mathematics

MDPI

Article

Robust Variable Selection for Single-Index Varying-Coefficient Model with Missing Data in Covariates

Yunquan Song, Yaqi Liu * and Hang Su

College of Science, China University of Petroleum, Qingdao 266580, China; syqfly1980@upc.edu.cn (Y.S.); z21090004@s.upc.edu.cn (H.S.)
* Correspondence: z20090005@s.upc.edu.cn

Abstract: As applied sciences grow by leaps and bounds, semiparametric regression analyses have broad applications in various fields, such as engineering, finance, medicine, and public health. Single-index varying-coefficient model is a common class of semiparametric models due to its flexibility and ease of interpretation. The standard single-index varying-coefficient regression models consist mainly of parametric regression and semiparametric regression, which assume that all covariates can be observed. The assumptions are relaxed by taking the models with missing covariates into consideration. To eliminate the possibility of bias due to missing data, we propose a probability weighted objective function. In this paper, we investigate the robust variable selection for a single-index varying-coefficient model with missing covariates. Using parametric and nonparametric estimates of the likelihood of observations with fully observed covariates, we examine the estimators for estimating the likelihood of observations. For variable selection, we use a weighted objective function penalized by a non-convex SCAD. Theoretical challenges include the treatment of missing data and a single-index varying-coefficient model that uses both the non-smooth loss function and the non-convex penalty function. We provide Monte Carlo simulations to evaluate the performance of our approach.

Keywords: single-index varying-coefficient model; missing data; variable selection; inverse probability weighting; sparsity

MSC: 62F12; 62G08; 62G20; 62J07T07

Citation: Song, Y.; Liu, Y.; Su, H. Robust Variable Selection for Single-Index Varying-Coefficient Model with Missing Data in Covariates. *Mathematics* **2022**, *10*, 2003. https://doi.org/10.3390/math10122003

Academic Editor: Christophe Chesneau

Received: 4 April 2022
Accepted: 1 June 2022
Published: 10 June 2022

1. Introduction

Traditional statistical techniques are based on completely observed data. However, in many scientific experiments, such as questionnaire survey, medical research and psychological science, respondents are unwilling to provide some information which the researchers need. In addition, there are many factors that cannot be controlled in the research process, and it is often impossible to obtain all the desired data. When data are missing, traditional statistical techniques cannot be directly applied. Some statisticians consider using the observed data to draw valid conclusions in this situation. Until now, in order to deal with missing data, various methods have been employed such as complete-case analysis (CC) (Yates [1] and Healy and Westmacott [2]), imputation and inverse probability weighting (IPW), and methods based on likelihood. The IPW method proposed by Horvitz and Thompson [3] a way to deal with the missing data problems, which selects the inverse of the probability as the estimated weight so that it is not distorted by random missing data. It has earned extensive attention in the field of missing data research. There are also some related literatures, such as Robins et al. [4], Wang et al. [5], Little and Rubin [6], Liang et al. [7], Tsiatis [8], etc. However, when the error distribution is highly tailed or skewed, the results of the two aforementioned methods are not stable because they are based on least squares (LS) method.

In most regression models, it is critical to choose the proper loss function $\rho(\cdot)$ to make the resulting estimator robust. Therefore, researchers pay more attention to loss functions that have higher robustness. The exponential squared loss that has robustness is defined as $\psi_\eta(t) = 1 - \exp(-t^2/\eta)$, where η is the tuning parameter that determines the robustness degree of the estimator. For large η, $\psi_\eta(t)$ is approximately equal to t^2/η. Thus the proposed estimator is the same as the LS estimator in some extreme circumstances. When η is small, observations with absolute values of $t_i = Y_i - \mathbf{x}_i^T\beta$ will lead to a great loss of $\psi_\eta(t_i)$, whose influence upon the estimate of β is insiganificant. Thus, making η smaller limits the impact of outliers on the estimator but also reduces the sensitivity of the estimator. Moreover, quantile regression (QR) has become an increasingly popular method because regression methods based on exponential squared loss are more resistant to the effects of outliers than LS. Such exponential loss functions have been used in classification problems in AdaBoost (Friedman et al. [9]) and variable selection in regression models (Wang et al. [10]).

As applied sciences grow, research on semiparametric models has been extensively developed due to the high degree of flexibility and ease of interpretation. The singleindex varying-coefficient model (SIVCM) is a common semiparametric model. The main advantage of the model is that it avoids the curse of dimensionality. Another is that it has the explanatory power like parametric models. Generally, it takes the following form

$$\boldsymbol{Y} = \boldsymbol{g}^T(\boldsymbol{\beta_0^T}X)Z + \boldsymbol{\varepsilon}, \tag{1}$$

where $\boldsymbol{Y}$ is the dependent variable, (X, Z) are the covariates and $(X, Z) : R^p \times R^q$. $\boldsymbol{g}(\cdot)$ and $\boldsymbol{\beta_0}$ represent the vector of unknown functions and unknown parameters, respectively, whose dimension are $q \times 1$ and $p \times 1$. $\boldsymbol{\varepsilon}$ is the disturbance term with zero mean and finite variance σ^2 which is independent of (X, Z). Furthermore, assume that the Euclidean norm of $\boldsymbol{\beta_0}$ is equal to 1 and its first component is positive. Moreover, in order to avoid the influence due to the lack of uniqueness of the index direction $\boldsymbol{\beta_0}$, $g(x)$ cannot take the form of $g(x) = \boldsymbol{\alpha}^T\boldsymbol{x}\boldsymbol{\beta_0^T}\boldsymbol{x} + \boldsymbol{\gamma}^T\boldsymbol{x} + c$, where α, γ, c are constants, $\boldsymbol{\alpha} \in R^p$, $\boldsymbol{\gamma} \in R^p$, $c \in R$ and $\boldsymbol{\beta_0}$ are not parallel to each other (Feng and Xue [11]; Xue and Pang [12]).

Model (1) is so flexible that it covers a class of significant statistical models. It becomes the standard single-index model (SIM) when $Z = 1$ and $q = 1$; for related literatures, see Hardle et al. [13] and Wu et al. [14]. When $\beta_0 = 1$ and $p = 1$, it is simplified to the varying coefficient models (VCM) proposed by Hastie and Tibshirani [15] and Fan and Zhang [16]. Consequently, it is easily interpretable and has broad applications in practice. In particular, Xia and Li [17] first studied Model (1) using the kernel smoothing method with the LS method. The empirical likelihood ratio method was proposed by Xue and Wang [18]. Based on estimating equations, the estimate of the parametric component was built by Xue and Pang [12]. Using the function approximation, Feng and Xue [11] investigated Model (1).

Variable selection is of great importance to statistical modeling. The reason is that it will cause seriously biased results if researchers ignore the significant variables, whereas including spurious variables suffers from substantial loss in estimation efficiency. Hence, there are many popular choices for penalty functions, such as least absolute shrinkage and selection operator (LASSO, Tibshirani [19]), bridge penalty, smoothly clipped absolute deviation (SCAD, Fan and Li [20]), and adaptive lasso (Zou [21]). In particular, the non-conave least-squares penalty method based on SCAD penalization in SIM has been proposed by Peng and Huang [22] using SCAD penalization; Yang and Yang [23] adopted the SCAD penalty to achieve efficient estimation and variable selection simultaneously in partially linear single-index models (PLSIM); Wang and Kulasekera [24] proposed the partial linear varying-coefficient model (PLVCM) based on adaptive lasso.

SIVCM is a common semiparametric model. The selection of variables in semiparametric models includes two parts: the selection of the model in the nonparametric part and the selection of significant variables in the parametric part. Classical variable selection procedure involves stepwise regression and optimal subsets selection. However, the nonparametric parts of each submodel need to be extracted separately, leading to high

computational cost. It is a great challenge to select variables in SIVCM for the reason that it has a complex multivariate nonlinear structure that incudes both a nonparametric function vector $\boldsymbol{g}(\cdot)$ and an unknown parameter vector $\boldsymbol{\beta}$. Based on the approximation of the SCAD function and penalties, Feng and Xue [11] developed a penalty method for SIVCM. The method they propose allows the selection of significant variables into parametric and nonparametric components. It should be noted that existing research adopts the LS or likelihood method and assume that the error follows a normal distribution. Therefore, when the error is highly tailed, it makes the method sensitive to outliers and it becomes inefficient. It is not robust to outliers in the dependent variable due to using least squares criterion. Yang and Yang [25] proposed an efficient iterative procedure for SIVCM based on quantile regression. The results indicate that the resulting estimator is robust without accounting for both outliers and errors of variation. However, all existing work on SIVCM assumes that all variables are fully observed. A robust variable selection approach for SIVCM with missing covariates has not yet been studied.

The following are the innovations of this paper:

1. For the case of missing covariates, we propose a robust variable selection approach based on exponential squared loss and adopt the IPW method to eliminate the latent bias due to the missing values in covariates.
2. We consider parametric and nonparametric methods to estimate the probabilistic model and propose a objective function with a weighted penalty for variable selection.
3. We also examine how to select the parameters η of the squared exponential loss function to ensure that the corresponding estimator is robust.

The rest of this article is organized as follows. Section 2 proposes an efficient iterative SIVCM method using exponential quadratic loss, and the SCAD penalty is applied to select both important parametric variables and nonparametric components. In addition, we discusses the implementation, including bandwidth selection and tuning parameters. Section 3 conducts several Monte Carlo experiments with different error distributions in order to show the finite sample performance of the proposed method. Section 4 concludes the paper briefly.

2. Methodology

Using the exponential squared loss functions, the basis function approximation, and the SCAD penalty function, a robust variable selection procedure for SIVCM with missing covariates is proposed. First, the unknown coefficient functions are approximated applying the B-spline function. Next, under the constraint of $\|\boldsymbol{\beta}\| = 1$, we use the 'delete-one-component' approach constructed by Yu and Ruppert [26] in order to establish the objective function of the penalized exponential squared loss.

2.1. Basis Function Expansion

Consider that $\{(\boldsymbol{X_i}, \boldsymbol{Z_i}, Y_i), 1 \leq i \leq n\}$ is a sample from model (1), i.e.,

$$Y_i = \boldsymbol{g}^T(\boldsymbol{\beta}^T \boldsymbol{X_i})\boldsymbol{Z_i} + \varepsilon_i, \;\; i = 1, \cdots, n, \tag{2}$$

where $\boldsymbol{X_i} = (X_{i1}, \cdots, X_{ip})^T$ and $\boldsymbol{Z_i} = (Z_{i1}, \cdots, Z_{iq})^T$ are p-dimensional and q-dimensional independent variables, respectively. The disturbance term ε_i is unobserved random variable with zero mean and finite variance σ^2. We assume that $\{\varepsilon_i, 1 \leq i \leq n\}$ are independent of $\{(\boldsymbol{X_i}, \boldsymbol{Z_i}), 1 \leq i \leq n\}$.

In order to get the unknown $\boldsymbol{g}(\cdot)$, according to He et al. [27], we use its basis function approximations to replace the original $\boldsymbol{g}(\cdot)$. More specifically, construct a B-spline basis function of order M+1, $B(u) = (B_1(u), \cdots, B_L(u))^T$, where $L = K + M + 1$, and K is the number of interior knots. We can approximate $g_k(u)$ as

$$g_k(u) \approx B^T(u)\gamma_k, \;\; k = 1, \cdots, q, \tag{3}$$

where γ_k is the vector of the spline coefficient. The following robust estimation procedure will be performed if all the data $\{(\boldsymbol{X_i}, \boldsymbol{Z_i}, Y_i), 1 \leq i \leq n\}$ could be collected.

$$\ell(\boldsymbol{\beta}, \boldsymbol{g}(\cdot)) = \sum_{i=1}^{n} \exp\{-(Y_i - \boldsymbol{g}^T(\boldsymbol{\beta}^T \boldsymbol{X_i})\boldsymbol{Z_i})^2/\eta_n\}, \tag{4}$$

where $\eta_n > 0$ is a tuning parameter. To prevent outliers from affecting the estimate, we introduce an exponential squared loss in (4). However, (4) cannot be directly optimized when $\boldsymbol{g}(\cdot)$ is unknown. After we replace the unknown $\boldsymbol{g}(\cdot)$ by its basis function approximations in (4), we get

$$\ell_n(\boldsymbol{\beta}, \boldsymbol{\gamma}) = \sum_{i=1}^{n} \exp\{-(Y_i - W_i^T(\boldsymbol{\beta})\boldsymbol{\gamma})^2/\eta_n\}, \tag{5}$$

where $\boldsymbol{\gamma} = (\boldsymbol{\gamma_1}^T, \cdots, \boldsymbol{\gamma_q}^T)^T$, $W_i(\beta) = I_p \otimes B(\boldsymbol{\beta}^T \boldsymbol{X_i}) \cdot Z_i$.

We first handle the constraints $||\boldsymbol{\beta}|| = 1$ and $\beta_1 > 0$ on the p-dimensional single index parameter vector $\boldsymbol{\beta}$ by reparametrization. Denote $\boldsymbol{\phi} = (\beta_2, \cdots, \beta_p)^T$ and define

$$\boldsymbol{\beta} = \boldsymbol{\beta}(\boldsymbol{\phi}) = (\sqrt{1 - ||\boldsymbol{\phi}||^2}, \boldsymbol{\phi}^T)^T. \tag{6}$$

The true parameter $\boldsymbol{\phi_0}$ must satisfy $||\boldsymbol{\phi_0}|| < 1$, which is an inequality constraint. Therefore, $\boldsymbol{\beta}(\boldsymbol{\phi})$ is infinitely differentiable with respect to $\boldsymbol{\phi}$. Therefore, the Jacobian matrix of $\boldsymbol{\beta}$ with respect to $\boldsymbol{\phi}$ is

$$J_{\boldsymbol{\phi}} = \begin{pmatrix} -(1 - ||\boldsymbol{\phi}||^2)^{-1/2}\boldsymbol{\phi}^T \\ I_{p-1} \end{pmatrix},$$

where I_q is the q-order identity matrix. As we can see, $\boldsymbol{\phi}$ is one dimension lower than $\boldsymbol{\beta}$, and the penalized robust regression with the exponential squared loss is converted to

$$\ell_n(\boldsymbol{\phi}, \boldsymbol{\gamma}) = \sum_{i=1}^{n} \exp\{-(Y_i - W_i^T(\boldsymbol{\phi})\boldsymbol{\gamma})^2/\eta_n\}, \tag{7}$$

where $W_i(\boldsymbol{\phi}) = W_i(\boldsymbol{\beta})$. By maximizing (7),we can get $\hat{\boldsymbol{\phi}}$ and $\hat{\boldsymbol{\gamma}} = (\hat{\gamma_1}^T, \cdots, \hat{\gamma_q}^T)^T$. Then, through (3) and (6), the robust regression estimator of $\boldsymbol{\beta}$ based on the exponential squared loss is

$$\hat{\boldsymbol{\beta}} = (\sqrt{1 - ||\hat{\boldsymbol{\phi}}||^2}, \hat{\boldsymbol{\phi}}^T)^T, \tag{8}$$

and the estimator of $g_k(u)$ can be procured by

$$\hat{g}_k(u) = B^T(u)\hat{\boldsymbol{\gamma_k}}. \tag{9}$$

2.2. Robust Estimation Based on Inverse Probability Weighting

We consider the case where a subset of covariates has missing values when estimating (5). Let $\mathbf{1_i} \in R^{p+q-k}$ be the vector of always obtained covariates and $\boldsymbol{m_i} \in R^k$ is a vector of covariates that may contain some missing parts from $\boldsymbol{X_i}$ or $\boldsymbol{Z_i}$. We define the vector of variables which can be always observed as $\boldsymbol{t_i} = (Y_i, \mathbf{1}_i^T)^T \in R^s$,and $s = p + q - k$. Based on each observation, the value of an indicator variable R is related to whether $\boldsymbol{m_i}$ is completely observed ,which can be obtained by the following formula

$$R_i = \begin{cases} 1, & \text{if } \boldsymbol{m_i} \text{ is observed,} \\ 0, & \text{otherwise.} \end{cases}$$

The missing mechanism we proposed satisfies:

$$P(R_i = 1|Y_i, \boldsymbol{X_i}, \boldsymbol{Z_i}) = P(R_i = 1|Y_i, \boldsymbol{m_i}, \boldsymbol{t_i}) = P(R_i = 1|\boldsymbol{t_i}) \equiv \pi(\boldsymbol{t_i}) \equiv \pi_i, \tag{10}$$

With this missing mechanism, under the condition of $\boldsymbol{t_i}$, we can ensure the event that $\boldsymbol{m_i}$ is missing has no connection with $(Y_i, \boldsymbol{m_i}^T)$. Although the response data are fully observed, the selection probability $\pi(\cdot)$ in (10) still only related to the observed covariates $\boldsymbol{t_i}$ instead of the observed response. Therefore, we conclude that the missing mechanism is different from the missing at random (MAR) mechanism. We need this missing mechanism in order to continue the theoretical research.

When faced with missing covariates, we estimate (5) with a naive approach; only observations with complete data are used to fit the model. The naive estimator is

$$(\hat{\boldsymbol{\phi}}^N, \hat{\boldsymbol{\gamma}}^N) = argmax \sum_{i=1}^{n} R_i \exp\{-(Y_i - W_i^T(\boldsymbol{\phi})\boldsymbol{\gamma})^2/\eta_n\}, \tag{11}$$

while all observations with missing data are dropped when we estimate the model. Under the assumption that it is not the MAR, this estimator will be asymptotically biased.

An objective function based on inverse probability weights (IPW) is proposed in order to reduce the potential error caused by missing data. The expression R_i/π_{i0} is used to weight the ith data point in the IPW method. The difference between IPW and naive method is that IPW provides different weights for records with fully observed data. The idea behind weighting is that for every fully observed data point with probability π_{i0} of being fully observed, $1/\pi_{i0}$ data points with the same covariates are expected if there were no missing data.

The weight $1/\pi_{i0}$ is usually unknown and needs to be estimated. We consider estimating the weights using a parametric model. The general parametric relationship of the parametric model is assumed as

$$\pi_{i0} \equiv \pi_i(\boldsymbol{t_i}, \eta_0).$$

Assuming the logistic relationship as an example

$$\pi_i(\boldsymbol{t_i}, \eta_0) = \frac{\exp\{(1, \boldsymbol{t_i})^T \eta_0\}}{1 + \exp\left\{(1, \boldsymbol{t_i})^T \eta_0\right\}}.$$

In practice $\pi_i(\boldsymbol{t_i}, \eta_0)$ is replaced with $\pi_i(\boldsymbol{t_i}, \hat{\eta}) \equiv \pi_i(\hat{\eta})$. The parametric model $P(R_i = 1|\boldsymbol{t_i})$ is used to estimate $\hat{\eta}$.

Throughout the paper $\pi_i(\hat{\eta})$ will denote the parametric estimate, $\hat{\pi}_i$ will denote a general estimate that could be parametric, and π_{i0} will denote the true probability when observation i has full data. The definition of our parametric robust regression estimator is

$$(\hat{\boldsymbol{\phi}}^L, \hat{\boldsymbol{\gamma}}^L) = argmax \sum_{i=1}^{n} \frac{R_i}{\pi_i(\hat{\eta})} \exp\{-(Y_i - W_i^T(\boldsymbol{\phi})\boldsymbol{\gamma})^2/\eta_n\}. \tag{12}$$

According to the above, through (3) and (6) and using the exponential squared loss,β can be robustly estimated by

$$\hat{\boldsymbol{\beta}}^L = (\sqrt{1 - ||\hat{\boldsymbol{\phi}}^L||^2}, (\hat{\boldsymbol{\phi}}^L)^T)^T. \tag{13}$$

Then, the estimator of $g_k(u)$ can be written as

$$\hat{g}_k^L(u) = B^T(u)\hat{\boldsymbol{\gamma}_k}^L. \tag{14}$$

2.3. The Penalized Robust Regression Estimator

Here we consider the variable selection problem when Model (2) has missing covariates. In order to improve the accuracy and interpretability of model fitting and ensure the identifiability of the model, the vector of the real regression coefficient $\boldsymbol{\beta}^*$ is generally set to a scattered state with only a small fraction of non-zeroes (Fan and Li [20]; Tibshirani [19]).

For the purpose of getting the true model and estimating $\boldsymbol{\beta}^*$ and $\boldsymbol{g}(\cdot)$, a penalized robust regression that uses exponential squared loss is as follows

$$\begin{aligned}\ell(\boldsymbol{\beta}, \boldsymbol{g}(\cdot)) = &\sum_{i=1}^{n} \frac{R_i}{\pi_i(\hat{\eta})} \exp\{-(Y_i - g^T(\boldsymbol{\beta}^T \mathbf{X_i})\mathbf{Z_i})^2/\eta_n\} \\ &- n\lambda_1 \sum_{k=1}^{q} p_{\lambda_{1k}}(||g_k(\cdot)||) - n\lambda_2 \sum_{l=1}^{p} p_{\lambda_{2l}}(||\beta_l||),\end{aligned} \tag{15}$$

where

$$||g_k(\cdot)|| = (\int g_k^2(u)du)^{1/2}.$$

The penalty function $p_\lambda(\cdot)$ is defined on the interval $[0, \infty)$ and the regularization parameter λ is non-negative. It is necessary to emphasize that the tuning parameters λ_1 and λ_2 have no need to be the same for all $g_k(\cdot)$ and $\boldsymbol{\beta_l}$. Our purpose of using exponential squared loss in (5) is to prevent outliers from affecting the estimation process. It is unrealistic to directly optimize (15) when $\boldsymbol{g}(\cdot)$ is unknown. To solve this problem, the unknown function $\boldsymbol{g}(\cdot)$ in (15) is replaced by its basis function approximation, which can be written as

$$\begin{aligned}\ell_n(\boldsymbol{\beta}, \boldsymbol{\gamma}) = &\sum_{i=1}^{n} \frac{R_i}{\pi_i(\eta)} \exp\{-(Y_i - W_i^T(\boldsymbol{\beta})\gamma)^2/\eta_n\} \\ &- n\lambda_1 \sum_{k=1}^{q} p_{\lambda_{1k}}(||\boldsymbol{\gamma_k}||_H) - n\lambda_2 \sum_{l=1}^{p} p_{\lambda_{2l}}(||\beta_l||),\end{aligned} \tag{16}$$

where $||\boldsymbol{\gamma_k}||_H = (\boldsymbol{\gamma_k}^T H \boldsymbol{\gamma_k})^{1/2}$, $H = \int B(u)B^T(u)du$.

When π_i's parametric estimate is $\pi_i(\hat{\eta})$, the parametric penalized robust regression with the exponential squared loss transforms to

$$\begin{aligned}\ell_n(\boldsymbol{\phi}, \boldsymbol{\gamma}) = &\sum_{i=1}^{n} \frac{R_i}{\pi_i(\hat{\eta})} \exp\{-(Y_i - W_i^T(\boldsymbol{\phi})\gamma)^2/\eta_n\} \\ &- n\lambda_1 \sum_{k=1}^{q} p_{\lambda_{1k}}(||\boldsymbol{\gamma_k}||_H) - n\lambda_2 \sum_{l=1}^{p-1} p_{\lambda_{2l}}(||\phi_l||),\end{aligned} \tag{17}$$

where $W_i(\boldsymbol{\phi}) = W_i(\boldsymbol{\beta})$. By maximizing (17), we can get the result $\hat{\boldsymbol{\phi}}^P$ and $\hat{\gamma}^P = (\hat{\boldsymbol{\gamma_1}}^T, \cdots, \hat{\boldsymbol{\gamma_q}}^T)^T$. Then, through (3) and (6), the penalized robust regression estimator of $\boldsymbol{\beta}$ based on the exponential squared loss is

$$\hat{\boldsymbol{\beta}}^P = (\sqrt{1 - ||\hat{\boldsymbol{\phi}}^P||^2}, (\hat{\boldsymbol{\phi}}^P)^T)^T, \tag{18}$$

and the estimator of $g_k(u)$ can be obtained by

$$\hat{g}_k^P(u) = B^T(u)\hat{\boldsymbol{\gamma_k}}^P. \tag{19}$$

2.4. Algorithm

A quadratic approximation is used to replace the loss function for the purpose of facilitating the computation. Let

$$\ell^*(\boldsymbol{\phi}, \boldsymbol{\gamma}) = \sum_{i=1}^{n} \frac{R_i}{\pi_i(\hat{\eta})} \exp\{-(Y_i - W_i^T(\boldsymbol{\phi})\gamma)^2/\eta_n\}.$$

When we get the initial estimator $(\tilde{\boldsymbol{\phi}}, \tilde{\gamma})$, then the loss function can be approximated as

$$\ell^*(\boldsymbol{\phi}, \boldsymbol{\gamma}) \approx \ell^*(\tilde{\boldsymbol{\phi}}, \tilde{\gamma}) + \frac{1}{2}\{(\boldsymbol{\phi}, \boldsymbol{\gamma}) - (\tilde{\boldsymbol{\phi}}, \tilde{\gamma})\}^T \nabla^2 \ell^*(\tilde{\boldsymbol{\phi}}, \tilde{\gamma})\{(\boldsymbol{\phi}, \boldsymbol{\gamma}) - (\tilde{\boldsymbol{\phi}}, \tilde{\gamma})\}.$$

What makes implementing the Newton–Raphson algorithm directly difficult is that the SCAD-penalty function is irregular at the origin. Now, we develop an iterative algorithm based on the local quadratic approximation of the penalty function $p_\lambda(\cdot)$ as in Fan and Li [20]. More specially, in a neighborhood of a given nonzero ω_0, an approximation of the penalty function at the value ω_0 can be given by

$$p_\lambda(|\omega|) \approx p_\lambda(|\omega_0|) + \frac{1}{2}\frac{\dot{p}_\lambda(\omega_0)}{|\omega_0|}(\omega^2 - \omega_0^2).$$

Hence, for the given initial value ϕ_l^0 with $|\phi_l^0| > 0, l = 1, \cdots, p-1$, and γ_k^0 with $||\gamma_k^0||_H > 0$, $k = 1, \cdots, q$, we have

$$p_{\lambda_{1k}}(||\gamma_k||_H) \approx p_{\lambda_{1k}}(||\gamma_k^0||_H) + \frac{1}{2}\frac{\dot{p}_{\lambda_{1k}}(||\gamma_k^0||_H)}{||\gamma_k^0||_H}(||\gamma_k||_H^2 - ||\gamma_k^0||_H^2),$$

$$p_{\lambda_{2l}}(|\phi_l|) \approx p_{\lambda_{2l}}(|\phi_l^0|) + \frac{1}{2}\frac{\dot{p}_{\lambda_{2l}}(|\phi_l^0|)}{|\phi_l^0|}(|\phi_l|^2 - |\phi_l^0|^2).$$

Let

$$\Sigma(\boldsymbol{\phi}, \boldsymbol{\gamma}) = \mathrm{diag}\{\frac{\dot{p}_{\lambda_{21}}(|\phi_1|)}{|\phi_1|}, \cdots, \frac{\dot{p}_{\lambda_{2,p-1}}(|\phi_{p-1}|)}{|\phi_{p-1}|}, \frac{\dot{p}_{\lambda_{11}}(||\boldsymbol{\gamma_1}||_H)}{||\boldsymbol{\gamma_1}||_H}H, \cdots, \frac{\dot{p}_{\lambda_{1q}}(||\boldsymbol{\gamma_q}||_H)}{||\boldsymbol{\gamma_q}||_H}H\}.$$

Then, in addition to the constant term, we maximize

$$\begin{aligned}\ell(\boldsymbol{\phi}, \boldsymbol{\gamma}) = &\frac{1}{2}\{(\boldsymbol{\phi}, \boldsymbol{\gamma}) - (\tilde{\boldsymbol{\phi}}, \tilde{\boldsymbol{\gamma}})\}^T \nabla^2 \ell^*(\tilde{\boldsymbol{\phi}}, \tilde{\boldsymbol{\gamma}})\{(\boldsymbol{\phi}, \boldsymbol{\gamma}) - (\tilde{\boldsymbol{\phi}}, \tilde{\boldsymbol{\gamma}})\} \\ &- \frac{n}{2}(\boldsymbol{\phi}^T, \boldsymbol{\gamma}^T)\Sigma(\boldsymbol{\phi}, \boldsymbol{\gamma})(\boldsymbol{\phi}^T, \boldsymbol{\gamma}^T)^T\end{aligned} \tag{20}$$

with respect to $\boldsymbol{\phi}$ and $\boldsymbol{\gamma}$, which brings about an approximated solution of (17). We can get estimates $\hat{\beta}$ and $\hat{g}_k(u)$ of $\boldsymbol{\beta}$ and $g_k(u)$ by solving for (3) and (6) respectively.

In order to implement the above method, we should correctly choose the number of interior knots K and make appropriate adjustments to the tuning parameters a, λ_1, λ_2 and η_n in the penalty function. Fan and Li [20] showed that the choice of $a = 3.7$ performs well in variety of situations. Hence, we also follow their setup in this article.

2.5. The Choice of the Regularization Parameter λ_1 and λ_2

We can choose the tuning parameters using a method that is similar to cross-validation. However, our penalty function contains too many tuning parameters, and higher-dimensional space makes it difficult to solve the minimization problem for the cross-validation score. To overcome this difficulty, similar to Zhao and Xue [28], we take the tuning parameters as

$$\lambda_1 = \frac{\lambda}{||\hat{\gamma_k}^u||_H}, \quad \lambda_2 = \frac{\lambda}{||\hat{\boldsymbol{\phi}_l}^u||}, \tag{21}$$

where $\hat{\gamma_k}^u$ and $\hat{\phi_l}^u$ are the unpenalized estimators of γ_k^u and ϕ_l^u, respectively. Then, we can estimate λ and K by minimizing the following cross-validation score:

$$CV(K, \lambda) = \sum_{i=1}^{n}\{Y_i - W_i^T(\hat{\boldsymbol{\phi}_{[i]}})\hat{\gamma_{[i]}}\}^2, \tag{22}$$

where $\hat{\boldsymbol{\phi}_{[i]}}$ and $\hat{\gamma_{[i]}}$ are the solutions ground on (17) after deleting the ith subject.

2.6. The Choice of the Regularization Parameter η_n

The tuning parameter η_n plays a decisive role in the degree of robustness and efficiency of the proposed robust regression estimators. A data-driven procedure is proposed to

choose the appropriate η_n, the new method yields both high efficiency and high robustness simultaneously. We first choose a series of the tuning parameters that makes the proposed penalized robust estimators have an asymptotic breakdown point at 1/2 and then use the maximum efficiency as a measure to select the tuning parameter.

The specific procedure steps are as follows:

Step 1 In this step, we will find the pseudo outlier set of the sample as in Wang et al. [10]. Let $D_i = (\boldsymbol{X_i}, \boldsymbol{Z_i}, Y_i)$ and $D = (D_1, \cdots, D_n)$. Calculate $r_i(\hat{\boldsymbol{\phi}}_{\boldsymbol{n}}, \hat{\boldsymbol{\gamma}}_{\boldsymbol{n}}) = Y_i - W_i^T(\hat{\boldsymbol{\phi}}_{\boldsymbol{n}})\hat{\boldsymbol{\gamma}}_{\boldsymbol{n}}$, $i = 1, \cdots, n$ and $S_n = 1.486 \times \text{median}_i|r_i(\hat{\boldsymbol{\phi}}_{\boldsymbol{n}}, \hat{\boldsymbol{\gamma}}_{\boldsymbol{n}}) - \text{median}_j(r_j(\hat{\boldsymbol{\phi}}_{\boldsymbol{n}}, \hat{\boldsymbol{\gamma}}_{\boldsymbol{n}})|$. Then, take the pseudo outlier set $D_m = \{(\boldsymbol{X_i}, \boldsymbol{Z_i}, Y_i) : |r_i(\hat{\boldsymbol{\phi}}_{\boldsymbol{n}}, \hat{\boldsymbol{\gamma}}_{\boldsymbol{n}})| > 2.5S_n\}$, set $m = \sharp\{1 \le i \le n : |r_i(\hat{\boldsymbol{\phi}}_{\boldsymbol{n}}, \hat{\boldsymbol{\gamma}}_{\boldsymbol{n}})| > 2.5S_n\}$, and $D_{n-m} = D_n/D_m$.

Step 2 In this step, we are going to update the tuning parameter η_n. Suppose there are m bad points and $n - m$ good points in D_n. Define the bad points by $D_m = (D_1, \cdots, D_m)$ and the good points by $D_{n-m} = (D_{m+1}, \cdots, D_n)$.

The proportion of bad points in D_n is m/n. The computation of the initial estimators $\tilde{\boldsymbol{\phi}}_{\boldsymbol{n}}$ and $\tilde{\boldsymbol{\gamma}}_{\boldsymbol{n}}$ is the first thing to do . For a contaminated sample D_n, let

$$\xi(\eta) = \frac{2m}{n} + \frac{2}{n}\sum_{m+1}^{n} \psi_\eta\{r_i(\tilde{\boldsymbol{\phi}}_{\boldsymbol{n}}, \tilde{\boldsymbol{\gamma}}_{\boldsymbol{n}})\}, \tag{23}$$

where $r_i(\boldsymbol{\phi}, \boldsymbol{\gamma}) = Y_i - W_i^T(\boldsymbol{\phi})\boldsymbol{\gamma}$. Let η_n be the minimizer of $\det(\hat{V}(\eta))$ in the set $G = \{\eta : \xi(\eta) \in (0, 1]\}$, where $\det(\cdot)$ indicate the determinant operator,

$$\hat{V}(\eta) = \{\hat{I}_1(\hat{\boldsymbol{\phi}}_{\boldsymbol{n}}, \hat{\boldsymbol{\gamma}}_{\boldsymbol{n}})\}^{-1}\tilde{\Sigma}_2\{\hat{I}_1(\hat{\boldsymbol{\phi}}_{\boldsymbol{n}}, \hat{\boldsymbol{\gamma}}_{\boldsymbol{n}})\}^{-1},$$

and

$$\hat{I}_1(\hat{\boldsymbol{\phi}}_{\boldsymbol{n}}, \hat{\boldsymbol{\gamma}}_{\boldsymbol{n}}) = \frac{2}{\eta}\left\{\frac{1}{n}\sum_{i=1}^{n}\exp(-r_i^2(\hat{\boldsymbol{\phi}}_{\boldsymbol{n}}, \hat{\boldsymbol{\gamma}}_{\boldsymbol{n}})/\eta)\left(\frac{2r_i^2(\hat{\boldsymbol{\phi}}_{\boldsymbol{n}}, \hat{\boldsymbol{\gamma}}_{\boldsymbol{n}})}{\eta} - 1\right)\right\} \times \left(\frac{1}{n}\sum_{i=1}^{n} W_i W_i^T\right),$$

$$\tilde{\Sigma}_2 = \text{cov}\left\{\exp(-r_1^2(\hat{\boldsymbol{\phi}}_{\boldsymbol{n}}, \hat{\boldsymbol{\gamma}}_{\boldsymbol{n}})/\eta)\frac{2r_1(\hat{\boldsymbol{\phi}}_{\boldsymbol{n}}, \hat{\boldsymbol{\gamma}}_{\boldsymbol{n}})}{\eta}W_1, \cdots, \exp(-r_n^2(\hat{\boldsymbol{\phi}}_{\boldsymbol{n}}, \hat{\boldsymbol{\gamma}}_{\boldsymbol{n}})/\eta)\frac{2r_n(\hat{\boldsymbol{\phi}}_{\boldsymbol{n}}, \hat{\boldsymbol{\gamma}}_{\boldsymbol{n}})}{\eta}W_n\right\}.$$

Step 3 The value of λ can be calculated from (22). Then, we can get the value of λ_1 and λ_2 by (21). Through fixed λ_1 and λ_2, and selected η_n in Step 2, $\hat{\boldsymbol{\phi}}_{\boldsymbol{n}}$ and $\hat{\boldsymbol{\gamma}}_{\boldsymbol{n}}$ can be updated by maximizing (17).

Step 4 We learn from Xue and Pang [12] to set the estimator $\tilde{\boldsymbol{\phi}}$ and $\tilde{\boldsymbol{\gamma}}$ as the initial estimate, which means $\hat{\boldsymbol{\phi}} = \tilde{\boldsymbol{\phi}}$ and $\hat{\boldsymbol{\gamma}} = \tilde{\boldsymbol{\gamma}}$. We then repeat Steps 1-3 until $\hat{\boldsymbol{\phi}}$, $\hat{\boldsymbol{\gamma}}$, and η_n converge.

Step 5 Using (3) and (6), we get the penalized robust regression estimator $\hat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$, and the estimator $\hat{g}_k(u)$ of $g_k(u)$.

3. Simulation

Here we compare the performance of the estimation and variable selection methods we propose for the finite samples with that of Yang and Yang [25] (QR), Xue and Wang [18] (EL), Xue and Pang [12] (EE) via some Monte Carlo simulations. In contrast, Xue and Wang [18] (EL) and Xue and Pang [12] (EE) fail to take into account the problem of selection of significant variables, so we introduced an adaptive penalty term into their objective function to ensure that significant variables are selected.

According to Yang and Yang [25], we choose the Gaussian kernel function in the simulations of the quantile regression method with $\tau = 0.5$. Evaluation of the performance of the estimators noted above is based on the following three criteria: (1) the average absolute deviations (AAD) of the estimated coefficients and the standard deviations (SD) for each; (2) mean absolute deviations (MAD) of $\hat{\boldsymbol{\beta}}$, which can be calculated by the expression $MAD(\hat{\boldsymbol{\beta}}) = E(||\hat{\boldsymbol{\beta}} - \boldsymbol{\beta_0}||_1)$, where $||\cdot||_p$ represents the p-norm; and (3) the square root

of the average square error (RASE) as a measure of the performance of estimator $\hat{g}_k(\cdot)$, calculated as follows:

$$RASE_k = \{\frac{1}{n_{grid}} \sum_{i=1}^{n_{grid}} (\hat{g}_k(u_i) - g_k(u_i))^2\}^{1/2}$$

for $k = 1, \cdots, q$, where $\{u_i, i = 1, 2, \cdots, n_{grid}\}$ denote the grid points used to assess the function $g_k(\cdot)$.

Additionally, in order to demonstrate the effectiveness of the variable selection procedure, the average number of real zero coefficients accurately identified as zero (NC), the average number of real non-zero coefficients mistakenly identified as zero (NIC), as well as the probability of correctly selecting the real model (PC) are presented in our simulation. The tuning parameter η is chosen for each simulation sample.

Example 1. In this example, we focus attention on the estimation of the proposed estimation procedure, and the following SIVCM is considered:

$$\boldsymbol{Y} = g_1(X^T\boldsymbol{\beta_0}) + g_2(X^T\boldsymbol{\beta_0})Z_1 + g_3(X^T\boldsymbol{\beta_0})Z_2 + \boldsymbol{\varepsilon}, \tag{24}$$

where $\boldsymbol{\beta_0} = (\frac{1}{3}, \frac{2}{3}, \frac{2}{3})^T$, $X = (\boldsymbol{X_1}, \boldsymbol{X_2}, \boldsymbol{X_3})^T$, and $Z = (\boldsymbol{Z_1}, \boldsymbol{Z_2})^T$ are jointly normally distributed with mean 0, variance 1 and correlation $0.5^{|i-j|}$, $g_1(u) = 2\cos(\pi u)$, $g_2(u) = 1 + u^2/2$ and $g_3(u) = \exp(-u)$. The error ε and $\boldsymbol{X_1}$, $\boldsymbol{X_2}$, $\boldsymbol{X_3}$, $\boldsymbol{Z_1}$, $\boldsymbol{Z_2}$ are independent; $\boldsymbol{X_1}$ may have missing values. The selection probability functions are given by:

$$\pi_1(\boldsymbol{X_2}, \boldsymbol{X_3}, \boldsymbol{Z_1}, \boldsymbol{Z_2}) = \{1 + \exp(-(\gamma_0 + \gamma_1\boldsymbol{X_2} + \gamma_2\boldsymbol{X_3} + \gamma_3\boldsymbol{Z_1} + \gamma_4\boldsymbol{Z_2}))\}^{-1}.$$

We consider π_1 with $(\gamma_0, \gamma_1, \gamma_2, \gamma_3, \gamma_4) = (1, 0.2, 0.2, 0.4, 0.5)$. The corresponding average missing rates are 25%. In our simulation, three different distributions of model error $\boldsymbol{\varepsilon}$ are considered:

case1: The standard normal distribution $N(0, 1)$.

case2: The centralized t-distribution with three degrees of freedom $t(3)$ that is used to generate heavy-tailed distribution.

case3: The mixture of normals $0.9N(0, 1) + 0.1N(0, 100)(MN(1, 100))$ which is used to produce the outliers.

Table 1 displays the average absolute deviations (AAD) and the standard deviations (SD), as well as the mean absolute deviations (MAD), for each case with sample sizes $n = 50, 200, 400$. It can be seen that when the errors are normally distributed, our proposed estimator, based on the exponential loss squared (ESL), has smaller AAD, SD, and MAD than the $QR_{0.5}$, the estimating equations (EE) and the empirical likelihood ratio (EL) methods for all sample sizes, which means that the proposed estimator performs better than the other three estimators. The proposed estimator also gives good results for the other two error distributions, $t(3)$ and $MN(1, 100)$. The significant improvement in the performance of our proposed estimator over the EE, EL, and $QR_{0.5}$ estimators indicates that our proposed estimation method ESL is robust to datasets with outliers or error distributions of response variables with high tails. More importantly, as the sample size n increases, the performance of the estimator $\hat{\boldsymbol{\beta}}$ tends to improve significantly.

The square root of average square error (RASE) of the estimator $\hat{g}_k(\cdot)$for the nonparametric function $g_k(\cdot)$ with sample sizes of n = 50, 200. and 400 is reported in Table 2. Table 2 gives results similar to those in Table 1. We note that no matter which of the above three distributions the error follows, our proposed estimator, compared with the other three estimators, has smaller RASE and performs better. That is, for the non-normal distributions, our proposed estimate method ESL is consistently superior to QR, EE, and EL. When the probability of selection $\pi(\cdot)$ is correctly specified and estimated using the parametric model, a clear pattern emerges: as the sample size n increases, the performance of the two estimators $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{g}}(\cdot)$ becomes greater and greater.

Table 1. Simulation results of AAD ($\times 10^2$), SD ($\times 10^2$), and MAD ($\times 10^2$) for the estimators of $\beta_i (i = 1, 2, 3)$.

Dist	*n*	Method	$\hat{\beta}_1$		$\hat{\beta}_2$		$\hat{\beta}_3$		MAD
			AD	SD	AD	SD	AD	SD	
$N(0,1)$	50	ESL	0.0862	0.0842	0.0934	0.0940	0.0916	0.0928	0.0923
		$QR_{0.5}$	0.0924	0.0908	0.0945	0.0966	0.0928	0.0936	0.09379
		EE	0.6386	0.6812	0.6806	0.7013	0.7047	0.7014	0.6734
		EL	0.6418	0.6826	0.6814	0.702	0.7069	0.7022	0.6742
	200	ESL	0.0521	0.0517	0.0568	0.0655	0.0656	0.0637	0.0623
		$QR_{0.5}$	0.0643	0.0635	0.0696	0.0743	0.0704	0.0725	0.0739
		EE	0.4993	0.5014	0.4884	0.5113	0.5025	0.5204	0.4997
		EL	0.4998	0.5022	0.4892	0.5121	0.5033	0.5212	0.4999
	400	ESL	0.0467	0.0475	0.0482	0.0499	0.0491	0.0493	0.0478
		$QR_{0.5}$	0.0473	0.0489	0.0495	0.0504	0.0493	0.0497	0.0484
		EE	0.4306	0.4682	0.4673	0.4809	0.4835	0.4824	0.4531
		EL	0.4312	0.4694	0.4682	0.4816	0.4847	0.4830	0.4538
$t(3)$	50	ESL	1.8642	1.9342	1.9575	2.0416	1.9464	1.9488	1.9240
		$QR_{0.5}$	2.0468	2.1726	2.1934	2.2682	2.2610	2.3208	2.2627
		EE	4.2203	4.8418	5.7262	5.9258	5.4436	6.0240	5.1639
		EL	4.2217	4.8446	5.7280	5.9264	5.4475	6.0264	5.1653
	200	ESL	0.4734	0.4892	0.4957	0.5044	0.4936	0.4978	0.4846
		$QR_{0.5}$	0.4902	0.5013	0.5075	0.5184	0.5118	0.5250	0.5129
		EE	2.3115	2.7526	3.2385	3.5381	3.1327	3.5504	2.9215
		EL	2.3121	2.7532	3.2390	3.5388	3.1331	3.5508	2.9217
	400	ESL	0.0643	0.0635	0.0696	0.0743	0.0704	0.0725	0.0739
		$QR_{0.5}$	0.0713	0.0762	0.0728	0.0801	0.0697	0.0755	0.0784
		EE	1.4832	1.8364	2.3734	2.4119	2.3658	2.5706	2.1897
		EL	1.4838	1.8370	2.3742	2.4126	2.3664	2.5712	2.1903
$MN(1,100)$	50	ESL	2.2328	2.3444	2.3727	2.4228	2.4053	2.4264	2.4306
		$QR_{0.5}$	2.7892	2.8526	2.8913	2.9726	2.9436	3.1175	2.8328
		EE	4.6206	5.2304	5.8304	7.4631	5.6529	6.4336	5.9205
		EL	4.6224	5.2120	5.8316	7.4655	5.6542	6.4368	5.9217
	200	ESL	0.5036	0.5158	0.5525	0.5844	0.5534	0.5812	0.5406
		$QR_{0.5}$	0.5142	0.5276	0.5697	0.5982	0.5680	0.5903	0.5534
		EE	2.5612	3.0546	3.5274	4.9437	3.4935	4.1178	3.6893
		EL	2.5616	3.0550	3.5278	4.9443	3.4940	4.1182	3.6899
	400	ESL	0.0565	0.0547	0.0585	0.0657	0.0646	0.0621	0.0633
		$QR_{0.5}$	0.0720	0.0755	0.0718	0.0762	0.0722	0.0719	0.0743
		EE	1.5764	1.8035	2.4832	2.7761	2.5167	2.6839	2.3106
		EL	1.5772	1.8043	2.4838	2.7767	2.5171	2.6842	2.3110

Table 2. Simulation results of RASE for the estimators of $g_i(\cdot)(i = 1, 2, 3)$.

Dist	*n*	Method	$\hat{g}_1$	$\hat{g}_2$	$\hat{g}_3$
			RASE	RASE	RASE
$N(0,1)$	50	ESL	0.3647	0.2281	0.3872
		$QR_{0.5}$	0.3893	0.2463	0.3969
		EE	0.3854	0.2358	0.3905
		EL	0.3872	0.2364	0.3916
	200	ESL	0.0941	0.0915	0.1030
		$QR_{0.5}$	0.1083	0.1041	0.1152
		EE	0.1034	0.0967	0.1096
		EL	0.1038	0.0969	0.1098
	400	ESL	0.0323	0.0304	0.0357
		$QR_{0.5}$	0.0447	0.0428	0.0483
		EE	0.0333	0.0314	0.0446
		EL	0.0339	0.0320	0.0448

Table 2. *Cont.*

Dist	*n*	Method	$\hat{g}_1$	$\hat{g}_2$	$\hat{g}_3$
			RASE	**RASE**	**RASE**
$t(3)$	50	ESL	0.4028	0.3884	0.3916
		$QR_{0.5}$	0.4264	0.4152	0.4237
		EE	1.6802	1.4716	1.7231
		EL	1.6826	1.4738	1.7242
	200	ESL	0.1052	0.1056	0.1104
		$QR_{0.5}$	0.1188	0.1170	0.1219
		EE	0.7308	0.6131	0.7463
		EL	0.7314	0.6138	0.7469
	400	ESL	0.0366	0.0340	0.0485
		$QR_{0.5}$	0.0492	0.0476	0.0513
		EE	0.4120	0.3493	0.5042
		EL	0.4124	0.3495	0.5048
$MN(1,100)$	50	ESL	0.4356	0.3751	0.3938
		$QR_{0.5}$	0.4682	0.4065	0.4175
		EE	1.5145	1.4127	1.6127
		EL	1.5163	1.4203	1.6343
	200	ESL	0.1102	0.1045	0.1146
		$QR_{0.5}$	0.1228	0.1137	0.1201
		EE	0.7089	0.6715	0.7141
		EL	0.7093	0.6719	0.7147
	400	ESL	0.0379	0.0384	0.0361
		$QR_{0.5}$	0.0483	0.0496	0.0489
		EE	0.3747	0.3572	0.5387
		EL	0.3751	0.3577	0.5393

Example 2. This example aims to study the variable selection performance of the index parameters in model (1). The model setup is similar to (24) except that $X = (\boldsymbol{X_1}, \boldsymbol{X_2}, \cdots, \boldsymbol{X_8})^T$ independently generated from $[-1,1]^8$ and $\boldsymbol{\beta_0} = (\frac{1}{3}, \frac{2}{3}, 0, 0, \frac{2}{3}, 0, 0, 0)^T$. As considered in Example 1, three different error distributions $N(0,1)$, $t(3)$, and $MN(1,25)$ are considered to show the robustness of the proposed estimator method based on the exponential squared loss (ESL). The error $\boldsymbol{\varepsilon}$ and $\boldsymbol{X_1}, \cdots, \boldsymbol{X_8}, \boldsymbol{Z_1}, \boldsymbol{Z_2}$ are independent; $\boldsymbol{X_1}$ may have missing values. The selection probability functions are given by:

$$\pi_2(\boldsymbol{X_2}, \boldsymbol{X_5}, \boldsymbol{Z_1}, \boldsymbol{Z_2}) = \{1 + \exp(-(\gamma_0 + \gamma_1 \boldsymbol{X_2} + \gamma_2 \boldsymbol{X_5} + \gamma_3 \boldsymbol{Z_1} + \gamma_4 \boldsymbol{Z_2}))\}^{-1}.$$

We consider π_1 with $(\gamma_0, \gamma_1, \gamma_2, \gamma_3, \gamma_4,) = (1, 0.2, 0.2, 0.4, 0.5)$. The corresponding average missing rates are 25%.

For each mechanism mentioned above, we compare the performance of four methods: our proposed method [ESL-SCAD], LSE-SCAD proposed by Feng and Xue [11], LAD-SCAD proposed by Yang and Yang [25], and EE-SCAD method based on Xue and Pang [12]. The results are reported in Table 3 and are similar to the conclusions of Example 1. Whether the error term follows the normal distribution, the centralized t-distribution, or the mixture of normals, our proposed method performs more efficiently in variable selection, which has larger NC and smaller NIC. When there exist outliers in the response variables or heavy-tailed error distributions, ESL-SCAD has an obviously better performance than LAD-SCAD, EE-SCAD, or LSE-SCAD estimators. For normal error, ESL-SCAD hardly loses any efficiency.

The proposed procedure is also competitive in terms of computational cost. The calculation was performed on a computer with AMD Ryzen processors, a 16 GB RAM, running a Windows 10 system, and only one CPU was used for fair comparisons. Results on computational efficiency of the our proposed method are presented in Tables 4 and 5, which show CPU times (in seconds) for different combinations of the full data size n and the number of covariates p. It is seen that the proposed algorithm is faster.

Table 3. Variable selection results and RASE of $\hat{g}_k(\cdot)$, $k = 1, 2, 3$ in Example 2.

Dist	*n*	Method	NC	NIC	PC	$\mathbf{RASE_1}$	$\mathbf{RASE_2}$	$\mathbf{RASE_3}$
$N(0,1)$	50	ESL-SCAD	4.850	0	0.948	0.2351	0.2306	0.2412
		LAD-SCAD	4.820	0	0.942	0.2364	0.2318	0.2430
		LSE-SCAD	4.865	0	0.950	0.2336	0.2140	0.2384
		EE-SCAD	4.855	0	0.944	0.2340	0.2153	0.2396
	200	ESL-SCAD	4.945	0	0.962	0.1143	0.1132	0.1228
		LAD-SCAD	4.940	0	0.960	0.1187	0.1145	0.1256
		LSE-SCAD	4.955	0	0.970	0.1132	0.1065	0.1182
		EE-SCAD	4.950	0	0.965	0.1138	0.1071	0.1194
	400	ESL-SCAD	5.000	0	1.000	0.0467	0.0432	0.0556
		LAD-SCAD	5.000	0	1.000	0.0545	0.0526	0.0581
		LSE-SCAD	5.000	0	1.000	0.0423	0.0404	0.0536
		EE-SCAD	5.000	0	1.000	0.0429	0.0408	0.0540
$t(3)$	50	ESL-SCAD	4.924	0.006	0.948	0.2262	0.2250	0.2333
		LAD-SCAD	4.916	0.009	0.922	0.2350	0.2344	0.2475
		LSE-SCAD	3.503	0.178	0.594	0.9616	0.9826	0.9688
		EE-SCAD	3.524	0.190	0.589	0.9624	0.9856	0.9723
	200	ESL-SCAD	4.946	0.003	0.962	0.1128	0.1117	0.1253
		LAD-SCAD	4.930	0.005	0.950	0.1290	0.1274	0.1321
		LSE-SCAD	3.765	0.160	0.690	0.7398	0.7015	0.7547
		EE-SCAD	3.775	0.175	0.675	0.7412	0.7035	0.7569
	400	ESL-SCAD	4.998	0	0.998	0.0466	0.0432	0.0585
		LAD-SCAD	4.990	0	0.995	0.0598	0.0584	0.0619
		LSE-SCAD	4.215	0.105	0.750	0.4206	0.3573	0.5134
		EE-SCAD	4.190	0.110	0.735	0.4224	0.3597	0.5148
$MN(1,25)$	50	ESL-SCAD	4.895	0	0.930	0.2268	0.2150	0.2269
		LAD-SCAD	4.880	0	0.922	0.2440	0.2312	0.2453
		LSE-SCAD	3.425	0.175	0.546	0.9367	0.9536	0.9516
		EE-SCAD	3.440	0.190	0.536	0.9435	0.9557	0.9535
	200	ESL-SCAD	4.940	0	0.955	0.1129	0.1063	0.1147
		LAD-SCAD	4.935	0	0.950	0.1335	0.1241	0.1305
		LSE-SCAD	3.805	0.155	0.685	0.7151	0.6803	0.7233
		EE-SCAD	3.815	0.165	0.660	0.7193	0.6819	0.7245
	400	ESL-SCAD	4.997	0	1.000	0.0414	0.0563	0.0467
		LAD-SCAD	4.995	0	1.000	0.0587	0.0601	0.0595
		LSE-SCAD	4.355	0.090	0.840	0.3823	0.3634	0.5467
		EE-SCAD	4.275	0.105	0.755	0.3851	0.3676	0.5493

Table 4. CPU times for different n in Example 1.

n	$N(0,1)$	$t(3)$	$MN(1,100)$
50	0.5609	0.6491	0.7832
200	0.7326	0.8169	0.8664
400	0.9528	0.9868	1.0610

Table 5. CPU times for different n in Example 2.

n	$N(0,1)$	$t(3)$	$MN(1,100)$
50	0.8702	0.9564	1.1062
200	1.1470	1.2235	1.3598
400	1.3682	1.4373	1.6476

4. Discussion

In this paper, we use penalized regression with exponential squared loss to propose a robust variable selection procedure for a single-index model along with missing data. The B-spline is a method that can estimate the relationship with the response. IPW is a

frequently used method dealing with the bias resulting from missing covariates, and the non-convex penalty method is used to estimate and select the variable at the same time. We examine the properties of sampling and robustness of our estimator. From theoretical and simulation study in this paper, the merits of our method are obvious. We also illustrate that the outcomes are good when using our method for actual data. In particular, we reveal that this estimator has the highest sample breakdown point, and the influence function for outliers are limited either in the response domain or in the covariate domain. In this paper, simulation studies and applications indicate the advantage of our method. When outliers are presented (regardless of the mechanism), EE-SCAD and LSE-SCAD are inferior in terms of non-caused selection rate.

Moreover, we can make further studies based on our proposed method. First, it is worth considering the goodness-of-fit test; in this paper we only study the sparse estimation and variable selection, however. Second, censoring can be examined based on this model. An investigation of the difficulties above is a portion of further study but is out of this paper's scope. In the proposed theory, internal knots are considered as fixed values. Finally, how to optimally select internal knots when data are missing is an interesting problem worthy of future research.

Author Contributions: Formal analysis, H.S.; Methodology, Y.S.; Software, Y.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by NNSF project (61503412) of China, NSF project (ZR2019MA016) of Shandong Province of China.

Conflicts of Interest: The authors declare that they have no competing interest.

References

1. Yates, F. The analysis of replicated experiments when the field results are incomplete. *Emp. J. Exp. Agric.* **1933**, *1*, 129–142.
2. Healy, M.; Westmacott, M. Missing values in experiments analysed on automatic computers. *J. R. Stat. Soc. Ser. B Methodol.* **1956**, *5*, 203–206. [CrossRef]
3. Horvitz, D.G.; Thompson, D.J. A generalization of sampling without replacement from a finite universe. *J. Am. Stat. Assoc.* **1952**, *47*, 663–685. [CrossRef]
4. Robins, J.M.; Rotnitzky, A.; Zhao, L.P. Estimation of regression coefficients when some regressors are not always observed. *J. Am. Stat. Assoc.* **1994**, *89*, 846–866. [CrossRef]
5. Wang, C.; Wang, S.; Zhao, L.P.; Ou, S.T. Weighted semiparametric estimation in regression analysis with missing covariate data. *J. Am. Stat. Assoc.* **1997**, *92*, 512–525. [CrossRef]
6. Little, R.J.; Rubin, D.B. *Statistical Analysis with Missing Data*; John Wiley & Sons: Hoboken, NJ, USA, 2019; Volume 793.
7. Liang, H.; Wang, S.; Robins, J.M.; Carroll, R.J. Estimation in partially linear models with missing covariates. *J. Am. Stat. Assoc.* **2004**, *99*, 357–367. [CrossRef]
8. Tsiatis, A.A. *Semiparametric Theory and Missing Data*; Springer: Berlin/Heidelberg, Germany, 2006.
9. Friedman, J.; Hastie, T.; Tibshirani, R. Additive logistic regression: A statistical view of boosting (with discussion and a rejoinder by the authors). *Ann. Stat.* **2000**, *28*, 337–407. [CrossRef]
10. Wang, X.; Jiang, Y.; Huang, M.; Zhang, H. Robust variable selection with exponential squared loss. *J. Am. Stat. Assoc.* **2013**, *108*, 632–643. [CrossRef]
11. Feng, S.; Xue, L. Variable selection for single-index varying-coefficient model. *Front. Math China* **2013**, *8*, 541–565. [CrossRef]
12. Xue, L.; Pang, Z. Statistical inference for a single-index varying-coefficient model. *Stat. Comput.* **2013**, *23*, 589–599. [CrossRef]
13. Hardle, W.; Hall, P.; Ichimura, H. Optimal smoothing in single-index models. *Ann. Stat.* **1993**, *21*, 157–178. [CrossRef]
14. Wu, T.Z.; Lin, H.; Yu, Y. Single-index coefficient models for nonlinear time series. *J. Nonparametr. Stat.* **2011**, *23*, 37–58. [CrossRef]
15. Hastie, T.; Tibshirani, R. Varying-coefficient models. *J. R. Stat. Soc. Ser. B Methodol.* **1993**, *55*, 757–779. [CrossRef]
16. Fan, J.; Zhang, W. Statistical estimation in varying coefficient models. *Ann. Stat.* **1999**, *27*, 1491–1518. [CrossRef]
17. Xia, Y.; Li, W.K. On single-index coefficient regression models. *J. Am. Stat. Assoc.* **1999**, *94*, 1275–1285. [CrossRef]
18. Xue, L.; Wang, Q. Empirical likelihood for single-index varying-coefficient models. *Bernoulli* **2012**, *18*, 836–856. [CrossRef]
19. Tibshirani, R. Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Ser. B Methodol.* **1996**, *58*, 267–288. [CrossRef]
20. Fan, J.; Li, R. Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Am. Stat. Assoc.* **2001**, *96*, 1348–1360. [CrossRef]
21. Zou, H. The adaptive lasso and its oracle properties. *J. Am. Stat. Assoc.* **2006**, *101*, 1418–1429. [CrossRef]
22. Peng, H.; Huang, T. Penalized least squares for single index models. *J. Stat. Plan. Inference* **2011**, *141*, 1362–1379. [CrossRef]

23. Yang, H.; Yang, J. A robust and efficient estimation and variable selection method for partially linear single-index models. *J. Multivar. Anal.* **2014**, *129*, 227–242. [CrossRef]
24. Wang, D.; Kulasekera, K. Parametric component detection and variable selection in varying-coefficient partially linear models. *J. Multivar. Anal.* **2012**, *112*, 117–129. [CrossRef]
25. Yang, J.; Yang, H. Quantile regression and variable selection for single-index varying-coefficient models. *Commun. Stat.-Simul. C* **2017**, *46*, 4637–4653. [CrossRef]
26. Yu, Y.; Ruppert, D. Penalized spline estimation for partially linear single-index models. *J. Am. Stat. Assoc.* **2002**, *97*, 1042–1054. [CrossRef]
27. He, X.; Zhu, Z.Y.; Fung, W.K. Estimation in a semiparametric model for longitudinal data with unspecified dependence structure. *Biometrika* **2002**, *89*, 579–590. [CrossRef]
28. Zhao, P.; Xue, L. Variable selection for semiparametric varying coefficient partially linear models. *Stat. Probab. Lett.* **2009**, *79*, 2148–2157. [CrossRef]