Journal of the Korean Statistical Society (2021) 50:163–184
https://doi.org/10.1007/s42952-020-00067-4

Online ISSN 2005-2863
Print ISSN 1226-3192

RESEARCH ARTICLE

Regularization statistical inferences for partially linear models with high dimensional endogenous covariates

Changqing Liu[1] · Peixin Zhao[2,3] · Yiping Yang[2]

Received: 30 September 2019 / Accepted: 1 April 2020 / Published online: 20 April 2020

Abstract
In this paper, we consider the statistical inferences for a class of partially linear models with high dimensional endogenous covariates, when high dimensional instrumental variables are also available. A regularized estimation procedure is proposed for identifying the optimal instrumental variables, and estimating covariate effects of the parametric and nonparametric components. Under some conditions, some theoretical properties are studied, such as the consistency of the optimal instrumental variable identification and significant covariate selection. Furthermore, some simulation studies and a real data analysis are carried out to examine the finite sample performance of the proposed method.

Keywords Partially linear model · High dimensional endogenous covariates · High dimensional instrumental variables · Regularized estimation

Mathematics Subject Classification 62G05 · 62G20

1 Introduction

Let Y_i be the response variable, and X_i and U_i be the corresponding covariates, then the partially linear model has the following structure:

$$Y_i = X_i^{\mathrm{T}}\beta + g(U_i) + \varepsilon_i, \quad i = 1, \ldots, n, \tag{1}$$

where $\beta = (\beta_1, \ldots, \beta_{p_n})^T$ is a p_n-dimensional vector of unknown parameters, $g(\cdot)$ is an unknown nonparametric function, and ε_i is the model error with $E(\varepsilon_i|X_i, U_i) = 0$.

✉ Peixin Zhao
zpx81@163.com

[1] College of Mathematics and Statistics, Baise University, Guangxi Baise 533000, China

[2] College of Mathematics and Statistics, Chongqing Technology and Business University, Chongqing 400067, China

[3] Chongqing key laboratory of social economy and applied statistics, Chongqing 400067, China

In this paper, we assume the dimension p_n can be diverging with the sample size n. Model (1) increases the flexibility of linear models by allowing the intercept to be a nonparametric function, and this model is one of the most popular semiparametric regression models in the literature. Due to the flexibility of model (1), it has attracted extensive attentions of many scholars recently, such as Fan and Li (2004), Xue and Zhu (2007), Xie and Huang (2009), Huang and Zhao (2017) and Liu et al. (2018), among others.

However, some covariates for regression modeling may be endogenous in practice (see Newhouse and McClellan 1998; Greenland 2000; Hernan and Robins 2006; Fan and Liao 2014). For such case, the estimation methods for model (1) listed above will give an endogeneity bias, and can not give a consistent estimator any more. For such models with endogenous covariates, the instrumental variable adjustment technology can provide a way to obtain a consistent estimation procedure. Therefore, the semiparametric instrumental variable models with endogenous covariates have received a great deal of attention recently, such as Cai and Xiong (2012), Zhao and Li (2013), Yang et al. (2017), Yuan et al. (2016), and Huang and Zhao (2018), among others. In these studies, an essential assumption is that the covariates and instrumental variables are both low-dimensional data with fixed dimension. However, high dimensional data frequently occur in practice.

For the partially linear model, defined by (1), with high dimensional endogenous covariates, Chen et al. (2016) proposed a penalized GMM estimation procedure to perform variable selection for covariates. However, the regularized estimation method proposed by Chen et al. (2016) does not exploit the sparsity of the instrumental variables, and then is still facing the dimensionality curse of high dimensional instrumental variables. Hence take this issue into account, in this paper, we consider the statistical inference for model (1) when some covariates are high dimensional endogenous covariates and a high dimensional set of instrumental variables is available. More specifically, we assume the covariate X in model (1) is an endogenous covariate, and satisfies the following structure

$$X = \Gamma Z + e, \tag{2}$$

where $Z = (Z_1, \ldots, Z_{q_n})^T$ is the corresponding q_n-dimensional vector of instrumental variables, Γ is a $p_n \times q_n$ matrix of unknown parameters, and e is the model error with $E\{e|Z, U\} = 0$. Furthermore, ε and e are assumed to be independent each other, and the dimension q_n also allows to be diverging with the sample size n.

As discussed above, in the following discussion, we are interested in making inference under the high dimensional setting of covariate X and instrumental variable Z. As is typical in high dimensional sparse modeling, we assume models (1) and (2) are both sparse in the sense that only a small subset of parameters in β and Γ are nonzero. Our goal is to identify the optimal instrumental variables, and propose a regularized estimation procedure for model (1) based on the selected optimal instrumental variables.

Recently, penalized methods have a great attraction and proved their efficiency for performing variable selection and parameter estimation simultaneously. Some of these methods are bridge penalty (see Frank and Friedman 1993), Lasso penalty (see

Tibshirani 1996), SCAD penalty (see Fan and Li 2001), MCP penalty (see Zhang 2010), and among others. In addition, Lee et al. (2019) present a systematic review on variable selection for high dimensional regression models. In most of the literature listed above, however, the data are assumed to be exogenous. For such high dimensional model with endogenous covariates, these variable selection procedure listed above will give an endogenity bias, and can not give a consistent variable selection result any more. Then, compared with existing estimation methods, our estimation method has the following improvements. Firstly, the proposed method can identify the optimal instrumental variables and important covariates simultaneously, and this is an essential improvement of the regularized estimation procedure proposed by Chen et al. (2016). Secondly, our regularized estimation method is constructed based on the penalized least absolute deviation estimation procedure. Hence, compared with the optimal instrumental variable identification method proposed by Lin et al. (2015), our regularized estimation is more robust. Lastly, the proposed regularized estimation for identifying optimal instrumental variables are constructed by using an auxiliary regression model, which is very different from the existing identification methods of optimal instrumental variables, such as Lin et al. (2015) and Windmeijer et al. (2019).

The rest of this paper is organized as follows. In Sect. 2, we propose an identification method of optimal instrumental variables based on the penalized least absolute deviation method and an auxiliary regression model constructed artificially, and demonstrate some theoretical properties of the proposed optimal instrumental variable identification method. In Sect. 3, we propose a variable selection method of significant covariates in model (1) with the selected optimal instrumental variables, and derive the estimators of model parametric and nonparametric components. In Sect. 4, we propose an iterative algorithm procedure for the proposed regularized estimation method based on the local linear approximation method. In Sect. 5, some simulation studies and a real data analysis are conducted to assess the performances of the proposed method. The technical proofs for all asymptotic results are presented in the Appendix.

2 Optimal instrumental variable identification

For the identification of models (1) and (2), similar to Cai and Xiong (2012), we first give some regularity conditions for models (1) and (2). More specifically, let $\tilde{Z}$ be a vector of true valid instrumental variables, which is a subset of $Z = (Z_1, \ldots, Z_{q_n})^T$. Then we assume that the dimensionality of $\tilde{Z}$ is larger than or equal to the dimensionality of X. Furthermore, we assume that the matrix Γ is a row full rank matrix. Obviously, these regularity conditions ensure that the models (1) and (2) are identifiable, and every endogenous variable $X_j, 1 \le j \le p_n$ has at least one valid instrumental variable. Because X is an endogenous covariate, and Z is the corresponding instrumental variable, we have $E(\varepsilon|X, U) \neq 0$ and $E(\varepsilon|Z, U) = 0$. In addition, note that U is an exogenous covariate, we assume that the instrumental variable Z is independent of U. Let Z_k be the kth component of instrumental variable $Z = (Z_1, \ldots, Z_{q_n})^T$, then invoking model (1), and some calculations yield

$$Cov(Y, Z_k) = \sum_{j=1}^{p_n} Cov(X_j, Z_k)\beta_j. \tag{3}$$

Didelez et al. (2010) point that if Z_k is an optimal instrumental variable, then Z_k should be significantly correlated with endogenous covariate X. That is, if for all $j \in \{1, \ldots, p_n\}$, we have $Cov(X_j, Z_k) = 0$, then Z_k is an invalid instrumental variable. Hence, (3) implies that if $Cov(Y, Z_k) \neq 0$ significantly hold, then Z_k is an optimal instrumental variable. Hence, based on this result, we can identify optimal instrumental variable based on the following auxiliary regression model

$$Y_i = Z_{i1}\theta_1 + Z_{i2}\theta_2 + \cdots + Z_{iq_n}\theta_{q_n} + \varepsilon_i, \quad i = 1, \ldots, n. \tag{4}$$

More specifically, if Z_k is an optimal instrumental variable, then the corresponding coefficient θ_k should be significantly nonzero. Hence, to recover the optimal instrumental variables, we define the following penalized objective function

$$Q_n(\theta) = \frac{1}{n}\sum_{i=1}^{n}\left|Y_i - Z_i^T\theta\right| + \sum_{k=1}^{q_n} p_{\lambda_{1n}}(|\theta_k|), \tag{5}$$

where $\theta = (\theta_1, \ldots, \theta_{q_n})^T$ is a q_n-dimensional parametric vector, $p_{\lambda_{1n}}(\cdot)$ is a specified penalized function, and λ_{1n} is a tuning parameter. In practice, there are many penalty functions can be used, such as the Lasso penalty proposed by Tibshirani (1996), the SCAD penalty proposed by Fan and Li (2001), and the MCP penalty proposed by Zhang (2010), among others.

Let $\widehat{\theta}$ be the solution of θ by minimizing (5), then we next study the asymptotic properties of the regularized estimator $\widehat{\theta}$. From the assumption of sparsity in instrumental variables, we know that only a small subset of components in θ is nonzero. Then, for convenience and simplicity, we let θ_0 be the true value of θ, $\mathscr{A}_1 = \{1 \le k \le q_n : \theta_{0k} \neq 0\}$ and $\mathscr{A}_2 = \{1 \le k \le q_n : \theta_{0k} = 0\}$. The corresponding optimal instrumental variables and coefficient matrix are denoted as $Z_{\mathscr{A}_1}$ and $\Gamma_{\mathscr{A}_1}$, respectively. Then model (2) can also be rewritten as

$$X = \Gamma_{\mathscr{A}_1} Z_{\mathscr{A}_1} + e. \tag{6}$$

Next, we demonstrate some asymptotic properties of the resulting estimator $\widehat{\theta}$. To establish the asymptotic properties, we first assume some regularity conditions as follows:

(C1) The nonparametric function $g(u)$ is rth continuously differentiable on (0, 1) with $r \ge 2$.
(C2) The error ε has continuous and symmetric density $f(\cdot)$. Moreover, the density function $f(\cdot)$ has finite derivatives in any neighborhood of zero.
(C3) Let $c_1, \ldots, c_K$ be the interior knots of [0, 1]. Furthermore, we let $c_0 = 0, c_{K+1} = 1$, $h_i = c_i - c_{i-1}$. Then, there exists a constant c such that $\max\{h_i\}/\min\{h_i\} \le c$ and $\max\{|h_{i+1} - h_i|\} = o(\kappa_n^{-1})$, where κ_n is the number of interior knots.
(C4) There exists a positive constant c such that $\max_{i,j}|X_{ij}| < c$, $\max_{i,k}|Z_{ik}| < c$ in probability, where $i = 1, \ldots, n$, $j = 1, \ldots, p_n$ and $k = 1, \ldots, q_n$.

(C5) The dimensions of covariate X and instrumental variable Z satisfy $p_n^3/n \to 0$ and $q_n^3/n \to 0$ as $n \to \infty$.

(C6) The matrix $\Gamma_{\mathscr{A}_1}$, defined in (6), is a row full rank matrix. In addition, let τ_{n1} and τ_{n2} be the smallest and largest eigenvalues of the matrix $E(ZZ^T)$, and ρ_{n1} and ρ_{n2} be the smallest and largest eigenvalues of the matrix $\Gamma_{\mathscr{A}_1}\Gamma_{\mathscr{A}_1}^T$, respectively. Then, there exist constants $0 < \rho_1 < \rho_2 < \infty$ and $0 < \tau_1 < \tau_2 < \infty$ such that $\rho_1 < \rho_{n1} < \rho_{n2} < \rho_2$ and $\tau_1 < \tau_{n1} < \tau_{n2} < \tau_2$.

(C7) The tuning parameter λ_{1n} satisfies $\lambda_{1n} \to 0$ and $\sqrt{n/q_n}\lambda_{1n} \to \infty$. In addition, we assume $\liminf_{n\to\infty} \liminf_{\theta_k \to 0^+} \lambda_{1n}^{-1} p'_{\lambda_{1n}}(|\theta_k|) > 0$.

(C8) Let $a_n = \max_k \left\{ |p'_{\lambda_{1n}}(|\theta_{0k}|)| : \theta_{0k} \neq 0 \right\}$, $b_n = \max_k \left\{ |p''_{\lambda_{1n}}(|\theta_{0k}|)| : \theta_{0k} \neq 0 \right\}$, then we assume that $n^{1/2}a_n \to 0$ and $b_n \to 0$ as $n \to \infty$.

Conditions C1 and C2 are common assumptions of nonparametric and semiparametric estimation technology. Condition C3 implies that $c_0, \ldots, c_{K+1}$ is a C_0-quasi-uniform sequence of partitions of [0, 1]. The assumptions of high order moment conditions in C4–C6 are standard assumptions for high dimensional semiparametric regression models in the literature, which ensure that the models (1) and (2) are identifiable, and the proposed regularized estimation procedure is consistent. Condition C7 and C8 are assumptions on the penalty function, which ensure that the proposed variable selection method is consistent, which are widely used in variable selection literature (see Fan and Li 2001, Li and Liang (2008), and Wang et al. (2008)). Under these regularity conditions, the following Theorem 1 shows that the resulting regularized estimator $\widehat{\theta}$ is consistent, and gives the convergence rate of $\widehat{\theta}$.

Theorem 1 *Suppose the regularity conditions (C1)–(C8) hold. Then we have that*

$$\|\widehat{\theta} - \theta_0\| = O_p(\sqrt{q_n/n}).$$

Furthermore, we show that such consistent estimators must possess the sparsity property, which is stated in the following Theorem 2.

Theorem 2 *Suppose the regularity conditions (C1)–(C8) hold. Then, with probability tending to* 1, *we have that*

$$\widehat{\theta}_k = 0, \quad k \in \mathscr{A}_2.$$

3 Regularized estimation for model parameters

We denote Z^* as the vector of selected optimal instrumental variables in Sect. 2. Note that Theorem 2 implies that the variable selection for optimal instrumental variables is consistent, then with probability tending to one, we have $Z_{\mathscr{A}_1} = Z^*$ when n is large enough. Then, invoking model (6), the moment estimator of $\Gamma_{\mathscr{A}_1}$ is defined by

$$\widehat{\Gamma} = \left(\sum_{i=1}^{n} X_i Z_i^{*T} \right) \left(\sum_{i=1}^{n} Z_i^* Z_i^{*T} \right)^{-1}. \tag{7}$$

Therefore, the optimal instrumental variable adjusted covariates is defined by $X^* = \widehat{\Gamma} Z^*$. Next, invoking the adjusted covariate X^*, we proceed to identify and estimate the nonzero effects of the covariates in model (1). The regularized estimation objective function is defined by

$$M_n(\beta, g(\cdot)) = \frac{1}{n} \sum_{i=1}^{n} \left| Y_i - X_i^{*T} \beta - g(U_i) \right| + \sum_{j=1}^{p_n} p_{\lambda_{2n}}(|\beta_j|).$$

Note that $g(\cdot)$ is a nonparametric function, $M_n(\beta, g(\cdot))$ is not ready for optimization. Then, invoking B-spline approximation technique (see Schumaker 1981), we replace $g(\cdot)$ in $M_n(\beta, g(\cdot))$ by its basis function approximations. More specifically, let $B(u) = (B_1(u), \ldots, B_{L_n}(u))^T$ be B-spline basis functions with the order of M, where $L = \kappa_n + M$, and κ_n is the number of interior knots. Then, $g(u)$ can be approximated by $g(u) \approx B(u)^T \gamma$, where $\gamma = (\gamma_1, \ldots, \gamma_{L_n})^T$ is a vector of basis functions coefficients. Substituting it into $M_n(\beta, g(\cdot))$, we can obtain that

$$M_n(\beta, \gamma) = \frac{1}{n} \sum_{i=1}^{n} \left| Y_i - X_i^{*T} \beta - W_i^T \gamma \right| + \sum_{j=1}^{p_n} p_{\lambda_{2n}}(|\beta_j|), \tag{8}$$

where $W_i = B(U_i)$. Let $\widehat{\beta}$ and $\widehat{\gamma}$ be the solution by minimizing (8), then $\widehat{\beta}$ is the optimal instrumental variable based estimator of β, and the estimator of $g(u)$ is given by $\widehat{g}(u) = B(u)^T \widehat{\gamma}$.

Remark 1 Although $M_n(\beta, g(\cdot))$ contains nonparametric function $g(\cdot)$, and cannot be minimized directly, we replace $g(\cdot)$ by its basis function approximations based on the B-spline approximation technique. Then, we can obtain the estimator of parametric component β and nonparametric component $g(\cdot)$ simultaneously. B-spline based estimation is a more effective nonparametric estimation method, which is widely used in the nonparametric and semiparametric regression literature. In addition, note that $M_n(\beta, g(\cdot))$ contains unknown parametric component β and nonparametric function $g(\cdot)$ simultaneously, then the proposed estimation procedure can be regarded as a semiparaemtric regularized estimation procedure.

Next, we study the asymptotic properties of the regularized estimator $\widehat{\beta}$ and $\widehat{g}(u)$. Similar to conditions (C7) and (C8), we give some conditions for the penalty function used in (8).

(C9) The tuning parameter λ_{2n} satisfies $\lambda_{2n} \to 0$ and $\sqrt{n/p_n}\lambda_{2n} \to \infty$. Furthermore, we assume $\liminf_{n\to\infty} \liminf_{\beta_j \to 0^+} \lambda_{2n}^{-1} p'_{\lambda_{2n}}(|\beta_j|) > 0$.

(C10) $a_n^* = \max_j \left\{ |p'_{\lambda_{2n}}(|\beta_{0j}|)| : \beta_{0j} \neq 0 \right\}$ and $b_n^* = \max_j \left\{ |p''_{\lambda_{2n}}(|\beta_{0j}|)| : \beta_{0j} \neq 0 \right\}$, then we assume that $n^{1/2} a_n^* \to 0$ and $b_n^* \to 0$ as $n \to \infty$.

In addition, for convenience and simplicity, we let β_0 be the true value of β, $\mathscr{B}_1 = \{1 \le j \le p_n : \beta_{0j} \ne 0\}$ and $\mathscr{B}_2 = \{1 \le j \le p_n : \beta_{0j} = 0\}$. Furthermore, we let γ_0 be the true value of γ. The following Theorem 3 shows the resulting estimator $\widehat{\beta}$ is consistent, and $\widehat{\beta}$ satisfies sparsity.

Theorem 3 *Suppose the regularity conditions (C1)–(C10) hold, and the number of knots κ_n satisfies $\kappa_n = O(n^{1/(2r+1)})$ and $\lambda_{2n}/\sqrt{\kappa_n/n} \to \infty$. Then we have that*

(i) $\|\widehat{\beta} - \beta_0\| = O_p(\sqrt{(p_n + \kappa_n)/n})$.
(ii) $\widehat{\beta}_j = 0$, $j \in \mathscr{B}_2$, *with probability tending to* 1.

Theorem 3 shows that, under some regularity conditions, the resulting estimator $\widehat{\beta}$ is consistent, and satisfies sparsity. This implies that the proposed regularized estimation method for β can be used to select important covariates in model (1). Furthermore, the following Theorem 4 shows that the estimator of nonparametric function $g(u)$ is also consistent, and achieves the optimal nonparametric convergence rate.

Theorem 4 *Suppose the regularity conditions (C1)–(C10) hold, and the number of knots κ_n satisfies $\kappa_n = O(n^{1/(2r+1)})$ and $\lambda_{2n}/\sqrt{\kappa_n/n} \to \infty$. Then we have that*

$$\|\widehat{g}(u) - g(u)\| = O_p(n^{-r/(2r+1)}),$$

where r is defined in condition (C1).

4 Iterative algorithms

In this section, we give an iterative algorithm procedure of the proposed estimation method in Sects. 2 and 3. Similar to Zou and Li (2008), we use the local linear approximation method to the penalty function $p_{\lambda_{1n}}(\cdot)$. Then, (5) can be rewritten as

$$Q_n(\theta) = \frac{1}{n}\sum_{i=1}^{n}\left|Y_i - Z_i^T\theta\right| + \sum_{k=1}^{q_n} p'_{\lambda_{1n}}\left(\left|\theta_k^{(0)}\right|\right)|\theta_k|, \tag{9}$$

where $p'_{\lambda_{1n}}(\cdot)$ is the first-order derivative of $p_{\lambda_{1n}}(\cdot)$, and $\theta^{(0)} = (\theta_1^{(0)}, \ldots, \theta_{q_n}^{(0)})^T$ is an initial estimator of θ, which is computed based on the least absolute deviation estimator without penalty. Furthermore, we construct an augmented data set $(\widetilde{Y}_i, \widetilde{Z}_i)$ with $i = 1, \ldots, n + q_n$ as follows

$$(\widetilde{Y}_i, \widetilde{Z}_i) = \begin{cases} (Y_i/n, Z_i/n), & 1 \le i \le n \\ \left(0, p'_{\lambda_{1n}}\left(\left|\theta_k^{(0)}\right|\right)\xi_{i-n}\right), & i = n+1, \ldots, n+q_n, \end{cases} \tag{10}$$

where ξ_k is the unit vector with the kth element being 1. Then, (9) can be rewritten as

$$Q_n(\theta) = \sum_{i=1}^{n+q_n} \left| \widetilde{Y}_i - \widetilde{Z}_i^T \theta \right|. \tag{11}$$

Hence, the penalized least absolute deviation estimator of θ can be easily calculated by the R package "quantreg" for quantile regression (see Koenker 2005). Similarly, we can easily obtain the penalized least absolute deviation estimator of β and γ based on (8). Then, our regularized algorithm has two stages. In the first stage, we identify the optimal instrumental variables. In the second stage, we identify the important covariates and estimate model parameters. In addition, it is worth mentioning that, as discussed above, both of the two stages can be calculated easily. More specifically, the two-stage regularized algorithm is as follows:

Stage 1. We identify the optimal instrumental variables based on model (4). Let $\widehat{\theta} = (\widehat{\theta}_1, \ldots, \widehat{\theta}_{q_n})^T$ be the solution by minimizing the objective function (5), and $\mathscr{A}_* = \{1 \le k \le q_n : \widehat{\theta}_k \ne 0\}$. Invoking model (4), we denote the corresponding instrumental variables of $\mathscr{A}_*$ as Z^*. Then based on the argument in Sect. 2, Z^* is a vector of identified optimal instrumental variables.

Stage 2. We define the instrumental variable adjusted covariates as $X^* = \widehat{\Gamma} Z^*$, where $\widehat{\Gamma}$ is defined by (7). Then, we identify the important covariates and estimate model parameters based on the objective function (8). Let $\widehat{\beta}$ and $\widehat{\gamma}$ be the solution by minimizing (8), then $\widehat{\beta}$ is the regularized estimator of β, and the estimator of $g(u)$ is given by $\widehat{g}(u) = B(u)^T \widehat{\gamma}$.

In addition, in the proposed iterative algorithm, the penalty parameter λ_{1n}, λ_{2n} and the number in interior knots κ_n should be chosen. Similar to Gao and Huang (2010) and Wang et al. (2015), we suggest choosing these parameters by using the Bayesian information criterion (BIC). More specifically, we estimate λ_{1n} by the following BIC function

$$BIC(\lambda_{1n}) = \log\left(\frac{1}{n} \sum_{i=1}^{n} \left| Y_i - Z_i^T \widehat{\theta} \right| \right) + d_{\lambda_{1n}} \frac{\log(n)}{n},$$

where $d_{\lambda_{1n}}$ is the number of nonzero coefficients in $\widehat{\theta}$, which is obtained by (5). Furthermore, λ_{2n} and κ_n are chosen by using the following BIC function

$$BIC(\lambda_{2n}, \kappa_n) = \log\left(\frac{1}{n} \sum_{i=1}^{n} \left| Y_i - X_i^{*T} \widehat{\beta} - W_i^T \widehat{\gamma} \right| \right) + d_{\kappa_n} \frac{\log(n)}{n},$$

where d_{κ_n} is the effective number of parameters in $\widehat{\beta}$ and $\widehat{\gamma}$, which are obtained by (8). Note that λ_{1n}, λ_{2n} and κ_n are all one-dimensional parameters, then we can minimize the BIC criterion by some grid points. More specifically, for parameter λ_{1n}, we first choose a 100×1 uniform grid points vector in the region [0.01, 0.99]. Secondly, we calculate the BIC values on each grid point. Then, based on the 100 BIC values, we choose the optimal λ_{1n} that corresponds to the minimum BIC value. Similarly, we can choose the optimal λ_{2n} and κ_n by using the grid search method.

5 Numerical results

In this section, we conduct several simulation experiments to illustrate the finite sample performances of the proposed method, and consider a real data set analysis for further illustration.

5.1 Simulation studies

In this section, we conduct some Monte Carlo simulations to evaluate finite sample performance of the proposed method. The main objective is to evaluate the performance of instrumental variable identification, and the effectiveness of instrumental variable based covariate adjustment technique. Then, the data are generated from the following model

$$\begin{cases} Y_i = X_i^T \beta + g(U_i) + \varepsilon_i \\ X_i = \Gamma Z_i + \alpha \varepsilon_i, \quad i = 1, \ldots, n, \end{cases}$$

where $g(u) = \sin(2\pi u)$ and $\beta = (2.5, 2, 1.5, 1, 0, \ldots, 0)^T$ is a p_n dimensional parametric vector. From the definition of β, we can see that X_{ij}, $j = 1, \ldots, 4$ are important covariates, and the others are unimportant covariates. Furthermore, we generate nonzero entries of the first four columns in Γ from the uniform distribution $U(0.75, 1)$, and the other columns are all set to zero, and the instrumental variables are generated by $Z_{ik} \sim N(1, 1.5)$, $k = 1, \ldots, q_n$. From the generative mechanism of instrumental variables, we can see that Z_{ik}, $k = 1, \ldots, 4$ are optimal instrumental variables, and the others are invalid instrumental variables. The exogenous covariate U_i is generated from $U_i \sim U(0, 1)$, and the endogenous covariates X_i and the response Y_i are generated according to the model with $\varepsilon \sim N(0, 0.5)$ and $\alpha = 0.2$ and 0.8, respectively, to represent different levels of endogeneity of covariates. This set up allows the covariate X_i is endogenous, because $E(X_i \varepsilon_i) \neq 0$.

In the following simulation, the penalty function is taken as the SCAD penalty, Lasso penalty and MCP penalty, respectively. The sample size is taken as $n = 200$, 400 and 600, respectively, the dimensionality of covariate and instrumental variable are taken as $(p_n, q_n) = (5\lfloor n^{1/5} \rfloor, 5\lfloor n^{1/4} \rfloor)$ for each sample size, and for each case, we take 1000 simulation runs.

We first evaluate the performance of the proposed optimal instrumental variable selection procedure. In this simulation, we present the number of true positive (TP), false positive (FP) and the false selection rate (FSR) as the effectiveness of the variable selection procedure, where the TP is the number of true optimal instrumental variables correctly set to optimal instrumental variables, the FP is the number of true invalid instrumental variables incorrectly set to optimal instrumental variables, and the FSR is defined as FSR = IN/TN, where "IN" is the number of the invalid instrumental variables incorrectly set to optimal instrumental variables, and "TN" is the total number set to optimal instrumental variables. In fact, FSR represents the proportion of falsely selected invalid instrumental variables among the total variables selected in the variable selection procedure. All these performance

indicators are averaged over all simulation runs. Based on 1000 simulation runs, the simulation results are reported in Table 1. From Table 1, we can make the following observations:

(i) For the given level of endogeneity of covariates, the FP and FSR decrease as the sample size n increases, and the TP tends to the true number 4 when the size of sample increases. This implies that the proposed identification method of optimal instrumental variables is consistent.
(ii) For given n, the proposed identification method performs similar in terms of TP, FP and FSR for both levels of endogeneity of covariates. This indicates that the proposed identification method can also select optimal instrumental variables when the level of endogeneity is lower.
(iii) For given size of sample and level of endogeneity of covariates, the simulation results are similar for different penalties, which means that the proposed method does not depend sensitively on the choice of penalty functions.

Next, we evaluate the performance of the proposed significant covariate selection procedure. In this simulation, we also present the number of true positive (TP), false positive (FP) and the false selection rate (FSR) as the effectiveness of the variable selection procedure. In addition, to evaluate the consistence of the resulting estimator of parametric component β, we define the generalized mean square error (GMSE) as follows:

$$GMSE = (\widehat{\beta} - \beta)^T \left[\frac{1}{n} \sum_{i=1}^{n} X_i^* X_i^{*T} \right] (\widehat{\beta} - \beta).$$

Based on 1000 simulation runs, we can obtain 1000 GMSE values. In the following simulations, we present the median of the 1000 GMSE values. The simulation results are reported in Table 2. From Table 2, we can see that the FP and FSR decrease as the sample size n increases, and the TP tends to the true number 4 when the size of sample increases. In addition, for given n, we also can see that the

Table 1 The identification results of optimal instrumental variables based on the proposed method under different cases

n	Method	$\alpha = 0.2$			$\alpha = 0.8$		
		TP	FP	FSR	TP	FP	FSR
200	LAD-SCAD	3.536	0.339	0.087	3.585	0.334	0.085
	LAD-Lasso	3.534	0.339	0.088	3.573	0.336	0.086
	LAD-MCP	3.538	0.336	0.086	3.564	0.337	0.086
400	LAD-SCAD	3.714	0.118	0.031	3.738	0.113	0.029
	LAD-Lasso	3.717	0.117	0.032	3.731	0.116	0.030
	LAD-MCP	3.718	0.117	0.031	3.724	0.112	0.029
600	LAD-SCAD	3.997	0.022	0.005	3.984	0.024	0.006
	LAD-Lasso	3.992	0.025	0.006	3.998	0.025	0.006
	LAD-MCP	3.988	0.025	0.006	3.984	0.027	0.007

Table 2 The variable selection results of significant covariates under different cases

n	Method	$\alpha = 0.2$				$\alpha = 0.8$			
		GMSE	TP	FP	FSR	GMSE	TP	FP	FSR
200	LAD-SCAD	0.184	2.938	0.331	0.101	0.185	2.935	0.334	0.102
	LAD-Lasso	0.189	2.934	0.331	0.101	0.186	2.935	0.333	0.102
	LAD-MCP	0.185	2.931	0.332	0.102	0.183	2.933	0.335	0.103
400	LAD-SCAD	0.133	3.647	0.219	0.057	0.133	3.632	0.214	0.056
	LAD-Lasso	0.137	3.642	0.212	0.055	0.132	3.632	0.216	0.056
	LAD-MCP	0.132	3.647	0.216	0.056	0.138	3.635	0.215	0.056
600	LAD-SCAD	0.072	3.986	0.036	0.009	0.076	3.982	0.034	0.008
	LAD-Lasso	0.072	3.981	0.033	0.008	0.077	3.983	0.035	0.009
	LAD-MCP	0.073	3.987	0.031	0.008	0.073	3.981	0.038	0.009

simulation results are similar in terms of TP, FP and FSR under different penalties, which implies that the proposed variable selection method for covariates does not depend sensitively on the choice of penalty functions.

Lastly, we evaluate the efficiency of the proposed instrumental variable adjustment mechanism. Here, two methods are compared: the instrumental variable adjustment based estimation (IVE) method proposed by this paper and the naive estimation (NE) method. The latter is neglecting the endogeneity of covariate, and minimizing the following objective function to obtain the estimator of β and γ.

$$\sum_{i=1}^{n} \left| Y_i - X_i^T \beta - W_i^T \gamma \right|^2 + \sum_{j=1}^{p_n} p_{\lambda_{2n}}(|\beta_j|),$$

where γ is a vector of basis function coefficients, which satisfies $g(u) \approx B(u)^T \gamma$, and $W_i = B(U_i)$.

For the parametric β, we only present the simulation results of nonzero component β_1 with SCAD penalty. The simulation results for other nonzero components are similar, and then are not shown. Based on 1000 simulation runs, the box-plots for 1000 values of absolute biases, defined by $|\widehat{\beta}_1 - \beta_1|$, are presented in Figs. 1 and

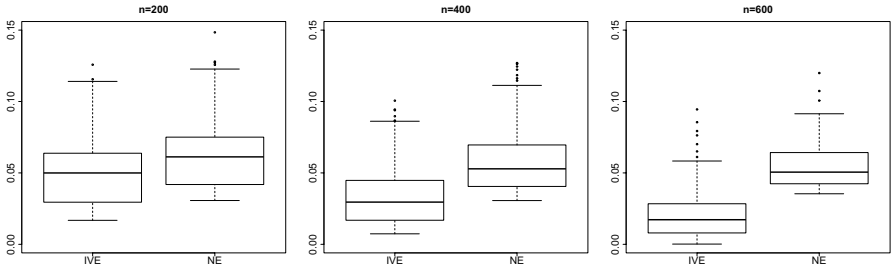

Fig. 1 The box-plots of 1000 absolute bias values for the estimator of β_1 based on the IVE method and the NE method under the endogenous level $\alpha = 0.2$

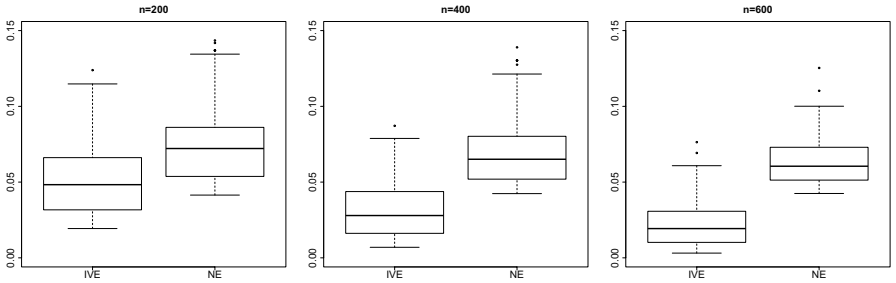

Fig. 2 The box-plots of 1000 absolute bias values for the estimator of β_1 based on the IVE method and the NE method under the endogenous level $\alpha = 0.8$

2, where Fig. 1 presents the simulation results under the endogenous level $\alpha = 0.2$, and Fig. 2 presents the simulation results under the endogenous level $\alpha = 0.8$. From Figs. 1 and 2, we can make the following observations:

(i) The absolute bias values, obtained by the proposed IVE method, decease as the sample size increases. However, the absolute bias values, obtained by the NE method, are still larger even though the sample size increases. This implies that the estimator based on the IVE method is consistent, and the estimator based on the NE method will give an endogeneity bias.
(ii) For given n, the IVE estimation procedure performs similar for different levels of endogeneity of covariates. This indicates that the proposed IVE estimation procedure can attenuate the effect of the endogeneity of covariates.

In addition, the simulation results for the nonparametric component $g(u)$ when $n = 400$ are shown in Fig. 3, where Fig3a presents the results under the endogenous level $\alpha = 0.2$, Fig. 3b presents the results under the endogenous level $\alpha = 0.8$. In Fig. 3, the dashed curve means the estimator based on the proposed IVE method, the dotted curve means the estimator based on the NE method, and the solid curve means the real curve of $g(u)$.

From Fig. 3, we can see that the estimator, obtained by the NE method, is biased, and the estimator, obtained by the proposed IVE method, can attenuate the endogenous

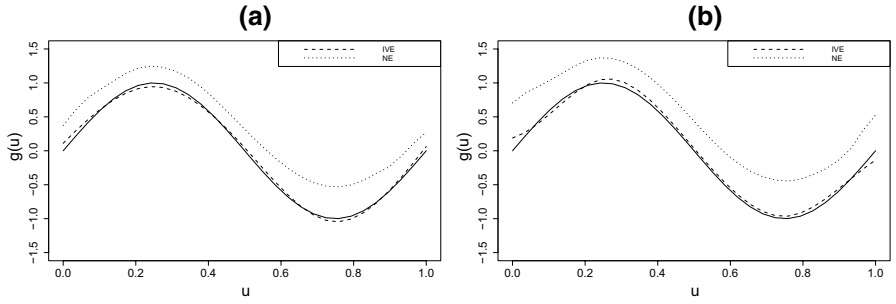

Fig. 3 The estimator of $g(u)$ under the endogenous levels $\alpha = 0.2$ (**a**) and $\alpha = 0.8$ (**b**), respectively

biases. In addition, we also can see that the estimators, obtained by the proposed IVE method, are similar for different levels of endogeneity of covariates. This indicates that the endogeneity in parametric component still affect the estimation for nonparametric components, and the proposed IVE method is also workable for the statistical inferences of the nonparametric component $g(u)$.

5.2 Real data analysis

We analyze a data set from the National Longitudinal Survey of Young Men (NLSYM) to illustrate the estimation procedure proposed by this paper. This data set contains 3010 observations from the NLSYM in 1976, and has been studied by many authors (see Card 1995; Zhao and Xue 2013; Huang and Zhao 2018). The objective of the study is to evaluate the effects of individual's education and work experience on individual's wage. More details for this data description and analysis can be seen in Card (1995).

Similar to Zhao and Xue (2013), we consider the following partially linear regression model

$$logwage = \beta_1 educ + \beta_2 educ^2 + \beta_3 educ^3 + \beta_4 black + \beta_5 south + \beta_6 smsa + g(exper) + \epsilon,$$

where *logwage* is the log of individual's hourly wage in cents, *educ* is the years of individual's schooling, $educ^2$ and $educ^3$ represent the quadratic and cubic effects, respectively, *exper* is the individual's work experience constructed as $age - educ - 6$, and *black*, *south*, and *smsa* (Standard Metropolitan Statistical Area) are dummy variables whose detailed description can be seen in Card (1995). In practice, the years of individual's schooling may be correlative with some factors in model errors, such as individual's intelligence quotient. Hence, similar to Card (1995), we take the variable *educ* as an endogenous variable, and use the proximity to a 4-year college as an instrumental variable for *educ*.

In addition, in order to demonstrate the performance of the proposed variable selection method for optimal instrumental variables, we add 99 invalid instrumental variables to have in total $q_n = 100$ instrumental variables. More specifically, we set $Z = (Z_1, \ldots, Z_{100})^T$, where Z_1 is the valid instrumental variable "*the proximity to a 4-year college*", and $Z_2, \ldots, Z_{100}$ are all invalid instrumental variables, which are independently sampled from the standard normal distribution $N(0, 1)$. The penalty function is taken as the SCAD penalty, Lasso penalty and MCP penalty, respectively.

Since some noise instrument variables are randomly generated, we repeated the optimal instrument variable selection procedure 1000 times. The simulation results for instrument variable selection are shown in Table 3, where "avg. num." means the average number of selected optimal instrument variables with 1000

Table 3 Application to NLSYM data. The identification of optimal instrumental variables under different penalties

penalty function	LAD-SCAD	LAD-Lasso	LAD-MCP
avg. num.	1.009	1.012	1.014
selected times	1000	1000	1000

Table 4 Application to NLSYM data. The regularized estimators of parametric component β based on the proposed method under different penalties

β	LAD-SCAD	LAD-Lasso	LAD-MCP
β_1	0.084	0.086	0.085
β_2	0	0	0
β_3	0	0	0
β_4	−0.187	−0.187	−0.183
β_5	−0.123	−0.125	−0.124
β_6	0.163	0.161	0.165

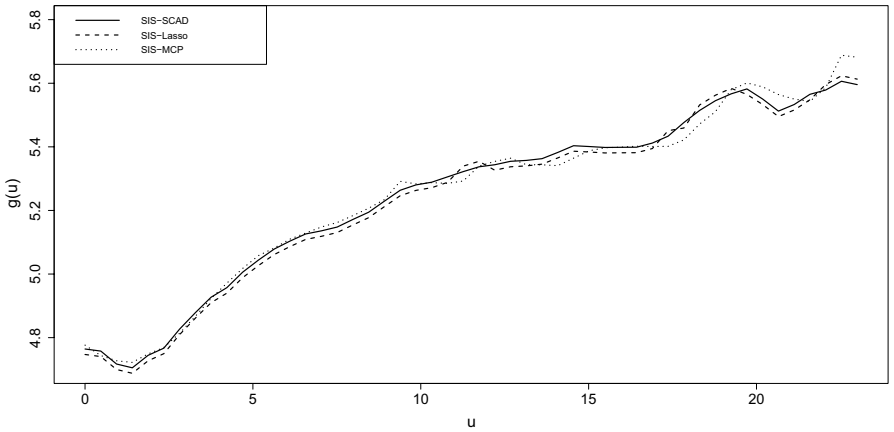

Fig. 4 The estimated curves of $g(u)$ with the SCAD penalty (solid curve), Lasso penalty (dashed curve) and MCP penalty (dotted curve), respectively

simulation runs, "selected times" means the times of the true optimal instrumental variables were selected in the final model over the 1000 simulation runs. From Table 3, we can see that average number of selected optimal instrument variables is very close to the true number 1 for all penalties, and the true optimal instrument variable can be always selected in all simulation runs. In addition, the simulation results under different penalties are similar in term of the identification of optimal instrument variables, which implies that the proposed method is insensitive to the penalty function.

The regularized estimators of parametric component β are shown in Table 4. From Table 4, we can see that β_2 and β_3 are zero, which indicates that the quadratic and cubic effects of *educ* have no significant impact on individual's wage. In addition, the estimated curves of $g(u)$ are shown in Fig. 4, and the results show that these estimated curves are also similar for differen penalties.

Acknowledgements This research is supported by the National Social Science Foundation of China (No. 18BTJ035).

Compliance with ethical standards

Conflicts of interest The authors declare that there is no conflict of interests regarding the publication of this paper.

Appendix. Proof of theorems

In this Appendix, we provide the proof details of Theorems 1–4 in this paper.

Proof of Theorem 1 Let $\delta_n = \sqrt{q_n/n}$ and $\theta = \theta_0 + \delta_n M$. We first show that, for any given $\varepsilon > 0$, there exists a large constant C such that

$$P\left\{ \inf_{\|M\|=C} Q_n(\theta) > Q_n(\theta_0) \right\} \geq 1 - \varepsilon. \tag{12}$$

Let $\Delta_n(\theta) = Q_n(\theta) - Q_n(\theta_0)$, then, invoking $\theta_{0k} = 0$ with $k \in \mathscr{A}_2$, $p_{\lambda_{1n}}(0) = 0$ and model (4), some simple calculations yield

$$\begin{aligned}
\Delta_n(\theta) =& \frac{1}{n}\sum_{i=1}^{n}\left|Y_i - Z_i^T\theta\right| - \frac{1}{n}\sum_{i=1}^{n}\left|Y_i - Z_i^T\theta_0\right| + \sum_{k=1}^{q_n}[p_{\lambda_{1n}}(|\theta_k|) - p_{\lambda_{1n}}(|\theta_{0k}|)] \\
=& \frac{1}{n}\sum_{i=1}^{n}\left|\varepsilon_i - \delta_n Z_i^T M\right| - \frac{1}{n}\sum_{i=1}^{n}|\varepsilon_i| + \sum_{k=1}^{q_n}[p_{\lambda_{1n}}(|\theta_k|) - p_{\lambda_{1n}}(|\theta_{0k}|)] \\
\geq& \frac{1}{n}\sum_{i=1}^{n}\left[\left|\varepsilon_i - \delta_n Z_i^T M\right| - |\varepsilon_i|\right] + \sum_{k\in\mathscr{A}_1}[p_{\lambda_{1n}}(|\theta_k|) - p_{\lambda_{1n}}(|\theta_{0k}|)] \\
\equiv& I_{n1} + I_{n2}.
\end{aligned} \tag{13}$$

We first consider I_{n1}. From Knight (1998), we have the following identity:

$$|a-b| - |a| = -b[I(a>0) - I(a<0)] + 2\int_0^b [I(a \leq s) - I(a \leq 0)]ds.$$

Hence, we have

$$\begin{aligned}
I_{n1} =& -\frac{1}{n}\sum_{i=1}^{n}\delta_n Z_i^T M[I(\varepsilon_i > 0) - I(\varepsilon_i < 0)] \\
&+ \frac{2}{n}\sum_{i=1}^{n}\int_0^{\delta_n Z_i^T M}[I(\varepsilon_i \leq s) - I(\varepsilon_i \leq 0)]ds \\
\equiv& I_{n3} + I_{n4}.
\end{aligned} \tag{14}$$

From condition (C3), we have $E[I(\varepsilon_i > 0) - I(\varepsilon_i < 0)] = 0$. Hence, invoking condition (C6), we can prove

$$E(I_{n3}) = -\frac{1}{n}\sum_{i=1}^{n}\delta_n E(Z_i)^T ME[I(\varepsilon_i > 0) - I(\varepsilon_i < 0)] = 0,$$
$$Var(I_{n3}) = \frac{\delta_n^2}{n^2}\sum_{i=1}^{n} M^T E(Z_i Z_i^T) ME[I(\varepsilon_i > 0) - I(\varepsilon_i < 0)]^2 \le \frac{\delta_n^2 \rho_2}{n}\|M\|^2.$$

Hence by the Markov inequality, we obtain

$$P(|I_{n3}| \ge \delta_n^2\|M\|) \le \frac{E(I_{n3}^2)}{\delta_n^4\|M\|^2} \le \frac{\delta_n^2\rho_2\|M\|^2}{n\delta_n^4\|M\|^2} \to 0.$$

This implies that

$$I_{n3} = o_p(\delta_n^2)\|M\|. \tag{15}$$

Next we consider I_{n4}. We denote

$$S_{ni} = \frac{2}{n}\int_0^{\delta_n Z_i^T M} [I(\varepsilon_i \le s) - I(\varepsilon_i \le 0)]ds.$$

Then

$$I_{n4} = \sum_{i=1}^{n} S_{ni} = \sum_{i=1}^{n}[S_{ni} - E(S_{ni})] + \sum_{i=1}^{n} E(S_{ni}) \equiv I_{n5} + I_{n6}. \tag{16}$$

Note that

$$\begin{aligned} nE(S_{ni}^2) =& n\frac{4}{n^2}E\left\{\left[\int_0^{\delta_n Z_i^T M}[I(\varepsilon_i \le s) - I(\varepsilon_i \le 0)]ds\right]^2\right\} \\ \le& \frac{4}{n}E\left\{\left[\int_0^{|\delta_n Z_i^T M|} ds\right]^2\right\} \\ =& \frac{4}{n}E\left\{|\delta_n Z_i^T M|^2\right\} \le \frac{4\delta_n^2}{n}M^T E(Z_i Z_i^T)M \le \frac{4\delta_n^2}{n}\rho_2\|M\|^2. \end{aligned}$$

Then we obtain

$$P(|I_{n5}| \ge \delta_n^2) \le \frac{Var(\sum_{i=1}^n S_{ni})}{\delta_n^4} \le \frac{nE(S_{ni}^2)}{\delta_n^4} \to 0.$$

This implies $I_{n5} = o_p(\delta_n^2)$. In addition, by the dominated convergence theorem, we can obtain

$$
\begin{aligned}
I_{n6} &= 2E\left\{\int_0^{\delta_n Z_i^T M}[I(\varepsilon_i \le s) - I(\varepsilon_i \le 0)]ds\right\} \\
&= 2E\left\{E\left[\int_0^{\delta_n Z_i^T M}[I(\varepsilon_i \le s) - I(\varepsilon_i \le 0)]ds \middle| Z_i\right]\right\} \\
&= 2E\left\{\int_0^{\delta_n Z_i^T M}[F(s) - F(0)]ds\right\} \\
&= 2f(0)E\left\{(1+o(1))\int_0^{\delta_n Z_i^T M} sds\right\} \\
&= f(0)(1+o(1))\delta_n^2 M^T E(Z_i Z_i^T)M \\
&= O_p(\delta_n^2)\|M\|^2.
\end{aligned} \tag{17}
$$

Next we consider the term I_{n2}. Invoking condition (C8), some calculations yield

$$
\begin{aligned}
|I_{n2}| &= \sum_{k\in\mathscr{A}_1}[p_{\lambda_{1n}}(|\theta_k|) - p_{\lambda_{1n}}(|\theta_{0k}|)] \\
&= \sum_{k\in\mathscr{A}_1}\delta_n p'_{\lambda_{1n}}(|\theta_{0k}|)sgn(\theta_{0k})|M_k| + \sum_{k\in\mathscr{A}_1}\delta_n^2 p''_{\lambda_{1n}}(|\theta_{0k}|)sgn(\theta_{0k})|M_k|^2(1+o(1)) \\
&\le \sqrt{s}\delta_n a_n\|M\| + \delta_n^2 b_n\|M\|^2 \\
&= o_p(\delta_n^2)\|M\|^2.
\end{aligned} \tag{18}
$$

Then, by choosing a large C, all terms I_{n2}, I_{n3} and I_{n5} are dominated by I_{n6} with $\|M\| = C$. Note that I_{n6} is positive, then invoking (13–18), we obtain that (12) holds. Furthermore, by the convexity of $Q_n(\cdot)$, we have

$$
P\left\{\inf_{\|M\|\le C} Q_n(\theta) > Q_n(\theta_0)\right\} \ge 1-\varepsilon.
$$

This implies, with probability at least $1-\varepsilon$, that there exists a local minimizer $\widehat{\theta}$ such that $\widehat{\theta} - \theta_0 = O_p(\delta_n)$, which completes the proof of Theorem 1. □

Proof of Theorem 2 For convenience and simplicity, let $\theta_0 = (\theta_{\mathscr{A}_1}^T, \theta_{\mathscr{A}_2}^T)^T$ with $\theta_{\mathscr{A}_1} = \{\theta_{0k} : k \in \mathscr{A}_1\}$ and $\theta_{\mathscr{A}_2} = \{\theta_{0k} : k \in \mathscr{A}_2\}$. The corresponding covariate is denoted by $Z_i = (Z_i^{(1)T}, Z_i^{(2)T})^T$. From the proof of Theorem 1, for a sufficiently large C, $\widehat{\theta}$ lies in the ball $\{\theta_0 + \delta_n M : \|M\| \le C\}$ with probability converging to 1, where $\delta_n = \sqrt{q_n/n}$. We denote $\theta_1 = \theta_{\mathscr{A}_1} + \delta_n M_1$ and $\theta_2 = \theta_{\mathscr{A}_2} + \delta_n M_2$ with $\|M_1\|^2 + \|M_2\|^2 \le C^2$, and $V_n(M_1, M_2) = Q_n(\theta_1, \theta_2) - Q_n(\theta_{\mathscr{A}_1}, 0)$, then the estimator $\widehat{\theta} = (\widehat{\theta}_1^T, \widehat{\theta}_2^T)^T$ can also be obtained by minimizing $V_n(M_1, M_2)$, except on an event with probability tending to zero. Hence, to prove this theorem, we only need to prove that, for any M_1 and M_2 satisfying $\|M_1\|^2 + \|M_2\|^2 \le C^2$, if $\|M_2\| > 0$, then with probability tending to 1, we have

$$V_n(M_1, M_2) - V_n(M_1, 0) > 0. \tag{19}$$

Note that

$$\begin{aligned}
&V_n(M_1, M_2) - V_n(M_1, 0) \\
&= Q_n(\theta_1, \theta_2) - Q_n(\theta_{\mathscr{A}_1}, 0) - [Q_n(\theta_1, 0) - Q_n(\theta_{\mathscr{A}_1}, 0)] \\
&= \left\{ \frac{1}{n}\sum_{i=1}^{n} \left| Y_i - Z_i^{(1)T}\theta_{\mathscr{A}_1} - \delta_n Z_i^{(1)T} M_1 - \delta_n Z_i^{(2)T} M_2 \right| - \frac{1}{n}\sum_{i=1}^{n} \left| Y_i - Z_i^{(1)T}\theta_{\mathscr{A}_1} \right| \right\} \\
&\quad - \left\{ \frac{1}{n}\sum_{i=1}^{n} \left| Y_i - Z_i^{(1)T}\theta_{\mathscr{A}_1} - \delta_n Z_i^{(1)T} M_1 \right| - \frac{1}{n}\sum_{i=1}^{n} \left| Y_i - Z_i^{(1)T}\theta_{\mathscr{A}_1} \right| \right\} \\
&\quad + \sum_{k=\mathscr{A}_2} p_{\lambda_{1n}}(|\theta_{2k}|) \\
&= \left\{ \frac{1}{n}\sum_{i=1}^{n} \left| \varepsilon_i - \delta_n Z_i^{(1)T} M_1 - \delta_n Z_i^{(2)T} M_2 \right| - \frac{1}{n}\sum_{i=1}^{n} |\varepsilon_i| \right\} \\
&\quad - \left\{ \frac{1}{n}\sum_{i=1}^{n} \left| \varepsilon_i - \delta_n Z_i^{(1)T} M_1 \right| - \frac{1}{n}\sum_{i=1}^{n} |\varepsilon_i| \right\} + \sum_{k\in\mathscr{A}_2} p_{\lambda_{1n}}(|\theta_{2k}|).
\end{aligned} \tag{20}$$

Similar to the proof of Theorem 1, we can obtain

$$\begin{aligned}
&\left\{ \frac{1}{n}\sum_{i=1}^{n} \left| \varepsilon_i - \delta_n Z_i^{(1)T} M_1 - \delta_n Z_i^{(2)T} M_2 \right| - \frac{1}{n}\sum_{i=1}^{n} |\varepsilon_i| \right\} \\
&\qquad - \left\{ \frac{1}{n}\sum_{i=1}^{n} \left| \varepsilon_i - \delta_n Z_i^{(1)T} M_1 \right| - \frac{1}{n}\sum_{i=1}^{n} |\varepsilon_i| \right\} \\
&\geq O_p(\delta_n^2) + \frac{f(0)}{2}\delta_n^2 M^T E(ZZ^T) M + O_p(\delta_n^2) - \frac{f(0)}{2}\delta_n^2 M_1^T E\left(Z^{(eq1)} Z^{(1)T} \right) M_1 \\
&= O_p(\delta_n^2) + O_p(\delta_n^2) f(0) \|M\|^2.
\end{aligned} \tag{21}$$

In addition, for $k \in \mathscr{A}_2$, we have $\theta_{0k} = 0$. Then invoking $p_{\lambda_n}(0) = 0$, we can derive

$$\begin{aligned}
\sum_{k\in\mathscr{A}_2} p_{\lambda_{1n}}(|\theta_{2k}|) &= \sum_{k\in\mathscr{A}_2} p_{\lambda_{1n}}(|\theta_{0k} + \delta_n M_{2k}|) \\
&= \sum_{k\in\mathscr{A}_2} p_{\lambda_{1n}}(|\theta_{0k}|) + \sum_{k\in\mathscr{A}_2} p'_{\lambda_{1n}}(|\theta_{0k}|)\delta_n |M_{2k}| + O_p(\delta_n^2) \sum_{k\in\mathscr{A}_2} |M_{2k}|^2 \\
&= \delta_n p'_{\lambda_{1n}}(0) \sum_{k\in\mathscr{A}_2} |M_{2k}| + O_p(\delta_n^2)\|M\|^2.
\end{aligned} \tag{22}$$

Hence, from (20–22), we have

$$V_n(M_1, M_2) - V_n(M_1, 0)$$
$$\geq \delta_n \lambda_{1n} \left(O_p(\sqrt{q_n/n}/\lambda_{1n}) f(0) \|M\|^2 + p'_{\lambda_{1n}}(0)/\lambda_{1n} \sum_{k \in \mathscr{A}_2} |M_{2k}| \right). \tag{23}$$

By conditions (C7) and (C8), we have $\sqrt{q_n/n}/\lambda_n \to 0$ and $p'_{\lambda_{1n}}(0)/\lambda_{1n} > 0$. Hence, (23) implies that (19) holds with probability tending to 1. This completes the proof of Theorem 2. □

Proof of Theorem 3 Note that Theorem 2 implies that the variable selection for optimal instrumental variables is consistent, then model (6) implies that, with probability tending to 1, we have $X_i = \Gamma_{\mathscr{A}_1} Z_i^* + e_i$, $i = 1, \ldots, n$. In addition, because $\widehat{\Gamma}$ is the moment estimator of $\Gamma_{\mathscr{A}_1}$, we can prove $\widehat{\Gamma} = \Gamma_{\mathscr{A}_1} + O_p(\sqrt{p_n/n})$. Hence, invoking $E(e_i) = 0$, a simple calculation yields

$$\begin{aligned} \frac{1}{n} \sum_{i=1}^{n} (X_i - X_i^*)^T \beta =& \frac{1}{n} \sum_{i=1}^{n} (\Gamma_{\mathscr{A}_1} Z_i^* + e_i - \widehat{\Gamma} Z_i^*)^T \beta \\ =& \frac{1}{n} \sum_{i=1}^{n} Z_i^{*T} (\Gamma_{\mathscr{A}_1} - \widehat{\Gamma})^T \beta + \frac{1}{\sqrt{n}} \left(\frac{1}{\sqrt{n}} \sum_{i=1}^{n} e_i^T \beta \right) \\ =& O_p(\|\Gamma_{\mathscr{A}_1} - \widehat{\Gamma}\|) + O_p(n^{-1/2}) = O_p(\sqrt{p_n/n}). \end{aligned} \tag{24}$$

Furthermore, we let β_0 and γ_0 be the true values of β and γ, respectively, and denote $R(U_i) = g(U_i) - W_i^T \gamma_0$. Then from Schumaker (1981), we have $\|R(U_i)\| = O_p(\kappa_n^{-r}) = O_p(\sqrt{\kappa_n/n})$. Hence, invoking (24), some calculations yield

$$\begin{aligned} M_n(\beta, \gamma) =& \frac{1}{n} \sum_{i=1}^{n} \Big| X_i^T (\beta_0 - \beta) + (X_i - X_i^*)^T \beta + W_i^T (\gamma_0 - \gamma) \\ & + R(U_i) + \varepsilon_i \Big| + \sum_{j=1}^{p_n} p_{\lambda_{2n}}(|\beta_j|) \\ =& \frac{1}{n} \sum_{i=1}^{n} \Big| X_i^T (\beta_0 - \beta) + W_i^T (\gamma_0 - \gamma) + \varepsilon_i \Big| + \sum_{j=1}^{p_n} p_{\lambda_{2n}}(|\beta_j|) \\ & + O_p\Big(\|\hat{\Gamma} - \Gamma_{\mathscr{A}_1}\| + \|R(U_i)\| \Big) \\ =& \frac{1}{n} \sum_{i=1}^{n} \Big| X_i^T (\beta_0 - \beta) + W_i^T (\gamma_0 - \gamma) + \varepsilon_i \Big| + \sum_{j=1}^{p_n} p_{\lambda_{2n}}(|\beta_j|) + O_p(\delta_n), \end{aligned} \tag{25}$$

where $\delta_n = \sqrt{(p_n + \kappa_n)/n}$. Furthermore, we denote $\alpha_0 = (\beta_0^T, \gamma_0^T)^T$ and $\alpha = (\beta^T, \gamma^T)^T$ with $\alpha = \alpha_0 + \delta_n M$, where M is a $(p_n + L_n)$ dimensional vector. Then (25) implies that

$$M_n(\beta,\gamma)=\frac{1}{n}\sum_{i=1}^{n}\left|\varepsilon_i-\delta_n\xi_i^TM\right|+\sum_{j=1}^{p_n}p_{\lambda_{2n}}\left(|\beta_j|\right)+O_p(\delta_n), \tag{26}$$

and

$$M_n(\beta_0,\gamma_0)=\frac{1}{n}\sum_{i=1}^{n}|\varepsilon_i|+\sum_{j=1}^{p_n}p_{\lambda_{2n}}\left(|\beta_{0j}|\right)+O_p(\delta_n), \tag{27}$$

where $\xi_i=(X_i^T,W_i^T)^T$. Furthermore, we let $\Delta_n(\beta,\gamma)=M_n(\beta,\gamma)-M_n(\beta_0,\gamma_0)$, then from (26) and (27), we have

$$\Delta_n(\beta,\gamma)=\frac{1}{n}\sum_{i=1}^{n}\left[|\varepsilon_i-\delta_n\xi_i^TM|-|\varepsilon_i|\right]+\sum_{j=1}^{p_n}\left[p_{\lambda_{2n}}(|\beta_j|)-p_{\lambda_{2n}}(|\beta_{0j}|)\right]+O_p(\delta_n). \tag{28}$$

Hence invoking (28), and using the similar arguments to the proof of (13), we have that, for any given $\varepsilon>0$, there exists a large constant C such that

$$P\left\{\inf_{\|M\|=C}M_n(\beta,\gamma)>M_n(\beta_0,\gamma_0)\right\}\geq 1-\varepsilon.$$

This implies, with probability at least $1-\varepsilon$, that there exists a local minimizer $\widehat{\beta}$ and $\widehat{\gamma}$, which satisfy $\widehat{\beta}-\beta_0=O_p(\sqrt{(p_n+\kappa_n)/n})$ and $\widehat{\gamma}-\gamma_0=O_p(\sqrt{(p_n+\kappa_n)/n})$. Then, we complete the proof of part (i) in Theorem 3. □

In addition, invoking the proof of part (i), and using the same arguments as the proof of Theorem 2, we can prove part (ii) in Theorem 3. Then we omit the proof procedure of part (ii) in detail.

Proof of Theorem 4 A simple calculation yields

$$\begin{aligned}\|\widehat{g}(u)-g(u)\|^2 &= \int_0^1\{\widehat{g}(u)-g(u)\}^2du\\ &= \int_0^1\{B^T(u)\widehat{\gamma}-B^T(u)\gamma_0+R(u)\}^2du\\ &\leq 2\int_0^1\{B^T(u)\widehat{\gamma}-B^T(u)\gamma_0\}^2du+2\int_0^1R(u)^2du\\ &= 2(\widehat{\gamma}-\gamma_0)^TH(\widehat{\gamma}-\gamma_0)+2\int_0^1R(u)^2du,\end{aligned} \tag{29}$$

where $R(u)=g(u)-B^T(u)\gamma_0$ and $H=\int_0^1B(u)B^T(u)du$. From the proof of Theorem 3, we can obtain $\|\widehat{\gamma}-\gamma_0\|=O_p(\sqrt{(p_n+\kappa_n)/n})$. Then from condition (C7) and $\kappa_n=O(1/(2r+1))$, we can prove $O_p(\sqrt{(p_n+\kappa_n)/n})=O_p(\sqrt{\kappa_n/n})=O_p(n^{-r/(2r+1)})$. Then, invoking $\|H\|=O(1)$, a simple calculation yields

$$(\widehat{\gamma}_k - \gamma_{k0})^T H(\widehat{\gamma}_k - \gamma_{k0}) = O_p\left(n^{\frac{-2r}{2r+1}}\right). \tag{30}$$

In addition, from conditions C1, C4 and Corollary 6.21 in Schumaker (1981), we can obtain $R(u) = O(\kappa_n^{-r}) = O(n^{-r/(2r+1)})$. Then, it is easy to show that

$$\int_0^1 R(u)^2 du = O_p\left(n^{\frac{-2r}{2r+1}}\right). \tag{31}$$

Invoking (29–31), we complete the proof Theorem 4. □

References

Cai, Z., & Xiong, H. (2012). Partially varying coefficient instrumental variables models. *Statistca Neerlandica*, *66*, 85–110.

Card, D. (1995). Using geographic variation in college proximity to estimate the return to schooling. In L. Christofides, E. Grant, & R. Swidinsky (Eds.), *Aspects of Labor Market Behaviour: Essays in Honour of John Vanderkamp* (pp. 201–222). Toronto: University of Toronto Press.

Chen, B. C., Liang, H., & Zhou, Y. (2016). GMM estimation in partial linear models with endogenous covariates causing an over-identified problem. *Communications in Statistics - Theory and Methods*, *45*, 3168–3184.

Didelez, V., Meng, S., & Sheehan, N. A. (2010). Assumptions of IV methods for observational epidemiology. *Statistical Science*, *25*, 22–40.

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

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

Fan, J. Q., & Liao, Y. (2014). Endogeneity in dimensions. *The Annals of Statistics*, *42*, 872–917.

Frank, I. E., & Friedman, J. H. (1993). A statistical view of some chemometrics regression tools. *Technometrics*, *35*, 109–135.

Gao, X., & Huang, J. (2010). Asymptotic analysis of high-dimensional lad regression with lasso. *Statistica Sinica*, *20*, 1485–1506.

Greenland, S. (2000). An introduction to instrumental variables for epidemiologists. *International Journal of Epidemiologists*, *29*, 722–729.

Hernan, M. A., & Robins, J. M. (2006). Instruments for causal inference-an epidemiologists dream? *Epidemiology*, *17*, 360–372.

Huang, J. T., & Zhao, P. X. (2017). QR decomposition based orthogonality estimation for partially linear models with longitudinal data. *Journal of Computational and Applied Mathematics*, *321*, 406–415.

Huang, J. T., & Zhao, P. X. (2018). Orthogonal weighted empirical likelihood based variable selection for semiparametric instrumental variable models. *Communications in Statistics-Theory and Methods*, *47*, 4375–4388.

Knight, K. (1998). Limiting distributions for L_1 regression estimators under general conditions. *The Annals of Statistics*, *26*, 755–770.

Koenker, R. (2005). *Quantile Regression*. Cambridge: Cambridge University Press.

Lee, E. R., Cho, J., & Yu, K. (2019). A systematic review on model selection in high-dimensional regression. *Journal of the Korean Statistical Society*, *48*, 1–12.

Lin, W., Feng, R., & Li, H. Z. (2015). Regularization methods for high-dimensional instrumental variables regression with an application to genetical genomics. *Journal of the American Statistical Association*, *110*, 270–288.

Liu, J. Y., Lou, L. J., & Li, R. Z. (2018). Variable selection for partially linear models via partial correlation. *Journal of Multivariate Analysis*, *167*, 418–434.

Newhouse, J. P., & McClellan, M. (1998). Econometrics in outcomes research: the use of instrumental variables. *Annual Review of Public Health*, *19*, 17–24.

Schumaker, L. L. (1981). *Spline Function*. New York: Wiley.
Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society Series B*, *58*, 267–288.
Wang, H., Li, G., & Jiang, G. (2007). Robust regression shrinkage and consistent variable selection through the lad-lasso. *Journal of Business & Economic Statistics*, *25*, 347–355.
Wang, M. Q., Song, L. X., & Tian, G. L. (2015). SCAD-penalized least absolute deviation regression in high-dimensional models. *Communications in Statistics-Theory and Methods*, *44*, 2452–2472.
Windmeijer, F., Farbmacher, H., Davies, N., & Smith, G. D. (2019). On the use of the Lasso for instrumental variables estimation with some invalid instruments. *Journal of the American Statistical Association*, *114*, 1339–1350.
Xue, L. G., & Zhu, L. X. (2007). Empirical likelihood semiparametric regression analysis for longitudinal data. *Biometrika*, *94*, 921–937.
Xie, H., & Huang, J. (2009). SCAD-penalized regression in high-dimensional partially linear models. *The Annals of Statistics*, *37*, 673–696.
Yang, Y. P., Chen, L. F., & Zhao, P. X. (2017). Empirical likelihood inference in partially linear single index models with endogenous covariates. *Communications in Statistics-Theory and Methods*, *46*, 3297–3307.
Yuan, J. Y., Zhao, P. X., & Zhang, W. G. (2016). Semiparametric variable selection for partially varying coefficient models with endogenous variables. *Computational Statistics*, *31*, 693–707.
Zhang, C. H. (2010). Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics*, *38*, 894–942.
Zhao, P. X., & Li, G. R. (2013). Modified SEE variable selection for varying coefficient instrumental variable models. *Statistical Methodology*, *12*, 60–70.
Zhao, P. X., & Xue, L. G. (2013). Empirical likelihood inferences for semiparametric instrumental variable models. *Journal of Applied Mathematics and Computing.*, *43*, 75–90.
Zou, H., & Li, R. (2008). One-step sparse estimates in nonconcave penalized likelihood models. *The Annals of Statistics*, *36*, 1509–1533.