Robust Variable Selection in Functional Linear Models

JASDEEP PANNU *, NEDRET BILLOR

*Corresponding author address: Department of Mathematics & Statistics, Auburn University, Auburn, AL 36830.

E-mail: jkp0008@auburn.edu
ABSTRACT

We consider the problem of selecting functional variables using the L1 regularization in a functional linear regression model with a scalar response and functional predictors in the presence of outliers. Since the LASSO is a special case of the penalized least squares regression with L1-penalty function it suffers from the heavy-tailed errors and/or outliers in data. Recently, the LAD regression and the LASSO methods have been combined (the LAD-LASSO regression method) to carry out robust parameter estimation and variable selection simultaneously for a multiple linear regression model. However variable selection of the functional predictor based on LASSO fails since multiple parameters exist for a functional predictor. Therefore group LASSO is used for selecting grouped variables rather than individual variables. In this study we extend the LAD-group LASSO to a functional linear regression model with a scalar response and functional predictors. We illustrate the LAD-group LASSO on both simulated and real data.

**Keywords:** Functional Regression Model; LASSO, LAD-LASSO, Outliers, Variable selection
1. **INTRODUCTION**

Functional data analysis has become increasingly frequent and important in diverse fields of sciences, engineering, and humanities because most of the data collected these days is functional in nature. For example, genomics data, fMRI data, DTI, weather data etc.. The need of analyzing Functional Data from different angles is increasing exponentially. There has been an evolving literature devoted to understanding the performance of estimation of functional predictors. Ramsay and Silverman (), Escabias et al. (), Muller and StadtMuller () discuss different approaches of parameter estimation. Denehe & Billor(), Boente &Fraiman (), Gervini (), Bali et al. (), Sawant et al. (), Goldsmith et al. () and Ogden and Reiss () proposed some robust parameter estimation techniques in functional logistic regression model, functional principal component analysis and generalized functional linear models.

Just as in ordinary data analysis, variable selection is also an important aspect of Functional Data Analysis. The 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 mis-identification. 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.

Not much work has been done in the area of variable selection for functional predictors in functional regression models. Gertheiss et. al (), Matusi & Konishi(), Lian(), Zhao et.al() and Zhu & Cox () propose some variable selection techniques for functional predictors via $L_1$ and $L_2$ regularizations for example using various roughness penalties like groupLASSO, Wavelet Based-LASSO, gSCAD etc. for the generalized functional linear models.

But these methods do not work well in the presence of functional outliers i.e. the observations that deviate from the overall pattern of the data. So since these variable selection techniques are not robust in nature thus there is a need for some robust variable selection method which is resistant to functional outliers.

In this article we propose a new methodology, by extending the ideas of functional group-LASSO from Gertheiss et al. (), that minimizes the effect of outlying curves in the estimation and selection of the functional covariates in
Functional Linear Models.

To our knowledge, there no work that has been done in this area of robust variable selection of the functional linear model. This paper is different from others because we propose a new robust functional variable selection technique for functional covariates in functional linear models.

This article is organized as follows. In section 2, we provide the methodology which covers the formulation of the functional linear regression in subsection (2.a), Penalty settings and selection in subsection (2.b) and Functional LAD-groupLASSO Criterion in subsection (2.c).

In section 3, Numerical studies are presented, under which a simulation study is provided in subsection (3.a), which assesses the proposed robust variable selection approach and in subsection (3.b) a real data application of the proposed method is provided.

The last section concludes with a summary and discussion.

We consider the problem of selecting functional variables using the L1 regularization in a functional linear regression model with a scalar response and functional predictors in the presence of outliers.

The first approach that we take in this paper is reducing the functional lin-
ear model to a multiple linear one by approximating the functional covariates as a linear combination of an appropriate basis as discussed in Ramsay and Silverman ()

2. METHODOLOGY

a. **Functional Regression Model**

Functional data are data that have been sampled discretely over a continuum, usually time. There is assumed to be an underlying curve describing the data. The functional data can be represented as:

\[(y_i, x_i(t)), \quad t \in T, \quad i = 1, \ldots, n,\]

We consider the response to be scalar. Let \(Y_i\) represent the scalar response for the \(i\)th subject and \(X_1, X_2, \ldots, X_p\) be the random curves where \(X_{i1}, X_{i2}, \ldots, X_{ip}\) denote their independent realizations respectively. The \(X_j\)'s are assumed to be real squared integrable with zero mean function. For the sake of simplicity, each \(X_{ij}\) is considered to be observed without measurement error at a dense grid of time points \(\{t_{j1}, t_{j2}, \ldots, t_{jN_j}\}\). The response \(Y_i\) is assumed to have distribution in the exponential family with linear predictor \(\eta_i\) and dispersion
parameter $\phi$

The linear predictor $\eta_i$ takes the following form:

$$\eta_i = \alpha + \sum_{j=1}^{p} \int X_{ij}(t)\beta_j(t)dt$$

where $h(.)$ is a known link function with $h^{-1}(\eta_i) = \mu_i = E[Y_i|X_{ij}]$

We get the following functional liner model if $h(\mu_i) = \mu_i$

$$Y_i = \alpha + \sum_{j=1}^{p} \int X_{ij}(t)\beta_j(t)dt + \epsilon_i$$

The random error terms $\epsilon_i$ are assumed to be independent Normally distributed with mean 0 and variance $\sigma^2$.

A multivariate generalization can be applied to this functional form of the model using two methods.

**METHOD I**

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

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

where, $\phi_{jb}(t)$ are the basis functions and $a_{ijb}$ are the coefficients.

The coefficient functions $\beta_j$ can also be expressed as linear combination of
some known basis functions.

\[ \beta_j(t) = \sum_{b=1}^{K} c_{jb} \phi_{jb}(t) \]  

(1)

where \( \phi_{jb}(t) \) are the known basis functions and \( c_{jb} \) are the unknown corresponding coefficients.

Most commonly used basis functions are Fourier basis, B-spline, wavelet or polynomial basis. The kind of basis to be used depends on the nature of the data. But as suggested by Mercer’s theorem and Karhunen-Loève theorem (), the orthonormal basis can also be taken to be the eigenfunctions of the covariance operator \( K \), where if \( k(s,t) = \text{Cov}(X(s),X(t)) \) then \( K \) is given by:

\[ KX(t) = \int X(t)k(s,t)ds \]

The coefficients \( a_{ijb} \) 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 Hall et.al () and Ramsay and Silverman().

In this method, the smoothness of the \( \beta_j \) is controlled by the number of basis functions used. The smaller it is the smoother the solution is.
METHOD II

The curves $X_{ij}(t)$ can be also discretized as Riemann Integration as below:

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

and the coefficient functions $\beta_j$ are defined the same way as in (1) i.e.

$$\beta_j(t) = \Sigma_{b=1}^q c_{jb}\phi_j(t)$$

We use method II in this paper and a pre-set basis to express $\beta_j$ in the above form. Also we assume no restriction on the number of basis function. As suggested by Gertheiss (), the smoothness of $\beta_j$ is controlled by smoothing parameter $\varphi$.

Then

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

where,

$$\delta_j = t_{jm} - t_{j,m-1}$$

$$c_j = (c_{j1}, \ldots, c_{jq})^T$$

$$\Phi_{ij} = (\Phi_{ij1}, \ldots, \Phi_{ijq})^T$$

$$\Phi_{ijb} = \delta_j \Sigma_m X_{ij}(t_{jm})\phi_j(t_{jm})$$
The new model that we get is:

\[ Y_i = \alpha + \sum_{j=1}^{p} \Phi_{ij}^T c_j + \epsilon_i \]

where \( \Phi_{ij} \) are known but \( \alpha \) and \( c_j \)'s are the unknown regression coefficients that need to be estimated.

**b. Penalty Settings**

The proposed penalty function is as used by J. Gerthesis et al. (2013)\(^1\)

\[ P_{\lambda,\varphi}(\beta_j) = \lambda(||\beta_j||^2 + \varphi||\beta_j'||^2)^{1/2} \]

where

\[ ||.||^2 = \int (.)^2 dt \]

is the \( L^2 \) norm and \( \beta_j'' \) is the second derivative of \( \beta_j \)

Then

\[ P_{\lambda,\varphi}(\beta_j) = \lambda(c_j^T(C_{\varphi,j})c_j)^{1/2} \]

where, \( C_{\varphi,j} = \Psi_j + \varphi\Omega_j \) is a \( l \times l \) symmetric and positive definite matrix.

\( \Psi_j \) is a \( l \times l \) matrix whose \((b,k)\)th element is \( \int \phi_{jb}(t)\phi_{jk}(t)dt \) and \( \Omega_j \) is a \( l \times l \) matrix whose \((b,k)\)th element is \( \int \phi_{jb}''(t)\phi_{jk}''(t)dt \) for \( b,k = 1,\ldots,l \)
Further $C_{\phi,j}$ can be decomposed using Cholesky decomposition as following:

$$C_{\phi,j} = L_{\phi,j}L_{\phi,j}^T$$

where $L_{\phi,j}$ is non-singular lower triangular matrix. Also let us define $\tilde{c}_j(t) = L_{\phi,j}^Tc_j$ and $\tilde{\Phi}_j = L_{\phi,j}^{-1}\Phi_j$

c. Functional LAD-groupLASSO Criterion

Now $\hat{\alpha}$ and $\hat{c}_j(t)$’s are the minimizers of

$$\sum_{i=1}^{n}|Y_i - \alpha - \tilde{\Phi}_{ij}^T\tilde{c}_j| + \lambda \sum_{j=1}^{p}||\tilde{c}_j||$$

Here $||.||$ is the Euclidean norm in $\mathbb{R}^d$ and $\hat{\beta}_j(t) = \sum_{b=1}^{d}\phi_{jb}(t)c_{jb}(t)$ for $j=1,...,p$

d. Selection of Tuning Parameters

$\lambda$ and $\varphi$ are selected via $K$-fold cross-validation. The most commonly used values of $K$ are 5 and 10. The given sample is partitioned into $K$ equal sized subsamples. $K$ estimates of the prediction error of the model are
obtained by treating each of the $k$th subsample as a validating data and the remaining $K-1$ as the training data. The average of these $K$ prediction error estimates is taken and the values of tuning parameters that minimize the overall prediction error are selected. The prediction accuracy is measured by sum of absolute errors $\frac{1}{n}\sum_i |Y_i - \hat{Y}_i|$

3. NUMERICAL EXPERIMENTS

a. Simulation Study

We take the following steps to carry out simulation study to investigate our proposed method.

1. Generating data as below:

STEP 1: Generating Functional Predictors $X_j(t)$

In the simulation study we consider only two functional covariates $X_1(t)$ and $X_2(t)$.

We generated 50 sample curves for each $X_j(t)$ which are observed at 50 equidistant time points.

Functional Predictors $X_j(t)$s are simulated similar to Tutz and Gerthesis (2010)[2]
\[ X_{ij}(t) = [\sigma(t)]^{-1} \sum_{r=1}^{5} (a_{ijr} \sin(\pi t (5 - a_{ijr})/150) - m_{ijr}) \]

where \( i = 1, \ldots, 50 \) and \( j = 1, 2 \)

Here \( a_{ijr} \sim U(0, 5), m_{ijr} \sim U(0, 2*\pi) \)

\( \sigma(t) \) is defined so that \( \text{var}[X_{ij}(t)] = 0.01 \)

50 sample curves of \( X_1(t) \) and \( X_2(t) \) are shown below respectively in Figure 1.

![Figure 1](image.png)

**Figure 1.** 50 Sample curves of \( X_1(t) \) and \( X_2(t) \) respectively.

**STEP 2: Generating Y**

Response \( Y \) is defined as:

\[ Y_i = \alpha + \int_0^{300} \beta_1(t) X_{ij}(t) dt + \epsilon_i \]
where $\epsilon_i \sim N(0,4)$

and the parameter function $\beta_1(t)$ has a sine-wave function shape as shown in Figure 1 below.

So the response depends only on the first functional predictor and not on the second.

![Figure 1](image1.png)

Figure 1. $\beta_1(t)$

2. Corrupting $X_{ij}(t)$

We consider corrupting only $X_{i1}(t)$ to produce functional outliers. $X_{i2}(t)$ is not corrupted.
The contamination process is carried out as described by Fraiman & Muniz() and Denhere& Billor(). The following five cases of contamination are considered:

- **Case (1):** No contamination $X_{i1}(t)$ are generated as described previously.

- **Case (2):** Asymmetric contamination $Z_{ij}(t) = X_{i1}(t) + cM$ where $c$ is 1 with probability $q$ and 0 with probability $1-q$ and $q = \{0\%; 5\%; 10\%; 15\%; 20\%\}$; $M$ is the contamination constant size equal to 10 and $X_{ij}(t)$ is as defined in Case(1).

- **Case (3):** Symmetric contamination $Z_{ij}(t) = X_{ij}(t) + c\sigma M$ where $X_{ij}(t)$, $c$ and $M$ are as defined before and $\sigma$ is a sequence of random variables independent of $c$ that takes the values 1 and -1 with probability 0.5.

- **Case (4):** Partial contamination $Z_{ij}(t) = X_{ij}(t) + c\sigma M$ if $t \leq T$ and $Z_{ij}(t) = X_{ij}(t)$ if $t > T$, where $T \sim U[0, 10]$

- **Case (5):** Peak contamination $Z_{ij}(t) = X_{ij}(t) + c\sigma M$ if $T \leq t \leq T + l$ and $Z_{ij}(t) = X_{ij}(t)$ if $t \notin [T, T + l]$ where $l = 2$ and $T \sim U[0, 10 - l]$

The effects of these different types of contamination are shown in Figure 3.
Figure 3. The contaminated $X_{i1}(t)$ curves for cases 2-5 at $q = 15\%$
As the sparseness parameter $\lambda$ increases, the estimated coefficient functions $\beta(t)$’s are shrunk and at some value, set to zero. As the smoothing parameter $\varphi$ increases, the departure from linearity is penalized stronger and thus the estimated curves become closer to a linear function. Smaller values for $\varphi$ result in very wiggly and difficult to interpret estimated coefficient functions. For optimal estimates (in terms of accuracy and interpretability), an adequate $(\lambda, \varphi)$ combination has to be chosen. Figure 4 below shows the comparison of the classical functional LASSO with new proposed method functional LAD-group LASSO which is robust in nature in the presence of functional outliers. The green curve is the true function $\beta_1(t)$. The red line in the plots represents the estimation done by classical functional LASSO and the blue line represents the estimation done by robust functional LAD-group LASSO at fixed combination of $(\lambda=0.4, \varphi=100)$. We can see from the figure that in the presence of outliers, classical method does poor estimation and shrinkage, whereas new robust method does good both in estimation and shrinkage.
Figure 4 Comparison on the effect of outlier curves on classical functional groupLasso and robust functional LAD-groupLASSO at 15% contamination.
4. Real Data Application

5. Summary and Discussion

6. Acknowledgement

Acknowledgments.

7. References

8. Figures and tables