Comput Stat (2016) 31:693–707
DOI 10.1007/s00180-015-0601-y

ORIGINAL PAPER

Semiparametric variable selection for partially varying coefficient models with endogenous variables

Jinyi Yuan[1] · **Peixin Zhao**[2] · **Weiguo Zhang**[1,3]

Received: 10 December 2014 / Accepted: 22 June 2015 / Published online: 6 July 2015

Abstract By using instrumental variable technology and the partial group smoothly clipped absolute deviation penalty method, we propose a variable selection procedure for a class of partially varying coefficient models with endogenous variables. The proposed variable selection method can eliminate the influence of the endogenous variables. With appropriate selection of the tuning parameters, we establish the oracle property of this variable selection procedure. A simulation study is undertaken to assess the finite sample performance of the proposed variable selection procedure.

Keywords Partially varying coefficient model · Variable selection · Endogenous variable

1 Introduction

In many disciplines, some covariates may be endogenous in regression modeling. In this situation, the estimator based on the classical method, such as the ordinary least squares method, is not consistent any more (see Newhouse and McClellan 1998; Greenland 2000; Hernan and Robins 2006). The instrumental variable method provides a way to correct the possible endogeneity between covariates and structural errors, and can obtain consistent parameter estimators. Recently, this method has been

✉ Peixin Zhao
zpx81@163.com

1 College of Economics and Business Administration, Chongqing University, Chongqing 400044, China

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

3 College of Economics and Management, Southwest University, Chongqing 400715, China

widely used in applied statistics, econometrics, and more generally related disciplines. Because the linear instrumental variable model, which assumes that the coefficients of all covariates are constant, is sometimes too restrictive for real economic models (see Schultz 1997; Card 2001), many papers have considered the statistical inferences for semiparametric models. For example, Yao (2012) considered the efficient estimation for partially linear instrumental variable models, and proposed a semiparametric instrumental variable estimation procedure. Zhao and Xue (2013) considered the confidence region construction for regression coefficients in partially linear instrumental variable models based on the empirical likelihood method. Zhao and Li (2013) considered the variable selection for varying coefficient instrumental variable models by using the smooth-threshold estimating equations method. The varying coefficient instrumental variable model allows the effect of endogenous covariates to be varying with a covariate, and is commonly used for analysis of data measured repeatedly over time, such as time series analysis, longitudinal data analysis and functional data analysis. In practice, however, only some of the coefficients vary with certain covariate, hence one useful extension of the varying coefficient instrumental variable model is the partially varying coefficient model with endogenous variables.

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

where $\theta(u) = (\theta_1(u), \ldots, \theta_p(u))^T$ is a $p \times 1$ vector of unknown functions, $\beta = (\beta_1, \ldots, \beta_q)^T$ is a $q \times 1$ vector of unknown parameters, Γ is a $q \times k$ matrix of unknown parameters. Y_i is the response variable, and ε_i and e_i are zero-mean model errors. Furthermore, we assume that X_i and U_i are exogenous covariates, Z_i is the endogenous covariate, and ξ_i is the corresponding instrumental variable. This implies that the covariate Z_i is correlated with the model error ε_i, but X_i, U_i and ξ_i are uncorrelated with ε_i. Then we have

$$E(\varepsilon_i | Z_i) \neq 0, \quad \text{and} \quad E(\varepsilon_i | X_i, U_i, \xi_i) = 0.$$

Model (1) is more flexible, and the linear instrumental variable model, the partially linear instrumental variable model and the varying coefficient instrumental variable model are all special cases of model (1). For model (1), Cai and Xiong (2012) considered the efficient estimation problem, and proposed a three-step estimation procedure to estimate the parametric components and the nonparametric components. However, when the number of covariates in model (1) is the large, an important problem is to select the important variables in such model.

Variable selection is a very important topic in modern statistical inference. Recently, based on some penalty methods, many variable selection procedures have been proposed. For example, Frank and Friedman (1993) proposed a variable selection procedure based on the bridge regression technology. Tibshirani (1996) proposed a variable selection procedure based on the least absolute shrinkage and selection operator (LASSO) technology. Fan and Li (2001) proposed a variable selection procedure based on smoothly clipped absolute deviation penalty (SCAD), which include bridge

regression and LASSO penalty. Wang et al. (2008) extended the SCAD variable selection method to the varying coefficient model, and proposed a group SCAD (gSCAD) variable selection procedure. Zhao and Xue (2009) proposed a partial gSCAD variable selection method for the varying coefficient partially linear model. Recently, many papers considered the variable selection for varying coefficient models with high dimensional data. For example, Lin and Yuan (2012) considered the variable selection for generalized varying coefficient partially linear models with diverging number of parameters. Lian (2012) considered the variable selection for high-dimensional generalized varying coefficient models. Wang et al. (2013) considered the polynomial spline estimation for generalized varying coefficient partially linear models with a diverging number of components. However, for the case that some covariates are endogenous, these variable selection methods are not consistent, and can not be directly used any more.

To overcome this problem, in this paper, we extend the partial gSCAD variable selection method, used by Zhao and Xue (2009), to the varying coefficient partially linear regression model with endogenous covariates. We propose an instrumental variable based partial gSCAD variable selection procedure which can select significant variables in the parametric components and nonparametric components simultaneously. With the proper choice of regularization parameters, we show that the variable selection procedure is consistent, and the penalized estimators have the oracle property in the sense of Fan and Li (2001). In addition, it is noteworthy that the proposed method can attenuate the effect of the endogeneity of covariates, which is an improvement of the variable selection method used in Zhao and Xue (2009).

The rest of this paper is organized as follows. In Sect. 2, we propose the instrumental variable based partial gSCAD variable selection procedure, and establish some asymptotic properties, including the consistency and the oracle property. In Sect. 3, based on the local quadratic approximation technology, we propose an iterative algorithm for finding the penalized estimators. In Sect. 4, some simulations are carried out to assess the performance of the proposed methods. Finally, the technical proofs of all asymptotic results are provided in "Appendix".

2 Methodology and main results

We let $B(u) = (B_1(u), \ldots, B_L(u))^T$ denote B-spline basis functions with the order of M, where $L = K + M + 1$, and K is the number of interior knots. Then, $\theta_k(u)$ can be approximated by

$$\theta_k(u) \approx B(u)^T \gamma_k, \quad k = 1, \ldots, p.$$

Substituting this into model (1), we can get

$$Y_i = W_i^T \gamma + Z_i^T \beta + \varepsilon_i, \tag{2}$$

where $W_i = I_p \otimes B(U_i) \cdot X_i$ and $\gamma = (\gamma_1^T, \ldots, \gamma_p^T)^T$. Model (2) is a standard linear regression model. Note that each function $\theta_k(u)$ in (1) is characterized by γ_k in (2).

Then, motivated by the idea of Zhao and Xue (2009), we propose the following partial gSCAD regularized estimation

$$Q(\gamma, \beta) = \sum_{i=1}^{n} \left\{ Y_i - W_i^T \gamma - Z_i^T \beta \right\}^2 + n \sum_{k=1}^{p} p_\lambda(\|\gamma_k\|_H) + n \sum_{l=1}^{q} p_\lambda(|\beta_l|), \quad (3)$$

where $\|\gamma_k\|_H = (\gamma^T H \gamma)^{1/2}$, $H = (h_{ij})_{L\times L}$ is a matrix with $h_{ij} = \int B_i(u) B_j(u) du$, and $p_\lambda(\cdot)$ is the SCAD penalty function with λ as a tuning parameter (see Fan and Li 2001), defined as

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

with $a > 2$, $w > 0$ and $p_\lambda(0) = 0$.

If $Z_i, i = 1, \ldots, n$ in model (1) are exogenous as well, then by Zhao and Xue (2009), it can be shown that we can get a consistent sparse solution by minimizing (3). However, Z_i, $i = 1, \ldots, n$ in model (1) are endogenous covariates, and then $E(\varepsilon_i | Z_i) \neq 0$. In this case, one can show that the resulting estimator, based on (3), is biased. Hence, (3) cannot be used directly to select the important variables and estimate regression coefficients any more.

Next, we propose an adjustment for (3) based on instrumental variables $\xi_i, i = 1, \ldots, n$. From model (1), we have $E(Z\xi^T) = \Gamma E(\xi\xi^T)$. Hence, the moment estimator of Γ can be given by

$$\hat{\Gamma} = \hat{\Gamma}_1 \hat{\Gamma}_2^{-1}$$

where

$$\hat{\Gamma}_1 = \frac{1}{n} \sum_{i=1}^{n} Z_i \xi_i^T, \quad \text{and} \quad \hat{\Gamma}_2 = \frac{1}{n} \sum_{i=1}^{n} \xi_i \xi_i^T,$$

By the proof in the "Appendix", we have $\hat{\Gamma} = \Gamma + o_p(1)$. Note that $E(Z_i|\xi_i) = \Gamma \xi_i$, then an unbiased adjustment of Z_i can be given by $\hat{Z}_i = \hat{\Gamma}\xi_i$. Hence, an instrumental variable based partial gSCAD regularized estimation function can be given by

$$\hat{Q}(\gamma, \beta) = \sum_{i=1}^{n} \left\{ Y_i - W_i^T \gamma - \hat{Z}_i^T \beta \right\}^2 + n \sum_{k=1}^{p} p_\lambda(\|\gamma_k\|_H) + n \sum_{l=1}^{q} p_\lambda(|\beta_l|). \quad (4)$$

Remark 1 Because the endogeneity of the covariate Z_i will result in the inconsistent estimation and variable selection, we replace Z_i in $Q(\gamma, \beta)$ by the adjustment $\hat{Z}_i$. Note that $\hat{\Gamma} = \Gamma + o_p(1)$, we have $\hat{Z}_i = \Gamma\xi_i + o_p(1)$. Hence, invoking that the instrumental variable ξ_i is an exogenous covariate, the following asymptotic results show that such an adjustment can attenuate the effect of endogenous covariates, and give a consistent regularity estimation procedure.

Let $\hat{\beta}$ and $\hat{\gamma} = (\hat{\gamma}_1^T, \ldots, \hat{\gamma}_p^T)^T$ be the solution by minimizing (4). Then, $\hat{\beta}$ is the penalized least squares estimator of β, and the estimator of $\theta_k(u)$ can be obtained by $\hat{\theta}_k(u) = B^T(u)\hat{\gamma}_k$.

Next, we study the asymptotic properties of the resulting penalized least squares estimators. Similar to Zhao and Xue (2009), we let $\theta_0(\cdot)$ and β_0 be the true value of $\theta(\cdot)$ and β respectively. Without loss of generality, we assume that $\beta_{l0} = 0,\ l = s+1, \ldots, q$, and $\beta_{l0},\ l = 1, \ldots, s$ are all nonzero components of β_0. Furthermore, we assume that $\theta_{k0}(\cdot) = 0,\ k = d+1, \ldots, p$, and $\theta_{k0}(\cdot),\ k = 1, \ldots, d$ are all nonzero components of $\theta_0(\cdot)$. Let

$$a_{1n} = \max_l \left\{ |p'_\lambda(|\beta_{l0}|)| : \beta_{l0} \neq 0 \right\}, \quad a_{2n} = \max_k \left\{ |p'_\lambda(\|\gamma_{k0}\|_H)| : \gamma_{k0} \neq 0 \right\},$$

and

$$b_{1n} = \max_l \left\{ |p''_\lambda(|\beta_{l0}|)| : \beta_{l0} \neq 0 \right\}, \qquad b_{2n} = \max_k \left\{ |p''_\lambda(\|\gamma_{k0}\|_H)| : \gamma_{k0} \neq 0 \right\}.$$

Furthermore, we let $a_n = \max\{a_{1n}, a_{2n}\}$ and $b_n = \max\{b_{1n}, b_{2n}\}$. Then, the following theorem gives the consistency of the penalized least squares estimators.

Theorem 1 *Suppose that the regularity conditions C1-C5 in "Appendix" hold and the number of knots $K = O_p(n^{1/(2r+1)})$, where r is defined in condition C1 in "Appendix". If $a_n \to 0$ and $b_n \to 0$, as $n \to \infty$, then,*

(i) $\|\hat{\beta} - \beta_0\| = O_p(n^{\frac{-r}{2r+1}} + a_n)$.
(ii) $\|\hat{\theta}_k(u) - \theta_{k0}(u)\| = O_p(n^{\frac{-r}{2r+1}} + a_n),\ \ k = 1, \ldots, p.$

Remark 2 For the SCAD penalty function that used in this paper, it is clear that $a_n = 0$ if $\lambda \to 0$ when n is large enough. Hence, under the regularity conditions defined in the "Appendix", the consistent penalized estimator indeed exists with probability tending to one.

Furthermore, under some conditions, we show that such consistent estimators must possess the sparsity property, which is stated as follows

Theorem 2 *Suppose that the regularity conditions in Theorem 1 hold, and*

$$\liminf_{n\to\infty} \liminf_{|\beta_l|\to 0} \lambda^{-1} p'_\lambda(|\beta_l|) > 0, \quad l = s+1, \ldots, q,$$
$$\liminf_{n\to\infty} \liminf_{\|\gamma_k\|_H \to 0} \lambda^{-1} p'_\lambda(\|\gamma_k\|_H) > 0, \quad k = d+1, \ldots, p.$$

If $n^{r/(2r+1)}\lambda \to \infty$ and $\lambda \to 0$, as $n \to \infty$. Then, with probability tending to 1, $\hat{\beta}$ and $\hat{\theta}(u)$ must satisfy

(i) $\hat{\beta}_l = 0, \quad l = s+1, \ldots, q.$
(ii) $\hat{\theta}_k(u) = 0, \quad k = d+1, \ldots, p.$

Remark 3 From remark 1 in Fan and Li (2001), we have that, if $\lambda \to 0$ as $n \to \infty$, then $a_n = 0$. Then from Theorems 1 and 2, it is clear that, by choosing a proper λ, the proposed variable selection method is consistent and the estimators achieve the convergence rate as if the subset of true zero coefficients is already known. This implies that the penalized estimators have the oracle property.

3 Algorithm

Note that the penalty function $p_\lambda(\cdot)$ in $\hat{Q}(\gamma, \beta)$ is irregular at the origin, then the classical gradient method can not be used to solve $\hat{Q}(\gamma, \beta)$. In this section, we give an iterative algorithm based on local quadratic approximation technology that used in Fan and Li (2001) and Zhao and Xue (2009). More specifically, for any given non-zero w_0, in a neighborhood of w_0, we have the following approximation

$$p_\lambda(|w|) \approx p_\lambda(|w_0|) + \frac{1}{2}\frac{p'_\lambda(|w_0|)}{|w_0|}(w^2 - w_0^2).$$

Hence, for the given initial value β_l^{ini} with $|\beta_l^{\text{ini}}| > 0, l = 1, \ldots, q$, and γ_k^{ini} with $\|\gamma_k^{\text{ini}}\|_H > 0, k = 1, \ldots, p$, we can obtain that

$$p_\lambda(|\beta_l|) \approx p_\lambda(|\beta_l^{\text{ini}}|) + \frac{1}{2}\frac{p'_\lambda(|\beta_l^{\text{ini}}|)}{|\beta_l^{\text{ini}}|}\left(|\beta_l|^2 - |\beta_l^{\text{ini}}|^2\right),$$

$$p_\lambda(\|\gamma_k\|_H) \approx p_\lambda(\|\gamma_k^{\text{ini}}\|_H) + \frac{1}{2}\frac{p'_\lambda(\|\gamma_k^{\text{ini}}\|_H)}{\|\gamma_k^{\text{ini}}\|_H}\left(\|\gamma_k\|_H^2 - \|\gamma_k^{\text{ini}}\|_H^2\right).$$

Let $\tilde{Z}_i = (\hat{Z}_i^T, W_i^T)^T$ and $\alpha = (\beta^T, \gamma^T)^T$ be $pL + q$-dimensional vectors. Furthermore, we let

$$\Sigma(\alpha^{\text{ini}}) = \text{diag}\left\{\frac{p'_\lambda(|\beta_1^{\text{ini}}|)}{|\beta_1^{\text{ini}}|}, \ldots, \frac{p'_\lambda(|\beta_q^{\text{ini}}|)}{|\beta_q^{\text{ini}}|}, \frac{p'_\lambda(\|\gamma_1^{\text{ini}}\|_H)}{\|\gamma_1^{\text{ini}}\|_H}H, \ldots, \frac{p'_\lambda(\|\gamma_p^{\text{ini}}\|_H)}{\|\gamma_p^{\text{ini}}\|_H}H\right\},$$

where $\alpha^{\text{ini}} = (\beta^{\text{ini}T}, \gamma^{\text{ini}T})^T$. Then, except for a constant term, $\hat{Q}(\gamma, \beta)$ that defined in (4) can be written as

$$\hat{Q}(\alpha) = \sum_{i=1}^{n}\{Y_i - \tilde{Z}_i^T\alpha\}^2 + \frac{n}{2}\alpha^T\Sigma(\alpha^{\text{ini}})\alpha.$$

It is clear that $\hat{Q}(\alpha)$ is a quadratic form, and it can be solved by

$$\left(\sum_{i=1}^{n}\tilde{Z}_i\tilde{Z}_i^T + \frac{n}{2}\Sigma(\alpha^{\text{ini}})\right)\alpha = \sum_{i=1}^{n}\tilde{Z}_i Y_i. \tag{5}$$

Hence, we can give an iterative algorithm as follows

S1. Initialize $\alpha^{(0)} = \alpha^{\text{ini}}$.
S2. Set $\alpha^{(0)} = \alpha^{(k)}$, solve $\alpha^{(k+1)}$ by Eq. (5).
S3. Iterate the step S2 until convergence, and denote the final estimator of α as $\hat{\alpha}$.

Then $\hat{\beta} = (I_{q\times q}, 0_{q\times pL})\hat{\alpha}$, and $\hat{\gamma} = (0_{pL\times q}, I_{pL\times pL})\hat{\alpha}$. In the initialization step, we obtain an initial estimator $\alpha^{\text{ini}} = (\beta^{\text{ini}T}, \gamma^{\text{ini}T})^T$ by using ordinary least squares method based on the following objective function

$$\hat{Q}^*(\gamma, \beta) = \sum_{i=1}^{n} \left\{ Y_i - W_i^T \gamma - \hat{Z}_i^T \beta \right\}^2.$$

Furthermore, to implement this method, the number of interior knots K, and the tuning parameters a and λ in the penalty function should be chosen. Fan and Li (2001) showed that the choice of $a = 3.7$ performs well in a variety of situations. Hence, we use this suggestion throughout this paper. In addition, we estimate λ and K by minimizing the following cross-validation score function

$$CV(K, \lambda) = \sum_{i=1}^{n} \left\{ Y_i - X_i^T \hat{\theta}_{[i]}(U_i) - \hat{Z}_i^T \hat{\beta}_{[i]} \right\}^2, \tag{6}$$

where $\hat{\theta}_{[i]}(\cdot)$ and $\hat{\beta}_{[i]}$ are estimators of $\theta(\cdot)$ and β respectively based on (4) after deleting the ith subject.

Although maybe some nonzero parameters will be incorrectly set to zeros in this algorithm, from the following simulation studies, we can see that the number of the nonzeros incorrectly set to zero is very small, and it decreases rapidly when the sample size n increases. This implies that the proposed iterative algorithm is workable.

4 Simulation studies

In this section, we conduct some Monte Carlo simulations to evaluate the finite sample performance of the proposed variable selection method. And as in Zhao and Xue (2009), the performance of estimator $\hat{\beta}$ will be assessed by using the generalized mean square error (GMSE), defined as

$$\text{GMSE} = (\hat{\beta} - \beta_0)^T E(ZZ^T)(\hat{\beta} - \beta_0).$$

The performance of estimator $\hat{\theta}(\cdot)$ will be assessed by using the square root of average square errors (RASE)

$$\text{RASE} = \left\{ \frac{1}{M} \sum_{s=1}^{M} \sum_{k=1}^{p} \left[\hat{\theta}_k(u_s) - \theta_{k0}(u_s) \right]^2 \right\}^{1/2},$$

where $u_s, s = 1, \ldots, M$ are the grid points at which the function $\hat{\theta}(u)$ are evaluated. In our simulation, $M = 200$ is used.

We simulate data from model (1), where $\beta = (\beta_1, \ldots, \beta_{10})^T$ with $\beta_1 = 3$, $\beta_2 = 2$, $\beta_3 = 1$ and $\beta_4 = 0.5$, and $\theta(u) = (\theta_1(u), \ldots, \theta_{10}(u))^T$ with $\theta_1(u) = 2.5 + 0.5\exp(2u - 1)$, $\theta_2(u) = 2 - \sin(\pi u)$ and $\theta_3(u) = 0.5 + 0.8u(1 - u)$. While the remaining coefficients, corresponding to the irrelevant variables, are given by zeros. To perform this simulation, we take the covariates $U \sim U(0, 1)$, $X_k \sim N(1, 1.5)$, and the instrumental variables $\xi_k \sim N(1, 1), k = 1, \ldots, 10$. The covariate $Z_k = \xi_k + \alpha\varepsilon$, where $\varepsilon \sim N(0, 0.5)$ and $\alpha = 0.2, 0.4$ and 0.6 to represent different levels of endogeneity of covariates. This setting up makes sure $E(Z_k\varepsilon) \neq 0$, which implies that the covariate Z_k is endogenous. In the following simulations, we use the quadratic B-splines, and the interior knots are taken equidistantly. Furthermore, the sample size is taken as $n = 100$, 200 and 300 respectively, and for each case, we take 1000 simulation runs.

To evaluate the performance of the proposed variable selection method, two methods are compared: the instrumental variable based partial gSCAD variable selection method (IV-gSCAD) based on Theorem 1, and the naive partial gSCAD variable selection method (Naive-gSCAD). The latter is neglecting the endogeneity of covariate Z_i, and using the partial gSCAD penalty method based on (3) directly. Based on the 1000 simulation runs, the average number of zero coefficients for parametric components is reported in Table 1, and the average number of zero coefficients for nonparametric components is reported in Table 2. In Tables 1 and 2, the column labeled "C" presents the average number of coefficients of the true zeros correctly set to zero, and the column labeled "I" presents the average number of the true nonzeros incorrectly set to zero. Tables 1 and 2 also present the average false selection rate (FSR), which is defined as FSR = IN/TN, where "IN" is the average number of the true zeros incorrectly set to nonzero, and "TN" is the average total number set to nonzero. In fact, FSR represents the proportion of falsely selected unimportant variables among the

Table 1 Variable selection results for parametric components based on different variable selection methods

n	α	IV-gSCAD				Naive-gSCAD			
		C	I	FSR	GMSE	C	I	FSR	GMSE
100	0.2	5.388	0.031	0.134	0.035	5.347	0.028	0.141	0.038
	0.4	5.374	0.034	0.136	0.038	4.423	0.036	0.285	0.154
	0.6	5.365	0.036	0.139	0.041	3.804	0.039	0.357	0.261
200	0.2	5.725	0.014	0.064	0.014	5.648	0.015	0.081	0.018
	0.4	5.726	0.016	0.064	0.016	4.583	0.018	0.262	0.143
	0.6	5.722	0.019	0.065	0.019	4.024	0.019	0.332	0.217
300	0.2	6	0.003	0	0.007	5.734	0.005	0.062	0.014
	0.4	5.980	0.004	0.005	0.009	5.176	0.008	0.171	0.082
	0.6	5.974	0.007	0.006	0.012	4.348	0.013	0.293	0.161

Table 2 Variable selection results for nonparametric components based on different variable selection methods

n	α	IV-gSCAD				Naive-gSCAD			
		C	I	FSR	RASE	C	I	FSR	RASE
100	0.2	6.136	0.042	0.226	0.084	6.127	0.049	0.228	0.095
	0.4	6.071	0.044	0.239	0.088	5.343	0.067	0.361	0.158
	0.6	5.962	0.049	0.260	0.094	4.836	0.088	0.426	0.272
200	0.2	6.768	0.022	0.072	0.036	6.534	0.034	0.136	0.056
	0.4	6.752	0.025	0.077	0.041	5.439	0.056	0.347	0.134
	0.6	6.744	0.028	0.079	0.043	5.127	0.079	0.391	0.247
300	0.2	6.993	0.015	0.002	0.019	6.763	0.025	0.074	0.041
	0.4	6.990	0.019	0.003	0.025	5.581	0.052	0.325	0.125
	0.6	6.985	0.021	0.005	0.028	5.287	0.072	0.369	0.232

total variables selected in the variable selection procedure. From Tables 1 and 2, we can make the following observations:

(i) The performances of IV-gSCAD method for parametric components and nonparametric components are both better than those of Naive-gSCAD method, and this is especially true when the level of endogeneity of covariates is large. Because the Naive-gSCAD variable selection method cannot eliminate some unimportant variables in the parametric and nonparametric components, and gives significantly larger model errors. This implies that the Naive-gSCAD variable selection procedure is biased.

(ii) For the given level of endogeneity of covariates, the GMSE, RASE and FSR, obtained by the IV-gSCAD variable selection method, all decrease as the sample size n increases. This implies that the proposed IV-gSCAD variable selection procedure is consistent.

(iii) For given n, the IV-gSCAD variable selection method performs similar in terms of model error and model complexity for all levels of endogeneity of covariates. This indicates that the proposed instrumental variable based variable selection can attenuate the effect of the endogeneity of covariates. In general, the proposed variable selection method works well in terms of model error and the model complexity.

Acknowledgments This paper is supported by the National Natural Science Foundation of China (11301569), the Higher-education Reform Project of Guangxi (2014JGA209), and the Project of Outstanding Young Teachers Training in Higher Education Institutions of Guangxi.

Appendix: Proof of theorems

For convenience and simplicity, let c denote a positive constant which may be different value at each appearance throughout this paper. Before we prove our main theorems, we list some regularity conditions which are used in this paper.

C1. $\theta(u)$ is rth continuously differentiable on $(0, 1)$, where $r > 1/2$.

C2. Let $c_1, \ldots, c_K$ be the interior knots of $[0, 1]$. Furthermore, we let $c_0 = 0$, $c_{K+1} = 1$, $h_i = c_i - c_{i-1}$ and $h = \max\{h_i\}$. Then, there exists a constant C_0 such that

$$\frac{h}{\min\{h_i\}} \leq C_0, \quad \max\{|h_{i+1} - h_i|\} = o\left(K^{-1}\right).$$

C3. The density function of U, says $f(u)$, is bounded away from zero and infinity on $[0, 1]$, and is continuously differentiable on $(0, 1)$.

C4. Let $G_1(u) = E\{ZZ^T|U = u\}$, $G_2(u) = E\{XX^T|U = u\}$ and $\sigma^2(u) = E\{\varepsilon^2|U = u\}$. Then, $G_1(u)$, $G_2(u)$ and $\sigma^2(u)$ are continuous with respect to u. Furthermore, for given u, $G_1(u)$ and $G_2(u)$ are positive definite matrix, and the eigenvalues of $G_1(u)$ and $G_2(u)$ are bounded.

C5. The penalty function $p_\lambda(\cdot)$ satisfies that

(i) $\lim_{n\to\infty} \lambda = 0$, and $\lim_{n\to\infty} \sqrt{n}\lambda = \infty$.

(ii) For any given non-zero w, $\lim_{n\to\infty} \sqrt{n} p'_\lambda(|w|) = 0$, and $\lim_{n\to\infty} p''_\lambda(|w|) = 0$.

(iii) $\lim_{n\to\infty} \sup_{|w|\leq cn^{-1/2}} p''_\lambda(|w|) = 0$, and $\lim_{n\to\infty} \lambda^{-1} \inf_{|w|\leq cn^{-1/2}} p'_\lambda(|w|) > 0$, where c is a positive constant.

These conditions are commonly adopted in the nonparametric literature and variable selection methodology. Conditions C1 is the continuity condition of nonparametric components which is common in the nonparametric literature. Condition C2 indicates that $c_0, \ldots, c_{K+1}$ is a C_0-quasi-uniform sequence of partitions of $[0, 1]$ (see Schumaker 1981, p. 216), and this assumption is used in Zhao and Li (2013), Zhao and Xue (2009), Wang et al. (2013). Conditions C3 and C4 are some regularity conditions for covariates, which are similar to those used in Zhao and Xue (2013), Cai and Xiong (2012), Wang et al. (2008). Condition C5 contains some assumptions for the penalty function. These conditions on the penalty function are similar to those used in Fan and Li (2001), Wang et al. (2008), Zhao and Xue (2009), and it is easy to show that the SCAD, Lasso penalty functions satisfy these conditions.

Proof of Theorem 1 Let $\delta = n^{-r/(2r+1)} + a_n$, $\beta = \beta_0 + \delta M_1$, $\gamma = \gamma_0 + \delta M_2$ and $M = (M_1^T, M_2^T)^T$. For part (i), we show that, for any given $\epsilon > 0$, there exists a large constant c such that

$$P\left\{\inf_{\|M\|=c} \hat{Q}(\gamma, \beta) > \hat{Q}(\gamma_0, \beta_0)\right\} \geq 1 - \epsilon. \tag{7}$$

Let $R_{k0}(u) = \theta_{k0}(u) - B(u)^T \gamma_{k0}$, then note that $W_i = I_p \otimes B(U_i) \cdot X_i$, we have

$$X_i^T \theta_0(U_i) - W_i^T \gamma_0 = X_i^T R_0(U_i), \tag{8}$$

where $R_0(U_i) = (R_{10}(U_i), \ldots, R_{p0}(U_i))^T$. Hence, we have

$$\begin{aligned}
&\sum_{i=1}^{n}\{Y_i - W_i^T\gamma_0 - \hat{Z}_i^T\beta_0\}^2 \\
&= \sum_{i=1}^{n}\{X_i^T\theta_0(U_i) + Z_i^T\beta_0 + \varepsilon_i - W_i^T\gamma_0 - \hat{Z}_i^T\beta_0\}^2 \\
&= \sum_{i=1}^{n}\{X_i^T R_0(U_i) + (Z_i - \hat{Z}_i)^T\beta_0 + \varepsilon_i\}^2. \qquad (9)
\end{aligned}$$

Then, invoking $\beta = \beta_0 + \delta M_1$, $\gamma = \gamma_0 + \delta M_2$ and (9), we can obtain that

$$\begin{aligned}
&\sum_{i=1}^{n}\{Y_i - W_i^T\gamma - \hat{Z}_i^T\beta\}^2 \\
&= \sum_{i=1}^{n}\{Y_i - W_i^T(\gamma_0 + \delta M_2) - \hat{Z}_i^T(\beta_0 + \delta M_1)\}^2 \\
&= \sum_{i=1}^{n}\{Y_i - W_i^T\gamma_0 - \hat{Z}_i^T\beta_0 - \delta W_i^T M_2 - \delta\hat{Z}_i^T M_1\}^2 \\
&= \sum_{i=1}^{n}\{X_i^T R_0(U_i) + (Z_i - \hat{Z}_i)^T\beta_0 + \varepsilon_i - \delta(\hat{Z}_i^T M_1 + W_i^T M_2)\}^2. \qquad (10)
\end{aligned}$$

By (9) and (10), and based on the formula $a^2 - b^2 = (a+b)(a-b)$, we have that

$$\begin{aligned}
&\sum_{i=1}^{n}\{Y_i - W_i^T\gamma - \hat{Z}_i^T\beta\}^2 - \sum_{i=1}^{n}\{Y_i - W_i^T\gamma_0 - \hat{Z}_i^T\beta_0\}^2 \\
&= \sum_{i=1}^{n}\{X_i^T R_0(U_i) + (Z_i - \hat{Z}_i)^T\beta_0 + \varepsilon_i - \delta(\hat{Z}_i^T M_1 + W_i^T M_2)\}^2 \\
&\quad - \sum_{i=1}^{n}\{X_i^T R_0(U_i) + (Z_i - \hat{Z}_i)^T\beta_0 + \varepsilon_i\}^2 \\
&= \sum_{i=1}^{n}[-\delta(\hat{Z}_i^T M_1 + W_i^T M_2)]\{2[X_i^T R_0(U_i) + (Z_i - \hat{Z}_i)^T\beta_0 + \varepsilon_i] - \delta(\hat{Z}_i^T M_1 + W_i^T M_2)\} \\
&= -2\delta\sum_{i=1}^{n}[\hat{Z}_i^T M_1 + W_i^T M_2][X_i^T R_0(U_i) + (Z_i - \hat{Z}_i)^T\beta_0 + \varepsilon_i] \\
&\quad + \delta^2\sum_{i=1}^{n}[\hat{Z}_i^T M_1 + W_i^T M_2]^2. \qquad (11)
\end{aligned}$$

Let $\Delta(\gamma, \beta) = K^{-1}\{\hat{Q}(\gamma, \beta) - \hat{Q}(\gamma_0, \beta_0)\}$, then from (11), we have that

$$\Delta(\gamma, \beta) = \frac{1}{K}\left\{\sum_{i=1}^{n}\{Y_i - W_i^T\gamma - \hat{Z}_i^T\beta\}^2 - \sum_{i=1}^{n}\{Y_i - W_i^T\gamma_0 - \hat{Z}_i^T\beta_0\}^2\right\}$$

$$
\begin{aligned}
&+\frac{n}{K}\sum_{l=1}^{q}\left[p_\lambda(|\beta_l|)-p_\lambda(|\beta_{l_0}|)\right]+\frac{n}{K}\sum_{k=1}^{p}\left[p_\lambda(\|\gamma_k\|_H)-p_\lambda(\|\gamma_{k_0}\|_H)\right]\\
&=-\frac{2\delta}{K}\sum_{i=1}^{n}\left[\varepsilon_i+X_i^TR(U_i)+(Z_i-\hat{Z}_i)^T\beta_0\right]\left[\hat{Z}_i^TM_1+W_i^TM_2\right]\\
&\quad+\frac{\delta^2}{K}\sum_{i=1}^{n}[\hat{Z}_i^TM_1+W_i^TM_2]^2+\frac{n}{K}\sum_{l=1}^{q}[p_\lambda(|\beta_l|)-p_\lambda(|\beta_{l_0}|)]\\
&\quad+\frac{n}{K}\sum_{k=1}^{p}[p_\lambda(\|\gamma_k\|_H)-p_\lambda(\|\gamma_{k_0}\|_H)]\\
&\equiv I_1+I_2+I_3+I_4.
\end{aligned}
$$

From conditions C1, C2 and Corollary 6.21 in Schumaker (1981), we get that $\|R(u)\|=O(K^{-r})$. Then, invoking condition C4, a simple calculation yields

$$\sum_{i=1}^{n}X_i^TR(U_i)[\hat{Z}_i^TM_1+W_i^TM_2]=O_p(nK^{-r}\|M\|). \tag{12}$$

Invoking $E\{\varepsilon_i|\xi_i,X_i\}=0$ and $\hat{Z}_i=\Gamma\xi_i+O_p(n^{-1/2})$, we can prove that

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\varepsilon_i[\hat{Z}_i^TM_1+W_i^TM_2]=O_p(\|M\|). \tag{13}$$

In addition, note that $Z_i-\hat{Z}_i=(\Gamma-\hat{\Gamma})\xi_i+e_i$, then invoking $\hat{\Gamma}=\Gamma+O_p(n^{-1/2})$ and $E\{e_i|\xi_i,X_i\}=0$, we can prove that

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(Z_i-\hat{Z}_i)^T\beta_0[\hat{Z}_i^TM_1+W_i^TM_2]=O_p(\|M\|). \tag{14}$$

Hence, from (12) to (14), it is easy to show that

$$I_1=O_p(nK^{-1-r}\delta)\|M\|+O_p(\sqrt{n}K^{-1}\delta)\|M\|=O_p(1+n^{\frac{r}{2r+1}}a_n)\|M\|.$$

Similarly, we can prove that

$$I_2=O_p(nK^{-1}\delta^2)\|M\|^2=O_p(1+2n^{\frac{r}{2r+1}}a_n)\|M\|^2.$$

By the condition C5(ii), we have that $\lim_{n\to\infty}\sqrt{n}p'_\lambda(|w|)=0$, for any given nonzero w. Then invoking the definition of a_n, we can obtain $\sqrt{n}a_n\to 0$ when n is large enough. Hence, we obtain that

$$n^{\frac{r}{2r+1}}a_n<\sqrt{n}a_n\to 0.$$

Hence we have $I_2/I_1 = O_p(1)\|M\|$. Then, by choosing a sufficiently large c, I_2 can dominate I_1 uniformly in $\|M\| = c$. Furthermore, invoking $p_\lambda(0) = 0$, and by the standard argument of the Taylor expansion, we get that

$$\begin{aligned} I_3 &\le K^{-1} n \sum_{l=1}^{s} [p_\lambda(|\beta_l|) - p_\lambda(|\beta_{l_0}|)] \\ &\le \sum_{l=1}^{s} [K^{-1} n\delta p'_\lambda(|\beta_{l0}|)\mathrm{sgn}(\beta_{l0})|M_{1l}| + K^{-1} n\delta^2 p''_\lambda(|\beta_{l0}|)|M_{1l}|^2\{1 + o(1)\}] \\ &\le \sqrt{s} K^{-1} n\delta a_n \|M\| + K^{-1} n\delta^2 b_n \|M\|^2 \\ &= O_p\left(n^{\frac{r}{2r+1}} a_n\right)\|M\| + O_p(1 + 2n^{\frac{r}{2r+1}} a_n) b_n \|M\|^2. \end{aligned}$$

Note that $n^{\frac{r}{2r+1}} a_n \to 0$ and $b_n \to 0$, we obtain that $I_3 = o_p(1)\|M\|^2$. Hence, we have that I_3 is dominated by I_2 uniformly in $\|M\| = c$. With the same argument, we can prove that I_4 is also dominated by I_2 uniformly in $\|M\| = c$. In addition, note that I_2 is positive, then by choosing a sufficiently large c, (7) holds.

By the continuity of $\hat{Q}(\cdot, \cdot)$, the inequality (7) implies that $\hat{Q}(\cdot, \cdot)$ should have a local minimum on $\{\|M\| \le c\}$ with probability greater than $1 - \epsilon$. Hence, there exists a local minimizer $\hat{\beta}$ such that $\|\hat{\beta} - \beta_0\| = O_p(\delta)$, which completes the proof of part (i).

Next, we prove part (ii). Note that

$$\begin{aligned} \|\hat{\theta}_k(u) - \theta_{k0}(u)\|^2 &= \int_0^1 \{\hat{\theta}_k(u) - \theta_{k0}(u)\}^2 du \\ &= \int_0^1 \{B^T(u)\hat{\gamma}_k - B^T(u)\gamma_k + R_k(u)\}^2 du \\ &\le 2\int_0^1 \{B^T(u)\hat{\gamma}_k - B^T(u)\gamma_k\}^2 du + 2\int_0^1 R_k(u)^2 du \\ &= 2\int_0^1 (\hat{\gamma}_k - \gamma_k)^T B(u) B^T(u)(\hat{\gamma}_k - \gamma_k) du + 2\int_0^1 R_k(u)^2 du. \end{aligned}$$

With the same arguments as the proof of part (i), we can get that $\|\hat{\gamma} - \gamma\| = O_p(n^{-r/(2r+1)} + a_n)$. Then, a simple calculation yields

$$\int_0^1 (\hat{\gamma}_k - \gamma_k)^T B(u) B^T(u)(\hat{\gamma}_k - \gamma_k) du = O_p\left\{(n^{\frac{-r}{2r+1}} + a_n)^2\right\}. \tag{15}$$

In addition, it is easy to show that

$$\int_0^1 R_k(u)^2 du = O_p(n^{\frac{-2r}{2r+1}}). \tag{16}$$

Invoking (15) and (16), we complete the proof of part (ii). □

Proof of Theorem 2 We first prove part (i). Invoking the condition $\lambda \to 0$, it is easy to show that $a_n = 0$ for large n. Then by Theorem 1, it is sufficient to show that, for any given $\beta_l, l = 1, \ldots, s$, which satisfy $\|\beta_l - \beta_{l0}\| = O_p(n^{-r/(2r+1)})$, and a small ϵ which satisfies $\epsilon = cn^{-r/(2r+1)}$, with probability tending to 1, we have

$$\frac{\partial \hat{Q}(\gamma, \beta)}{\partial \beta_l} > 0, \quad \text{for } 0 < \beta_l < \epsilon, \ \ l = s+1, \ldots, q, \tag{17}$$

and

$$\frac{\partial \hat{Q}(\gamma, \beta)}{\partial \beta_l} < 0, \quad \text{for } -\epsilon < \beta_l < 0, \ \ l = s+1, \ldots, q. \tag{18}$$

Thus, (17) and (18) imply that the minimizer attains at $\beta_l = 0, l = s+1, \ldots, q$.

By a similar the proof of Theorem 1, we have that

$$\begin{aligned}
\frac{\partial \hat{Q}(\gamma, \beta)}{\partial \beta_l} &= \sum_{i=1}^{n} \hat{Z}_{il}(Y_i - \hat{Z}_i^T \beta - W_i^T \gamma) + np'_\lambda(|\beta_l|)\mathrm{sgn}(\beta_l) \\
&= -2\sum_{i=1}^{n} \hat{Z}_{il}[\varepsilon_i + X_i^T R(U_i)] - 2\sum_{i=1}^{n} \hat{Z}_{il} Z_i^T (\beta_0 - \beta) \\
&\quad -2\sum_{i=1}^{n} \hat{Z}_{il}(Z_i - \hat{Z}_i)^T \beta - 2\sum_{i=1}^{n} \hat{Z}_{il} W_i^T (\gamma_0 - \gamma) + np'_\lambda(|\beta_l|)\mathrm{sgn}(\beta_l) \\
&= n\lambda\{\lambda^{-1} p'_\lambda(|\beta_l|)\mathrm{sgn}(\beta_l) + O_p(\lambda^{-1} n^{-\frac{r}{2r+1}})\}.
\end{aligned}$$

Since $\lim_{n\to\infty} \liminf_{\beta_l \to 0} \lambda^{-1} p'_\lambda(|\beta_l|) > 0$ and $\lambda n^{\frac{r}{2r+1}} \to \infty$, the sign of the derivative is completely determined by the sign of β_l, then (17) and (18) hold. This completes the proof of part (i).

Applying the similar techniques as in the analysis of part (i) in this theorem, we have, with probability tending to 1, that $\hat{\gamma}_k = 0, k = d+1, \ldots, p$. Then, the result of this theorem is immediately achieved form $\hat{\theta}_k(u) = B^T(u)\hat{\gamma}_k$. □

References

Cai Z, Xiong H (2012) Partially varying coefficient instrumental variables models. Stat Neerl 66:85–110

Card D (2001) Estimating the return to schooling: progress on some persistent econometric problems. Econometrica 69:1127–1160

Fan JQ, Li R (2001) Variable selection via nonconcave penalized likelihood and its oracle properties. J Am Stat. Assoc 96:1348–1360

Frank IE, Friedman JH (1993) A statistical view of some chemometrics regression tool (with discussion). Technometrics 35:109–148

Greenland S (2000) An introduction to instrumental variables for epidemiologists. Int J Epidemiol 29:722–729

Hernan MA, Robins JM (2006) Instruments for causal inference—an epidemiologists dream? Epidemiology 17:360–372

Lian H (2012) Variable selection for high-dimensional generalized varying-coefficient models. Stat Sinica 22:1563–1588
Lin ZY, Yuan YZ (2012) Variable selection for generalized varying coefficient partially linear models with diverging number of parameters. Acta Math Appl Sinica Eng Ser 28(2):237–246
Newhouse JP, McClellan M (1998) Econometrics in outcomes research: the use of instrumental variables. Annu Rev Public Health 19:17–24
Schultz TP (1997) Human capital, schooling and health. IUSSP, XXIII, General Population Conference. Yale University
Schumaker LL (1981) Spline functions. Wiley, New York
Tibshirani R (1996) Regression shrinkage and selection via the Lasso. J R Stat Soc Ser B 58:267–288
Wang L, Li H, Huang JZ (2008) Variable selection in nonparametric varying-coefficient models for analysis of repeated measurements. J Am Stat Assoc 103:1556–1569
Wang LC, Lai P, Lian H (2013) Polynomial spline estimation for generalized varying coefficient partially linear models with a diverging number of components. Metrika 76:1083–1103
Yao F (2012) Efficient semiparametric instrumental variable estimation under conditional heteroskedasticity. J Quant Econ 10:32–55
Zhao PX, Li GR (2013) Modified SEE variable selection for varying coefficient instrumental variable models. Stat Methodol 12:60–70
Zhao PX, Xue LG (2009) Variable selection for semiparametric varying coefficient partially linear models. Stat Probab Lett 79:2148–2157
Zhao PX, Xue LG (2013) Empirical likelihood inferences for semiparametric instrumental variable models. J Appl Math Comput 43:75–90