Rank-Based Methods for Single-Index Varying Coefficient Models

by

Wei Sun

A dissertation submitted to the Graduate Faculty of
Auburn University
in partial fulfillment of the
requirements for the Degree of
Doctor of Philosophy

Auburn, Alabama
August 5, 2017

Keywords: Wilcoxon pseudo-norm, asymptotic normality, local linear estimation, robust estimation, variable selection

Approved by

Asheber Abebe, Chair, Professor of Mathematics and Statistics
Nedret Billor, Co-Chair ,Professor of Mathematics and Statistics
Peng Zeng, Associate Professor of Mathematics and Statistics
Guanqun Cao, Assistant Professor of Mathematics and Statistics
Huybrechts F.Bindele, Assistant Professor of Mathematics and Statistics

Abstract

The single-index varying coefficient model has received much attention due to its flexibility and interpretability in recent years. This dissertation is mainly concerned with the rank-based estimation and variable selection in single-index varying coefficient models.

In the first part of this dissertation, we consider a rank-based estimation of the index parameter and the coefficient functions for single-index varying coefficient model. The consistency and asymptotic normality of the proposed estimators are established. An extensive Monte-Carlo simulation study demonstrates the robustness and the efficiency of the proposed estimators compared to the least squares estimators. The rank-based approach was motivated by a problem from fisheries ecology where it is used to provide accurate estimates of interspecies dependence along an environmental gradient. We use a real data example to show that the classical approach is highly affected by outliers in response space but not the rank-based method we proposed in this dissertation.

The second part of this dissertation is based on variable selection method for single-index varying coefficient model. We develop a LASSO-type rank-based variable selection procedure to select and estimate coefficient functions. A Monte-Carlo simulation study shows that the proposed method is highly robust and efficient compared to least squares type approaches. Our method can be easily applied to single-index and varying coefficient models since they are special cases of single-index varying coefficient model.

Acknowledgments

This dissertation would not have been possible without the support of many people.

First of all, I wish to take this opportunity to express my sincere gratitude to my supervisor, Dr. Asheber Abebe who was abundantly helpful and offered invaluable assistance, support and guidance. I benefit so much from his deep insight into statistics, his professionalism and his constant encouragement. I feel extremely lucky to have him as my thesis advisor.

I am also deeply grateful to my Co-advisor Dr. Nedret Billor for her invaluable advice and helpful comments. Deepest gratitude are also due to my members of the supervisory committee, Drs. Peng Zeng, Guanqun Cao and Huybrechts F.Bindele without whose knowledge and assistance this study would not have been successful.

I wish to thank all professors who taught me in the Mathematics and Statistics department for their excellent lectures, particularly, Drs. Asheber Abebe, Nedret Billor, Peng Zeng, Ming Liao and Jessica McDonald.

I appreciate the friendship with my classmates and friends, including Dr. Yi Xu, Dr. Zhefeng Jiang, Dr. Zhifeng He, Xuyu Wang, Kefan Xiao, Yu Wang, Mingjie Feng, Ninkai Tang, Dr. Hao Wu, Chi Xu, Hannah Correia, Dr. Jessica Busler, and many others.

Finally, I would like to dedicate this thesis to my parents for their boundless love.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Background

The single-index varying coefficient model (SIVCM) is studied by many researchers due to its flexibility and interpretability. The model has been applied for addressing problems in areas such as finance, ecology, and public health among others. One important feature that makes the SIVCM attractive is the ability to overcome the "curse of dimensionality" often encountered in nonparametric modeling of multivariate data. Suppose y_i is the response variable, $X_i = (\mathrm{x}_{0i}, \ldots, \mathrm{x}_{pi})^T$ with $\mathrm{x}_{0i} = 1$, and $Z = (z_{1i}, \ldots, z_{qi})^T$ are predictor variables. The single-index varying coefficient model (SIVCM) has the following form

$$y_i = \{G(\boldsymbol{\theta}_0^T Z_i)\}^T X_i + \varepsilon_i, \quad i = 1, \ldots, n \tag{1.1}$$

where $\boldsymbol{\theta}_0$ is a $q-$vector of unknown regression parameters representing the single-index direction; $G(\cdot) = (g_0(\cdot), \ldots, g_p(\cdot))^T$ is a $p-$ vector of unknown coefficient functions; and ε_i are random errors with finite Fisher information. For model identifiability, it is assumed that $\|\boldsymbol{\theta}_0\| = 1$ and the first component of $\boldsymbol{\theta}_0$ is positive. Model (1.1) includes a class of important semiparametric models known as single-index models (SIM) by setting $p = 0$. When $q = 1$ and $\boldsymbol{\theta}_0 = 1$, model (1.1) is reduced to varying coefficient models (VCM) (Trevor Hastie (1993)), which has been widely used in application. The historical estimation approach for model (1.1) is based on least squares (LS) methods.

For a general regression model: $y_i = f(X_i, \boldsymbol{\theta}_0) + \varepsilon_i, \ i = 1, \ldots, n$, LS procedure minimizes the sum of square errors: $\sum_{i=1}^n (y_i - f(X_i, \boldsymbol{\theta}))^2$. Such estimators are computationally simple and possess general optimality properties. However, they are known to be sensitive to outliers, model contaminations, and/or heavy-tail error distributions. Some approaches has been take to mitigate the effect of these abnormalities. In 1960s, Huber (1964) proposed so-called M-estimators by minimizing $\rho\big(\dfrac{y_i - f(x_i, \boldsymbol{\theta})}{\widehat{\sigma}_i}\big)$, where $\rho(\cdot)$ is a symmetric function and $\widehat{\sigma}_i$ is an estimate of the standard deviation of the errors ε_i. One type of M-estimator that has been widely used is least absolute deviation (L_1) estimator, which minimizes $\sum_{i=1}^{n} |y_i - f(X_i, \boldsymbol{\theta})|$. When Huber and others were developing the theory of M estimators, rank-based (R) estimation methods were not considered to be as generalizable as M estimators. They were used for simple problems such as location comparisons for two-sample problems. Later Jaeckel (1972), Hettmansperger & McKean (1998) and others showed that R estimators, some times called Wilcoxon estimators can be obtained by minimizing $\sum_{i=1}^{n} a\Big(R(e_i(\boldsymbol{\theta}))\Big)e_i(\boldsymbol{\theta})$, where $R(e_i(\boldsymbol{\theta}))$ is the rank of $e_i(\boldsymbol{\theta}) = y_i - f(X_i, \boldsymbol{\theta})$ and $a(\cdot)$ is some score function. The R estimator can be applied in any general linear model, and it is well known that R estimator outperforms LS estimator when the data deviate from normality and/or contain outliers. However, the original M estimators and R estimators can be affected by outliers in X space in regression models. A generalized M-estimators (Krasker & Welsch (1982)) and a weight Wilcoxon procedure (Sievers (1983)) were later developed by introducing weights to take care of the leverage points.

1.2 Motivation

Our consideration of the model (1.1) and its robust estimation were primarily motivated by an ecological problem that involves high-dimensional environmental predictors as well as interacting groundfish species. This is a part of a large project where we consider a subset of fisheries data obtained from the NOAA Marine Ecology Stock Assessment (MESA) Program paired with environmental data from the NOAA National Data Buoy Center (NDBC) to understand interactions between groundfish predator species in the Gulf of Alaska. Much of the use of the SIVCM in ecology has focused on predator-prey models (Fan *et al.*, 2003; Xia *et al.*, 2007). Another reasonable application of the SICVM is to better quantify interspecific competition in a complex food web. The Gulf of Alaska food web is of particular interest to fisheries management, as several commercially important species of fish are found there. Accurate management of these fisheries is aided by the understanding of predator-prey interactions and competition of food resources (Gaichas *et al.*, 2010). Since 1979, the MESA Program has conducted annual longline surveys on seven groundfish species along the coast of Alaska (AFSC, 2015). For each species, a catch per unit effort (CPUE) is calculated based on a catch rate standardized for size of the geographic area. These locations ("stations") are each sampled once a year and are repeated each year from May to October. We focused on two groupings of stations: six stations located near Kodiak Island were aggregated to create a median CPUE for the 2° by 4° latitude-longitude block, and five stations near the Aleutians were treated in the same manner. The NDBC manages stationary and floating buoys deployed by various public organizations and downloads measurements from these buoys for public access (NOAA, 2015). The ones we focus on here are maintained by the NOAA and collect environmental measures including wind, wave, pressure, and temperature variables. These buoys sample every six or twelve hours, so in order to match the environmental data with MESA catch data a summer coefficient of variation was calculated for each variable: $c_v = \sigma/\mu$, where μ is the variable's mean taken from values in May to October and σ is its standard deviation. In this analysis we used data from two anchored NOMAD buoy located off the coast of Alaska, one near the Kodiak Island MESA stations and the other near the Aleutian Island stations. The MESA data was paired with environmental data from the NOMAD buoys, providing a yearly median CPUE for each of seven groundfish and yearly summer coefficients of variation for each environmental variable. Research on the stomach contents of Pacific halibut (Best & St-Pierre, 1986; Yang *et al.*, 2006; Gaichas *et al.*, 2010) reveals that it is a top predator in the Gulf of Alaska and may prey opportunistically on sablefish and Pacific cod while all three species share a common preferred prey - walleye pollock (Moukhametov *et al.*, 2008; Yang *et al.*, 2006).

As a simple preliminary analysis, we performed principal components analysis on seven environmental variables from NDBC and retained the first PC as an indicator of environmental condition. We then split the data into two at the median of the first PC indicating two environmental regimes. This analysis is rather naïve but sufficient for an initial descriptive discussion of the data. We will re-analyze the data more rigorously later. CPUE values were log-transformed and scaled as is common in practice (Phillips *et al.*, 2014). We identify two outliers in the CPUE (log-transformed) of Pacific halibut based on Figure 2.8 in Section 2.3.2.

For now, we simply plotted CPUE values of Pacific halibut versus CPUE values of Pacific cod and sablefish, respectively. We also superimposed these plots with loess fits to detect any nonlinearities. These are given in the top panels of Figure 1.1. It is evident that the relationships between Pacific cod and sablefish with Pacific halibut are nonlinear and dependent on the environmental regimes. This process was repeated after removing two outliers, corresponding to unusually high Pacific halibut catch values, identified using residual diagnostics on the basis of our proposed methodology. The plots are displayed in the bottom panels of Figure 1.1.

It appears that sablefish and Pacific cod tend to prefer different environmental regimes and their relationship with Pacific halibut depends on the environment. In both cases, the first PC captures only about 40% of the environmental variability. So, a varying coefficient model relating Pacific halibut to Pacific cod and sablefish using just the first PC would not be sufficient. Building a varying coefficient model using all the environmental variables is also not realistic as it requires a seven dimensional smoother. This suggests that the model is

$$y_i = g_0(\boldsymbol{\theta}_0^T Z) + g_1(\boldsymbol{\theta}_0^T Z)\mathrm{x}_{1i} + g_2(\boldsymbol{\theta}_0^T Z)\mathrm{x}_{2i} + \varepsilon_i \tag{1.2}$$

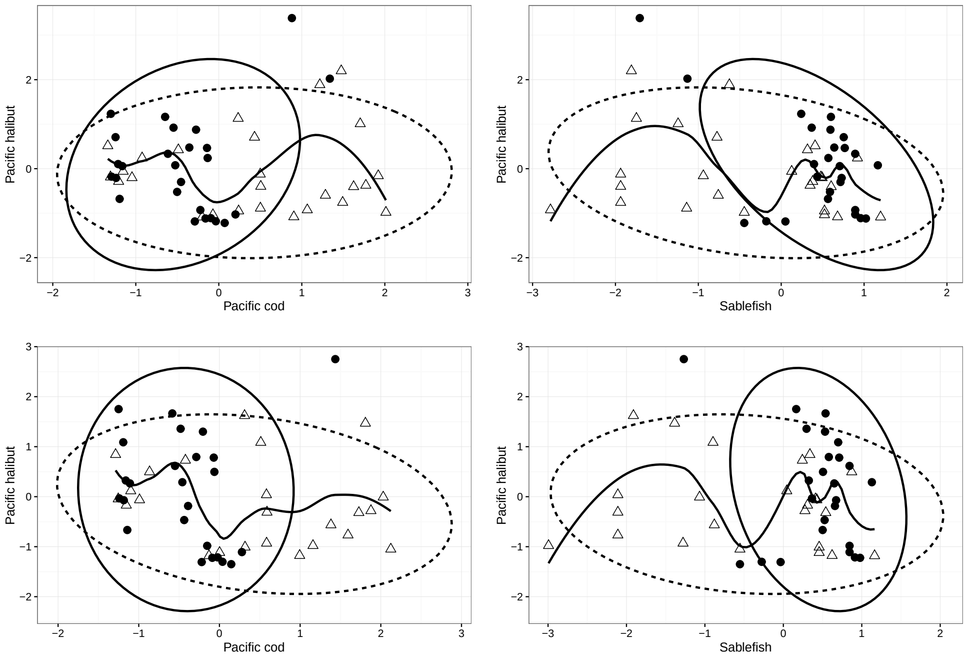

Figure 1.1: Plots of groudfish categorized by environmental regime. *Left*: Pacific halibut vs Pacific cod. *Right*: Pacific halibut vs sablefish. *Top*: Original Data. *Bottom*: Two outliers removed. Filled circle: first PC < median; Open triangle: first PC > median

where y_i is the CPUE (log-transformed) of Pacific halibut for $i = 1, 2, \ldots, 52$; x_{1i} and x_{2i} are the CPUE of Pacific cod and the CPUE of sablefish, respectively. The matrix $Z = (z_1, z_2, z_3, z_4, z_5, z_6, z_7)^T$ contains the summer coefficient of variation of seven buoy environmental variables supposed to have an impact on fish population numbers: wind direction (z_1), wind speed(z_2), significant wave height (z_3), dominant wave period (z_4), average wave period(z_5), sea level pressure (z_6), and sea surface temperature (z_7), respectively. Robust fitting is required to mitigate the effect of the possible outlying points on our fit.

1.3 Contribution

In Chapter 2, we propose a general R estimation procedure that is robust and more efficient alternative to the least squares method for fitting model (1.1) when dealing with contaminated and heavy-tailed model error distributions, or when data contain outliers in the response space. The development of R estimation procedure for model (1.1) is motivated by an ecological problem described in Section 1.2. We show that for these data in Section 1.2 it is indeed the case that the classical approach is highly affected by the outliers but not the robust method proposed in this dissertation. For computational purposes, we propose a backfitting type algorithm by iterating between the coefficient functions through local linear procedure and estimating $\boldsymbol{\theta}_0$ by one-step R estimation. We demonstrated that the resulting estimators are robust and asymptotically efficient compared to LS estimation when the data contain outliers. The consistency and asymptotic normality of the proposed estimators are established.

In Chapter 3, we propose a robust two-stage procedure to select coefficient functions using group LASSO and estimate index parameters using general local rank estimation. We showed that our procedure are highly efficient in both function selection and index estimation compare to LS when the error distribution are not normal and performs as well as LS under normal error distribution. The R estimation and variable selection method for SIVCM we developed in this dissertation can be easily extended on SIM and VCM since they are special cases of SIVCM. We also provided a Monte Carlo simulation study to show our method outperforms LS for VCM when the error distribution are not normal.

Chapter 2

General Local Rank Estimation for Single-index Varying Coefficient Models

2.1 Introduction

Suppose y_i is the response variable, $X = (\mathrm{x}_{0i}, \ldots, \mathrm{x}_{pi})^T$ with $\mathrm{x}_{0i} = 1$, and $Z = (z_{1i}, \ldots, z_{qi})^T$ are predictor variables. The single-index varying coefficient model (SIVCM) is defined as

$$y_i = \{G(\boldsymbol{\theta}_0^T Z_i)\}^T X_i + \varepsilon_i \quad i = 1, \ldots, n \tag{2.1}$$

where $\boldsymbol{\theta}_0$ is a $q-$vector of unknown regression parameters representing the single-index direction; $G(\cdot) = (g_0(\cdot), \ldots, g_p(\cdot))^T$ is a $p-$ vector of unknown coefficient functions; and ε_i are random errors with finite Fisher information. For model identifiability, it is assumed that the $\|\boldsymbol{\theta}_0\| = 1$ and first component of $\boldsymbol{\theta}_0$ is positive. Model (2.1) was first considered by Xia & Li (1999) who proposed estimating $\boldsymbol{\theta}_0$ via an L_2-cross validation approach following ideas of Härdle *et al.* (1993). They also established the $\sqrt{n}$-consistency and asymptotic normality of their proposed estimator under some mild conditions. Setting $Z = (\mathrm{x}_{1i}, \ldots, \mathrm{x}_{pi})^T$ in model (2.1), Fan *et al.* (2003) proposed a computationally efficient estimation approach based on a profile least squares (LS) local linear regression, from which they also discussed how to select locally significant variables based on the t-statistic and the Akaike information criterion. Motivated by the "remove-one-component" approach proposed in Yu & Ruppert (2002), Xue & Pang (2013) provided the estimation of $\boldsymbol{\theta}_0$ and the coefficient functions. In an effort to construct a robust confidence region for $\boldsymbol{\theta}_0$, Xue & Wang (2012) studied model (2.1) using an empirical likelihood approach.

However, all the estimation methods above are LS-type methods, which are known to be sensitive to outliers, model contamination, and/or heavy-tail error distributions. To mitigate the effect of these abnormalities, it is imperative to develop robust and efficient estimation procedures. Yao *et al.* (2012) proposed a local modal estimation procedure for nonparametric regression models using an EM algorithm. Their estimator was shown to be more efficient compared to ordinary local polynomial estimators when dealing with outliers in the response space and for heavy tailed model error distributions. They also showed that their estimator is as asymptotically efficient as the local polynomial regression estimator under the condition that there are no outliers or the errors are from the normal distribution. The concept of using local modal estimation to get robust estimators has been extended to semiparametric partial linear varying coefficient models, single-index models and SIVCMs by Zhang *et al.* (2013), Liu *et al.* (2013) and Yang *et al.* (2014), respectively. Feng *et al.* (2012) used the Wilcoxon rank-based method of Hettmansperger & McKean (2011) to produce a robust estimator $\boldsymbol{\theta}_0$ for the single-index model. Although their simulation study includes categorical predictors, this is not justified theoretically, as their approach relies on taking derivatives with respect to covariates. For varying coefficient models, Wang *et al.* (2009) proposed a local rank estimation method which is based on the objective function of Jaeckel (1972). Their approach was shown to have several advantages compared to the LS-type approaches.

In this chapter, we propose a general rank-based (R) estimation procedure for model (2.1). Our approach will include that of Wang *et al.* (2009) as a particular case. As in Wang *et al.* (2009), the motivation behind considering the R estimation relies on the fact that as for the LS approach, it has a simple geometric interpretability and it results in robust and more efficient estimators compared to those obtained via many of the method of moments type estimation approaches that include the LS and the least absolute deviation (LAD) approaches as particular cases (Hettmansperger & McKean, 2011). For computing the regression estimators, we propose a backfitting type algorithm by iterating between the coefficient functions through local linear procedure and estimating $\boldsymbol{\theta}_0$ by a one-step R estimation. The local linear estimation of $g_j(\cdot)$, $j = 0, \ldots, p$, involves bandwidth selection. This is done via a leave-one-out R cross-validation. We demonstrate that the resulting estimators are robust and asymptotically efficient compared to LS estimators when the data contain outliers.

The remainder of Chapter 2 is organized as follows: Section 2.2 presents the estimation procedures for the parameter $\boldsymbol{\theta}_0$ and the function $G(\cdot) = (g_0(\cdot), \ldots, g_p(\cdot))^T$. The computational algorithm for obtaining the rank-based estimators of $\boldsymbol{\theta}_0$ and $G(\cdot)$ is also provided in Section 2.2. In Section 2.3, an extensive Monte Carlo simulation study and an illustrative real data example are presented to demonstrate the advantage of the proposed rank-based estimation method. Section 2.4 discusses the asymptotic properties of the proposed estimators. A brief conclusion is provided in Section 2.5. Proofs of some theoretical results are given in Section 2.6.

2.2 General Local Rank Estimation

Suppose that $\{(X_i, Z_i, y_i), i = 1, \ldots, n\}$ is a random sample from model (2.1). Define the residuals as $\eta_i(\boldsymbol{\theta}) = y_i - \sum_{k=0}^{p} g_k(\boldsymbol{\theta}^T Z_i)\mathrm{x}_{ji} = y_i - \{G(\boldsymbol{\theta}^T Z_i)\}^T X_i$, and consider the following general rank objective function introduced by Jaeckel (1972)

$$D_n(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n} \varphi\Big(\frac{R(\eta_i(\boldsymbol{\theta}))}{n+1}\Big)\eta_i(\boldsymbol{\theta}), \tag{2.2}$$

where $R(\eta_i(\boldsymbol{\theta}))$ is the rank of $\eta_i(\boldsymbol{\theta})$ among $\eta_1(\boldsymbol{\theta}), \ldots, \eta_n(\boldsymbol{\theta})$, and φ is a general bounded nondecreasing score function defined on $(0, 1)$. Since it was proposed, the objective function given by equation (2.2) has captured a lot of attention, as its minimization results in a robust and efficient estimator of $\boldsymbol{\theta}_0$. Although, our interest is placed in the estimation of $\boldsymbol{\theta}_0$, it is worth pointing out that in the residuals defined above, both $\boldsymbol{\theta}$ and $g_k(\cdot)$, $k = 0, \ldots, p$, are unknown. Thus $D_n(\boldsymbol{\theta})$ is a function of two unknown parameters: the index parameter $\boldsymbol{\theta}$, and the functional parameters $g_k(\cdot)$. To this end, we first consider estimating $g_k(\cdot)$ based on a local linear estimator (Fan & Gijbels, 1996). Under the smoothness assumption on $g_k(\cdot)$ and applying the mean value theorem to $g_k(t)$, for $t \in \mathcal{A}$, one can approximate $g_k(\boldsymbol{\theta}^T Z)$ as $g_k(\boldsymbol{\theta}^T Z_i) \approx g_k(\boldsymbol{\theta}^T z) + g_k'(\boldsymbol{\theta}^T z)\boldsymbol{\theta}^T(Z_i - z)$ for any z satisfying $\|Z_i - z\| \to 0$. Let $G(\boldsymbol{\theta}^T z_0) = (g_0(\boldsymbol{\theta}^T z), \ldots, g_p(\boldsymbol{\theta}^T z))^T$ and $G'(\boldsymbol{\theta}^T z) = (g_0'(\boldsymbol{\theta}^T z), \ldots, g_p'(\boldsymbol{\theta}^T z_0))^T$. Denote $a = G(\boldsymbol{\theta}^T z)$ and $b = G'(\boldsymbol{\theta}^T z)$. Define $\eta_i(\boldsymbol{\theta}, a, b)$ as $\eta_i(\boldsymbol{\theta}, a, b) = y_i - X_i^T a - X_i^T b Z_{i0}^T \boldsymbol{\theta}$, where $Z_{i0} = Z_i - z$; then minimizing $D_n(\boldsymbol{\theta})$ in equation (2.2) will be equivalent to minimizing $L_n(\boldsymbol{\theta}, a, b)$ defined by

$$L_n(\boldsymbol{\theta}, a_j, b_j) = \frac{1}{n(n-1)} \sum_{j=1}^{n} \sum_{i=1}^{n} \varphi\left(\frac{R(\eta_{ij}(\boldsymbol{\theta}, a_j, b_j))}{n^2+1}\right) \eta_{ij}(\boldsymbol{\theta}, a_j, b_j) w_{ij}, \tag{2.3}$$

where the weight function w_{ij} is defined by $w_{ij} = K_h(\boldsymbol{\theta}^T Z_{ij}) / \sum_{j=1}^{n} K_h(\boldsymbol{\theta}^T Z_{ij})$, with $K_h(\cdot) = K(\cdot/h)$, K a kernel function defined on the real line and h the corresponding bandwidth.

Remark 1. *Note that for each fixed j, $a_j = (a_{j0}, \ldots, a_{jp})^T$ and $b_j = (b_{j0}, \ldots, b_{jp})^T$. Thus, when minimizing $L_n(\boldsymbol{\theta}, a, b)$, one might face an over-parametrization problem. To overcome this issue, an alternating estimation procedure can be used. That is, starting with a $\sqrt{n}$-consistent estimator of $\boldsymbol{\theta}_0$, say $\widetilde{\boldsymbol{\theta}}$, we can obtain $(\widehat{a}_j, \widehat{b}_j)$ as $(\widehat{a}_j, \widehat{b}_j) = \operatorname{Argmin} \ell_n(a_j, b_j)$, where*

$$\ell_n(a_j, b_j) = \frac{1}{n} \sum_{i=1}^{n} \varphi\left(\frac{R(\nu_i(a_j, b_j))}{n+1}\right) \nu_i(a_j, b_j),$$

$\nu_i(a_j, b_j) = y_i - (a_j - b_j \widetilde{\boldsymbol{\theta}}^T Z_{ij}) x_{ji}$. Once (a_j, b_j) have been estimated, we then move to finding $\widehat{\boldsymbol{\theta}}_n$ as $\widehat{\boldsymbol{\theta}}_n = \operatorname*{Argmin}_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} L_n(\boldsymbol{\theta}, \widehat{a}, \widehat{b})$.

We provide a computational algorithm that will achieve this estimation procedure below:

2.2.1 Computational Algorithm

The outline of the algorithm pertaining to the estimation of $\boldsymbol{\theta}_0$, $g_j(\cdot)$ and $g_j'(\cdot)$ is as follows:

Step 0: (Initialization): Specify an initial value of $\boldsymbol{\theta}_0$ with first component 1 or $\|\boldsymbol{\theta}_0\| = 1$, and denote the initial estimate as $\widetilde{\boldsymbol{\theta}}$.

Step 1: Given $\widetilde{\boldsymbol{\theta}}$ and for each fixed j, estimate $g_j(\cdot)$ and $g_j'(\cdot)$ by

$$\underset{a_j,b_j}{\text{Argmin}} \, \frac{1}{n} \sum_{j=1}^{n} \varphi\left(\frac{R(\nu_i(a_j,b_j))}{n+1}\right) \nu_i(a_j,b_j) w_{ij}$$

where $\nu_i(a_j,b_j) = y_i - X_i^T a_j - {X_i}^T b_j Z_{ij}^T \widetilde{\boldsymbol{\theta}}$.

Step 2: Let $\widehat{a}_j$ and $\widehat{b}_j$ be the estimates of a_j and b_j. Once $\widehat{a}_j$ and $\widehat{b}_j$ are obtained in **Step 1**, estimate $\boldsymbol{\theta}$ by

$$\underset{\boldsymbol{\theta}}{\text{Argmin}} \, \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j=1}^{n} \varphi\left(\frac{R(e_{ij}(\boldsymbol{\theta}))}{n^2+1}\right) e_{ij}(\theta) w_{ij}$$

where $e_{ij}(\boldsymbol{\theta}) = (y_i - \widehat{a}_j^T X_i) - X_i^T \widehat{b}_j Z_{ij}^T \boldsymbol{\theta}$

Step 3: Repeat **Step 1** and **Step 2** until convergence.

Step 4: Once the final estimate of $\boldsymbol{\theta}_0$, say $\widehat{\boldsymbol{\theta}}_n$, is obtained from **Step 3**, use it to get the final estimate of $g_j(\cdot)$, $j = 0, \ldots, p$.

Remark 2. *We used a sliced inverse regression (SIR) of Li (1991) to obtain the initial estimate of $\boldsymbol{\theta}_0$ as in Xia (2006).*

2.3 Simulation and Real Data Analysis

2.3.1 Simulation

To demonstrate the performance of the rank-based estimation approach, an extensive simulation study for the estimation of the index $\boldsymbol{\theta}_0$ and the coefficients $g_j(\cdot)$ was conducted. We considered the following SIVCM defined as

$$y_i = 3\exp\{-(\boldsymbol{\theta}_0^T Z_i)^2\} + 0.8\{\boldsymbol{\theta}_0^T Z_i\}\mathrm{x}_{1i} + 1.5\sin(\pi\boldsymbol{\theta}_0^T Z_i)\mathrm{x}_{3i} + \varepsilon_i,$$

where, for $i = 1,\dots,n$, $Z_i = (z_{1i}, z_{2i}, z_{3i}, z_{4i})^T$ are independent random vectors uniformly distributed on $[-1,1]^{\otimes 4}$, $X_i = (\mathrm{x}_{1i}, \mathrm{x}_{2i}, \mathrm{x}_{3i}, \mathrm{x}_{4i})^T$ with x_{li}, $l = 1,\dots,4$, being independent standard normal random variables, and $\boldsymbol{\theta}_0 = (\theta_{01}, \theta_{02}, \theta_{03}, \theta_{04})^T = (1/3, 2/3, 0, 2/3)^T$. The score function φ that appears in the objective function (2.2) is taken to be the Wilcoxon score function $\varphi(u) = \sqrt{12}(u - 1/2)$. Also, the kernel function K is chosen to be the Gaussian kernel $K(u) = (1/\sqrt{2\pi})\exp(-u^2/2)$. From 200 replications, the bias, the standard deviation (SD), and mean absolute deviations of the coefficient functions and their overall mean absolute deviation (MAD) are calculated under different sample sizes (n = 50, 100 and 200). Six different model error (ε) distributions are considered: the standard normal distribution ($N(0,1)$); the t-distribution with 3 degrees of freedom (t_3); the contaminated normal distribution ($\mathcal{CN}$) with contamination rate 0.05, given as $\mathcal{CN}(0.95) = 0.95N(0,1) + 0.05N(0,100)$; the Laplace distribution; the log-normal distribution; and the Cauchy distribution. These choices are motivated to show the robustness of the proposed method compared to the least squares (LS) approach in the presence of gross outliers and/or under heavy tailed model error distributions. The performance of the rank-based estimator of $\boldsymbol{\theta}_0$ is assessed based on its bias, SD, and mean square error (MSE) and compared to those obtained via the LS approach. When it comes to assessing the performance of the estimator of $g_j(\cdot)$, $j = 0,\dots,3$, we consider mean absolute deviation of each coefficient function (MAD_j), defined as

$$\mathrm{MAD}_j = n_{grid}^{-1} \sum_{k=1}^{n_{grid}} |\widehat{g}_j(u_k) - g_j(u_k)|, \quad j = 0, \ldots, 3 \tag{2.4}$$

where u_k, $k = 1, \ldots, n_{grid}$ are the grid points and the functions $\widehat{g}_j(\cdot)$ are the estimates. The overall performance is assessed via the mean absolute deviation of all estimated coefficient functions defined by

$$\mathrm{MAD} = \frac{1}{n_{grid} \times p} \sum_{j=0}^{p-1} \sum_{k=1}^{n_{grid}} |\widehat{g}_j(u_k) - g_j(u_k)| \,.$$

This same criteria was used to assess the performance of their proposed estimators in Fan *et al.* (2003). It is worth pointing out that in the process of estimating $g_j(\cdot)$ which involves kernel smoothing, the bandwidth selection is very crucial. The optimal bandwidth, say $\widehat{h}_{opt}$, can be obtained as

$$\widehat{h}_{opt} = \underset{h}{\mathrm{Argmin}} \, \frac{1}{n} \sum_{i=1}^{n} \varphi\left(\frac{R(\widehat{\eta}_{-i}(h))}{n+1} \right) \widehat{\eta}_{-i}(h),$$

where $\widehat{\eta}_{-i}(h)$ is the leave-one-out version of $\widehat{\eta}_i(h) = y_i - \left\{ \widehat{G}\left(\widehat{\boldsymbol{\theta}}^T Z_i, h \right) \right\}^T X_i$. Following similar arguments and ideas in Delecroix *et al.* (2006), it can be demonstrated that $\widehat{h}_{opt}$ is proportional to $n^{1/(2r+1)}$, where r is the order of smoothness of $g_j(\cdot)$. The results of the entire simulation study are displayed in Tables 2.1 − 2.2. We report estimated functions for different error distributions in Figures 2.1 − 2.6, respectively.

Considering the estimated index (Table 2.1), we observe that the LS estimator slightly outperforms the rank-based approach for normal distribution, as expected, by providing slightly smaller biases, standard deviations, and mean squared errors. The rank-based approach, however, was superior for all other considered distributions, especially for larger sample sizes. The same observation is made for the estimated coefficient functions as can be seen in Table 2.2 and Figures 2.1 and 2.2.

Table 2.1: Bias ($\times 10^2$), standard deviation ($\times 10^2$) and MSE ($\times 10^2$) of the true index $\boldsymbol{\theta}_0$.

			θ_1			θ_2			θ_3			θ_4		
ε	n	method	Bias	SD	MSE	Bias	SD	MSE	Bias	SD	MSE	Bias	SD	MSE
$N(0,1)$	100	LS	0.033	7.422	0.548	-1.592	10.617	1.147	0.862	8.802	0.778	-1.341	11.850	1.415
		R	-0.105	8.444	0.710	-1.796	10.538	1.137	0.765	10.010	1.003	-1.421	12.235	1.510
	200	LS	0.216	3.421	0.117	-0.180	2.862	0.082	-0.351	4.276	0.183	-0.271	2.729	0.075
		R	0.033	3.873	0.149	-0.220	3.360	0.113	-0.181	5.114	0.261	-0.265	3.203	0.103
	400	LS	0.448	2.505	0.064	-0.010	1.717	0.029	0.058	2.575	0.066	-0.359	1.811	0.034
		R	0.634	2.799	0.082	-0.016	2.086	0.043	0.076	2.998	0.089	-0.497	2.112	0.047
t_3	100	LS	1.269	15.021	2.261	-10.753	31.590	11.086	0.857	21.273	4.510	-14.210	35.532	14.581
		R	1.402	13.647	1.873	-8.160	27.337	8.101	2.922	18.111	3.349	-8.083	26.038	7.399
	200	LS	-0.486	8.039	0.645	-3.109	16.453	2.790	-1.483	10.068	1.031	-1.825	15.540	2.436
		R	0.021	5.768	0.331	-0.681	4.415	0.199	-0.476	5.810	0.338	-0.122	4.352	0.189
	400	LS	-0.322	5.212	0.271	-1.197	10.632	1.139	0.816	7.308	0.538	-0.964	10.694	1.147
		R	-0.102	3.417	0.116	-0.217	2.470	0.061	0.251	3.695	0.136	-0.011	2.429	0.059
$\mathcal{CN}(0.95)$	100	LS	3.888	22.981	5.406	-49.457	53.171	52.591	-0.994	41.378	17.046	-48.892	54.537	53.499
		R	1.654	16.959	2.889	-23.208	43.977	24.629	-1.132	26.126	6.804	-19.320	41.992	21.278
	200	LS	2.505	19.474	3.836	-29.435	48.036	31.624	-0.360	31.297	9.747	-32.017	49.933	35.059
		R	-1.302	8.252	0.695	-0.364	5.753	0.331	0.823	10.383	1.079	-0.792	5.485	0.306
	400	LS	0.291	13.576	1.835	-8.712	26.796	7.903	-0.104	16.500	2.709	-7.478	29.032	8.945
		R	0.296	4.329	0.187	-0.538	3.014	0.093	0.127	4.851	0.234	-0.064	3.026	0.091
Cauchy	100	LS	8.128	23.573	6.190	-69.728	51.203	74.706	1.654	51.541	26.459	-66.490	49.764	68.849
		R	3.250	22.506	5.146	-34.000	49.099	35.547	0.052	37.213	13.779	-39.464	50.862	41.314
	200	LS	11.492	25.507	7.794	-58.873	48.514	58.079	-1.732	48.292	23.235	-62.738	51.137	65.379
		R	0.330	17.258	2.965	-24.006	45.406	26.277	-3.260	26.788	7.246	-22.236	44.837	24.948
	400	LS	13.794	24.526	7.888	-61.827	49.726	62.829	-2.944	45.942	21.087	-62.387	50.820	64.619
		R	-2.599	11.057	1.284	-3.987	23.311	5.566	-0.402	11.998	1.434	-4.493	21.396	4.757
Log Normal	100	LS	1.999	18.607	3.485	-21.110	40.865	21.072	0.994	29.253	8.524	-25.218	45.939	27.358
		R	0.253	12.414	1.534	-5.996	22.720	5.496	0.963	16.956	2.870	-9.406	31.035	10.468
	200	LS	0.352	12.561	1.571	-6.075	24.945	6.561	-0.883	15.314	2.341	-7.797	26.863	7.788
		R	-0.274	4.787	0.229	-0.358	4.050	0.164	0.184	6.973	0.484	-0.276	3.878	0.150
	400	LS	0.726	6.886	0.477	-0.070	6.944	0.480	0.644	8.471	0.718	-2.438	10.634	1.185
		R	0.633	3.034	0.096	-0.016	2.149	0.046	-0.146	3.320	0.110	-0.529	2.242	0.053
Laplace	100	LS	-1.362	13.424	1.812	-7.348	27.316	7.964	-0.859	17.829	3.170	-10.711	33.087	12.040
		R	-1.348	12.438	1.557	-6.383	27.102	7.716	0.518	15.133	2.281	-8.462	29.094	9.138
	200	LS	-0.325	6.446	0.414	-0.111	4.383	0.191	0.031	6.144	0.376	-0.623	4.602	0.215
		R	-0.119	5.603	0.312	0.002	4.136	0.170	0.483	5.932	0.352	-0.730	4.591	0.215
	400	LS	0.862	3.522	0.131	-0.340	2.799	0.079	-0.150	3.564	0.127	-0.405	2.845	0.082
		R	0.687	3.122	0.102	-0.190	2.466	0.061	0.157	3.240	0.105	-0.401	2.493	0.063

Table 2.2: Mean absolute deviations ($\times 10^2$) of the coefficient functions and the overall mean absolute deviation (MAD).

ε	n	method	g_0	g_1	g_2	g_3	MAD
$N(0,1)$	100	LS	23.376	19.798	19.564	36.736	23.712
		R	22.744	20.400	20.482	41.151	24.974
	200	LS	16.093	12.568	12.464	25.618	15.835
		R	14.737	12.921	12.949	29.708	16.723
	400	LS	11.938	8.986	8.903	19.245	11.568
		R	11.473	9.575	9.421	23.205	12.576
t_3	100	LS	39.663	32.834	28.929	57.756	37.943
		R	31.189	25.954	25.187	55.323	32.672
	200	LS	24.096	20.812	20.877	35.793	24.475
		R	18.774	15.771	16.189	36.467	20.785
	400	LS	17.017	14.030	13.955	26.311	17.188
		R	12.407	10.819	10.515	27.371	14.400
$\mathcal{CN}(0.95)$	100	LS	67.822	64.561	55.824	100.601	69.053
		R	41.688	31.022	26.219	71.488	39.474
	200	LS	49.272	44.877	34.871	74.831	47.876
		R	32.105	15.617	14.201	49.065	25.271
	400	LS	31.858	25.395	23.887	47.697	30.457
		R	21.144	10.249	10.189	37.886	17.886
Cauchy	100	LS	760.087	438.748	794.890	822.496	653.728
		R	62.240	52.355	51.598	98.545	62.295
	200	LS	1109.948	713.621	696.488	648.632	785.181
		R	52.914	32.729	24.622	79.545	42.653
	400	LS	909.510	496.278	1049.644	545.006	722.578
		R	80.485	15.043	14.616	68.010	38.433
Log Normal	100	LS	62.522	38.510	34.749	69.376	48.094
		R	36.505	23.190	22.106	55.330	31.665
	200	LS	55.383	24.777	22.323	45.042	34.161
		R	25.106	12.642	12.483	37.999	20.220
	400	LS	56.239	16.569	17.854	31.355	27.610
		R	16.227	9.305	8.952	29.150	14.409
Laplace	100	LS	31.056	27.790	25.158	51.300	32.022
		R	28.034	26.182	23.087	52.367	30.410
	200	LS	20.525	16.559	17.177	29.745	20.250
		R	17.063	14.648	14.723	32.712	18.799
	400	LS	14.699	12.607	11.854	23.118	14.910
		R	11.590	10.125	9.710	25.345	13.343

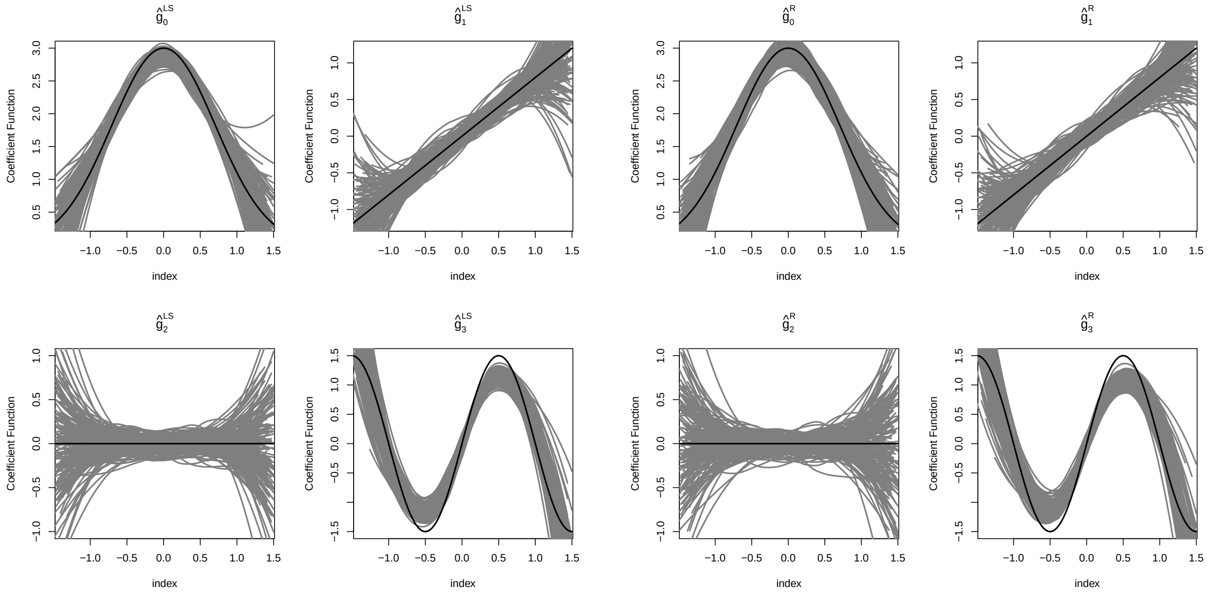

Figure 2.1: Estimated coefficient functions under the standard normal error distribution. *Left panel*: LS. *Right panel*: Rank

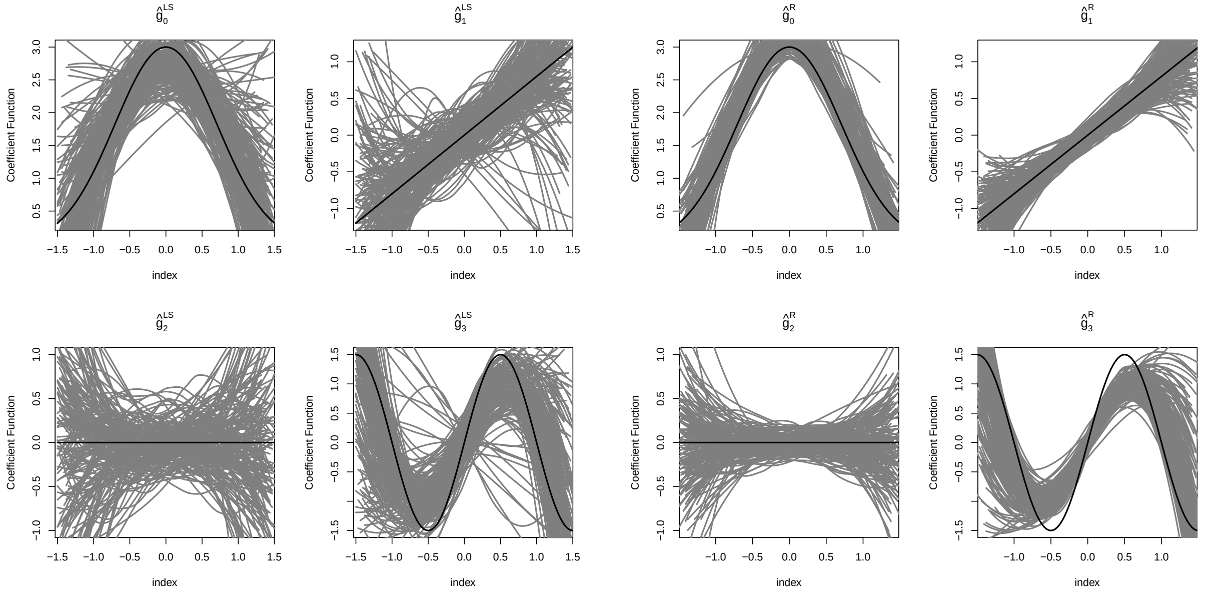

Figure 2.2: Estimated coefficient functions under the contaminated normal error distribution with contaminated rate 5%. *Left panel*: LS. *Right panel*: Rank

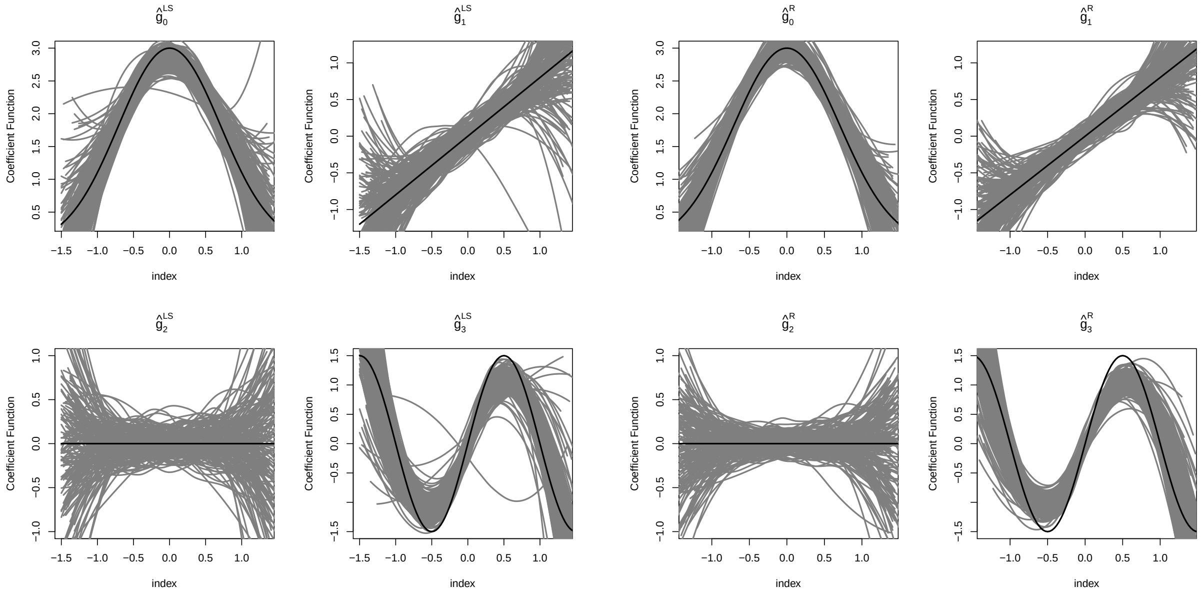

Figure 2.3: Estimated coefficient functions under the t_3 error distribution. *Left panel*: LS. *Right panel*: Rank

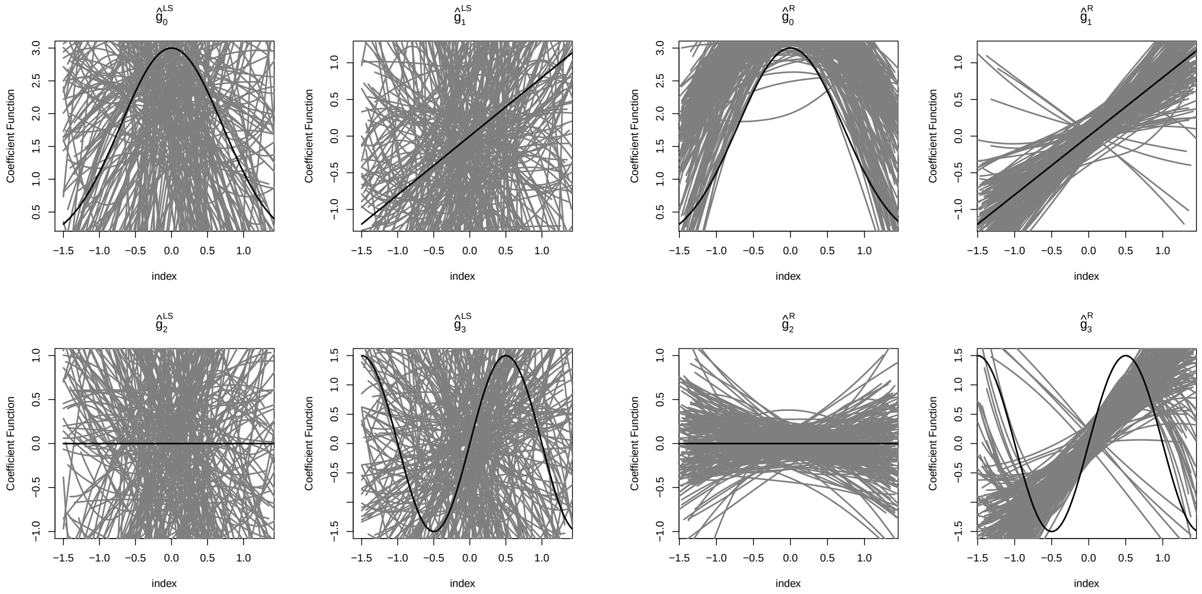

Figure 2.4: Estimated coefficient functions under Cauchy error distribution. *Left panel*: LS. *Right panel*: Rank

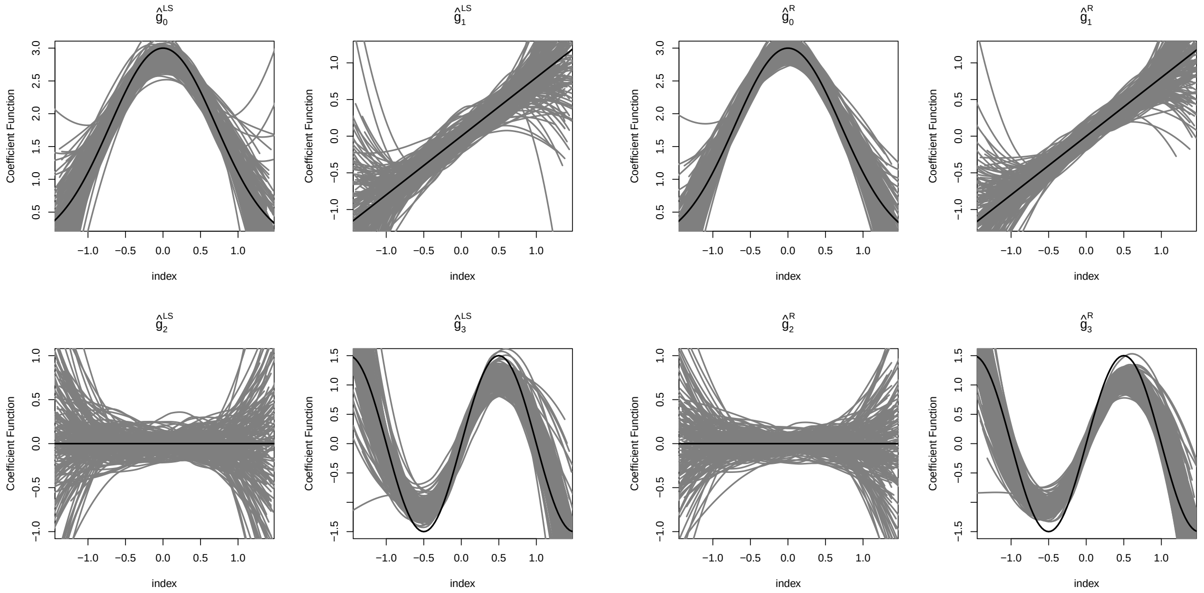

Figure 2.5: Estimated coefficient functions under Laplace error distribution. *Left panel*: LS. *Right panel*: Rank

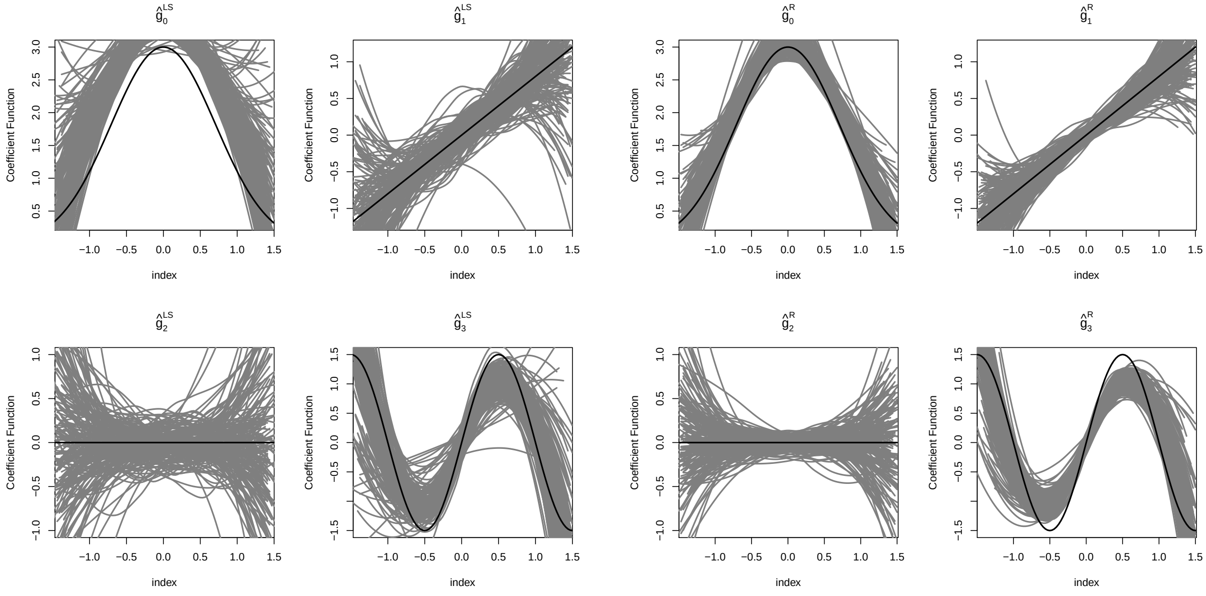

Figure 2.6: Estimated coefficient functions under Log Normal error distribution. *Left panel*: LS. *Right panel*: Rank

2.3.2 Real Data Example

In this section, we consider the fisheries data that motivated the development of our SIVCM estimation and discussed in the Section 1.2. The data were obtained to study interactions between groundfish predator species in the Gulf of Alaska. We are interested in investigating interspecific competition among these three predators while also considering Pacific halibut's role as a predator on Pacific cod and sablefish. We chose the response to be the CPUE of Pacific halibut to model how the CPUE of the other groundfish predators with similar diet preferences responded to the population numbers of an apex predator. The model we use is given in Equation (1.2). CPUEs and covariates in the matrix Z were centered and scaled to have mean zero and variance 1. We fit model (1.2) using both the LS and R methods for the full data, and this same process was then repeated after removing two identified outliers.

Considering the full data, one can see that the estimated coefficient functions (Figure 2.7) and the estimated index parameter (Table 2.3) from the LS and R methods are quite different. Also, from Figure 2.8, whether we consider the LS or the R fits, there are two apparent outliers, which makes the LS analysis inefficient. To evaluate the performance of the LS and R methods, a leave-one-out cross-validation was performed and prediction errors were computed. For the full data, as can be seen in Table 2.3, the R method provides more consistent and accurate estimates, and results in a smaller prediction error compared to the LS method. However, after removing the two outliers, the prediction error for R did not change (1.4%) while the one for LS changed quite substantially (62.5%). Moreover, the residuals look approximately normally distributed for both the LS and R methods. The estimated index parameters for the two approaches are similar for the two methods; however, LS now provides better performance in terms of prediction error (Table 2.4, Figure 2.9 and Figure 2.10). This is not surprising as for model errors with distribution close to normal, we expect the LS to have better performance than the R method, as was shown in the simulation study. This real data example demonstrates that while the LS fit is highly affected by the identified outliers, the R fit on the other hand shows its robustness, as the R estimates and the corresponding predictor error are very similar for data with and without outliers (Tables 2.3 and 2.4).

Table 2.3: Estimated value of $\boldsymbol{\theta}_0$ and prediction error (LOOCV) for the model using full data

$\widehat{\boldsymbol{\theta}}$	$\widehat{\theta}_1$	$\widehat{\theta}_2$	$\widehat{\theta}_3$	$\widehat{\theta}_4$	$\widehat{\theta}_5$	$\widehat{\theta}_6$	$\widehat{\theta}_7$	Pred. Err
LS	0.0707	0.1601	-0.1492	-0.4929	0.8233	-0.0033	-0.1621	1.3108
R	0.0432	0.1908	0.1216	-0.7240	0.5983	-0.0447	-0.2505	0.7227

Table 2.4: Estimated value of $\boldsymbol{\theta}_0$ and prediction error (LOOCV) for the model using data without outliers

$\widehat{\boldsymbol{\theta}}$	$\widehat{\theta}_1$	$\widehat{\theta}_2$	$\widehat{\theta}_3$	$\widehat{\theta}_4$	$\widehat{\theta}_5$	$\widehat{\theta}_6$	$\widehat{\theta}_7$	Pred. Err
LS	0.1048	0.1334	0.2249	-0.7716	0.4650	-0.0459	-0.3271	0.4922
R	0.0876	0.1501	0.1656	-0.7848	0.4989	-0.0669	-0.2704	0.7129

It is well known that outliers are a common feature to environmental data, and thus, robust methods should be required for better prediction with such data. Our simulation study and illustrative real data example demonstrate that the proposed R method is able to better handle contaminated and heavy-tailed model error distributions, or data containing outliers in the response space compared to the LS method. The estimated coefficient functions shown in Figures 2.7 and 2.9 share a change in functional pattern around $\theta^T Z = -1$ for Pacific cod (g_1) and sablefish (g_2). Overall Pacific halibut CPUE is decreasing along the estimated environmental matrix according to the direction of g_0. The function g_1 is sharply decreasing when $\theta^T Z < -1$ but increases when $\theta^T Z > 0$, indicating that Pacific cod CPUE negatively affects the CPUE of Pacific halibut when below a threshold of environmental factors but positively affect Pacific halibut CPUE when above a threshold. The function g_2 is increasing until a leveling-off occurs around $\theta^T Z = 1$ followed by a slight decline, suggesting that the CPUE of sablefish has a positive effect on the CPUE of Pacific halibut.

Temperature of surrounding waters has been known to affect the ability of Pacific halibut and sablefish to detect and strike bait in laboratory studies (Stoner *et al.*, 2006; Stoner & Sturm, 2004), where increased temperature was shown to increase the ability of both species to locate, attack, and consume baits prepared with squid. The MESA studies also bait with squid to survey groundfish and obtain CPUE measures for population approximations. Less is known about the response of groundfish to other environmental factors, particularly outside of laboratory conditions. The above model and results imply a hidden threshold variable consisting of multiple environmental factors that impact the CPUE dynamics of Pacific halibut. Note also that the environmental threshold for Pacific cod CPUE is different from that of sablefish CPUE, highlighting the variable response of different fish species to the same environmental factors. This is consistent with our preliminary observation given in Section 2.1. The models presented here motivate further study on these environmental variables and their affect on groundfish ability to detect and consume prey. Robust estimation methods such as the rank-based technique we have presented in this paper allow models such as these to more accurately predict multiple organisms' responses to environmental variations in the presence of outliers common to ecological data.

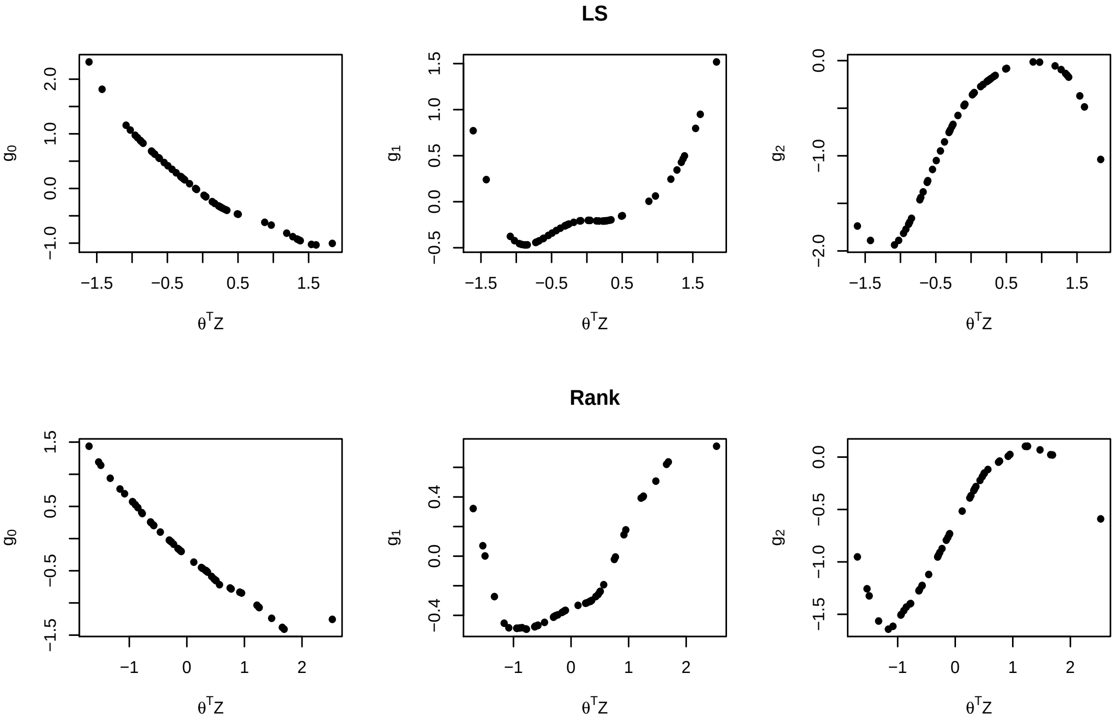

Figure 2.7: Estimated coefficient functions for the log transformed data with outliers. *Top*: LS estimators. *Bottom*: Rank estimators.

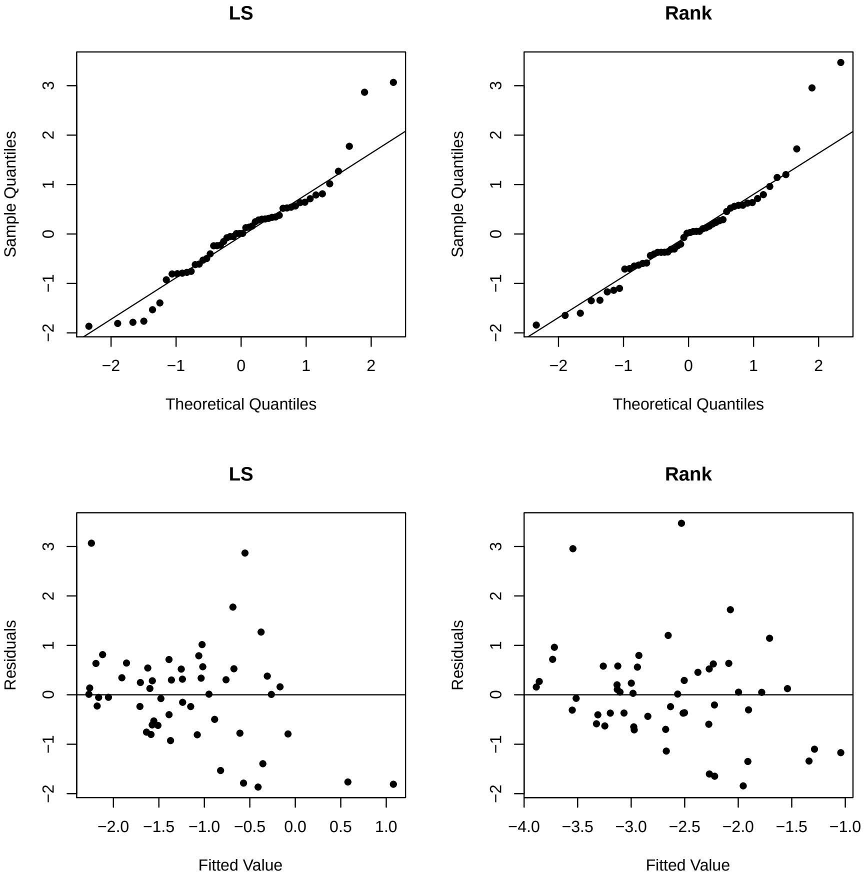

Figure 2.8: QQ plot of the residuals for the log transformed data with outliers. *Left*: LS method. *Right*: Rank method.

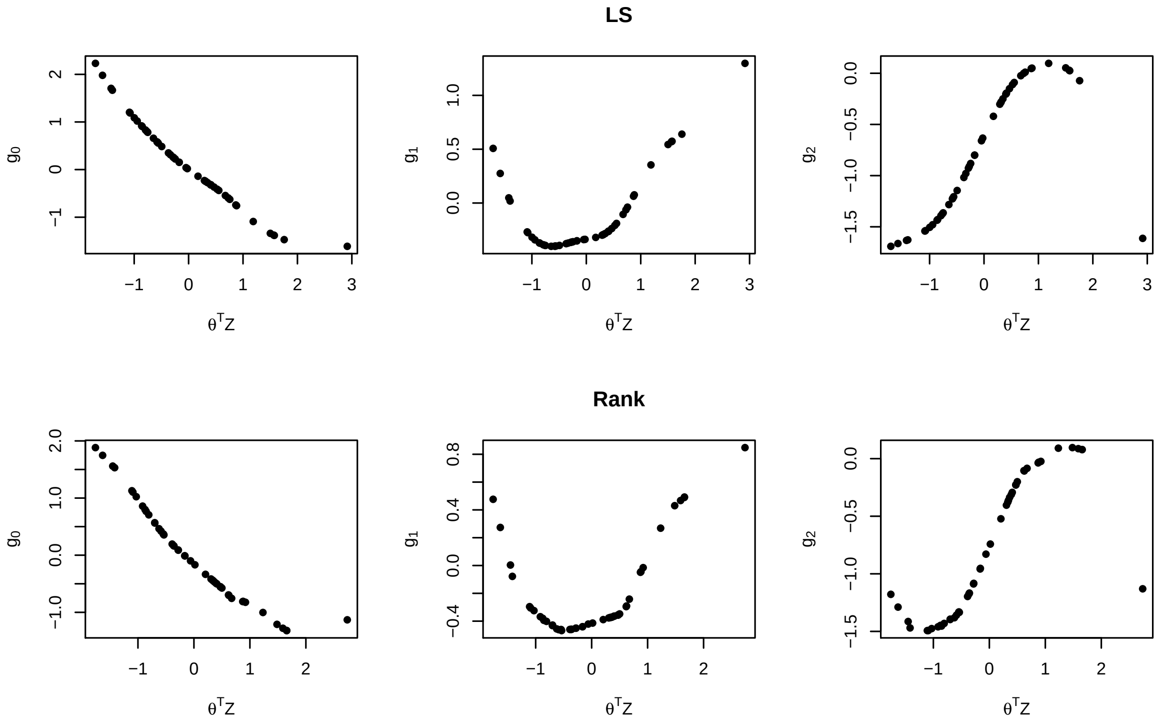

Figure 2.9: Estimated coefficient functions for the log transformed data without outliers. *Top*: LS estimators. *Bottom*: Rank estimators.

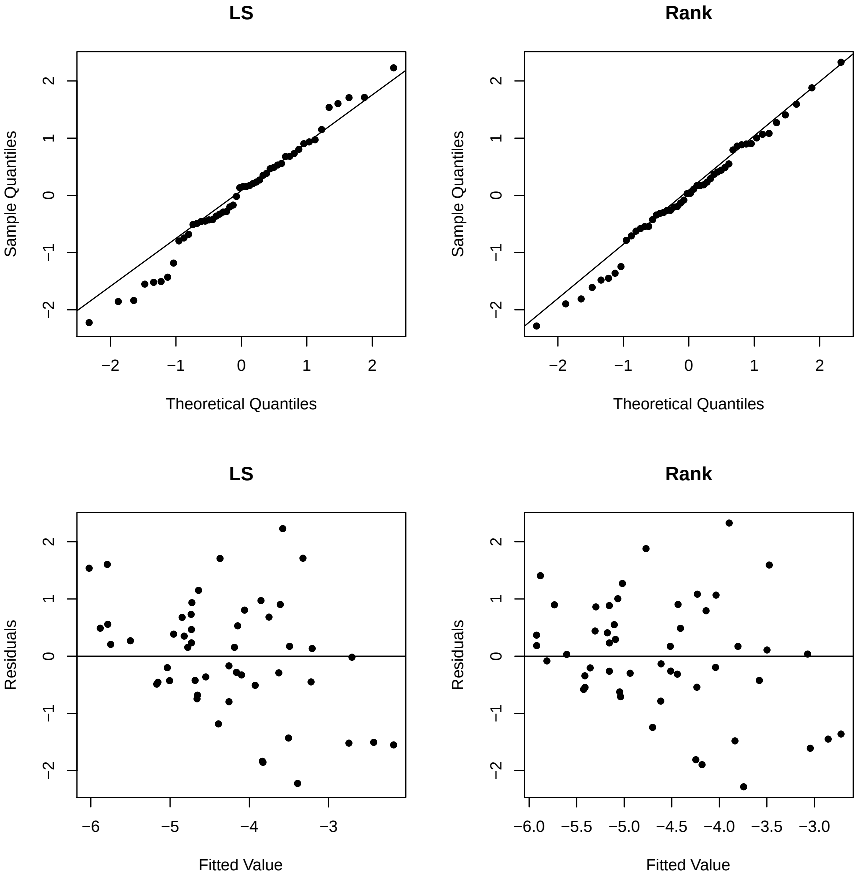

Figure 2.10: QQ plot of the residuals of the log transformed data without outliers. *Left*: LS method. *Right*: Rank method.

2.4 Asymptotic Properties of the rank-based estimators

Under the assumptions given below, Lemmas A1-A3 given in Xia *et al.* (2007) hold and can be used to prove that $D_n(\boldsymbol{\theta})$ and $L_n(\boldsymbol{\theta}, a, b)$ are asymptotically equivalent in the sense that

$$\lim_{n\to\infty} \sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta},a,b} |L_n(\boldsymbol{\theta}, a, b) - D_n(\boldsymbol{\theta})| = 0 \ a.s.$$

These lemmas will not be included here, and readers seeking for more details are referred to the aforementioned paper. To this end, we establish the asymptotic properties of $\widehat{\boldsymbol{\theta}}_n$ based on $D_n(\boldsymbol{\theta})$, while computations are performed using $L_n(\boldsymbol{\theta}, a, b)$. Let (X_i, Z_i, y_i), $i = 1, \ldots, n$ be a random sample with X_i and Z_i being i.i.d. and X_i independent of Z_i. Throughout this paper, we consider the following assumptions:

(I_1) φ is a nondecreasing, bounded and twice continuously differentiable score function with bounded derivatives, defined on $(0, 1)$, and assume that φ can be standardized as

$$\int_0^1 \varphi(u)du = 0 \quad and \quad \int_0^1 \varphi^2(u)du = 1.$$

(I_2) ε_i, $i = 1, \ldots, n$, are continuous errors with common distribution F and finite Fisher information.

(I_3) $g_j(\cdot)$, $j = 0, \ldots, p$, is a function defined on $\mathcal{A} = \{\boldsymbol{\theta}^T Z : \ \boldsymbol{\theta} \in \boldsymbol{\Theta}, \ Z \in \mathbb{R}^p\}$, where $\boldsymbol{\Theta}$ is a compact subspace of $\mathbb{R}^p$. There exists a function $J_j(\cdot)$ not necessarily the same, independent of $\boldsymbol{\theta}$ such that $\|\nabla^r_{\boldsymbol{\theta}} \kappa_j(\boldsymbol{\theta}, Z)\| \leqslant J_j(Z)$, for $r = 0, 1, 2, 3$, $\kappa_j(\boldsymbol{\theta}, Z) = g_j(Z^\tau \boldsymbol{\theta})$ is three times continuously differentiable with respect to $\boldsymbol{\theta}$, and $E[J_j^\alpha(Z)] < \infty$, for some $\alpha \geqslant 1$. Also, the function $t \mapsto G(t)$ is twice differentiable for any $t \in \mathcal{A}$. Moreover, for identifiability reasons, as assumed in Theorem 1 of Fan *et al.* (2003), for $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)^T$ with $\theta_p \neq 0$, we assume that $g_p(\cdot) \equiv 0$.

(I_4) Letting m and $m_{\boldsymbol{\theta}}$ be the density functions of $\boldsymbol{\theta}_0^T Z$ and $\boldsymbol{\theta}^T Z$, respectively, we assume that $m_{\boldsymbol{\theta}}$ have bounded continuous derives up to order 2. Also, $\inf\limits_{t \in \mathcal{A}} m_{\boldsymbol{\theta}}(t) > \alpha$, with $\alpha > 0$, for all $\boldsymbol{\theta} \in \boldsymbol{\Theta}$.

(I_5) $K(\cdot)$ is a regular kernel function with bandwidth h_n satisfying $h_n \to 0$ and $nh_n^{p+2}/\log n \to \infty$ as $n \to \infty$.

(I_6) $\sup\limits_{x,z} E[|Y|^r | X = x, Z = z] < \infty$, where r is the order of smoothness of $g_j(\cdot)$. Also,

$$E[\exp\{\lambda \|X\|\}] < \infty \quad \text{and} \quad E[\exp\{\lambda \|Z\|\}] < \infty, \text{ for some } \lambda > 0.$$

(I_7) $\boldsymbol{\theta}_0 \in Int(\boldsymbol{\Theta})$ and for fixed n, there exists a unique $\boldsymbol{\theta}_{0,n} \in Int(\boldsymbol{\Theta})$, a minimizer of $E[D_n(\boldsymbol{\theta})]$ such that $\boldsymbol{\theta}_0 = \lim\limits_{n \to \infty} \boldsymbol{\theta}_{0,n}$.

(I_8) Set $A_i = \nabla_{\boldsymbol{\theta}_0}[G(\boldsymbol{\theta}_0^T Z_i)] X_i$ and assume that $n^{-1}\mathbf{A}\mathbf{A}^T = n^{-1}\sum_{i=1}^n A_i A_i^T \to \boldsymbol{\Sigma} = E[AA^T]$ is positive definite matrix, where $\mathbf{A} = (A_1, \ldots, A_n)$.

Remark 3. *(I_1) and (I_2) are regular assumptions in the framework of rank-based estimation; see Hettmansperger & McKean (2011). Assumptions $(I_3) - (I_6)$ on the other hand, are regular assumption for estimation problems involving single-index models and ensure the strong consistency of the estimator of $G(\cdot)$; see Hansen (2008) and Gu & Yang (2015). (I_6) is the identifiability assumption from which, together with the previous assumptions, ensure the strong consistency of the proposed estimator that is established in Theorem 2.1. (I_7) and (I_8) together with the previous assumptions are used to establish the asymptotic distribution of the proposed estimator.*

2.4.1 Consistency

From assumption (I_2), when $\boldsymbol{\theta} \neq \boldsymbol{\theta}_0$, the $\eta_i(\boldsymbol{\theta})$ are still independent but not necessarily identically distributed. Also, under the local linear approximation, $\eta_i(\boldsymbol{\theta}, a, b) \approx \eta_i(\boldsymbol{\theta})$. Let

F_i be the probability distribution of $\eta_i(\boldsymbol{\theta})$. The following theorem, whose proof is provide in the Appendix, relies on the next lemma, and gives the strong consistency $\widehat{\boldsymbol{\theta}}_n$ with respect to $\boldsymbol{\theta}_0$.

Theorem 2.1. *Under $(I_1) - (I_3)$ and $(I_6) - (I_7)$, $\widehat{\boldsymbol{\theta}}_n \to \boldsymbol{\theta}_0$ a.s. as $n \to \infty$.*

Lemma 1. *Let $\{A_n(\boldsymbol{\theta})\}_{n\geqslant 1}$ be a random objective function defined on a compact space $\boldsymbol{\Theta}$ such that $\widehat{\boldsymbol{\theta}}_n = \underset{\boldsymbol{\theta}\in\boldsymbol{\Theta}}{\operatorname{Argmin}} A_n(\boldsymbol{\theta})$ and for fixed n, there is a unique $\boldsymbol{\theta}_{0,n} \in \boldsymbol{\Theta}$ that satisfies $\boldsymbol{\theta}_{0,n} = \underset{\boldsymbol{\theta}\in\boldsymbol{\Theta}}{\operatorname{Argmin}} E(A_n(\boldsymbol{\theta}))$, with $E(A_n(\boldsymbol{\theta}))$ being continuous with respect to $\boldsymbol{\theta}$. Furthermore, assume that for $\boldsymbol{\theta}_0 \in \boldsymbol{\Theta}$, $\boldsymbol{\theta}_0 = \lim_{n\to\infty} \boldsymbol{\theta}_{0,n}$.*

(i) If $\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} |A_n(\boldsymbol{\theta}) - E(A_n(\boldsymbol{\theta}))| \to 0$ a.s. as $n \to \infty$, then, $\widehat{\boldsymbol{\theta}}_n \to \boldsymbol{\theta}_0$ a.s. as $n \to \infty$.

(ii) If for every $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, $A_n(\boldsymbol{\theta}) - E(A_n(\boldsymbol{\theta})) \to 0$ a.s. as $n \to \infty$ and $A_n(\boldsymbol{\theta})$ is stochastically equicontinuous then, $\widehat{\boldsymbol{\theta}}_n \to \boldsymbol{\theta}_0$ a.s. as $n \to \infty$.

The proof of this lemma can be found in Andrews (1994), Newey & McFadden (1994), and Rao *et al.* (2014), so for the sake of brevity it will not be included here.

2.4.2 Asymptotic Normality

Let $\nabla_{\boldsymbol{\theta}} = \big(\partial/\partial\theta_i\big)_i$ and $\nabla^2_{\boldsymbol{\theta}} = \big(\partial^2/\partial\theta_i\partial\theta_j\big)_{ij}$, for $\boldsymbol{\theta} = (\theta_1, \cdots, \boldsymbol{\theta}_p)^\tau$, $i, j = 1, \ldots, p$, denote the gradient and Hessian operators, respectively. Also, $\nabla^r_{\boldsymbol{\xi}}[G(\boldsymbol{\xi}^T Z)] = \nabla^r_{\boldsymbol{\theta}}[G(\boldsymbol{\theta}^T Z)]|_{\boldsymbol{\theta}=\boldsymbol{\xi}}$ for some arbitrary $\boldsymbol{\xi}$ and $r = 1, 2, 3$. Under the smoothness assumption on φ and g_j, $k = 1, \ldots, p$, $D_n(\boldsymbol{\theta})$ is weakly differentiable. From now, set $S_n(\boldsymbol{\theta}) = -\nabla_{\boldsymbol{\theta}} D_n(\boldsymbol{\theta})$. With $\widehat{\boldsymbol{\theta}}_n$ being a minimizer of $D_n(\boldsymbol{\theta})$, we have $S_n(\widehat{\boldsymbol{\theta}}_n) = 0$. Explicitly, $S_n(\boldsymbol{\theta})$ is given by $S_n(\boldsymbol{\theta}) = n^{-1}\sum_{i=1}^n \nabla_{\boldsymbol{\theta}}[G(\boldsymbol{\theta}^T Z_i)] X_i \varphi\left(R(\eta_i(\boldsymbol{\theta}))/(n+1)\right)$.

At the true parameter $\boldsymbol{\theta}_0$, $S_n(\boldsymbol{\theta}_0) = n^{-1}\sum_{i=1}^n \nabla_{\boldsymbol{\theta}_0}[G(\boldsymbol{\theta}_0^T Z_i)] X_i \varphi\left(R(\varepsilon_i)/(n+1)\right)$. Also, define $T_n(\boldsymbol{\theta}_0)$ as

$$T_n(\boldsymbol{\theta}_0) = \frac{1}{n}\sum_{i=1}^n \nabla_{\boldsymbol{\theta}_0}[G(\boldsymbol{\theta}_0^T Z_i)] X_i \varphi\left(F(\varepsilon_i)\right).$$

The following theorem gives the equivalence of the two estimating functions.

Theorem 2.2. *Under assumptions* $(I_1)-(I_8)$, $\sqrt{n}\big(S_n(\boldsymbol{\theta}_0)-T_n(\boldsymbol{\theta}_0)\big) \xrightarrow{P} 0$ *and* $\sqrt{n}T_n(\boldsymbol{\theta}_0) \xrightarrow{\mathcal{D}} N(\mathbf{0},\boldsymbol{\Sigma})$ *as* $n \to \infty$. *Moreover,* $\lim_{n\to\infty} \sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \|S_n(\boldsymbol{\theta}) - T_n(\boldsymbol{\theta})\| = 0$ *a.s.*

This theorem implies that $\sqrt{n}S_n(\boldsymbol{\theta}_0)$ and $\sqrt{n}T_n(\boldsymbol{\theta}_0)$ have the same asymptotic distribution. On the other hand, with probability 1, $S_n(\boldsymbol{\theta}) = T_n(\boldsymbol{\theta}) + o(1)$. A Taylor expansion of $T_n(\boldsymbol{\theta})$ around $\boldsymbol{\theta}_0$ gives

$$T_n(\boldsymbol{\theta}) = T_n(\boldsymbol{\theta}_0) + (\boldsymbol{\theta}-\boldsymbol{\theta}_0)^{\tau}\nabla_{\boldsymbol{\theta}}T_n(\boldsymbol{\theta}_0) + \frac{1}{2}(\boldsymbol{\theta}-\boldsymbol{\theta}_0)^{\tau}\nabla^2_{\boldsymbol{\theta}}T_n(\boldsymbol{\xi})(\boldsymbol{\theta}-\boldsymbol{\theta}_0),$$

where $\boldsymbol{\xi}$ belongs in the line segment joining $\boldsymbol{\theta}_0$ and $\boldsymbol{\theta}$. Thus, with probability 1,

$$S_n(\boldsymbol{\theta}) = T_n(\boldsymbol{\theta}_0) + (\boldsymbol{\theta}-\boldsymbol{\theta}_0)^{\tau}\nabla_{\boldsymbol{\theta}}T_n(\boldsymbol{\theta}_0) + \frac{1}{2}(\boldsymbol{\theta}-\boldsymbol{\theta}_0)^{\tau}\nabla^2_{\boldsymbol{\theta}}T_n(\boldsymbol{\xi})(\boldsymbol{\theta}-\boldsymbol{\theta}_0) + o(1).$$

$\widehat{\boldsymbol{\theta}}_n$ being a solution of $S_n(\boldsymbol{\theta}) = 0$, we have

$$0 = S_n(\widehat{\boldsymbol{\theta}}_n) = T_n(\boldsymbol{\theta}_0) + \nabla_{\boldsymbol{\theta}}T_n(\boldsymbol{\theta}_0)\cdot(\widehat{\boldsymbol{\theta}}_n-\boldsymbol{\theta}_0) + \frac{1}{2}(\widehat{\boldsymbol{\theta}}_n-\boldsymbol{\theta}_0)^{\tau}\cdot\nabla^2_{\boldsymbol{\theta}}T_n(\boldsymbol{\xi}_n)\cdot(\widehat{\boldsymbol{\theta}}_n-\boldsymbol{\theta}_0) + o(1), \quad (2.5)$$

where $\boldsymbol{\xi}_n = \lambda\boldsymbol{\theta}_0 + (1-\lambda)\widehat{\boldsymbol{\theta}}_n$.

Theorem 2.3. *Under assumptions* $(I_1)-(I_8)$, *the following hold:*

a. $\nabla_{\boldsymbol{\theta}}T_n(\boldsymbol{\theta}_0) \to \mathbf{W}$ *a.s., where* $\mathbf{W} = -E\{AA^{\tau}f(\varepsilon)\varphi'(F(\varepsilon))\} + E\{\nabla^2_{\boldsymbol{\theta}_0}[G(\boldsymbol{\theta}_0^T Z)]X\varphi(F(\varepsilon))\}$ *is a positive definite matrix, and*

b. $\nabla^2_{\boldsymbol{\theta}}T_n(\boldsymbol{\xi}_n)$ *is almost surely bounded.*

Note, if we assume that ε is independent of (X,Z), $\mathbf{W}$ can be expressed using the rank scale parameter as $\mathbf{W} = \gamma_{\varphi}^{-1}\boldsymbol{\Sigma}$ similar to the linear model case, where

$$\gamma_{\varphi}^{-1} = \int_0^1 \varphi(u)\varphi_f(u)du \quad \text{with} \quad \varphi_f(u) = \frac{f'(F^{-1}(u))}{f(F^{-1}(u))}.$$

To this end, the asymptotic normality distribution of the rank estimator is now obtained from that of $\sqrt{n}S_n(\boldsymbol{\theta}_0)$ and given in the following theorem:

Theorem 2.4. *Under assumptions $(I_1)-(I_8)$, we have $\sqrt{n}(\widehat{\boldsymbol{\theta}}_n-\boldsymbol{\theta}_0) \xrightarrow{\mathcal{D}} N\big(\mathbf{0},\mathbf{W}^{-1}\mathbf{\Sigma}\mathbf{W}^{-1}\big)$. Moreover, if ε is independent of (X,Z), we have $\sqrt{n}(\widehat{\boldsymbol{\theta}}_n-\boldsymbol{\theta}_0) \xrightarrow{\mathcal{D}} N\big(\mathbf{0},\gamma_\varphi^2\mathbf{\Sigma}^{-1}\big)$*

From now, define $\widehat{G}(t)=(\widehat{g}_0(t),\ldots,\widehat{g}_p(t))^T$ and $\widehat{G}''(t)=(\widehat{g}''_0(t),\ldots,\widehat{g}''_p(t))^T$, for $t\in\widetilde{\mathcal{A}}=\{t=\widehat{\boldsymbol{\theta}}_n^T Z\}$. The following theorem gives the asymptotic distribution of $\widehat{G}(t)$. The proof can be obtained in a similar way as given in Xia *et al.* (2007), and is therefore omitted here.

Theorem 2.5. *Let $\omega_n=\sqrt{\log n/nh_n}$, and suppose that the derivative with respect to t of the function*

$$t\mapsto\mathcal{W}_0(t)=-E\{XX^\tau f(\varepsilon)\varphi'(F(\varepsilon))|\boldsymbol{\theta}_0^T Z=t\}+E\{X\varphi(F(\varepsilon))|\boldsymbol{\theta}_0^T Z=t\}$$

exists. Then, under assumptions $(I_1)-(I_8)$, we have

$$\sup_{|t|\leqslant c}\|\widehat{G}(t)-G(t)\|=O_p((\omega_n+h_n^2)/c). \tag{2.6}$$

Moreover,

$$(nh_n)^{1/2}\{\widehat{G}(t)-G(t)-\boldsymbol{\eta}(t)h_n^2\}\xrightarrow{\mathcal{D}}N\Big(0,m^{-1}(t)\mathcal{W}_0^{-1}(t))\mathcal{W}_1(t)\mathcal{W}_0^{-1}(t)\int K^2(u)du\Big), \tag{2.7}$$

where

$$\boldsymbol{\eta}(t)=\frac{1}{2}G''(t)\int u^2K(u)du,\quad \textit{and}\quad \mathcal{W}_1(t)=E\{XX^T\varphi^2(F(\varepsilon))|\boldsymbol{\theta}_0^T Z=t\}$$

2.5 Discussion

This paper provides a rank-based procedure that is a robust and more efficient alternative to the least squares method for fitting the SIVCM when dealing with contaminated and heavy-tailed model error distributions, or when data contain outliers in the response space. It is worth pointing out that for high leverage points (outliers in the design space), the performance of the proposed method can be affected. When the design space is well controlled (in the absence of high leverage points) with outliers in the response space, we recommend the use of the proposed procedure. For cases where there are obvious outliers in the predictor variables, a weighted version of the considered rank objective function could be derived following ideas similar to those in Naranjo & Hettmansperger (1994), Chang *et al.* (1999), and Bindele & Abebe (2012).

2.6 Proofs of Theorems 2.1-2.4

This section presents proofs of theoretical results established in the paper.

Proofs

Proof of Theorem 2.1. From the fact that φ has a bounded first derivative, $\varphi \in Lip(1)$. Moreover, since g_k, $k = 0, \ldots, p$ are bounded on $\mathcal{A}$ and $\eta_i(\boldsymbol{\theta})$ depend on $\boldsymbol{\theta}$ only through g_k, we have $Var(\eta_i(\boldsymbol{\theta})) < \infty$ for all i and $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ by (I_5). Then

$$\sum_{i=1}^{n} \frac{Var(\eta_i(\boldsymbol{\theta}))}{n^2} \leqslant \frac{\sigma^2_{max}(\boldsymbol{\theta})}{n} = O(1/n),$$

where $\sigma^2_{max}(\boldsymbol{\theta}) = \max\{Var(\eta_1(\boldsymbol{\theta})), \ldots, Var(\eta_n(\boldsymbol{\theta}))\}$. Setting $\alpha_n = 1/n$ and $\beta = 1$ in the theorem of Xiang (1995), we find that for every $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, $D_n(\boldsymbol{\theta}) - E\{D_n(\boldsymbol{\theta})\} \to 0$ *a.s.* To complete the proof, we have to show that $\{D_n(\boldsymbol{\theta})\}_{n \geqslant 1}$ is stochastically equicontinuous. To that end, taking $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \in \boldsymbol{\Theta}$ and setting $a_{in}(\boldsymbol{\theta}) = R(\eta_i(\boldsymbol{\theta}))/(n+1)$ we have

$$\begin{aligned} D_n(\boldsymbol{\theta}_1) - D_n(\boldsymbol{\theta}_2) &= \frac{1}{n}\sum_{i=1}^{n}[\varphi(a_{in}(\boldsymbol{\theta}_1))\eta_i(\boldsymbol{\theta}_1) - \varphi(a_{in}(\boldsymbol{\theta}_2))\eta_i(\boldsymbol{\theta}_2)] \\ &= \frac{1}{n}\sum_{i=1}^{n}\varphi(a_{in}(\boldsymbol{\theta}_1))[\eta_i(\boldsymbol{\theta}_1) - \eta_i(\boldsymbol{\theta}_2)] + \frac{1}{n}\sum_{i=1}^{n}\left[\varphi\left(a_{in}(\boldsymbol{\theta}_1)\right) - \varphi\{F_i(\eta_i(\boldsymbol{\theta}_1))\}\right]\eta_i(\boldsymbol{\theta}_2) \\ &\quad + \frac{1}{n}\sum_{i=1}^{n}\left[\varphi\{F_i(\eta_i(\boldsymbol{\theta}_1))\} - \varphi\{F_i(\eta_i(\boldsymbol{\theta}_2))\}\right]\eta_i(\boldsymbol{\theta}_2) \\ &\quad + \frac{1}{n}\sum_{i=1}^{n}\left[\varphi\{F_i(\eta_i(\boldsymbol{\theta}_2))\} - \varphi\left(a_{in}(\boldsymbol{\theta}_2)\right)\right]\eta_i(\boldsymbol{\theta}_2). \end{aligned}$$

Note that $\eta_i(\boldsymbol{\theta}_1) - \eta_i(\boldsymbol{\theta}_2) = \{G(\boldsymbol{\theta}_2^T Z_i) - G(\boldsymbol{\theta}_1^T Z_i)\}^T X_i$. Since for $j = 0, \ldots, p$, $\kappa_j(\boldsymbol{\theta}, Z)$ is differentiable with respect to $\boldsymbol{\theta}$, G is differentiable with respect to $\boldsymbol{\theta}$, by assumption (I_3). Thus, from the mean value theorem on the vector function G there exists $\boldsymbol{\xi} = \lambda\boldsymbol{\theta}_1 + (1-\lambda)\boldsymbol{\theta}_2$ for some $\lambda \in (0,1)$ such that $G(\boldsymbol{\theta}_1^\tau Z_i) - G(\boldsymbol{\theta}_2^\tau Z_i) = \nabla_{\boldsymbol{\xi}}[G(\boldsymbol{\xi}^\tau Z_i)](\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2)$. Thus by assumption (I_3)

$$\|G(\boldsymbol{\theta}_1^\tau Z_i) - G(\boldsymbol{\theta}_2^\tau Z_i)\| = \|\nabla_{\boldsymbol{\xi}}[G(\boldsymbol{\xi}^\tau Z_i)](\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2)\| \leqslant \sum_{j=1}^{p} J_j(Z_i)\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|.$$

Furthermore, set $h_i(\boldsymbol{\theta}) = \varphi\{F_i(\eta_i(\boldsymbol{\theta}))\} = \varphi\{F_i(Y_i - [G(\boldsymbol{\theta}^\tau Z_i)]^\tau X_i)\}$, where F_i is a cumulative distribution function of $\eta_i(\boldsymbol{\theta})$ and therefore almost surely differentiable. So by the mean value theorem, there exists $\boldsymbol{\zeta} = \lambda\boldsymbol{\theta}_1 + (1-\lambda)\boldsymbol{\theta}_2$ for $\lambda \in (0,1)$ such that $h_i(\boldsymbol{\theta}_1) - h_i(\boldsymbol{\theta}_2) = h_i'(\boldsymbol{\zeta})(\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2)$, with $h_i'(\boldsymbol{\zeta}) = -\nabla_{\boldsymbol{\zeta}}[G(\boldsymbol{\zeta}^\tau Z_i)]X_i f_i(\eta_i(\boldsymbol{\zeta}))\varphi'\{F_i(\eta_i(\boldsymbol{\zeta}))\}$ and $f_i(t) = dF_i(t)/dt$. It is worth pointing out that f_i being a density, is almost surely bounded. By assumption (I_3) again together with the boundedness of φ', we have $\|h_i'(\boldsymbol{\zeta})\| \leqslant M\sum_{j=1}^{p} J_j(Z_i)$ *a.s.*, where M is such that $|f_i(\eta_i(\boldsymbol{\zeta}))\varphi'\{F_i(\eta_i(\boldsymbol{\zeta}))\}| \leqslant M$ *a.s.* On the other hand, for $i = 1, \ldots, n$, $F_i(\eta_i(\boldsymbol{\theta}))$ being independent uniformly distributed in the interval $(0,1)$, for all $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, following Hájek & Šidák (1967) in chapter 6, it is obtained that $a_{ni}(\boldsymbol{\theta}) - F_i(\eta_i(\boldsymbol{\theta})) \to 0$ *a.s.*, for all $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ and for each i. By continuity of φ, we have $\varphi\left(a_{ni}(\boldsymbol{\theta})\right) - \varphi\{F_i(\eta_i(\boldsymbol{\theta}))\} \to 0$ *a.s.*, for all $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ and for each i. Thus,

$\max\limits_{1\leqslant i\leqslant n} |\varphi\left(a_{ni}(\boldsymbol{\theta})\right) - \varphi\{F_i(\eta_i(\boldsymbol{\theta}))\}| \to 0$ $a.s.$, for all $\boldsymbol{\beta} \in \boldsymbol{\Theta}$. To this end,

$$\begin{aligned} \left|\frac{1}{n}\sum_{i=1}^{n} \varphi(a_{in}(\boldsymbol{\theta}_1))[\eta_i(\boldsymbol{\theta}_1) - \eta_i(\boldsymbol{\theta}_2)]\right| &\leqslant \frac{1}{n}\sum_{i=1}^{n} |\varphi(a_{in}(\boldsymbol{\theta}_1))\|G(\boldsymbol{\theta}_1^\tau Z_i) - G(\boldsymbol{\theta}_2^\tau Z_i)\|\|X_i\| \\ &\leqslant \|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|\frac{L}{n}\sum_{i=1}^{n}\sum_{j=1}^{p} \|X_i\|J_j(Z_i), \end{aligned}$$

where L is such that $|\varphi(t)| \leqslant L$, for all $t \in (0,1)$. Also, with probability 1, we have

$$\begin{aligned} & \left|\frac{1}{n}\sum_{i=1}^{n} \left[\varphi\{F_i(\eta_i(\boldsymbol{\theta}_1))\} - \varphi\{F_i(\eta_i(\boldsymbol{\theta}_2))\}\right]\eta_i(\boldsymbol{\theta}_2)\right| \\ \leqslant\ & \|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|\frac{M}{n}\sum_{i=1}^{n}\sum_{j=1}^{p} \|X_i\|J_j(Z_i)|\eta_i(\boldsymbol{\theta}_2)| \\ \leqslant\ & \|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|M\left(\frac{1}{n}\sum_{i=1}^{n}\left\{\sum_{j=1}^{p}\|X_i\|J_j(Z_i)\right\}^2\right)^{1/2}\left(\frac{1}{n}\sum_{i=1}^{n}|z_i(\boldsymbol{\beta}_2)|^2\right)^{1/2} \\ \leqslant\ & \|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|M\left(\frac{1}{n}\sum_{i=1}^{n}\left\{\sum_{j=1}^{p}\|X_i\|J_j(Z_i)\right\}^2\right)^{1/2}\left(\frac{1}{n}\sum_{i=1}^{n}\left[|Y_i| + \sum_{j=1}^{p}\|X_i\|J_j(Z_i)\right]^2\right)^{1/2}. \end{aligned}$$

Moreover,

$$\begin{aligned} & \left|\frac{1}{n}\sum_{i=1}^{n}\left[\varphi\left(a_{in}(\boldsymbol{\theta}_1)\right) - \varphi\{F_i(\eta_i(\boldsymbol{\theta}_1))\}\right]\eta_i(\boldsymbol{\theta}_2)\right| \\ \leqslant\ & \frac{1}{n}\sum_{i=1}^{n}\left|\varphi\left(a_{in}(\boldsymbol{\theta}_1)\right) - \varphi\{F_i(\eta_i(\boldsymbol{\theta}_1))\}\right||\eta_i(\boldsymbol{\theta}_2)| \\ \leqslant\ & \left(\max_{1\leqslant i\leqslant n}\left|\varphi\left(a_{in}(\boldsymbol{\theta}_1)\right) - \varphi\{F_i(\eta_i(\boldsymbol{\theta}_1))\}\right|^2\right)^{1/2}\left(\frac{1}{n}\sum_{i=1}^{n}\left[|Y_i| + \sum_{j=1}^{p}\|X_i\|J_j(Z_i)\right]^2\right)^{1/2} \to 0\ a.s., \end{aligned}$$

as $\max\limits_{1\leqslant i\leqslant n} |\varphi\left(a_{ni}(\boldsymbol{\beta}_1)\right) - \varphi\{F_i(z_i(\boldsymbol{\beta}_1))\}|^2 \to 0 \quad a.s.$ and $\left(n^{-1}\sum_{i=1}^{n}\left[|Y_i| + \sum_{j=1}^{p}\|X_i\|J_j(Z_i)\right]^2\right)^{1/2}$ which converges almost surely to a finite quantity by the strong law of large numbers under assumptions (I_3) and (I_4). Similarly, $\left|n^{-1}\sum_{i=1}^{n}\left[\varphi\{F_i(\eta_i(\boldsymbol{\theta}_2))\} - \varphi\left(a_{in}(\boldsymbol{\theta}_2)\right)\right]\eta_i(\boldsymbol{\theta}_2)\right|$ converges almost surely to zero. Hence, with probability 1, we have $|D_n(\boldsymbol{\beta}_1) - D_n(\boldsymbol{\beta}_2)| \leqslant B_n\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|$, where

$$\begin{aligned} B_n \quad =: & \quad \frac{L}{n}\sum_{i=1}^{n}\sum_{j=1}^{p}\|X_i\|J_j(Z_i) \\ & +M\left(\frac{1}{n}\sum_{i=1}^{n}\left\{\sum_{j=1}^{p}\|X_i\|J_j(Z_i)\right\}^2\right)^{1/2}\left(\frac{1}{n}\sum_{i=1}^{n}\left[|Y_i| + \sum_{j=1}^{p}\|X_i\|J_j(Z_i)\right]^2\right)^{1/2} + o(1). \end{aligned}$$

For n large enough, B_n is independent of $\boldsymbol{\theta}$, and from the fact that all terms in the definition of B_n converge almost surely to a finite quantity, so does B_n. Therefore, $\{D_n(\boldsymbol{\theta})\}_{n\geqslant 1}$ is stochastically equicontinuous (Rao *et al.*, 2014), and so, the proof is complete. □

Proof of Theorem 2.2. $E\{S_n(\boldsymbol{\theta}_0) - T_n(\boldsymbol{\theta}_0)\} = \dfrac{1}{n}\sum_{i=1}^{n}\nabla_{\boldsymbol{\theta}_0}[G(\boldsymbol{\theta}_0^T Z_i)]X_i E\left(\varphi\left(\dfrac{R(\varepsilon_i)}{n+1}\right) - \varphi\left(F(\varepsilon_i)\right)\right)$. By Schwartz inequality, we have

$$E\{S_n(\boldsymbol{\theta}_0) - T_n(\boldsymbol{\theta}_0)\} \leqslant \left\{\frac{1}{n}\sum_{i=1}^{n}\{\nabla_{\boldsymbol{\theta}_0}[G(\boldsymbol{\theta}_0^T Z_i)]X_i\}^2\right\}^{1/2}\left\{\frac{1}{n}\sum_{i=1}^{n}\left(E\left(\varphi\left(\frac{R(\varepsilon_i)}{n+1}\right) - \varphi\left(F(\varepsilon_i)\right)\right)\right)^2\right\}^{1/2}.$$

Also, by Jensen's inequality, we have

$$\left(E\left(\varphi\left(\frac{R(\varepsilon_i)}{n+1}\right) - \varphi\left(F(\varepsilon_i)\right)\right)\right)^2 \leqslant E\left[\left(\varphi\left(\frac{R(\varepsilon)}{n+1}\right) - \varphi\left(F(\varepsilon)\right)\right)^2\right].$$

Then,

$$\begin{aligned} E\{S_n(\boldsymbol{\theta}_0) - T_n(\boldsymbol{\theta}_0)\} \quad & \leqslant \quad \left\{\frac{1}{n}\sum_{i=1}^{n}\{\nabla_{\boldsymbol{\theta}_0}[G(\boldsymbol{\theta}_0^T Z_i)]X_i\}^2\right\}^{1/2}\left\{\frac{1}{n}\sum_{i=1}^{n}E\left[\left(\varphi\left(\frac{R(\varepsilon_i)}{n+1}\right) - \varphi\left(F(\varepsilon_i)\right)\right)^2\right]\right\}^{1/2} \\ & \leqslant \quad \left\{\frac{1}{n}\sum_{i=1}^{n}\{\nabla_{\boldsymbol{\theta}_0}[G(\boldsymbol{\theta}_0^T Z_i)]X_i\}^2\right\}^{1/2}\left\{E\left[\left(\varphi\left(\frac{R(\varepsilon)}{n+1}\right) - \varphi\left(F(\varepsilon)\right)\right)^2\right]\right\}^{1/2}. \quad (2.8) \end{aligned}$$

By the continuity of φ and the fact that $R(\varepsilon)/(n+1) \to F(\varepsilon)$ *a.s.* as $n \to \infty$ (Hájek & Šidák, 1967), applying the Dominated Convergence Theorem gives $E\{[\varphi\left(R(\varepsilon)/(n+1)\right) - \varphi\left(F(\varepsilon)\right)]^2\} \to 0$. On the other hand, by assumption (I_3), $\|\nabla_{\boldsymbol{\theta}_0}[G(\boldsymbol{\theta}_0^T Z_i)]\| \leqslant \sum_{j=0}^{p} J_j(Z_i)$. Then,

$$\frac{1}{n}\sum_{i=1}^{n}\{\nabla_{\boldsymbol{\theta}_0}[G(\boldsymbol{\theta}_0^T Z_i)]X_i\}^2 \leqslant \frac{1}{n}\sum_{i=1}^{n}\left\{\sum_{j=0}^{p}\|X_i\|J_j(Z_i)\right\}^2 \to E\left[\|X\|^2\left\{\sum_{j=0}^{p}J_j(Z)\right\}^2\right] < \infty \; a.s.,$$

by (I_6) and the strong law of large numbers. Thus, $E\{S_n(\boldsymbol{\theta}_0) - T_n(\boldsymbol{\theta}_0)\} \to 0$ as $n \to \infty$. By Chebychev's inequality, for any $\epsilon > 0$, we have

$$P\left(\sqrt{n}\big(S_n(\boldsymbol{\theta}_0) - T_n(\boldsymbol{\theta}_0)\big) > \epsilon\right) \leqslant \frac{1}{\epsilon^2}E\left[n\big(S_n(\boldsymbol{\theta}_0) - T_n(\boldsymbol{\theta}_0)\big)^2\right].$$

To complete the proof, it sufficies to show that $E\big[n\big(S_n(\boldsymbol{\theta}_0) - T_n(\boldsymbol{\theta}_0)\big)^2\big] \to 0$ as $n \to \infty$. Indeed,

$$\begin{aligned} E\left[n\big(S_n(\boldsymbol{\theta}_0) - T_n(\boldsymbol{\theta}_0)\big)^2\right] &= \frac{1}{n}E\left[\left\{\sum_{i=1}^{n}\nabla_{\boldsymbol{\theta}_0}[G(\boldsymbol{\theta}_0^T Z_i)]X_i\left(\varphi\left(\frac{R(\varepsilon_i)}{n+1}\right) - \varphi\left(F(\varepsilon_i)\right)\right)\right\}^2\right] \\ &\leqslant \frac{n}{n-1}\left\{\frac{1}{n}\sum_{i=1}^{n}\{\nabla_{\boldsymbol{\theta}_0}[G(\boldsymbol{\theta}_0^T Z_i)]X_i\}^2\right\}E\left[\left(\varphi\left(\frac{R(\varepsilon)}{n+1}\right) - \varphi\left(F(\varepsilon)\right)\right)^2\right] \end{aligned}$$

Note that from the discussion following equation (2.8), the right hand side of the above inequality converges to zero as $n \to \infty$. Thus, $E\left[n\big(S_n(\boldsymbol{\theta}_0) - T_n(\boldsymbol{\theta}_0)\big)^2\right] \to 0$ as $n \to \infty$ and therefore,

$$\lim_{n\to\infty} P\left(\sqrt{n}\big(S_n(\boldsymbol{\theta}_0) - T_n(\boldsymbol{\theta}_0)\big) > \epsilon\right) = 0.$$

Now, by assumption (I_2), we have $E\{T_n(\boldsymbol{\theta}_0)\} = 0$. To obtain the asymptotic distribution of $T_n(\boldsymbol{\theta}_0)$, we employ the Cramér-Wold device (Serfling, 1980). To this end, set

$$U = n^{-1/2}\sum_{i=1}^{n}\mathbf{a}^\tau\nabla_{\boldsymbol{\theta}_0}[G(\boldsymbol{\theta}_0^T Z_i)]X_i\varphi[F(\varepsilon_i)],$$

where $\mathbf{a} \in \mathbb{R}^q$. Since F is the distribution of ε and $\int_0^1 \varphi(t)dt = 0$, we have $E(U) = 0$. Also, since $\int_0^1 \varphi^2(t)dt = 1$,

$$\begin{aligned} Var(U) &= \frac{1}{n}\sum_{i=1}^n \left(\mathbf{a}^\tau \nabla_{\boldsymbol{\theta}_0}[G(\boldsymbol{\theta}_0^T Z_i)]X_i\right)^2 E\left\{\varphi^2(F(\varepsilon))\right\} \\ &= \frac{1}{n}\sum_{i=1}^n \left(\mathbf{a}^\tau \nabla_{\boldsymbol{\theta}_0}[G(\boldsymbol{\theta}_0^T Z_i)]X_i\right)^2 \rightarrow \mathbf{a}^\tau \boldsymbol{\Sigma} \mathbf{a} \quad a.s. \end{aligned}$$

Note that U is the sum of independent functions of random variables which may not be necessarily identically distributed. Hence, the limiting distribution is established by verifying the Lindeberg-Feller condition for the applicability of the Central Limit Theorem. To this end, set $\sigma_n^2 = Var(U)$. We need to show that

$$\lim_{n\to\infty} \frac{1}{\sigma_n^2}\sum_{i=1}^n E\left[\frac{1}{n}\left(\mathbf{a}^\tau \nabla_{\boldsymbol{\theta}_0}[G(\boldsymbol{\theta}_0^T Z_i)]X_i\right)^2 \varphi^2[F(\varepsilon_i)]\right] \times$$
$$I\left(\left|\frac{1}{\sqrt{n}}\left(\mathbf{a}^\tau \nabla_{\boldsymbol{\theta}_0}[G(\boldsymbol{\theta}_0^T Z_i)]X_i\right)\varphi[F(\varepsilon_i)]\right| > \epsilon\sigma_n\right) = 0.$$

To this end, we have $n^{-1/2}|\mathbf{a}^\tau \nabla_{\boldsymbol{\theta}_0}[G(\boldsymbol{\theta}_0^T Z_i)]X_i| \leqslant n^{-1/2}\|\mathbf{a}\|\|X_i\|\sum_{j=0}^p J_j(Z_i)$. By assumptions (I_3) and (I_6), $\|\mathbf{a}\|\|X_i\|\sum_{j=0}^p J_j(Z_i)$ is bounded in probability, for all i. Thus, with probability 1, we have

$$\frac{1}{\sqrt{n}}|\mathbf{a}^\tau \nabla_{\boldsymbol{\theta}_0}[G(\boldsymbol{\theta}_0^T Z_i)]X_i| \rightarrow 0 \quad as \;\; n \rightarrow \infty.$$

Set $\lambda_n = \left[\max_{1\leqslant i\leqslant n} n^{-1}\|X_i\|\sum_{j=0}^p J_j(Z_i)\right]^{1/2}\|\mathbf{a}\|$. Then $\lambda_n \rightarrow 0$ as $n \rightarrow \infty$, and, is independent of i. Since σ_n^2 converges to a positive quantity, the ratio $\sigma_n/\lambda_n \rightarrow \infty$ as $n \rightarrow \infty$. Now conditioning on Z_i and X_i, it is easy to see that

$$\sum_{i=1}^n E\Big[\frac{1}{n}(\mathbf{a}^\tau \nabla_{\boldsymbol{\theta}_0}[G(\boldsymbol{\theta}_0^T Z_i)]X_i)^2\varphi^2[F(\varepsilon_i)]I\Big(\Big|\frac{1}{\sqrt{n}}(\mathbf{a}^\tau \nabla_{\boldsymbol{\theta}_0}[G(\boldsymbol{\theta}_0^T Z_i)]X_i)\varphi[F(\varepsilon_i)]\Big| > \epsilon\sigma_n\Big)\Big]$$

is less than or equal to

$$E\Big[\varphi^2[F(\varepsilon)]I\Big(\big|\varphi[F(\varepsilon)]\big| > \epsilon\sigma_n/\lambda_n\Big)\Big] \times \frac{1}{n}\sum_{i=1}^{n} E(\mathbf{a}^\tau \nabla_{\boldsymbol{\theta}_0}[G(\boldsymbol{\theta}_0^T Z_i)]X_i)^2.$$

In this expression, the second term

$$\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n} E(\mathbf{a}^\tau \nabla_{\boldsymbol{\theta}_0}[G(\boldsymbol{\theta}_0^T Z_i)]X_i)^2 \ < \infty \quad \text{by } (I_3) \text{ and } (I_6).$$

From the boundedness of φ and applying the Dominated Convergence Theorem to the first term, we have

$$E\Big[\varphi^2[F(\varepsilon)]I\Big(\big|\varphi[F(\varepsilon)]\big| > \epsilon\sigma_n/\lambda_n\Big)\Big] \to 0 \quad as \quad n \to \infty.$$

This shows that the limit in (2.9) goes to zero as $n \to \infty$ and thus, the Central Limit Theorem gives $\sqrt{n}T_n(\boldsymbol{\theta}_0) \ \xrightarrow{\mathcal{D}}\ N(\mathbf{0}, \boldsymbol{\Sigma})$ as $n \to \infty$.

Next,

$$S_n(\boldsymbol{\theta}) - T_n(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n} \nabla_{\boldsymbol{\theta}_0}[G(\boldsymbol{\theta}_0^T Z_i)]X_i\left[\varphi\left(\frac{R(\eta_i(\boldsymbol{\theta}))}{n+1}\right) - \varphi(F_i(\eta_i(\boldsymbol{\theta})))\right].$$

By Cauchy-Schwartz inequality, we have

$$\begin{aligned}\|S_n(\boldsymbol{\theta}) - T_n(\boldsymbol{\theta})\| &\leqslant \left[\frac{1}{n}\sum_{i=1}^{n}\|\nabla_{\boldsymbol{\theta}_0}[G(\boldsymbol{\theta}_0^T Z_i)]X_i\|^2\right]^{1/2}\left[\frac{1}{n}\sum_{i=1}^{n}\left|\varphi\left(\frac{R(\eta_i(\boldsymbol{\theta}))}{n+1}\right) - \varphi(F_i(\eta_i(\boldsymbol{\theta})))\right|^2\right]^{1/2}\\ &\leqslant \left[\frac{1}{n}\sum_{i=1}^{n}\|X_i\|^2\left\{\sum_{j=0}^{p}J_j(Z_i)\right\}^2\right]^{1/2}\left[\max_{1\leqslant i\leqslant n}\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}}\left|\varphi\left(\frac{R(\eta_i(\boldsymbol{\theta}))}{n+1}\right) - \varphi(F_i(\eta_i(\boldsymbol{\theta})))\right|^2\right]^{1/2}\end{aligned}$$

Once again, by continuity of φ and the fact that $R(\eta_i(\boldsymbol{\theta}))/(n+1) - F_i(\eta_i(\boldsymbol{\theta})) \to 0$ *a.s.*, for all i and all $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, we have $\max_{1\leqslant i\leqslant n}\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}}\left|\varphi\left(\frac{R(\eta_i(\boldsymbol{\theta}))}{n+1}\right) - \varphi(F_i(\eta_i(\boldsymbol{\theta})))\right|^2 \to 0$ *a.s.* On the other hand, the strong law of large numbers gives

$$\frac{1}{n}\sum_{i=1}^{n}\|X_i\|^2\left\{\sum_{j=0}^{p}J_j(Z_i)\right\}^2 \to E\left[\|X\|^2\left\{\sum_{j=0}^{p}J_j(Z)\right\}^2\right] < \infty \ a.s.,$$

by assumption (I_3) and (I_5). Thus, $\lim_{n\to\infty}\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}}\|S_n(\boldsymbol{\theta}) - T_n(\boldsymbol{\theta})\| = 0 \ a.s.$ □

Proof of Theorem 2.3. a. Considering the definition of $T_n(\boldsymbol{\theta})$, we have

$$\nabla_{\boldsymbol{\theta}}T_n(\boldsymbol{\theta}_0) = -\frac{1}{n}\sum_{i=1}^{n}A_iA_i^{\tau}f(\varepsilon_i)\varphi'(F(\varepsilon_i)) + \frac{1}{n}\sum_{i=1}^{n}\nabla^2_{\boldsymbol{\theta}_0}[G(\boldsymbol{\theta}_0^T Z_i)]X_i\varphi(F(\varepsilon_i)).$$

The strong law of large numbers gives $n^{-1}\sum_{i=1}^{n}A_iA_i^{\tau}f(\varepsilon_i)\varphi'(F(\varepsilon_i)) \to E\{AA^{\tau}f(\varepsilon)\varphi'(F(\varepsilon))\}$ *a.s.*, and $n^{-1}\sum_{i=1}^{n}\nabla^2_{\boldsymbol{\theta}_0}[G(\boldsymbol{\theta}_0^T Z_i)]X_i\varphi(F(\varepsilon_i)) \to E\{\nabla^2_{\boldsymbol{\theta}_0}[G(\boldsymbol{\theta}_0^T Z)]X\varphi(F(\varepsilon))\}$ *a.s.* Thus,

$$\nabla_{\boldsymbol{\theta}}T_n(\boldsymbol{\theta}_0) \to \mathbf{W} = -E\{AA^{\tau}f(\varepsilon)\varphi'(F(\varepsilon))\} + E\{\nabla^2_{\boldsymbol{\theta}_0}[G(\boldsymbol{\theta}_0^T Z)]X\varphi(F(\varepsilon))\} \ a.s.$$

If we were to assume that ε is independent of (Z, X), we have

$$E\{AA^{\tau}f(\varepsilon)\varphi'(F(\varepsilon))\} = \boldsymbol{\Sigma} \times E\{f(\varepsilon)\varphi'(F(\varepsilon))\}.$$

But

$$E[f(\varepsilon)\varphi'(F(\varepsilon))] = \int_{-\infty}^{\infty} f(\varepsilon)\varphi'(F(\varepsilon))dF(\varepsilon) = -\int_{-\infty}^{\infty} f'(\varepsilon)\varphi(F(\varepsilon))d\varepsilon,$$

from the integration by parts, since $f(\varepsilon)\varphi(F(\varepsilon)) \to 0$ as $\varepsilon \to \pm\infty$. Now, putting $u = F(\varepsilon)$, we have

$$\int_{-\infty}^{\infty} f'(\varepsilon)\varphi(F(\varepsilon))d\varepsilon = -\int_{0}^{1}\varphi(u)\varphi_f(u)du = -\gamma_{\varphi}^{-1}.$$

Also, $E\{\nabla^2_{\boldsymbol{\theta}_0}[G(\boldsymbol{\theta}_0^T Z)]X\varphi(F(\varepsilon))\} = E\{\nabla^2_{\boldsymbol{\theta}_0}[G(\boldsymbol{\theta}_0^T Z)]X\}E\{\varphi(F(\varepsilon))\}$, and by assumption (I_1), $E\left[\varphi\big(F(\varepsilon)\big)\right] = \int_0^1 \varphi(t)dt = 0$. Thus, $\mathbf{W} = \gamma_\varphi^{-1}\boldsymbol{\Sigma}$.

b. Taking the second derivative of $T_n(\boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$, we have

$$\begin{aligned}\nabla^2_{\boldsymbol{\theta}} T_n(\boldsymbol{\xi}_n) &= -\frac{3}{n}\sum_{i=1}^n \nabla^3_{\boldsymbol{\xi}_n}[G(\boldsymbol{\xi}_n^T Z_i)]X_i f_i(\eta_i(\boldsymbol{\xi}_n))\varphi'(F_i(\eta_i(\boldsymbol{\xi}_n))) \\ &\quad + \frac{1}{n}\sum_{i=1}^n \left(\nabla_{\boldsymbol{\xi}_n}[G(\boldsymbol{\xi}_n^T Z_i)]X_i\right)^3 f_i'(\eta_i(\boldsymbol{\xi})_n)\varphi'(F_i(\eta_i(\boldsymbol{\xi}_n))) \\ &\quad + \frac{1}{n}\sum_{i=1}^n \left(\nabla_{\boldsymbol{\xi}_n}[G(\boldsymbol{\xi}_n^T Z_i)]X_i\right)^3 f_i^2(\eta_i(\boldsymbol{\xi}_n))\varphi''(F_i(\eta_i(\boldsymbol{\xi}_n))) \\ &\quad + \frac{1}{n}\sum_{i=1}^n \{\nabla^3_{\boldsymbol{\xi}_n}[G(\boldsymbol{\xi}_n^T Z_i)]X_i\varphi(F_i(\eta_i(\boldsymbol{\xi}_n))).\end{aligned}$$

From this, taking into account assumptions (I_3) and (I_5), it can be shown that each term to the right hand side of this equation converges almost surely to a finite quantity and therefore is almost surely bounded. Thus, $\nabla^2_{\boldsymbol{\theta}} T_n(\boldsymbol{\xi}_n)$ is almost surely bounded. □

Proof of Theorem 2.4. From equation (2.5) and the results of Theorem 2.3, we have

$$\sqrt{n}(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) = -\mathbf{W}^{-1}\sqrt{n}S_n(\boldsymbol{\theta}_0) + o_p(1).$$

Considering the fact that $\sqrt{n}T_n(\boldsymbol{\theta}_0) \xrightarrow{\mathcal{D}} N(\mathbf{0}, \boldsymbol{\Sigma})$ together with Slutsky Lemma, the result follows. □

Chapter 3

Rank Based Variable Selection for Single-index Varying Coefficient Models

3.1 Introduction

Suppose y_i is the response variable, $X = (\mathrm{x}_{0i}, \ldots, \mathrm{x}_{pi})^T$ with $\mathrm{x}_{0i} = 1$, and $Z = (z_{1i}, \ldots, z_{qi})^T$ are predictor variables. The single-index varying coefficient model (SIVCM) has the following form

$$y_i = \{G(\boldsymbol{\theta}_0^T Z_i)\}^T X_i + \varepsilon_i \quad i = 1, \ldots, n \tag{3.1}$$

where $\boldsymbol{\theta}_0$ is a $q-$vector of unknown regression parameters representing the single-index direction; $G(\cdot) = (g_0(\cdot), \ldots, g_p(\cdot))^T$ are unknown coefficient functions; and ε are random errors with finite Fisher information. For model identifiability, it is assumed that the $\|\boldsymbol{\theta}_0\| = 1$ and first component of $\boldsymbol{\theta}_0$ is positive.

Variable selection is an important topic in statistical modeling. It is common in practice that some variables we included in our model are redundant and they increase the model complexity without improving the accuracy of prediction. In linear regression setting, traditional variable selection criteria, such as Akaike information criterion (AIC) and Bayesian information criterion (BIC) for best subset variable selection have been extensively used in practice. However, these methods are unstable and they suffer expensive computationally cost (Breiman (1995); Tibshirani (1996); Fan & Li (2001)). Shrinkage methods such as the least absolute shrinkage and selection operator (LASSO) (Tibshirani (1996); Zou (2006)) and smoothly clipped absolute deviation (SCAD) (Fan & Li (2001)) have received much attention in last two decade.

Comparing to parametric models, applying shrinkage method to semiparametric model is much more challenging since it includes selection of significant variables in the nonparametric component, which involves other type of regularization parameters (i.e., the smoothing parameters).

Various works have been done to extend the shrinkage method to semiparametric models. Wang & Yin (2008) proposed sparse MAVE method, which combine the dimension reduction method MAVE (Xia *et al.* (2002)) with LASSO. Zeng *et al.* (2012) further explore the idea of combining MAVE and LASSO in SIM, and proposed to penalize the index vector $\boldsymbol{\theta}$ and the norm of the derivative of unknown function $g(\cdot)$ simultaneously. Wang & Xia (2009) extended LASSO to VCM with local constant kernel estimation. By combining local constant kernel estimation and SCAD penalty, Cai *et al.* (2015) proposed a two-stage approach to select coefficient functions and index parameters for time series SIVCM. Feng & Xue (2015) proposed to penalize not only the coefficient functions but also their derivatives by using SCAD penalty to detect zero and constant functions, they also penalized index vector using SCAD to select index parameters. All the above methods are based on least squares method, which is sensitive to outliers. Feng *et al.* (2015) developed a robust variable selection method for VCM by combining a rank-based spline loss function and SCAD penalty. Song *et al.* (2016) used the exponential squared loss with SCAD penalty to perform robust variable selection for SIVCM. However, in addition to three tuning parameters for controlling smoothness, coefficient functions and index parameter selection, exponential square loss comes with an extra tuning parameter that controls the degree of robustness and efficiency of their proposed estimators. The cross validation approach proposed by them could be computational demanding in practice.

To the best of our knowledge, all existing variable selection methods for SIVCM use SCAD penalty. LASSO has been used widely in practice, especially after an efficient algorithm was developed for computing its entire solution path (Osborne *et al.* (2000) and

Efron *et al.* (2004)). It is worthwhile to develop a LASSO based variable selection procedure for SIVCM. We propose to combine rank-based spline estimation and group LASSO (RS-GLASSO) to select coefficient functions and use R estimation procedure to estimate index parameters. We refer the LS version of our method as LSSGLASSO.

3.2 Rank Based Variable Selection

Model (3.1) is a semiparametric model. The parametric estimators $\widehat{\boldsymbol{\theta}}$ has faster convergence rate than nonparametric estimators $\widehat{G}(\cdot)$. It is common to use a backfitting approach Fan *et al.* (2003) to estimate $\boldsymbol{\theta}_0$ and $G(\cdot)$. We propose a rank-based (R) procedure to select $G(\cdot)$ and estimate $\boldsymbol{\theta}_0$ in two stages. In stage one, for given $\boldsymbol{\theta}$, we replace $g(\cdot)$ by its basis function approximation and reformulate (3.1) as linear model. Since each function is a linear combination of basis functions, applying group LASSO (Yuan & Lin (2006)) with general rank loss function to the reformulated model achieves robust function selection in (3.1). Although kernel smoothing can also be used for functions approximation, using basis function expansion is more straightforward in selecting coefficient functions. In stage two, we exclude the coefficient functions that are not selected in step one and use R estimation procedure to estimate $\boldsymbol{\theta}_0$ and $G(\cdot)$. It is worth pointing out that we use local rank estimation in step two and re-estimate the selected coefficient functions we obtain from step one. Since B-spline smoothing suffers from boundary effects (Hastie *et al.* (2001)), re-estimating coefficient functions can help us improve the estimation near boundaries.

3.2.1 Variable Selection for Coefficient Functions

Suppose that $\{(X_i, Z_i, y_i),\ i = 1, \ldots, n\}$ is a random sample from model (3.1). Let $B(\cdot) = (B_1(\cdot), \ldots, B_L(\cdot))^T$ be the B-spline basis functions with a fixed degree and knot sequence, $g(\cdot)_k$ can be approximated by

$$g_k(\boldsymbol{\theta}^T Z_i) \approx \sum_{j=1}^{L} B_j(\boldsymbol{\theta}^T Z_i)\gamma_{kj} = \{B(\boldsymbol{\theta}^T Z_i)\}^T \gamma_k, \quad k = 0, \ldots, p.$$

Model (3.1) can be written as

$$y_i \approx \{V_i(\boldsymbol{\theta})\}^T \boldsymbol{\gamma} + \boldsymbol{\varepsilon}_i \tag{3.2}$$

where $\boldsymbol{\gamma} = (\gamma_0, \ldots, \gamma_p)^T$, $V_i(\boldsymbol{\theta}) = I_{p+1} \otimes B(\boldsymbol{\theta}^T Z_i) \cdot X_i$.

Define the residuals as $e_i(\boldsymbol{\theta}, \gamma) = y_i - \{G(\boldsymbol{\theta}^T Z_i)\}^T X_i = y_i - \{V_i(\boldsymbol{\theta})\}^T \boldsymbol{\gamma}$, and we define the local rank objective function to be

$$L_n(\boldsymbol{\theta}, \boldsymbol{\gamma}) = \frac{1}{n} \sum_{i=1}^{n} \varphi\Big(\frac{R(e_i(\boldsymbol{\theta}, \boldsymbol{\gamma}))}{n+1}\Big) e_i(\boldsymbol{\theta}, \boldsymbol{\gamma}) \tag{3.3}$$

where $R(e_i(\boldsymbol{\theta}, \boldsymbol{\gamma}))$ is the rank of $e_i(\boldsymbol{\theta}, \boldsymbol{\gamma})$ among $e_1(\boldsymbol{\theta}, \boldsymbol{\gamma}), \ldots, e_n(\boldsymbol{\theta}, \boldsymbol{\gamma})$, and φ is a general bounded nondecreasing score function defined on $(0, 1)$. To select the coefficient functions robustly, we combine (3.3) with group LASSO penalty and achieve variable selection for coefficient functions by solving the following minimization problem,

$$\min_{\boldsymbol{\theta}, \boldsymbol{\gamma}} \frac{1}{n} \sum_{i=1}^{n} \varphi\Big(\frac{R(e_i(\boldsymbol{\theta}, \boldsymbol{\gamma}))}{n+1}\Big) e_i(\boldsymbol{\theta}, \boldsymbol{\gamma}) + \lambda \sum_{j=1}^{p} \|\gamma_j\| \tag{3.4}$$

Note that we do not penalize $g_0(\cdot)$ since it is the intercept coefficient function and it always stays in the model.

3.2.2 Estimation for Index Parameter

Suppose that $\{(X_i, Z_i, y_i),\ i = 1, \ldots, n\}$ is a random sample from model (3.1). For Z_i in a neighborhood of any given z, we can locally approximate the coefficient function using Taylor expansion $g_k(\boldsymbol{\theta}^T Z_i) \approx g_k(\boldsymbol{\theta}^T Z_i) + g'_k(\boldsymbol{\theta}^T Z_i)\boldsymbol{\theta}^T Z_{i0}$, $k = 0, \ldots, p$, where $Z_{i0} = Z_i - z$.

Let $G(\boldsymbol{\theta}^T z_0) = (g_0(\boldsymbol{\theta}^T z), \ldots, g_p(\boldsymbol{\theta}^T z))^T$ and $G'(\boldsymbol{\theta}^T z) = (g_0'(\boldsymbol{\theta}^T z), \ldots, g_p'(\boldsymbol{\theta}^T z_0))^T$. Denote $a = G(\boldsymbol{\theta}^T z)$ and $b = G'(\boldsymbol{\theta}^T z)$.

For Z_i close to z, we define the residual as $e_i(\boldsymbol{\theta}, a, b) = y_i - X_i^T a - X_i^T b Z_{i0}^T \boldsymbol{\theta}$. The local rank objective function is

$$L_n(\boldsymbol{\theta}, a_j, b_j) = \frac{1}{n(n-1)} \sum_{j=1}^{n} \sum_{i=1}^{n} \varphi\Big(\frac{R(e_{ij}(\boldsymbol{\theta}, a_j, b_j))}{n^2+1}\Big) e_{ij}(\boldsymbol{\theta}, a_j, b_j) w_{ij} \tag{3.5}$$

where $R(e_i(\boldsymbol{\theta}, a, b))$ is the rank of $e_i(\boldsymbol{\theta}, a, b)$ among $e_1(\boldsymbol{\theta}, a, b), \ldots, e_n(\boldsymbol{\theta}, a, b)$, φ is a general bounded nondecreasing score function defined on $(0,1)$, $w_{ij} = K_h(\boldsymbol{\theta}^T Z_{ij}) / \sum_{j=1}^{n} K_h(\boldsymbol{\theta}^T Z_{ij})$ and $Z_{ij} = Z_i - Z_j$. We can get the R estimator for $\boldsymbol{\theta}_0$, a and b by minimizing Equation 3.5.

3.2.3 Computational Algorithm

Here, we provide a detailed computational algorithm to implement the estimation procedure in subsections 3.2.1 and 3.2.2. The algorithm contains two stages. In **Stage 1** (**Step 0** - **Step 2**), we select $g_j(\cdot)$, $j = 1, \ldots, p$. In **Stage 2** (**Step 3** - **Step 6**), we use R estimation procedure to estimate $\boldsymbol{\theta}_0$. Re-estimate $g_0(\cdot)$ and the non-zero $g_j(\cdot)$, $j \in \{1, \ldots, p\}$ selected in **Stage 1** using local rank estimation method.

Step 0: (Initialization): Input data $\{(X_i, Z_i, y_i), i = 1, \ldots, n\}$. Use sliced inverse regression (SIR) to get initial value of $\boldsymbol{\theta}_0$ with $\|\boldsymbol{\theta}_0\| = 1$, and denote the initial estimate as $\widetilde{\boldsymbol{\theta}}^{(0)}$.

Step 1: Given $\widetilde{\boldsymbol{\theta}}^{(0)}$, estimate $\boldsymbol{\gamma}$ by

$$\underset{\boldsymbol{\gamma}}{\text{Argmin}} \frac{1}{n} \sum_{i=1}^{n} \varphi\Big(\frac{R(e_i(\boldsymbol{\gamma}))}{n+1}\Big) e_i(\boldsymbol{\gamma}) + \lambda \sum_{j=1}^{p} \|\gamma_j\|$$

where $e_i(\gamma) = y_i - \{V_i(\widetilde{\boldsymbol{\theta}}^{(0)}\}^T \gamma$

Step 2: If $g_j(\cdot)$, $j = 1, \ldots, p$ is not selected, delete j_{th} column of X_i and denote it as X_i^{glasso}.

Step 3: Input data $\{(X_i^{glasso}, Z_i, y_i),\ i = 1, \ldots, n\}$. Use SIR to get initial value of $\boldsymbol{\theta}$ with $\|\boldsymbol{\theta}\| = 1$, and denote the initial estimate as $\widetilde{\boldsymbol{\theta}}^{(1)}$.

Step 4: Given $\widetilde{\boldsymbol{\theta}}^{(1)}$ and for each fixed j, estimate $g_j(\cdot) = a_j$ and $g_j'(\cdot) = b_j$ by

$$\operatorname*{Argmin}_{a_j, b_j} \frac{1}{n} \sum_{j=1}^{n} \varphi\left(\frac{R(e_i(a_j, b_j))}{n+1}\right) e_i(a_j, b_j) w_{ij}$$

where $e_i(a_j, b_j) = y_i - X_i^T a_j - {X_i}^T b_j Z_{ij}^T \widetilde{\boldsymbol{\theta}}^{(1)}$. Denote the estimates of a_j and b_j as $\widehat{a}_j$ and $\widehat{b}_j$.

Step 5: Given $\widehat{a}_j$ and $\widehat{b}_j$, estimate $\boldsymbol{\theta}_0$ by

$$\operatorname*{Argmin}_{\boldsymbol{\theta}} \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j=1}^{n} \varphi\left(\frac{R(e_{ij}(\boldsymbol{\theta}))}{n^2+1}\right) e_{ij}(\theta) w_{ij}$$

where $e_{ij}(\boldsymbol{\theta}) = (y_i - \widehat{a}_j^T X_i) - X_i^T \widehat{b}_j Z_{ij}^T \boldsymbol{\theta}$. Denote the estimate of $\boldsymbol{\theta}$ as $\widehat{\boldsymbol{\theta}}$. Update $\widetilde{\boldsymbol{\theta}}^{(0)}$ to $\widehat{\boldsymbol{\theta}}$.

Step 6: Repeat **Step 1-Step 6** until two successive values of $\widehat{\boldsymbol{\theta}}$ differ insignificantly.

3.2.4 Tuning Parameter Selection for Group LASSO

Tuning parameter, or regularization parameter, plays a critical role in shrinkage and variable selection methods. Generalized cross validation (GCV) has been extensively used to choose tuning parameters. However, Wang *et al.* (2007) showed that optimal tuning parameter chosen by GCV tends to produce overfitted models. And they proposed a BIC-type selection criterion. We choose the tuning parameter for group LASSO using a BIC-type selection criterion defined as

$$BIC_\lambda = \log\Big(\frac{1}{n}\sum_{i=1}^{n} \varphi\Big(\frac{R(\widehat{e}_i(\boldsymbol{\gamma}_\lambda))}{n+1}\Big)\widehat{e}_i(\boldsymbol{\gamma}_\lambda)\Big) + df_\lambda \frac{log(n)}{n}, \tag{3.6}$$

where $\widehat{e}_i(\boldsymbol{\gamma}_\lambda) = y_i - \{V_i(\widehat{\boldsymbol{\theta}})\}^T \boldsymbol{\gamma}_\lambda$ and df_λ is the number of nonzero coefficient functions. The optimal tuning parameter can be obtained as

$$\widehat{\lambda} = \underset{\lambda}{\operatorname{Argmin}}\, BIC_\lambda.$$

3.2.5 Smoothing Parameter Selection for Function Estimation

Smoothing parameters such as number of knots in basis expansion and bandwidth in local linear kernel estimation play an important role in nonparametric smoothing. In stage 1 of our proposed procedure, we use B-spline to approximate the coefficient functions. In general, the number of knots can be chosen by GCV, BIC or some type of cross-validation methods. In our simulation, we set the number of knots to be a fixed number for two reasons. First, choosing number of knots are computationally expensive and it is not easy to decide the range of number for the knots to be chosen. Secondly, the estimation of our coefficient functions does not rely on the number of knots since we use local linear kernel estimation to re-estimate the selected coefficient functions in stage 2. We only use basis expansion for the purpose of selecting the coefficient functions and we find using different number of knots do not affect the results of selection significantly.

Following the same procedure in subsection 2.3.1, we use cross-validation method to choose the optimal bandwidth for local linear kernel smoothing.

3.3 Simulation

To assess the performance of rank-based variable selection method, we conduct a finite sample Monte Carlo simulations. We use the Gaussian kernel in our calculation. The algorithm that was described in subsection 3.2.3 is used to perform simultaneous variable selection and estimation for coefficient functions and estimation for index parameters. We stop the iteration when either the two successive values of $\boldsymbol{\theta}$ differ less than 0.001 or the number of iteration exceeds 30. All computations in this section were performed using the R software environment. Our method in subsection 3.2.1 were implemented using the "grpLasso" R package of Meier *et al.* (2008).

Example 1. Consider the model

$$y_i = (1 + 3u_i^2) + 3\exp(-u_i^2)\mathrm{x}_{1i} + 1.5\sin(\pi u_i)\mathrm{x}_{2i} + 0.8(u_i)\mathrm{x}_{3i} + \varepsilon_i. \tag{3.7}$$

where, for $i = 1, \dots n$, $g_0(u_i) = 1 + 3(u_i)^2$, $g_1(u_i) = 3\exp\{-(u_i)^2\}$, $g_2(u_i) = 1.5\sin(\pi u_i)$, $g_3(u_i) = 0.8(u_i)$ and $g_4(u_i) = \cdots = g_7(u_i) = 0$. u_i is generated from Uniform$(-1, 1)$. $X_i = (\mathrm{x}_1, \dots, \mathrm{x}_7)^T$ follow the multivariate normal distribution $N(0, \Sigma)$ with mean 0 and $Cov(Z_k, Z_l) = 0.5^{|k-l|}$. Three different model error (ε) distributions are considered: the standard normal distribution ($N(0, 1)$); the t-distribution with 3 degrees of freedom (t_3); the contaminate normal distribution ($\mathcal{CN}(0.95)$) with contamination rate 0.05, given as $\mathcal{CN}(0.95) = 0.95N(0, 1) + 0.05N(0, 10^2)$. We simulate 500 samples for sample size of 200 and 400, respectively. We report the true positive rates (TPR), the false positive rates (FPR), the percentage of correct models identified, model size and oracle values for all criteria to assess the proposed performance of variable selection procedure for coefficient functions. When it comes to assessing the individual and overall performance of the estimator of $g_j(\cdot)$, $j = 0, \dots, 7$, we use mean absolute deviation of each coefficient function MAD$_j$ and mean absolute deviation of all estimated coefficient functions MAD defined in Section 2.3.1. Model (3.7) is a VCM, which is a special case of SIVCM. We use (3.7) to evaluate the methods proposed in section 3.2.1, which correspond to **Stage 1** in section subsection 3.2.3.

From Table 3.1 and Table 3.2, we can observe that the performance of the RSGLASSO is similar to LSSGLASSO for normal error, but much better than the LSSGLASSO for both

t_3 and $\mathcal{CN}(0.95)$. RSGLASSO always gives higher percentage of correct fit and even outperforms LSSGLASSO under normal error for large sample. Figure 3.1- Figure 3.6 plot the estimated coefficient functions (g_0-g_7) for 500 simulations when $n = 400$. We can clearly observe that RSGLASSO estimator has smaller variance and outperforms LSSGLASSO in terms of selection when error distribution are not normal and the performance for RSGLASSO and LSSGLASSO are almost the same when we have normal error distribution.

Table 3.1: The simulation results are based on 500 runs. TPR is the average true positive rate; FPR is the average false positive rate; correct fit % is the proportion of times the correct model is selected; and model size is the average number of nonzero functions in the model.

ε	n	Method	TPR	FPR	Correct Fit (%)	Model Size
$N(0,1)$	200	LS	0.979	0.058	0.766	4.170
		R	0.977	0.053	0.772	4.144
	400	LS	1.000	0.026	0.906	4.104
		R	0.999	0.002	0.986	4.006
		Oracle	1.000	0.000	1.000	4.000
t_3	200	LS	0.841	0.037	0.476	3.670
		R	0.857	0.015	0.562	3.630
	400	LS	0.968	0.037	0.790	4.054
		R	0.969	0.001	0.904	3.912
		Oracle	1.000	0.000	1.000	4.000
$\mathcal{CN}(0.95)$	200	LS	0.643	0.025	0.218	3.030
		R	0.692	0.000	0.340	3.078
	400	LS	0.859	0.032	0.496	3.706
		R	0.981	0.000	0.942	3.942
		Oracle	1.000	0.000	1.000	4.000

Table 3.2: Mean absolute deviations of the coefficient functions and the overall mean absolute deviation (MAD).

ε	n	Method	g_0	g_1	g_2	g_3	g_4	g_5	g_6	g_7	MAD
$N(0,1)$	200	LS	0.973	0.672	1.131	0.541	0.027	0.007	0.003	0.002	0.419
		R	0.995	0.663	1.126	0.542	0.024	0.007	0.002	0.001	0.420
	400	LS	0.977	0.667	1.127	0.538	0.012	0.001	0.001	0.000	0.415
		R	0.986	0.660	1.114	0.539	0.001	0.000	0.000	0.000	0.413
t_3	200	LS	0.978	0.701	1.162	0.515	0.027	0.008	0.002	0.002	0.424
		R	1.002	0.683	1.144	0.510	0.011	0.000	0.000	0.000	0.419
	400	LS	0.983	0.685	1.141	0.544	0.025	0.004	0.002	0.001	0.423
		R	0.987	0.666	1.123	0.533	0.001	0.000	0.000	0.000	0.414
$\mathcal{CN}(0.95)$	200	LS	0.986	0.811	1.154	0.475	0.015	0.002	0.000	0.000	0.430
		R	1.031	0.741	1.132	0.461	0.000	0.000	0.000	0.000	0.421
	400	LS	0.985	0.692	1.175	0.520	0.027	0.003	0.000	0.000	0.425
		R	0.995	0.661	1.121	0.540	0.000	0.000	0.000	0.000	0.415

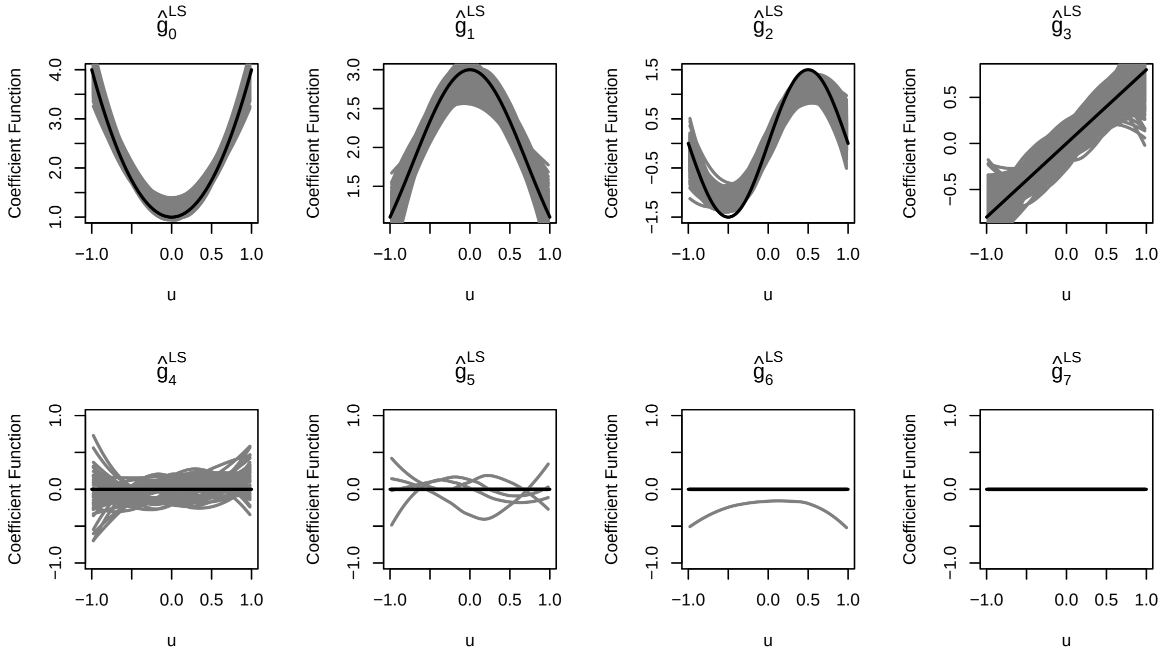

Figure 3.1: Coefficient functions estimated by LSSGLASSO under standard normal error distribution

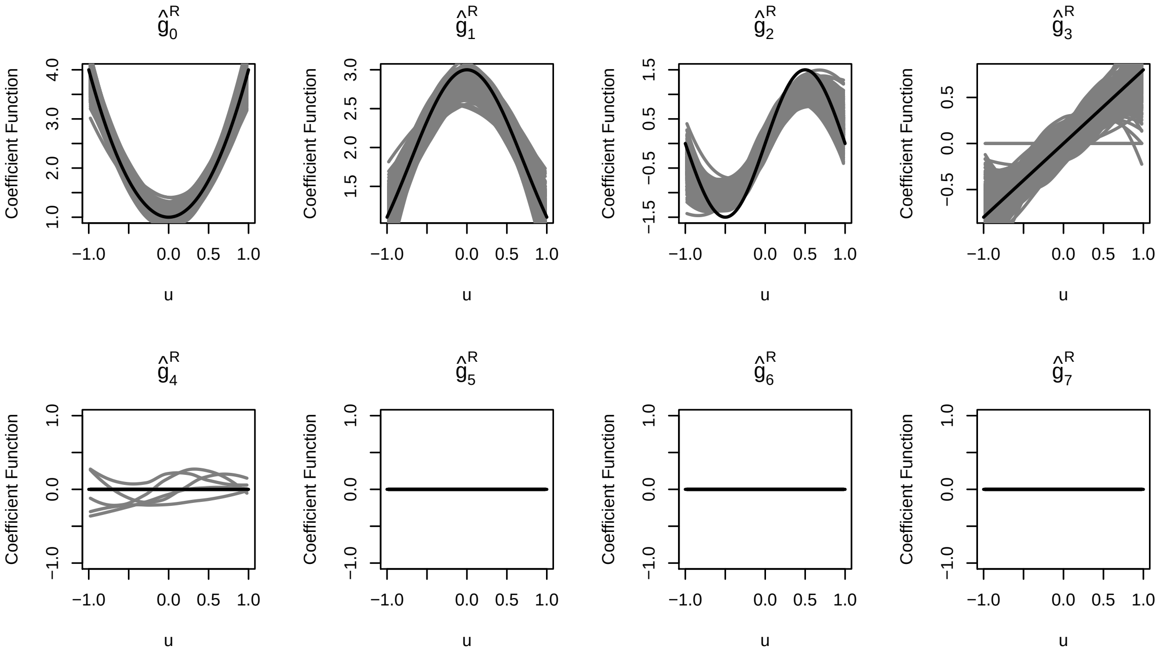

Figure 3.2: Coefficient functions estimated by RSGLASSO under standard normal error distribution

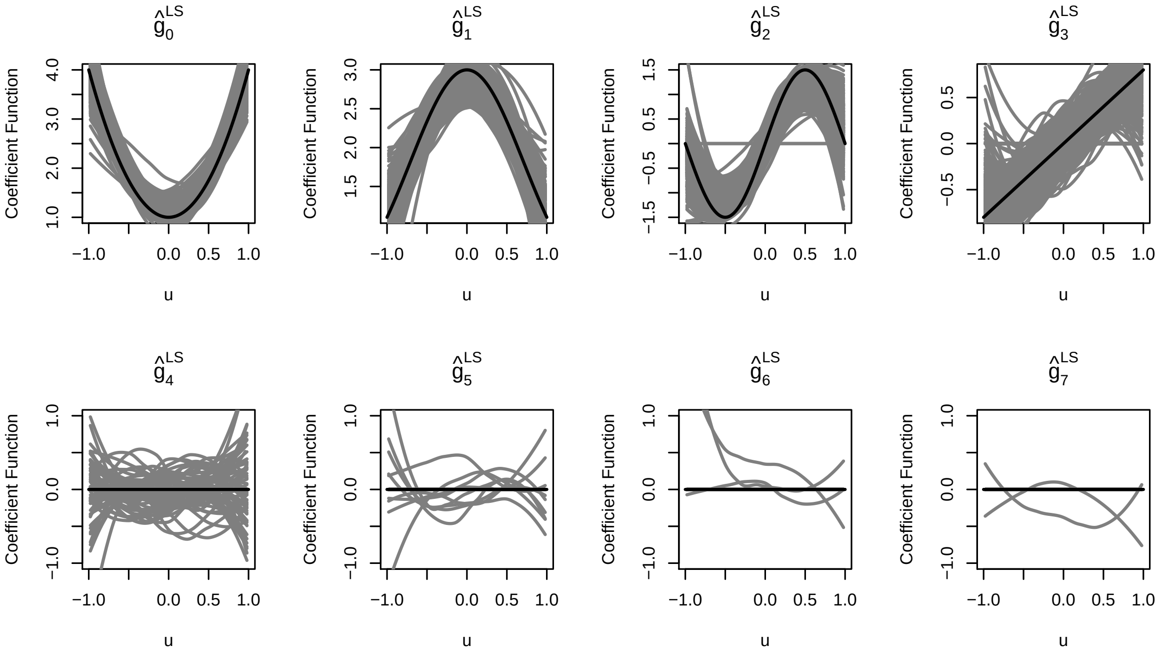

Figure 3.3: Coefficient functions estimated by LSSGLASSO under t_3 error distribution

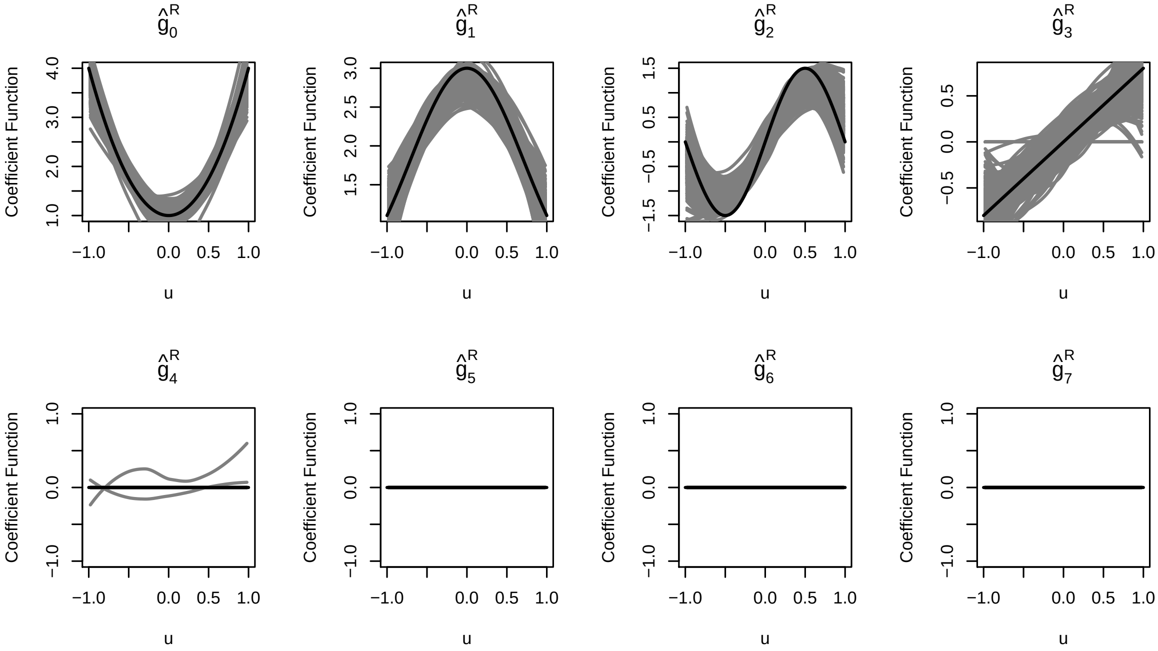

Figure 3.4: Coefficient functions estimated by RSGLASSO under t_3 error distribution

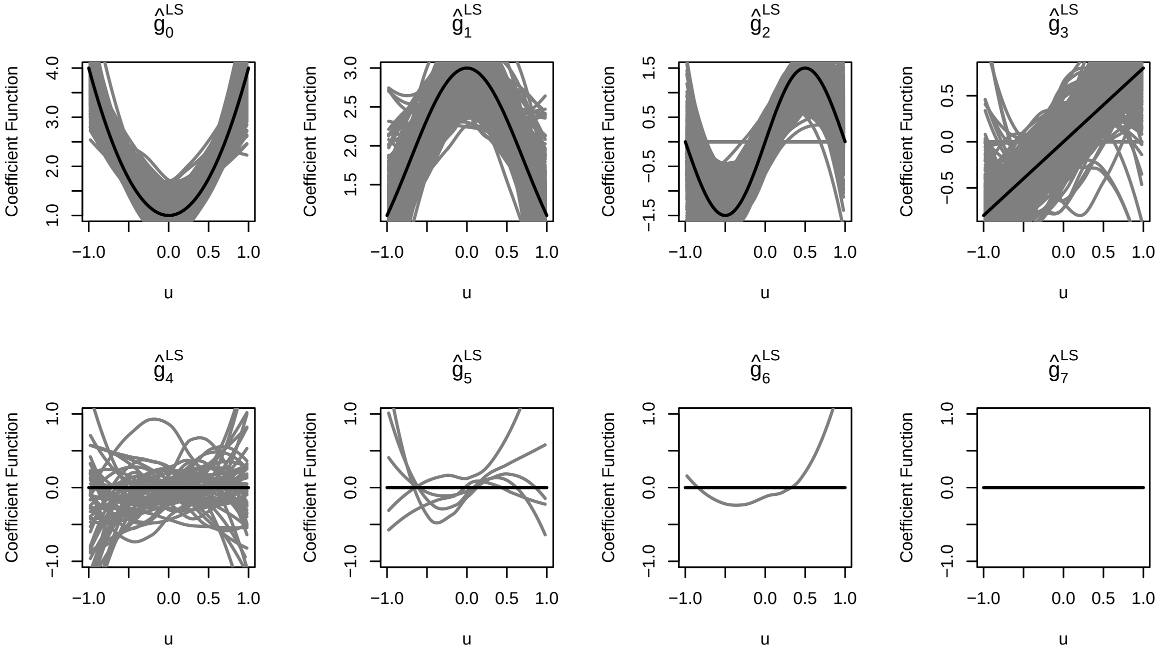

Figure 3.5: Coefficient functions estimated by LSSGLASSO under $\mathcal{CN}(0.95)$ error distribution

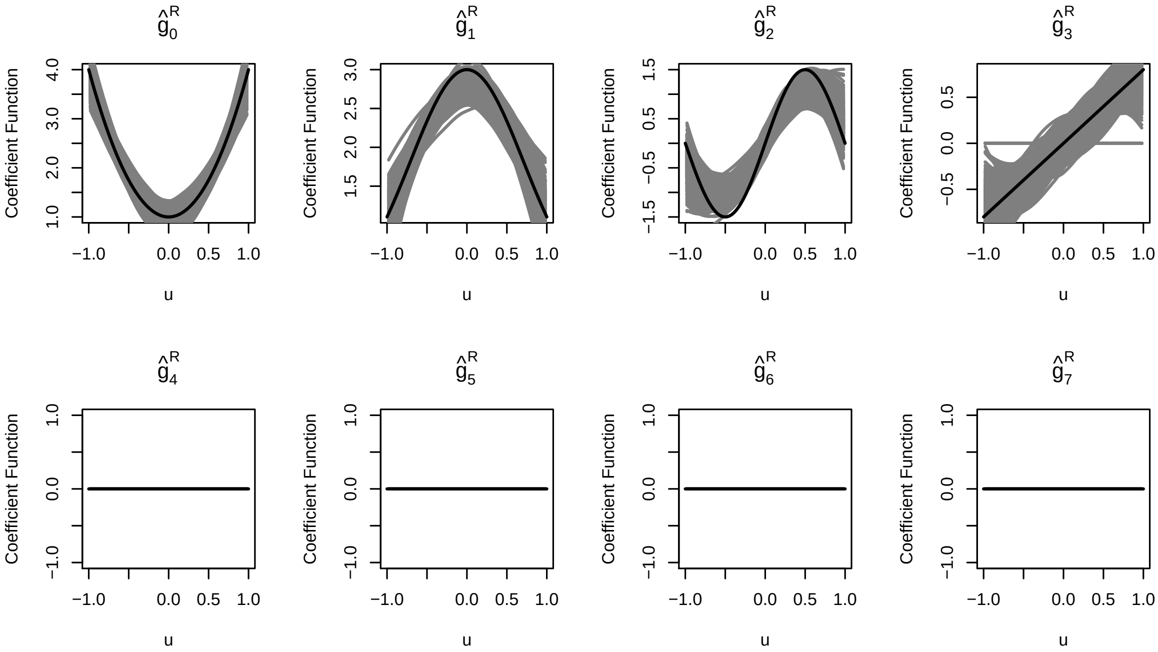

Figure 3.6: Coefficient functions estimated by RSGLASSO under $\mathcal{CN}(0.95)$ error distribution

Example 2. Consider the model

$$y_i = \{1 + 3(\boldsymbol{\theta}_0^T Z_i)^2)\} + 3\exp\{-(\boldsymbol{\theta}_0^T Z_i)^2\}\mathrm{x}_{1i} + 1.5\sin(\pi\boldsymbol{\theta}_0^T Z_i)\mathrm{x}_{2i} + 0.8(\boldsymbol{\theta}_0^T Z_i)\mathrm{x}_{3i} + \varepsilon_i. \quad (3.8)$$

where, for $i = 1, \dots n$, $g_0(\boldsymbol{\theta}_0^T Z_i) = 1+3(\boldsymbol{\theta}_0^T Z_i)^2)$, $g_1(\boldsymbol{\theta}_0^T Z_i) = 3\exp\{-(\boldsymbol{\theta}_0^T Z_i)^2\}$, $g_2(\boldsymbol{\theta}_0^T Z_i) = 1.5\sin(\pi\boldsymbol{\theta}_0^T Z_i)$, $g_3(\boldsymbol{\theta}_0^T Z_i) = 0.8(\boldsymbol{\theta}_0^T Z_i)$ and $g_4(\boldsymbol{\theta}_0^T Z_i) = \cdots = g_7(\boldsymbol{\theta}_0^T Z_i) = 0$. $Z_i = (\mathrm{z}_1, \dots, \mathrm{z}_7)^T$ are independent random vectors with each component uniformly distributed on $(-1, 1)$. $X_i = (\mathrm{x}_1, \dots, \mathrm{x}_7)^T$ follow the multivariate normal distribution $N(0, \Sigma)$ with mean 0 and $Cov(Z_k, Z_l) = 0.5^{|k-l|}$, and $\boldsymbol{\theta} = (1/3, 2/3, 0, 2/3)^T$. Three different model error (ε) distributions are considered: the standard normal distribution ($N(0, 1)$); the t-distribution with 3 degrees of freedom (t_3); the contaminate normal distribution (CN) with contamination rate 0.05, given as $CN(0.95) = 0.95N(0, 1) + 0.05N(0, 10^2)$. We simulate 200 samples for sample size to 200 and 400, respectively. We use the same criteria used in example 1 to assess the proposed performance of variable selection procedure for coefficient functions. Following Zeng *et al.* (2012), the performance of estimation for $\boldsymbol{\theta}_0$ is assessed by the angle (in degree) between $\widehat{\boldsymbol{\theta}}$ and $\boldsymbol{\theta}_0$. The angle is defined as $A(\widehat{\boldsymbol{\theta}}, \boldsymbol{\theta}_0) = (180/\pi)\arccos|\widehat{\boldsymbol{\theta}}^T\boldsymbol{\theta}_0|$. $A(\widehat{\boldsymbol{\theta}}, \boldsymbol{\theta}_0) \in [0, 90]$, with small values indicating good performance. We use MAD_j defined in example 1 to assess performance of function estimation. Table 3.3 report the performance of variable selection of the coefficient functions. Table 3.4 and Table 3.5 summarized the performance of estimation for $\boldsymbol{\theta}$ and $G(\cdot)$ and we can see RSGLASSO performs almost as well as LSSGLASSO when the error distribution are normal. For other error distributions RSGLASSO outperforms LSSGLASSO. Figure 3.7 - Figure 3.12 plot the estimated coefficient functions (g_0-g_7) for 200 simulations when $n = 400$ and the results are consistent with those in Table 3.4 and Table 3.5.

Table 3.3: The simulation results are based on 200 runs. TPR is the average true positive rate; FPR is the average false positive rate; correct fit % is the proportion of times the correct model is selected; and model size is the average number of nonzero functions in the model.

ε	n	Method	TPR	FPR	Correct Fit (%)	Model Size
$N(0,1)$	200	LS	0.967	0.062	0.680	4.150
		R	0.963	0.061	0.680	4.135
	400	LS	1.000	0.044	0.840	4.175
		R	1.000	0.044	0.835	4.175
		Oracle	1.000	0.000	1.000	4.000
t_3	200	LS	0.827	0.029	0.470	3.595
		R	0.857	0.012	0.580	3.620
	400	LS	0.953	0.045	0.750	4.040
		R	0.968	0.016	0.860	3.970
		Oracle	1.000	0.000	1.000	4.000
$\mathcal{CN}(0.95)$	200	LS	0.613	0.044	0.150	3.015
		R	0.695	0.009	0.315	3.120
	400	LS	0.827	0.019	0.460	3.555
		R	0.905	0.000	0.725	3.715
		Oracle	1.000	0.000	1.000	4.000

3.4 Discussion

We propose RSGLASSO procedure that is robust and efficient comparing to LSS-GLASSO in both selecting and estimating the coefficient functions. Our method includes Feng *et al.* (2015)'s rank-based spline SCAD (RSSCAD) as a special case. Comparing to RSSCAD, our one-step variable selection method do not need iteration and we can estimate $\boldsymbol{\theta}_0$ for SIVCM at the same time. Therefore, our method is more general. It is worth pointing out that the performance of the proposed method can be affected by high leverage points (outliers in the design space). The proposed method do not perform variable selection for index parameters and that needs to be further studied.

Table 3.4: The mean and standard deviation of the $A(\widehat{\boldsymbol{\theta}}, \boldsymbol{\theta}_0)$. The simulation results are based on 200 runs.

ε	n	Method	Mean	Std. dev
$N(0,1)$	200	LS	2.646	1.189
		R	3.022	1.363
	400	LS	1.707	0.721
		R	1.989	0.816
t_3	200	LS	5.176	6.174
		R	3.827	1.855
	400	LS	3.028	2.488
		R	2.343	1.012
$\mathcal{CN}(0.95)$	200	LS	12.237	17.725
		R	4.612	8.147
	400	LS	4.733	4.476
		R	2.257	0.975

Table 3.5: Mean absolute deviations of the coefficient functions and the overall mean absolute deviation (MAD).

ε	n	Method	g_0	g_1	g_2	g_3	g_4	g_5	g_6	g_7	MAD
$N(0,1)$	200	LS	1.193	0.733	1.126	0.530	0.041	0.002	0.002	0.001	0.453
		R	1.210	0.730	1.114	0.531	0.042	0.004	0.000	0.000	0.454
	400	LS	1.211	0.740	1.134	0.538	0.021	0.002	0.000	0.000	0.456
		R	1.221	0.737	1.119	0.539	0.024	0.001	0.000	0.000	0.455
t_3	200	LS	1.189	0.762	1.149	0.509	0.021	0.002	0.000	0.000	0.454
		R	1.206	0.743	1.125	0.496	0.008	0.000	0.000	0.000	0.447
	400	LS	1.212	0.748	1.150	0.535	0.028	0.009	0.006	0.004	0.462
		R	1.219	0.734	1.119	0.526	0.009	0.002	0.000	0.000	0.451
$\mathcal{CN}(0.95)$	200	LS	1.223	0.862	1.132	0.441	0.022	0.007	0.010	0.008	0.463
		R	1.249	0.786	1.109	0.444	0.004	0.001	0.001	0.001	0.449
	400	LS	1.229	0.764	1.160	0.495	0.018	0.001	0.000	0.000	0.459
		R	1.226	0.733	1.118	0.498	0.000	0.000	0.000	0.000	0.447

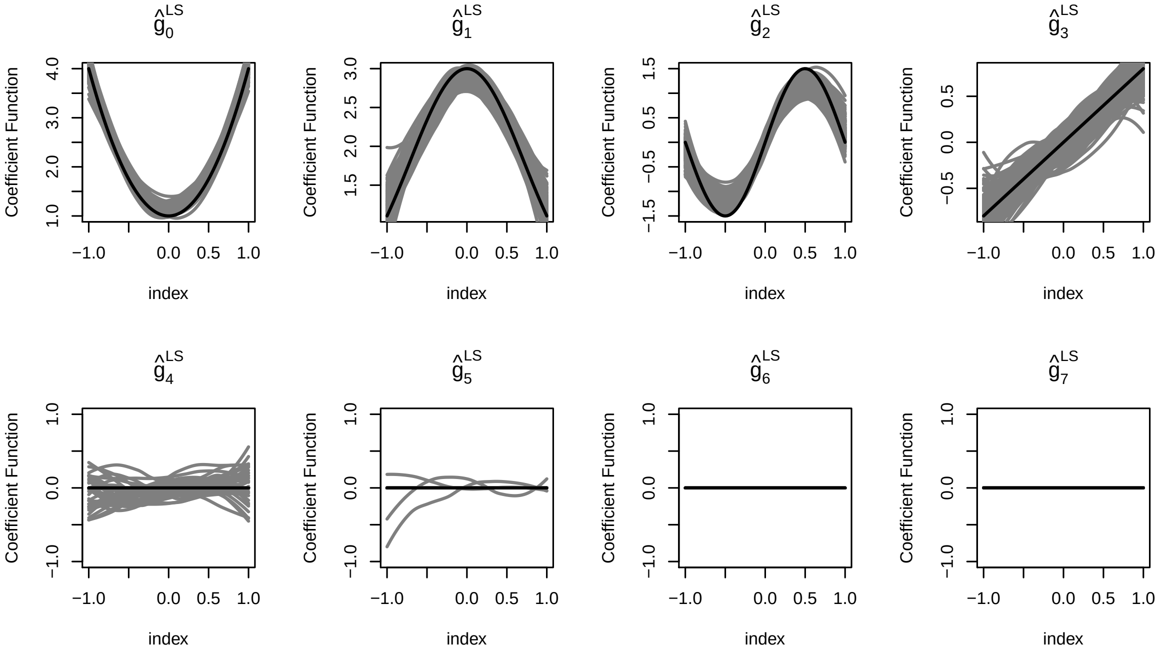

Figure 3.7: Coefficient functions estimated by LSSGLASSO under standard normal error distribution

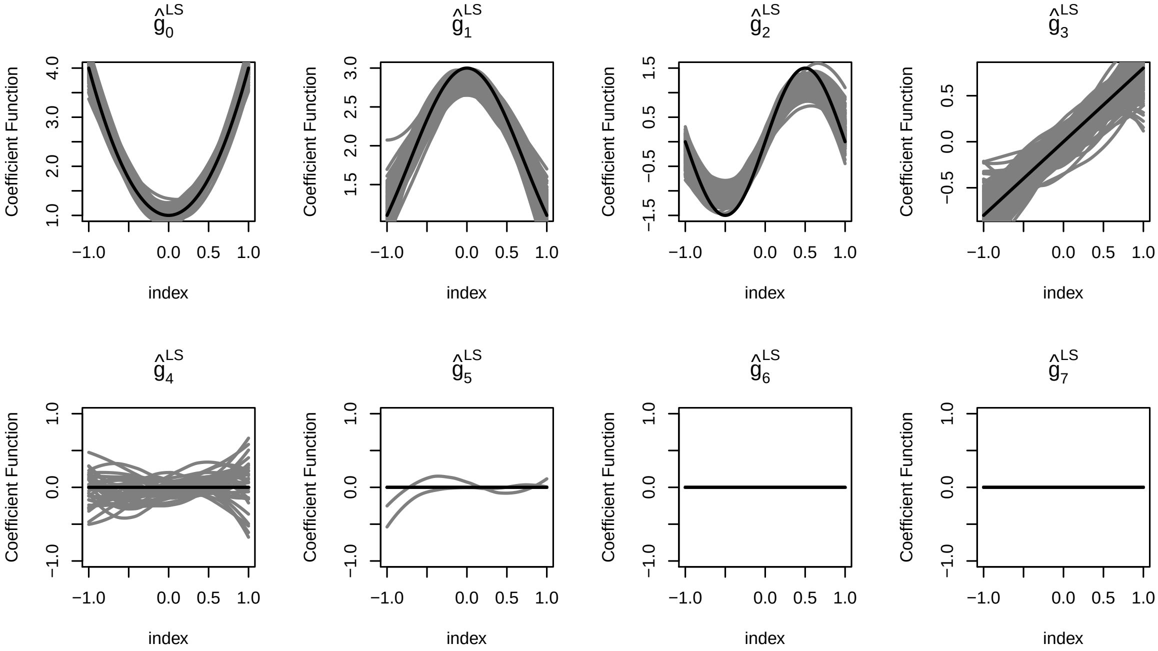

Figure 3.8: Coefficient functions estimated by RSGLASSO under standard normal error distribution

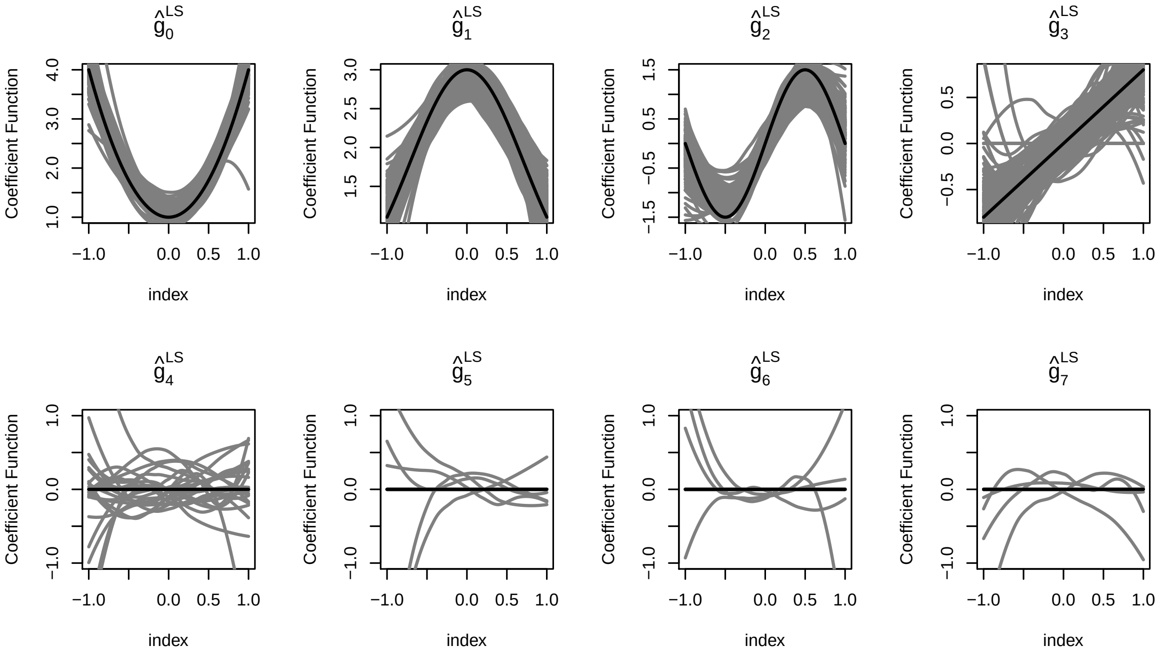

Figure 3.9: Coefficient functions estimated by LSSGLASSO under t_3 error distribution

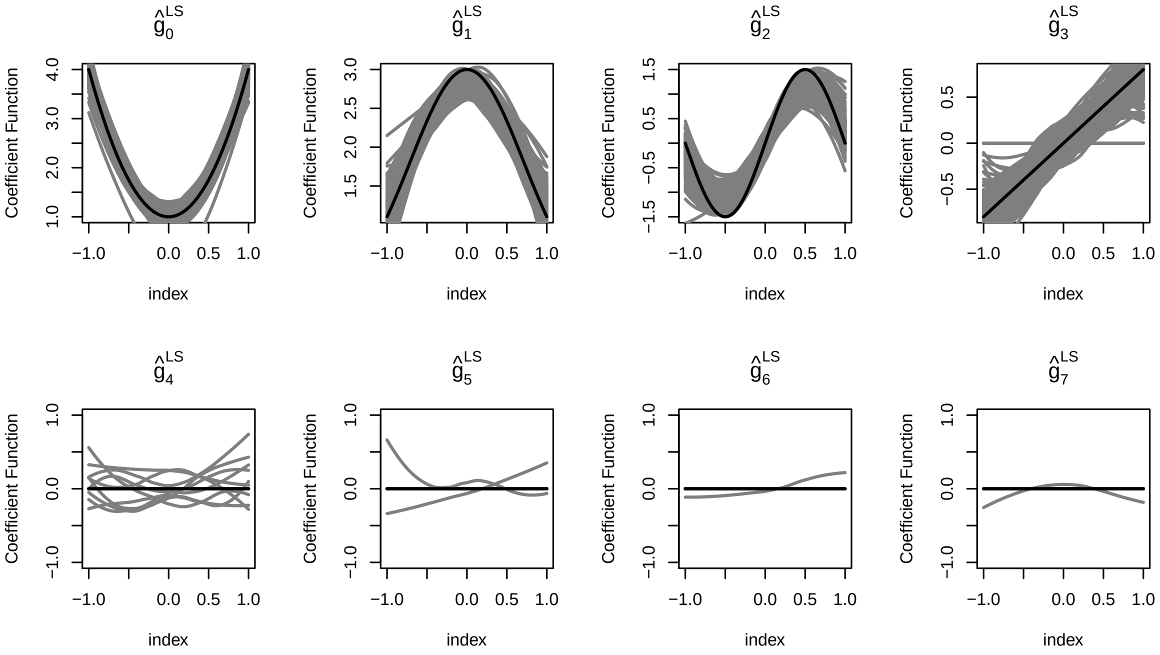

Figure 3.10: Coefficient functions estimated by RSGLASSO estimation under t_3 error distribution

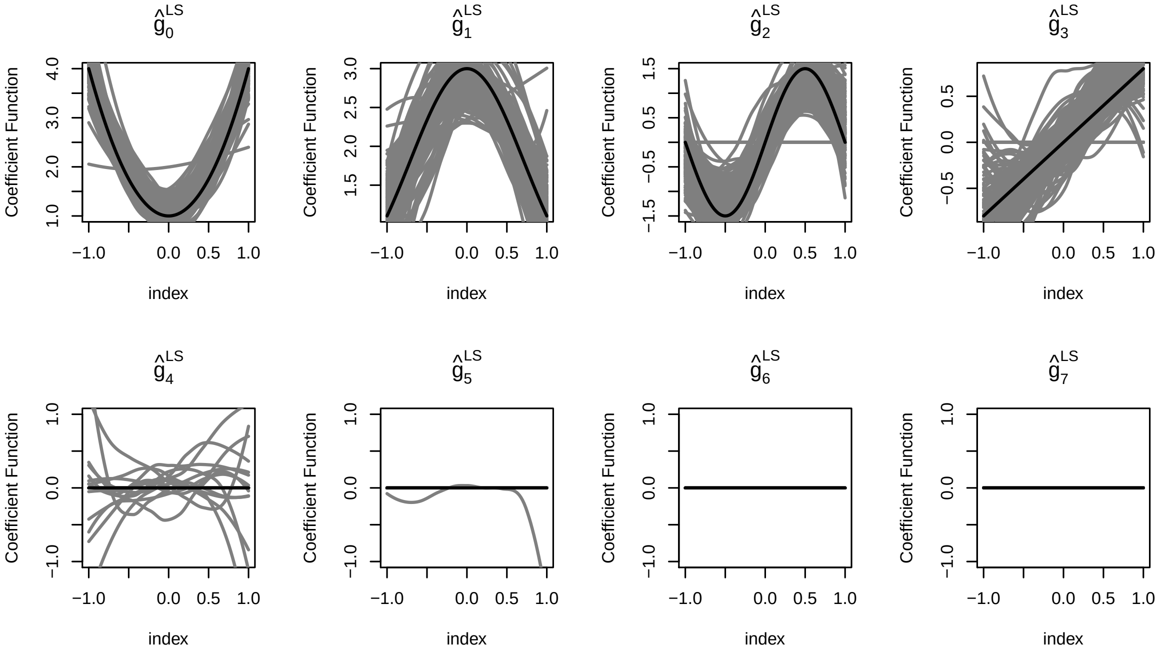

Figure 3.11: Coefficient functions estimated by LSSGLASSO estimation under $\mathcal{CN}(0.95)$ error distribution

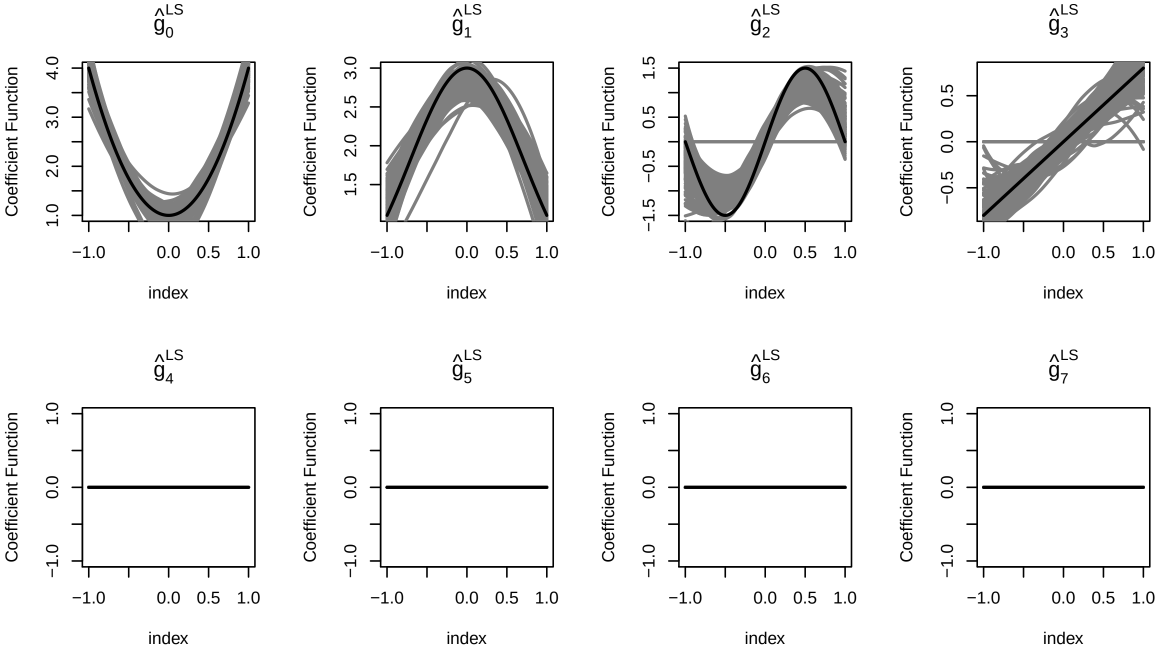

Figure 3.12: Coefficient functions estimated by RSGLASSO under $\mathcal{CN}(0.95)$ error distribution

References

AFSC. 2015. *NOAA MESA: Longline Survey.* Online. Accessed: 2014-04-14.

Andrews, Donald W. K. 1994. Asymptotics for Semiparametric Econometric Models Via Stochastic Equicontinuity. *Econometrica*, **62**(1), pp. 43–72.

Best, E.A., & St-Pierre, Gilbert. 1986. *Pacific halibut as predator and prey.* Technical Report 21. International Pacific Halibut Commission.

Bindele, Huybrechts F., & Abebe, Asheber. 2012. Bounded influence nonlinear signed-rank regression. *Canadian Journal of Statistics*, **40**(1), 172–189.

Breiman, L. 1995. Better subset selection using nonnegative garrote. *Techonometrics*, **37**, 373–384.

Cai, Zongwu, Juhl, Ted, & Yang, Bingduo. 2015. Functional index coefficient models with variable selection. *Journal of Econometrics*, **189**(2), 272 – 284. Frontiers in Time Series and Financial Econometrics.

Chang, William H., McKean, Joseph W., Naranjo, Joshua D., & Sheather, Simon J. 1999. High-breakdown rank regression. *J. Amer. Statist. Assoc.*, **94**(445), 205–219.

Delecroix, Michel, Hristache, Marian, & Patilea, Valentin. 2006. On semiparametric - estimation in single-index regression. *Journal of Statistical Planning and Inference*, **136**(3), 730 – 769.

Efron, Bradley, Hastie, Trevor, Johnstone, Iain, & Tibshirani, Robert. 2004. Least angle regression. *Ann. Statist.*, **32**(2), 407–499.

Fan, Jianqin, & Li, Ruize. 2001. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, **96**(456), 1348–1360.

Fan, Jianqing, & Gijbels, Irene. 1996. *Local polynomial modelling and its applications: monographs on statistics and applied probability 66*. Vol. 66. CRC Press.

Fan, Jianqing, Yao, Qiwei, & Cai, Zongwu. 2003. Adaptive varying-coefficient linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **65**(1), 57–80.

Feng, Long, Zou, Changliang, & Wang, Zhaojun. 2012. Rank-based inference for the single-index model. *Statistics & Probability Letters*, **82**(3), 535 – 541.

Feng, Long, Zou, Changliang, Wang, Zhaojun, Wei, Xianwu, & Chen, Bin. 2015. Robust spline-based variable selection in varying coefficient model. *Metrika*, **78**(1), 85–118.

Feng, Sanying, & Xue, Liugen. 2015. Model detection and estimation for single-index varying coefficient model. *Journal of Multivariate Analysis*, **139**, 227 – 244.

Gaichas, Sarah K.GaichasS.K., Aydin, Kerim Y.AydinK.Y., & Francis, Robert C.FrancisR.C. 2010. Using food web model results to inform stock assessment estimates of mortality and production for ecosystem-based fisheries management. *Canadian Journal of Fisheries and Aquatic Sciences*, **67**(9), 1490–1506.

Gu, Lijie, & Yang, Lijian. 2015. Oracally efficient estimation for single-index link function with simultaneous confidence band. *Electronic Journal of Statistics*, **9**(1), 1540–1561.

Hájek, Jaroslav, & Šidák, Zbyněk. 1967. *Theory of rank tests*. New York: Academic Press.

Hansen, Bruce E. 2008. Uniform convergence rates for kernel estimation with dependent data. *Econometric Theory*, **24**(03), 726–748.

Härdle, Wolfgang, Hall, Peter, & Ichimura, Hidehiko. 1993. Optimal Smoothing in Single-Index Models. *The Annals of Statistics*, **21**(1), pp. 157–178.

Hastie, Trevor, Tibshirani, Robert, & Friedman, Jerome. 2001. *The Elements of Statistical Learning*. Springer Series in Statistics. New York, NY, USA: Springer New York Inc.

Hettmansperger, T. P., & McKean, J. W. 1998. *Robust nonparametric statistical methods*. Kendall's Library of Statistics, vol. 5. London: Edward Arnold.

Hettmansperger, Thomas P., & McKean, Joseph W. 2011. *Robust nonparametric statistical methods*. Second edn. Monographs on Statistics and Applied Probability, vol. 119. CRC Press, Boca Raton, FL.

Huber, Peter J. 1964. Robust Estimation of a Location Parameter. *The Annals of Mathematical Statistics*, **35**(1), 73–101.

Jaeckel, Louis A. 1972. Estimating regression coefficients by minimizing the dispersion of the residuals. *Ann. Math. Statist.*, **43**, 1449–1458.

Krasker, William S., & Welsch, Roy E. 1982. Efficient Bounded-Influence Regression Estimation. *Journal of the American Statistical Association*, **77**(379), 595–604.

Li, Ker-Chau. 1991. Sliced inverse regression for dimension reduction. *Journal of the American Statistical Association*, **86**(414), 316–327.

Liu, Jicai, Zhang, Riquan, Zhao, Weihua, & Lv, Yazhao. 2013. A robust and efficient estimation method for single index models. *Journal of Multivariate Analysis*, **122**, 226 – 238.

Meier, Lukas, Van De Geer, Sara, & Bühlmann, Peter. 2008. The group lasso for logistic regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **70**(1), 53–71.

Moukhametov, I.N., Orlov, A.M., & Leaman, B.M. 2008. *Diet of Pacific halibut (Hippoglossus stenolepis) in the northwestern Pacific Ocean.* Technical Report 52. International Pacific Halibut Commission.

Naranjo, J. D., & Hettmansperger, T. P. 1994. Bounded influence rank regression. *J. Roy. Statist. Soc. Ser. B*, **56**(1), 209–220.

Newey, Whitney K, & McFadden, Daniel. 1994. Large sample estimation and hypothesis testing. *Handbook of econometrics*, **4**, 2111–2245.

NOAA. 2015. *National Data Buoy Center*. Online. Accessed: 2015-09-18.

Osborne, MR, Presnell, B, & Turlach, BA. 2000. A new approach to variable selection in least squares problems. *IMA Journal of Numerical Analysis*, **20**(3), 389–404.

Phillips, A Jason, Ciannelli, Lorenzo, Brodeur, Richard D, Pearcy, William G, & Childers, John. 2014. Spatio-temporal associations of albacore CPUEs in the Northeastern Pacific with regional SST and climate environmental variables. *ICES Journal of Marine Science: Journal du Conseil*, **71**(7), 1717–1727.

Rao, Tata Subba, Das, Sourav, & Boshnakov, Georgi N. 2014. A frequency domain approach for the estimation of parameters of spatio-temporal stationary random processes. *Journal of Time Series Analysis*, **35**(4), 357–377.

Serfling, Robert J. 1980. *Approximation theorems of mathematical statistics.* New York: John Wiley & Sons Inc. Wiley Series in Probability and Mathematical Statistics.

Sievers, Gerald L. 1983. A weighted dispersion function for estimation in linear models. *Communications in Statistics - Theory and Methods*, **12**(10), 1161–1179.

Song, Yunquan, Jian, Ling, & Lin, Lu. 2016. Robust exponential squared loss-based variable selection for high-dimensional single-index varying-coefficient model. *Journal of Computational and Applied Mathematics*, **308**, 330 – 345.

Stoner, Allan W, & Sturm, Erick A. 2004. Temperature and hunger mediate sablefish (Anoplopoma fimbria) feeding motivation: implications for stock assessment. *Canadian Journal of Fisheries and Aquatic Sciences*, **61**(2), 238–246.

Stoner, Allan W., Ottmar, Michele L., & Hurst, Thomas P. 2006. Temperature affects activity and feeding motivation in Pacific halibut: Implications for bait-dependent fishing. *Fisheries Research*, **81**(2-3), 202 – 209.

Tibshirani, Robert. 1996. Regression shrinkage and selection via the LASSO. *Journal of the Royal Statistical Society, Series B(Methodological)*, **58**(1), 267–288.

Trevor Hastie, Robert Tibshirani. 1993. Varying-Coefficient Models. *Journal of the Royal Statistical Society. Series B (Methodological)*, **55**(4), 757–796.

Wang, Hansheng, & Xia, Yingcun. 2009. Shrinkage estimation of the varying coefficient model. *Journal of the American Statistical Association*, **104**(486), 747–757.

Wang, Hansheng, Li, Runze, & Tsai, Chih-Ling. 2007. Tuning parameter selectors for the smoothly clipped absolute deviation method. *Biometrika*, **94**(3), 553.

Wang, Lan, Kai, Bo, & Li, Runze. 2009. Local Rank Inference for Varying Coefficient Models. *Journal of the American Statistical Association*, **104**(488), 1631–1645. PMID: 20657760.

Wang, Qin, & Yin, Xiangrong. 2008. A nonlinear multi-dimensional variable selection method for high dimensional data: Sparse MAVE. *Computational Statistics and Data Analysis*, **52**, 4512 – 4520.

Xia, Yingcun. 2006. Asymptotic distribution for two estimators of the single-index model. *Econometric Theory*, **22**(11), 1112–1137.

Xia, Yingcun, & Li, W. 1999. On Single-Index Coefficient Regression Models. *Journal of the American Statistical Association*, **94**(448), 1275–1285.

Xia, Yingcun, Tong, Howell, Li, W. K., & Zhu, Li-Xing. 2002. An adaptive estimation of dimension reduction space. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **64**(3), 363–410.

Xia, Yingcun, Li, W, & Tong, Howell. 2007. Threshold variable selection using nonparametric methods. *Statistica Sinica*, **17**(1), 265.

Xiang, Xiaojing. 1995. A strong law of large number for L-statistics in the non-i.d. case. *Comm. Statist. Theory Methods*, **24**(7), 1813–1819.

Xue, Liugen, & Pang, Zhen. 2013. Statistical inference for a single-index varying-coefficient model. *Statistics and Computing*, **23**(5), 589–599.

Xue, Liugen, & Wang, Qihua. 2012. Empirical likelihood for single-index varying-coefficient models. *Bernoulli*, **18**(3), 836–856.

Yang, Hu, Guo, Chaohui, & Lv, Jing. 2014. A robust and efficient estimation method for single-index varying-coefficient models. *Statistics & Probability Letters*, **94**, 119–127.

Yang, M-S., Dodd, K., Hibpshman, R., & Whitehouse, A. 2006 (May). *Food Habits of Groundfishes in the Gulf of Alaska in 1999 and 2001.* NOAA Technical Memorandum 164. U.S. Department of Commerce, NOAA, NMFS, AFSC.

Yao, Weixin, Lindsay, Bruce G., & Li, Runze. 2012. Local modal regression. *Journal of Nonparametric Statistics*, **24**(3), 647–663.

Yu, Yan, & Ruppert, David. 2002. Penalized Spline Estimation for Partially Linear Single-Index Models. *Journal of the American Statistical Association*, **97**(460), pp. 1042–1054.

Yuan, Ming, & Lin, Yi. 2006. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **68**(1), 49–67.

Zeng, Peng, He, Tianhong, & Zhu, Yu. 2012. A Lasso type approach for estimation and variable selection in single-index models. *Journal of Computational and Graphical Statistics*, **21**, 92 – 109.

Zhang, Riquan, Zhao, Weihua, & Liu, Jicai. 2013. Robust estimation and variable selection for semiparametric partially linear varying coefficient model based on modal regression. *Journal of Nonparametric Statistics*, **25**(2), 523–544.

Zou, Hui. 2006. The Adaptive Lasso and Its Oracle Properties. *Journal of the American Statistical Association*, **101**(476), 1418–1429.