Scientific Research Publishing

Advances in Pure Mathematics, 2025, 15(3), 220-234
https://www.scirp.org/journal/apm
ISSN Online: 2160-0384
ISSN Print: 2160-0368

Robustness of Variable Selection Based on Functional Response Regression Model

Ziqi Yan, Wanzhou Ye

Department of Mathematics, College of Science, Shanghai University, Shanghai, China
Email: ziqiyan1@126.com, wzhy@shu.edu.cn

How to cite this paper: Yan, Z.Q. and Ye, W.Z. (2025) Robustness of Variable Selection Based on Functional Response Regression Model. *Advances in Pure Mathematics*, **15**, 220-234.
https://doi.org/10.4236/apm.2025.153010

Received: February 18, 2025
Accepted: March 17, 2025
Published: March 20, 2025

Abstract

Function-on-scalar regression is a type of function response regression used to analyze the relationship between function response and a set of scalar predictor factors. The variable selection methods of FOSR models mostly focus on the linear effects of scalar predictor factors. Therefore, in this paper, we perform robust variable selection for nonlinear FOSR models with the presence of multiple continuous covariates to further explain the behavior of function response over time. We project functional data into a low dimensional principal component space via the principal component score of the function response, in order to use the principal component score for variable selection and regression modeling. In this paper, we develop a regularized iterative algorithm based on exponential squared loss and group smoothly clipped absolute for predicting estimates of scalar factors and function coefficients, and use tuning parameters chosen by data-driven methods to achieve robustness in variable selection. The robustness of the proposed method is verified through simulation studies and demonstrated on real datasets.

Keywords

Functional Data Analysis, Functional Response Regression, B-Splines, Robust Estimation, Variable Selection

1. Introduction

With the advancement of modern technology, more and more data are recorded continuously at intervals of time, or intermittently at several discrete points in time, often stored in the form of curves and images. Functional data are essentially infinite dimensional its basic unit is the function. A common type of functional data are curves recorded during a time interval. Functional Data Analysis (FDA) includes statistical methods for such data. FDA deals with the analysis of data

which is in the form of functions. The functions could be curves, images, or other types of forms, depending on the set on which the functions are recorded. Functional data analysis is mainly divided into the following three regression models: 1) scalar-on-function regression: a continuous response variable regressed on functional covariates; 2) function-on-function regression: a functional response regressed on functional predictors; 3) Function-on-scalar regression: a functional response regressed on scalar predictors. The aim of this paper is to provide robust variable selection for function-on-scalar regression models containing nonlinear scalar predictors.

Function-on-scalar regression (FOSR) is an increasingly popular area of research in the analysis of functional data, e.g. Barber *et al.* [1] for genome-wide association studies (GWAS), Chen *et al.* [2] for the effect of stroke severity on motor control, Goldsmith *et al.* [3] for physical activity (PA), and so on. In FOSR, a continuous functional response is typically modeled by the additive effects of a set of scalar predictors of interest, which are captured using smooth univariate coefficient functions. This has led to several methods of variable selection in the FOSR model. Chen *et al.* [2] used a pre-whitening approach to account for temporal correlation within curves and proposed a variable selection procedure via group minima and maxima concave penalties (MCP) regularization with a penalized least squares technique. Barber *et al.* [1] used a penalized least squares objective function with the LASSO penalty on the basis coefficients and proposed a function-on-scalar LASSO (FSL). Fan and Reimherr [4] further developed an adaptive FSL in high dimensions. Parodi and Reimherr [5] proposed a functional linear adaptive mixed estimation method to achieve variable selection and smoothing for function-on-scalar regression.

The vast majority of FOSR methods focus on linear effects of scalar predictors, but this may not necessarily hold true in many applications. Ghosal and Maity [6] proposed a nonlinear function-on-scalar regression approach considering the additional nonparametric effects of the scalar predictors into the FOSR model. This paper is based on this nonlinear model for robust variable selection. It is well known that the least squares method is sensitive to outliers in the data, however, the variable selection methods discussed above are based on the least squares criterion. Therefore, it is necessary to propose a more robust variable selection method to replace the least squares method in the presence of outliers. Many loss functions are resistant to outliers, such as minimum absolute deviation loss, quantile loss, or Huber loss, which can be used for robust estimation and variable selection. In this paper, we introduce the exponential squared loss (ESL) introduced by Wang *et al.* [7] in nonlinear FOSR to reduce the effect of outliers. ESL is a bounded continuous function, and we can control the robustness of the estimator by selecting the appropriate adjustment parameter h, that is, reducing the influence of outliers that lead to large errors. The ESL function is defined as $\phi_h(t) = 1 - \exp\left(-t^2/h\right)$. When h is very large, $\phi_h(x) \approx x^2/h$, and therefore the proposed estimator is similar to the least squares estimator in the extreme case.

When h is small, observations with larger absolute values will have a small impact on the loss function $\phi_h(x)$ and therefore have a smaller impact on the estimator.

In this paper, we investigate the use of functional principal component analysis (FPCA) to project the functional response into a low-dimensional principal component space in a nonlinear FOSR model. The principal component scores are used as target variables in a regression model to analyze the effect of scalar covariates on the functional response and perform robust variable selection for the nonlinear covariates. The nonlinear covariates are approximated by a B-spline basis function, and the estimation of the coefficient function combines the FPCA eigenfunctions and the B-spline basis function. We integrate the exponential squared loss function into the nonlinear FOSR model and propose a robust variable selection process using the group SCAD regularization method. By locally linearly approximating the SCAD penalty function, it is replaced by a set of weighted LASSO penalty functions. We utilize a data-driven approach to select the smoothing parameters, which achieves robustness and establishes consistency of the proposed estimators. The estimated coefficients are projected back using univariate basis functions and eigenfunctions of the functional response to recover the effects of the scalar predictors. Finally, we show through simulations and real data applications that the proposed method performs well. When outliers are present in the dataset, our method is robust to outliers. Meanwhile, in the absence of outliers, the proposed estimation method is comparable to the least squares estimation method.

The rest of this paper is organized as follows. In Section 2, we describe the estimation methods for nonlinear function-on-scalar regression and discuss the theoretical properties of the proposed estimators. In Section 3, we present the implementation algorithm for computing the estimator based on nonlinear FOSR and the criterion used to select the smoothing parameter. To demonstrate the superiority of the proposed method, in Section 4, we report and compare the simulation results and illustrate the proposed method with a real diffusion tensor imaging data example. Finally, in Section 5, the summary and conclusion of the paper are given in Section 5.

2. Estimation Method and Theoretical Properties

2.1. Estimation Method

Suppose that the observed data for the ith subject is $\{Y_i(t), M_{i1}, M_{i2}, \cdots, M_{iq}\}$ $(i = 1, 2, \cdots, n)$ where $Y_i(t)$ represents the functional response and $M_{i1}, M_{i2}, \cdots, M_{iq}$ are the corresponding scalar predictors of interest. Sometimes there might be additional control or confounding variables $X_{i1}, X_{i2}, \cdots, X_{ip}$ which we want to adjust for. In this paper, we assume the functions are observed on a dense and regular grid of points $S = \{t_1, t_2, \cdots, t_m\} \subset T = [a, b]$ for some $a, b \in R$ where the functional response is observed on an irregular and sparse domain. Now, the linear FOSR model is extended to the nonlinear dynamic effect of scalar

predictors, and we use the following generalized function on scalar regression (GFOSR) model proposed by Ghosal and Maity [6]:

$$Y_i(t)=\mu(t)+\sum_{j=1}^{p}X_{ij}\beta_j(t)+\sum_{l=1}^{q}\theta_l(M_{il},t)+\varepsilon_i(t), \tag{1}$$

Here, $\beta_j(t)$ represents the functional effect of the scalar predictor X_{ij}, and $\theta_l(,)$ is an unknown smooth function on $R\times T$ (twice differentiable with respect to both parameters), used to capture the functional effect of the predictor M_{il}. The coefficient functions $\beta_j(\cdot)$, $\theta_l(\cdot,\cdot)$, functional response $Y_i(\cdot)$, and the error process $\varepsilon_i(\cdot)$ are assumed to lie in a real separable Hilbert space [1]. In this paper, we focus our attention to $L^2(T)$. We assume that the covariates are centered and $\mu(\cdot)$ is a smooth function capturing marginal mean of the functional response $Y_i(\cdot)$, where the error functions $\varepsilon_i(\cdot)$ are i.i.d. of $\varepsilon_i(\cdot)$ which is a mean zero stochastic process with unknown nontrivial covariance structure.

Let $Y_1(t),\cdots,Y_n(t)$ be independent realizations of a smooth random function $Y(t)$ with mean function $\mu(t)$ and covariance function $\operatorname{cov}\{Y(t),Y(s)\}=C(t,s)$. By the Karhunen-Loève theorem by Ash and Gardner [8], the random function $Y_i(t)=\mu(t)+\sum_{k=1}^{\infty}\xi_{ik}\phi_k(t)$ where $\mu(t)=E\{Y_i(t)\}$ is the overall mean function; $\phi_k(t)$ is the kth orthonormal eigenfunction of the covariance function, satisfying $\int\phi_k(t)\phi_j(t)\mathrm{d}t=1$ if $j=k$ and 0 otherwise; ξ_{ik} denotes FPC scores with $E(\xi_{ik})=0$, $\operatorname{var}(\xi_{ik})=\rho_k$ and $\operatorname{cov}(\xi_{ij},\xi_{ik})=0$ if $j\neq k$, and ρ_k is the eigenvalue corresponding to the eigenfunction $\phi_k(t)$. Since $\sup_{t\in[0,1]}E\left\{\sum_{k=1}^{\infty}\xi_{ik}\phi_k(t)-\sum_{k=1}^{K_n}\xi_{ik}\phi_k(t)\right\}^2\to 0$, one can reasonably suppose that

$$Y_i(t)\approx\mu(t)+\sum_{k=1}^{K_n}\xi_{ik}\phi_k(t) \tag{2}$$

as $K_n\to\infty$. The approximation of (2) with a fixed K_n is commonly adopted for longitudinal and functional data analysis (see Yao *et al.* [9]). To be more flexible, Hall and Mohammad [10] considered model (2) with $K_n\to\infty$ as $n\to\infty$. In this paper, we adopt the FPCA method of Yao *et al.* [9], which is used to estimate the FPCA score ξ_{ik}, the mean function $\mu(t)$, and the eigenfunction $\phi_k(t)$. Specifically, a spectral decomposition of the covariance function yields $\mu(t)$ and $\phi_k(t)$, and thus ξ_{ik} is estimated by numerical integration. Thus, the model in (1) becomes:

$$\begin{aligned}\xi_{ik}&=\int_T\left(Y_i(t)-\mu(t)\right)\phi_k(t)\mathrm{d}t\\&=\sum_{j=1}^{p}X_{ij}\int_T\beta_j(t)\phi_k(t)\mathrm{d}t+\sum_{l=1}^{q}\int_T\theta_l(M_{il},t)\phi_k(t)\mathrm{d}t+\int_T\varepsilon_i(t)\phi_k(t)\mathrm{d}t\\&=\sum_{j=1}^{p}X_{ij}\beta_{jk}+\sum_{l=1}^{q}\theta_{lk}(M_{il})+\varepsilon_{ik}\end{aligned} \tag{3}$$

for $k=1,\cdots,K$, where we define $\beta_{jk}=\int_s\beta_j(t)\phi_k(t)\mathrm{d}t$ and $\theta_{lk}(u)=\int_T\theta_l(u,t)\phi_k(t)\mathrm{d}t$.

We project the nonlinear variable M_{il} onto the B-spline basis function space, and expand the nonparametric function $\theta_l(\cdot)$ with B-spline basis functions as $\theta_{lk}(u)=\int_T \theta_l(u,t)\phi_k(t)\mathrm{d}t$. Let $B_j(t)=\left(B_j(t):1\le j\le J_n\right)^{\mathrm{T}}$ denote the qth-order B-spline basis functions, where $J_n=N_n+q$, and N_n is the number of interior knots for a knot sequence $\alpha_1=\cdots=0=\alpha_q<\alpha_{q+1}<\cdots<\alpha_{N_n+q}<1=\alpha_{N_n+q+1}=\cdots=\alpha_{N_n+2q}$. So, model (3) becomes as follows:

$$\xi_{ik}=X_i^{\mathrm{T}}\beta_K+\sum_{l=1}^{q}\sum_{j=1}^{J_n}B_j(M_{il})\delta_{l,k,j}+\varepsilon_{ik} \tag{4}$$

where we define $X_i^{\mathrm{T}}=\left(X_{i1},\cdots,X_{ip}\right)$, $\beta_k^{\mathrm{T}}=\left(\beta_{1k},\cdots,\beta_{pk}\right)$, $\delta_{\cdot k\cdot}^{\mathrm{T}}=\left(\delta_{1,k,1},\cdots,\delta_{1,k,J_n},\cdots,\delta_{q,k,J_n}\right)$, $\beta_k^{\mathrm{T}}=\left(\beta_{1k},\cdots,\beta_{pk}\right)$, and $\eta_i^{\mathrm{T}}=\left(B_1(M_{i1}),\cdots,B_1(M_{iq}),B_2(M_{i1}),\cdots,B_{J_n}(M_{iq})\right)$. Let $Z_i[U_i,V_i]$ dimension of $K\times\left(Kp+KqJ_n\right)$, $\gamma=(\beta,\delta)$, where $U_i=I_{K\times K}\otimes X_i^{\mathrm{T}}$ dimension of $K\times(Kp)$, $V_i=I_{K\times K}\otimes\eta_i^{\mathrm{T}}$ dimension of $K\times\left(KqJ_n\right)$, $\delta=\left(\delta_{\cdot 1}^{\mathrm{T}},\cdots,\delta_{\cdot k}^{\mathrm{T}}\right)^{\mathrm{T}}$, $\xi_1=\left(\xi_{i1},\xi_{i2},\cdots,\xi_{ik}\right)$, and $\varepsilon_1=\left(\varepsilon_{i1},\varepsilon_{i2},\cdots,\varepsilon_{ik}\right)$. Thus, we can write the FOSR model as,

$$\xi_i=Z_i\gamma+\varepsilon_i. \tag{5}$$

Then, the robust estimator of $\gamma=(\beta,\delta)$ can be obtained by minimizing

$$Q(\gamma)=\sum_{i=1}^{n}\phi_h(\xi_i-Z_i\gamma), \tag{6}$$

where $\phi_h(\cdot)$ is the ESL function. We select variables for all scalar predictors, including linear and nonlinear variables, and group them. We group the variables γ to be selected into, $\gamma=(\gamma_1,\cdots,\gamma_{l+1})$ where each γ_i represents a group of variables. Specifically, when $l=0$, $\gamma_1=(\beta_1,\cdots,\beta_p)$, the effect of all control or confounding variables $X_j, j=1,\cdots,p$, is divided into a group that is either simultaneously selected or simultaneously 0. When $l=1,\cdots,q$, $\gamma_{l+1}=\left(\delta_{l,1,1},\cdots,\delta_{l,k,J_n}\right)$, $l=0,1,\cdots,q$ each nonlinear predictor M_l is divided into a group, with the projection coefficients under the B-spline basis forming a group for variable selection. We use the group penalty function $p_\lambda\left(\|\gamma_l\|_2\right)$ for group variable selection, where $\|\gamma_l\|_2=\sqrt{\sum_{j=1}^{d}\gamma_{l,j}^2}$, where d is the number of variables within the lth group. In this paper, we use the group smoothly clipped absolute deviation (SCAD) regularization method, thus proposing to minimizing:

$$L(\gamma)=\sum_{i=1}^{n}\phi_h(\xi_i-Z_i\gamma)+n\sum_{l=0}^{q}p_\lambda\left(\|\gamma_l\|\right) \tag{7}$$

The SCAD penalty of Fan and Li [11] is defined in the following way,

$$p_\lambda(\theta)=\begin{cases}\lambda|\theta|, |\theta|\le\lambda,\\ -\dfrac{\theta^2-2a\lambda|\theta|+\lambda^2}{2(a-1)}, \lambda<|\theta|\le a\lambda,\\ \dfrac{(a+1)\lambda^2}{2}, |\theta|>a\lambda.\end{cases}$$

for some $a>2$. Where $\|\gamma_j\|$ is the Euclidean norm of γ_j, and $p_\lambda\left(\|\gamma_j\|\right)$ is the penalty function with λ as a regularization parameter. When the loss function $\phi_h(x)=x^2$ is the least square loss, it is the least squares (LS) method proposed by Wang, *et al.* [12]. When the loss function $\phi_h(x)=|x|$ is the least absolute deviation loss, there is a LAD estimation method.

2.2. Theoretical Properties

In this subsection, we discuss the theoretical properties of the proposed estimators. We apply local linear smoothing to the mean function μ. Let $\hat{C}(s,t)$ be the estimated smooth surface of $C(t,s)=\operatorname{cov}\{Y(t),Y(s)\}$ obtained through local quadratic fitting, and h_μ and h_G be the bandwidths of the estimated values $\hat{\mu}$ and $\hat{G}$, respectively. We use the symbol $\|\cdot\|$ to denote the Euclidean norm for vectors and the $L^2(T)$ norm for functions defined on T. Let f' and f'' present the first and second derivatives of f, respectively. We impose the following regularity conditions:

(C1) The mean function $\mu(t)$ and the covariance function $G(s,t)$ are smooth, and the estimation error satisfies: $\sup_{t\in T}\left|\hat{\mu}-\mu(t)\right|=O_p\left(1/\sqrt{nh_\mu}\right)$, $\sup_{t\in T}\left|\hat{G}(s,t)-G(s,t)\right|=O_p\left(1/\sqrt{nh_G^2}\right)$, thus FPCA estimation characteristic function $\hat{\phi}_k(t)$ converge to the real characteristic function: $\sup\left|\hat{\phi}_k(t)-\phi_k(t)\right|=O_p\left(1/\sqrt{nh_\mu}\right)$.

(C2) Error function $\varepsilon(t)$ is a finite variance with a mean of zero, satisfy the $\operatorname{cov}\left(\varepsilon_i(t),\varepsilon_j(t)\right)=0$, $\forall j\neq i$, so that the FPCA score ξ_{ik} satisfies consistency: $\hat{\xi}_{ik}\xrightarrow{p}\xi_{ik}$.

(C3) The matrix $E\left(ZZ^{\mathrm{T}}\right)$ is positively definite and the eigenvalues are bounded, where the dimension of Z is fixed. There exists a positive constant M such that $|Z_j|\leq M$, for all $1\leq j\leq p$.

(C4) The functional coefficient $\delta_l(\cdot)\in C^r[0,1]$ for some integer $r\geq 2$, and the spline order q satisfies $q\geq r$. $C^r[0,1]=\left\{\varphi\middle|\varphi^{(r)}\in C[0,1]\right\}$ means the space of r-th order smooth function.

(C5) The distance between neighboring knots $H_i=\tau_{i+1}-\tau_i$, and $H=\max_{1\leq i\leq 1+N_n}H_i$. For some constant $0<C<\infty$ such that $H/\min_{1\leq i\leq 1+N_n}H_i\leq C$ and $\max_{1\leq i\leq N_n}\{H_{i+1}-H_i\}=o\left(N_n^{-1}\right)$.

(C6) $E\left[\phi_h'(\varepsilon(t))\right]=0$, and $E\left[\phi_h''(\varepsilon(t))\right]>0$, for all t and any $h>0$.

Remark 1 The conditions (C1) and (C2) are the consistency and asymptotic results derived by Yao *et al.* [9] for estimating the FPC score $\hat{\xi}_{ik}$. Therefore, in this paper, the function FPCA scores ξ_{ik} can be used as the target variable, and the estimated coefficients can be projected back using the univariate basis functions and the eigenfunctions of the functional response to recover the influence of the scalar predictor. The conditions (C3)-(C5) are standard in spline estimation literature, where the number of nodes J_n for B-splines should satisfy $\lim_{n\to\infty}J_n/n^{1/(2n+1)}=C$, where C is some nonzero constant. Condition (C6) is in-

troduced by Yao *et al.* [13] as an additional tuning parameter to automatically select based on observed data, ensuring that the objective function has a local minimum.

Theorem 1 Suppose that the conditions (C1)-(C6) are satisfied, if $\lambda_n \to \infty$ and $n^{r/(2r+1)}\lambda_n \to \infty$ as $n \to \infty$, then we have

$$\left\|\hat{\gamma}_i(t) - \gamma_i(t)\right\| = O_p\left(n^{-r/(2r+1)}\right), j = 1, \cdots, p+q.$$

Remark 2 Based on the convergence rate in Theorem 1, we can see that our proposed method provides the optimal rate of convergence $O_p\left(n^{-r/(2r+1)}\right)$ established by Stone [14] for estimating the functional coefficients. When $r=2$, the optimal convergence rate becomes $O_p\left(n^{-2/5}\right)$. In the studies of Fan and Li [11] and Wang *et al.* [7], similar variable selection methods have been proven to be consistent and asymptotically normal. Therefore, the asymptotic property of our method still holds if the above conditions are satisfied.

3. Computational Algorithm and Tuning Parameter Selection

3.1. Computational Algorithm

In this paper, we develop a new iterative algorithm using the local linear approximation (LLA) [15] to the group SCAD method and combining it with the exponential square loss function. The minimization problem of the objective function is solved through iterative reweighted least squares. Since the SCAD penalty function itself is non-convex, the LLA approximation is used to transform it into a weighted LASSO, which penalizes different groups of variables with different weights, and the weights are updated at each iteration, thereby enhancing the robustness of variable selection. The ESL function is applied to the residuals to reduce the effect of outliers on the objective function.

In the LLA method, we perform a local linear approximation on the SCAD penalty function: $p_\lambda\left(\|\gamma_l\|_2\right) \approx p'_\lambda\left(\left\|\gamma_l^{(m)}\right\|_2\right)\|\gamma_l\|_2$, where:

$$w_l = p'_\lambda\left(\left\|\gamma_l^{(m)}\right\|_2\right) = \begin{cases} \lambda, \text{ if } \left\|\gamma_l^{(m)}\right\|_2 \le \lambda, \\ \dfrac{a\lambda - |\gamma|}{a-1}, \text{ if } \lambda < \left\|\gamma_l^{(m)}\right\|_2 \le a\lambda, \\ 0, \text{ if } \left\|\gamma_l^{(m)}\right\|_2 > a\lambda, \end{cases}$$

Thus, the SCAD penalty function is approximated as a weighted group LASSO: $\sum_{l=0}^{q} p_\lambda\left(\|\gamma_l\|_2\right) \approx \sum_{l=0}^{q} w_l \|\gamma_l\|_2$ where w_l is the weight updated at each iteration, let $W_{SCAD} = diag\left(\dfrac{w_1}{\|\gamma_l\|_2}, \cdots, \dfrac{w_l}{\|\gamma_l\|_2}\right)$. Then our objective Function (7) becomes:

$$L(\gamma) = \sum_{i=1}^{n} \phi_h\left(\xi_i - Z_i\gamma\right) + n\sum_{l=0}^{q} w_l \|\gamma_l\|_2 \tag{8}$$

Derive this equation with respect to γ and set it to 0, resulting in the following equation:

$$-\sum_{i=1}^{n} Z_i^{\mathrm{T}} \phi_h'\left(\xi_i - Z_i\gamma\right) + nW_{SCAD}\gamma = 0. \tag{9}$$

Under ESL loss, let the weight matrix W_{ESL} be calculated from the derivatives of ESL: $w_i = \frac{\phi_h'(r_i)}{r_i} = \frac{2}{h}\exp\left(-\frac{r_i^2}{h}\right)$, where $r_i = \xi_i - Z_i^{\mathrm{T}}\gamma$, $i = 1, \cdots, n$. With $W_{ESL} = diag\left(w_1, \cdots, w_n\right)$. Then Equation (9) becomes:

$$\left(Z^{\mathrm{T}} W_{ESL} Z + nW_{SCAD}\right)\gamma = Z^{\mathrm{T}} W_{ESL}\xi \tag{10}$$

Update γ as in each iteration step:

$$\hat{\gamma} = \left(Z^{\mathrm{T}} W_{ESL} Z + nW_{SCAD}\right)^{-1} Z^{\mathrm{T}} W_{ESL}\xi.$$

Therefore, the iterative calculation process for $\hat{\gamma} = \left(\hat{\gamma}_0^{\mathrm{T}}, \cdots, \hat{\gamma}_q^{\mathrm{T}}\right)^{\mathrm{T}}$ is as follows:

Step 1: Obtain initial values using least squares estimation $\hat{\gamma}^{(0)} = \left(Z^{\mathrm{T}} Z\right)^{-1} Z^{\mathrm{T}}\xi$.

Step 2: Given $\hat{\gamma}^{(m)}$: Calculate the residual weight matrix $\hat{W}^{(m)}$, where the residual is $\hat{\gamma}_i^{(m)} = \xi_i - Z_i^{\mathrm{T}}\hat{\gamma}^{(r)}$.

Step 3: Given $\hat{\gamma}^{(m)}$, calculate the LLA approximation weights for SCAD $\hat{w}^{(m)}$.

Step 4: Update $\hat{\gamma}^{(m+1)}$ through iterative reweighted least squares: $\hat{\gamma}^{(m+1)} = \left(Z^{\mathrm{T}} W_{ESL} Z + nW_{SCAD}\right)^{-1} Z^{\mathrm{T}} W_{ESL}\xi$.

Step 5: Repeat steps 2 to 4 until convergence. If $\left\|\hat{\gamma}^{(m+1)} - \hat{\gamma}^{(m)}\right\|$ is less than a predetermined threshold, the final estimate of γ is obtained.

3.2. The Choice of Tuning Parameters

The parameter K is the number of eigenbases in the function principal component analysis, and the parameter J is the number of B-spline bases used to approximate the nonlinear parameters as a linear model. To select the number of these two basis functions more robustly, we use the following weighted generalized cross-validation (GCV) statistic:

$$GCV\left(K, J\right) = \frac{\left(\xi - Z\eta\right)^{\mathrm{T}} \hat{W}\left(Y - Z\eta\right)/nk}{\left(1 - t\left(K, J\right)/nk\right)^2},$$

where $t\left(K, J\right) = trace\left\{Z\left(Z^{\mathrm{T}}\hat{W}Z\right)^{-1} Z^{\mathrm{T}}\hat{W}\right\}$. Thus, the optimal number of K and J is obtained by $\left(\hat{K}, \hat{J}\right) = \min GCV\left(K, J\right)$. In order to achieve consistency in variable selection, we use the extended Bayesian information criterion (EBIC) [16] of a corresponding Gaussian likelihood for choosing the penalty parameter λ. For the parameter a in the SCAD penalty function, following the suggestion of Fan and Li [11], we adopt $a = 3.7$ for implementation. The tuning parameter h impacts the efficiency of the proposed estimator. To attain high efficiency, we adopt the approach of Yao *et al.* [13] to select h.

4. Simulation and Application

4.1. Simulation Studies

In this subsection, we implement several simulation studies to investigate the fi-

nite sample performance of the proposed method. We take the sample sizes of $n=50,100,200$ and generate the data from the nonlinear FOSR model:

$$Y_i(t)=\mu(t)+\sum_{j=1}^{2}X_{ij}\beta_j(t)+\sum_{l=1}^{20}\theta_l(M_{il},t)+\varepsilon_i(t),$$

Here, we have the coefficient functions given by $\mu(t)=8\sin(\pi t/50)$, $\beta_1(t)=3+5t/100$, $\beta_2(t)=4\sin(\pi t/50)+4\cos(\pi t/50)$, the nonparametric functions are given by $\theta_1(x,t)=2x^3\exp(t/100)$, $\theta_2(x,t)=5(x+x^3)\sin(\pi t/100)$, $\theta_3(x,t)=16\sin(xt/50)$. The rest of the nonlinear functions $\theta_l(x,t)$ are set to zero. The exogenous covariates $X_{ij}\sim Unif(-2j,-2i)$, and the predictors of interest $M_{ij}\sim N(0,1^2)$. The error process $\varepsilon_i(t)$ is generated as $\varepsilon_i(t)=\xi_{i1}\cos(t)+\xi_{i2}\sin(t)+N(0,1^2I_m)$, where $\xi_{i1}\sim N(0,0.5^2)$ and $\xi_{i2}\sim N(0,0.75^2)$. The response $Y_i(t)$ is observed on a grid of $m=80$. The values of K and J (number of B-spline basis) were chosen from a two-dimensional grid with $K\in\{3,4,5,6,7\}$ and $J\in\{5,7,9\}$.

In order to examine the accuracy of the proposed method, we use the standard deviation of the integrated squared error (ISE), which is defined as $\mathrm{ISE}=\sum_{j=1}^{p}\int_T\{\hat{\eta}_j(t)-\eta_j(t)\}^2\,\mathrm{d}t$. Besides, we use the mean square prediction error (MSPE) to assess the accuracy of prediction. MSPE is given by $\mathrm{MSPE}=\frac{1}{N}\sum_{i=1}^{N}\int_T\left\{\hat{Y}_i^*(t)-\sum_{j=1}^{p}\eta_j(t)Z_{ij}^*\right\}^2\mathrm{d}t$, where the independent samples $Z_i=[X_i,M_i]$ are generated for the test set, $\hat{Y}_i^*(t)=\sum_{j=1}^{2}X_{ij}\beta_j(t)+\sum_{l=1}^{20}\theta_l(M_{il},t)$ are estimated from the training data.

To demonstrate the effectiveness of the algorithm described in Section 4, we generated data with a sample size of $n=100$ and used cubic splines of order $q=4$ to approximate the nonlinear predictor variables. The tuning parameters K,J and h selected by the proposed process are approximately 5, 7, and 3.5, respectively.

In order to show the robustness of the proposed approach, we simulated three types of outlier contamination: outliers in the function response, outliers in the predictor variables, and outliers in both the function response and scalar predictor. To introduce outliers in the function response, we assumed that 10% of the response curves of the original sample were contaminated by outlier curves, denoted as $\hat{Y}_i^*(t)$, so that the contaminated function response data $Y_i^0(t)=(1-R_i)Y_i(t)+R_iY_i^*(t)$ is treated as a numerical simulation, where R_i is a Bernoulli random variable such that $P(R_i=1)=0.1$. To introduce outliers in the predictor variables, we randomly selected subsets of the predictors from the original sample and contaminated them with peaks values. The detail mechanisms for producing the outlier response $Y_i^*(t)$ and outliers in scalar predictors are described as follows:

Scenario 1. No outliers. In this setting, $Y_i^*(t)=Y_i(t)$

Scenario 2. When the outlier is in the function response. The outlier curve $Y_i^*(t)$ is generated as $Y_i^*(t_k) = Y_i(t_k) + r_{ik} I_{[U_i, U_i+v]}(t_k), k = 1, \cdots, m$, where r_{ik} are random values taken from a uniform distribution on $[-r_u, -r_l] \cup [r_l, r_u]$ and $I(\cdot)$ is the indicator function, $[U_i, U_i + v]$ be an interval in $[0,1]$ for the ith subject, where U_i is randomly chosen from $[0, 1-v]$, and v is a constant. The r_{ik} control the strength of outliers while the constant v determines the contamination length of each outlier curve. We set $r_l = 3$, $r_u = 10$ and $v = 0.5$ in the scenario.

Scenario 3. When the outliers are in the scalar predictors. We randomly select 5% of the predictors that are contaminated by the high leverage outliers $\left(X_{i1}^*, \cdots, X_{ip}^*\right) = \left(X_{i1}, \cdots, X_{ip}\right) + 5$.

Scenario 4. When the outliers are both in the function response and in the scalar predictor. We consider the contaminated error distribution $0.8N\left(0, 0.5^2\right) + 0.2Cauchy(0,1)$, which is used to produce outliers with a heavy-tailed error distribution for data set.

For each scenario, our method is compared with the other two methods. Since the method proposed in this paper is based on the exponential square loss function $\phi_h(t) = 1 - \exp\left(-t^2/h\right)$, our model is called ESL. When $\phi_h(x)$ is the square loss function $\phi_h(x) = x^2$, it is the least squares (LS) method proposed by Wang *et al.* [12]. When $\phi_h(x) = |x|$, it leads to the least absolute deviation method (LAD). For each scenario, 100 repetitions are used, and the numerical results are recorded in the following four tables. First, as shown in **Table 1**, when there are no outliers in the dataset and the error distribution is normal, the ISE and MSPE values of the ESL estimator are close to those of the LS estimator. This means that for datasets without outliers, the performance of these two methods is comparable. Moreover, the LS estimator performs best because its sample mean and the standard deviations of ISE and MSPE are smaller than those of the proposed ESL and LAD estimators. Second, from **Tables 2-4**, we can see that the ESL estimator produces smaller ISE and MSPE values than the LS and LAD estimators. This indicates that for datasets with outliers, the proposed ESL method outperforms the LS and LAD methods. This is mainly because when there are some very large outliers in the dataset, the proposed ESL method places more weight on the "most likely" data around the true value, thereby achieving a robust estimator. Finally, the ISE and MSPE values of the ESL estimator decrease with increasing sample size n, which confirms the consistency result established in Theorem 1. In summary, the proposed ESL method has certain advantages in variable selection and estimation.

4.2. Application

In this subsection, we apply the proposed method to analyze the real DTI dataset collected from the NIH Alzheimer's Disease Neuroimaging Initiative (ADNI) study, which is available at http://adni.loni.usc.edu/. The DTI data includes 213 subjects. For each subject, we calculate the fractional anisotropy (FA) curve along the midsagittal skeleton of the corpus callosum at m = 83 positions, as shown in

Table 1. Simulation results of the averages ($\times 10^{-2}$) and the standard deviations ($\times 10^{-2}$) of integrated squared error (ISE) and mean square prediction error (MSPE) under Scenario 1.

n	Methods	ISE		MSPE	
		Aver.	std	Aver.	std
50	ESL	0.532	0.214	1.763	0.473
	LS	0.438	0.186	1.516	0.432
	LAD	0.732	0.278	1.928	0.675
100	ESL	0.273	0.106	0.316	0.127
	LS	0.258	0.103	0.308	0.119
	LAD	0.376	0.145	0.531	0.193
200	ESL	0.124	0.032	0.078	0.026
	LS	0.121	0.032	0.076	0.025
	LAD	0.183	0.054	1.117	0.041

Table 2. Simulation results of the averages ($\times 10^{-2}$) and the standard deviations ($\times 10^{-2}$) of integrated squared error (ISE) and mean square prediction error (MSPE) under Scenario 2.

n	Methods	ISE		MSPE	
		Aver.	std	Aver.	std
50	ESL	0.574	0.203	1.837	0.413
	LS	3.486	2.693	13.723	7.516
	LAD	0.904	0.523	3.841	1.967
100	ESL	0.241	0.098	0.564	0.139
	LS	2.773	1.392	5.103	2.877
	LAD	0.339	0.124	1.072	0.438
200	ESL	0.134	0.045	0..341	0.113
	LS	1.747	1.219	2.914	1.485
	LAD	0.202	0.074	0.517	0.149

Table 3. Simulation results of the averages ($\times 10^{-2}$) and the standard deviations ($\times 10^{-2}$) of integrated squared error (ISE) and mean square prediction error (MSPE) under Scenario 3.

n	Methods	ISE		MSPE	
		Aver.	std	Aver.	std
50	ESL	0.728	0.165	2.137	0.413
	LS	87.72	31.46	486.7	103.2
	LAD	14.47	16.82	63.59	89.74
100	ESL	0.341	0.105	0.536	0.153
	LS	83.45	16.82	253.6	41.67
	LAD	12.97	8.183	39.26	23.92
200	ESL	0.179	0.063	0.218	0.072
	LS	76.81	11.74	165.3	17.23
	LAD	11.34	3.769	14.35	6.138

Table 4. Simulation results of the averages ($\times 10^{-2}$) and the standard deviations ($\times 10^{-2}$) of integrated squared error (ISE) and mean square prediction error (MSPE) under Scenario 4.

n	Methods	ISE		MSPE	
		Aver.	std	Aver.	std
50	ESL	0.931	0.273	2.928	0.773
	LS	673.5	703.4	2156	2013
	LAD	32.18	132.6	107.3	388.5
100	ESL	0.452	0.156	0.763	0.248
	LS	756.1	514.9	1235	891.4
	LAD	15.14	14.42	42.37	28.36
200	ESL	0.237	0.083	0.153	0.076
	LS	548.6	413.7	409.6	314.9
	LAD	12.16	6.297	13.54	9.433

Figure 1. We refer to Smith *et al.* [17] for a detailed description of the application of this data.

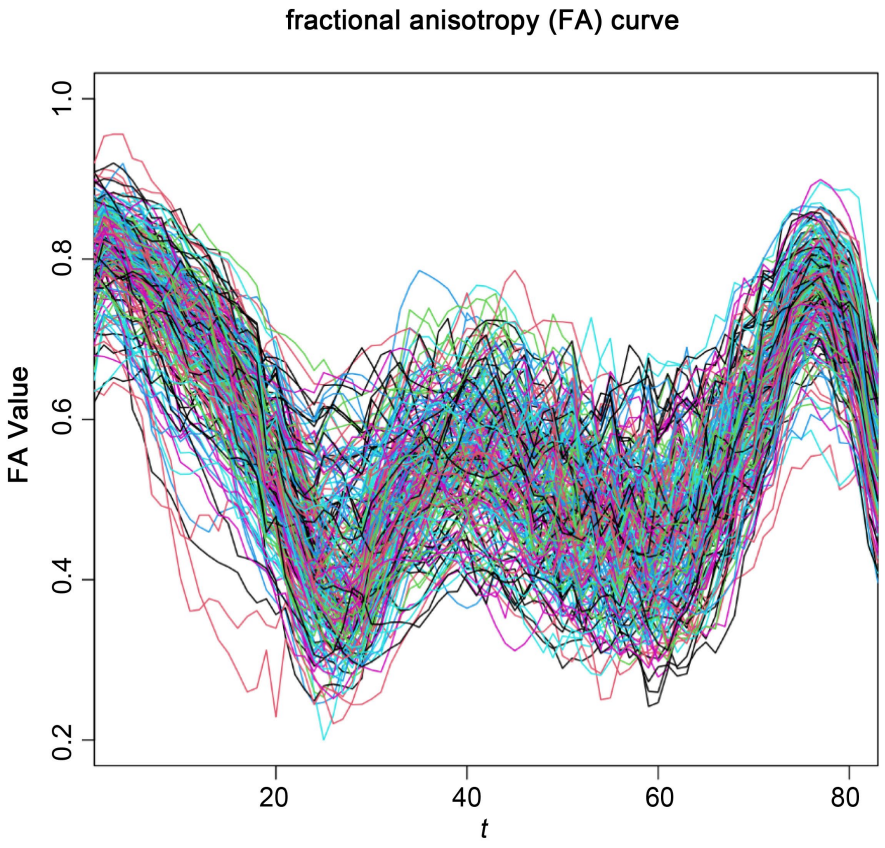

Figure 1. For 213 subjects, fractional anisotropy (FA) curves were calculated for $m = 83$ locations along the sagittal skeleton in the corpus callosum.

We applied the model proposed in this paper, $Y_i(t)=\mu(t)+\sum_{j=1}^{p}X_{ij}\beta_j(t)+\sum_{l=1}^{q}\theta_l(M_{il},t)+\varepsilon_i(t)$, to this dataset. Here, X_{ij} are the categorical predictors (dummy encoded) with linear effects $\beta_j(t)$, and M_{il} are the continuous predictors. In this study, we regarded gender (coded by a dummy variable indicating males), handedness (coded by a dummy variable indicating left-handedness), education level, Alzheimer's disease (AD) status, and mild cognitive impairment (MCI) status as categorical predictors, and age as a continuous variable. We centered all the functional responses and scalar predictors at zero mean. Then, we use the proposed ESL method to analyze the ADNI data, which approximates the nonlinear scalar function using cubic splines of order $q=4$. The first five panels of Figure 2 show the estimated function coefficients of these five linear variables. We note that the function coefficient corresponding to gender takes positive values for the most part, while the functional coefficients of the other four predictors take negative values. This indicates that males tend to have higher FA values than females, and that patients with AD and MCI have lower FA values. Left-handedness or higher education level may lead to smaller FA values. The last graph shows the effect of continuous predicted age on FA values, which decrease with age and decline more rapidly than FA values during normal aging. These results are consistent with the previous analysis by Cai *et al.* [18].

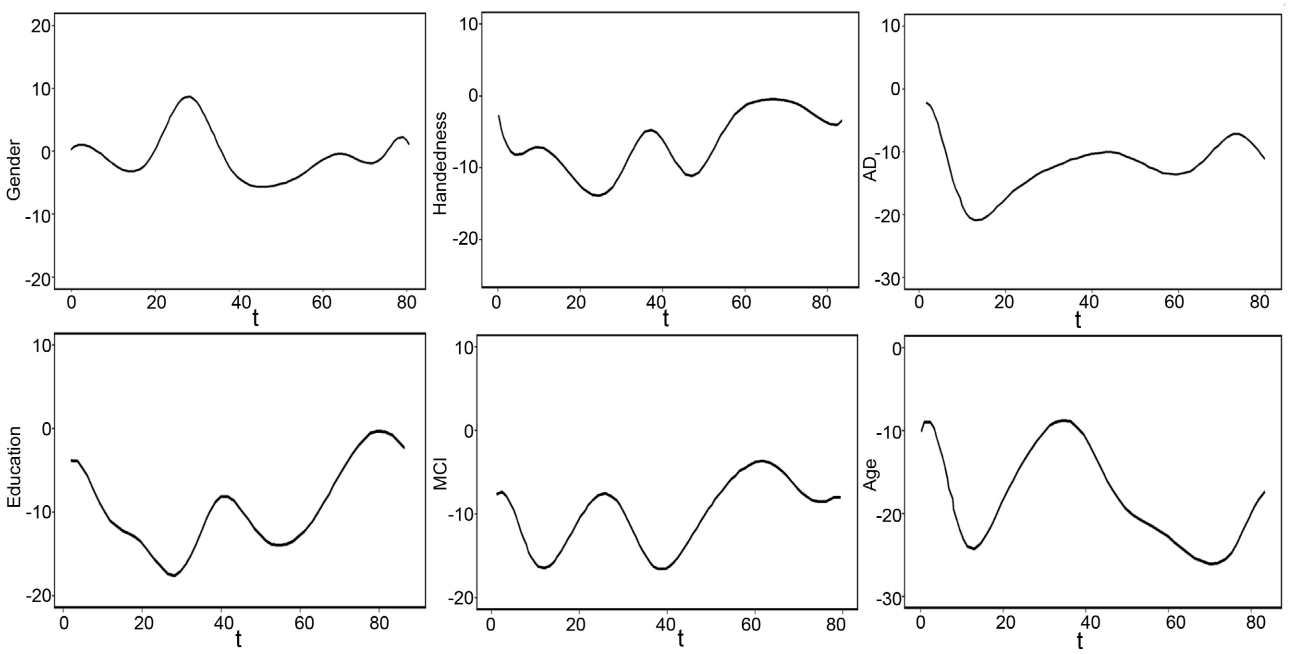

Figure 2. Estimated functional coefficients ($\times 10^{-3}$) for gender, handedness, education, AD, age and MCI for the ADNI data.

To evaluate the predictive performance of the model, we randomly divided 213 samples into a training set containing 149 samples and a testing set containing 64 samples. We use the training set to estimate the function coefficients and then predict the response in the test set. We use mean squared prediction error (MSPE)

to verify the quality of the prediction, and the smaller the MSPE, the better the prediction. Based on 100 random splits, we obtained LS estimation average MSPEs and their standard deviations of 0.74×10^{-2} and 0.21×10^{-2}, respectively. We obtained the average MSPEs and standard deviation of our proposed ESL estimator, which are 0.69×10^{-2} and 0.18×10^{-2}, respectively. These results indicate that our proposed method is preferred when considering the robustness of predictive performance.

5. Conclusions

In order to better explain the behavior of the function response over time, this paper develops a robust variable selection procedure for the nonlinear FOSP model containing continuous type. The function response is projected into a low-dimensional principal component space by using functional principal component analysis, and the principal component scores are used for variable selection and regression modeling. A robust variable selection for all scalar predictors is achieved by combining the ESL function with the group SCAD regularization method. To implement the computational procedure of the proposed method, we propose an iterative algorithm based on the LLA method. The tuning parameters are selected through a data driven program to achieve the robustness of variable selection. Simulation studies and real data analysis show that when there are outliers in the function response or scalar predictors, the proposed method can select the relevant predictors.

There are still some problems worth solving in the future research. In this paper, all linear covariates are divided into a group, which can either be selected or rejected at the same time. This condition is too strict and has limitations in practical application. Therefore, the variable selection of paper for scalar predictors with nonlinear covariates needs further study. When nonlinear covariates have high dimensional characteristics, robust variable selection for functional response regression is a significant topic for continued research.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] Barber, R.F., Reimherr, M. and Schill, T. (2017) The Function-on-Scalar LASSO with Applications to Longitudinal GWAS. *Electronic Journal of Statistics*, **11**, 1351-1389. https://doi.org/10.1214/17-ejs1260

[2] Chen, Y., Goldsmith, J. and Ogden, R.T. (2016) Variable Selection in Function-on-Scalar Regression. *Stat*, **5**, 88-101. https://doi.org/10.1002/sta4.106

[3] Goldsmith, J., Liu, X., Jacobson, J.S. And Rundle, A. (2016) New Insights into Activity Patterns in Children, Found Using Functional Data Analyses. *Medicine & Science in Sports & Exercise*, **48**, 1723-1729. https://doi.org/10.1249/mss.0000000000000968

[4] Fan, Z. and Reimherr, M. (2017) High-Dimensional Adaptive Function-on-Scalar Regression. *Econometrics and Statistics*, **1**, 167-183.

https://doi.org/10.1016/j.ecosta.2016.08.001

[5] Parodi, A. and Reimherr, M. (2018) Simultaneous Variable Selection and Smoothing for High-Dimensional Function-On-Scalar Regression. *Electronic Journal of Statistics*, **12**, 4602-4639. https://doi.org/10.1214/18-ejs1509

[6] Ghosal, R. and Maity, A. (2021) Variable Selection in Nonlinear Function-on-Scalar Regression. *Biometrics*, **79**, 292-303. https://doi.org/10.1111/biom.13564

[7] Wang, X., Jiang, Y., Huang, M. and Zhang, H. (2013) Robust Variable Selection with Exponential Squared Loss. *Journal of the American Statistical Association*, **108**, 632-643. https://doi.org/10.1080/01621459.2013.766613

[8] Ash, R.B. and Gardner, M.F. (1978) Topics in Stochastic Processes. Academic Press.

[9] Yao, F., Müller, H. and Wang, J. (2005) Functional Data Analysis for Sparse Longitudinal Data. *Journal of the American Statistical Association*, **100**, 577-590. https://doi.org/10.1198/016214504000001745

[10] Hall, P. and Hosseini-Nasab, M. (2005) On Properties of Functional Principal Components Analysis. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, **68**, 109-126. https://doi.org/10.1111/j.1467-9868.2005.00535.x

[11] Fan, J. and Li, R. (2001) Variable Selection via Nonconcave Penalized Likelihood and Its Oracle Properties. *Journal of the American Statistical Association*, **96**, 1348-1360. https://doi.org/10.1198/016214501753382273

[12] Wang, L., Chen, G. and Li, H. (2007) Group SCAD Regression Analysis for Microarray Time Course Gene Expression Data. *Bioinformatics*, **23**, 1486-1494. https://doi.org/10.1093/bioinformatics/btm125

[13] Yao, W., Lindsay, B.G. and Li, R. (2012) Local Modal Regression. *Journal of Nonparametric Statistics*, **24**, 647-663. https://doi.org/10.1080/10485252.2012.678848

[14] Stone, C.J. (1982) Optimal Global Rates of Convergence for Nonparametric Regression. *The Annals of Statistics*, **10**, 1040-1053. https://doi.org/10.1214/aos/1176345969

[15] Zou, H. and Li, R. (2008) One-Step Sparse Estimates in Nonconcave Penalized Likelihood Models. *The Annals of Statistics*, **36**, 1509-1533. https://doi.org/10.1214/009053607000000802

[16] Chen, J. and Chen, Z. (2008) Extended Bayesian Information Criteria for Model Selection with Large Model Spaces. *Biometrika*, **95**, 759-771. https://doi.org/10.1093/biomet/asn034

[17] Smith, S.M., Jenkinson, M., Johansen-Berg, H., Rueckert, D., Nichols, T.E., Mackay, C.E., *et al.* (2006) Tract-Based Spatial Statistics: Voxelwise Analysis of Multi-Subject Diffusion Data. *NeuroImage*, **31**, 1487-1505. https://doi.org/10.1016/j.neuroimage.2006.02.024

[18] Cai, X., Xue, L., Pu, X. and Yan, X. (2020) Efficient Estimation for Varying-Coefficient Mixed Effects Models with Functional Response Data. *Metrika*, **84**, 467-495. https://doi.org/10.1007/s00184-020-00776-0