The Canadian Journal of Statistics
Vol. 50, No. 1, 2022, Pages 162–179
La revue canadienne de statistique

Robust estimation and variable selection for function-on-scalar regression

Xiong CAI[1], Liugen XUE[2], and Jiguo CAO[3*], for the Alzheimer's Disease Neuroimaging Initiative[†]

[1]*School of Statistics and Mathematics, Nanjing Audit University, Nanjing 211815, China*
[2]*College of Statistics and Data Science, Faculty of Science, Beijing University of Technology, Beijing 100124, China*
[3]*Department of Statistics and Actuarial Science, Simon Fraser University, Burnaby, BC V5A1S6, Canada*

Key words and phrases: Functional data analysis; group SCAD; oracle property; robust estimation; variable selection.

MSC 2020: Primary 62H99; secondary 62G08.

Abstract: Function-on-scalar regression is commonly used to model the dynamic behaviour of a set of scalar predictors of interest on the functional response. In this article, we develop a robust variable selection procedure for function-on-scalar regression with a large number of scalar predictors based on exponential squared loss combined with the group smoothly clipped absolute deviation regularization method. The proposed procedure simultaneously selects relevant predictors and provides estimates for the functional coefficients, and achieves robustness and efficiency using tuning parameters selected by a data-driven procedure. Under reasonable conditions, we establish the asymptotic properties of the proposed estimators, including estimation consistency and the oracle property. The finite-sample performance of the proposed method is investigated with simulation studies. The proposed method is also demonstrated with a real diffusion tensor imaging data example. *The Canadian Journal of Statistics* 50: 162–179; 2022 © 2021 Statistical Society of Canada

Résumé: La régression fonction sur scalaire est communément utilisée pour modéliser le lien entre une variable réponse fonctionnelle et un ensemble de prédicteurs scalaires d'intérêt. C'est dans ce cadre que les auteurs de ce travail proposent une procédure de sélection de variables lorsque le nombre de prédicteurs scalaires est grand. La procédure en question est robuste et fait usage de l'exponentielle de l'erreur quadratique couplée avec une méthode de régularisation basée sur une pénalité lisse coupée de la déviation absolue (SCAD). En choisissant des paramètres d'ajustement basés sur les données, cette procédure permet de simultanément sélectionner les prédicteurs pertinents et d'estimer les coefficients fonctionnels de manière robuste et efficace. Le comportement asymptotique des estimateurs proposés, dont la convergence et la propriété d'oracle, est exploré sous des conditions de régularité raisonnables. Quant à leurs propriétés à distance finie, elles sont illustrées à laide de simulations numériques et un exemple pratique de données d'imagerie de diffusion par tenseurs. *La revue canadienne de statistique* 50: 162–179; 2022 © 2021 Société statistique du Canada

Additional Supporting Information may be found in the online version of this article at the publisher's website.
[*] *Corresponding author: jiguo_cao@sfu.ca*

[†]Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf.

1. INTRODUCTION

With the development of technology, a number of approaches in biomedical and life sciences involve dynamic measurements. As a result, the collected data are not limited to isolated observations but are functions over space and/or time, such as diffusion tensor imaging (DTI) and positron emission tomography (Zhu et al., 2007; Friston, 2009). This practical problem inspires us to capture the dynamic behaviour of a set of scalar predictors of interest on the functional response. Function-on-scalar regression, which characterizes the relationship between a functional response and a set of scalar predictors, is an integral part of functional data analysis (Ramsay & Silverman, 2005; Ferraty & Vieu, 2006; Cao & Ramsay, 2010; Ainsworth, Routledge & Cao, 2011; Liu, Wang & Cao, 2017; Lin et al., 2017; Guan, Lin & Cao, 2020; Jiang et al., 2020; Cai, Xue & Cao, 2021). Function-on-scalar regression has become increasingly popular in the analysis of gene expression data and imaging data (see, e.g., Wang, Chen & Li, 2007; Li, Huang & Zhu, 2017).

Let $Y(t)$ be a functional response defined on a closed interval $\mathcal{T}$, and $\boldsymbol{X} = (X_1, \dots, X_p)^T$ be a p-dimensional vector of scalar predictors. A function-on-scalar model is given as

$$Y(t) = \boldsymbol{X}^T \boldsymbol{\beta}(t) + \epsilon(t), \quad t \in \mathcal{T}, \tag{1}$$

where $\boldsymbol{\beta}(t) = (\beta_1(t), \dots, \beta_p(t))^T$ is a p-dimensional vector of functional coefficients, and $\epsilon(t)$ is a mean-zero random error function independent of $\boldsymbol{X}$. The model given by Equation (1) has been extensively studied and developed for longitudinal and functional data. A large number of studies in the literature have explored various approaches to estimate the functional coefficients and to construct test statistics for hypotheses regarding $\boldsymbol{\beta}(\cdot)$ (see, e.g., Chiang, Rice & Wu, 2001; Zhang & Chen, 2007; Zhu, Li & Kong, 2012; Li, Huang & Zhu, 2017; Cai et al., 2020; and the references therein). In practice, it is common to collect data with a large number of predictors, in which case some of the predictors may have no effect on the functional response. Therefore, with high-dimensional predictors, it is important to identify and incorporate only relevant scalar predictors into function-on-scalar regressions to enhance model prediction and interpretability. Under a framework of parametric models with only scalar variables, various regularization methods have been developed for variable selection, including LASSO (Tibshirani, 1996), SCAD (Fan & Li, 2001) and minimax concave penalty (MCP; Zhang, 2010).

The literature on variable selection for function-on-scalar regression models is growing. Wang, Chen & Li (2007) introduced a group SCAD penalized estimation procedure for variable selection in the context of function-on-scalar regression and applied their proposed procedure to analyze real gene expression data. Chen, Goldsmith & Ogden (2016) incorporated within-subject correlation into function-on-scalar models and proposed a variable selection procedure via group MCP regularization coupled with the generalized least squares technique. Barber, Reimherr & Schill (2017) extended the group LASSO method to the model given by Equation (1) and proposed a function-on-scalar LASSO (FSL) procedure. Fan & Reimherr (2017) further developed an adaptive FSL procedure for the model given by Equation (1). Parodi & Reimherr (2018) presented a functional linear adaptive mixed estimation method to achieve variable selection and smoothing for function-on-scalar regression.

However, to our knowledge, the variable selection methods discussed above are based on the least squares criterion. It is well known that the least squares method is sensitive to outliers in the data, and hence the efficiency of methods based on least squares in variable selection can be seriously affected by outliers. Therefore, it is desirable to substitute the least squares criterion with one that is more robust to the presence of outliers. Many outlier-resistant loss functions can be used to conduct robust estimation and variable selection, such as the least absolute deviation loss, the quantile loss or Huber's loss; see Maronna, Martin & Yohai

(2006) for a detailed introduction. Here, we consider the exponential squared loss (ESL) introduced by Wang et al. (2013) in the context of the function-on-scalar regression. The ESL function is defined as $\phi_h(x) = 1 - \exp(-x^2/h)$, where h is a tuning parameter which controls the degree of robustness of the estimators. ESL is a bounded continuous function and we can control the impact of outliers that cause relatively large errors by selecting a proper tuning parameter h. When h is large, $\phi_h(x) \approx x^2/h$, and therefore the proposed estimator is similar to the least squares estimator in the extreme case. For a small h, large absolute values of x will have a small impact on the loss function $\phi_h(x)$ and further lead to a small impact on the estimator. In other words, a small h can restrict the influence of outliers on the estimator.

In this article, we employ ESL in the model given by Equation (1) and present a robust variable selection procedure using the group SCAD regularization method, where the functional coefficients are approximated by B-spline basis functions. The proposed procedure simultaneously selects relevant predictors and provides estimates for the functional coefficients at their best convergence rates. An iterative algorithm based on the local quadratic approximation (LQA) procedure is provided for implementing the proposed method. Our method can achieve robustness and efficiency using the appropriate tuning parameters, which are selected by a data-driven procedure. The oracle property of the proposed method is also established. Simulation studies and analysis of real DTI data are conducted to illustrate the finite-sample performance of the proposed method. Simulation results show that our method is robust to outliers, performing much better than the least absolute deviation estimator and least squares estimator when there are outliers in the dataset. Meanwhile, the proposed estimator works comparably to the least squares estimator in the absence of outliers.

The rest of the article is organized as follows. In Section 2, we describe the estimation method for function-on-scalar regression. In Section 3, we establish the consistency and oracle property of the proposed method. In Section 4, we present the implementation algorithm and tuning parameter selection. In Section 5, we report and compare simulation results. In Section 6, we illustrate the proposed method through an analysis of DTI data. Section 7 summarizes the conclusions of the article. The proofs are given in the Supplementary Material.

2. ESTIMATION METHOD

Suppose that $\{(Y_i(t), \boldsymbol{X}_i), t \in \mathcal{T}, i = 1, \ldots, n, \}$ is a random sample from the population $\{(Y(t), \boldsymbol{X}), t \in \mathcal{T}\}$. In this article, we consider a predictor set with fixed dimension p. Without loss of generality, we set $\mathcal{T} = [0, 1]$ and assume that only the first d scalar predictors are relevant while the rest are not, that is, $\beta_j(t) \equiv 0$ for $j = d+1, \ldots, p$. We first represent the functional coefficients $\beta_j(t)$ using B-spline basis functions. Let $\boldsymbol{B}_q(t) = (B_{s,q}(t) : 1 \leq s \leq q + N_n)^T$ denote the qth-order B-spline basis functions with knot sequence $\{\tau_s\}$ satisfying $\tau_1 = \cdots = 0 = \tau_q < \tau_{q+1} < \cdots < \tau_{q+N_n} < 1 = \tau_{N_n+q+1} = \cdots = \tau_{N_n+2q}$, where N_n is the number of interior knots. For $q \leq s \leq q + N_n$, let $H_s = \tau_{s+1} - \tau_s$ be the distance between neighbouring knots, and let $H = \max_{q \leq s \leq q+N_n} H_s$. To study the asymptotic properties of the spline estimator of $\beta_j(\cdot)$, we assume that $\max_{q \leq s \leq q+N_n-1} \left|H_{s+1} - H_s\right| = o(N_n^{-1})$ and $H / \min_{q \leq s \leq q+N_n} H_s \leq C$ for some constant $0 < C < \infty$. The above assumption on the distances between neighbouring knots is typical in polynomial spline regression literature (see, e.g., Huang, 2003). Let $J_n = N_n + q$. Then, the functional coefficients $\beta_j(t)$ can be approximated by the B-spline basis functions

$$\beta_j(t) \approx \sum_{s=1}^{J_n} B_{s,q}(t)\gamma_{js} = \boldsymbol{B}_q^T(t)\boldsymbol{\gamma}_j, \quad j = 1, \ldots, p, \tag{2}$$

where $\boldsymbol{\gamma}_j = (\gamma_{j1}, \ldots, \gamma_{jJ_n})^T$.

In practice, functions are observed on a set of discrete grids. For simplicity, we assume that $Y_i(t)$ is observed at the same dense points $t_1 = 0 \leq \cdots \leq t_m = 1$, for all i. Let $Y_{ik} = Y_i(t_k)$ be the kth observation for the ith subject, and $\epsilon_{ik} = \epsilon_i(t_k)$. Following Equations (1) and (2), we have

$$Y_{ik} \approx \sum_{j=1}^{p} \sum_{s=1}^{J_n} X_{ij} B_{s,q}(t_k) \gamma_{js} + \epsilon_{ik}.$$

Then, the robust estimator of $\boldsymbol{\gamma} = (\boldsymbol{\gamma}_1^T, \ldots, \boldsymbol{\gamma}_p^T)^T$ can be obtained by minimizing

$$Q(\boldsymbol{\gamma}) = \sum_{i=1}^{n} \sum_{k=1}^{m} \phi_h \left(Y_{ik} - \sum_{j=1}^{p} \sum_{s=1}^{J_n} X_{ij} B_{s,q}(t_k) \gamma_{js} \right),$$

where $\phi_h(\cdot)$ is the ESL function.

Note that setting $\beta_j(t) = 0$ is equivalent to setting all the entries of $\boldsymbol{\gamma}_j$ to zero. To achieve variable selection, we treat each coefficient vector $\boldsymbol{\gamma}_j$ as a group and adopt the general form of a group penalty, $\sum_{j=1}^{p} p_{\lambda_n}(\|\boldsymbol{\gamma}_j\|)$, where $\|\boldsymbol{\gamma}_j\|$ is the Euclidean norm of $\boldsymbol{\gamma}_j$, and $p_{\lambda_n}(\cdot)$ is the penalty function with λ_n as a regularization parameter. There exist various penalty functions for conducting variable selection. In this article, we consider the SCAD penalty of Fan & Li (2001), defined as

$$p_\lambda(\theta) = \begin{cases} \lambda|\theta|, & |\theta| \leq \lambda, \\ -\left(\theta^2 - 2a\lambda|\theta| + \lambda^2\right)/[2(a-1)], & \lambda < |\theta| \leq a\lambda, \\ (a+1)\lambda^2/2, & |\theta| > a\lambda \end{cases}$$

for some $a > 2$. The SCAD penalty function possesses some desirable properties: it enables us to obtain consistent variable selection and the oracle property, and it results in estimates that are almost unbiased for large coefficients. Following the suggestion of Fan & Li (2001), we adopt $a = 3.7$ for implementation.

To achieve robust estimation and variable selection, we propose minimizing

$$\mathcal{L}(\boldsymbol{\gamma}) = \sum_{i=1}^{n} \sum_{k=1}^{m} \phi_h \left(Y_{ik} - \sum_{j=1}^{p} \sum_{s=1}^{J_n} X_{ij} B_{s,q}(t_k) \gamma_{js} \right) + nm \sum_{j=1}^{p} p_{\lambda_n}(\|\boldsymbol{\gamma}_j\|) \tag{3}$$

with respect to $\boldsymbol{\gamma}$. It is important to note that when we set $\phi_h(\cdot)$ as the squared loss function, i.e., $\phi_h(x) = x^2$, the minimizer of Equation (3) gives the least squares group SCAD (LS-gSCAD) estimator (Wang, Chen & Li, 2007). When $\phi_h(x) = |x|$, it leads to the least absolute deviation group SCAD (LAD-gSCAD) estimator. Let $\hat{\boldsymbol{\gamma}} = \left(\hat{\boldsymbol{\gamma}}_1^T, \ldots, \hat{\boldsymbol{\gamma}}_p^T\right)^T$ be the solution to minimizing Equation (3). Then, the penalized robust estimators of $\beta_j(t)$, for $j = 1, \ldots, p$ are given by $\hat{\beta}_j(t) = \boldsymbol{B}_q^T(t)\hat{\boldsymbol{\gamma}}_j$.

3. ASYMPTOTIC PROPERTIES

In this section, we explore the asymptotic properties of the proposed estimators. We introduce some notation before stating the result. Denote the space of rth-order smooth functions as $C^{(r)}([0,1]) = \{\varphi \mid \varphi^{(r)} \in C([0,1])\}$. Let $\|\cdot\|$ denote the Euclidean norm for a vector or the $L_2(\mathcal{T})$ norm for a function defined on $\mathcal{T}$. Let f' and f'' represent the first and second derivatives of f, respectively. Denote the smallest and largest eigenvalues of a symmetric matrix

$\boldsymbol{A}$ by $\rho_{\min}(\boldsymbol{A})$ and $\rho_{\max}(\boldsymbol{A})$, respectively. For positive numbers c_n and d_n, let $c_n \asymp d_n$ denote that $\lim_{n\to\infty} c_n/d_n = C$, where C is some nonzero constant. We need the following regularity conditions:

- (C1) The matrix $\boldsymbol{\Sigma} = \mathbb{E}(\boldsymbol{X}\boldsymbol{X}^T)$ is positive definite and its eigenvalues are uniformly bounded away from 0 and infinity. In addition, assume that the dimension of $\boldsymbol{X}$ is fixed, and there exists a positive constant M such that $|X_j| \leq M$, for all $1 \leq j \leq p$.
- (C2) For every $1 \leq j \leq p$, the functional coefficient $\beta_j(\cdot) \in C^{(r)}[0,1]$ for some integer $r \geq 2$, and the spline order q satisfies $q \geq r$.
- (C3) The number of knots J_n satisfies $J_n \asymp n^{1/(2r+1)}$.
- (C4) $\mathbb{E}\left[\phi_h'(\epsilon(t))\right] = 0$ and $\mathbb{E}\left[\phi_h''(\epsilon(t))\right] > 0$, for all t and any $h > 0$.
- (C5) Define $\boldsymbol{\epsilon}_i = (\epsilon_{i1}, \dots, \epsilon_{im})^T$, $\phi_h'(\boldsymbol{\epsilon}_i) = \left(\phi_h'(\epsilon_{i1}), \dots, \phi_h'(\epsilon_{im})\right)^T$ and $\boldsymbol{G} = \mathbb{E}\left[\phi_h'(\boldsymbol{\epsilon}_i)(\phi_h'(\boldsymbol{\epsilon}_i))^T\right]$. Assume that $\rho_{\min}(\boldsymbol{G}) \geq \rho_0 > 0$.

Condition (C1) is similar to the conditions (C2) and (C3) in Huang, Wu & Zhou (2002, 2004) and Wang, Li & Huang (2008). Condition (C2) gives smoothness conditions for the functional coefficients, which are commonly assumed in the spline-smoothing literature, such as Zhou, Shen & Wolfe (1998). Condition (C3) is a common assumption on the number of knots. Condition (C4) is similar to Assumption (A6) in Yao, Lindsay & Li (2012), which ensures that there exists a local minimizer of the objective function given by Equation (3). For $\epsilon(t)$ satisfying this assumption, we can cite, for instance, $\epsilon(t)$ following a zero-mean Gaussian process. Condition (C5) is the same as Assumption (A6) of Lima, Cao & Billor (2019), which is used for obtaining the asymptotic normality of the proposed estimator.

Theorem 1. *Assume that conditions (C1)–(C4) hold. If $\lambda_n \to 0$ and $n^{\frac{r}{2r+1}}\lambda_n \to \infty$ as $n \to \infty$, then there exists a local minimizer $\hat{\boldsymbol{\gamma}}$ of $\mathcal{L}(\boldsymbol{\gamma})$ such that $\hat{\beta}_j(t) = \boldsymbol{B}_q^T(t)\hat{\boldsymbol{\gamma}}_j$ and*

(i) $\|\hat{\beta}_j(t) - \beta_j(t)\| = O_p\left(n^{\frac{-r}{2r+1}}\right)$, $j = 1, \dots, p$;
(ii) $\hat{\beta}_j(\cdot) = 0$, $j = d+1, \dots, p$, with probability approaching one.

Remark 1. *The assumptions on the regularization parameter λ_n are general in the variable selection literature (see, e.g., Fan & Li, 2001; Wang et al., 2013). It is worth noting that if we choose $r = 2$, the resulting convergence rate is then $O_p(n^{-2/5})$, which is optimal for nonparametric regression (Stone, 1982).*

Let $\boldsymbol{X}_i^{(1)} = (X_{i1}, \dots, X_{id})^T$ be the vector of predictors corresponding to the nonzero functional coefficients. Denote $\boldsymbol{Z}_{ik}^{(1)} = \boldsymbol{X}_i^{(1)} \otimes \boldsymbol{B}_q(t_k)$, $\boldsymbol{Z}_i^{(1)} = \left(\boldsymbol{Z}_{i1}^{(1)}, \dots, \boldsymbol{Z}_{im}^{(1)}\right)^T$ and $\boldsymbol{\Lambda}^{(1)} = \mathbb{E}\left[(\boldsymbol{Z}_i^{(1)})^T \boldsymbol{G} \boldsymbol{Z}_i^{(1)}\right]$, where $\otimes$ represents the Kronecker product. Define

$$F(t) = \mathbb{E}\left[\phi_h''(\epsilon(t))\right], \quad \boldsymbol{u}^{(1)} = \mathbb{E}\left(\boldsymbol{X}_i^{(1)}\right), \quad \mathbf{z}^{(1)}(t) = \boldsymbol{u}^{(1)} \otimes \boldsymbol{B}_q(t), \text{and}$$

$$\boldsymbol{\Sigma}^{(1)} = \mathbb{E}\left[\boldsymbol{X}_i^{(1)}\left(\boldsymbol{X}_i^{(1)}\right)^T\right], \quad \mathbf{B}^*(t) = \boldsymbol{B}_q(t)\boldsymbol{B}_q^T(t), \quad \boldsymbol{\Delta}^{(1)}(t) = \boldsymbol{\Sigma}^{(1)} \otimes \mathbf{B}^*(t).$$

Let $\boldsymbol{e}_j$ denote the d-dimensional vector with the jth element taken to be 1 and zero elsewhere. Set $\boldsymbol{A}_j(t) = \boldsymbol{e}_j \otimes \boldsymbol{B}_q(t)$. The following theorem gives the asymptotic bias, asymptotic variance and point-wise asymptotic distribution of $\hat{\beta}_j(t)$, for $j = 1, \dots, d$.

Theorem 2. *Supposing that the conditions in Theorem 1 and condition (C5) hold, we have the following results:*

(i) For $j = 1, \dots, d$, the asymptotic bias of $\hat{\beta}_j(t)$, denoted by $b_j(t)$, is given by

$$b_j(t) = J_n^{-r+1/2} \boldsymbol{A}_j^T(t) \left\{ \sum_{k=1}^{m} F(t_k) \boldsymbol{\Delta}^{(1)}(t_k) \right\}^{-1} \times \left\{ \sum_{k=1}^{m} F(t_k) \mathbf{z}^{(1)}(t_k) \right\} \{1 + o(1)\}.$$

(ii) For $j = 1, \dots, d$, the asymptotic variance of $\hat{\beta}_j(t)$ is given by

$$\mathrm{var}\left\{\hat{\beta}_j(t)\right\} = n^{-1} \boldsymbol{A}_j^T(t) \left\{ \sum_{k=1}^{m} F(t_k) \boldsymbol{\Delta}^{(1)}(t_k) \right\}^{-1} \boldsymbol{\Lambda}^{(1)} \times \left\{ \sum_{k=1}^{m} F(t_k) \boldsymbol{\Delta}^{(1)}(t_k) \right\}^{-1} \boldsymbol{A}_j(t) \{1 + o(1)\}.$$

(iii) If $\lim_{n\to\infty} J_n m/n = 0$, then, for any fixed $t \in [0, 1]$, the estimate $\hat{\beta}_j(t)$ has the following asymptotic distribution:

$$\frac{\hat{\beta}_j(t) - \beta_j(t) - b_j(t)}{\sqrt{\mathrm{var}\left\{\hat{\beta}_j(t)\right\}}} \xrightarrow{D} N(0, 1), \quad j = 1, \dots, d.$$

Remark 2. *Theorem 2(iii) gives the point-wise asymptotic normality of the proposed estimator $\hat{\beta}_j(t)$, for $j = 1, \dots, d$. This result is similar to Theorem 3.1 in Zhou, Shen & Wolfe (1998). The condition $\lim_{n\to\infty} J_n m/n = 0$ in Theorem 2(iii), which ensures normality, was also used in Huang, Wu & Zhou (2004) and Wang, Li & Huang (2008).*

4. ESTIMATION ALGORITHM AND TUNING PARAMETER SELECTION

4.1. Estimation Algorithm

Note that the traditional computation method is not applicable for the minimization problem of Equation (3) since the SCAD penalty functions do not have continuous second-order derivatives. We adopt the LQA method (Fan & Li, 2001) to approximate the penalty function $p_{\lambda_n}(\cdot)$ and develop an iterative algorithm to solve the minimization problem of Equation (3). Specifically, given an initial value $\boldsymbol{\gamma}^0$ that is close to $\boldsymbol{\gamma}$, when $\boldsymbol{\gamma}_j \neq 0$, LQA of the SCAD function $p_{\lambda_n}(\|\boldsymbol{\gamma}_j\|)$ is

$$p_{\lambda_n}(\|\boldsymbol{\gamma}_j\|) \approx p_{\lambda_n}(\|\boldsymbol{\gamma}_j^0\|) + \frac{1}{2}\left\{p'_{\lambda_n}(\|\boldsymbol{\gamma}_j^0\|)/\|\boldsymbol{\gamma}_j^0\|\right\}\left(\|\boldsymbol{\gamma}_j\|^2 - \|\boldsymbol{\gamma}_j^0\|^2\right).$$

Let $\boldsymbol{X}_i = (X_{i1}, \dots, X_{ip})^T$, $\boldsymbol{Z}_{ik} = \boldsymbol{X}_i \otimes \boldsymbol{B}_q(t_k)$, and

$$\boldsymbol{\Sigma}(\boldsymbol{\gamma}^0) = \mathrm{diag}\left\{\left(p'_{\lambda_n}(\|\boldsymbol{\gamma}_1^0\|)/\|\boldsymbol{\gamma}_1^0\|\right)\boldsymbol{I}_{J_n}, \dots, \left(p'_{\lambda_n}(\|\boldsymbol{\gamma}_p^0\|)/\|\boldsymbol{\gamma}_p^0\|\right)\boldsymbol{I}_{J_n}\right\},$$

where $\boldsymbol{I}_{J_n}$ is the $J_n \times J_n$ identity matrix. Then, the expression of $\mathcal{L}(\boldsymbol{\gamma})$ in Equation (3) can be approximated by

$$\mathcal{L}_0(\boldsymbol{\gamma}) = \sum_{i=1}^{n}\sum_{k=1}^{m} \phi_h\left(Y_{ik} - \boldsymbol{Z}_{ik}^T\boldsymbol{\gamma}\right) + \frac{1}{2}nm\boldsymbol{\gamma}^T\boldsymbol{\Sigma}(\boldsymbol{\gamma}^0)\boldsymbol{\gamma}. \tag{4}$$

Differentiating Equation (4) with respect to $\boldsymbol{\gamma}$ and setting it to zero, we have

$$-\sum_{i=1}^{n}\sum_{k=1}^{m} \phi_h'\left(Y_{ik} - \boldsymbol{Z}_{ik}^T\boldsymbol{\gamma}\right)\boldsymbol{Z}_{ik} + nm\boldsymbol{\Sigma}(\boldsymbol{\gamma}^0)\boldsymbol{\gamma} = 0. \tag{5}$$

Set $w_{ik} = \phi_h'(r_{ik})/r_{ik}$ with $r_{ik} = Y_{ik} - \boldsymbol{Z}_{ik}^T\boldsymbol{\gamma}$, for $i = 1, \dots, n$, $k = 1, \dots, m$. Let $\boldsymbol{W} = \mathrm{diag}(w_{11}, \dots, w_{1m}, w_{21}, \dots, w_{nm})$, $\boldsymbol{Y} = (Y_{11}, \dots, Y_{1m}, Y_{21}, \dots, Y_{nm})^T$ and $\boldsymbol{Z} = (\boldsymbol{Z}_{11}, \dots, \boldsymbol{Z}_{1m}, \boldsymbol{Z}_{21}, \dots, \boldsymbol{Z}_{nm})^T$. Then, Equation (5) can be rewritten as

$$\boldsymbol{Z}^T\boldsymbol{W}\boldsymbol{Z}\boldsymbol{\gamma} + nm\boldsymbol{\Sigma}(\boldsymbol{\gamma}^0)\boldsymbol{\gamma} = \boldsymbol{Z}^T\boldsymbol{W}\boldsymbol{Y}. \tag{6}$$

This yields the solution

$$\hat{\boldsymbol{\gamma}} = \left\{\boldsymbol{Z}^T\boldsymbol{W}\boldsymbol{Z} + nm\boldsymbol{\Sigma}(\boldsymbol{\gamma}^0)\right\}^{-1}\boldsymbol{Z}^T\boldsymbol{W}\boldsymbol{Y}.$$

The iterative algorithm to compute $\hat{\boldsymbol{\gamma}} = \left(\hat{\boldsymbol{\gamma}}_1^T, \dots, \hat{\boldsymbol{\gamma}}_p^T\right)^T$ is as follows:

Step 1: Obtain an initial estimate of $\boldsymbol{\gamma}$, say $\hat{\boldsymbol{\gamma}}^{(0)}$.

Step 2: Given $\hat{\boldsymbol{\gamma}}^{(r)}$, compute the residuals $\hat{r}_{ik}^{(r)} = Y_{ik} - \boldsymbol{Z}_{ik}^T\hat{\boldsymbol{\gamma}}^{(r)}$ and the weight matrix $\widehat{\boldsymbol{W}}^{(r)}$ with elements $\hat{w}_{ik}^{(r)} = \phi_h'\left(\hat{r}_{ik}^{(r)}\right)/\hat{r}_{ik}^{(r)}$, for $i = 1, \dots, n$, $k = 1, \dots, m$, and obtain an updated estimate, $\hat{\boldsymbol{\gamma}}^{(r+1)} = \left\{\boldsymbol{Z}^T\widehat{\boldsymbol{W}}^{(r)}\boldsymbol{Z} + nm\boldsymbol{\Sigma}(\hat{\boldsymbol{\gamma}}^{(r)})\right\}^{-1}\boldsymbol{Z}^T\widehat{\boldsymbol{W}}^{(r)}\boldsymbol{Y}$.

Step 3: Repeat Step 2 until convergence. The final estimate of $\boldsymbol{\gamma}$ is obtained when all elements in $\hat{\boldsymbol{\gamma}}$ change less than a pre-specified threshold.

In practice, we set the least squares estimator as the initial estimate, that is $\hat{\boldsymbol{\gamma}}^{(0)} = \left(\boldsymbol{Z}^T\boldsymbol{Z}\right)^{-1}\boldsymbol{Z}^T\boldsymbol{Y}$. At any iteration of Step 2, if some $\|\hat{\boldsymbol{\gamma}}_j^{(r)}\|$ is smaller than a pre-specified cut-off value (10^{-3} is used in our implementation), we then set $\hat{\boldsymbol{\gamma}}_j = 0$ and treat the corresponding predictor as irrelevant in subsequent iterations.

4.2. The Choice of Tuning Parameters

To implement the estimation algorithm, we need to choose the number of interior knots N_n, the regularization parameter λ_n, and the tuning parameter h. Since these three parameters depend on each other, it could be treated as a ternary optimization problem. We consider two data-driven procedures to select N_n, λ_n and h.

4.2.1. Choice of N_n and λ_n

The parameter N_n controls the dimensions of the spline spaces used to approximate the true functional coefficients, and λ_n regulates the shrinkage strength. Some robust selection criteria, including the robust cross-validation (CV) procedure proposed by Yao & Wang (2013) and the weighted leave-out-one-column CV employed in Lee, Shin & Billor (2013), can be used to select these parameters. To simplify the computation, we present a weighted generalized CV (WGCV) criterion to select N_n and λ_n. Note that the solution of Equation (6) is equal to the minimizer of the penalized weighted least squares criterion

$$\tilde{\mathcal{L}}_0(\boldsymbol{\gamma}) = (\boldsymbol{Y} - \boldsymbol{Z}\boldsymbol{\gamma})^T \boldsymbol{W} (\boldsymbol{Y} - \boldsymbol{Z}\boldsymbol{\gamma}) + nm\boldsymbol{\gamma}^T \boldsymbol{\Sigma}(\boldsymbol{\gamma}^0)\boldsymbol{\gamma}.$$

We suggest the following WGCV criterion:

$$\text{WGCV}(N_n, \lambda_n) = \frac{n^{-1}m^{-1}(\boldsymbol{Y} - \boldsymbol{Z}\hat{\boldsymbol{\gamma}})^T \widehat{\boldsymbol{W}} (\boldsymbol{Y} - \boldsymbol{Z}\hat{\boldsymbol{\gamma}})}{\left(1 - n^{-1}m^{-1}\text{trace}\{\mathbf{H}(N_n, \lambda_n)\}\right)^2},$$

where $\mathbf{H}(N_n, \lambda_n) = \boldsymbol{Z}\{\boldsymbol{Z}^T \widehat{\boldsymbol{W}} \boldsymbol{Z} + nm\boldsymbol{\Sigma}(\hat{\boldsymbol{\gamma}})\}^{-1}\boldsymbol{Z}^T \widehat{\boldsymbol{W}}$ is the hat matrix.

4.2.2. Choice of h

The tuning parameter h impacts the efficiency of the proposed estimator. To attain high efficiency, we adopt the approach of Yao, Lindsay & Li (2012) to select h by minimizing the asymptotic variances of the proposed estimators. Specifically, let $\tilde{\boldsymbol{A}}(t) = (\boldsymbol{B}_q^T(t), \dots, \boldsymbol{B}_q^T(t))^T$ be a $(p \times J_n)$-dimensional vector valued function $\boldsymbol{Z}_i = (\boldsymbol{Z}_{i1}, \dots, \boldsymbol{Z}_{im})^T$, and $\widehat{\boldsymbol{G}}_i = \phi_h'(\hat{\boldsymbol{\epsilon}}_i)[\phi_h'(\hat{\boldsymbol{\epsilon}}_i)]^T$, where $\hat{\boldsymbol{\epsilon}}_i = (\hat{\epsilon}_{i1}, \dots, \hat{\epsilon}_{im})^T$ and $\hat{\epsilon}_{ik} = Y_{ik} - \boldsymbol{X}_i^T \hat{\boldsymbol{\beta}}(t_k)$. Based on Theorem 2, define $\hat{V}(h) = \sum_{k=1}^m \hat{\boldsymbol{\Xi}}_h(t_k)$, where $\hat{\boldsymbol{\Xi}}_h(t) = \tilde{\boldsymbol{A}}^T(t)\hat{\boldsymbol{\Gamma}}_h^{-1}\hat{\boldsymbol{\Lambda}}_h\hat{\boldsymbol{\Gamma}}_h^{-1}\tilde{\boldsymbol{A}}(t)$, and

$$\hat{\boldsymbol{\Gamma}}_h = \frac{1}{n}\sum_{i=1}^n \sum_{k=1}^m \phi_h''(\hat{\epsilon}_{ik})\boldsymbol{Z}_{ik}\boldsymbol{Z}_{ik}^T, \text{ and } \hat{\boldsymbol{\Lambda}}_h = \frac{1}{n}\sum_{i=1}^n \boldsymbol{Z}_i^T \widehat{\boldsymbol{G}}_i \boldsymbol{Z}_i.$$

Then, we select the tuning parameter h that minimizes $\hat{V}(h)$.

5. SIMULATION STUDIES

In this section, we conduct several simulation studies to illustrate the finite-sample performance of the proposed method. Specifically, we generate the functional response $Y_i(t)$ from Equation (1) with $p = 6$ scalar predictors. In all simulation studies, the functional response is defined on $\mathcal{T} = [0, 1]$ with $m = 50$ equally spaced observation points. We generate predictors $\{X_{ij}\}_{j=1}^p$ from a multivariate normal distribution with mean zero and covariance $\text{cov}(X_{ij}, X_{ij'}) = 0.5^{|j-j'|}$. The functional coefficients are set as $\beta_1(t) = 2t^2$, $\beta_2(t) = \cos(3\pi t/2 + \pi/2)$, $\beta_3(t) = \sqrt{2}\sin(\pi t/2) + 3\sqrt{2}\sin(3\pi t/2)$ and $\beta_4(t) = \beta_5(t) = \beta_6(t) = 0$. The random error functions are generated as $\epsilon_i(t) = \sum_{l=1}^2 \xi_{il}\phi_l(t) + \varepsilon_i(t)$, $i = 1, \dots, n$, where $\phi_1(t) = -\cos\{\pi(t - 1/2)\}$, $\phi_2(t) = \sin\{\pi(t - 1/2)\}$, $\xi_{il} \sim N(0, 0.1^2)$ and $\varepsilon_i(t_k)$ are simulated i.i.d. $N(0, 0.5^2)$.

We simulate 100 datasets from the model given by Equation (1) with sample sizes of $n = 50$, 100 and 200. To demonstrate the robust nature of the proposed approach over the least squares method, we simulate several contamination situations including outliers in the functional response, outliers in the scalar predictors, and outliers in both the functional response and scalar predictors. For outliers occurring in the functional response case, we assume that 10% response curves from the original sample are contaminated with outlier curves, recorded as $Y_i^o(t)$; that is, the contaminated functional response data have the form $Y_i^*(t) = (1 - E_i)Y_i(t) + E_iY_i^o(t)$, where E_i is a Bernoulli random variable that takes a value of one with probability 0.1. Similarly, we simulate outliers in scalar predictors by randomly selecting a subset of predictors from the original sample and contaminating them with peak values. The detailed mechanisms for producing the outlier response $Y_i^o(t)$ and outliers in scalar predictors are described as follows:

(I) *No outliers*. In this setting, $Y_i^o(t) = Y_i(t)$. This is used to evaluate the performance of the proposed robust estimator compared to the least squares estimator when there are no outliers.

(II) *Bump outliers in the response*. This type of outlier is characterized by contamination occurring in a subinterval of $[0, 1]$ in the functional response. Specifically, let $[U_i, U_i + v]$ 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 outlier curve $Y_i^o(t)$ is generated as $Y_i^o(t_k) = Y_i(t_k) + e_{ik} I_{[U_i, U_i+v]}(t_k),\ \ k = 1, \dots, m$, where e_{ik} are random values taken from a uniform distribution on $[-e_u, -e_l] \cup [e_l, e_u]$ and $I(\cdot)$ is the indicator function. In practice, the constant v determines the contamination length of each outlier curve, and the parameters e_l and e_u control the strength of outliers. We set $e_l = 6$, $e_u = 10$ and $v = 0.5$ in the simulation.

(III) *Shifted outliers in the response*. This is introduced to simulate an outlier with a shifted influence in the functional response. We generate each outlier curve by adding a random value at each observation point on the corresponding original curve, i.e., $Y_i^o(t_k) = Y_i(t_k) + \zeta_{ik}$, $k = 1, \dots, m$, where ζ_{ik} are random values taken from a uniform distribution on $[-a_u, -a_l] \cup [a_l, a_u]$. In the simulation, we set $a_l = 4$ and $a_u = 6$.

(IV) *Outliers in the predictors*. To generate outliers in the scalar predictors, we randomly select 5% of predictors to be contaminated with high-leverage outliers $\left(X_{i1}^o, \dots, X_{ip}^o\right) = (X_{i1}, \dots, X_{ip}) + 4$.

(V) *Outliers in both the response and predictors*. The noise terms $\varepsilon_i(t_k)$ follow a mixture of a normal distribution and a Cauchy distribution: $0.8N(0, 0.5^2) + 0.2Cauchy(0, 1)$. Five percent of predictors are randomly selected to be contaminated with high-leverage outliers $\left(X_{i1}^o, \dots, X_{ip}^o\right) = (X_{i1}, \dots, X_{ip}) + 4$.

The finite-sample performance is evaluated by the positive selection rate (PSR; the proportion of causal features selected by one method in all causal features) and the noncausal selection rate (NSR; the average restricted only to the true zero functional coefficients), which were advocated by Wang et al. (2013), as well as the average and standard deviation of the integrated squared error (ISE)

$$\mathrm{ISE} = \sum_{j=1}^{p} \int_{\mathcal{T}} \left\{\hat{\beta}_j(t) - \beta_j(t)\right\}^2 dt$$

over 100 simulated datasets. All integrations required in the simulations are approximated by Riemann sums. For each dataset, we generate an independent predictor sample $\{X_{ij}^*, j = 1, \dots, p, i = 1, \dots, N\}$ with the sample size $N = 200$. Besides PSR, NSR and ISE, we use the mean square prediction error (MSPE) to assess the accuracy of prediction. MSPE is given by

$$\mathrm{MSPE} = \frac{1}{N} \sum_{i=1}^{N} \int_{\mathcal{T}} \left\{ \hat{Y}_i^*(t) - \sum_{j=1}^{p} \beta_j(t) X_{ij}^* \right\}^2 dt,$$

where $\hat{Y}_i^*(t) = \sum_{j=1}^{p} \hat{\beta}_j(t) X_{ij}^*$, and $\hat{\beta}_j(t)$, for $j = 1, \dots, p$, are estimated from the training data.

We first take Setting (III) as an example to illustrate the utility of the procedures described in Section 4.2. Specifically, we generate data with sample size $n = 100$ under Setting (III). We use

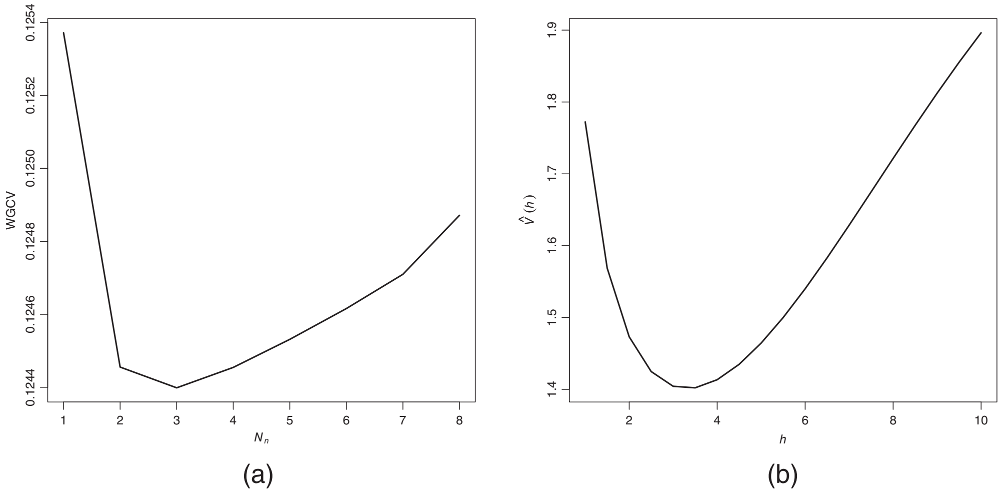


FIGURE 1: (a) WGCV against N_n given $\lambda_n = 0.3$ and $h = 3.5$; (b) $\hat{V}(h)$ against h given $N_n = 3$ and $\lambda_n = 0.3$.

TABLE 1: Simulation results when the sample size $n = 100$ under Setting (III), including positive selection rate (PSR), noncausal selection rate (NSR), and averages (with standard deviations in parentheses) of the integrated squared error (ISE) and mean square prediction error (MSPE), when varying h from 0.5 to 14 or choosing the optimal h, denoted by h_{opt}, using the proposed procedure described in Section 4.2.

h	PSR	NSR	ISE ($\times 10^{-2}$)	MSPE ($\times 10^{-2}$)
0.5	0.56	0.91	1143 (0.052)	1869 (193)
1	0.71	0.94	1143 (0.028)	1869 (192)
2	1.00	0.97	0.259 (0.124)	0.356 (0.150)
3	1.00	0.97	0.254 (0.123)	0.349 (0.150)
4	1.00	0.96	0.259 (0.131)	0.355 (0.159)
5	1.00	0.99	0.250 (0.112)	0.342 (0.130)
6	1.00	1.00	0.255 (0.102)	0.350 (0.116)
8	1.00	1.00	0.306 (0.138)	0.419 (0.165)
10	1.00	1.00	0.411 (0.229)	0.560 (0.285)
14	1.00	1.00	0.793 (0.590)	1.071 (0.762)
h_{opt}	1.00	1.00	0.239 (0.095)	0.327 (0.103)

cubic splines with order $q = 4$ to approximate the functional coefficients. The tuning parameters N_n, λ_n, and h selected by the proposed procedures are around 3, 0.3 and 3.5, respectively. Figure 1 displays the plots of WGCV against N_n given $\lambda_n = 0.3$, $h = 3.5$ and $\hat{V}(h)$ against h given $N_n = 3$ and $\lambda_n = 0.3$.

To illustrate the effect of the choice of h, in Table 1 we show results for the proposed estimator with h varying from 0.5 to 14 for Setting (III). The other two tuning parameters N_n and λ_n are selected using the WGCV criterion. It can be seen that the selection of predictors is

TABLE 2: Simulation results under Setting (I) including positive selection rate (PSR), noncausal selection rate (NSR), and averages (with standard deviations in parentheses) of the integrated squared error (ISE) and mean square prediction error (MSPE).

n	Methods	PSR	NSR	ISE ($\times 10^{-2}$)	MSPE ($\times 10^{-2}$)
50	ESL-gSCAD	1.000	1.000	0.423 (0.180)	1.158 (0.473)
	LS-gSCAD	1.000	0.997	0.410 (0.175)	1.120 (0.454)
	LAD-gSCAD	1.000	0.990	0.741 (0.298)	1.958 (0.744)
	Oracle	1.000	1.000	0.431 (0.184)	1.180 (0.482)
100	ESL-gSCAD	1.000	1.000	0.220 (0.095)	0.308 (0.112)
	LS-gSCAD	1.000	1.000	0.212 (0.090)	0.298 (0.108)
	LAD-gSCAD	1.000	0.997	0.370 (0.168)	0.497 (0.181)
	Oracle	1.000	1.000	0.224 (0.096)	0.314 (0.113)
200	ESL-gSCAD	1.000	1.000	0.108 (0.036)	0.077 (0.026)
	LS-gSCAD	1.000	1.000	0.107 (0.036)	0.076 (0.025)
	LAD-gSCAD	1.000	0.993	0.167 (0.057)	0.115 (0.038)
	Oracle	1.000	1.000	0.109 (0.037)	0.078 (0.026)

quite accurate and stable for a wide range of h, but with a relatively small PSR when $h \leq 1$. ISE and MSPE achieve their minima when $h = 5$, and appear stable for h ranging from 2 to 6 but then deteriorate as h gets too large. When we choose h using the proposed method described in Section 4.2.2, it yields the optimal PSR, NSR, ISE and MSPE.

For each setting, we compare the performance of our method (ESL-gSCAD) with three other approaches, namely the LS-gSCAD method introduced by Wang, Chen & Li (2007), the LAD-gSCAD method and the oracle method based on the MM estimator. We use cubic splines with order $q = 4$ to approximate the functional coefficients $\beta_j(\cdot)$. The LS-gSCAD estimator is implemented by the group coordinate descent algorithm (Breheny & Huang, 2015). The LAD-gSCAD estimator is implemented using the iterative coordinate descent algorithm (QICD; Peng & Wang, 2015). The tuning parameters N_n, λ_n and h of the proposed method are selected using the procedures presented in Section 4.2.

Tables 2–6 summarize the PSR, NSR and the averages and standard deviations of ISE and MSPE over 100 simulated datasets for Settings (I)–(V), respectively. Several observations can be made from Tables 2–6. First, when there are no outliers in the datasets and the error distribution is normal, the values of PSR, NSR, ISE and MSPE of the ESL-gSCAD estimator are close to those of the LS-gSCAD estimator. This implies that the performances of these two methods are comparable for the datasets in the absence of outliers. Second, the PSR and NSR of the ESL-gSCAD estimator are around 1 in all settings, while the PSR of the LS-gSCAD estimator for Setting (V) and the NSR of the LS-gSCAD estimator for Settings (III), (IV) and (V) are further away from 1. These results suggest that the proposed ESL-gSCAD estimator leads to accurate variable selection in situations where outliers result from either the functional response or the scalar predictor domain. Third, the ESL-gSCAD estimator yields smaller ISE and MSPE than the LS-gSCAD estimator and LAD-gSCAD estimator for Settings (II)–(V). This indicates that the proposed ESL-gSCAD method is superior to the LS-gSCAD and LAD-gSCAD methods in terms of ISE and MSPE for the datasets in the presence of outliers. Moreover, it is interesting to see that the superiority of ESL-gSCAD becomes increasingly clear when outliers

TABLE 3: Simulation results under Setting (II) including positive selection rate (PSR), noncausal selection rate (NSR), and averages (with standard deviations in parentheses) of the integrated squared error (ISE) and mean square prediction error (MSPE).

n	Methods	PSR	NSR	ISE ($\times 10^{-2}$)	MSPE ($\times 10^{-2}$)
50	ESL-gSCAD	1.000	1.000	0.437 (0.165)	1.226 (0.417)
	LS-gSCAD	1.000	0.997	4.368 (2.759)	11.716 (7.056)
	LAD-gSCAD	1.000	1.000	0.940 (0.503)	2.520 (1.246)
	Oracle	1.000	1.000	0.447 (0.166)	1.250 (0.418)
100	ESL-gSCAD	1.000	1.000	0.201 (0.071)	0.580 (0.196)
	LS-gSCAD	1.000	1.000	1.774 (1.129)	4.967 (2.780)
	LAD-gSCAD	1.000	0.997	0.349 (0.130)	0.970 (0.340)
	Oracle	1.000	1.000	0.205 (0.070)	0.593 (0.194)
200	ESL-gSCAD	1.000	1.000	0.118 (0.049)	0.320 (0.113)
	LS-gSCAD	1.000	1.000	1.046 (0.530)	2.841 (1.389)
	LAD-gSCAD	1.000	1.000	0.193 (0.075)	0.515 (0.174)
	Oracle	1.000	1.000	0.120 (0.050)	0.326 (0.116)

TABLE 4: Simulation results under Setting (III) including positive selection rate (PSR), noncausal selection rate (NSR), and averages (with standard deviations in parentheses) of the integrated squared error (ISE) and mean square prediction error (MSPE).

n	Methods	PSR	NSR	ISE ($\times 10^{-2}$)	MSPE ($\times 10^{-2}$)
50	ESL-gSCAD	1.000	1.000	0.512 (0.228)	1.389 (0.565)
	LS-gSCAD	1.000	0.620	45.55 (36.78)	123.2 (90.38)
	LAD-gSCAD	1.000	0.990	2.148 (2.390)	6.157 (6.810)
	Oracle	1.000	1.000	0.507 (0.225)	1.375 (0.558)
100	ESL-gSCAD	1.000	1.000	0.239 (0.095)	0.327 (0.103)
	LS-gSCAD	1.000	0.567	20.40 (15.76)	26.33 (17.77)
	LAD-gSCAD	1.000	0.987	0.720 (0.478)	0.952 (0.549)
	Oracle	1.000	1.000	0.237 (0.094)	0.325 (0.102)
200	ESL-gSCAD	1.000	1.000	0.123 (0.044)	0.087 (0.027)
	LS-gSCAD	1.000	0.607	8.718 (7.891)	5.832 (4.471)
	LAD-gSCAD	1.000	0.980	0.329 (0.193)	0.231 (0.126)
	Oracle	1.000	1.000	0.121 (0.043)	0.086 (0.027)

exist in the predictors or in both the response and predictors. This is mainly because the proposed ESL-gSCAD method puts more weight on the "most likely" data around the true value when there are some very large outliers in the datasets; as a result, a robust and efficient estimator is realized. Fourth, the performance of the ESL-gSCAD estimator is comparable to the oracle

TABLE 5: Simulation results under Setting (IV) including positive selection rate (PSR), noncausal selection rate (NSR), and averages (with standard deviations in parentheses) of the integrated squared error (ISE) and mean square prediction error (MSPE).

n	Methods	PSR	NSR	ISE ($\times 10^{-2}$)	MSPE ($\times 10^{-2}$)
50	ESL-gSCAD	1.000	1.000	0.486 (0.160)	1.520 (0.537)
	LS-gSCAD	1.000	0.423	95.36 (29.47)	525.3 (103.5)
	LAD-gSCAD	1.000	0.813	10.74 (16.28)	57.87 (91.10)
	Oracle	1.000	1.000	0.477 (0.161)	1.484 (0.531)
100	ESL-gSCAD	1.000	1.000	0.304 (0.099)	0.536 (0.151)
	LS-gSCAD	1.000	0.283	80.22 (14.56)	273.6 (39.51)
	LAD-gSCAD	1.000	0.630	12.78 (8.820)	37.12 (26.31)
	Oracle	1.000	1.000	0.298 (0.098)	0.523 (0.148)
200	ESL-gSCAD	1.000	1.000	0.184 (0.055)	0.190 (0.050)
	LS-gSCAD	1.000	0.153	73.88 (11.74)	135.4 (16.26)
	LAD-gSCAD	1.000	0.473	12.10 (4.969)	19.34 (8.310)
	Oracle	1.000	1.000	0.180 (0.054)	0.185 (0.048)

estimator in all settings. The LAD-gSCAD estimator performs poorly when there are outliers in the predictors. Finally, the ISE and MSPE values of the ESL-gSCAD estimator decrease as the sample size n increases, which corroborates our consistency results established in Theorem 1. In summary, the proposed ESL-gSCAD method performs as well as or better than the LS-gSCAD and LAD-gSCAD approaches for variable selection and estimation.

6. APPLICATION

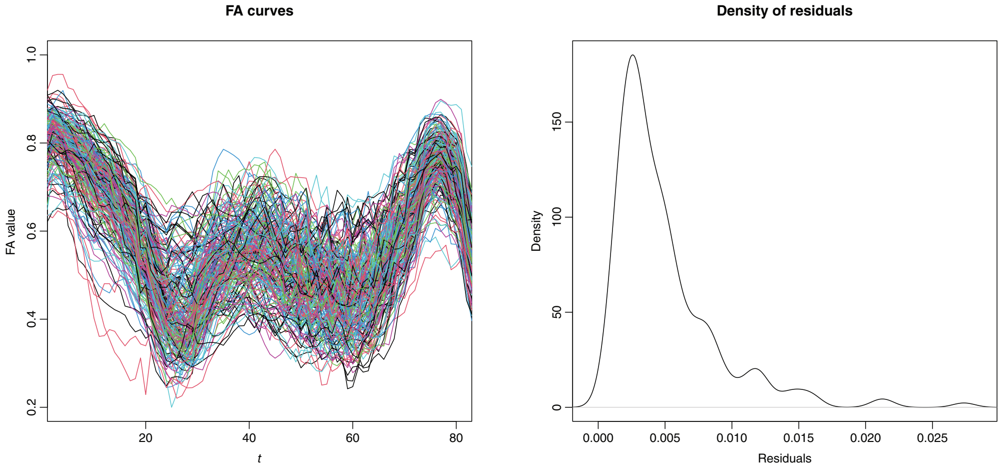
In this section, we apply the proposed method to analyze a real DTI dataset collected from the NIH Alzheimer's Disease Neuroimaging Initiative (ADNI) study. This dataset consists of 213 subjects and can be downloaded from the publicly available ADNI database (http://adni.loni.usc.edu/). For each subject, we compute a fractional anisotropy (FA) curve at $m = 83$ location points along the midsagittal skeleton of the corpus callosum, as displayed in Figure 2a. These FA curves, which quantify the directional strength of white matter tract structure at particular locations, are one of the most used measures in DTI data analysis and have been widely applied to statistical analyses in imaging studies. Detailed descriptions for calculating FA can be found in Smith et al. (2006) and Li, Huang & Zhu (2017).

Our goal is to investigate the association between FA ($Y(t)$) and seven scalar covariates, namely gender (123 male and 90 female, coded by a dummy variable indicating males), age in years (ranges from 48.4 to 90.4), handedness (193 right-handed and 20 left-handed, coded by a dummy variable indicating left-handedness), education level in years (ranges from 9 to 20), Alzheimer's disease (AD) status (19.6% of participants), mild cognitive impairment (MCI) status (55.1% of participants), and Mini-Mental State Exam (MMSE) score. Without loss of generality, we centralize all the functional response and scalar predictors to have mean zero, and apply the model given by Equation (1) with $p = 7$ to the dataset.

First, we fit the model using the LS-gSCAD method. Figure 2b shows the estimated density of the integrated squared residuals for each of the 213 FA curves based on LS-gSCAD method, in which we see that the resulting residuals are slightly right-skewed with possible outliers. This

TABLE 6: Simulation results under Setting (V) including positive selection rate (PSR), noncausal selection rate (NSR), and averages (with standard deviations in parentheses) of the integrated squared error (ISE) and mean square prediction error (MSPE).

n	Methods	PSR	NSR	ISE ($\times10^{-2}$)	MSPE ($\times10^{-2}$)
50	ESL-gSCAD	1.000	1.000	0.750 (0.295)	2.362 (0.860)
	LS-gSCAD	0.447	0.820	709.0 (783.2)	2308 (2083)
	LAD-gSCAD	0.987	0.903	25.14 (114.7)	106.1 (369.0)
	Oracle	1.000	1.000	0.748 (0.300)	2.329 (0.876)
100	ESL-gSCAD	1.000	1.000	0.394 (0.142)	0.722 (0.258)
	LS-gSCAD	0.490	0.777	639.1 (509.4)	1053 (847.2)
	LAD-gSCAD	0.997	0.760	14.86 (14.42)	40.31 (31.12)
	Oracle	1.000	1.000	0.384 (0.139)	0.691 (0.246)
200	ESL-gSCAD	1.000	1.000	0.258 (0.077)	0.280 (0.076)
	LS-gSCAD	0.457	0.827	666.2 (440.1)	546.4 (349.5)
	LAD-gSCAD	1.000	0.667	13.08 (7.141)	18.91 (9.811)
	Oracle	1.000	1.000	0.245 (0.075)	0.258 (0.073)



(a) (b)

FIGURE 2: (a) Raw FA curves measured at 83 grid points. (b) Estimated density of the integrated squared residuals obtained from LS-gSCAD method.

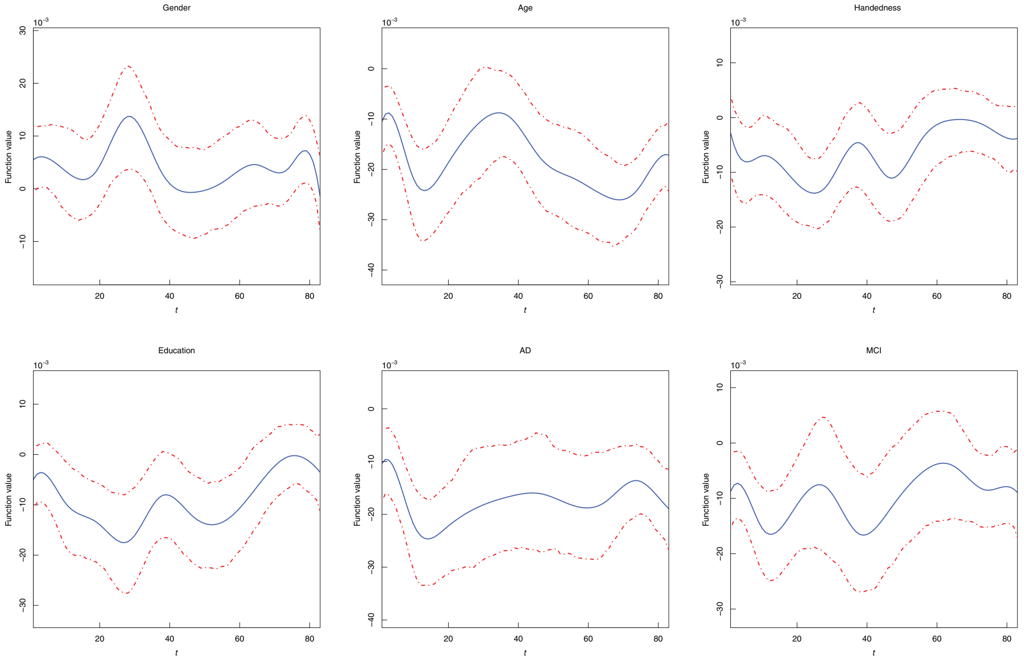


FIGURE 3: Estimated functional coefficients for gender, age, handedness, education, AD and MCI for the ADNI data. Solid lines are the estimates and dashed lines are the 95% bootstrap point-wise confidence intervals.

motivates us to apply our ESL-gSCAD approach as a robust alternative. Next, we apply the ESL-gSCAD approach to analyze the ADNI data. With the proposed ESL-gSCAD procedure, we use cubic splines with order $q = 4$ to approximate the functional coefficients, and select the tuning parameters N_n, λ_n and h using the proposed procedures described in Section 4.2. Our method selects gender, age, handedness, education, AD and MCI as relevant predictors and treats the MMSE as irrelevant. Figure 3 displays the estimated functional coefficients for these six relevant variables with their associated 95% point-wise confidence intervals. The point-wise confidence intervals are obtained with the standard error based on bootstrap resampling 500 times. From Figure 3, we see that the functional coefficient for gender takes positive values at most of the grid points, while the functional coefficients for the other five predictors take negative values. This indicates that males tend to have higher FA values than females, and AD and MCI patients tend to have lower FA values than participants without AD or MCI. In addition, being older, left-handed, or more educated may lead to smaller FA values. These results coincide with those of the previous analysis in Cai et al. (2020).

To further demonstrate the effectiveness of the proposed method, we compare our ESL-gSCAD method with the LS-gSCAD method in terms of prediction accuracy. Specifically, we randomly split the 213 samples into a training set with 149 samples and a test set with 64 samples. We use the training set to estimate the functional coefficients, and then predict the responses in the test set. The prediction accuracy is assessed by using the average squared prediction error (ASPE) defined as $\sum_{i=1}^{64}\sum_{j=1}^{83}\left\{Y_i^{\text{pred}}(t_j) - Y_i^{\text{test}}(t_j)\right\}^2/(64\times 83)$, where $\{Y_i^{\text{test}}(t_j), 1 \leq j \leq 83, 1 \leq i \leq 64\}$ are the responses in the test set, and $\left\{Y_i^{\text{pred}}(t_j), 1 \leq j \leq 83, 1 \leq i \leq 64\right\}$ are the corresponding predicted values. The averages (and standard deviations) of the ASPEs based on 100 replications are 0.489×10^{-2} (0.043×10^{-2}) for the proposed ESL-gSCAD method and 0.524×10^{-2} (0.050×10^{-2}) for the LS-gSCAD method. The proposed ESL-gSCAD method yields a much smaller ASPE than the LS-gSCAD method.

7. DISCUSSION

We have developed a robust variable selection procedure in the context of function-on-scalar models through the ESL function coupled with the group SCAD regularization method. The proposed procedure simultaneously identifies important predictors and provides estimates of the functional coefficients at their best convergence rates. The oracle property of the proposed method was also established. To implement the proposed ESL-gSCAD procedure, we presented an iterative algorithm based on the LQA method. The tuning parameters of the proposed method are selected via two data-driven procedures to achieve robustness and efficiency. The merits of our method were illustrated with simulation studies and an analysis of a real DTI data example. Specifically, we showed that our method could achieve high efficiency in estimating the functional coefficients and in selecting the relevant predictors in the presence of outliers in either the functional response or scalar predictors.

There are still some important issues worth addressing in future research. How to select the tuning parameter h in a data-driven way to achieve both robustness and efficiency is an important problem. In this article, we selected the tuning parameter h through minimizing the asymptotic variances of the proposed estimators to attain high efficiency; however, this procedure falls short in controlling the robustness of the estimators. Our future work will be focused on exploring more effective selection criteria to balance the robustness and efficiency of the proposed estimators. Investigating the robust variable selection problem for function-on-scalar regression with high-dimensional covariates is another meaningful topic for further study. It would also be valuable to extend the proposed method to longitudinal or sparse functional data and incorporate within-subject covariance into the estimation process to improve efficiency.

ACKNOWLEDGEMENTS

The authors would like to thank the editor, the associate editor and two anonymous referees for many insightful comments. These comments were very helpful for us to improve our work. This work was supported by the National Natural Science Foundation of China (11971001), the Beijing Natural Science Foundation (1182002) and the Natural Sciences and Engineering Research Council of Canada (NSERC) Discovery grant (2018-06008).

REFERENCES

Ainsworth, L. M., Routledge, R., & Cao, J. (2011). Functional data analysis in ecosystem research: The decline of Oweekeno Lake sockeye salmon and Wannock River flow. *Journal of Agricultural, Biological, and Environmental Statistics*, 16, 282–300.

Barber, R. F., Reimherr, M., & Schill, T. (2017). The function-on-scalar LASSO with applications to longitudinal GWAS. *Electronic Journal of Statistics*, 11, 1351–1389.

Breheny, P. & Huang, J. (2015). Group descent algorithms for nonconvex penalized linear and logistic regression models with grouped predictors. *Statistics and Computing*, 25, 173–187.

Cai, X., Xue, L., & Cao, J. (2021). Variable selection for multiple function-on-function linear regression. *Statistica Sinica*, https://doi.org/10.5705/ss.202020.0473.

Cai, X., Xue, L., Pu, X., & Yan, X. (2020). Efficient estimation for varying-coefficient mixed effects models with functional response data. *Metrika*, 84, 467–495.

Cao, J. & Ramsay, J. O. (2010). Linear mixed-effects modeling by parameter cascading. *Journal of the American Statistical Association*, 105, 365–374.

Chen, Y., Goldsmith, J., & Ogden, R. T. (2016). Variable selection in function-on-scalar regression. *Stat*, 5, 88–101.

Chiang, C.-T., Rice, J. A., & Wu, C. O. (2001). Smoothing spline estimation for varying coefficient models with repeatedly measured dependent variables. *Journal of the American Statistical Association*, 96, 605–619.

Fan, J. & Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96, 1348–1360.

Fan, Z. & Reimherr, M. (2017). High-dimensional adaptive function-on-scalar regression. *Econometrics and Statistics*, 1, 167–183.

Ferraty, F. & Vieu, P. (2006). *Nonparametric Functional Data Analysis: Theory and Practice*. Springer, New York.

Friston, K. J. (2009). Modalities, modes, and models in functional neuroimaging. *Science*, 326, 399–403.

Guan, T., Lin, Z., & Cao, J. (2020). Estimating truncated functional linear models with a nested group bridge approach. *Journal of Computational and Graphical Statistics*, 29, 620–628.

Huang, J. Z. (2003). Local asymptotics for polynomial spline regression. *The Annals of Statistics*, 31, 1600–1635.

Huang, J. Z., Wu, C. O., & Zhou, L. (2002). Varying-coefficient models and basis function approximations for the analysis of repeated measurements. *Biometrika*, 89, 111–128.

Huang, J. Z., Wu, C. O., & Zhou, L. (2004). Polynomial spline estimation and inference for varying coefficient models with longitudinal data. *Statistica Sinica*, 14, 763–788.

Jiang, F., Baek, S., Cao, J., & Ma, Y. (2020). A functional single index model. *Statistica Sinica*, 30, 303–324.

Lee, S., Shin, H., & Billor, N. (2013). M-type smoothing spline estimators for principal functions. *Computational Statistics & Data Analysis*, 66, 89–100.

Li, J., Huang, C., & Zhu, H. (2017). A functional varying-coefficient single-index model for functional response data. *Journal of the American Statistical Association*, 112, 1169–1181.

Lima, I. R., Cao, G., & Billor, N. (2019). M-based simultaneous inference for the mean function of functional data. *Annals of the Institute of Statistical Mathematics*, 71, 577–598.

Lin, Z., Cao, J., Wang, L., & Wang, H. (2017). Locally sparse estimator for functional linear regression models. *Journal of Computational and Graphical Statistics*, 26, 306–318.

Liu, B., Wang, L., & Cao, J. (2017). Estimating functional linear mixed-effects regression models. *Computational Statistics & Data Analysis*, 106, 153–164.

Maronna, R. A., Martin, R. D., & Yohai, V. J. (2006). *Robust Statistics: Theory and Methods*. Wiley, New York.

Parodi, A. & Reimherr, M. (2018). Simultaneous variable selection and smoothing for high-dimensional function-on-scalar regression. *Electronic Journal of Statistics*, 12, 4602–4639.

Peng, B. & Wang, L. (2015). An iterative coordinate descent algorithm for high-dimensional nonconvex penalized quantile regression. *Journal of Computational and Graphical Statistics*, 24, 676–694.

Ramsay, J. O. & Silverman, B. W. (2005). *Functional Data Analysis*, 2nd ed., Springer, New York.

Smith, S. M., Jenkinson, M., Johansen-Berg, H., Rueckert, D., Nichols, T. E., Mackay, C. E., Watkins, K. E., et al. (2006). Tract-based spatial statistics: Voxelwise analysis of multi-subject diffusion data. *Neuroimage*, 31, 1487–1505.

Stone, C. J. (1982). Optimal global rates of convergence for nonparametric regression. *The Annals of Statistics*, 10, 1040–1053.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B*, 58, 267–288.

Wang, L., Chen, G., & Li, H. (2007). Group SCAD regression analysis for microarray time course gene expression data. *Bioinformatics*, 23, 1486–1494.

Wang, L., Li, H., & Huang, J. Z. (2008). Variable selection in nonparametric varying-coefficient models for analysis of repeated measurements. *Journal of the American Statistical Association*, 103, 1556–1569.

Wang, X., Jiang, Y., Huang, M., & Zhang, H. (2013). Robust variable selection with exponential squared loss. *Journal of the American Statistical Association*, 108, 632–643.

Yao, W., Lindsay, B. G., & Li, R. (2012). Local modal regression. *Journal of Nonparametric Statistics*, 24, 647–663.

Yao, W. & Wang, Q. (2013). Robust variable selection through MAVE. *Computational Statistics & Data Analysis*, 63, 42–49.

Zhang, C.-H. (2010). Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics*, 38, 894–942.

Zhang, J.-T. & Chen, J. (2007). Statistical inferences for functional data. *The Annals of Statistics*, 35, 1052–1079.

Zhou, S., Shen, X., & Wolfe, D. A. (1998). Local asymptotics for regression splines and confidence regions. *The Annals of Statistics*, 26, 1760–1782.

Zhu, H., Li, R., & Kong, L. (2012). Multivariate varying coefficient model for functional responses. *The Annals of Statistics*, 40, 2634–2666.

Zhu, H., Zhang, H., Ibrahim, J. G., & Peterson, B. S. (2007). Statistical analysis of diffusion tensors in diffusion-weighted magnetic resonance imaging data. *Journal of the American Statistical Association*, 102, 1085–1102.

Received 2 August 2020
Accepted 22 May 2021

DOI: 10.1002/cjs.11661