.3 shows the fitting results of functional *LAD-gLASSO* and classical functional *gLASSO* method. We use *rq.fit.lasso ()* function from the R package *quantreg* to implement our proposed method and the R package *grplasso* for the classical *gLASSO*. In Figure 3.3, the green curves display the true functions $\beta_1(t)$ and $\beta_2(t)$; the red and blue

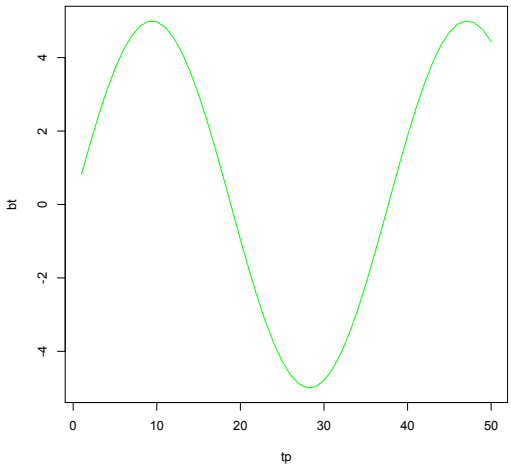


Figure 3.1: $\beta_1(t)$.

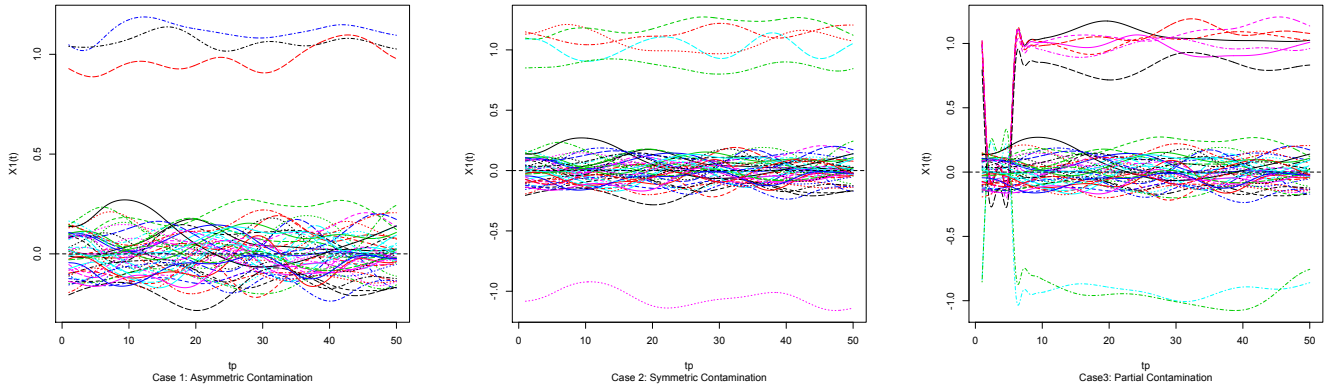


Figure 3.2: The contaminated $X_{i1}(t)$ curves for contamination cases 1- 3 ($q = 15\%$).

dashed lines display the estimations done by classical functional *gLASSO* and the functional *LAD- gLASSO*, respectively. The combination of λ and φ for boh functional *LAD-gLASSO* and the classical functional *gLASSO* is ($\lambda = 10$, $\varphi = 10$). We can see in Figure 3.3 that both methods estimate the relevant coefficient $\beta_1(t)$ close to its true value and exclude the irrelevant coefficient $\beta_2(t)$ from the model.

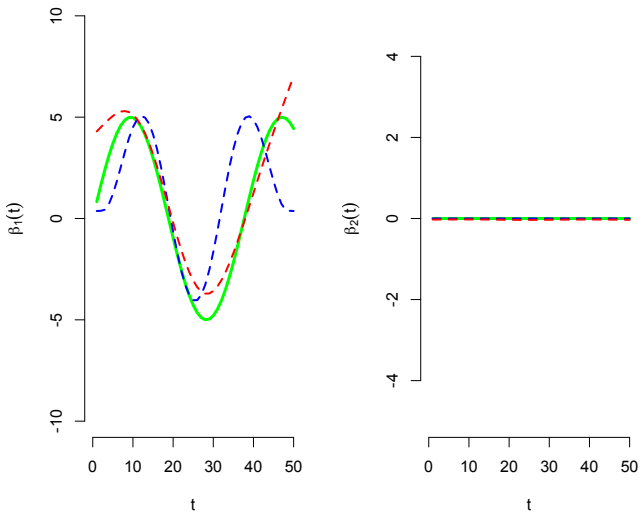


Figure 3.3: Fitting results for the comparison of functional *LAD-gLASSO* (blue) and classical functional *gLASSO* (red) for Model (0)(0% contamination).

Model (1): Presence of outliers in the scalar response Y only.

Secondly, we apply our proposed method to the Model (1). Model (1) has outliers only in scalar response Y. The functional predictors $X_1(t)$ and $X_2(t)$ are free of outliers. Also the response Y depends only the first predictor $X_1(t)$ and not on $X_2(t)$. Since $X_2(t)$ is irrelevant to the true model, so it should be excluded from the model by the applied method. Figure 3.4 shows the comparison of the classical functional *gLASSO* with the new proposed method functional *LAD-gLASSO*. R package *quantreg* was employed again to execute our proposed method. In Figure 3.4, the green solid curves display the true functions $\beta_1(t)$ and $\beta_2(t)$, the red and blue dashed lines display the estimations done by classical functional *gLASSO* ($\lambda = 10$, $\varphi = 10$) and the functional *LAD- gLASSO*($\lambda = 10$, $\varphi = 100$), respectively. Figure 3.4 shows that our proposed method is not only able to exclude the irrelevant predictor $X_2(t)$ from the model, but is also able to provide good estimation of relevant predictor $X_1(t)$ compared to the classical method. In short, the classical method does poor estimation and shrinkage, in its comparison to the our proposed robust method.

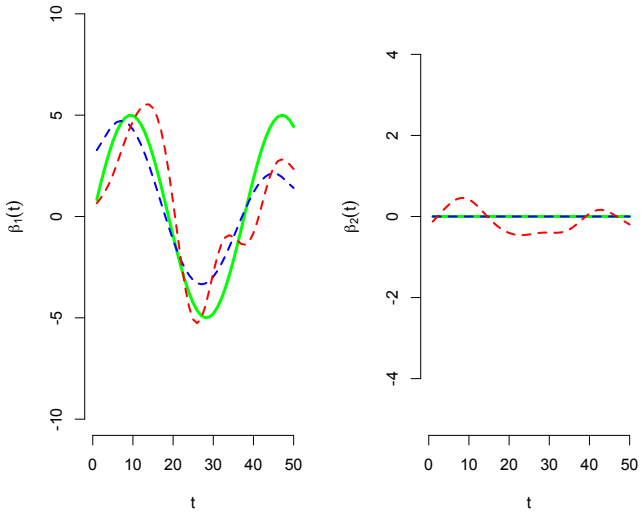


Figure 3.4: Fitting results for the comparison of functional *LAD-gLASSO* (blue) and classical functional *gLASSO* (red) for Model (1)(15% contamination).

Model (2): Presence of outliers in both scalar response Y and functional predictors $X(t)$.

Thirdly, we apply our proposed method to Model (2). Model (2) has outliers both in scalar response Y and functional predictors $X_1(t)$ and $X_2(t)$. All three cases of contamination Case 1(Asymmetric contamination), Case 2 (Symmetric contamination) and Case 3 (Partial contamination) are considered for both functional covariates. The contamination of functional covariates is done as described above. Also only the first covariate $X_1(t)$ is relevant to the true model and $X_2(t)$ being irrelevant should be excluded from the model. Figure 3.5 shows the fitting results of classical functional *gLASSO* and new functional *LAD-g LASSO* method. Again R package *quantreg* is utilized to execute the proposed method. In Figure 3.5, the green curves are the true coefficient functions $\beta_1(t)$ and $\beta_2(t)$, the red dashed lines represent the estimation done by classical functional *gLASSO* and the blue lines represent the estimation done by the functional *LAD- gLASSO*. Figure 3.5 shows that the performance of functional *LAD- gLASSO* reduces in the presence of outliers in explanatory variables, as

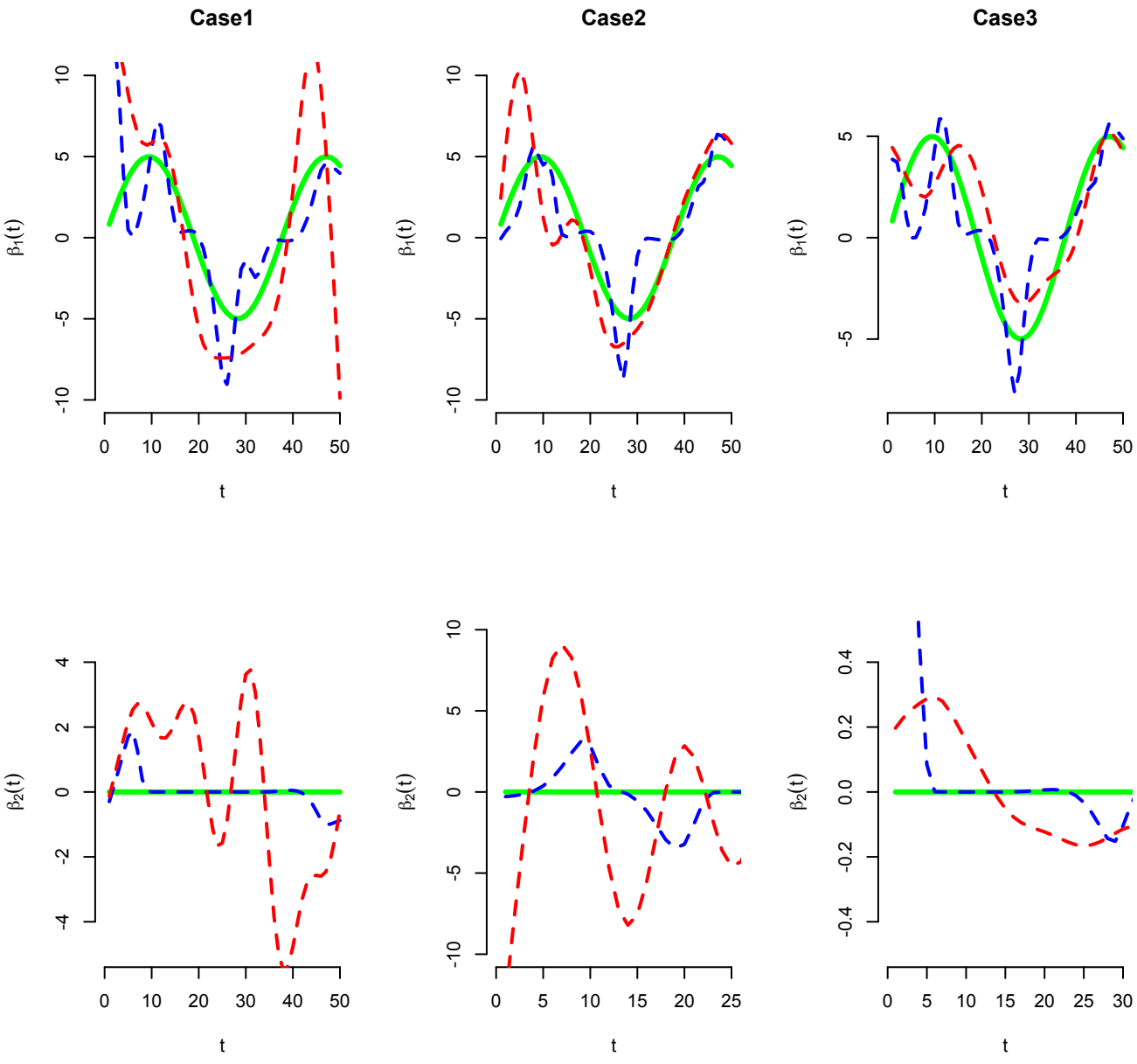


Figure 3.5: Fitting results for the comparison functional *LAD-gLASSO* (blue) and classical functional *gLASSO* (red) for Model (2)(15% contamination).

expected. The combinations of (λ, φ) for Case 1, Case2 and Case 3 contamination cases for functional *LAD- gLASSO* are ($\lambda = 10$, $\varphi = 10$), ($\lambda = 1$, $\varphi = 10$) and ($\lambda = 10$, $\varphi = 100$), respectively. The combinations of (λ, φ) for Case 1, Case2 and Case 3 contamination cases for classical functional *gLASSO* are ($\lambda = 1$, $\varphi = 10$), ($\lambda = 10$, $\varphi = 10$) and ($\lambda = 10$, $\varphi = 10$), respectively. Both methods functional *LAD- gLASSO* and classical functional *gLASSO* perform poorly in both variable estimation and selection.

3.3.2 Simulation Study

To elucidate the performance of the proposed method functional *LAD- gLASSO*, we conduct simulation study in a variety of settings. We use the same technique as described above to generate as well as contaminate the data for simulation study. We consider the following:

1) 300 observations for the scalar response Y.

2) Ten functional predictors are considered. We generate 300 sample curves for each $X_j(t)$ which are observed at 300 equidistant time points.

3) The true model is

$$Y_i = \alpha + \sum_{j=1}^{5} \int_0^{300} \beta_j(t) X_{ij}(t) dt + \epsilon_i. \tag{3.21}$$

where, $\epsilon_i \sim$ N(0,4), and the parameter functions $\beta_j(t)$ are observed at 300 equidistant points in (0, 300). The shapes of $\beta_j(t)$ are as shown in Figure 3.6. We can see in Figure 3.6 that the $\beta_6(t)$ - $\beta_{10}(t)$ are essentially 0. The true model in (3.21) depends only on $\beta_1(t)$ - $\beta_5(t)$.

In simulation study, we compare the performance of the proposed method functional *LAD-gLASSO* with classical functional *gLASSO* in terms of estimation and selection of variables for three different model scenarios Model (0), Model (1) and Model (2), as described previously. Again the contamination is done for 15% in Model (1) and Model (2).

First we consider the squared errors (SE) to assess the performance of the proposed method. The Squared Error is

$$SE = \int (\hat{\beta}_j(t) - \beta_j(t))^2 dt \tag{3.22}$$

where $\hat{\beta}_j(t)$ and $\beta_j(t)$ are the estimated and true coefficient functions, respectively. Squared errors are observed in 50 independent simulation runs for Model (0), Model (1) and Model (2). Figures 3.6 and 3.7 show the boxplots of the squared errors for Model (0) and Model (1), respectively. Figures 3.8 - 3.10 show the boxplots for all three cases of contamination for Model (2). The blue and red boxplots in these figures correspond to the functional *LAD-gLASSO* and the classical functional *gLASSO*, respectively.

Then we consider the Mean Squared Errors (MSE) and the Mean Absolute Error (MAD) of prediction to assess the predictive ability of the proposed method. The Mean Squared Errors of prediction is

$$MSE = \frac{1}{n} \sum_i (Y_i - \hat{Y}_i)^2 \tag{3.23}$$

The Mean Absolute Error of prediction is

$$MAD = \frac{1}{n} \sum_i |Y_i - \hat{Y}_i| \tag{3.24}$$

For MSE and MAD of prediction, we generate data with 5000 observations. Mean Squared Errors and the Mean Absolute Errors are observed in 50 independent simulation runs for Model (0), Model (1) and Model (2). Figure 3.11 shows the boxplots of the mean squared errors and mean absolute errors for Model (0). Figure 3.12 shows the boxplots of the mean squared errors and mean absolute errors for Model (1) at 15% of the response variable. Figures 3.13 and 3.14 show the boxplots for all three cases of contamination for Model (2).

We can see in Figures 3.6 and 3.11 that functional *LAD- gLASSO*, which is represented by blue color, performs equally well as classical functional *gLASSO*, which is represented by red color for Model (0) settings, that is when there are no outliers in the data. Also we can see in Figures 3.7 and 3.12 that the proposed method functional *LAD- gLASSO* performs better than classical functional *gLASSO* for Model (1) settings, that is when there are outliers in response variable only. Furthermore, we notice in Figures 3.8- 3.10 and Figures 3.13- 3.14, that our proposed method functional *LAD- gLASSO* does not perform any better than classical functional *gLASSO* for Model (2) settings, that is when there are outliers in both response and predictor variables.

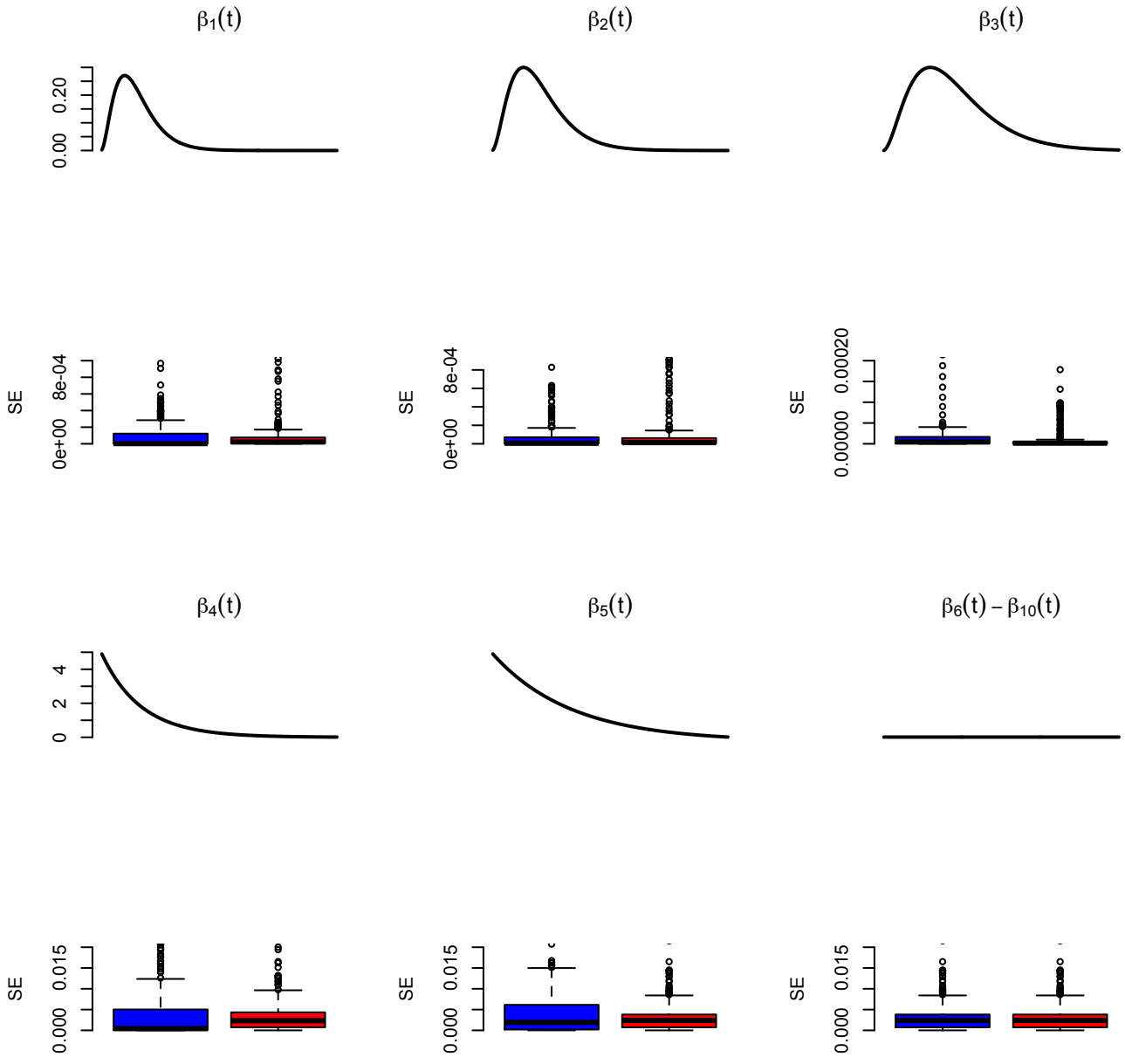


Figure 3.6: Comparison of SE of functional *LAD-gLASSO* (blue) and classical functional *gLASSO* (red) at 0% contamination for Model(0).

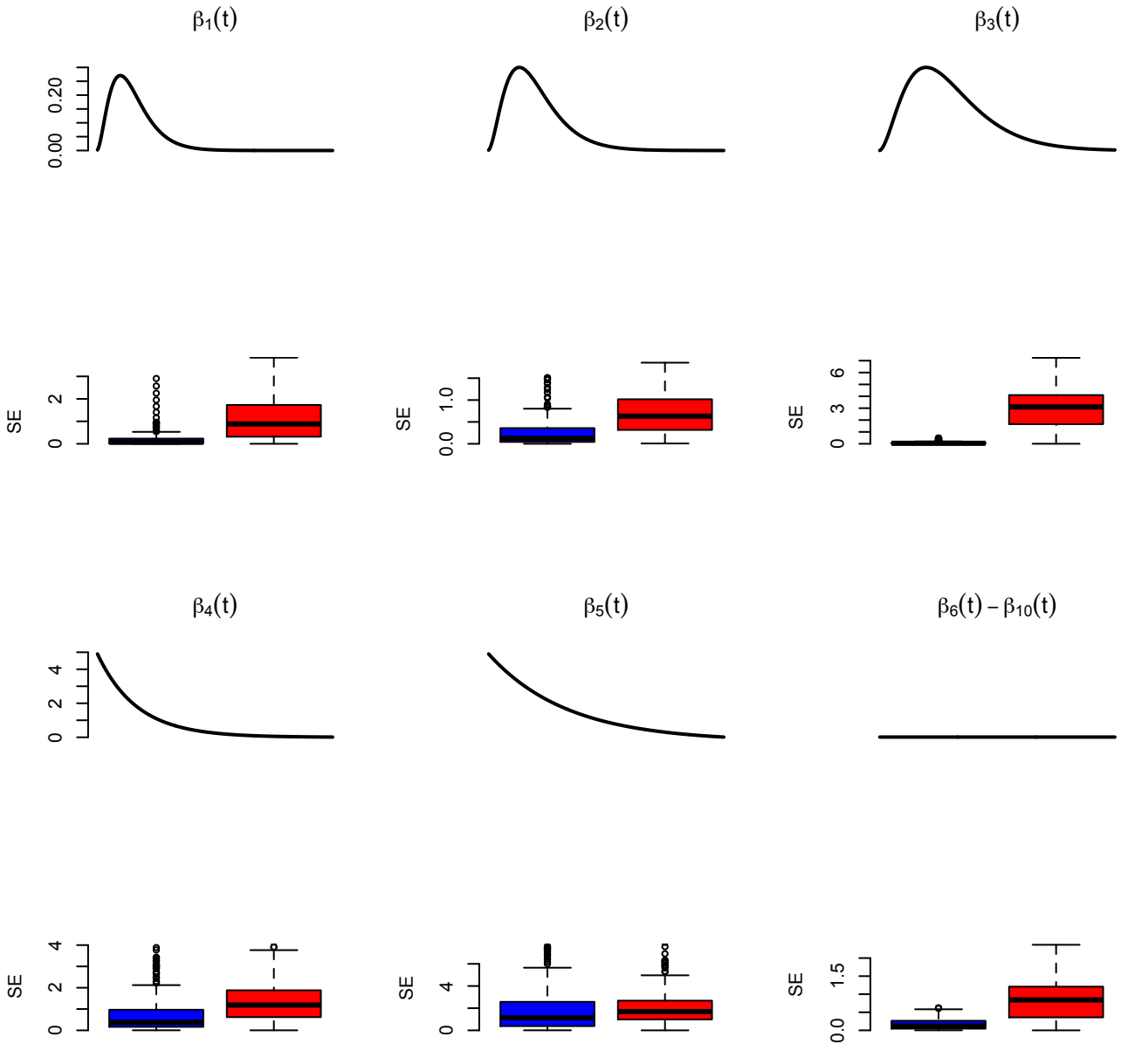


Figure 3.7: Comparison of SE of functional *LAD-gLASSO* (blue) and classical functional *gLASSO* (red) at 15% contamination for Model(1).

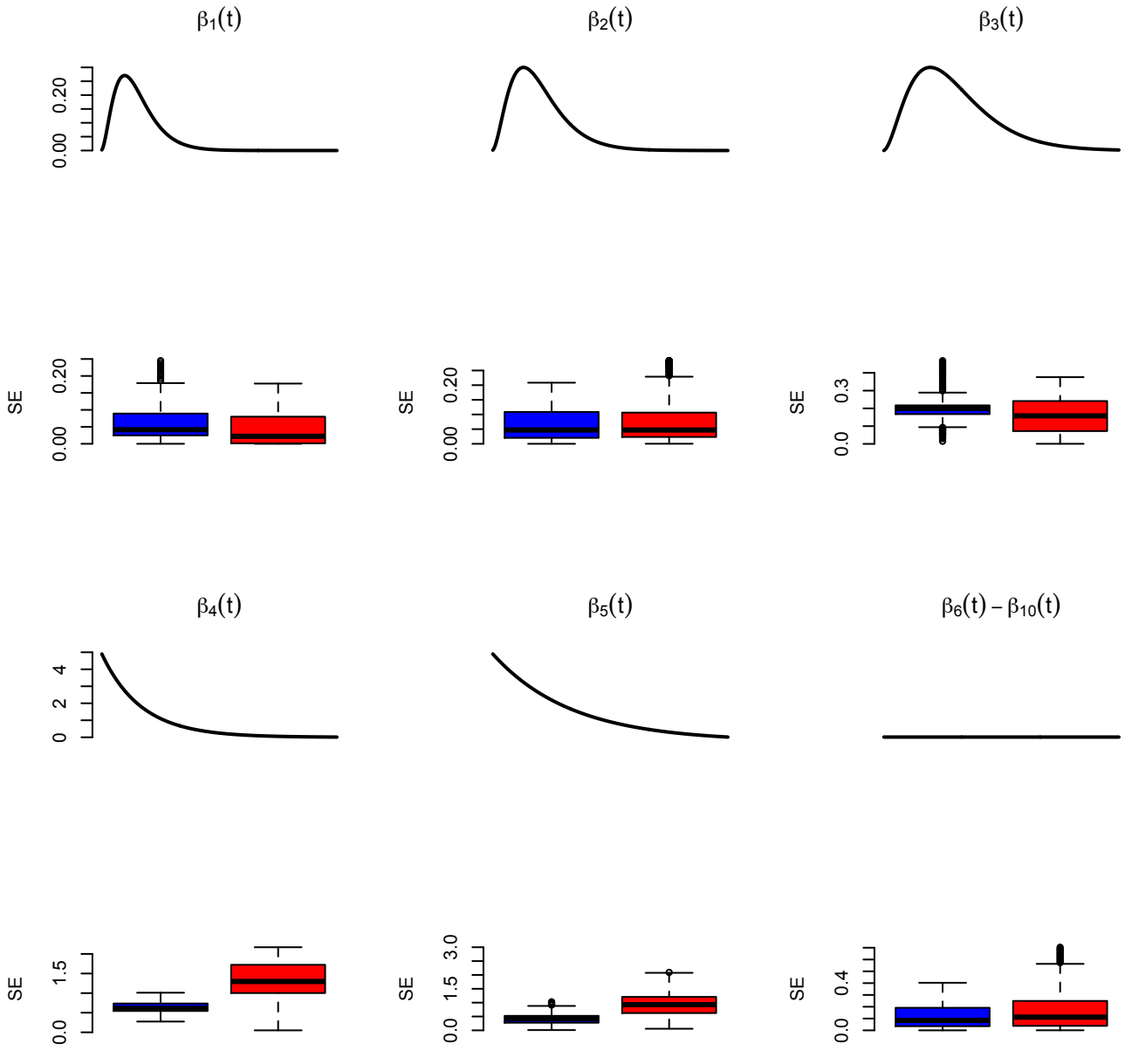


Figure 3.8: Comparison of SE of functional *LAD-gLASSO* (blue) and classical functional *gLASSO* (red) at 15% asymmetric contamination (Case 1) for Model(2).

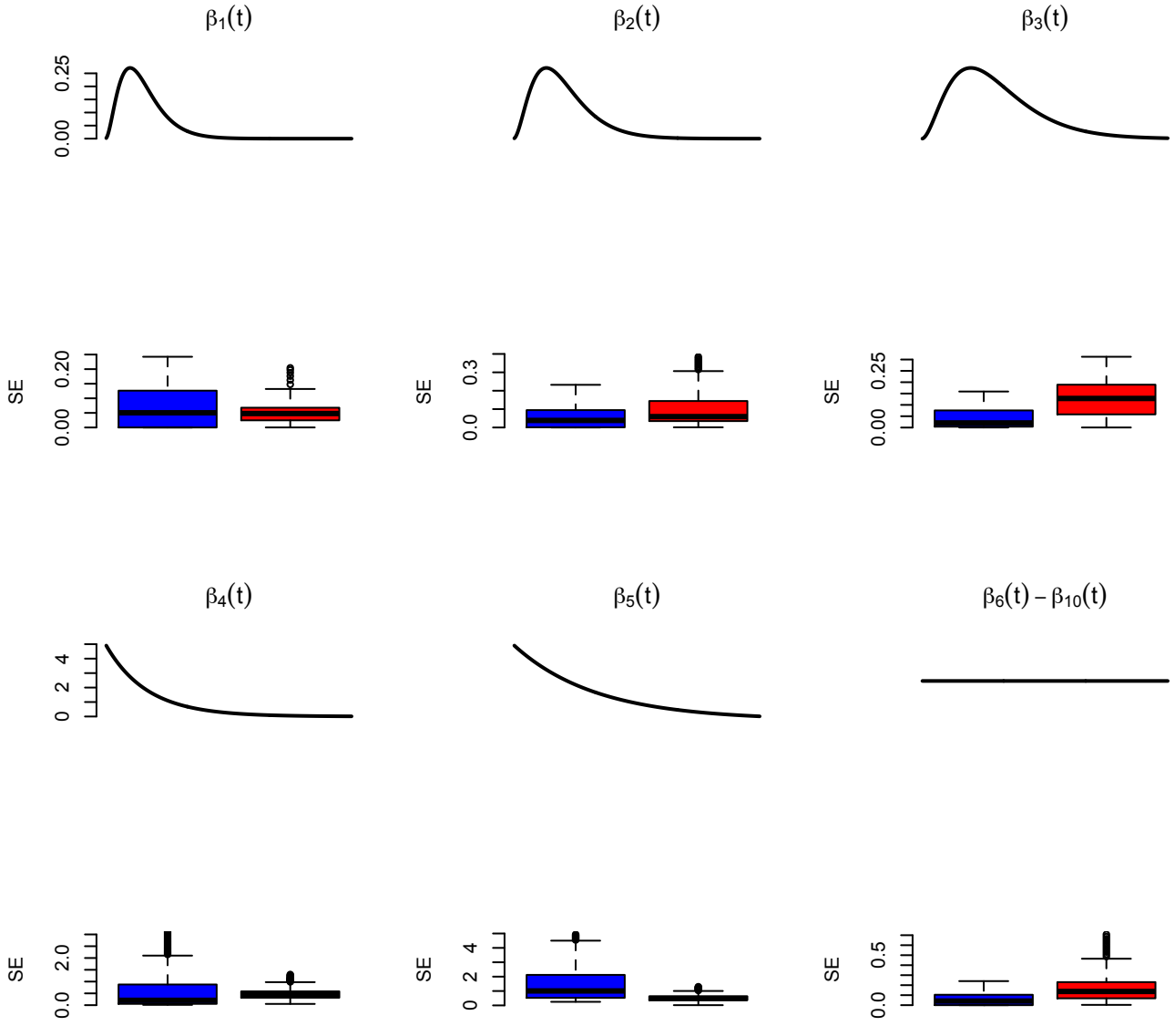


Figure 3.9: Comparison of SE of functional *LAD-gLASSO* (blue) and classical functional *gLASSO* (red) at 15% symmetric contamination (Case 2) for Model(2).

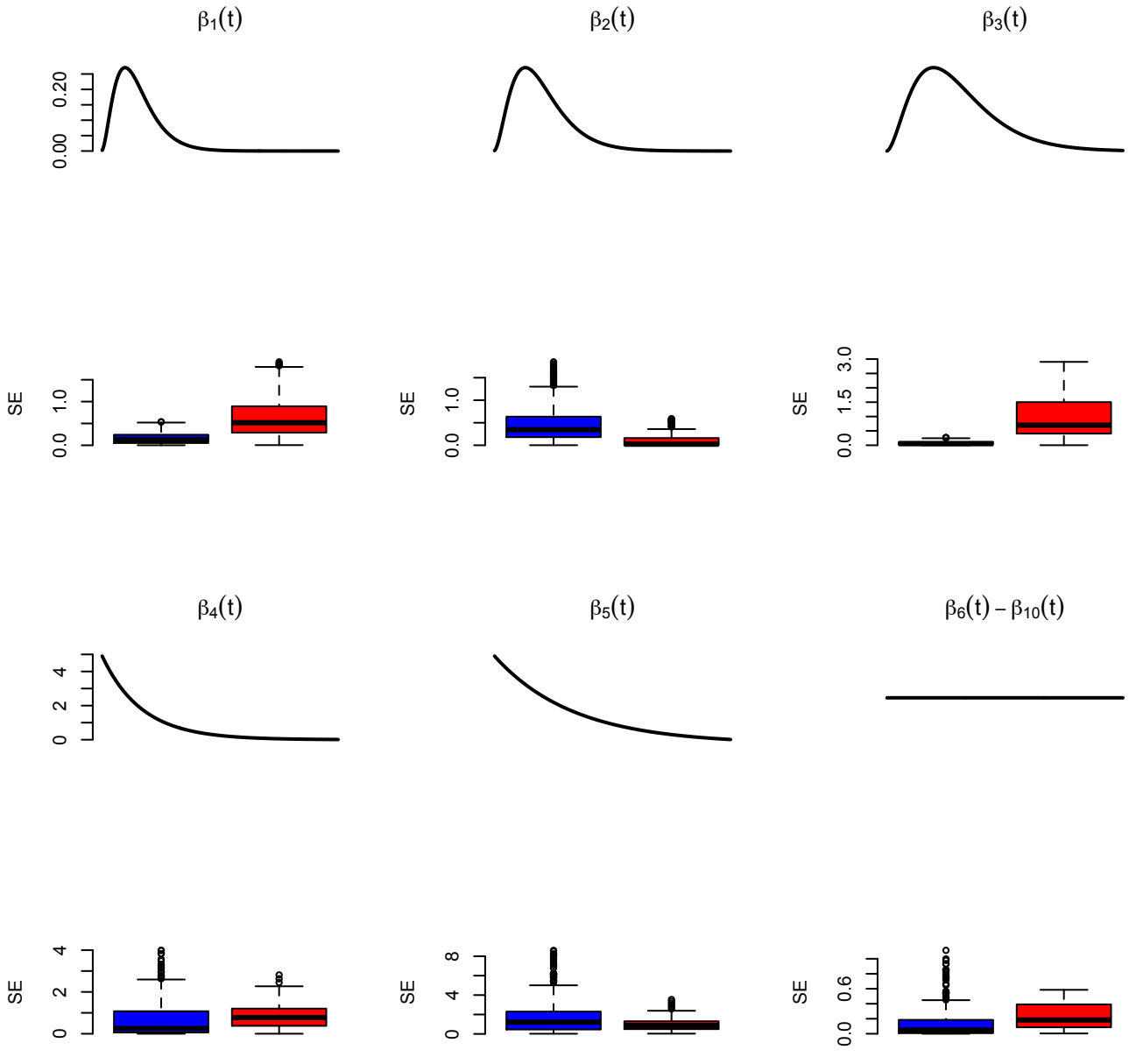


Figure 3.10: Comparison of SE of functional *LAD-gLASSO* (blue) and classical functional *gLASSO* (red) at 15% partial contamination (Case 3) for Model(2).

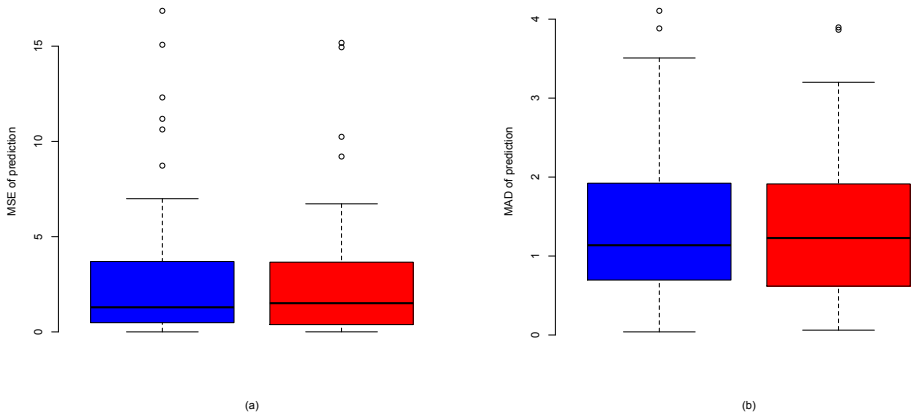


Figure 3.11: Comparison of MSE (Fig. (a)) and MAD (Fig. (b)) of functional *LAD-gLASSO* (blue) and classical functional *gLASSO* (red) at 0% contamination for Model(0).

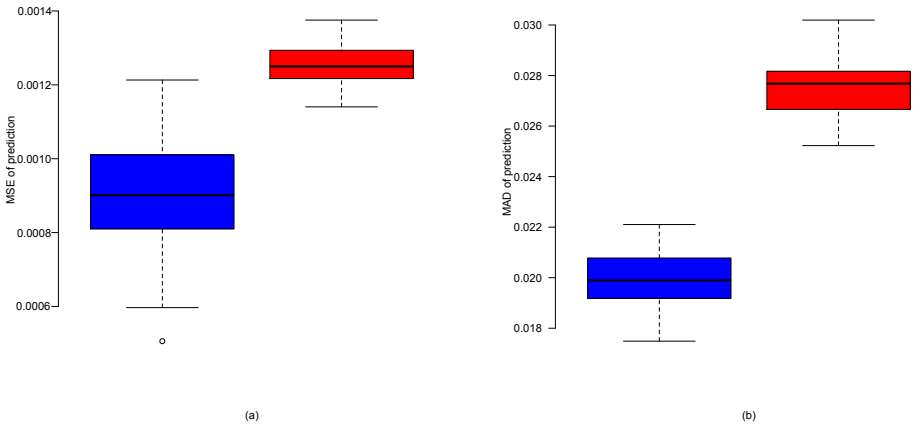


Figure 3.12: Comparison of MSE (Fig. (a)) and MAD (Fig. (b)) of functional *LAD-gLASSO* (blue) and classical functional *gLASSO* (red) at 15% contamination of Y for Model(1).

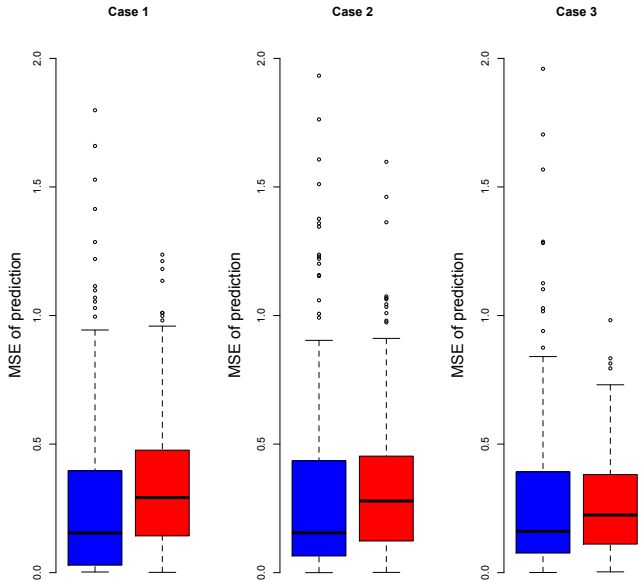


Figure 3.13: Comparison of MSE of functional *LAD-gLASSO* (blue) and classical functional *gLASSO* (red) at 15% contamination for Model(2).

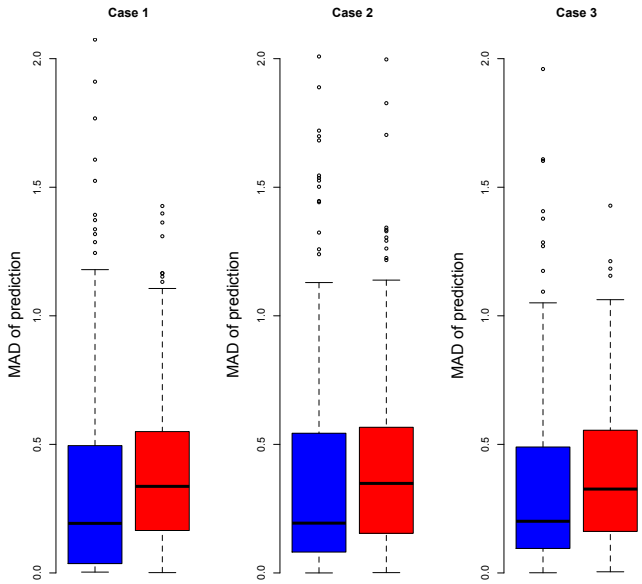


Figure 3.14: Comparison of MAD of functional *LAD-gLASSO* (blue) and classical functional *gLASSO* (red) at 15% contamination for Model(2).

Furthermore, we examined how many times each variable is selected in 50 independent simulation runs for Model (0), Model (1) and Model (2) using both functional *LAD- gLASSO* and classical functional *gLASSO*. For Model (2) setting we considered only Case 1 (Asymmetric contamination). Tables 3.1 and 3.2 show the proportion of simulation runs for which each predictor is selected using functional *LAD- gLASSO* and classical functional *gLASSO*, respectively. We see in Tables 3.1 and 3.2 that the true predictors $X_1(t) - X_5(t)$ are selected more frequently and predictors $X_6(t) - X_{10}(t)$ which are irrelevant to the true model are less frequently selected by the functional *LAD- gLASSO* compared to the classical functional *gLASSO* for Model (1) settings, that is, when the outliers are present in response variable only. That is, the percentage of true positives and true negatives is higher for functional *LAD- gLASSO* for Model (1) compared to classical functional *gLASSO*. We also note that functional *LAD- gLASSO* and classical functional *gLASSO* perform equally well for Model (0) and Model (2) settings. That is, both methods give same performance in both situations when the data are free of outliers and also when data has outliers in the functional predictors.

	$X_1(t)$	$X_2(t)$	$X_3(t)$	$X_4(t)$	$X_5(t)$	$X_6(t)$	$X_7(t)$	$X_8(t)$	$X_9(t)$	$X_{10}(t)$	$Avg.modelsize$
Model (0)	1	1	1	0.98	0.96	0.58	0.20	0.26	0.22	0.20	6.40
Model (1)	1	1	1	0.96	0.98	0.70	0.26	0.20	0.22	0.14	6.46
Model (2)	1	0.95	1	0.94	0.93	0.83	0.82	0.83	0.82	0.70	9.36

Table 3.1: Proportions of runs with respective functional predictor being selected and average model size using functional *LAD-gLASSO*.

	$X_1(t)$	$X_2(t)$	$X_3(t)$	$X_4(t)$	$X_5(t)$	$X_6(t)$	$X_7(t)$	$X_8(t)$	$X_9(t)$	$X_{10}(t)$	$Avg.modelsize$
Model (0)	1	1	1	0.92	0.94	0.50	0.24	0.30	0.32	0.26	6.48
Model (1)	1	1	1	0.90	0.94	0.98	0.72	0.80	0.76	0.82	8.92
Model (2)	1	0.84	1	1	0.86	0.96	0.94	0.98	0.92	0.90	9.40

Table 3.2: Proportions of runs with respective functional predictor being selected and average model size using classical functional *gLASSO*.

Furthermore, we consider 25% and 40% contamination levels for Model (1) and Model (2) settings for Toy example and simulation study presented above. We see that our proposed method functional *LAD- gLASSO* still performs better than classical functional *gLASSO* at 25%, but breaks down empirically at 40%. That is, functional *LAD- gLASSO* performs no better than classical functional *gLASSO* at contamination level of 40%.

Numerical Study for functional *LAD- agLASSO*

Next we perform numerical study for functional *LAD- agLASSO*. In the numerical study for functional *LAD- gLASSO*, we note that it does not perform well for Model (2) settings, that is, when both response and explanatory variables have outliers. Therefore, for the numerical study of functional *LAD- agLASSO* we consider only Model (1) settings, that is, when outliers are present in the response variable only. Also functional *LAD- agLASSO* is supposed to give better results compared to functional *LAD- gLASSO* as discussed in section 3.2.2.

We consider the following Model (1):

Model (1): Presence of outliers both in the scalar response Y only.

Data are generated as described in (3.18) and (3.19). Response Y is contaminated at 15 % level using the same method described previously. For this numerical study we consider four functional covariates $X_j(t)$. 100 curves for each these predictors are generated and each curve is observed at 50 equidistant points in (0, 50). The true model is as:

$$Y_i = \alpha + \int_0^{50} \beta_1(t) X_{i1}(t) dt + \int_0^{50} \beta_3(t) X_{i3}(t) dt + \epsilon_i \tag{3.25}$$

where $i = 1,\ldots,100$ and $\epsilon_i \sim N(0, 4)$.

The shapes of parameter functions are as shown in Figure 3.15. The model is set up where the response is related only to $X_1(t)$ and $X_3(t)$.

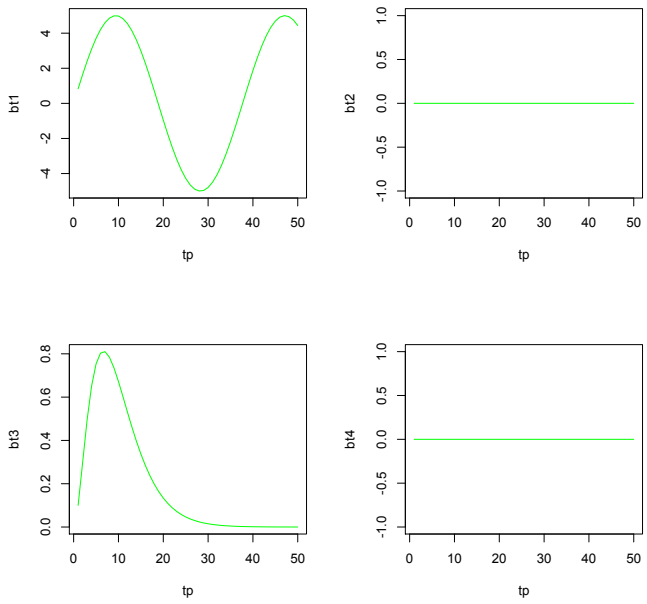


Figure 3.15: $\beta_1(t)$, $\beta_2(t)$, $\beta_3(t)$ and $\beta_4(t)$, respectively.

We apply all three versions Adapt 1, Adapt 2 and Adapt 3 of our second proposed method functional *LAD- agLASSO* (as described in section 3.2.2) to Model (1). Model (1) has outliers only in scalar response Y and the functional predictors $X_j(t)$s are free of outliers. Also the response Y depends only on the first predictor $X_1(t)$ and the third predictor $X_3(t)$. Second predictor $X_2(t)$ and fourth predictor $X_4(t)$ are irrelevant to the true model and should be excluded from the model by the applied method. We compare functional *LAD-agLASSO* with functional *LAD-gLASSO* and classical functional *agLASSO*. Classical functional *agLASSO* is discussed in section 2.2. Figures 3.16 - 3.18 show the comparison of the functional *LAD-agLASSO* (blue), the classical functional *agLASSO* (black) and functional *LAD-gLASSO* (red).

The initial estimates $\ddot{\beta}_j$ of the functional predictors that are required for adaptive *LASSO* are obtained using *pfr()* function in the *R* package *refund*. *R* package *quantreg* is employed again to execute our proposed method. The green solid curves in Figures 3.16 - 3.18 display the true functions $\beta_j(t)$. We can see in these figures that all three versions of functional *LAD-agLASSO* reduce the number of false positives compared to both functional *LAD-gLASSO*

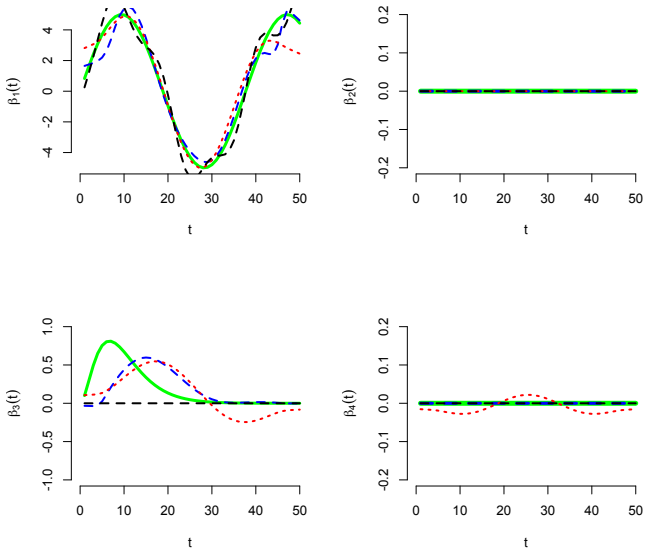


Figure 3.16: Fitting results of true beta functions (green) using classical functional *agLASSO* (black), functional *LAD-gLASSO* (red) and functional *LAD-agLASSO* (blue) (Adapt 1) at 15% contamination of Y for Model (1).

and classical functional *agLASSO*. Also the comparison of three versions of functional *LAD-agLASSO* indicates that Adapt 3 performs better than Adapt 1 and Adapt 2 in terms of estimation of true positives.

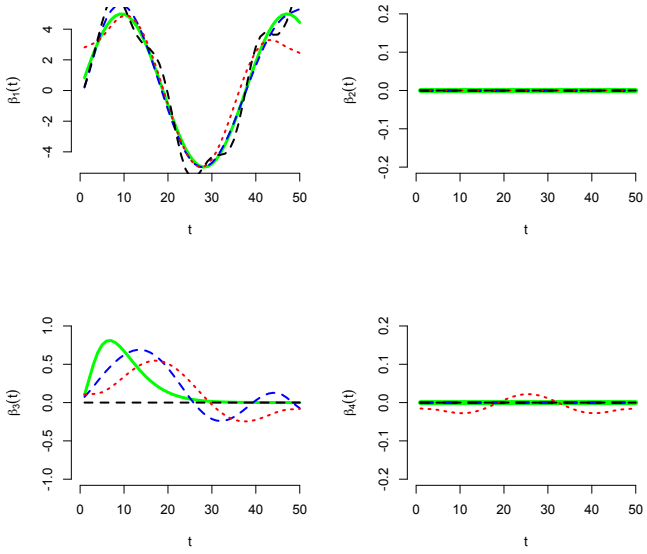


Figure 3.17: Fitting results of true beta functions (green) using classical functional *agLASSO* (black), functional *LAD-gLASSO* (red) and functional *LAD-agLASSO* (blue) (Adapt 2) at 15% contamination of Y for Model (1).

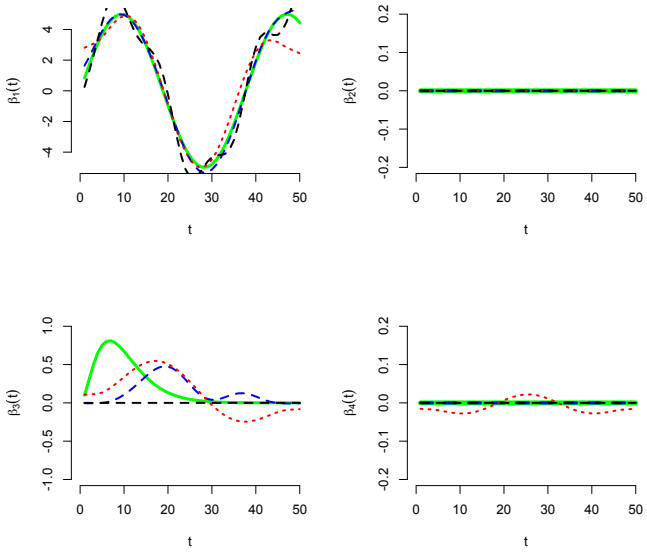


Figure 3.18: Fitting results of true beta functions (green) using classical functional *agLASSO* (black), functional *LAD-gLASSO* (red) and functional *LAD-agLASSO* (blue) (Adapt 3) at 15% contamination of Y for Model (1).

Simulation Study

Next we perform a simulation study to assess the performance of the functional *LAD-agLASSO*. This time we consider only Adapt 3, as it performs better than Adapt 1 and Adapt 2. We observe MSE and MAD of prediction along with SE in 50 independent simulation runs for Model (1). Figures (3.19), (3.20) and (3.21) show the boxplots of SE, MSE and MAD, respectively. In these figures blue boxplots correspond to functional *LAD-agLASSO*, red boxplots correspond to functional *LAD-gLASSO* and yellow boxplots correspond to classical functional *agLASSO*. All these figures reveal that functional *LAD-agLASSO* outperforms both functional *LAD-gLASSO* and classical functional *agLASSO*.

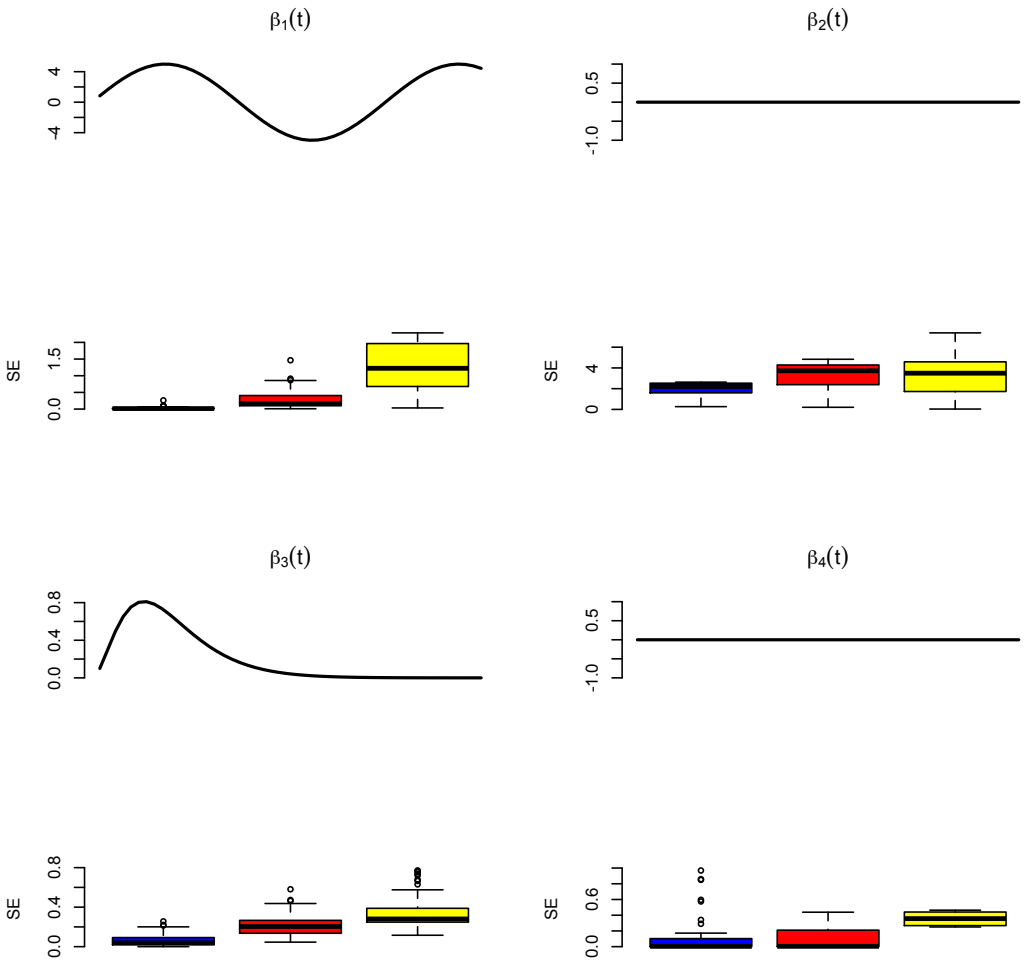


Figure 3.19: Comparison of SE of functional *LAD-agLASSO* (blue), functional *LAD-gLASSO* (red) and classical functional *agLASSO* (yellow) at 15% contamination for Model(1).

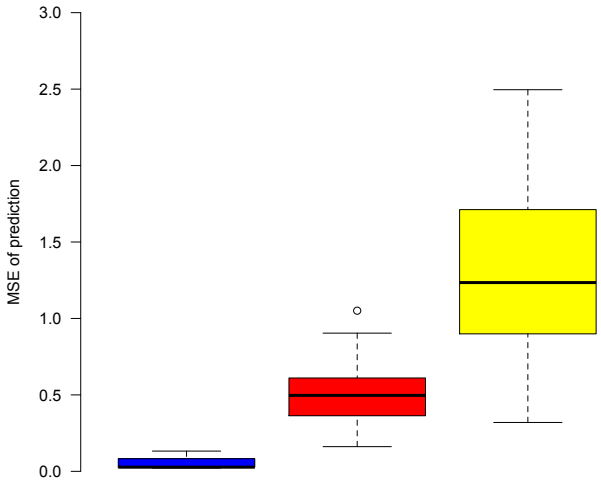


Figure 3.20: Comparison of MSE of prediction for functional *LAD-agLASSO* (blue), functional *LAD-gLASSO* (red) and classical functional *agLASSO* (yellow) at 15% contamination for Model(1).

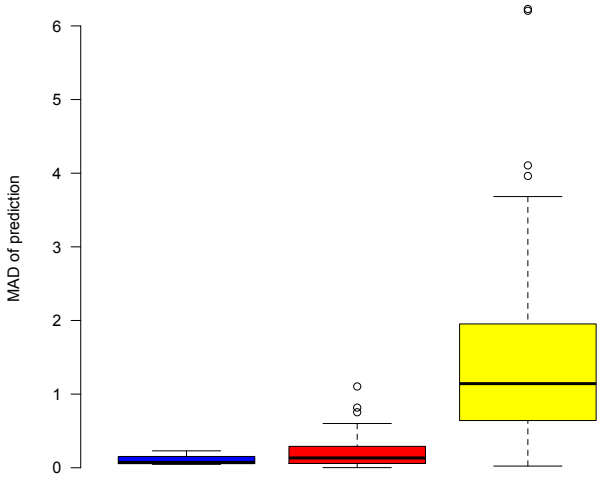


Figure 3.21: Comparison of MAD of prediction for functional *LAD-agLASSO* (blue), functional *LAD-gLASSO* (red) and classical functional *agLASSO* (yellow) at 15% contamination for Model(1).

Furthermore, Table 3.3 shows the proportions of 50 simulation runs with the respective functional predictor being selected and average model size using functional *LAD- agLASSO*, functional *LAD- gLASSO* and classical functional *gLASSO*. We see in Table 3.3 that the true predictors $X_1(t)$ and $X_3(t)$ are selected most frequently and predictors $X_2(t)$ and $X_4(t)$ which are irrelevant to the true model are less frequently selected by the functional *LAD- agLASSO* compared to functional *LAD- gLASSO* and classical functional *gLASSO*. To summarize, the percentage of false positives and false negatives reduces when functional *LAD- agLASSO* is used.

	$X_1(t)$	$X_2(t)$	$X_3(t)$	$X_4(t)$	Avg. Model Size
Functional *LAD-agLASSO*	1	0.36	1	0.28	2.64
Functional *LAD-gLASSO*	1	0.38	0.94	0.46	2.78
Classical functional *agLASSO*	1	0.66	0.86	0.78	3.30

Table 3.3: Proportions of runs with respective functional predictor being selected and average model size.

Additionally, we examine functional *LAD- agLASSO* at 25% and 40% contamination levels for Model (1) settings for simulation study presented above. We see that, this method also breaks down at 40% contamination level, but performs better than both functional *LAD-gLASSO* and classical functional *agLASSO* at contamination level of 25%.

3.4 Real Data Application

We apply our methods that are proposed in this chapter to the analysis of weather data used by Matsui & Konishi [22], available in Chronological Scientific Tables 2005, selecting variables concerning weather information. We use weather data observed at 79 stations in Japan. The data set includes monthly and annual total observations averaged from 1971 to 2000: monthly observed average temperatures (TEMP), average atmospheric pressure (PRESSURE), time of daylight (DAYLIGHT), average humidity (HUMIDITY) and annual total precipitation. The aim of the analysis is to select and estimate the variables that have

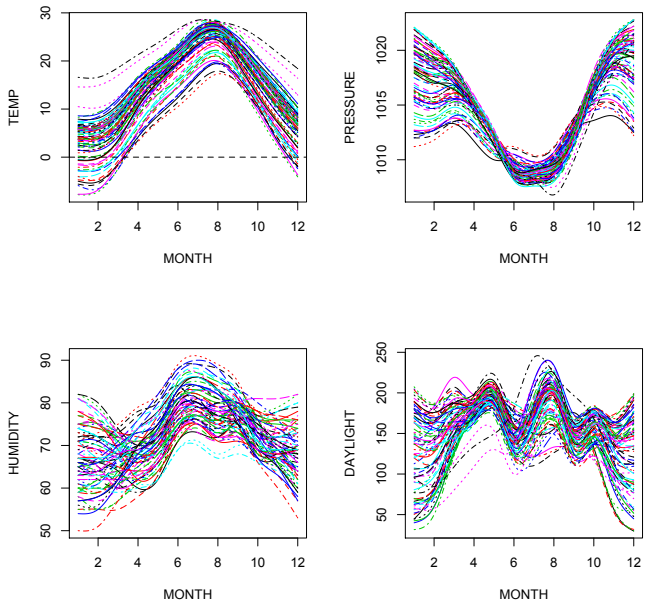
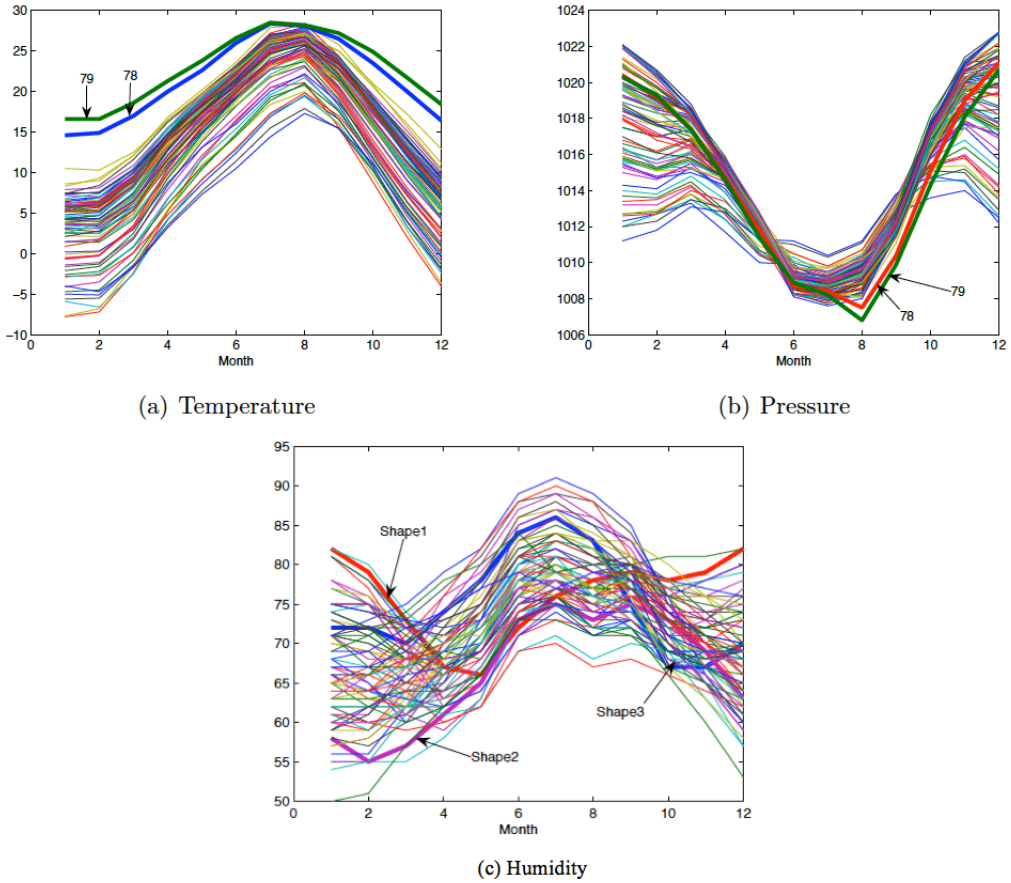


Figure 3.22: Weather Data.

a relationship with the response variable, annual total precipitation. Since the data are collected over time for 12 months, it can be treated as functional data. Figure 3.22 shows predictors in weather data, represented by functions observed at 12 points. In Figure 3.22, the group of curves shows presence of a few outliers, that is trajectories that are in some way different from the rest in the predictor variables. Specifically Figure 3.24 shows that curves 78 and 79 in both TEMP and PRESSURE variables and curves with shapes 1, 2 and 3 in the HUMIDITY variable are the outliers, as detected by Sawant [29] using robust functional principal component analysis. Figure 3.24 shows an outlier in the scalar response, annual total precipitation. To summarize this data have outliers in both functional predictors and the scalar response (annual total precipitation).

The response (annual total precipitation) is continuous. We use the functional linear model (3.1) with our proposed approaches to determine the most useful variables. But since our methods, which are proposed in this chapter, work better when there are no outliers in the predictors, so we consider removing outliers from the functional predictors before applying



(a) Temperature

(b) Pressure

(c) Humidity

Figure 3.23: Outliers in Weather Data.

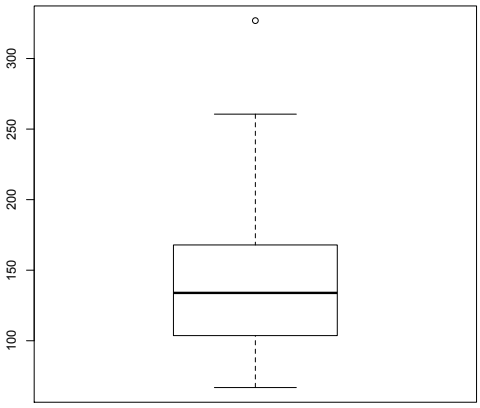


Figure 3.24: Outliers in response, annual total precipitation.

our proposed methodologies (functional *LAD-gLASSO* and functional *LAD- agLASSO*). After removing outliers from functional predictors, only the response variable is left with an outlier. Now we can safely use our proposed methods.

First we apply our proposed method functional *LAD- gLASSO* to the weather data. Figure 3.25 shows the estimated coefficient functions when using the functional *LAD- gLASSO* method. The PRESSURE variable is excluded from the model. According to these results, the PRESSURE variable does not seem to have a significant relationship with the precipitation. Secondly, we apply our proposed method functional *LAD- agLASSO* to the weather data. The fitting results are shown in Figure 3.26. The PRESSURE and DAYLIGHT are excluded from the model. The results indicate that there is no significant relationship between these variables and the precipitation. The remaining variables, TEMP and HUMIDITY, can be considered to relate significantly to the precipitation. We also apply classical functional group LASSO to this data set. The results are shown in Figure 3.27. It is clear from this figure that classical functional group LASSO is not able to exclude any variable(s) from the model, in the presence of outliers in the response variable.

Furthermore, we generate 50 bootstrap samples from the weather data. For each bootstrap sample, functional regression modeling is performed using functional *LAD- gLASSO*, functional *LAD-agLASSO* and classical functional *gLASSO*. We examine how many times each variable is selected. The results are shown in Table 3.4. These results reveal that functional *LAD- agLASSO* gives us the smallest model size and classical functional *gLASSO* gives the highest model size among three methods. Also the mean TEMP is selected most frequently among the four variables, followed by the HUMIDITY by functional *LAD-agLASSO*. This reveals significant relationships of these variables to the precipitation. On the other hand, the average PRESSURE and DAYLIGHT are less frequently selected by functional *LAD-agLASSO*. From the results, there seems to be less of a significant relationship between these variables and the precipitation.

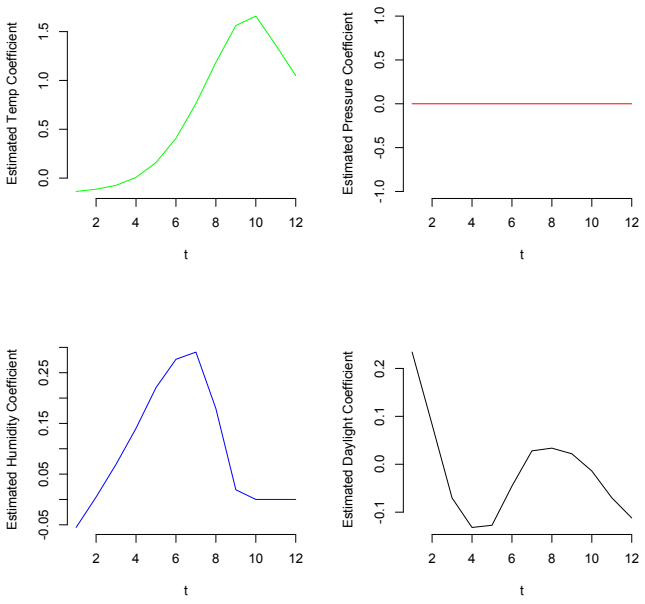


Figure 3.25: Estimated Variable Coefficients for Weather data using functional *LAD-gLASSO*.

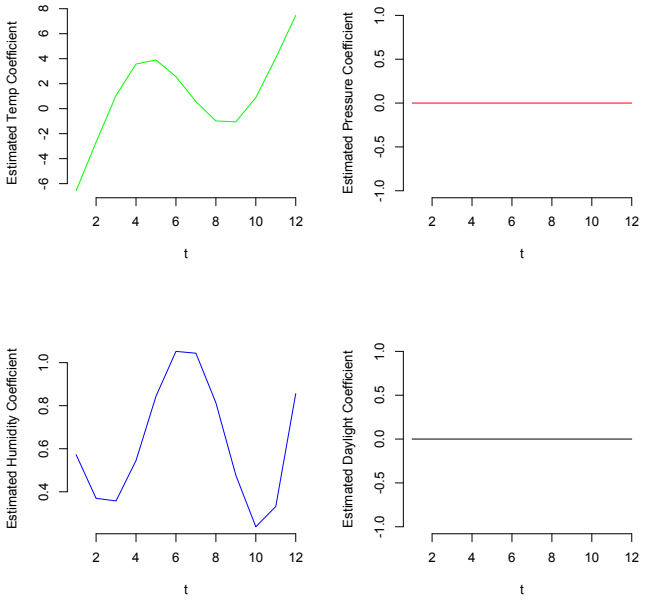


Figure 3.26: Estimated Variable Coefficients for Weather data using functional *LAD-agLASSO*.

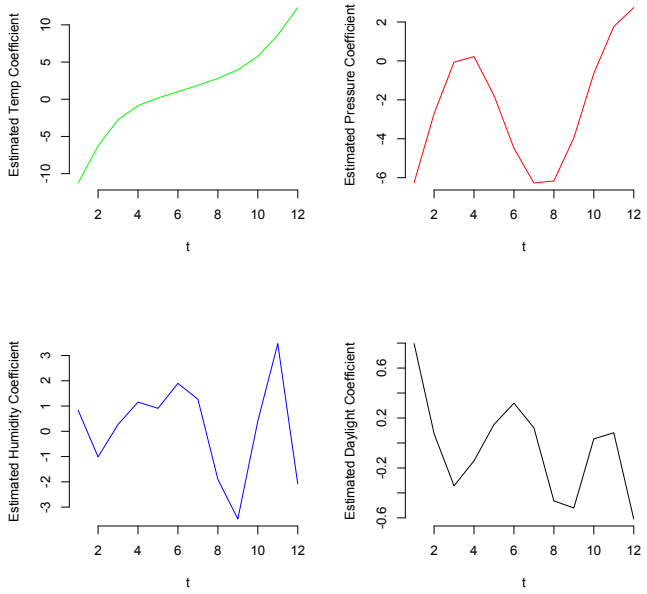


Figure 3.27: Estimated Variable Coefficients for Weather data using classical functional *gLASSO*.

	TEMP	PRESSURE	HUMIDITY	DAYLIGHT	Avg. Model Size
Functional *LAD-gLASSO*	1	0.58	0.94	0.92	3.44
Functional *LAD-agLASSO*	1	0.46	0.98	0.38	2.82
Classical functional *gLASSO*	0.90	1	0.96	0.98	3.84

Table 3.4: Proportions of runs with the respective functional predictor being selected and average model size.

3.5 Summary and Discussion

We considered two robust variable selection procedures for functional linear regression models in the presence of outliers, where various functional predictors are considered but only a few of these are actually related to the scalar response. Typical variable selection procedures for functional models do not consider the issue of outliers while selecting the useful predictors, and thus may suffer from wrong models. Our proposed procedures simultaneously select and estimate the important functional variables.

We found that our proposed methods perform well in terms of prediction error as well as mean squared errors for the estimated coefficient functions compared to classically fitting a

model without taking outliers into consideration. The false positive and false negative rates are also quite low for our methods. We also noted that our proposed method functional *LAD- agLASSO* performs better than functional *LAD- gLASSO* and among three versions of functional *LAD- agLASSO*, Adapt 3 performs the best. We also notice that our proposed methods still perform better than classical methods at 25% contamination level, but break down empirically at contamination level of 40%. That is, functional *LAD- gLASSO* and functional *LAD- agLASSO* perform no better than classical functional *gLASSO* and classical functional *agLASSO*, respectively at contamination level of 40%.

Furthermore, our proposed methods do not work perform better when there are outliers only in response variable compared to when there are outliers in both response and functional predictors.

In the following chapter, we propose two methodologies *WLAD- gLASSO* and *WLAD- ag LASSO*, that take into account the effect of outliers in functional predictors to overcome the limitations of our methods proposed in this chapter.

Chapter 4

Robust Group Variable Selection Methods for Multiple Functional Regression Model in the Presence of Outliers in the Response and Explanatory Variables

4.1 Introduction

In Chapter 3 we considered the problem of variable selection for functional regression model in the presence of outliers. We developed two robust functional variable selection techniques called functional *LAD- gLASSO* and functional *LAD- agLASSO*, which perform better than classical variable selection method, functional *gLASSO*.

But these robust methods give better results when outliers are present in y direction only. These robust techniques being based on simple *LAD* are highly sensitive to outliers in the x direction, therefore necessitating a different type of approach to handle this issue. Weighted *LAD* regression estimation has been proposed by Ellis and Morgenthaler [6], Hubert and Rousseeuw [17], Giloni et al. [13] and Giloni et al. [14], to deal with outliers in predictors for ordinary multiple regression model. Recently, Arslan [1] has proposed Weighted $LAD-LASSO$ (*WLAD-LASSO*) as a robust variable selection method to handle the issue of outliers in response and explanatory variables for ordinary multiple regression model. But to our knowledge no such method exists for functional regression model. Therefore, in this chapter, we consider a new criterion called functional *Weighted LAD- group LASSO*, abbreviated as *WLAD- gLASSO* that takes into account the effect of outliers in both y and x direction for functional regression model with a scalar response and functional predictors. It is a weighted version of functional *LAD- gLASSO*. This method is not only resistant to outliers in the response variable but also minimizes the effect of outliers in explanatory variable (leverage points), by introducing weights which are dependent on the explanatory variables only. These weights are introduced to downweight the leverage points and thus reducing

their effect on the estimation process. We also provide an adaptive version of this method in which adaptive *LASSO* penalty criterion is used to assign different weights to different coefficients to penalize them differently. We call this method functional *Weighted LAD-Adaptive group LASSO* abbreviated as (*WLAD- agLASSO*).

4.2 Methodology

Reconsider a functional linear model with the scalar response and p-functional predictors from (3.1):

$$Y_i = \alpha + \Sigma_{j=1}^{p} \int_{\mathcal{T}_I} X_{ij}(t)\beta_j(t)dt + \epsilon_i, \quad i = 1, \ldots, N. \tag{4.1}$$

where, the random error terms ϵ_i are assumed to be independent normally distributed with mean 0 and variance σ^2. α is a scalar parameter and $\beta_j(t)$ is a parameter function for $j = 1, \ldots, p$.

We apply the same method described in section 3.2 to the model in (4.1), to overcome the inherent infinite dimensionality problem and reformulate it as an ordinary multiple regression model. This gives us the same model in (3.5):

$$Y_i = \alpha + \Sigma_{j=1}^{p} \mathbf{\Phi_{ij}^{T}} \mathbf{c_j} + \epsilon_i, \quad i = 1, \ldots, N. \tag{4.2}$$

where $\mathbf{\Phi_{ij}}$ are known and α and $\mathbf{c_j}$'s are the unknown regression coefficients that need to be estimated.

Next we propose our method *WLAD- gLASSO* by modifying our previously proposed method functional *LAD- gLASSO* in Chapter 3.

4.2.1 Functional WLAD- groupLASSO

Reconsider the objective function for *LAD- gLASSO* in (3.8)

$$\Sigma_{i=1}^{n} |Y_i - \alpha - \Sigma_{j=1}^{p} \mathbf{\Phi_{ij}}^{T} \mathbf{c_j}| + P_{\lambda,\varphi}(\beta_j). \tag{4.3}$$

where, $P_{\lambda,\varphi}(\beta_j)$ is the penalty function as introduced by Meier et al. [23].

For *WLAD- gLASSO* criterion we introduce weights w_i to the function in (4.3), which are determined by a robust measure of predictors and are chosen to downweight the leverage points. Then the objective function for *WLAD- gLASSO* is given by:

$$\Sigma_{i=1}^{n} w_i |Y_i - \alpha - \Sigma_{j=1}^{p} \mathbf{\Phi_{ij}}^T \mathbf{c_j}| + P_{\lambda,\varphi}(\beta_j). \tag{4.4}$$

The penalty function $P_{\lambda,\varphi}(\beta_j)$ is also modified using the same method described in section 3.2.1, which reduces the objective function in (4.4) for *WLAD- gLASSO* criterion to the following:

$$\Sigma_{i=1}^{n} w_i |Y_i - \alpha - \tilde{\Phi}_{ij}^T \tilde{\mathbf{c}}_\mathbf{j}| + \lambda \Sigma_{j=1}^{p} ||\tilde{\mathbf{c}}_\mathbf{j}||. \tag{4.5}$$

Now $\hat{\alpha}$ and $\hat{\mathbf{c}}_\mathbf{j}$'s are the minimizers of (4.5). The tuning parameters λ and φ are chosen via K fold cross-validation as described in section 3.2.3. The weights w_i in (4.5) are obtained using the robust distances of the predictors so that the outlying observations in the x direction will have large distances and the corresponding weights will be small. Therefore, it is expected that the resulting regression estimator will be robust against the outliers in the response variable and leverage points.

The weights are computed using the weight definition given in Hubert and Rousseeuw [17]. Specifically the algorithm to find the weights is as following:

1. Calculate the robust location and scatter estimates, $\tilde{\mu}$ and $\tilde{\Sigma}$ for the location vector and the scatter matrix of the data $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n \in \Re^p$. One can use high breakdown point location and scatter estimators such as MCD (Minimum Covariance Determinant). The idea behind MCD is to find observations whose empirical covariance matrix has the smallest determinant, yielding a pure subset of observations from which to compute standards estimates of location and covariance. The Minimum Covariance Determinant estimator (MCD) has been introduced by Rousseeuw in [27]. The implementation in

R package *rrcov* uses the Fast MCD algorithm of Rousseeuw and Driessen [28] to approximate the minimum covariance determinant estimator.

2. Compute the robust distances: $RD(\boldsymbol{x}_i) = (\boldsymbol{x}_i - \tilde{\mu})^T \tilde{\Sigma}^{-1} (\boldsymbol{x}_i - \tilde{\mu})$.

3. Calculate the weights $w_i = min\left\{1, \frac{p}{RD(\boldsymbol{x}_i)}\right\}$ for $i = 1, \ldots, n$.

4.2.2 Functional WLAD- Adaptive groupLASSO

In this section we consider an adaptive penalty function for functional *WLAD- gLASSO* to allow for different shrinkage and smoothness for the different covariates. The penalty in (4.4) penalizes the coefficient functions by the same amount but the adaptive version of this penalty will reflect some subjectivity about the true parameter functions. This criterion that incorporates adaptive penalty is called functional *WLAD- agLASSO*.
For this, reconsider equation (4.4):

$$\Sigma_{i=1}^{n} w_i |Y_i - \alpha - \Sigma_{j=1}^{p} \mathbf{\Phi_{ij}}^T \mathbf{c_j}| + P_{\lambda,\varphi}(\beta_j). \tag{4.6}$$

Here $P_{\lambda,\varphi}(\beta_j$ is the same adaptive $LASSO$ penalty given in (3.14), that is,

$$P_{\lambda,\varphi}(\beta_j) = \lambda(\kappa_j ||\beta_j||_2^2 + \nu_j \varphi ||\beta_j''||_2^2)^{1/2}. \tag{4.7}$$

where $||.||^2 = \int (.)^2 dt$ is the L^2 norm, β_j'' is the second derivative of β_j, κ_j and ν_j are the data adaptive weights. The weights κ_j and ν_j are calculated the same way as described in section 3.2.2. Hence, the objective function minimized by functional *WLAD- agLASSO* is:

$$\Sigma_{i=1}^{n} w_i |Y_i - \alpha - \Sigma_{j=1}^{p} \mathbf{\Phi_{ij}}^T \mathbf{c_j}| + \lambda(\kappa_j ||\beta_j||_2^2 + \nu_j \varphi ||\beta_j''||_2^2)^{1/2}. \tag{4.8}$$

Next, in order to show the optimality of the proposed methods we perform a numerical study in which a simulation study is conducted. Furthermore, we show a real data application of the proposed methods.

4.3 Numerical Study

In this section we provide the numerical performances of the methods proposed in this chapter. We consider the following model, Model (2) from section 3.3, for the numerical studies:

- **Model (2)**: Presence of outliers in both scalar response Y and functional predictors $X(t)$.

First we present the numerical study for our proposed method functional *WLAD- gLASSO* and then for functional *WLAD- agLASSO*.

4.3.1 Numerical Study for functional WLAD- gLASSO

For the numerical study of functional *WLAD- gLASSO* we generate two functional covariates $X_1(t)$ and $X_2(t)$ from (3.18). 100 replications of each of $X_j(t)$ are observed at 50 equidistant time points in (0, 50). Response Y is generated for 100 functional curves from (3.19). The true model is as:

$$Y_i = \alpha + \int_0^{50} \beta_1(t) X_{i1}(t) dt + \epsilon_i. \tag{4.9}$$

where $i = 1, \ldots, 100$ and $\epsilon_i \sim$ N(0,4).

The parameter function $\beta_1(t)$ has an exponential function shape and the parameter function $\beta_2(t)$ is essentially zero, as shown in Figure 4.1. The model is set up where the response is related only to $X_1(t)$. Both response Y and functional predictors $X_j(t)$ are contaminated at 15 % level using the same method described in section 3.3. All three cases of contamination, Case 1(Asymmetric contamination), Case 2 (Symmetric contamination) and Case 3 (Partial contamination) of functional predictors $X_j(t)$ are taken into account. We apply functional *WLAD-gLASSO* to Model (0), Model (1) and Model (2) setting and compare it with functional *LAD-gLASSO*, proposed in Chapter 3 and with classical functional *gLASSO*.

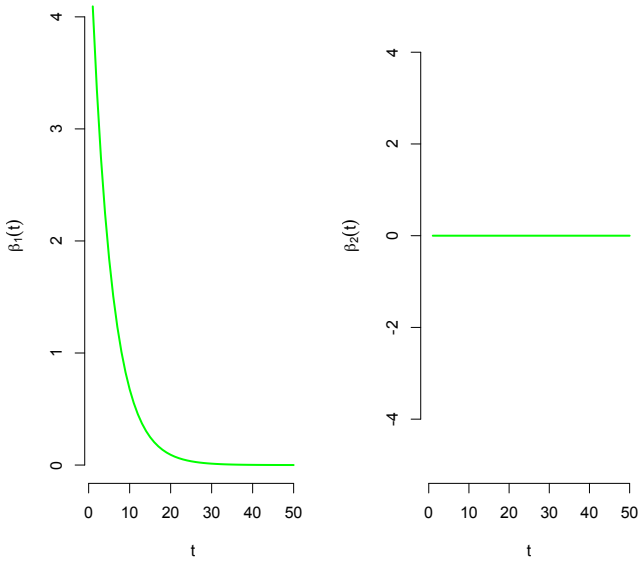


Figure 4.1: $\beta_1(t)$ and $\beta_2(t)$, respectively.

The results are as following.

Model (0): No outliers in the scalar response Y and the functional predictors $X(t)$.

First we apply our proposed method functional *WLAD-gLASSO* to Model (0) and compare it with functional *LAD-gLASSO* and classical functional *gLASSO*. Model (0) has neither outliers in scalar response Y nor in the functional predictors $X_1(t)$ and $X_2(t)$. The response Y is dependent only on the first predictor $X_1(t)$. Figure 4.2 shows the fitting results of functional *WLAD-gLASSO*, functional *LAD-gLASSO* and classical functional *gLASSO* method. We use functions *rq.fit.lasso ()* and *CovMcd ()* from *R* packages *quantreg* and *rrcov*, respectively to execute our proposed method, functional *WLAD-gLASSO*. *R* package *quantreg* is also used for functional *LAD-gLASSO*. For classical functional *gLASSO*, we use *R* package *grplasso*. In Figure 4.2, the green curves display the true functions $\beta_1(t)$ and $\beta_2(t)$; the purple, blue and red dashed lines display the estimations done by functional *WLAD-gLASSO*, functional *LAD-gLASSO* and classical functional *gLASSO*, respectively. The combinations

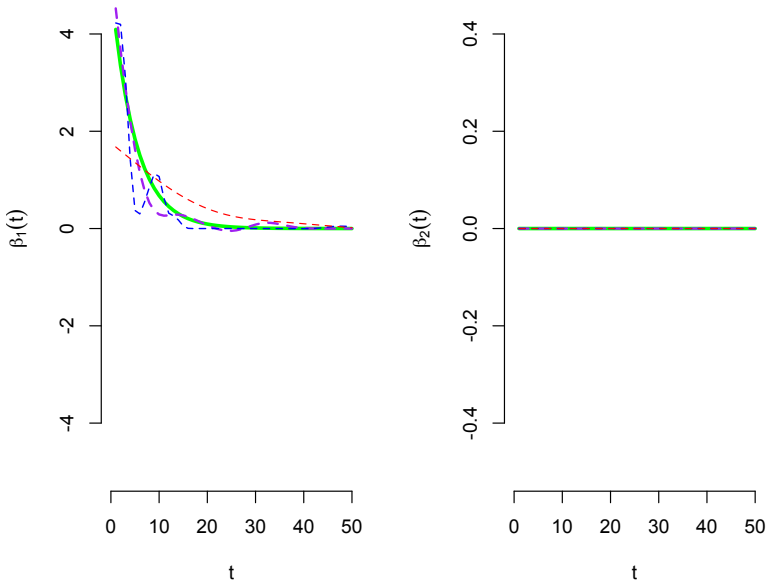


Figure 4.2: Fitting results for the comparison of functional *WLAD-gLASSO* (purple), functional *LAD-gLASSO* (blue) and classical functional *gLASSO* (red) for Model (0) (0% contamination).

of λ and φ for functional *WLAD-gLASSO*, functional *LAD-gLASSO* and the classical functional *gLASSO* are ($\lambda = 1$, $\varphi = 10$), ($\lambda = 10$, $\varphi = 10$) and ($\lambda = 10$, $\varphi = 100$), respectively. We can see in Figure 4.2 that all methods estimate the relevant coefficient $\beta_1(t)$ close to its true value and exclude the irrelevant coefficient $\beta_2(t)$ from the model.

Model (1): Presence of outliers in the scalar response Y only.

Secondly, we apply our proposed method to the Model (1). Model (1) has outliers only in scalar response Y. The functional predictors $X_1(t)$ and $X_2(t)$ are free of outliers. Also the response Y depends only the first predictor $X_1(t)$ and not on $X_2(t)$. Since $X_2(t)$ is irrelevant to the true model, so it should be excluded from the model by the applied method. Figure 4.3 shows the comparison of functional *WLAD-gLASSO*, functional *LAD-gLASSO* and classical functional *gLASSO* method. We use functions *rq.fit.lasso ()* and *CovMcd ()*

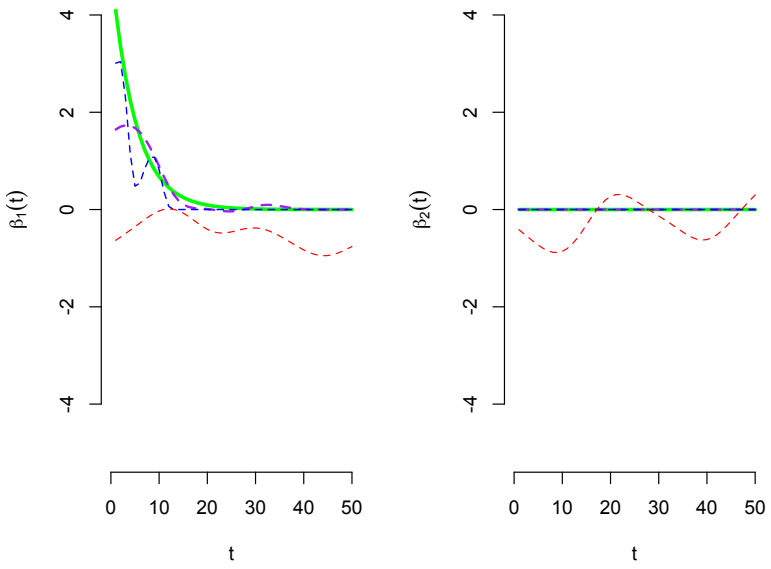


Figure 4.3: Fitting results for the comparison of functional *WLAD-gLASSO* (purple), functional *LAD-gLASSO* (blue) and classical functional *gLASSO* (red) for Model (1) (15% contamination).

from R packages *quantreg* and *rrcov*, respectively to execute functional *WLAD-gLASSO*. R package *quantreg* is used for functional *LAD-gLASSO*. For classical functional *gLASSO*, we use R package *grplasso*. In Figure 4.3, the green curves display the true functions $\beta_1(t)$ and $\beta_2(t)$; the purple, blue and red dashed lines display the estimations done by functional *WLAD-gLASSO*, functional *LAD-gLASSO* and classical functional *gLASSO*, respectively. The combinations of λ and φ for functional *WLAD-gLASSO*, functional *LAD-gLASSO* and the classical functional *gLASSO* are ($\lambda = 10$, $\varphi = 10$), ($\lambda = 1$, $\varphi = 10$) and ($\lambda = 10$, $\varphi = 100$), respectively. Figure 4.3 shows that functional *WLAD-gLASSO* and functional *LAD-gLASSO* not only exclude the irrelevant predictor $X_2(t)$ from the model, but also estimate relevant coefficient $\beta_1(t)$ close to its true value, compared to the classical functional *gLASSO*.

Model (2): Presence of outliers in both scalar response Y and functional predictors $X(t)$.

Finally, we apply our proposed method functional *WLAD- gLASSO* to Model (2). Model (2) has outliers both in scalar response Y and the functional predictors $X_1(t)$ and $X_2(t)$. All three cases of contamination, Case 1(Asymmetric contamination), Case 2 (Symmetric contamination) and Case 3 (Partial contamination) are considered for functional covariates. Also the first covariate $X_1(t)$ being relevant to the true model should be kept and $X_2(t)$ being irrelevant should be excluded from the model, by the applied method. Figure 4.4 shows the fitting results of functional *WLAD- gLASSO*, functional *LAD-gLASSO* and classical functional *gLASSO*. We use functions *rq.fit.lasso ()* and *CovMcd ()* from R packages *quantreg* and *rrcov*, respectively to execute our proposed method, functional *WLAD-gLASSO*. R package *quantreg* is also used for functional *LAD-gLASSO*. For classical functional *gLASSO*, we use R package *grplasso*. In Figure 4.4, the green curves are the true coefficient functions $\beta_1(t)$ and $\beta_2(t)$, the purple lines represent the estimation done by functional *WLAD- gLASSO*, the blue lines represent the estimation done by functional *LAD- gLASSO* and the red lines represent the estimation done by classical functional *gLASSO*. We can see in Figure 4.4 that the functional *WLAD- gLASSO* excludes the irrelevant predictor $X_2(t)$ from the estimated model, and estimates relevant predictor $X_1(t)$ close to its true value at fixed combinations of $(\lambda = 10, \varphi = 10^2), (\lambda = 1, \varphi = 10)$ and $(\lambda = 10, \varphi = 10)$ for the three cases of contamination, Case 1(Asymmetric contamination), Case 2 (Symmetric contamination) and Case 3 (Partial contamination), respectively. In contrast to functional *WLAD- gLASSO*, both classical functional *gLASSO* and functional *LAD- gLASSO* perform poorly in both variable selection and estimation.

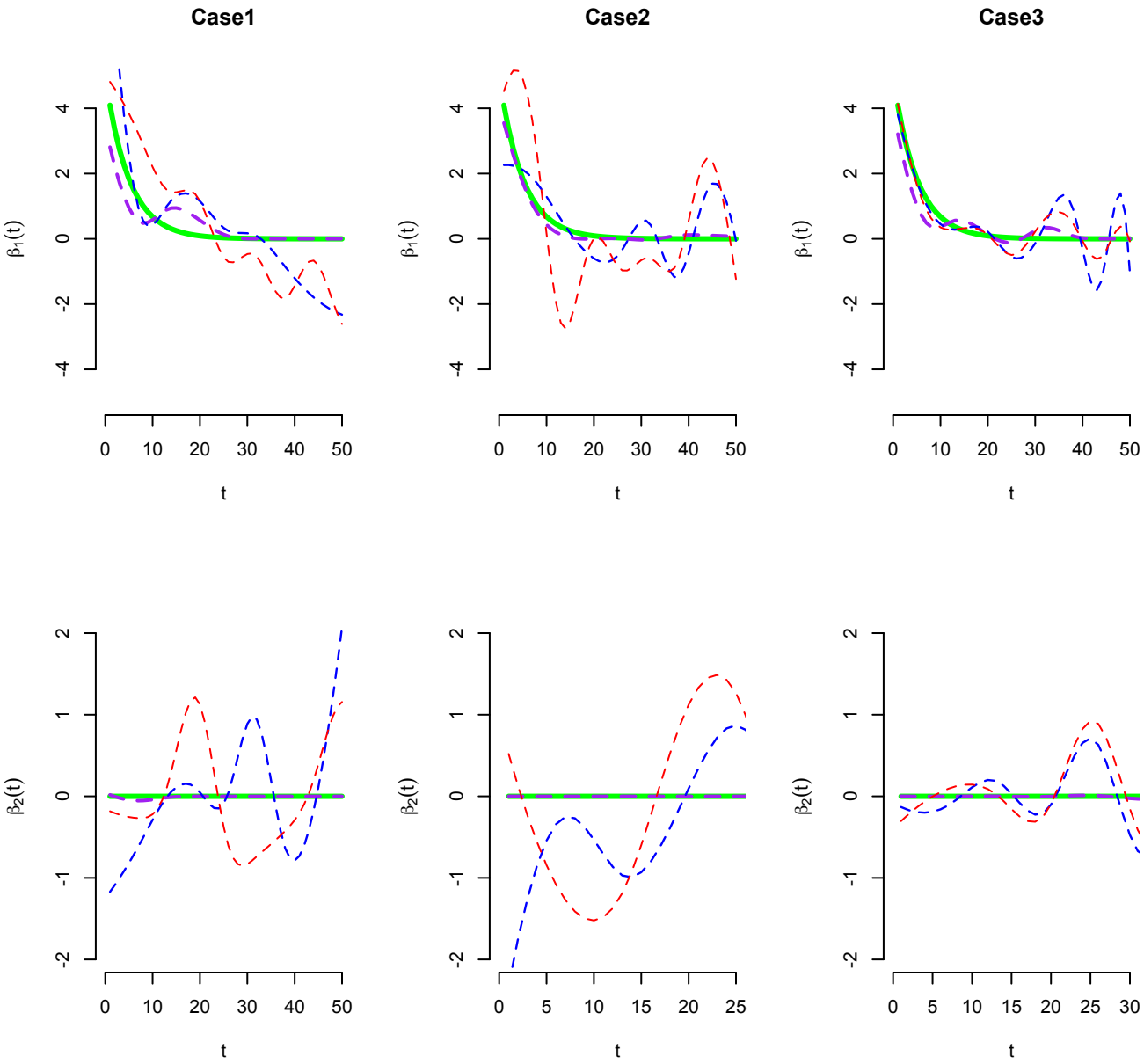


Figure 4.4: Fitting results for the comparison of functional *WLAD-gLASSO* (purple), functional *LAD-gLASSO* (blue) and classical functional *gLASSO* (red) for Model (2) (15% contamination).

Simulation Study

Next we conduct a simulation study to assess the performance of functional *WLAD- gLASSO*. The results of this simulation study reveals better performance of functional *WLAD- gLASSO* compared to functional *LAD- gLASSO* and classical functional *gLASSO*. For the simulation study the data are generated from (3.18) and (3.19). Specifically, we consider the following:
1) 1000 observations for the scalar response Y.
2) Two functional predictors are considered. We generate 1000 sample curves for each $X_j(t)$ which are observed at 300 equidistant time points in (0, 50).
3) The true model is

$$Y_i = \alpha + \int_0^{50} \beta_1(t) X_{ij}(t) dt + \epsilon_i. \tag{4.10}$$

where, $i = 1, \ldots, 1000$ and $\epsilon_i \sim$ N(0,4). The parameter function $\beta_1(t)$ is observed at 300 points in (0, 50). The shapes of $\beta_j(t)$ are as shown in Figure 4.3. We can see in Figure 4.5 that $\beta_2(t)$ is essentially 0. The true model in (4.6) depends only on $\beta_1(t)$.

In simulation study, we compare the performance of the proposed method functional *WLAD-gLASSO* with functional *LAD-gLASSO* and classical functional *gLASSO*, in terms of estimation and selection of variables for three different cases of contamination of Model (2). For simulation study we only consider Model (2), as functional *WLAD-gLASSO* performs best in the presence of leverage points, compared to functional *LAD-gLASSO* and classical functional *gLASSO*. The contamination is done for 15% in Model (2) using the method described in Section 3.3. The response Y is contaminated at 15% and functional predictors X_1 and X_2 are also contaminated at 15% for three cases of contamination, Case 1(Asymmetric contamination), Case 2 (Symmetric contamination) and Case 3 (Partial contamination).

First we consider the squared errors (SE) described in (3.22) to assess the performance of the proposed method. Squared errors are observed in 100 independent simulation runs for three cases of contamination of Model (2). Figure 4.5 shows the boxplots of the squared

errors for Case 1 (Asymmetric contamination), Case 2 (Symmetric contamination) and Case 3 (Partial contamination) for Model (2). The purple, blue and red boxplots in this figure correspond to functional *WLAD-gLASSO*, functional *LAD-gLASSO* and classical functional *gLASSO*, respectively.

Then we consider the Mean Squared Errors (MSE) and the Mean Absolute Error (MAD) of prediction described in (3.23) and (3.24), respectively to assess the predictive ability of the proposed method. Mean Squared Errors and the Mean Absolute Errors are observed in 150 independent simulation runs for three cases of contamination for Model (2). Figures 4.6 and 4.7 show the boxplots of MSE and MAD of prediction for all three cases of contamination for Model (2), respectively. In these figures, purple, blue and red boxplots correspond to functional *WLAD-gLASSO*, functional *LAD-gLASSO* and classical functional *gLASSO*, respectively.

We see in Figures 4.5 - 4.7, that the proposed method functional *WLAD- gLASSO* (purple) performs better than functional *LAD-gLASSO* (blue) and classical functional *gLASSO* (red) for all three cases of contamination for Model (2) setting, that is when there are outliers in both response variable and functional predictors. Also once again, we notice in these figures that our method functional *LAD- gLASSO* purposed in Chapter 3 does not perform any better than classical functional *gLASSO* for Model (2) settings, that is when there are outliers in both response and predictor variables.

To summarize, the proposed method functional *WLAD- gLASSO* performs better when there are outliers both in response and explanatory variables compared to both functional *LAD-gLASSO* and classical functional *gLASSO*.

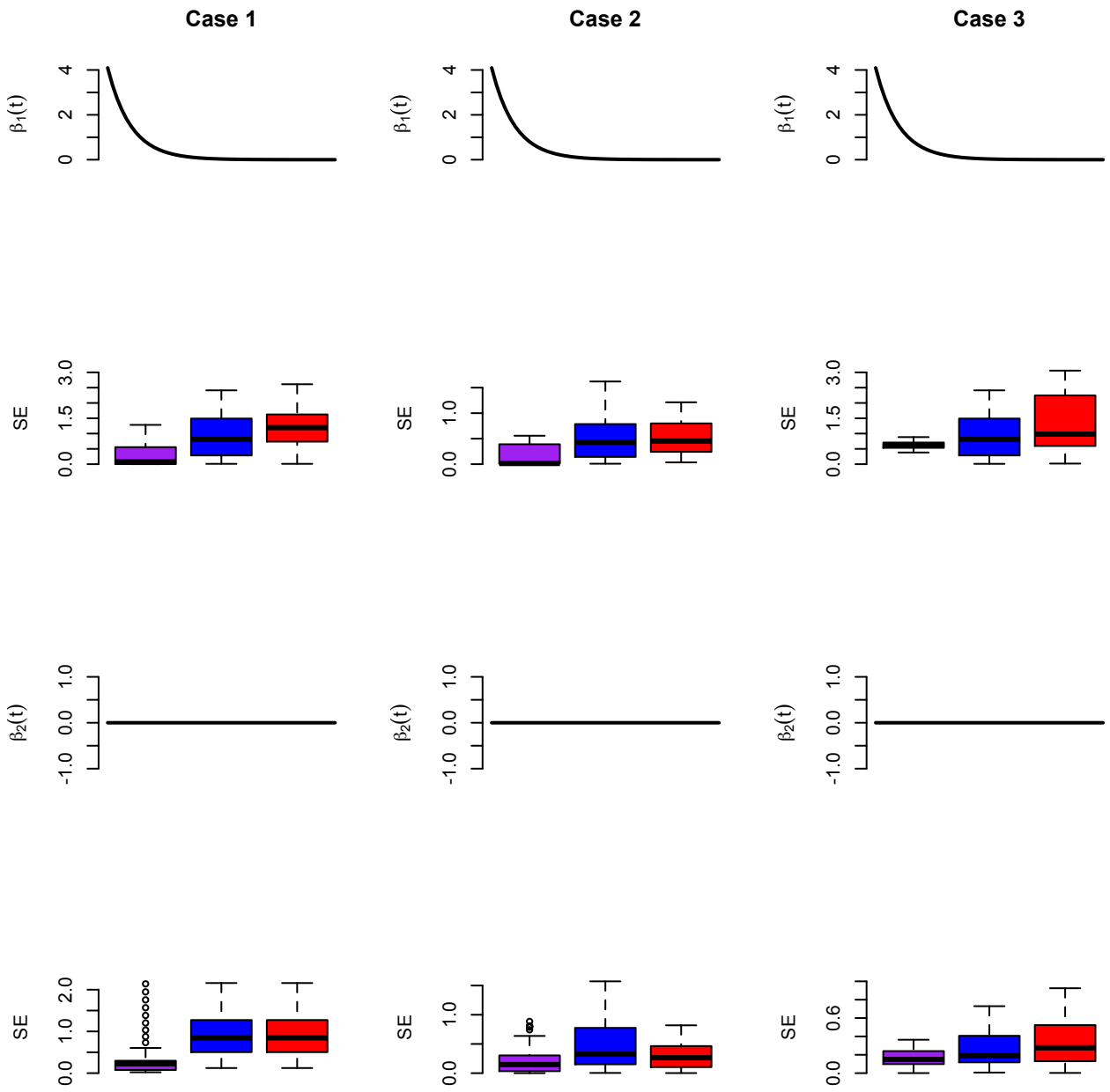


Figure 4.5: SE for the comparison of functional *WLAD-gLASSO* (purple), functional *LAD-gLASSO* (blue) and classical functional *gLASSO* (red) for Model (2) (15% contamination).

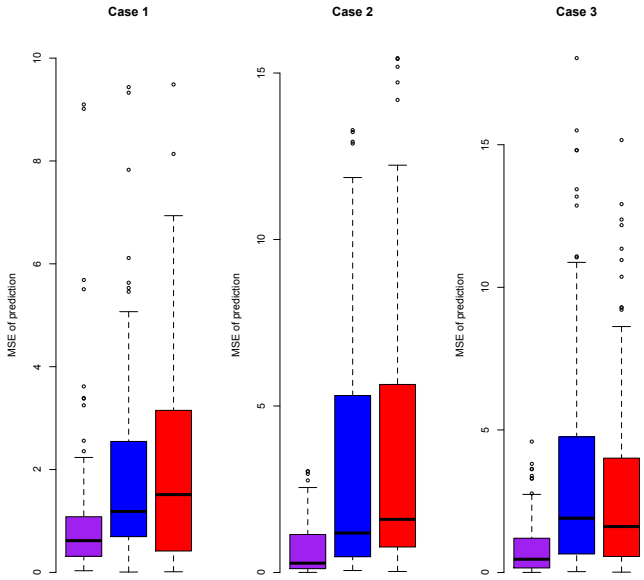


Figure 4.6: MSE of prediction for the comparison of functional *WLAD-gLASSO* (purple), functional *LAD-gLASSO* (blue) and classical functional *gLASSO* (red) for Model (2) (15% contamination).

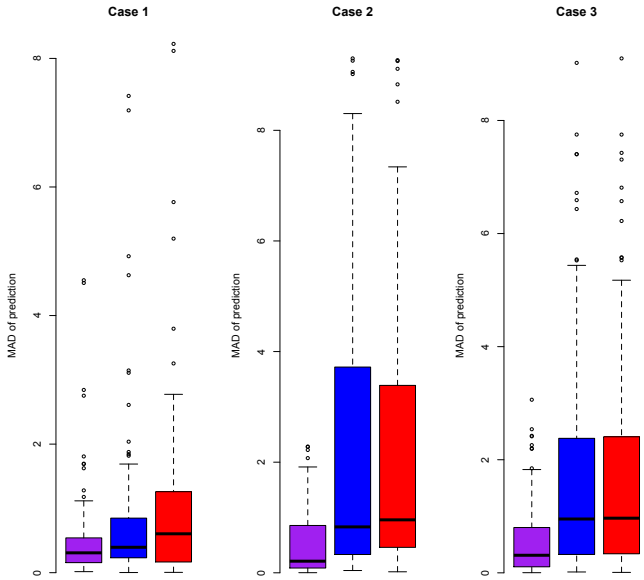


Figure 4.7: MAD of prediction for the comparison of functional *WLAD-gLASSO* (purple), functional *LAD-gLASSO* (blue) and classical functional *gLASSO* (red) for Model (2) (15% contamination).

4.3.2 Numerical Study for functional WLAD- agLASSO

Next we perform numerical study for functional *WLAD-agLASSO*. The performance of functional *WLAD- agLASSO* is assessed by simulation study. In this section we only consider the following Model (2) from Section 3 in Chapter 3, as functional *WLAD-gLASSO* performs best in the presence of outliers in x direction, compared to functional *LAD-gLASSO* and classical functional *gLASSO*:

- **Model (2)**: Presence of outliers in both scalar response Y and the functional predictors $X(t)$.

We consider two functional covariates $X_1(t)$ and $X_2(t)$ each of which is generated from (3.18). 100 curves for each of $X_j(t)$ are observed at 50 equidistant points in (0, 50). Response Y is generated for 100 functional curves from (3.19). The true model is as:

$$Y_i = \alpha + \int_0^{50} \beta_1(t) X_{i1}(t) dt + \epsilon_i. \tag{4.11}$$

where, $i = 1, \ldots 100$ and $\epsilon_i \sim$ N(0,4).

The shapes of parameter functions $\beta_1(t)$ and $\beta_2(t)$ are shown in Figure 4.8. The model is set up where the response is related only to $X_1(t)$.

Both response Y and functional predictors $X_j(t)$ are contaminated at 15 % level using the methods described in Section 3.3. All three cases of contamination, Case 1(Asymmetric contamination), Case 2 (Symmetric contamination) and Case 3 (Partial contamination), of functional predictors $X_j(t)$ are considered.

We apply our proposed method functional *WLAD- agLASSO* to Model (2). Model (2) has outliers both in scalar response Y and the functional predictors $X_1(t)$ and $X_2(t)$. All three cases of contamination, Case 1(Asymmetric contamination), Case 2 (Symmetric contamination) and Case 3 (Partial contamination), are considered for the functional covariates. Also only the first covariate $X_1(t)$ is relevant to the true model and $X_2(t)$ being irrelevant should

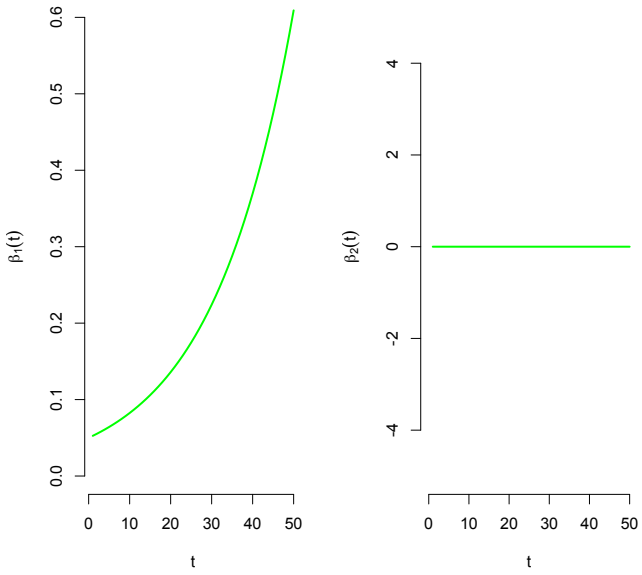


Figure 4.8: $\beta_1(t)$ and $\beta_2(t)$, respectively.

be excluded from the model. We compare functional *WLAD- agLASSO* with functional *WLAD- gLASSO* and classical functional *agLASSO*. Figure 4.9 shows the fitting results of functional functional *WLAD- agLASSO* (blue), *WLAD-g LASSO* (purple) and classical functional *agLASSO* (red). In Figure 4.9, the green curves are the true coefficient functions $\beta_1(t)$ and $\beta_2(t)$.

The proposed method functional *WLAD-agLASSO* is executed using *R* packages *quantreg*, *rrcov* and *refund*. Specifically, functions *CovMcd ()* in *R* package *rrcov* and *pfr ()* in *R* package *refund* are used to compute weights w_i and the initial estimates $\ddot{\beta}_j$ of the coefficients, respectively. *R* package *grplasso* is used for the execution of classical functional *agLASSO*. Figure 4.9 shows that the functional *WLAD- agLASSO* estimates relevant predictor $X_1(t)$ close to its true value at fixed combinations of ($\lambda = 10$, $\varphi = 10$), ($\lambda = 1$, $\varphi = 10$) and ($\lambda = 10$, $\varphi = 10^2$) for Case 1(Asymmetric contamination), Case 2 (Symmetric contamination) and Case 3 (Partial contamination), respectively compared to functional *WLAD- gLASSO* and classical functional *agLASSO*.

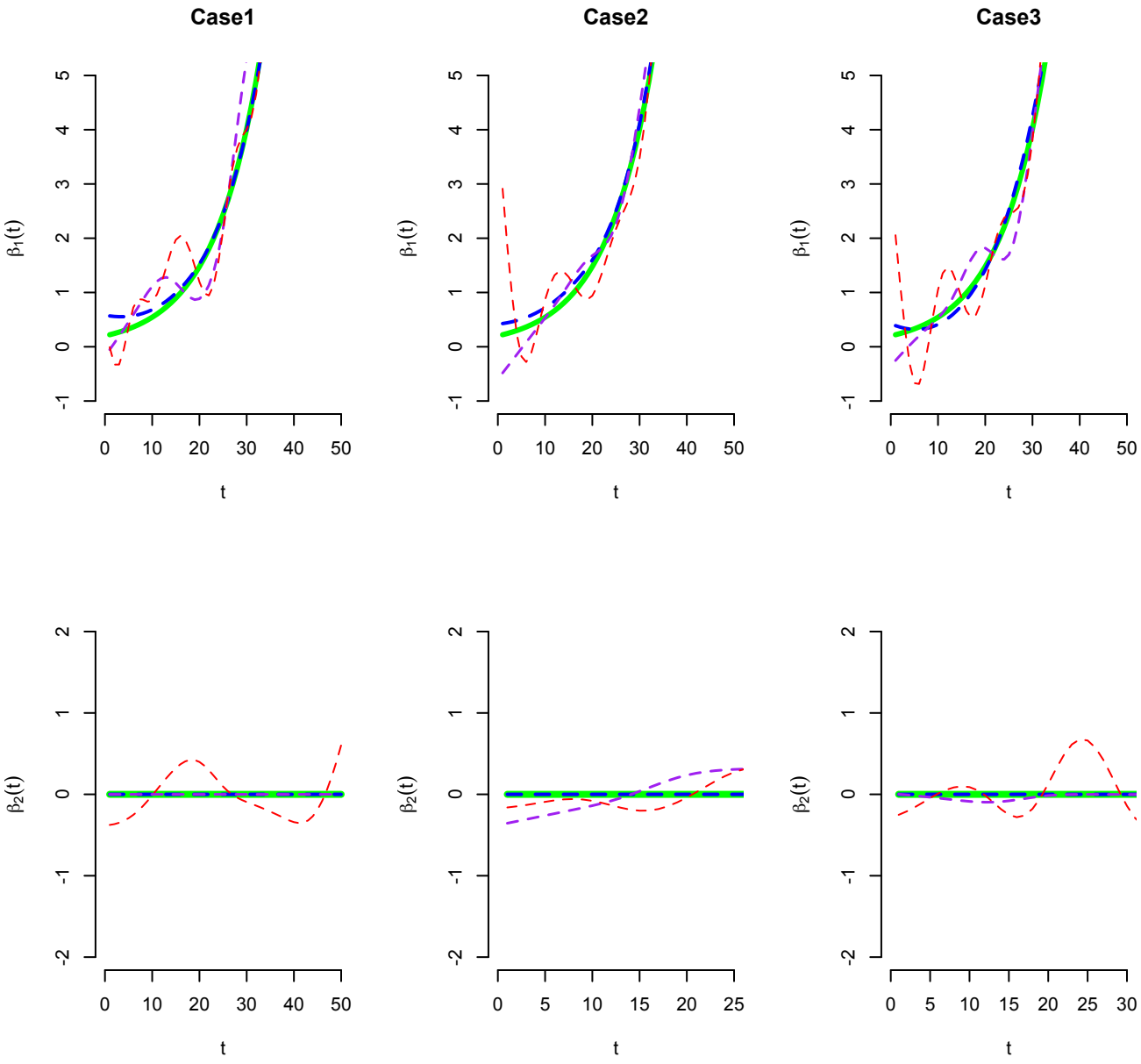


Figure 4.9: Fitting results for the comparison of functional *WLAD- agLASSO* (blue), functional *WLAD-gLASSO* (purple) and classical functional *agLASSO* (red) for Model (2) (15% contamination).

Simulation Study

Furthermore, we conduct a simulation study for functional *WLAD- agLASSO*, where it is compared with functional *WLAD-gLASSO* and classical functional *agLASSO*. Model (2) setting is considered. The data are generated and contaminated at 15% using the methods described in Section 3.3. We generate two functional predictors and 1500 observations for each predictor. Each curve is observed at 200 equidistant points in (0, 100).

First we consider the squared errors (SE) described in (3.22) to assess the performance of the proposed method. Squared errors are observed in 100 independent simulation runs for three cases of contamination of Model (2). Figure 4.10 shows the boxplots of the squared errors for Case 1 (Asymmetric contamination), Case 2 (Symmetric contamination) and Case 3 (Partial contamination) for Model (2). The blue, purple and red boxplots in this figure correspond to functional *WLAD- agLASSO*, functional *WLAD-gLASSO* and classical functional *agLASSO*, respectively.

Secondly, we consider the Mean Squared Errors (MSE) and the Mean Absolute Error (MAD) of prediction described in (3.23) and (3.24), respectively to assess the predictive ability of the proposed method. Mean Squared Errors and the Mean Absolute Errors are observed in 150 independent simulation runs for three cases of contamination for Model (2). Figures 4.11 and 4.12 show the boxplots of MSE and MAD of prediction for all three cases of contamination for Model (2), respectively. Functional *WLAD- agLASSO*, functional *WLAD-gLASSO* and classical functional *agLASSO* are represented by blue, purple and red boxplots, respectively, in these figures.

We see in Figures 4.10 - 4.12, that the proposed method functional *WLAD- agLASSO* (blue) performs better than functional *WLAD-gLASSO* (purple) and classical functional *agLASSO* (red) for all three cases of contamination for Model (2) setting, that is when there are outliers in both response variable and predictors. We also notice that classical functional *agLASSO* performs worse among the three methods.

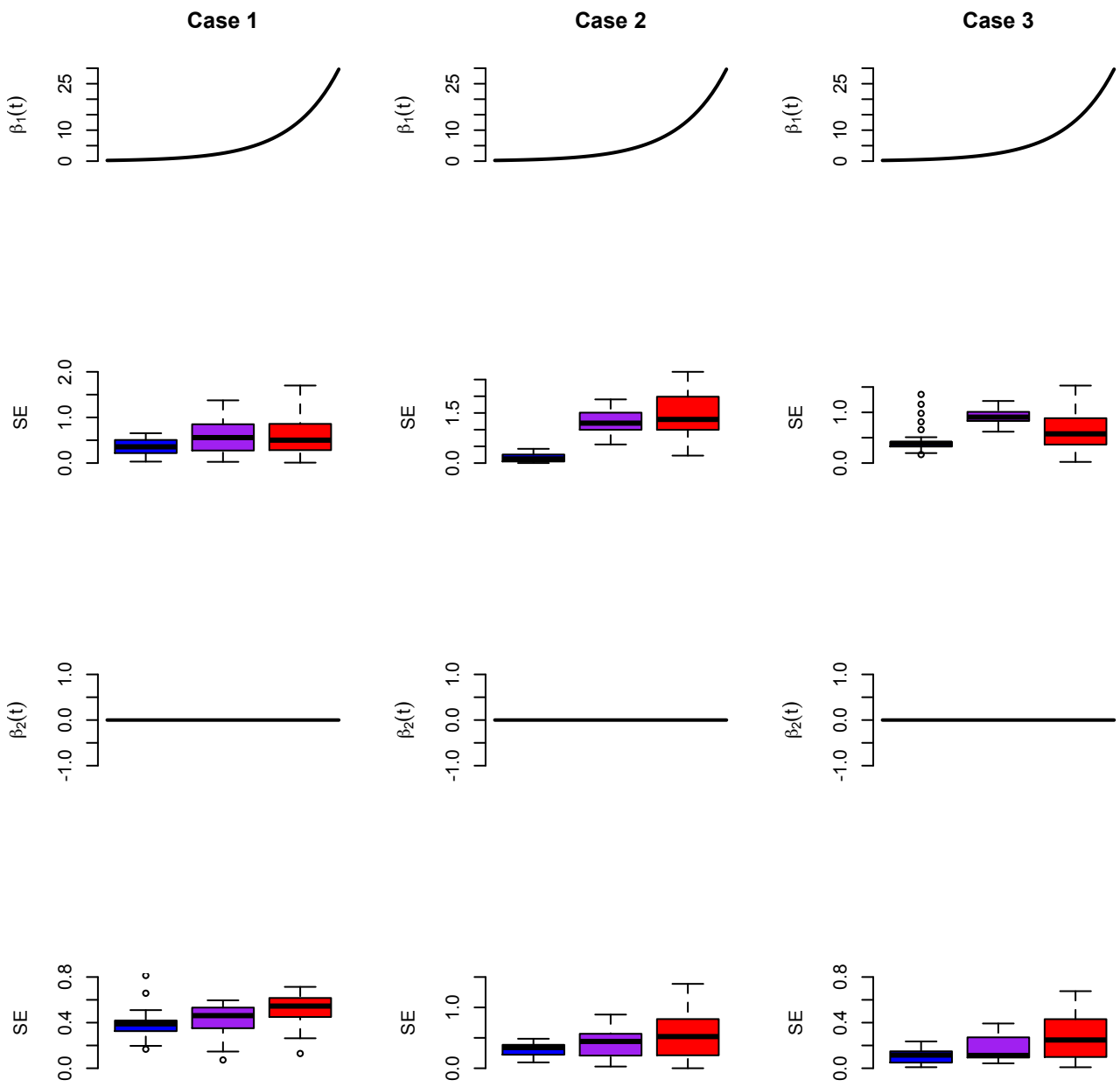


Figure 4.10: SE for the comparison of functional *WLAD-agLASSO* (blue), functional *WLAD-gLASSO* (purple) and classical functional *agLASSO* (red) for Model (2) (15% contamination).

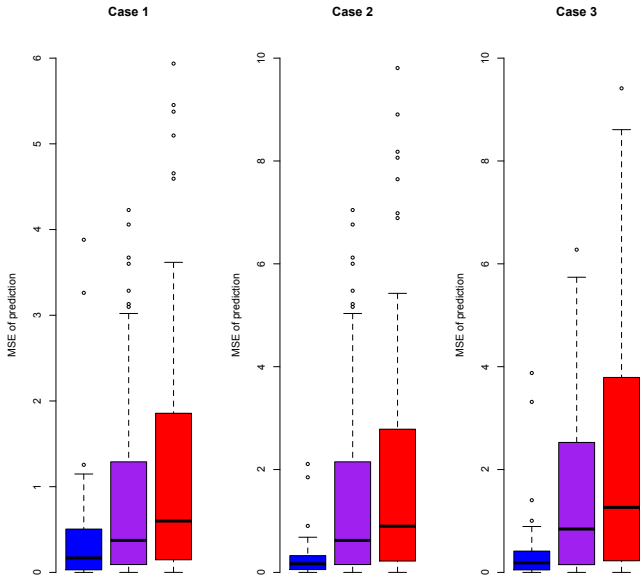


Figure 4.11: MSE of prediction for the comparison of functional *WLAD-agLASSO* (blue), functional *WLAD-gLASSO* (purple) and classical functional *agLASSO* (red) for Model (2) (15% contamination).

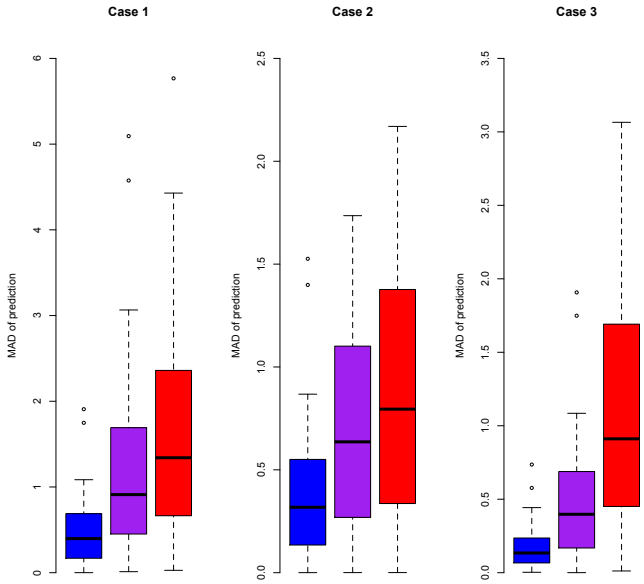


Figure 4.12: MAD of prediction for the comparison of functional *WLAD-agLASSO* (blue), functional *WLAD-gLASSO* (purple) and classical functional *agLASSO* (red) for Model (2) (15% contamination).

Furthermore, Tables 4.1 - 4.3 show the proportions of 50 simulation runs with the respective functional predictor being selected and average model size using functional *WLAD- agLASSO* and functional *WLAD- gLASSO* for three cases of contamination, Case 1(Asymmetric contamination), Case 2 (Symmetric contamination) and Case 3 (Partial contamination), respectively. For this we generate 10 functional predictors, which are observed at 300 equidistant time points in (0, 100). Each predictor has 500 replications. The shape of the corresponding 10 coefficients is same as given in Figure 3.6, which shows that only first 5 predictors are relevant to the true model and remaining 5 are irrelevant to the true model. We see in these tables that the true predictors $X_1(t) - X_5(t)$ are selected most frequently and predictor $X_6(t) - X_{10}(t)$ which are irrelevant to the true model are less frequently selected by the functional *WLAD- agLASSO* compared to functional *WLAD- gLASSO*. In other words, the percentage of false positives and false negatives reduces when functional *WLAD- agLASSO* is used.

	$X_1(t)$	$X_2(t)$	$X_3(t)$	$X_4(t)$	$X_5(t)$	$X_6(t)$	$X_7(t)$	$X_8(t)$	$X_9(t)$	$X_{10}(t)$	Avg. Model Size
Functional *WLAD- agLASSO*	1	1	1	0.98	0.98	0.28	0.34	0.22	0.26	0.38	6.44
Functional *WLAD- gLASSO*	1	1	1	0.98	0.96	0.26	0.38	0.30	0.22	0.48	6.58

Table 4.1: Proportions of runs with respective functional predictor being selected and average model size for Case 1(Asymmetric contamination).

	$X_1(t)$	$X_2(t)$	$X_3(t)$	$X_4(t)$	$X_5(t)$	$X_6(t)$	$X_7(t)$	$X_8(t)$	$X_9(t)$	$X_{10}(t)$	Avg. Model Size
Functional *WLAD- agLASSO*	1	1	1	0.96	0.98	0.30	0.28	0.32	0.34	0.28	6.46
Functional *WLAD- gLASSO*	1	1	1	0.94	0.98	0.32	0.36	0.32	0.30	0.56	6.78

Table 4.2: Proportions of runs with respective functional predictor being selected and average model size for Case 2 (Symmetric contamination).

	$X_1(t)$	$X_2(t)$	$X_3(t)$	$X_4(t)$	$X_5(t)$	$X_6(t)$	$X_7(t)$	$X_8(t)$	$X_9(t)$	$X_{10}(t)$	Avg. Model Size
Functional *WLAD- agLASSO*	1	1	1	0.98	0.96	0.28	0.28	0.32	0.24	0.36	6.42
Functional *WLAD- gLASSO*	1	1	1	0.92	0.94	0.32	0.32	0.38	0.36	0.30	6.54

Table 4.3: Proportions of runs with respective functional predictor being selected and average model size for Case 3 (Partial contamination).

4.4 Real Data Application

We notice that the weather data set considered in Chapter 3, has outliers in both response (precipitation) and functional variables (TEMP, PRESSURE, HUMIDITY and DAYLIGHT). Therefore, this data set is a good candidate for the real data application of our methods proposed in this chapter. Unlike the weather data analysis we presented in Chapter 3, we do not need to remove any outliers before applying our proposed methods functional *WLAD-gLASSO* and *WLAD- agLASSO*.

First we apply our proposed method functional *WLAD- gLASSO* to the weather data. Figure 4.13 shows the resulting estimated coefficient functions. The PRESSURE variable is excluded from the model. According to the resulting model, PRESSURE variable does not seem to have a significant relationship with the precipitation. Secondly, we apply our proposed method functional *WLAD- agLASSO* to the weather data. The fitting results are shown in Figure 4.14. This time PRESSURE and DAYLIGHT are excluded from the model. The results indicate that there is no significant relationship between these variables and the precipitation. The remaining variables, TEMP and HUMIDITY, may relate to the precipitation.

We also apply functional *LAD- gLASSO* to weather data set to compare the resulting model with the model that it provided in Section 3.4. In Section 3.4, outliers were removed from functional predictors before applying functional *LAD- gLASSO*, so that the data has outliers only in response. But in this section we apply functional *LAD- gLASSO* to the original weather data, that is in the presence of outliers in both response and predictors. The results are shown in Figure 4.15. On comparison of Figures 4.15 and 3.25, it is clear that functional *LAD- gLASSO* is not able to exclude any variable(s) from the model, in the presence of outliers both in response variable and functional predictors. Furthermore we apply functional *LAD- agLASSO* to the weather data set. The results are shown in Figure 4.16, in which we see that PRESSURE is excluded from the model. This shows that functional *LAD- agLASSO* gives smaller model size compared to functional *LAD- gLASSO* even in the

presence of outliers in both response and functional predictors.

Additionally, we generate 50 bootstrap samples from the weather data. For each bootstrap sample, functional regression modeling is performed using functional *WLAD- gLASSO*, functional *WLAD- agLASSO*, functional *LAD- gLASSO* and functional *LAD- agLASSO*. We examine how many times each variable is selected. The results are shown in Table 4.4. The table shows that functional *WLAD- agLASSO* gives us the smallest model size and functional *LAD- gLASSO* gives the highest model size among four methods. Also the mean TEMP is selected most frequently among the four variables, followed by HUMIDITY by functional *WLAD- agLASSO*. This reveals significant relationships of these variables to the precipitation. On the other hand, the average PRESSURE and DAYLIGHT are less frequently selected by functional *WLAD- agLASSO*. From the results, there seems to be less of a significant relationship between these variables and the precipitation.

	TEMP	PRESSURE	HUMIDITY	DAYLIGHT	Avg. Model Size
Functional *WLAD- agLASSO*	1	0.36	0.98	0.40	2.74
Functional *WLAD- gLASSO*	1	0.38	0.96	0.66	3.00
Functional *LAD- gLASSO*	1	0.94	0.98	0.96	3.88
Functional *LAD- agLASSO*	1	0.90	0.98	0.92	3.80

Table 4.4: Proportions of runs with the respective functional predictor being selected and average model size.

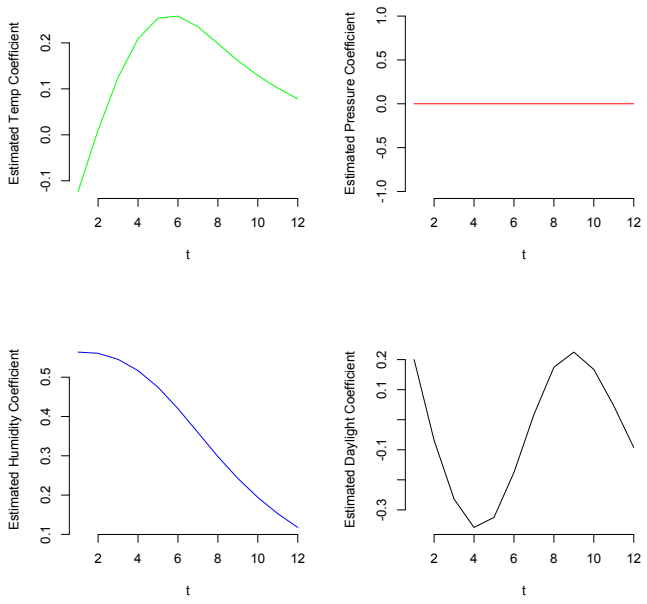


Figure 4.13: Estimated Variable Coefficients for Weather data using functional *WLAD-gLASSO*.

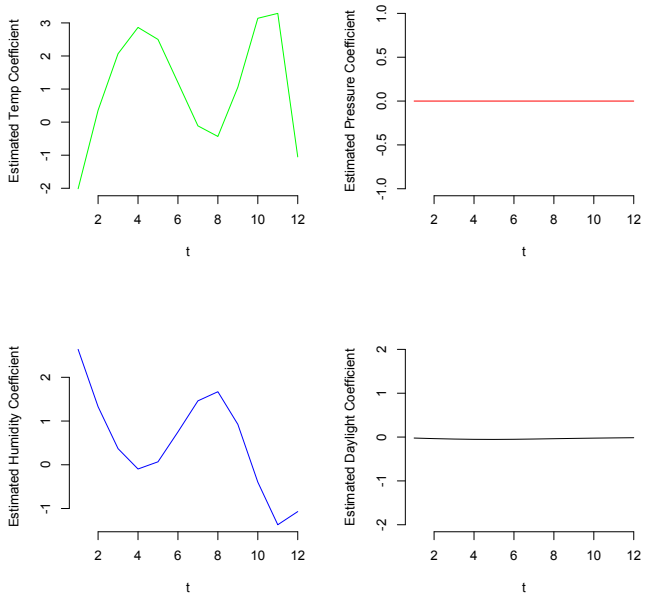


Figure 4.14: Estimated Variable Coefficients for Weather data using functional *WLAD-agLASSO*.

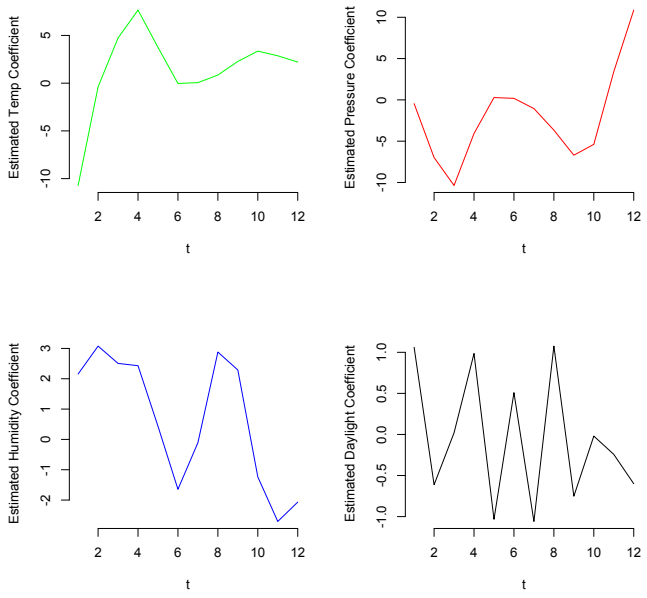


Figure 4.15: Estimated Variable Coefficients for Weather data using functional *LAD-gLASSO*.

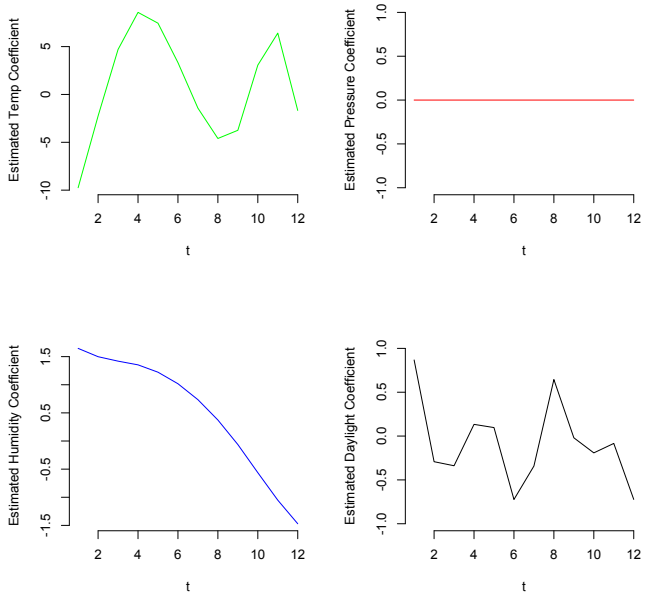


Figure 4.16: Estimated Variable Coefficients for Weather data using functional *LAD-agLASSO*.

4.5 Summary and Discussion

In this chapter we considered two robust variable selection procedures, functional *WLAD- gLASSO* and functional *WLAD- agLASSO*, for functional linear regression models in the presence of outliers both in response and explanatory variables. These methods fix the limitations of the methods proposed in Chapter 3. We found that our proposed methods perform well in terms of prediction error as well as mean squared errors for the estimated coefficient functions compared to fitting a model without taking outliers in x direction into consideration. We also notice that the false positive and false negative rates are low for functional *WLAD- agLASSO* compared to functional *WLAD- gLASSO*.

Further the examination of our proposed methods at 25% and 40% contamination levels, reveals that the proposed methods still perform better than other methods at 25% contamination level, but break down empirically at contamination level of 40%. That is, functional *WLAD- gLASSO* and functional *WLAD- agLASSO* perform no better than functional *LAD-gLASSO* at contamination level of 40%.

Furthermore, in the following chapter, we explore the theoretical properties of our proposed methods.

Chapter 5

Theoretical Properties

5.1 Introduction

As pointed out in Chapter 2, in general a good penalty function should result in an estimator with the following three desired properties.

1. Unbiasedness: The resulting estimator is nearly unbiased when the true unknown parameter is large to avoid unnecessary modeling bias.

2. Sparsity: The resulting estimator is a thresholding rule, which automatically sets small estimated coefficients to zero to reduce model complexity.

3. Continuity: The resulting estimator is continuous to avoid instability in model prediction.

Since the properties of consistency, sparsity, and the oracle property do not hold for the group $LAD-LASSO$ and $WLAD-LASSO$ (Fan and Li [8]), but hold for their adaptive versions (Lilly [20]) in multiple regression model we will not attempt to explore these properties for the functional $LAD-gLASSO$ and $WLAD-gLASSO$. In this study we only focus on the behavior of our proposed estimator functional $LAD-agLASSO$.

Consistency properties of estimation and shrinkage with adaptive group $LASSO$ penalty have been established in the literature (Wang and Leng [34]) for ordinary multiple regression. In the functional regression framework, Zhaoa et al. [37] have established the estimation consistency property of Wavelet- based $LASSO$ estimator and Lian [18] has proved the estimation and selection consistency properties for functional $gSCAD$ estimator. Lian [18] and Zhaoa et al. [37] have the same functional model settings as ours, that is a model with scalar response and functional predictors. Furthermore, Wang et al. [35] have established the oracle

property of functional *gSCAD* estimators for functional response model with time varying coefficients. Lian [18] also points out that estimation and selection consistency properties can also be established for functional Adaptive $LASSO$ based estimator obtained from the method based on combining PC estimation with and $gSCAD$. The main theoretical properties like estimation and selection consistency, of functional *LAD-agLASSO* estimator are considered in this chapter.

We should show that in our context the estimation procedure functional *LAD- agLASSO* can consistently estimate the functional coefficients as well as consistently identify the true model. However, extending these theoretical results to multiple functional regression is not trivial.

5.2 Preliminary Study for Consistency Properties of functional $LAD-agLASSO$

In this section we study the theoretical properties mainly, the estimation consistency and selection consistency of functional *LAD- agLASSO* estimator.
Reconsider the equation (3.7):

$$\Sigma_{i=1}^{n}|Y_i - \alpha - \Sigma_{j=1}^{p}\mathbf{\Phi_{ij}}^T\mathbf{c_j}| + P_{\lambda,\varphi}(\beta_j) \tag{5.1}$$

here, $P_{\lambda,\varphi}(\beta_j)$ is the adaptive $LASSO$ penalty function as discussed in Chapter 3. Specifically,

$$P_{\lambda,\varphi}(\beta_j) = \lambda(\kappa_j||\beta_j||^2 + \nu_j\varphi||\beta_j^{''}||^2)^{1/2} \tag{5.2}$$

where $||.||^2 = \int(.)^2dt$ is the L^2 norm, $\beta_j^{''}$ is the second derivative of β_j, κ_j and ν_j are the data adaptive weights.

Then we redefine the adaptive $LASSO$ penalty function $P_{\lambda,\varphi}(\beta_j)$ as,

$$P_{\lambda,\varphi}(\beta_j) = \lambda(\mathbf{c_j}^T(\kappa_j\Psi_j + \nu_j\varphi\Omega_j)\mathbf{c_j})^{1/2} \tag{5.3}$$

where Ψ_j, Ω_j and $\mathbf{c_j}$ are the same as defined in (3.10) .

Further we can simplify (5.3) as

$$P_{\lambda,\varphi}(\beta_j) = \lambda(\kappa_j\mathbf{c_j}^T(\Psi_j + \frac{\nu_j}{\kappa_j}\varphi\Omega_j)\mathbf{c_j})^{1/2} \tag{5.4}$$

$$P_{\lambda,\varphi}(\beta_j) = \lambda\sqrt{\kappa_j}(\mathbf{c_j}^T(\Psi_j + \frac{\nu_j}{\kappa_j}\varphi\Omega_j)\mathbf{c_j})^{1/2} \tag{5.5}$$

$$P_{\lambda,\varphi}(\beta_j) = \lambda\kappa_j^{'}(\mathbf{c_j}^T(\Psi_j + \nu_j^{'}\Omega_j)\mathbf{c_j})^{1/2} \tag{5.6}$$

where $\kappa_j^{'} = \sqrt{\kappa_j}$ and $\nu_j^{'} = \frac{\nu_j}{\kappa_j}\varphi$

$$P_{\lambda,\varphi}(\beta_j) = \lambda\kappa_j^{'}(\mathbf{c_j}^T(\tilde{C_j})\mathbf{c_j})^{1/2} \tag{5.7}$$

where $\tilde{C_j} = \Psi_j + \nu_j^{'}\Omega_j$ is a l x l symmetric and positive definite matrix. Further $\tilde{C_j}$ can be decomposed using Cholesky decomposition as following:

$$\tilde{C_j} = R_jR_j^T \tag{5.8}$$

where R_j is non-singular lower triangular matrix. Now using (5.7) and (5.8), our model in (5.1) reduces to the following:

$$\Sigma_{i=1}^n|Y_i - \alpha - Z_{ij}^T\mathbf{b_j}| + \lambda\Sigma_{j=1}^p\kappa_j^{'} \parallel \mathbf{b_j} \parallel \tag{5.9}$$

where $\mathbf{b_j} = R_j^T \mathbf{c_j}$ and $Z_{ij} = R_j^{-1} \Phi_{ij}$.

Now $\hat{\alpha}$ and $\hat{\mathbf{b_j}}$'s are the minimizers of (5.9).

For simplicity, define the response vector as $Y = (y_1, \ldots, y_n)^T$, the design matrix as $Z = (Z_1, \ldots, Z_n)^T$ and let $\alpha = 0$. Then the objective function $J(b)$ for functional *LAD- agLASSO* estimator can be written in the matrix form as:

$$J(b) = \| Y - Z\mathbf{b} \|_1 + \lambda \sum_{j=1}^{p} \kappa_j^{'} \| b_j \| \tag{5.10}$$

where $\| . \|_1$ is the L_1 norm.

We denote the true regression coefficients by $\beta = ((\beta^{(1)})^T, (\beta^{(2)})^T)^T$ with $\beta^{(1)} = (\beta_1, \ldots, \beta_s)^T$, $s \leq p$ containing all non vanishing components of β and $\beta_{s+1} = \ldots = \beta_p \equiv 0$.

We want to prove that under certain assumptions, functional *LAD- agLASSO* estimator has

1. (Estimation consistency) $\| \hat{\beta}_j - \beta_j \| = o_p(1)$, $1 \leq j \leq p$.

2. (Selection consistency) $\hat{\beta}_{s+1} = \ldots = \hat{\beta}_p \equiv 0$ with probability converging to 1.

Note that the study of optimal convergence rates for multiple functional regression problem is more complicated and is not attempted here. Also the objective function $J(b)$ is merely the counterpart of functional regression model in ordinary multiple regression. Therefore, to establish the consistency properties for functional *LAD- agLASSO* estimator, we believe that the ideas presented in Wang and Leng [34], Meier et al. [23] and Lilly [20] where they prove the consistency properties of *agLASSO* estimator and $LAD - agLASSO$ estimator for ordinary multiple regression model, respectively. Furthermore, the consistency properties can also be established for functional $WLAD - agLASSO$ estimator using the ideas from Wang and Leng [33], Giloni et al [13], Giloni et al [14] and Lilly [20]. These properties will be studied in details as a future research.

Chapter 6

Conclusion

In this dissertation, we explored the area of robust variable selection for the functional regression model, which has functional predictors and a scalar response. An ample amount of work has been done in various areas of functional data analysis but the area of functional variable selection is seldom discussed. But just as in ordinary multiple regression analysis, variable selection is an important problem in the functional regression framework. Especially, robust variable selection methods for functional regression model do not exist in literature to our knowledge. Therefore, in this dissertation we considered the problem of robust variable selection for functional regression model in the presence of outliers. Essentially, we considered ways that minimize the effect of outliers on the parameter estimator and selector, since the classical existing functional variable selection methods are all based on minimizing the penalized residual sum of squares, which is non- robust in nature, in the presence of outliers. Also since multiple parameters exist for a functional predictor so group variable selection methods are used for selecting functional predictors that select grouped variables rather than individual variables. In this work, we proposed robust variable selection methods using the L1 regularization for functional regression model with a scalar response and the functional predictors in the presence of outliers.

Firstly, we proposed a robust variable selection technique functional *LAD-group LASSO* (*LAD- gLASSO*), which uses a combination of a well known robust loss function *LAD* (Least Absolute Deviation) and equally known group *LASSO* (Least Absolute Shrinkage and Selection) penalty function, for simultaneously estimating and selecting significant functional predictors in a functional regression model in the presence of outliers. However, in

this method same amount of penalty is applied to all the parameters. In order to reflect some subjectivity about the true parameter functions and to allow for different shrinkage and smoothness for the different functional predictors, we then proposed another method functional *LAD- Adaptive group LASSO* (*LAD- agLASSO*), which uses an alternative penalty function based on adaptive weights for the penalized estimation criterion.

Secondly, we propose another robust variable selection technique, functional *Weighted LAD-group LASSO* (*WLAD- gLASSO*), which is a weighted version of the functional *LAD-gLASSO* method. It is well known that the *LAD* based method is only resistant to the outlier in the response variable, but not resistant to the outliers in the explanatory variables, which means functional *LAD- gLASSO* method remains robust only when outliers are present in y direction. Therefore, to deal with the outliers in the functional explanatory variables we proposed functional *WLAD- gLASSO*. This method is not only resistant to outliers in the response variable but also minimizes the effect of the leverage points by introducing weights which are only dependent on the explanatory variables. We also provided an adaptive version of the functional *WLAD-gLASSO* in which adaptive *LASSO* penalty criterion is used to assign different weights to different coefficients to penalize them differently. This method is called functional *Weighted LAD- Adaptive group LASSO* (*WLAD- agLASSO*).

We presented an extensive simulation studies and a real world example to illustrate the performances of the proposed estimators. We also provide preliminary study for the Consistency property of one our proposed methods, functional *LAD- agLASSO*.

6.1 Future Work

In this dissertation, we have shown promising results for functional *LAD- gLASSO* , functional *LAD- agLASSO*, functional *WLAD- gLASSO* and functional *WLAD- agLASSO* for the functional regression model with a scalar response and multiple functional predictors, in the presence of outliers. We would like to generalize these methods for generalized functional linear model where any link function can be used. We would also like to apply our proposed methodologies to an imaging dataset such as fMRI data. We also want to study the robustness properties of functional *LAD- agLASSO* and *WLAD- agLASSO* and prove their theoretical properties as well.

Bibliography

[1] Arslan O., "Weighted LAD-LASSO method for robust parameter estimation and variable selection in regression", *Computational Statistics & Data Analysis*, 56, 1952- 1965, 2012.

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

[3] Boente, G. and Fraiman, R., "Robust principal components for functional data", *Test*, 8, 1- 73, 1999.

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

[5] Denhere, M. and Billor, N., "Robust principal component functional logistic regression", *Communications in Statistics - Simulation and Computation*, 10.1080/03610918.2013.861628, 2014.

[6] Ellis, S. and Morgenthaler, "Leverage and breakdown in L1 regression", *Journal of the American Statistical Association*, 87, 143- 148, 1992.

[7] Escabias, M., Aguilera, A.M. and Valderrama, M.J., "Principal component estimation of functional logistic regression: Discussion of two different approaches", *Journal of Nonparametric Statistics*, 16, 365- 384, 2004.

[8] Fan, J., and Li, R., "Variable Selection via Nonconcave Penalized Like-lihood and Its Oracle Properties", *Journal of the American Statistical Association*, 96, 1348- 1360, 2001.

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

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

[11] Gertheiss, J., Maity, A. and Staicu, A.M., "Variable Selection in Generalized Functional Linear Models", *Stat*, 2, 86- 101, 2013.

[12] Gervini, D., "Robust functional estimation using the median and spherical principal components", *Biometrika*, 95, 587- 600, 2008.

[13] Giloni, A. Simonoff, J. and Sengupta, B., "Robust weighted LAD regression", *Computational Statistics Data Analysis*, 50, 3124- 3140, 2006.

[14] Giloni, A., Sengupta, B. and Simonoff, J., "A Mathematical Programming Approach for Improving the Robustness of Least Sum of Absolute Deviations Regression", *Wiley InterScience http://dx.doi.org/10.1002/nav.20139*, 2006.

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

[16] Hoerl, A.E. and Kennard, R., "Ridge regression: Biased estimation for nonorthogonal problems", *Technometrics*, 12, 55- 67, 1970.

[17] Hubert, M. and Rousseeuw, P., "Robust regression with both continuous and binary regressors", *Journal of Statistical Planning and Inference*, 57, 153- 163, 1997.

[18] Lian, H., "Shrinkage estimation and selection for multiple functional regression", *Statistica Sinica*, 23, 51-74, 2013.

[19] Lilly, K. and Billor, N., "A robust variable selection method for grouped data", *JSM Proceedings, Section on Statistical Learning and Data Mining. Alexandria, VA: American Statistical Association*, 3334-3341, 2013.

[20] Lilly, K., "Robust variable selection methods for grouped data", Unpublished doctoral dissertation, Auburn University.

[21] Maronna, R. A. and Yohai, V. J., "Robust functional linear regression based on splines", *Computational Statistics and Data Analysis*, 65, 46-55, 2011.

[22] Matsui, H. and Konishi, S., "Variable Selection for Functional Regression Models via the L1 Regularization", *Computational Statistics and Data Analysis*, 55, 3304- 3310, 2011.

[23] Meier, L., Van de Geer, S. and Bühlmann, P., "High- dimensional additive modeling", *The Annals of Statistics*, 37, 3779- 3821, 2009.

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

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

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

[27] Rousseeuw, P. J., "Least median of squares regression", *Stat Ass*, 79-871, 1984.

[28] Rousseeuw, P. J. and Driessen, K. V., "A fast algorithm for the minimum covariance determinant estimator", *Technometrics*, 41, 212 - 223, 1999.

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

[30] Tibshirani, R., "Regression shrinkage and selection via the lasso", *Journal of the Royal Statistical Society B*, 58, 267- 288, 1996.

[31] Tutz, G. and Gertheiss, J., "Feature extraction in signal regression: A boosting technique for functional data", *Journal of Computational and Graphical Statistics*, 19, 154-174, 2010.

[32] Viviani, R., Grön, G. and Spitzer, M., "Functional Principal Component Analysis of fMRI Data", *Human Brain Mapping*, 24, 109- 129, 2005.

[33] Wang, H. and Leng, C., "Unified lasso estimation by least squares approximation", *Journal of the American Statistical Association*, 102, 1039-1048, 2007.

[34] Wang, H. and Leng, C., "A note on adaptive group lasso", *Computational Statistics and Data Analysis*, 52, 5277- 5286, 2008.

[35] Wang, L., Chen, G., Li, H., "Group SCAD regression analysis for microarray time course gene expression data", *Bioinformatics*, 23, 1486- 1494, 2007.

[36] Yuan, M. and Lin, Y., "Model selection and estimation in regression with grouped variables", *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68, 49- 67, 2006

[37] Zhao, Y., Ogdenb, R. T., and Reissc, P. T., "Wavelet- Based LASSO in Functional Linear Regression", *Journal of Computational and Graphical Statistics*, 21, 600- 617, 2012

[38] Zhu, H. and Cox, D. D., "A functional generalized linear model with curve selection in cervical pre-cancer diagnosis using fluorescence spectroscopy", *IMS Lecture Notes, Monograph Series - Optimality: The Third Erich L. Lehmann Symposium*, 57, 173- 189, 2009.

[39] Zou, H., "The adaptive lasso and its oracle properties", *Journal of the American Statistical Association*, 101, 1418- 1429, 2006.