REVSTAT – Statistical Journal
Volume 0, Number 0, Month 0000, 000–000
https://doi.org/00.00000/revstat.v00i0.000

Variable Selection and Estimation for Partially Linear Single-Index Errors-in-Variables Model

Authors: Zhaoliang Wang ✉
– School of Mathematics and Information Science, Henan Polytechnic University, Jiaozuo, Henan, P.R. China
wangzhaoliang@hpu.edu.cn

Xuyang Zhang
– School of Mathematics and Information Science, Henan Polytechnic University, Jiaozuo, Henan, P.R. China
212110010016@home.hpu.edu.cn

Received: Month 0000 Revised: Month 0000 Accepted: Month 0000

Abstract:

- This paper focuses on variable selection and estimation for the partially linear single-index model, considering the presence of measurement errors in all variables. Based on local linear regression, SIMEX technique and profile least square method, we employ the smoothly clipped absolute deviation (SCAD) penalty method to simultaneously estimate parameters and select important variables. Under some regularity conditions, the asymptotic distributions and oracle property of the proposed estimators are obtained. Meanwhile, we discuss the implementation algorithm of the estimation and the selection of bandwidth and tuning parameters. Monte Carlo simulation studies are carried out to evaluate the finite sample behaviour of the proposed method. The results show that the variable selection and parameter estimation are effective.

Keywords:

- *partially linear model; single-index model; measurement error; variable selection; oracle property*

AMS Subject Classification:

- 62F10, 62G08.

✉ Corresponding author

1. INTRODUCTION

The partially linear model combines the traditional linear model with the nonparametric regression model, which is widely concerned because it makes the model more flexible. However, nonparametric components often suffer from the "curse of dimensionality" and apply only to the low-dimensional covariates. To overcome the problem, the partially linear single-index model (PLSIM) has played an important role in the studies. In this paper, we consider the partially linear single-index model of the form

$$Y = g(\boldsymbol{X}^\top \boldsymbol{\theta}) + \boldsymbol{Z}^\top \boldsymbol{\beta} + \varepsilon, \tag{1.1}$$

where Y is the response variable, $\boldsymbol{X} \in \mathbb{R}^p$ and $\boldsymbol{Z} \in \mathbb{R}^q$ are two disjoint groups of covariates, and ε is the error term satisfying $E(\varepsilon|\boldsymbol{X}, \boldsymbol{Z}) = 0$ and $E(\varepsilon^2|\boldsymbol{X}, \boldsymbol{Z}) = \sigma^2 < \infty$; $g(\cdot)$ is the nonparametric link function which has a continuous second derivative, and $\boldsymbol{\theta} \in \mathbb{R}^p$ and $\boldsymbol{\beta} \in \mathbb{R}^q$ are the index parameter and link parameter vectors. For identifiability purpose, we assume that $\|\boldsymbol{\theta}\| = 1$ and, without loss of generality, the the first nonzero component of $\boldsymbol{\theta}$ is positive, where $\|\cdot\|$ denotes the Euclidean metric.

Model (1.1) is a fairly general and flexible model. Because model (1.1) considers the features of the single-index model and the linear model, it enjoys the merits of both. Subsequently, PLSIM has stronger interpretability and broader application prospect. Moreover, the PLSIM is not only less restrictive than the parametric model but also free from the "curse of dimensionality" which is often encountered in multivariate nonparametric settings. Among the available semiparametric models, the PLSIM plays an important role and has been widely applied in practice. Various methods have been proposed in the literature for estimating the unknown parameters and the nonparametric link function in the PLSIM. For example, Carroll et al (1997) proposed the backfitting algorithm. However, the resulting estimators may be unstable and undersmoothing the nonparametric function is necessary to reduce the bias of the parametric estimators. Accordingly, Yu and Ruppert (2002) proposed the penalized spline approach, which is computationally fast and stable. Xia and Härdle (2006) developed the well-known minimizing average variance estimation (MAVE) method for dimensionality reduction, which is an estimation method based on local linear smoothing and a modified form of least squares in Härdle et al (1993). Wang et al (2010) proposed a two-stage estimation procedure for the partially linear single-index model, which proves the asymptotic normality of the estimators for the parametric components. Although Yu and Ruppert (2002)'s procedure is useful, it may not yield efficient estimators; that is, the asymptotic covariance of their estimators does not reach the semiparametric efficiency bound. Therefore, we use the profile least-squares approach, which obtains an effective estimators and provides an effective bound. Other latest interesting works in semiparamatric models include Zhu and Xue (2006), Yan et al (2020), Liu et al (2021), Zou et al (2021), Cai and Wang (2023), etc.

Although the PLSIM, as discussed earlier, stands out due to its interpretability and broad applicability in various settings, these discussions often assume direct observability of covariates. However in practical problems, covariates usually cannot be measured directly, such as blood pressure, intelligence and obesity, etc, which will be affected to a certain extent in the measurement process, resulting in measurement errors. If we ignore the measurement errors, the estimators and inference may be biased and inconsistent. Hence, while the PLSIM

has been widely applied, its reliance on directly observable covariates becomes a limitation in the presence of measurement errors. In this paper, we are interested in variable selection and estimation of $\boldsymbol{\theta}$, $\boldsymbol{\beta}$ and the nonparametric function $g(\cdot)$, in the presence of measurement errors in both the parametric and the nonparametric part. More specifically, we assume an additive measurement error model as

$$\begin{aligned} \boldsymbol{W} &= \boldsymbol{X} + \boldsymbol{U}, \\ \boldsymbol{T} &= \boldsymbol{Z} + \boldsymbol{V}, \end{aligned} \tag{1.2}$$

where $\boldsymbol{W}$ and $\boldsymbol{T}$ are the observed surrogates, and $\boldsymbol{U}$ and $\boldsymbol{V}$ are measurement errors, independent of $(\boldsymbol{X}, \boldsymbol{Z}, Y)$, satisfying $E(\boldsymbol{U}) = \mathbf{0}_p$, $E(\boldsymbol{V}) = \mathbf{0}_q$, $\mathrm{Cov}(\boldsymbol{U}) = \Sigma_u$ and $\mathrm{Cov}(\boldsymbol{V}) = \Sigma_v$. For simplicity, we focus on the situation where Σ_u and Σ_v are known. Otherwise, Σ_u and Σ_v need to be first estimated, e.g., by the replication experiments method in Liang and Carroll (1999); Carroll et al (2006).

To eliminate the effects of measurement error, Cook and Stefanski (1994) developed the SIMEX method to correct the estimates in the presence of additive measurement error. Liang and Wang (2005) considered the partially linear single-index measurement error models with the linear part containing the measurement error, where they applied the correction for attenuation approach to obtain the efficient estimators of the parameters of interest. We note, however, that the above methods are not applicable for the occurrence with measurement errors in the nonparametric part of partially linear single-index models. In view of this, Lin et al (2022) considered the PLSIM with measurement errors possibly in all the variables, and proposed a new efficient estimation procedure based on the local linear smoothing and the SIMEX, and further establish the asymptotic normality. More related works include Chen and Cui (2009), Yang et al (2019), Huang and Zhao (2019), etc.

In practice, many explanatory variables are generally collected and need to be assessed during the initial analysis. Deciding which covariates to keep in the final statistical model and which variables are non-informative is practically interesting, but is always a tricky task for data analysis. Variable selection is therefore of fundamental interest in statistical modeling and analysis of data, and has become an integral part in most of the widely used statistics packages. No doubt variable selection will continue to be an important basic strategy for data analysis. Various methods for various models have been studied, for example, the bridge regression proposed in Frank and Friedman (1993), the least absolute shrinkage and selection operator (LASSO) proposed by Tibshirani (1996), the smoothly clipped absolute deviation (SCAD) proposed by Fan and Li (2001), the adaptive LASSO proposed in Zhou (2006), and so forth.

Tibshirani (1996) introduced LASSO to shrink the estimated coefficients of superfluous variables to zero in linear regression models, thus achieving the selection of significant variables and the corresponding parameter estimators at the same time. Subsequently, Fan and Li (2001) pointed out that Lasso method compresses model coefficients with large absolute values too much, which may cause unnecessary bias of the model, proved that Lasso estimation method does not have oracle properties, and proposed the SCAD approach that not only selects important variables consistently, but also produces parameter estimators as efficiently as if the true model were known, a property not possessed by the LASSO. In their research, Fan and Li (2001) proposed that a good penalty function should result in an estimator with three properties: sparsity, unbiased and continuity. The SCAD penalty estimator satisfies

these properties and has the oracle property that the zero coefficient can be estimated to be zero with probability one, and the nonzero coefficient estimator is asymptotically normal and has the same variance structure as the estimator obtained from the real model. Hence, we prefer to use the SCAD method for variable selection and estimation of the model. With respect to the partially linear model with measurement errors, Liang and Li (2009) proposed the penalized least squares estimation and penalized quantile estimation for the parameters through applying the SCAD penalty function. Their variable selection procedures were proven to be consistent and the resulting estimators share the oracle property. Liang et al (2010) applied the profile likelihood to solve the PLSIM and also employ the SCAD approach to simultaneously select variables and estimate regression coefficients, and possess the oracle property.

In the partially linear single-index model, there are few literatures on variable selection and estimation considering the presence of measurement errors in all variables. The main content of this paper is to construct SCAD estimators of parameter vectors $\boldsymbol{\theta}$ and $\boldsymbol{\beta}$ in this situation. We combine the SIMEX method, the local linear smoothing and the bias-corrected profile least-squares approach to obtain the SCAD estimators for both parameter vectors θ and β, and reduce the bias of the estimators, filter out the variables that have little or no obvious influence on the model, and improve the simplicity and interpretability of the model. Furthermore, we establish the asymptotic results of the SCAD estimators, which include the consistency and oracle properties. Simulation results are consistent with theoretical findings.

The rest of the paper is organized as follows. Section 2 presents the estimation method and gives the asymptotic properties of the obtained estimators. Some implementing issues including the a specific iterative algorithm and the choice of tuning parameters are discussed in Section 3. In Section 4, Monte Carlo simulations are carried out to assess the performance of the proposed estimation procedure. Section 5 is conclusion remark. Proofs of the main results are placed in the Appendix.

2. METHODOLOGY AND ASYMPTOTIC PROPERTIES

In this section, we study the SCAD variable selection method for the partially linear single-index model with measurement errors in both the parametric and the nonparametric part, and then we establish the asymptotic properties of the estimators.

2.1. Methods

For the model (1.1) with measurement errors in (1.2), the naive estimator that ignores measurement error often leads to inconsistent estimation. In this subsection, we extend the SIMEX method to estimate $g(\cdot)$ and also apply the penalized least-squares approach to simultaneously estimate parameters and select important variables. Cook and Stefanski (1994) proposed the well-known SIMEX for simulating and extrapolation. The SIMEX method consists of the simulation step, the estimation step, and extrapolation step. It entailed additional measurement error to the data with known increments, computing estimates from the addi-

tionally contaminated data, establishing a trend between these estimates and the variance of the addition errors, and extrapolation this trend back to the case without measurement error.

Assume that $\{(\boldsymbol{X}_i, \boldsymbol{Z}_i, Y_i, \boldsymbol{W}_i, \boldsymbol{T}_i, \boldsymbol{U}_i, \boldsymbol{V}_i), i = 1, 2, \cdots, n\}$ is independent and identically distributed (i.i.d.) copies of $(\boldsymbol{X}, \boldsymbol{Z}, Y, \boldsymbol{W}, \boldsymbol{T}, \boldsymbol{U}, \boldsymbol{V})$, where $(\boldsymbol{X}_i, \boldsymbol{Z}_i, Y_i)$ satisfies the model (1.1), and $(\boldsymbol{W}_i, \boldsymbol{T}_i)$ satisfies the model (1.2). For each $i = 1, 2, \cdots, n$, we generate a sequence of variables

$$\boldsymbol{W}_{ib} = \boldsymbol{W}_i + (\omega\Sigma_u)^{1/2}\varepsilon_{ib}, \qquad b = 1, 2, \cdots, B, \tag{2.1}$$

where $\varepsilon_{ib} \sim N(0, \mathbb{I}_p)$, $\mathbb{I}_p$ is a $p \times p$ identity matrix, B is a large but fixed integer, $\omega > 0$ and $\omega \in \mathcal{A} = \{\omega_1, \omega_2, \cdots, \omega_M\}$ is the grid of ω in the extrapolation step. This is the simulation component of our method. Here ω controls how much additional independent measurement error is added to the original $\boldsymbol{W}_i$. Simulation evidence suggests that the extrapolantation should be fitted for ω in a range of $[0, \omega_M]$ with $1 \leq \omega_M \leq 2$, see Carroll et al (2006).

Because $g(\cdot)$ is nonparametric modeling, it is natural to consider local linear smoothness. However, efficient estimation of the global parameters $\boldsymbol{\theta}$ and $\boldsymbol{\beta}$ require the use of all data, so their estimation relies on local linear smoothness. Suppose that $g(\cdot)$ has a continuous second derivative. For any v in a neighborhood of u, we apply the linear approximation of $g(\cdot)$ as

$$g(v) \approx g(u) + g'(u)(v - u) \equiv a_0 + a_1(v - u),$$

where $a_0 = g(u)$, $a_1 = g'(u)$.

With the simulated $\boldsymbol{W}_{ib}$, for t in the domain of $g(\cdot)$ and any given $\boldsymbol{\theta}$ and $\boldsymbol{\beta}$, we minimize

$$\sum_{i=1}^{n}[Y_i - \boldsymbol{T}_i^\top\boldsymbol{\beta} - a_0 - a_1(\boldsymbol{W}_{ib}^\top\boldsymbol{\theta} - t)]^2 K_h(\boldsymbol{W}_{ib}^\top\boldsymbol{\theta} - t) \tag{2.2}$$

with respect to $(a_0, a_1)^\top$, and denote the resulted minimizer as $\hat{a}_0 = \hat{g}_{\omega,b}(t; \boldsymbol{\theta}, \boldsymbol{\beta})$ and $\hat{a}_1 = \hat{g}'_{\omega,b}(t; \boldsymbol{\theta}, \boldsymbol{\beta})$. In (2.2), $K_h(\cdot) = h^{-1}K(\cdot/h)$, $K(\cdot)$ is a kernel function on $\mathbb{R}$ and $h > 0$ is a bandwidth. Exact determination of $\hat{g}_\omega(t; \boldsymbol{\theta}, \boldsymbol{\beta})$ for $\omega > 0$ is generally not feasible, but it can always be estimated arbitrarily well by generating a large number of independent measurement error vectors, $\{\varepsilon_{ib}, b = 1, \cdots, B\}$, computing $\hat{g}_{\omega,b}(t; \boldsymbol{\theta}, \boldsymbol{\beta})$ and approximating $\hat{g}_\omega(t; \boldsymbol{\theta}, \boldsymbol{\beta})$ by the sample mean of $\hat{g}_{\omega,b}(t; \boldsymbol{\theta}, \boldsymbol{\beta})$ for $b = 1, 2, \cdots, B$, that is

$$\hat{g}_\omega(t; \boldsymbol{\theta}, \boldsymbol{\beta}) = \frac{1}{B}\sum_{b=1}^{B}\hat{g}_{\omega,b}(t; \boldsymbol{\theta}, \boldsymbol{\beta}). \tag{2.3}$$

This is the estimation component of our method.

The extrapolation step of the proposal entails fitting a regression model of $\hat{g}_\omega(t; \boldsymbol{\theta}, \boldsymbol{\beta})$ on ω for $\omega \geq 0$ and using the model to extrapolate back to $\omega = -1$. This yield the simulation-extrapolation estimator,

$$\hat{g}_{\text{SIMEX}}(\boldsymbol{W}_i^\top\boldsymbol{\theta}; \boldsymbol{\theta}, \boldsymbol{\beta}) = \hat{g}_{-1}(\boldsymbol{W}_i^\top\boldsymbol{\theta}; \boldsymbol{\theta}, \boldsymbol{\beta}) + \hat{g}'_{-1}(\boldsymbol{W}_i^\top\boldsymbol{\theta}; \boldsymbol{\theta}, \boldsymbol{\beta})(\boldsymbol{W}_i^\top\boldsymbol{\theta}),$$

where $\hat{g}_{-1}(\boldsymbol{W}_i^\top\boldsymbol{\theta}; \boldsymbol{\theta}, \boldsymbol{\beta})$ is extrapolation back to $\omega = -1$ for average value $\hat{g}_{\omega,b}(\boldsymbol{W}_i^\top\boldsymbol{\theta}; \boldsymbol{\theta}, \boldsymbol{\beta})$ and similarly we obtain $\hat{g}'_{-1}(\boldsymbol{W}_i^\top\boldsymbol{\theta}; \boldsymbol{\theta}, \boldsymbol{\beta})$ for extrapolation for average value of $\hat{g}'_{\omega,b}(\boldsymbol{W}_i^\top\boldsymbol{\theta}; \boldsymbol{\theta}, \boldsymbol{\beta})$. This is the extrapolation component of our method.

Given the estimator $\hat{g}_{\text{SIMEX}}(t;\boldsymbol{\theta},\boldsymbol{\beta})$, the bias-corrected profile least-squares estimators of $(\boldsymbol{\theta}^\top,\boldsymbol{\beta}^\top)^\top$ is obtained by minimizing the following function

$$Q(\boldsymbol{\theta},\boldsymbol{\beta}) = \sum_{i=1}^{n}(Y_i - \hat{g}_{\text{SIMEX}}(\boldsymbol{W}_i^\top\boldsymbol{\theta};\boldsymbol{\theta},\boldsymbol{\beta}) - \boldsymbol{T}_i^\top\boldsymbol{\beta})^2 - n\bar{g}'^2(\Lambda)\boldsymbol{\theta}^\top\Sigma_u\boldsymbol{\theta} - n\boldsymbol{\beta}^\top\Sigma_v\boldsymbol{\beta}, \tag{2.4}$$

where $\bar{g}'^2(\Lambda) = n^{-1}\sum_{i=1}^{n}\hat{g}'^2(\boldsymbol{W}_i^\top\boldsymbol{\theta})$.

In practice, the true model is often unknown a priori. An underfitted model can yield biased estimates and predicted values, while an overfitted model can degrade the efficiency of the parameter estimates and predictions. This motivates us to apply the penalized approach to simultaneously estimate parameters and select important variables. To this end, we consider a penalized bias-corrected profile least-squares function

$$\mathcal{L}(\boldsymbol{\theta},\boldsymbol{\beta}) = Q(\boldsymbol{\theta},\boldsymbol{\beta}) + n\sum_{j=1}^{p}p_{\lambda_1}(|\theta_j|) + n\sum_{k=1}^{q}p_{\lambda_2}(|\beta_k|), \tag{2.5}$$

where $Q(\boldsymbol{\theta},\boldsymbol{\beta})$ is defined in (2.4), the penalty functions $p_{\lambda_1}(\cdot)$ and $p_{\lambda_2}(\cdot)$ in (2.5) are not necessarily the same for all $\boldsymbol{\theta}$ and $\boldsymbol{\beta}$, respectively. For example, we may wish to keep important predictors in a parametric model and hence not be willing to penalize their corresponding parameters. For simplicity of presentation, we assume that the penalty function for all coefficients are the same. For the purpose of selecting $\boldsymbol{X}$-variable only, we simply set $p_{\lambda_2}(\cdot) = 0$ and the resulting penalized bias-corrected profile least-squares function becomes

$$\mathcal{L}(\boldsymbol{\theta},\boldsymbol{\beta}) = Q(\boldsymbol{\theta},\boldsymbol{\beta}) + n\sum_{j=1}^{p}p_{\lambda_1}(|\theta_j|).$$

Similarly, if we are only interested in selecting $\boldsymbol{Z}$-variable, then we set $p_{\lambda_1}(\cdot) = 0$ so that

$$\mathcal{L}(\boldsymbol{\theta},\boldsymbol{\beta}) = Q(\boldsymbol{\theta},\boldsymbol{\beta}) + n\sum_{k=1}^{q}p_{\lambda_2}(|\beta_k|).$$

There are various penalty functions available in the literature. To obtain the oracle property of Fan and Li (2001), we adopt their SCAD penalty, whose first derivative is

$$p'_\lambda(\tau) = \lambda\left\{I(\tau\le\lambda) + \frac{(a\lambda-\tau)_+}{(a-1)\lambda}I(\tau>\lambda)\right\},$$

and where $p_\lambda(0) = 0$, $a = 3.7$ and $(t)_+ = \max(t,0)$. For the given tuning parameters, we obtain the penalized the estimators by minimizing $\mathcal{L}(\boldsymbol{\theta},\boldsymbol{\beta})$ in (2.5) with respect to $\boldsymbol{\theta}$ and $\boldsymbol{\beta}$. For the sake of simplicity, we denote the resulting estimators by $\hat{\boldsymbol{\theta}}$ and $\hat{\boldsymbol{\beta}}$.

2.2. Properties

In this section, we study the theoretical properties of the penalized profile bias-corrected least-squares estimators with the SCAD penalty function. The following regularity conditions are needed:

(C1) The density function of $\boldsymbol{X}^\top\boldsymbol{\theta}$, $f(\cdot)$, is Lipschitz continuous, bounded away from 0 and has a continuous second derivative on its support.

(C2) The function $g(\cdot)$ has bounded continuous second derivative.

(C3) The kernel $K(\cdot)$ is a bounded, continuous and symmetric density function with a bounded support satisfying the Lipschitz condition of order 1 and $\int u^2K(u)du \neq 0$.

(C4) The extrapolation function is theoretically exact.

(C5) $nh^2/(\log n)^2 \to \infty$ and $nh^8 \to 0$ as $n \to \infty$.

(C6) The error term satisfies $\sup E(\varepsilon^2|\boldsymbol{X},\boldsymbol{Z}) < \infty$ and $\sup E(\varepsilon^4|\boldsymbol{X},\boldsymbol{Z}) < \infty$.

(C7) The matrices $E\{[\boldsymbol{X}-E(\boldsymbol{X}|\boldsymbol{W}^\top\boldsymbol{\theta})][\boldsymbol{X}-E(\boldsymbol{X}|\boldsymbol{W}^\top\boldsymbol{\theta})]^\top\}$ and $E\{[\boldsymbol{Z}-E(\boldsymbol{Z}|\boldsymbol{W}^\top\boldsymbol{\theta})][\boldsymbol{Z}-E(\boldsymbol{Z}|\boldsymbol{W}^\top\boldsymbol{\theta})]^\top\}$ are positive-definite.

(C8) $\liminf_{n\to\infty}\liminf_{\theta\to 0^+} p'_{\lambda_1}(\theta)/\lambda_1 > 0$ and $\liminf_{n\to\infty}\liminf_{\beta\to 0^+} p'_{\lambda_2}(\beta)/\lambda_2 > 0$.

The above conditions are imposed for mathematical simplicity and may be modified if necessary. Conditions (C1) and (C2) are commonly used in the literature of the single-index regression analysis. Condition (C3) is the usual assumption for the kernel function. Condition (C4) is needed for the SIMEX method. Condition (C5) gives more regular conditions to choose the bandwidths. Conditions (C6)–(C8) are the necessary conditions for deriving the asymptotic normality and oracle property for the proposed estimators.

Let

$$\boldsymbol{\theta}_0 = (\theta_{10},\cdots,\theta_{p0})^\top = (\boldsymbol{\theta}_{10}^\top,\boldsymbol{\theta}_{20}^\top)^\top,\ \boldsymbol{\beta}_0 = (\beta_{10},\cdots,\beta_{q0})^\top = (\boldsymbol{\beta}_{10}^\top,\boldsymbol{\beta}_{20}^\top)^\top$$

be the true value of $\boldsymbol{\theta}$ and $\boldsymbol{\beta}$, respectively. Without loss of generality, assume that $\boldsymbol{\theta}_{10}$ and $\boldsymbol{\beta}_{10}$ consist of all nonzero components of $\boldsymbol{\theta}_0$ and $\boldsymbol{\beta}_0$, respectively, and $\boldsymbol{\theta}_{20}=\mathbf{0}$, $\boldsymbol{\beta}_{20}=\mathbf{0}$. Let s and t respectively denote the dimension of $\boldsymbol{\theta}_{10}$ and $\boldsymbol{\beta}_{10}$. Denote

$$\begin{aligned} a_n &= \max_{1\le j\le p}\{|p'_{\lambda_1}(|\theta_{j0}|)| : \theta_{j0}\neq 0\}\ ,\ c_n = \max_{1\le k\le q}\{|p'_{\lambda_2}(|\beta_{k0}|)| : \beta_{k0}\neq 0\},\\ b_n &= \max_{1\le j\le p}\{|p''_{\lambda_1}(|\theta_{j0}|)| : \theta_{j0}\neq 0\}\ ,\ d_n = \max_{1\le k\le q}\{|p''_{\lambda_2}(|\theta_{k0}|)| : \beta_{k0}\neq 0\}, \end{aligned} \tag{2.6}$$

and

$$\begin{aligned} \boldsymbol{R}_1 &= (p'_{\lambda_1}(|\theta_{10}|)\mathrm{sgn}(\theta_{10}),\cdots,p'_{\lambda_1}(|\theta_{s0}|)\mathrm{sgn}(\theta_{s0}))^\top,\\ \boldsymbol{R}_2 &= (p'_{\lambda_2}(|\beta_{10}|)\mathrm{sgn}(\beta_{10}),\cdots,p'_{\lambda_2}(|\beta_{t0}|)\mathrm{sgn}(\beta_{t0}))^\top,\\ \Sigma_1 &= \mathrm{diag}\{p''_{\lambda_1}(|\theta_{10}|),\cdots,p''_{\lambda_1}(|\theta_{s0}|)\},\\ \Sigma_2 &= \mathrm{diag}\{p''_{\lambda_2}(|\beta_{10}|),\cdots,p''_{\lambda_2}(|\beta_{t0}|)\}. \end{aligned} \tag{2.7}$$

In what follows, $\boldsymbol{A}^{\otimes^2} = \boldsymbol{A}\boldsymbol{A}^\top$ for any vector $\boldsymbol{A}$. Let $\tilde{\xi}_i = \xi_i - E[\xi|\boldsymbol{W}^\top\boldsymbol{\theta} = \boldsymbol{W}_i^\top\boldsymbol{\theta}]$ for any random variable or vector ξ. For example, $\tilde{\boldsymbol{T}}_i = \boldsymbol{T}_i - E[\boldsymbol{T}|\boldsymbol{W}^\top\boldsymbol{\theta} = \boldsymbol{W}_i^\top\boldsymbol{\theta}]$. Let Σ_{u11} be the (s,s) - left upper submatrix of Σ_u and Σ_{v11} be the (t,t) - left upper submatrix of Σ_v. In addition, we define $(\boldsymbol{W}_1,\boldsymbol{X}_1)$ and $(\boldsymbol{T}_1,\boldsymbol{Z}_1)$ in such a way that they consist of the first s and t elements of $(\boldsymbol{W},\boldsymbol{X})$ and $(\boldsymbol{T},\boldsymbol{Z})$ respectively. We define $(\tilde{\boldsymbol{W}}_1,\tilde{\boldsymbol{X}}_1)$ and $(\tilde{\boldsymbol{T}}_1,\tilde{\boldsymbol{Z}}_1)$ analogously. We have the following theorem, whose proof is given in the Appendix.

Theorem 2.1. *Suppose that the regularity conditions (C1)–(C8) hold. If $b_n \to 0$ and $d_n \to 0$ as $n \to \infty$, then there exists a local minimizer $(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\beta}})$ of (2.5) such that the rate is $O_p(n^{-1/2} + a_n + c_n)$, where a_n, c_n and b_n, d_n are given by (2.6).*

It is clear from Theorem 2.1 that the rate of convergence of the estimators $\hat{\boldsymbol{\theta}}$ and $\hat{\boldsymbol{\beta}}$ depends on λ_1 and λ_2. As long as the appropriate λ_1 and λ_2 are chosen, such that $a_n = c_n = O(n^{-1/2})$, there exists a root-n consistent penalized estimators.

Theorem 2.2. *Suppose that the regularity conditions (C1)–(C8) hold. If $\lambda_1 \to 0$, $\sqrt{n}\lambda_1 \to \infty$ and $\lambda_2 \to 0$, $\sqrt{n}\lambda_2 \to \infty$ as $n \to \infty$, then with probability tending to 1, the penalized estimators $\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{\theta}}_1^\top, \hat{\boldsymbol{\theta}}_2^\top)^\top$ and $\hat{\boldsymbol{\beta}} = (\hat{\boldsymbol{\beta}}_1^\top, \hat{\boldsymbol{\beta}}_2^\top)^\top$ in Theorem 2.1 must satisfy:*

(i) *Sparsity: $\hat{\boldsymbol{\theta}}_2 = 0$ and $\hat{\boldsymbol{\beta}}_2 = 0$.*

(ii) *Asymptotic normality:*

$$\sqrt{n}\left\{\Gamma_4 - \Gamma_{\tilde{W}_1\tilde{Z}_1}\Gamma_3^{-1}\Gamma_{\tilde{W}_1\tilde{Z}_1}^\top\right\}\left(\hat{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}_{10}\right) + \frac{1}{\sqrt{n}}\left\{\Gamma_{\tilde{W}_1\tilde{Z}_1}\Gamma_3^{-1}\boldsymbol{R}_2 - \boldsymbol{R}_1\right\} \xrightarrow{L} N\left(0, \Omega_\theta\right),$$

$$\sqrt{n}\left\{\Gamma_3 - \Gamma_{\tilde{W}_1\tilde{Z}_1}^\top\Gamma_4^{-1}\Gamma_{\tilde{W}_1\tilde{Z}_1}\right\}\left(\hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_{10}\right) + \frac{1}{\sqrt{n}}\left\{\Gamma_{\tilde{W}_1\tilde{Z}_1}^\top\Gamma_4^{-1}\boldsymbol{R}_1 - \boldsymbol{R}_2\right\} \xrightarrow{L} N\left(0, \Omega_\beta\right),$$

where "$\xrightarrow{L}$" is convergence in distribution, and all expected values are well defined. $\Gamma_{\tilde{W}_1} = E(g'(\boldsymbol{W}_1^\top\boldsymbol{\theta}_{10})\tilde{\boldsymbol{W}}_1)^{\otimes^2}$, $\Gamma_{\tilde{W}_1\tilde{Z}_1} = E(g'(\boldsymbol{W}_1^\top\boldsymbol{\theta}_{10})\tilde{\boldsymbol{W}}_1\tilde{\boldsymbol{Z}}_1)$, $\Gamma_{\tilde{X}_1} = E(g'(\boldsymbol{W}_1^\top\boldsymbol{\theta}_{10})\tilde{\boldsymbol{X}}_1)^{\otimes^2}$, $\Gamma_{\tilde{Z}_1} = E(\tilde{\boldsymbol{Z}}_1^\top\tilde{\boldsymbol{Z}}_1)$, $\Gamma_{U_1} = E(g'(\boldsymbol{W}_1^\top\boldsymbol{\theta}_{10})\boldsymbol{U}_1)^{\otimes^2}$, $\Gamma_1 = E(\bar{g}'^2(\Lambda)(\boldsymbol{U}_1\boldsymbol{U}_1^\top - \Sigma_{u11})\boldsymbol{\theta}_{10})^{\otimes^2}$, $\Gamma_2 = E((\boldsymbol{V}_1\boldsymbol{V}_1^\top - \Sigma_{v11})\boldsymbol{\beta}_{10})^{\otimes^2}$, $\Gamma_3 = \Gamma_{\tilde{Z}_1} + \Sigma_2$, $\Gamma_4 = \Gamma_{\tilde{W}_1} - \bar{g}'^2(\Lambda)\Sigma_{u11} + \Sigma_1$, $\Gamma_{\Delta_1} = E(\varepsilon - g'(\boldsymbol{X}_1^\top\boldsymbol{\theta}_{10})\boldsymbol{U}_1^\top\boldsymbol{\theta}_{10} - \boldsymbol{V}_1^\top\boldsymbol{\beta}_{10})^{\otimes^2}$, $\Gamma_{\tilde{X}_1}^* = \Gamma_{\tilde{X}_1} - \Gamma_{\tilde{W}_1\tilde{Z}_1}\Gamma_3^{-1}\Gamma_{\tilde{Z}_1}\Gamma_3^{-1}\Gamma_{\tilde{W}_1\tilde{Z}_1}^\top$, $\Omega_\theta = \Gamma_{\tilde{X}_1}^*\Gamma_{\Delta_1} + \Gamma_{U_1}\sigma^2 + \Gamma_1 + \Gamma_{\tilde{W}_1\tilde{Z}_1}\Gamma_3^{-1}\{\Sigma_{v11}\sigma^2 + \Gamma_2\}\Gamma_3^{-1}\Gamma_{\tilde{W}_1\tilde{Z}_1}^\top$, $\Gamma_{\tilde{Z}_1}^* = \Gamma_{\tilde{Z}_1} - \Gamma_{\tilde{W}_1\tilde{Z}_1}\Gamma_4^{-1}\Gamma_{\tilde{X}_1}\Gamma_4^{-1}\Gamma_{\tilde{W}_1\tilde{Z}_1}$, $\Omega_\beta = \Gamma_{\tilde{Z}_1}^*\Gamma_{\Delta_1} + \Sigma_{v11}\sigma_2 + \Gamma_2 + \Gamma_{\tilde{W}_1\tilde{Z}_1}^\top\Gamma_4^{-1}\{\Gamma_{u11}\sigma^2 + \Gamma_1\}\Gamma_4^{-1}\Gamma_{\tilde{W}_1\tilde{Z}_1}$.

3. IMPLEMENTATION

3.1. Algorithm

In the previous section, the estimation method for the partially linear single-index model with the SCAD penalty function, considering measurement errors, is presented. Since $\boldsymbol{\theta}$ and $\boldsymbol{\beta}$ have no explicit solutions, the traditional iterative algorithm can not be directly used in optimization problem (2.5), but the optimization can still be achieved with this algorithm after appropriate adjustment. The SCAD penalty functions are singular at the origin, and they do not have continuous second order derivatives. Therefore, we can adopt the local quadratic approximation of the penalty function proposed by Fan and Li (2001) to replace the penalty function.

More specifically, given the initial value $\hat{\boldsymbol{\beta}}^{(0)}$, and a specified small positive number ϵ. For $k = 1, \cdots, q$, when $|\hat{\beta}_k^{(0)}| < \epsilon$, let $\hat{\beta}_k = 0$; when $|\hat{\beta}_k^{(0)}| \geq \epsilon$, we have

$$p_{\lambda_2}(|\beta_k|) \approx p_{\lambda_2}(|\hat{\beta}_k^{(0)}|) + \frac{p'_{\lambda_2}(|\hat{\beta}_k^{(0)}|)}{2|\hat{\beta}_k^{(0)}|}\left(\beta_k^2 - \hat{\beta}_k^{(0)2}\right). \tag{3.1}$$

Analogously, given the initial value $\hat{\boldsymbol{\theta}}^{(0)}$, if $|\hat{\theta}_k^{(0)}| < \epsilon$, $k = 1, \cdots, p$, then let $\hat{\theta}_k = 0$; otherwise, we have

$$p_{\lambda_1}(|\theta_j|) \approx p_{\lambda_1}(|\hat{\theta}_j^{(0)}|) + \frac{p'_{\lambda_1}(|\hat{\theta}_j^{(0)}|)}{2|\hat{\theta}_j^{(0)}|}\left(\theta_j^2 - \hat{\theta}_j^{(0)2}\right).$$

With the aid of the local quadratic approximation, minimizing (2.5) is equivalent to minimizing

$$\begin{aligned}
&\sum_{i=1}^{n}\left\{Y_i - \hat{g}(\boldsymbol{W}_i^\top\hat{\boldsymbol{\theta}}^{(0)};\hat{\boldsymbol{\theta}}^{(0)},\hat{\boldsymbol{\beta}}^{(0)}) - \hat{g}'(\boldsymbol{W}_i^\top\hat{\boldsymbol{\theta}}^{(0)};\hat{\boldsymbol{\theta}}^{(0)},\hat{\boldsymbol{\beta}}^{(0)})\left(\boldsymbol{W}_i^\top\boldsymbol{\theta} - \boldsymbol{W}_i^\top\hat{\boldsymbol{\theta}}^{(0)}\right) - \boldsymbol{T}_i^\top\boldsymbol{\beta}\right\}^2 \\
&- n\bar{g}'^2(\Lambda)\boldsymbol{\theta}^\top\Sigma_u\boldsymbol{\theta} - n\boldsymbol{\beta}^\top\Sigma_v\boldsymbol{\beta} + n\sum_{j=1}^{p}\frac{p'_{\lambda_1}(|\hat{\theta}_j^{(0)}|)}{|\hat{\theta}_j^{(0)}|}\theta_j^2 + n\sum_{k=1}^{q}\frac{p'_{\lambda_2}(|\hat{\beta}_k^{(0)}|)}{|\hat{\beta}_k^{(0)}|}\beta_k^2 \\
=&\sum_{i=1}^{n}\left\{\tilde{Y}_i - \left(\hat{g}'(\boldsymbol{W}_i^\top\hat{\boldsymbol{\theta}}^{(0)})\boldsymbol{W}_i^\top \;\; \boldsymbol{T}_i^\top\right)\begin{pmatrix}\boldsymbol{\theta}\\ \boldsymbol{\beta}\end{pmatrix}\right\}^2 \\
&+ n\left(\boldsymbol{\theta}^\top \;\; \boldsymbol{\beta}^\top\right)\begin{pmatrix}\Sigma_{\lambda_1}(\hat{\boldsymbol{\theta}}^{(0)}) - \bar{g}'^2(\Lambda)\Sigma_u & 0\\ 0 & \Sigma_{\lambda_2}(\hat{\boldsymbol{\beta}}^{(0)}) - \Sigma_v\end{pmatrix}\begin{pmatrix}\boldsymbol{\theta}\\ \boldsymbol{\beta}\end{pmatrix}
\end{aligned} \tag{3.2}$$

where $\tilde{Y}_i = Y_i - \hat{g}(\boldsymbol{W}_i^\top\hat{\boldsymbol{\theta}}^{(0)}) + \hat{g}'(\boldsymbol{W}_i^\top\hat{\boldsymbol{\theta}}^{(0)})\boldsymbol{W}_i^\top\hat{\boldsymbol{\theta}}^{(0)}$, where $\hat{g}$ and $\hat{g}'$ are part of the SIMEX estimator. $\Sigma_{\lambda_1}(\hat{\boldsymbol{\theta}}^{(0)}) = \mathrm{diag}\left\{\frac{p'_{\lambda_1}(|\hat{\theta}_1^{(0)}|)}{|\hat{\theta}_1^{(0)}|}, \cdots, \frac{p'_{\lambda_1}(|\hat{\theta}_p^{(0)}|)}{|\hat{\theta}_p^{(0)}|}\right\}$, $\Sigma_{\lambda_2}(\hat{\boldsymbol{\beta}}^{(0)}) = \mathrm{diag}\left\{\frac{p'_{\lambda_2}(|\hat{\beta}_1^{(0)}|)}{|\hat{\beta}_1^{(0)}|}, \cdots, \frac{p'_{\lambda_2}(|\hat{\beta}_q^{(0)}|)}{|\hat{\beta}_q^{(0)}|}\right\}$. Note that (3.2) is an approximation of (2.5) by first-order Taylor expansion. On this basis, optimization problem (3.2) is further implemented iteratively using the following algorithm.

Step 1 The initial estimators $\hat{\boldsymbol{\theta}}_1^{(0)}$ and $\hat{\boldsymbol{\beta}}^{(0)}$ are obtained by the method of the unpenalized profile least-squares, where we ignore measurement errors, replace $(\boldsymbol{X}, \boldsymbol{Z})$ by $(\boldsymbol{W}, \boldsymbol{T})$, and set $\hat{\boldsymbol{\theta}}^{(0)} = \hat{\boldsymbol{\theta}}_1^{(0)}/\|\hat{\boldsymbol{\theta}}_1^{(0)}\|$ and $l = 0$.

Step 2 Use the current estimators $\hat{\boldsymbol{\theta}}^{(l)}$ and $\hat{\boldsymbol{\beta}}^{(l)}$, find $\hat{a}_0(u;\hat{\boldsymbol{\theta}}^{(l)},\hat{\boldsymbol{\beta}}^{(l)})$ and $\hat{a}_1(u;\hat{\boldsymbol{\theta}}^{(l)},\hat{\boldsymbol{\beta}}^{(l)})$ by minimizing (2.2), averaging as in (2.3), and extrapolating back to $\omega = -1$, with $\hat{\boldsymbol{\theta}}^{(l)}$ and $\hat{\boldsymbol{\beta}}^{(l)}$ replacing $\boldsymbol{\theta}$ and $\boldsymbol{\beta}$.

Step 3 If $\hat{\theta}_j^{(l)}$ and $\hat{\beta}_k^{(l)}$ are close to zero, then set $\hat{\theta}_j = 0$ and $\hat{\beta}_k = 0$. Otherwise, by minimizing (3.2), we update the estimators of $(\boldsymbol{\theta}^\top, \boldsymbol{\beta}^\top)^\top$ by

$$\begin{aligned}
\begin{pmatrix}\hat{\boldsymbol{\theta}}^{(l+1)}\\ \hat{\boldsymbol{\beta}}^{(l+1)}\end{pmatrix} =& \left\{\sum_{i=1}^{n}\begin{pmatrix}\hat{g}'(\boldsymbol{W}_i^\top\hat{\boldsymbol{\theta}}^{(l)})\boldsymbol{W}_i\\ \boldsymbol{T}_i\end{pmatrix}\left(\hat{g}'(\boldsymbol{W}_i^\top\hat{\boldsymbol{\theta}}^{(l)})\boldsymbol{W}_i^\top \;\; \boldsymbol{T}_i^\top\right)\right. \\
&\left. + n\begin{pmatrix}\Sigma_{\lambda_1}(\hat{\boldsymbol{\theta}}^{(l)}) - \bar{g}'^2(\Lambda)\Sigma_u & 0\\ 0 & \Sigma_{\lambda_2}(\hat{\boldsymbol{\beta}}^{(l)}) - \Sigma_v\end{pmatrix}\right\}^{-1} \times \sum_{i=1}^{n}\begin{pmatrix}\hat{g}'(\boldsymbol{W}_i\hat{\boldsymbol{\theta}}^{(l)})\boldsymbol{W}_i\\ \boldsymbol{T}_i\end{pmatrix}\tilde{Y}_i.
\end{aligned}$$

Step 4 Iterate Step 2 and Step 3 until convergence. We obtain the estimators of $\boldsymbol{\theta}$ and $\boldsymbol{\beta}$.

Step 5 With $\left(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\beta}}\right)$ from Step 4, the final estimate of $g(\cdot)$ can be obtained by carrying out Step 2.

3.2. Tuning parameters selection

The bandwidth h is selected by applying leave-one-out cross validation method, which minimizes

$$\mathrm{CV}(h) = \frac{1}{n}\sum_{i=1}^{n}\left\{Y_i - \hat{g}_{(-i)}(\boldsymbol{W}_i^\top\hat{\boldsymbol{\theta}}_{(-i)}) - \boldsymbol{T}_i^\top\hat{\boldsymbol{\beta}}_{(-i)}\right\}^2,$$

over a grid of h's, where $\hat{\boldsymbol{\theta}}_{(-i)}$, $\hat{\boldsymbol{\beta}}_{(-i)}$ and $\hat{g}_{(-i)}(\cdot)$ are the proposed estimators of $\boldsymbol{\theta}$, $\boldsymbol{\beta}$, and $g(\cdot)$ with the ith observation deleted. The CV bandwidth h_{CV} is selected to minimize $\mathrm{CV}(h)$, that is $h_{\mathrm{CV}} = \min \mathrm{CV}(h)$.

Theorem 2.2 indicates that the proposed variable selection procedure possesses the oracle property. However, this attractive feature relies on the tuning parameters. After the selection of bandwidth h, we select tuning parameters λ_1 and λ_2. To this end, we describe the tuning parameters selection procedure in detail. Because it is computationally expensive to minimize BIC, defined below, with respect to the $(p+q)$-dimensional tuning parameters, we follow the approach of Fan and Li (2004) to set $\lambda_1 = \lambda\mathrm{SE}\left(\hat{\theta}_j^u\right)$ and $\lambda_2 = \lambda\mathrm{SE}\left(\hat{\beta}_k^u\right)$, where λ is the tuning parameter, and $\mathrm{SE}\left(\hat{\theta}_j^u\right)$ and $\mathrm{SE}\left(\hat{\beta}_k^u\right)$ are the standard errors of the unpenalized profile least-squares estimators of θ_j and β_k, respectively, for $j = 1, \cdots, p$ and $k = 1, \cdots, q$. Let the resulting SCAD estimators be $\hat{\boldsymbol{\theta}}_\lambda$ and $\hat{\boldsymbol{\beta}}_\lambda$. We then select λ by minimizing the objective function

$$\mathrm{BIC}(\lambda) = \log\left(\mathrm{MSE}_\lambda\right) + \mathrm{DF}_\lambda \log(n)/n, \tag{3.3}$$

where $\mathrm{MSE}_\lambda = \frac{1}{n}\sum_{i=1}^{n}\left\{\left(Y_i - \hat{g}(\boldsymbol{W}_i^\top\hat{\boldsymbol{\theta}}_\lambda) - \boldsymbol{T}_i^\top\hat{\boldsymbol{\beta}}_\lambda\right)^2 - \bar{g}'^2(\Lambda)\hat{\boldsymbol{\theta}}_\lambda^\top\Sigma_u\hat{\boldsymbol{\theta}}_\lambda - \hat{\boldsymbol{\beta}}_\lambda^\top\Sigma_v\hat{\boldsymbol{\beta}}_\lambda\right\}$ and DF_λ is the number of nonzero coefficients of both $\hat{\boldsymbol{\theta}}_\lambda$ and $\hat{\boldsymbol{\beta}}_\lambda$. Thus, the minimization problem over λ will reduce to a one-dimensional minimization problem concerning. In special cases, λ can be selected by minimizing $\mathrm{BIC}(\lambda)$ in the set of grid points on the bounded interval $[0, \lambda_{\max}]$, where $\lambda_{\max}/\sqrt{n} \to 0$ as $n \to \infty$. In practice, a plot of the $\mathrm{BIC}(\lambda)$ against λ can be used to determine an appropriate λ_{max} to ensure that the $\mathrm{BIC}(\lambda)$ reaches its minimum around the middle of the range of λ. Then take the grid points of λ on $[0, \lambda_{\max}]$, from which the tuning parameters are selected, which can avoid excessive calculation in the minimization of $\mathrm{BIC}(\lambda)$.

4. SIMULATIONS

In this section, we evaluate the finite sample performance of the proposed procedure through Monte Carlo simulations. We compare our SCAD penalty method (denoted as SIMEX+SCAD) with unpenalized estimators for complexity with considering measurement errors (denoted as SIMEX). In our simulations, we also compare our proposed methods with the naive estimator (denoted as Naive) and penalized naive estimator (denoted as Naive+SCAD) that ignore measurement errors with a direct replacement of $\boldsymbol{X}$ and $\boldsymbol{Z}$ by $\boldsymbol{W}$ and $\boldsymbol{T}$. As a benchmark, the oracle estimators (denoted as Oracle) are computed and

used for comparison, where the zero components are known a priori and $\boldsymbol{X}$ and $\boldsymbol{Z}$ can be observed. It was computed as gold standard although it is unachievable in practice.

To evaluate the performance of the proposed variable selection estimation method, we used the following criterions.

- The square of the R statistic: $R^2_{\boldsymbol{\theta}} = |\hat{\boldsymbol{\theta}}^\top \boldsymbol{\theta}_0|^2 / |\boldsymbol{\theta}_0^\top \boldsymbol{\theta}_0|^2$ and $R^2_{\boldsymbol{\beta}} = |\hat{\boldsymbol{\beta}}^\top \boldsymbol{\beta}_0|^2 / |\boldsymbol{\beta}_0^\top \boldsymbol{\beta}_0|^2$.
- The numbers of zero coefficients and nonzero coefficients obtained by different methods: "TN" was the average number of zero coefficients correctly estimated as zero, and "TP" was the number of nonzero coefficients identified as nonzero.

The performance of the estimate of the link function $g(\cdot)$ is assessed by using the square root of average square errors (RASE), defined by

$$\text{RASE}(\hat{g}(\cdot)) = \left[\frac{1}{n_{\text{grid}}} \sum_{j=1}^{n_{\text{grid}}} (\hat{g}(t_j) - g(t_j))^2 \right]^{1/2}, \tag{4.1}$$

where $\{t_j, j = 1, \cdots, n_{\text{grid}}\}$ is a set of grid points at which the functions $\hat{g}(\cdot)$ are evaluated. We considered $n_{\text{grid}} = 200$.

The data are generated from the following model

$$\begin{cases} Y = g(\boldsymbol{X}^\top \boldsymbol{\theta}_0) + \boldsymbol{Z}^\top \boldsymbol{\beta}_0 + \varepsilon, \\ \boldsymbol{W} = \boldsymbol{X} + \boldsymbol{U}, \\ \boldsymbol{T} = \boldsymbol{Z} + \boldsymbol{V}, \end{cases} \tag{4.2}$$

where $\boldsymbol{\theta}_0 = \frac{1}{\sqrt{2}}(0, 0, 0, 1, 1)^\top$, $\boldsymbol{\beta}_0 = (3, 2, 1, 0, 0, \cdots, 0)^\top$ and $g(x) = e^x$. The dimension of $\boldsymbol{\theta}_0$ and $\boldsymbol{\beta}_0$ are p and q, respectively. Denoted $\boldsymbol{X} = (X_1, \cdots, X_p)^\top$ and $\boldsymbol{Z} = (Z_1, \cdots, Z_q)^\top$, where $p = 5$ and $q = 10$. We generated the covariates $X_1, \cdots, X_p$ from uniform distribution $U(0, 1)$ independently, and the covariates $Z_1, \cdots, Z_q$ were generated from the multivariate normal distribution with mean vector zero and the pairwise correlation between Z_j and Z_k being $\rho^{|j-k|}$ with $\rho = 0.5$. Then the linear covariates were not independent and highly correlated. The error ε is generated from normal distribution $N(0, 0.2^2)$. The measurement error $\boldsymbol{U}$ is normally distributed $N(0, \sigma_u^2 \mathbb{I}_p)$, and $\boldsymbol{V}$ is normally distributed $N(0, \sigma_v^2 \mathbb{I}_q)$, where the measurement error variance $\sigma_u^2 = \sigma_v^2 = 0.2^2$.

Using the SIMEX algorithm, we consider widely used quadratic extrapolation function $a + b\omega + c\omega^2$ and take $\omega = 0, 0.2, \cdots, 2$ and $B = 50$. We use the Gaussian kernel $K(u) = \frac{1}{\sqrt{2\pi}} e^{-\frac{u^2}{2}}$. The sample sizes for the simulated data are $n = 100$, 200 and 400. For each setting, we simulate 500 times to assess the performance. We report the results in Tables 1 and 2 .

From Table 1, one may have the following observations. (i) The proposed estimators outperform SIMEX and two naive estimators. It can be observed that the proposed estimators are close to the oracle estimators in terms of $R^2_{\boldsymbol{\theta}}$ and $R^2_{\boldsymbol{\beta}}$, which are close to 1. (ii) Generally, with the sample size increasing, the proposed method's performance improves. The proposed method obtain the exactly same $\text{TN}_{\boldsymbol{\theta}}$, $\text{TP}_{\boldsymbol{\theta}}$, $\text{TN}_{\boldsymbol{\beta}}$, $\text{TP}_{\boldsymbol{\beta}}$ as the gold standard with sample size is not small. (iii) The purpose of dimension reduction cannot be reached for Naive and SIMEX. The proposed method performed variable selection and parameter estimation

simultaneously, which was able to delete most of the nonsignificant variables and achieve the goal of dimension reduction. (iv) As a gold standard, the oracle estimators give the perfect values of $\mathrm{TN}_{\boldsymbol{\theta}}$, $\mathrm{TP}_{\boldsymbol{\theta}}$, $\mathrm{TN}_{\boldsymbol{\beta}}$, $\mathrm{TP}_{\boldsymbol{\beta}}$.

From Table 2, one may have the following observations. (i) All the cases indicate that the SIMEX method reduce the biases observed in the Naive method and the SIMEX+SCAD

Table 1: Simulation results over 500 replications when $\sigma_u = \sigma_v = 0.2$ (the values in the parentheses are the corresponding standard errors).

n	Method	$R^2_{\boldsymbol{\theta}}$	$\mathrm{TN}_{\boldsymbol{\theta}}$	$\mathrm{TP}_{\boldsymbol{\theta}}$	$R^2_{\boldsymbol{\beta}}$	$\mathrm{TN}_{\boldsymbol{\beta}}$	$\mathrm{TP}_{\boldsymbol{\beta}}$	RASE
100	SIMEX	0.9710(0.0208)	0.028	2	1.0089(0.0692)	0.072	3	0.3534(0.1895)
	SIMEX+SCAD	0.9809(0.0270)	2.828	2	1.0001(0.0530)	6.998	3	0.3258(0.2158)
	Naive	0.9513(0.0416)	0.026	2	0.9467(0.0520)	0.070	3	0.4135(0.1484)
	Naive+SCAD	0.9884(0.0194)	2.936	2	0.9576(0.0449)	7.000	3	0.4094(0.1080)
	Oracle	0.9989(0.0018)	3.000	2	0.9994(0.0168)	7.000	3	0.1181(0.0764)
200	SIMEX	0.9878(0.0091)	0.038	2	1.0094(0.0428)	0.088	3	0.2623(0.1178)
	SIMEX+SCAD	0.9900(0.0143)	2.978	2	1.0043(0.0364)	7.000	3	0.2476(0.0971)
	Naive	0.9772(0.0168)	0.030	2	0.9521(0.0354)	0.138	3	0.3924(0.0752)
	Naive+SCAD	0.9944(0.0080)	2.996	2	0.9597(0.0323)	7.000	3	0.3981(0.0663)
	Oracle	0.9994(0.0008)	3.000	2	0.9999(0.0118)	7.000	3	0.0873(0.0477)
400	SIMEX	0.9948(0.0036)	0.066	2	0.9979(0.0289)	0.172	3	0.2270(0.0740)
	SIMEX+SCAD	0.9955(0.0063)	3.000	2	0.9975(0.0256)	7.000	3	0.2253(0.0771)
	Naive	0.9896(0.0080)	0.040	2	0.9462(0.0244)	0.156	3	0.3915(0.0505)
	Naive+SCAD	0.9975(0.0039)	3.000	2	0.9552(0.0222)	7.000	3	0.3975(0.0474)
	Oracle	0.9997(0.0004)	3.000	2	0.9992(0.0086)	7.000	3	0.0689(0.0354)

Table 2: Biases and Standard errors of the estimators of nonzero coefficients when $\sigma_u = \sigma_v = 0.2$ (the values in the parentheses are the corresponding standard errors)

n	Method	θ_4	θ_5	β_1	β_2	β_3
100	SIMEX	−0.0080(0.0603)	−0.0128(0.0606)	0.0036(0.0910)	0.0012(0.0961)	0.0067(0.0980)
	SIMEX+SCAD	−0.0058(0.0991)	−0.0080(0.0977)	0.0033(0.0847)	−0.0070(0.1102)	0.0027(0.0867)
	Naive	−0.0142(0.0831)	−0.0210(0.0841)	−0.0256(0.0691)	−0.0202(0.0739)	−0.0364(0.0721)
	Naive+SCAD	−0.0016(0.0766)	−0.0067(0.0767)	−0.0246(0.0664)	−0.0179(0.0725)	−0.0225(0.0642)
	Oracle	0.0000(0.0235)	−0.0007(0.0236)	−0.0003(0.0231)	−0.0009(0.0258)	0.0002(0.0232)
200	SIMEX	−0.0032(0.0405)	−0.0054(0.0408)	0.0029(0.0539)	0.0067(0.0613)	−0.0379(0.0607)
	SIMEX+SCAD	−0.0089(0.0716)	0.0018(0.0699)	0.0012(0.0500)	0.0056(0.0566)	−0.0270 (0.0491)
	Naive	−0.0112(0.0556)	−0.0050(0.0551)	−0.0257(0.0457)	−0.0096(0.0499)	0.0038(0.0489)
	Naive+SCAD	−0.0063(0.0531)	0.0023(0.0526)	−0.0255(0.0434)	−0.0089(0.0476)	−0.0008(0.0425)
	Oracle	−0.0004(0.0167)	0.0000(0.0166)	−0.0012(0.0165)	0.0008(0.0186)	0.0002 (0.0161)
400	SIMEX	−0.0033(0.0259)	−0.0005(0.0259)	0.0002(0.0371)	0.0009(0.0410)	−0.0045(0.0391)
	SIMEX+SCAD	−0.0062(0.0474)	0.0030(0.0468)	0.0000(0.0355)	0.0005(0.0392)	−0.0045(0.0336)
	Naive	−0.0068(0.0364)	−0.0006(0.0361)	−0.0272(0.0317)	−0.0130(0.0336)	−0.0419(0.0330)
	Naive+SCAD	−0.0043(0.0357)	0.0026(0.0350)	−0.0263(0.0310)	−0.0131(0.0328)	−0.0288(0.0292)
	Oracle	−0.0010(0.0119)	0.0008(0.0119)	0.0003(0.0111)	−0.0002(0.0123)	−0.0013(0.0115)

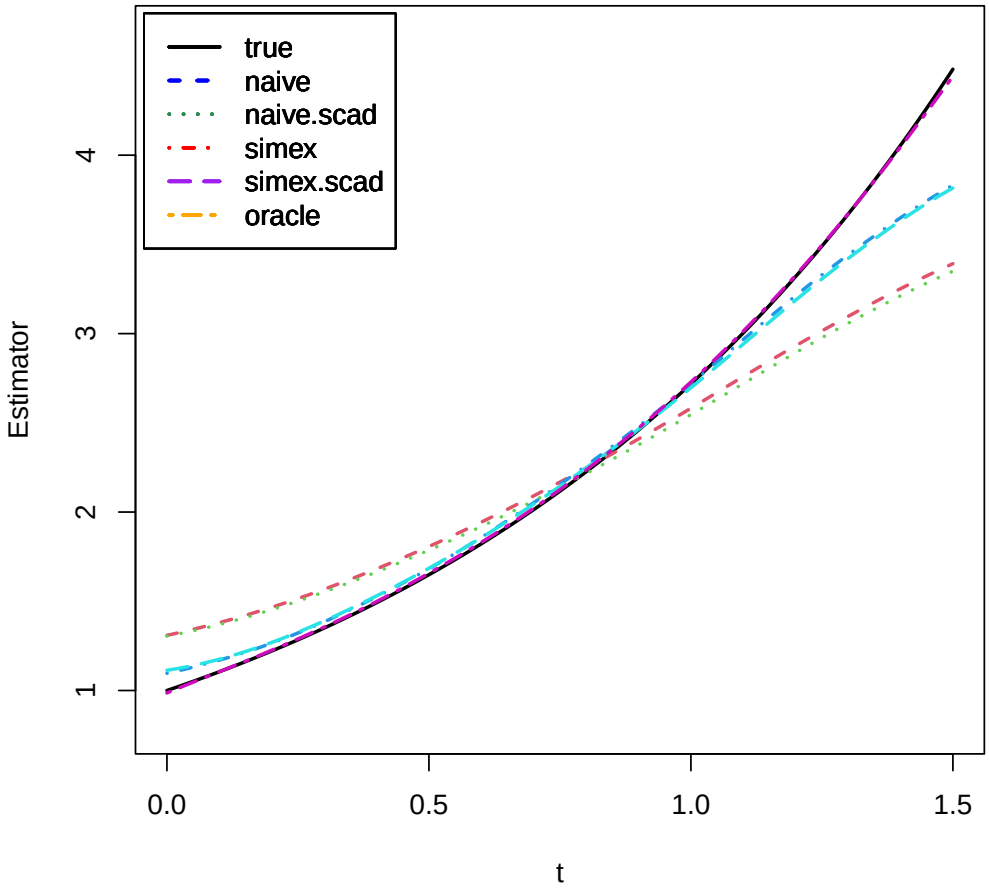

Figure 1: When $\sigma_u = \sigma_v = 0.2$ and $n = 200$, the nonparametric estimates of the link function $g(t)$.

is as good as Oracle. (ii) the bias and SE decrease as n increases. (iii) However, the standard errors based on the SIMEX method are larger than those based on the naive method. An intuitive explanation can be found in Yang et al (2019),

The nonparametric estimates of the link function $g(t)$ with $\sigma_u = \sigma_v = 0.2$ and $n = 200$ are provided in Figure 1, and other cases are similar. From Figure 1, we see that the estimated SIMEX curves are closer to the real link function curves than the estimated naive curves. The SE of the SIMEX and naive estimators for the link function are not large, but the SE of the SIMEX estimators are slightly larger than the naive estimators.

In summary, the proposed method performs well in both variable selection and parameter estimation.

5. CONCLUSION

We studied variable selection and estimation in the partially linear single-index model with measurement errors in all variables. By using local linear regression, SIMEX technique and profile least square method, the SCAD penalty is successfully introduced to achieve efficient selection of variables. The introduction of SCAD penalty term not only performs well in variable selection, but also achieves remarkable results in estimating parameters and non-parametric link function. When the regularity condition is satisfied, the obtained estimators have oracle property. Our research provides an innovative solution to the challenge of measurement errors in multivariate analysis.

In this paper we do not consider the cases of the response variables are missing and more complex cases, for example, the partially nonlinear single-index measurement error models or partially linear multiple-index measurement error models, etc. These are further studies and beyond the scope of this paper.

ACKNOWLEDGMENTS

This work was supported by the Humanities and Social Sciences Research Projects of the Ministry of Education of China (20YJC910010), Henan Province "double first-class" discipline establishment engineering cultivation project (GCCYJ202428), and the Doctoral Foundation of Henan Polytechnic University (B2020-37). We also acknowledge the valuable suggestions from the referees.

REFERENCES

Cai, L. and Wang, S. (2023). Simultaneous confidence bands and global inferences for extended partially linear single-index models. *Applied Mathematical Modelling*, 113:30–43.

Carroll, R., Fan, J., Gijbels, I. and Wand, M. (1997). Generalized partially linear single-index models. *Journal of the American Statistical Association*, 92(438):477–489.

Carroll, R., Ruppert, D., Stefanski, L. and Crainiceanu, C. (2006). *Measurement Error in Nonlinear Model*, 2nd edn. Chapman & Hall, London.

Chen, X. and Cui, H. (2009). Empirical likelihood for partially linear single-index errors-in-variables model. *Communications in Statistics-Theory and Methods*, 38(15):2498–2514.

Cook, J. and Stefanski, L. (1994). Simulation-extrapolation estimation in parametric measurement error models. *Journal of the American Statistical Association*, 89(428):1314–1328.

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

Fan, J. and Li, R. (2004). New estimation and model selection procedures for semiparametric modeling in longitudinal data analysis. *Journal of the American Statistical Association*, 99(467):710–723.

Frank, I. and Friedman, J. (1993). A statistical view of some chemometrics regression tools. (with discussion). *Technometrics*, 35(2):109–135.

Härdle, W., Hall, P. and Ichimura, H. (1993). Optimal smoothing in single-index models. *The Annals of Statistics*, 21(1):157–178.

Huang, Z. and Zhao, X. (2019). Statistical estimation for a partially linear single-index model with errors in all variables. *Communications in Statistics-Theory and Methods*, 48(5):1136–1148.

Liang, H. and Carroll, R. (1999). Estimation in a semiparametric partially linear errors-in-variables model, *The Annals of Statistics*, 27(5):1519–1535.

Liang, H. and Li, R. (2009). Variable selection for partially linear models with measurement errors. *Journal of the American Statistical Association*, 104(485):234–248.

Liang, H., Liu, X., Li, R. and Tsai, C. (2010). Estimation and testing for partially linear single-index models. *The Annals of Statistics*, 38(6):3811–3836.

Liang, H. and Wang, N. (2005). Partially linear single-index measurement error models. *Statistica Sinica*, 15(1):99–116.

Lin, H., Shi, J., Tong, T. and Zhang, R. (2022). Partially linear single-index model in the presence of measurement error. *Journal of Systems Science and Complexity*, 35(6):2361–2380.

Liu, Y., Ren, M. and Zhang, S. (2021). Empirical likelihood test for regression coefficients in high dimensional partially linear models. *Journal of Systems Science and Complexity*, 34(3):1135–1155.

Liu, Y., Zou, J. H, Zhao, S. and Yang, Q. (2022). Model averaging estimation for varying-coefficient single-index models. *Journal of Systems Science and Complexity*, 35(1):264–282.

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

Wang, J., Xue, L., Zhu, L. and Chong, Y. (2010). Estimation for a partially linear single-index model. *The Annals of Statistics*, 38(1):246–274.

Xia, Y. and Härdle, W. (2006). Semi-parametric estimation of partially linear single-index models. *Journal of Multivariate Analysis*, 97(5):1162–1184.

Yan, X., Pu, X., Zhou, Y. and Xun, X. (2020). Convergence rate of principal component analysis with local linear smoother for functional data under a unified weighing scheme. *Statistical Theory and Related Fields*, 4(1):55–65.

Yang, Y., Tong, T. and Li, G. (2019). Simex estimation for single-index model with covariate measurement error. *AStA-Advances in Statistical Analysis*, 103(1):137–161.

Yu, Y. and Ruppert, D. (2002). Penalized spline estimation for partially linear single-index models. *Journal of the American Statistical Association*, 97(460):1042–1054.

Zhu, L. and Xue, L. (2006). Empirical likelihood confidence regions in a partially linear single-index model. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 68(3):549–570.

Zou, H. (2006). The adaptive LASSO and its oracle properties. *Journal of the American Statistical Association*, 101(476):1418–1429.

APPENDIX

Proof of Theorem 2.1

Proof: Let $\gamma_n = n^{-1/2}+a_n+c_n$, $\boldsymbol{v}_1 = (v_{11}, \cdots, v_{1p})^\top$, $\boldsymbol{v}_2 = (v_{21}, \cdots, v_{2q})^\top$. We want to show that for any given $\epsilon > 0$, there exists a large constant C that satisfies $\|\boldsymbol{v}_1\| = \|\boldsymbol{v}_2\| = C$ such that

$$P\left\{\inf_{\|\boldsymbol{v}_1\|=\|\boldsymbol{v}_2\|=C} \mathcal{L}(\boldsymbol{\theta}_0 + \gamma_n\boldsymbol{v}_1, \boldsymbol{\beta}_0 + \gamma_n\boldsymbol{v}_2) > \mathcal{L}(\boldsymbol{\theta}_0, \boldsymbol{\beta}_0)\right\} \geq 1 - \epsilon. \tag{0.1}$$

This implies with probability at least $1-\epsilon$ that there exists a local minimum in the ball $\{\boldsymbol{\theta}_0 + \gamma_n\boldsymbol{v}_1, \boldsymbol{\beta}_0 + \gamma_n\boldsymbol{v}_2 : \|\boldsymbol{v}_1\| = \|\boldsymbol{v}_2\| \leq C\}$. Hence, there exists a local minimum such that the rate of convergence of $(\boldsymbol{\theta}^\top, \boldsymbol{\beta}^\top)^\top$ is $O_p(n^{-1/2} + a_n + c_n)$. Let

$$Q_1(\boldsymbol{\theta}, \boldsymbol{\beta}) = \sum_{i=1}^{n}\left(\tilde{Y}_i - \hat{g}'(\boldsymbol{W}_i^\top\hat{\boldsymbol{\theta}}; \boldsymbol{\theta}, \boldsymbol{\beta})\boldsymbol{W}_i^\top\boldsymbol{\theta} - \boldsymbol{T}_i^\top\boldsymbol{\beta}\right)^2 - n\bar{g}'^2(\Lambda)\boldsymbol{\theta}^\top\Sigma_u\boldsymbol{\theta} - n\boldsymbol{\beta}^\top\Sigma_v\boldsymbol{\beta}, \tag{0.2}$$

where $\tilde{Y}_i = Y_i - \hat{g}(\boldsymbol{W}_i^\top\hat{\boldsymbol{\theta}}) + \hat{g}'(\boldsymbol{W}_i^\top\hat{\boldsymbol{\theta}})\boldsymbol{W}_i^\top\hat{\boldsymbol{\theta}}$, where $\hat{\boldsymbol{\theta}}$ and $\hat{g}'$ are part of the SIMEX estimator. Note that (0.2) is an approximation of (2.4) by first-order Taylor expansion. Denote

$$\begin{aligned}
D_{n,1} =& Q_1(\boldsymbol{\theta}_0 + \gamma_n\boldsymbol{v}_1, \boldsymbol{\beta}_0 + \gamma_n\boldsymbol{v}_2) - Q_1(\boldsymbol{\theta}_0, \boldsymbol{\beta}_0)\\
=& \sum_{i=1}^{n}\left\{\left(\tilde{Y}_i - \hat{g}'(\boldsymbol{W}_i^\top\hat{\boldsymbol{\theta}})\boldsymbol{W}_i^\top(\boldsymbol{\theta}_0 + \gamma_n\boldsymbol{v}_1) - \boldsymbol{T}_i^\top(\boldsymbol{\beta}_0 + \gamma_n\boldsymbol{v}_2)\right)^2 - \left(\tilde{Y}_i - \hat{g}'(\boldsymbol{W}_i^\top\hat{\boldsymbol{\theta}})\boldsymbol{W}_i^\top\boldsymbol{\theta}_0 - \boldsymbol{T}_i^\top\boldsymbol{\beta}_0\right)^2\right\}\\
& - n\bar{g}'^2(\Lambda)\left\{(\boldsymbol{\theta}_0 + \gamma_n\boldsymbol{v}_1)^\top\Sigma_u(\boldsymbol{\theta}_0 + \gamma_n\boldsymbol{v}_1) - \boldsymbol{\theta}_0^\top\Sigma_u\boldsymbol{\theta}_0\right\}\\
& - n\left\{(\boldsymbol{\beta}_0 + \gamma_n\boldsymbol{v}_2)^\top\Sigma_v(\boldsymbol{\beta}_0 + \gamma_n\boldsymbol{v}_2) - \boldsymbol{\beta}_0^\top\Sigma_v\boldsymbol{\beta}_0\right\}\\
=& \sum_{i=1}^{n}\left\{\left(\hat{g}'(\boldsymbol{W}_i^\top\hat{\boldsymbol{\theta}})\boldsymbol{W}_i^\top\boldsymbol{v}_1\gamma_n + \boldsymbol{T}_i^\top\boldsymbol{v}_2\gamma_n\right)^2 - 2\left(\hat{g}'(\boldsymbol{W}_i^\top\hat{\boldsymbol{\theta}})\boldsymbol{W}_i^\top\boldsymbol{v}_1\gamma_n + \boldsymbol{T}_i^\top\boldsymbol{v}_2\gamma_n\right)\varepsilon_i\right\}\\
& - n\left\{\gamma_n^2\left(\bar{g}'^2(\Lambda)\boldsymbol{v}_1^\top\Sigma_u\boldsymbol{v}_1 + \boldsymbol{v}_2^\top\Sigma_v\boldsymbol{v}_2\right) + 2\gamma_n\left(\bar{g}'^2(\Lambda)\boldsymbol{\theta}_0^\top\Sigma_u\boldsymbol{v}_1 + \boldsymbol{\beta}_0^\top\Sigma_v\boldsymbol{v}_2\right)\right\} + o_p(1)
\end{aligned} \tag{0.3}$$

and

$$D_{n,2} = n\sum_{j=1}^{s}\left\{p_{\lambda_1}(|\boldsymbol{\theta}_{j0} + \gamma_n\boldsymbol{v}_{1j}|) - p_{\lambda_1}(|\boldsymbol{\theta}_{j0}|)\right\} + n\sum_{k=1}^{t}\left\{p_{\lambda_2}(|\boldsymbol{\beta}_{k0} + \gamma_n\boldsymbol{v}_{2k}|) - p_{\lambda_2}(|\boldsymbol{\beta}_{k0}|)\right\}. \tag{0.4}$$

Moreover, applying the Taylor expansion and the Cauchy – Schwarz inequality, we are able to show that $n^{-1}D_{n,2}$ is bounded by

$$\sqrt{s}\gamma_n a_n\|\boldsymbol{v}_1\| + \gamma_n^2 b_n\|\boldsymbol{v}_1\|^2 + \sqrt{t}\gamma_n c_n\|\boldsymbol{v}_2\| + \gamma_n^2 d_n\|\boldsymbol{v}_2\|^2 \le C\gamma_n^2(\sqrt{s} + b_n C + \sqrt{t} + d_n C).$$

When b_n and d_n tend to 0 and C is sufficiently large, the second term on the right-hand side of $D_{n,1}$ in (0.3) dominates the first term uniformly in $\|\boldsymbol{v}_1\| = \|\boldsymbol{v}_2\| = C$. And $D_{n,2}$ in (0.4) is also dominated by the second term of $D_{n,1}$ in (0.3). Hence, by choosing a sufficiently large C, (0.1) holds. This completes the proof of the theorem. □

Proof of Theorem 2.2

Proof: To prove Theorem 2.2, we now first prove the sparsity. It is sufficient to show that with probability tending to 1 as $n \to \infty$, for any $\boldsymbol{\theta}_1$ and $\boldsymbol{\beta}_1$ satisfy $\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_{10}\| = O_p(n^{-1/2})$ and $\|\boldsymbol{\beta}_1 - \boldsymbol{\beta}_{10}\| = O_p(n^{-1/2})$, respectively. We next show that

$$\mathcal{L}\left\{\begin{pmatrix}\boldsymbol{\theta}_1\\0\end{pmatrix}, \begin{pmatrix}\boldsymbol{\beta}_1\\0\end{pmatrix}\right\} = \min_{(\boldsymbol{\theta}_2,\boldsymbol{\beta}_2)\in C} \mathcal{L}\left\{\begin{pmatrix}\boldsymbol{\theta}_1\\\boldsymbol{\theta}_2\end{pmatrix}, \begin{pmatrix}\boldsymbol{\beta}_1\\\boldsymbol{\beta}_2\end{pmatrix}\right\}, \tag{0.5}$$

where $C = \left\{\|\boldsymbol{\theta}_2\| \le C^* n^{-1/2}, \|\boldsymbol{\beta}_2\| \le C^* n^{-1/2}\right\}$ and C^* is a positive constant.

Consider $\beta_k \in \left(-C^* n^{-1/2}, C^* n^{-1/2}\right)$ for $k = t+1, \cdots, q$. When $\beta_k \neq 0$, we have

$$\frac{\partial \mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\beta})}{\partial \beta_k} = l_k(\boldsymbol{\theta}, \boldsymbol{\beta}) + n p'_{\lambda 2}(|\beta_k|)\mathrm{sgn}(\beta_k),$$

where

$$\begin{aligned} l_k(\boldsymbol{\theta}, \boldsymbol{\beta}) &= \sum_{i=1}^{n} \left\{-2\left(Y_i - \hat{g}(\boldsymbol{W}_i^\top \boldsymbol{\theta}; \boldsymbol{\theta}, \boldsymbol{\beta}) - \boldsymbol{T}_i^\top \boldsymbol{\beta}\right)\left(\boldsymbol{T}_{ik}^\top + \frac{\partial \hat{g}(\boldsymbol{W}_i^\top \boldsymbol{\theta}; \boldsymbol{\theta}, \boldsymbol{\beta})}{\partial \beta_k}\right) - \Sigma_v \boldsymbol{\beta}\right\} \\ &= \sum_{i=1}^{n} \left\{2\left(\hat{g}(\boldsymbol{W}_i^\top \boldsymbol{\theta}; \boldsymbol{\theta}, \boldsymbol{\beta}) - g(\boldsymbol{W}_i^\top \boldsymbol{\theta}_0) + \boldsymbol{T}_i^\top(\boldsymbol{\beta} - \boldsymbol{\beta}_0) - \varepsilon_i\right)\left(\boldsymbol{T}_{ik}^\top + \frac{\partial \hat{g}(\boldsymbol{W}_i^\top \boldsymbol{\theta}; \boldsymbol{\theta}, \boldsymbol{\beta})}{\partial \beta_k}\right) - \Sigma_v \boldsymbol{\beta}\right\}. \end{aligned}$$

Using the assumptions that $\|\boldsymbol{\theta} - \boldsymbol{\theta}_0\| = O_p(n^{-1/2})$ and $\|\boldsymbol{\beta} - \boldsymbol{\beta}_0\| = O_p(n^{-1/2})$, we have that $n^{-1} l_k(\boldsymbol{\theta}, \boldsymbol{\beta})$ is of the order $O_p(n^{-1/2})$. Therefore,

$$\frac{\partial \mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\beta})}{\partial \beta_k} = n\lambda_2\left\{\lambda_2^{-1} p'_{\lambda 2}(|\beta_k|)\mathrm{sgn}(\beta_k) + O_p(n^{-1/2}/\lambda_2)\right\}.$$

Because of $\liminf_{n\to\infty} \liminf_{\beta_k \to 0^+} \lambda_2^{-1} p'_{\lambda 2}(|\beta_k|) > 0$ and $n^{-1/2}/\lambda_2 \to 0$, the sign of $\partial \mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\beta})/\partial \beta_k$ is solely determined by the sign of $\beta_k \in \left(-C^* n^{-1/2}, C^* n^{-1/2}\right)$ and hence the signs are the same.

Analogously, we can show that $\partial \mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\beta})/\partial \theta_j$ and θ_j have same signs when $\theta_j \in \left(-C^* n^{-1/2}, C^* n^{-1/2}\right)$ for $j = s+1, \cdots, p$. Consequently, the minimum is attained at $\boldsymbol{\theta}_2 = 0$ and $\boldsymbol{\beta}_2 = 0$. This completes the proof of (0.5).

We now demonstrate the asymptotic normality of $\hat{\boldsymbol{\theta}}_1$ and $\hat{\boldsymbol{\beta}}_1$ given that sparsity holds. It follows from (2.5) that $\hat{\boldsymbol{\theta}}_1$ and $\hat{\boldsymbol{\beta}}_1$ satisfy

$$\begin{pmatrix}\frac{\partial \mathcal{L}(\hat{\boldsymbol{\theta}}_1, \hat{\boldsymbol{\beta}}_1)}{\partial \boldsymbol{\theta}_1}\\ \frac{\partial \mathcal{L}(\hat{\boldsymbol{\theta}}_1, \hat{\boldsymbol{\beta}}_1)}{\partial \boldsymbol{\beta}_1}\end{pmatrix} = l(\hat{\boldsymbol{\theta}}_1, \hat{\boldsymbol{\beta}}_1) + \begin{pmatrix}\boldsymbol{R}_1 - \Sigma_1(\hat{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}_{10})\\ \boldsymbol{R}_2 - \Sigma_2(\hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_{10})\end{pmatrix} = 0, \tag{0.6}$$

where

$$l(\hat{\boldsymbol{\theta}}_1,\hat{\boldsymbol{\beta}}_1) =\frac{1}{n}\sum_{i=1}^{n}\begin{pmatrix} g'(\boldsymbol{W}_{i,1}^\top\boldsymbol{\theta}_{10})\boldsymbol{W}_{i,1} \\ \boldsymbol{T}_{i,1}\end{pmatrix}\left\{\varepsilon_i-\left(\hat{g}(\boldsymbol{W}_{i,1}^\top\hat{\boldsymbol{\theta}}_1)-g(\boldsymbol{X}_{i,1}^\top\boldsymbol{\theta}_{10})\right)-\boldsymbol{T}_{i,1}^\top(\hat{\boldsymbol{\beta}}_1-\boldsymbol{\beta}_{10})-\boldsymbol{V}_{i,1}^\top\boldsymbol{\beta}_{10}\right\}$$

(0.7)

$$+\begin{pmatrix}\bar{g}'^2(\Lambda)\Sigma_{u11}\hat{\boldsymbol{\theta}}_1 \\ \Sigma_{v11}\hat{\beta}_1\end{pmatrix}.$$

We have the asymptotic expansion of $\hat{g}(t;\hat{\boldsymbol{\theta}},\hat{\boldsymbol{\beta}})$ as

$$\begin{aligned}&\hat{g}(t;\hat{\boldsymbol{\theta}},\hat{\boldsymbol{\beta}})-g(t)\\ =\;&\frac{1}{nf(t)}\sum_{i=1}^{n}K_h(\boldsymbol{W}_i^\top\boldsymbol{\theta}-t)(Y_i-\boldsymbol{T}_i^\top\boldsymbol{\beta}-g(\boldsymbol{W}_i^\top\boldsymbol{\theta}))\\ &-(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})^\top E[\boldsymbol{T}|\boldsymbol{W}^\top\boldsymbol{\theta}=t]\\ &-(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta})^\top E[g'(t)\boldsymbol{W}|\boldsymbol{W}^\top\boldsymbol{\theta}=t]+o_p(n^{-1/2})+o_p(h^2).\end{aligned}\tag{0.8}$$

Note that the second and third terms on the right-hand side of (0.8) will converge to zero faster than the first term, provided that the $\boldsymbol{\theta}$ and $\boldsymbol{\beta}$ are consistent in a rate faster than $\sqrt{nh}+h^2$, where h is the bandwidth used in estimating g. This property has been used by Carroll et al (1997) in their proofs implicitly. Following (0.7) and (0.8), we can further derive that

$$\begin{aligned}l(\hat{\boldsymbol{\theta}}_1,\hat{\boldsymbol{\beta}}_1) =&\frac{1}{n}\sum_{i=1}^{n}\begin{pmatrix} g'(\boldsymbol{W}_{i,1}^\top\boldsymbol{\theta}_{10})\boldsymbol{W}_{i,1} \\ \boldsymbol{T}_{i,1}\end{pmatrix}\Big\{(\varepsilon_i-g'(\boldsymbol{X}_{i,1}^\top\boldsymbol{\theta}_{10})\boldsymbol{U}_{i,1}^\top\boldsymbol{\theta}_{10}-\boldsymbol{V}_{i,1}^\top\boldsymbol{\beta}_{10})\\ &-\frac{1}{nf(\boldsymbol{W}_{i,1}^\top\boldsymbol{\theta}_{10})}\sum_{j=1}^{n}K_h(\boldsymbol{W}_{j,1}^\top\boldsymbol{\theta}_{10}-\boldsymbol{W}_{i,1}^\top\boldsymbol{\theta}_{10})\{Y_j-\boldsymbol{T}_{j,1}^\top\boldsymbol{\beta}_{10}-g(\boldsymbol{W}_{j,1}^\top\boldsymbol{\theta}_{10})\}\end{aligned}$$

(0.9)

$$-g'(\boldsymbol{W}_{i,1}^\top\boldsymbol{\theta}_{10})\tilde{\boldsymbol{W}}_{i,1}^\top(\hat{\boldsymbol{\theta}}_1-\boldsymbol{\theta}_{10})-\tilde{\boldsymbol{T}}_{i,1}^\top(\hat{\boldsymbol{\beta}}_1-\boldsymbol{\beta}_{10})\Big\}+\begin{pmatrix}\bar{g}'^2(\Lambda)\Sigma_{u11}\hat{\boldsymbol{\theta}}_1 \\ \Sigma_{v11}\hat{\boldsymbol{\beta}}_1\end{pmatrix}+o_p(n^{-\frac{1}{2}})+o_p(h^2),$$

By interchanging the summations, kernel density estimation and Taylor expansion, the second terms of the right-hand side equals

$$\frac{1}{n}\sum_{i=1}^{n}\left(Y_i-\boldsymbol{T}_{i,1}^\top\boldsymbol{\beta}_{10}-g(\boldsymbol{W}_{i,1}^\top.\boldsymbol{\theta}_{10})\right)\frac{1}{n}\sum_{j=1}^{n}\begin{pmatrix} g'(\boldsymbol{W}_{j,1}^\top\boldsymbol{\theta}_{10})\boldsymbol{W}_{j,1} \\ \boldsymbol{T}_{j,1}\end{pmatrix}\frac{K_h(\boldsymbol{W}_{j,1}^\top\boldsymbol{\theta}_{10}-\boldsymbol{W}_{i,1}^\top\boldsymbol{\theta}_{10})}{f(\boldsymbol{W}_{j,1}^\top\boldsymbol{\theta}_{10})},$$

which is asymptotically equivalent to

$$\begin{aligned}&\frac{1}{n}\sum_{i=1}^{n}\begin{pmatrix} E[g'(\boldsymbol{W}^\top\boldsymbol{\theta})\boldsymbol{W}|\boldsymbol{W}^\top\boldsymbol{\theta}=\boldsymbol{W}_{i,1}^\top\boldsymbol{\theta}_{10}] \\ E[\boldsymbol{T}|\boldsymbol{W}^\top\boldsymbol{\theta}=\boldsymbol{W}_{i,1}^\top\boldsymbol{\theta}_{10}]\end{pmatrix}\left\{\varepsilon_i-g'(\boldsymbol{X}_{i,1}^\top\boldsymbol{\theta}_{10})\boldsymbol{U}_{i,1}^\top\boldsymbol{\theta}_{10}-\boldsymbol{V}_{i,1}^\top\boldsymbol{\beta}_{10}\right\}\\ &\left\{1+O_p(h^2)+o_p([nh]^{-1/2})\right\}.\end{aligned}\tag{0.10}$$

A combination of (0.9) and (0.10) yields

$$\begin{aligned}l(\hat{\boldsymbol{\theta}}_1,\hat{\boldsymbol{\beta}}_1) =&\frac{1}{n}\sum_{i=1}^{n}\begin{pmatrix} g'(\boldsymbol{W}_{i,1}^\top\boldsymbol{\theta}_{10})\boldsymbol{W}_{i,1} \\ \boldsymbol{T}_{i,1}\end{pmatrix}\left(\varepsilon_i-g'(\boldsymbol{X}_{i,1}^\top\boldsymbol{\theta}_{10})\boldsymbol{U}_{i,1}^\top\boldsymbol{\theta}_{10}-\boldsymbol{V}_{i,1}^\top\boldsymbol{\beta}_{10}\right)\\ &-\frac{1}{n}\sum_{i=1}^{n}\begin{pmatrix} E[g'(\boldsymbol{W}^\top\boldsymbol{\theta})\boldsymbol{W}|\boldsymbol{W}^\top\boldsymbol{\theta}=\boldsymbol{W}_{i,1}^\top\boldsymbol{\theta}_{10}] \\ E[\boldsymbol{T}|\boldsymbol{W}^\top\boldsymbol{\theta}=\boldsymbol{W}_{i,1}^\top\boldsymbol{\theta}_{10}]\end{pmatrix}\left(\varepsilon_i-g'(\boldsymbol{X}_{i,1}^\top\boldsymbol{\theta}_{10})\boldsymbol{U}_{i,1}^\top\boldsymbol{\theta}_{10}-\boldsymbol{V}_{i,1}^\top\boldsymbol{\beta}_{10}\right)\\ &-\frac{1}{n}\sum_{i=1}^{n}\begin{pmatrix} g'(\boldsymbol{W}_{i,1}^\top\boldsymbol{\theta}_{10})\boldsymbol{W}_{i,1} \\ \boldsymbol{T}_{i,1}\end{pmatrix}\left\{g'(\boldsymbol{W}_{i,1}^\top\boldsymbol{\theta}_{10})\tilde{\boldsymbol{W}}_{i,1}^\top(\hat{\boldsymbol{\theta}}_1-\boldsymbol{\theta}_{10})-\tilde{\boldsymbol{T}}_{i,1}^\top(\hat{\boldsymbol{\beta}}_1-\boldsymbol{\beta}_{10})\right\}+\begin{pmatrix}\bar{g}'^2(\Lambda)\Sigma_{u11}\hat{\boldsymbol{\theta}}_1 \\ \Sigma_{v11}\hat{\boldsymbol{\beta}}_1\end{pmatrix}.\end{aligned}\tag{0.11}$$

It follows that

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}\begin{pmatrix} g'(\boldsymbol{W}_{i,1}^{\top}\boldsymbol{\theta}_{10})\boldsymbol{W}_{i,1} \\ \boldsymbol{T}_{i,1}\end{pmatrix}\left\{g'(\boldsymbol{W}_{i,1}^{\top}\boldsymbol{\theta}_{10})\tilde{\boldsymbol{W}}_{i,1}^{\top}(\hat{\boldsymbol{\theta}}_1-\boldsymbol{\theta}_{10})-\tilde{\boldsymbol{T}}_{i,1}^{\top}(\hat{\boldsymbol{\beta}}_1-\boldsymbol{\beta}_{10})\right\}-\begin{pmatrix}\bar{g}'^2(\Lambda)\Sigma_{u11}\hat{\boldsymbol{\theta}}_1\\ \Sigma_{v11}\hat{\boldsymbol{\beta}}_1\end{pmatrix}\\
=&\frac{1}{n}\sum_{i=1}^{n}\begin{pmatrix} g'(\boldsymbol{W}_{i,1}^{\top}\boldsymbol{\theta}_{10})\boldsymbol{W}_{i,1} \\ \boldsymbol{T}_{i,1}\end{pmatrix}\begin{pmatrix} g'(\boldsymbol{W}_{i,1}^{\top}\boldsymbol{\theta}_{10})\tilde{\boldsymbol{W}}_{i,1} & \tilde{\boldsymbol{T}}_{i,1}\end{pmatrix}\begin{pmatrix}\hat{\boldsymbol{\theta}}_1-\boldsymbol{\theta}_{10}\\ \hat{\boldsymbol{\beta}}_1-\boldsymbol{\beta}_{10}\end{pmatrix}\\
&-\begin{pmatrix}\bar{g}'^2(\Lambda)\Sigma_{u11} & 0\\ 0 & \Sigma_{v11}\end{pmatrix}\begin{pmatrix}\hat{\boldsymbol{\theta}}_1-\boldsymbol{\theta}_{10}\\ \hat{\boldsymbol{\beta}}_1-\boldsymbol{\beta}_{10}\end{pmatrix}-\begin{pmatrix}\bar{g}'^2(\Lambda)\Sigma_{u11}\boldsymbol{\theta}_{10}\\ \Sigma_{v11}\boldsymbol{\beta}_{10}\end{pmatrix}\\
=&\frac{1}{n}\sum_{i=1}^{n}\begin{pmatrix} g'^2(\boldsymbol{W}_{i,1}^{\top}\boldsymbol{\theta}_{10})\boldsymbol{W}_{i,1}\tilde{\boldsymbol{W}}_{i,1}^{\top}-\bar{g}'^2(\Lambda)\Sigma_{u11} & g'(\boldsymbol{W}_{i,1}^{\top}\boldsymbol{\theta}_{10})\boldsymbol{W}_{i,1}\tilde{\boldsymbol{T}}_{i,1}^{\top}\\ g'(\boldsymbol{W}_{i,1}^{\top}\boldsymbol{\theta}_{10})\boldsymbol{T}_{i,1}\tilde{\boldsymbol{W}}_{i,1}^{\top} & \boldsymbol{T}_{i,1}\tilde{\boldsymbol{T}}_{i,1}^{\top}-\Sigma_{v11}\end{pmatrix}\begin{pmatrix}\hat{\boldsymbol{\theta}}_1-\boldsymbol{\theta}_{10}\\ \hat{\boldsymbol{\beta}}_1-\boldsymbol{\beta}_{10}\end{pmatrix}\\
&-\begin{pmatrix}\bar{g}'^2(\Lambda)\Sigma_{u11}\boldsymbol{\theta}_{10}\\ \Sigma_{v11}\boldsymbol{\beta}_{10}\end{pmatrix}.
\end{aligned}$$

So, when the LLN is applied, we have

$$\begin{aligned}
l(\hat{\boldsymbol{\theta}}_1,\hat{\boldsymbol{\beta}}_1)=&-\frac{1}{n}\sum_{i=1}^{n}\begin{pmatrix} g'^2(\boldsymbol{W}_{i,1}^{\top}\boldsymbol{\theta}_{10})\boldsymbol{W}_{i,1}\tilde{\boldsymbol{W}}_{i,1}^{\top}-\bar{g}'^2(\Lambda)\Sigma_{u11} & g'(\boldsymbol{W}_{i,1}^{\top}\boldsymbol{\theta}_{10})\boldsymbol{W}_{i,1}\tilde{\boldsymbol{T}}_{i,1}^{\top}\\ g'(\boldsymbol{W}_{i,1}^{\top}\boldsymbol{\theta}_{10})\boldsymbol{T}_{i,1}\tilde{\boldsymbol{W}}_{i,1}^{\top} & \boldsymbol{T}_{i,1}\tilde{\boldsymbol{T}}_{i,1}^{\top}-\Sigma_{v11}\end{pmatrix}\begin{pmatrix}\hat{\boldsymbol{\theta}}_1-\boldsymbol{\theta}_{10}\\ \hat{\boldsymbol{\beta}}_1-\boldsymbol{\beta}_{10}\end{pmatrix}\\
&+\frac{1}{n}\sum_{i=1}^{n}\begin{pmatrix} g'(\boldsymbol{W}_{i,1}^{\top}\boldsymbol{\theta}_{10})\tilde{\boldsymbol{W}}_{i,1}\Delta_{i,1}+\bar{g}'^2(\Lambda)\Sigma_{u11}\boldsymbol{\theta}_{10}\\ \tilde{\boldsymbol{T}}_{i,1}\Delta_{i,1}+\Sigma_{v11}\boldsymbol{\beta}_{10}\end{pmatrix}+o_p(1),
\end{aligned}$$

where $\Delta_{i,1}=\varepsilon_i-g'(\boldsymbol{X}_{i,1}^{\top}\boldsymbol{\theta}_{10})\boldsymbol{U}_{i,1}^{\top}\boldsymbol{\theta}_{10}-\boldsymbol{V}_{i,1}^{\top}\boldsymbol{\beta}_{10}$. Moreover, the summand of the matrix over n in the first term of the above equation converges to

$$\begin{pmatrix}\Gamma_{\tilde{\boldsymbol{W}}_1}-\bar{g}'^2(\Lambda)\Sigma_{u11} & \Gamma_{\tilde{\boldsymbol{W}}_1\tilde{Z}_1}\\ \Gamma_{\tilde{\boldsymbol{W}}_1\tilde{Z}_1}^{\top} & \Gamma_{\tilde{Z}_1}\end{pmatrix},$$

where $\Gamma_{\tilde{W}_1}=E(g'(\boldsymbol{W}_1^{\top}\boldsymbol{\theta}_{10})\tilde{\boldsymbol{W}}_1)^{\otimes^2}$, $\Gamma_{\tilde{\boldsymbol{W}}_1\tilde{\boldsymbol{Z}}_1}=E(g'(\boldsymbol{W}_1^{\top}\boldsymbol{\theta}_{10})\tilde{\boldsymbol{W}}_1\tilde{\boldsymbol{Z}}_1^{\top})$ and $\Gamma_{\tilde{\boldsymbol{Z}}_1}=E(\tilde{\boldsymbol{Z}}_1\tilde{\boldsymbol{Z}}_1^{\top})$. These results, together with (0.6), lead to

$$\begin{aligned}
&\sqrt{n}\begin{pmatrix}\Gamma_{\tilde{\boldsymbol{W}}_1}-\bar{g}'^2(\Lambda)\Sigma_{u11}+\Sigma_1 & \Gamma_{\tilde{\boldsymbol{W}}_1\tilde{Z}_1}\\ \Gamma_{\tilde{\boldsymbol{W}}_1\tilde{Z}_1}^{\top} & \Gamma_{\tilde{Z}_1}+\Sigma_2\end{pmatrix}\begin{pmatrix}\hat{\boldsymbol{\theta}}_1-\boldsymbol{\theta}_{10}\\ \hat{\boldsymbol{\beta}}_1-\boldsymbol{\beta}_{10}\end{pmatrix}-\frac{1}{\sqrt{n}}\begin{pmatrix}\boldsymbol{R}_1\\ \boldsymbol{R}_2\end{pmatrix}\\
=&\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\begin{pmatrix} g'(\boldsymbol{W}_{i,1}^{\top}\boldsymbol{\theta}_{10})\tilde{\boldsymbol{W}}_{i,1}\Delta_{i,1}+\bar{g}'^2(\Lambda)\Sigma_{u11}\boldsymbol{\theta}_{10}\\ \tilde{\boldsymbol{T}}_{i,1}\Delta_{i,1}+\Sigma_{v11}\boldsymbol{\beta}_{10}\end{pmatrix}+o_p(1).
\end{aligned}$$

It follows that

$$\begin{aligned}
&\sqrt{n}\left(\Gamma_{\tilde{\boldsymbol{W}}_1}-\bar{g}'^2(\Lambda)\Sigma_{u11}+\Sigma_1\right)\left(\hat{\boldsymbol{\theta}}_1-\boldsymbol{\theta}_{10}\right)+\sqrt{n}\Gamma_{\tilde{\boldsymbol{W}}_1\tilde{\boldsymbol{Z}}_1}\left(\hat{\boldsymbol{\beta}}_1-\boldsymbol{\beta}_{10}\right)-\frac{1}{\sqrt{n}}\boldsymbol{R}_1\\
=&\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left\{g'(\boldsymbol{W}_{i,1}^{\top}\boldsymbol{\theta}_{10})\tilde{\boldsymbol{W}}_{i,1}\Delta_{i,1}+\bar{g}'^2(\Lambda)\Sigma_{u11}\boldsymbol{\theta}_{10}\right\}+o_p(1)
\end{aligned}$$

and

$$\begin{aligned}
&\sqrt{n}\Gamma_{\tilde{\boldsymbol{W}}_1\tilde{\boldsymbol{Z}}_1}^{\top}\left(\hat{\boldsymbol{\theta}}_1-\boldsymbol{\theta}_{10}\right)+\sqrt{n}\left(\Gamma_{\tilde{\boldsymbol{Z}}_1}+\Sigma_2\right)\left(\hat{\boldsymbol{\beta}}_1-\boldsymbol{\beta}_{10}\right)-\frac{1}{\sqrt{n}}\boldsymbol{R}_2\\
=&\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left\{\tilde{\boldsymbol{T}}_{i,1}\Delta_{i,1}+\Sigma_{v11}\boldsymbol{\beta}_{10}\right\}+o_p(1).
\end{aligned}$$

After simplification, we have

$$\begin{aligned}
&\sqrt{n}\left\{\Gamma_4-\Gamma_{\tilde{\boldsymbol{W}}_1\tilde{\boldsymbol{Z}}_1}\Gamma_3^{-1}\Gamma_{\tilde{\boldsymbol{W}}_1\tilde{Z}_1}^{\top}\right\}\left(\hat{\boldsymbol{\theta}}_1-\boldsymbol{\theta}_{10}\right)+\frac{1}{\sqrt{n}}\left\{\Gamma_{\tilde{\boldsymbol{W}}_1\tilde{\boldsymbol{Z}}_1}\Gamma_3^{-1}\boldsymbol{R}_2-\boldsymbol{R}_1\right\}\\
=&\ \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left\{\left(g'(\boldsymbol{W}_{i,1}^{\top}\boldsymbol{\theta}_{10})\tilde{\boldsymbol{W}}_{i,1}\Delta_{i,1}+\bar{g}'^2(\Lambda)\Sigma_{u11}\boldsymbol{\theta}_{10}\right)-\left(\Gamma_{\tilde{\boldsymbol{W}}_1\tilde{Z}_1}\Gamma_3^{-1}\right)\left(\tilde{\boldsymbol{T}}_{i,1}\Delta_{i,1}+\Sigma_{v11}\boldsymbol{\beta}_{10}\right)\right\}+o_p(1)\\
=&\ \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\Big\{(g'(\boldsymbol{W}_{i,1}^{\top}\boldsymbol{\theta}_{10})\tilde{\boldsymbol{X}}_{i,1}-\Gamma_{\tilde{\boldsymbol{W}}_1\tilde{Z}_1}\Gamma_3^{-1}\tilde{\boldsymbol{Z}}_{i,1})\Delta_{i,1}+g'(\boldsymbol{W}_{i,1}^{\top}\boldsymbol{\theta}_{10})\boldsymbol{U}_{i,1}\varepsilon_i+\Gamma_{\tilde{\boldsymbol{W}}_1\tilde{\boldsymbol{Z}}_1}\Gamma_3^{-1}(\boldsymbol{V}_{i,1}\boldsymbol{V}_{i,1}^{\top}-\Sigma_{v11})\boldsymbol{\beta}_{10}\\
&-\bar{g}'^2(\Lambda)(\boldsymbol{U}_{i,1}\boldsymbol{U}_{i,1}^{\top}-\Sigma_{u11})\boldsymbol{\theta}_{10}-\Gamma_{\tilde{\boldsymbol{W}}_1\tilde{\boldsymbol{Z}}_1}\Gamma_3^{-1}\boldsymbol{V}_{i,1}\varepsilon_i\Big\}
\end{aligned}\tag{0.12}$$

and

$$\begin{aligned}
&\sqrt{n}\left\{\Gamma_3-\Gamma_{\tilde{\boldsymbol{W}}_1\tilde{\boldsymbol{Z}}_1}^{\top}\Gamma_4^{-1}\Gamma_{\tilde{\boldsymbol{Z}}_1\tilde{\boldsymbol{Z}}_1}\right\}\left(\hat{\boldsymbol{\beta}}_1-\boldsymbol{\beta}_{10}\right)+\frac{1}{\sqrt{n}}\left\{\left(\Gamma_{\tilde{\boldsymbol{W}}_1\tilde{\boldsymbol{Z}}_1}^{\top}\Gamma_4^{-1}\right)\boldsymbol{R}_1-\boldsymbol{R}_2\right\}\\
=&\ \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left\{\left(\tilde{\boldsymbol{T}}_{i,1}\Delta_{i,1}+\Sigma_{v11}\boldsymbol{\beta}_{10}\right)-\left(\Gamma_{\tilde{W}_1\tilde{Z}_1}^{\top}\Gamma_4^{-1}\right)\left(g'(\boldsymbol{W}_{i,1}^{\top}\boldsymbol{\theta}_{10})\tilde{\boldsymbol{W}}_{i,1}\Delta_{i,1}+\bar{g}'^2(\Lambda)\Sigma_{u11}\boldsymbol{\theta}_{10}\right)\right\}+o_p(1)\\
=&\ \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\Big\{(\tilde{\boldsymbol{Z}}_{i,1}-\Gamma_{\tilde{\boldsymbol{W}}_1\tilde{\boldsymbol{Z}}_1}^{\top}\Gamma_4^{-1}g'(\boldsymbol{W}_{i,1}^{\top}\boldsymbol{\theta}_{10})\tilde{\boldsymbol{X}}_{i,1})\Delta_{i,1}+\boldsymbol{V}_{i,1}\varepsilon_i-(\boldsymbol{V}_{i,1}\boldsymbol{V}_{i,1}^{\top}-\Sigma_{v11})\boldsymbol{\beta}_{10}\\
&+\Gamma_{\tilde{\boldsymbol{W}}_1\tilde{\boldsymbol{Z}}_1}^{\top}\Gamma_4^{-1}\bar{g}'^2(\Lambda)(\boldsymbol{U}_{i,1}\boldsymbol{U}_{i,1}^{\top}-\Sigma_{u11})\boldsymbol{\theta}_{10}-\Gamma_{\tilde{W}_1\tilde{Z}_1}^{\top}\Gamma_4^{-1}g'(\boldsymbol{W}_{i,1}^{\top}\boldsymbol{\theta}_{10})\boldsymbol{U}_{i,1}\varepsilon_i\Big\}+o_p(1).
\end{aligned}\tag{0.13}$$

Equations (0.12) and (0.13), together with Slutsky's theorem and the central limit theorem, yield that

$$\sqrt{n}\left\{\Gamma_4-\Gamma_{\tilde{\boldsymbol{W}}_1\tilde{Z}_1}\Gamma_3^{-1}\Gamma_{\tilde{\boldsymbol{W}}_1\tilde{\boldsymbol{Z}}_1}^{\top}\right\}\left(\hat{\boldsymbol{\theta}}_1-\boldsymbol{\theta}_{10}\right)+\frac{1}{\sqrt{n}}\left\{\left(\Gamma_{\tilde{\boldsymbol{W}}_1\tilde{Z}_1}\Gamma_3^{-1}\right)\boldsymbol{R}_2-\boldsymbol{R}_1\right\}\xrightarrow{L}N(0,\Omega_\theta)$$

and

$$\sqrt{n}\left\{\Gamma_3-\Gamma_{\tilde{\boldsymbol{W}}_1\tilde{\boldsymbol{Z}}_1}^{\top}\Gamma_4^{-1}\Gamma_{\tilde{\boldsymbol{W}}_1\tilde{\boldsymbol{Z}}_1}\right\}\left(\hat{\boldsymbol{\beta}}_1-\boldsymbol{\beta}_{10}\right)+\frac{1}{\sqrt{n}}\left\{\left(\Gamma_{\tilde{\boldsymbol{W}}_1\tilde{Z}_1}^{\top}\Gamma_4^{-1}\right)\boldsymbol{R}_1-\boldsymbol{R}_2\right\}\xrightarrow{L}N(0,\Omega_\beta)$$

where

$$\begin{aligned}
\Omega_\theta=&\left\{E\left(g'(\boldsymbol{W}_1^{\top}\boldsymbol{\theta}_{10})\tilde{\boldsymbol{X}}_1\right)^{\otimes^2}-\Gamma_{\tilde{\boldsymbol{W}}_1\tilde{Z}_1}\Gamma_3^{-1}E\{\tilde{\boldsymbol{Z}}_1\tilde{\boldsymbol{Z}}_1^{\top}\}\Gamma_3^{-1}\Gamma_{\tilde{\boldsymbol{W}}_1\tilde{\boldsymbol{Z}}_1}^{\top}\right\}E(\Delta_1)^{\otimes^2}\\
&+E\left(g'(\boldsymbol{W}_1^{\top}\boldsymbol{\theta}_{10})\boldsymbol{U}_1\right)^{\otimes^2}\sigma^2+E\left(\bar{g}'^2(\Lambda)(\boldsymbol{U}_1\boldsymbol{U}_1^{\top}-\Sigma_{u11})\boldsymbol{\theta}_{10}\right)^{\otimes^2}\\
&+\Gamma_{\tilde{\boldsymbol{W}}_1\tilde{\boldsymbol{Z}}_1}\Gamma_3^{-1}\left\{\Sigma_{v11}\sigma^2+E\left((\boldsymbol{V}_1\boldsymbol{V}_1^{\top}-\Sigma_{v11})\boldsymbol{\beta}_{10}\right)^{\otimes^2}\right\}\Gamma_3^{-1}\Gamma_{\tilde{\boldsymbol{W}}_1\tilde{\boldsymbol{Z}}_1}^{\top},
\end{aligned}$$

and

$$\begin{aligned}
\Omega_{\boldsymbol{\beta}}=&\left\{E(\tilde{\boldsymbol{Z}}_1\tilde{\boldsymbol{Z}}_1^{\top})-\Gamma_{\tilde{\boldsymbol{W}}_1\tilde{\boldsymbol{Z}}_1}^{\top}\Gamma_4^{-1}E\left(g'(\boldsymbol{W}_1^{\top}\boldsymbol{\theta}_{10})\tilde{\boldsymbol{X}}_1\right)^{\otimes^2}\Gamma_4^{-1}\Gamma_{\tilde{\boldsymbol{W}}_1\tilde{\boldsymbol{Z}}_1}\right\}E(\Delta_1)^{\otimes^2}\\
&+E\left((\boldsymbol{V}_1\boldsymbol{V}_1^{\top}-\Sigma_{v11})\boldsymbol{\beta}_1\right)^{\otimes^2}+\Sigma_{v11}\sigma^2\\
&+\Gamma_{\tilde{\boldsymbol{W}}_1\tilde{\boldsymbol{Z}}_1}^{\top}\Gamma_4^{-1}\left\{E(g'(\boldsymbol{W}_1^{\top}\boldsymbol{\theta}_{10})\boldsymbol{U}_1)^{\otimes^2}\sigma^2+E(\bar{g}'^2(\Lambda)(\boldsymbol{U}_1\boldsymbol{U}_1^{\top}-\Sigma_{u11})\boldsymbol{\theta}_{10})^{\otimes^2}\right\}\Gamma_4^{-1}\Gamma_{\tilde{\boldsymbol{W}}_1\tilde{Z}_1}.
\end{aligned}$$

Because each element of $\sqrt{n}\Sigma_1$, $\sqrt{n}\Sigma_2$, $\sqrt{n}\boldsymbol{R}_1$ and $\sqrt{n}\boldsymbol{R}_2$ tends to zero, then we complete the proof. □