Hacettepe Journal of Mathematics and Statistics
Volume 48 (1) (2019), 213 – 229

Variable selection for high dimensional partially linear varying coefficient errors-in-variables models

Zhaoliang Wang*[†] and Liugen Xue[‡]

Abstract

In this paper, we consider variable selection procedure for the high dimensional partially linear varying coefficient models where the parametric part covariates are measured with additive errors. The penalized bias-corrected profile least squares estimators are conducted, and their asymptotic properties are also studied under some regularity conditions. The rate of convergence and the asymptotic normality of the resulting estimates are established. We further demonstrate that, with proper choices of the penalty functions and the regularization parameter, the resulting estimates perform asymptotically as well as an oracle property. Choice of smoothing parameters is also discussed. Finite sample performance of the proposed variable selection procedures is assessed by Monte Carlo simulation studies.

Keywords: High dimensionality, Measurement error, Semiparametric models, Local linear regression, Variable selection.

2000 AMS Classification: 62G05, 62G20

Received : 09.06.2017 *Accepted* : 28.06.2018 *Doi* : 10.15672/HJMS.2018.603

1. Introduction

With the development of applied sciences, semiparametric regression models have been well researched and popularly used for their flexibility and interpretability. [16] present diverse semiparametric regression models along with their inference procedures and applications. Of particular interests to us in this paper is the partially linear varying coefficient (PLVC) model. Let $\{(Y_i, X_i, Z_i, T_i), i = 1, \ldots, n\}$ be an iid copies of (Y, X, Z, T), where

*School of Mathematics and Information Science, Henan Polytechnic University and College of Applied Sciences, Beijing University of Technology, Email: `wangzhaolinag@hpu.edu.cn`

[†]Corresponding Author.

[‡]College of Applied Sciences, Beijing University of Technology, Email: `lgxue@bjut.edu.cn`

Y is a scalar response variable and $(X, Z, T) \in \mathbb{R}^p \times \mathbb{R}^q \times \mathbb{R}$ is its associated regressors. The PLVC models take the form

$$(1.1) \qquad Y_i = X_i^\top \beta + Z_i^\top \alpha(T_i) + \varepsilon_i,$$

where $\beta = (\beta_1, \ldots, \beta_p)^\top$ is a p-dimensional vector of unknown parameters, $\alpha(\cdot) = (\alpha_1(\cdot), \ldots, \alpha_q(\cdot))^\top$ is an q-dimensional vector of unknown coefficient functions, and ε_i's are iid model error with $\mathrm{E}(\varepsilon_i | X_i, Z_i, T_i) = 0$. In this model, the dependence of $\alpha(\cdot)$ on T implies a special kind of interaction between the covariate Z and T. Due to the curse of dimensionality, we assume, for simplicity, that T is univariate. This model presents a novel and general structure, which indeed covers many well-studied, important semiparametric regression models, e.g. linear model, partially linear model and varying coefficient model.

Model (1.1) has been studied by many authors recently. Examples include but are not limited to [1, 26, 13, 12, 10, 3, 23]. An essential assumption in their papers is that all data can be observed directly. However, measurement error data are often encountered in many fields, including engineering, economics, biomedical sciences and epidemiology. Simply ignoring measurement errors, known as the naive method, will result in biased estimators. There is a long standing literature on statistical modeling subject to measurement errors. Comprehensive reviews can be found in [2, 7]. PLVC models have been used to study measurements with errors, see, for instance, [21, 8, 20, 19, 6].

Concerns about model bias often prompt us to build models that contain many variables, especially when the sample size becomes large. A reasonable way to capture such a tendency is to consider the situation where the dimension of the parameter increases along with the sample size. On the other hand, to enhance predictability and to select significant variables is practically interesting, but is always a tricky task for data analysis. When the number of covariates is large, traditional variable selection methods such as stepwise regression and best subset selection is computationally infeasible and statistical properties of the estimators are difficult to analyze, as argued in [14], this is part of the reason why penalization based method (e.g., Lasso [17], Elastic net [28], Adaptive Lasso [27], SCAD [4], MCP [22], among others) has gained popularity in recent years. There has been much work on variable selection for semiparametric regression models. In particular, examples for fixed dimensional PLVC models include [25, 24, 11, 18] and references therein.

In these studies, however, high dimensional vector X, variable selection in X and measurement error problem were not considered at the same time. The goal of this paper intends to develop an unified estimation and variable selection method for high dimensional PLVC errors-in-variables models. To be precise, we allow $p \to \infty$ as the sample size $n \to \infty$ and denote it by p_n whenever necessary, but q is a fixed and finite integer in (1.1). In addition, the covariate X is measured with additive errors, while Z and T are errors free. More specific, we cannot observe X_i but we can observe W_i with

$$(1.2) \qquad W_i = X_i + U_i,$$

and U_i's are iid measurement error, which is independent of $(X_i, Z_i, T_i, \varepsilon_i)$, and has mean zero and the known covariance $\mathrm{Cov}(U_i) = \Sigma_U$ (for simplicity). If Σ_U is unknown, its estimation usually requires multiple observations of W or instrumental variables, see [15] for details. We term (1.1) and (1.2) with PLVCE models. To our best knowledge, variable selection for PLVCE models with high dimension has not been systematically studied yet.

We propose penalized bias-corrected profile least squares estimator and systematically study the asymptotic properties of the estimators. It is worth pointing out that theoretic results in this paper provide explicit results on the asymptotic properties under the setting in which both the dimension of the true non-zero components of β and the total length of β tend to infinity as n goes to infinity. This resonates with the perspective that a more complex statistical model can be fit when more data are collected. The issue of a diverging number of parameters has also been considered in [5] in the context of penalized likelihood. This advances the results in current literature, where estimation and inference are studied only for fixed finite dimensional parameters for measurement error models. We demonstrate how the convergence rate of the resulting estimator depends on the regularization parameter. Furthermore, with a proper choice of the regularization parameters and the penalty function, we show that this variable selection procedure is consistent, and the regularized estimators of the regression coefficients have oracle property. This indicates that the penalized estimators work as well as if the subset of true zero coefficients were already known. In addition, we address issues of practical implementation of the proposed methodology. Monte-Carlo simulation studies are conducted to assess finite sample performance.

The rest of this paper is organized as follows. A variable selection procedure for PLVCE models is proposed in Section 2, assumptions and the asymptotic properties of the proposed estimators are given in this section. We give the computational algorithms and discuss the selections of tuning parameters in Section 3. In Section 4, some simulations are conducted to illustrate the performance of our methodology. Given in Section 5 are conclusions. All technical proofs are relegated to Section 6.

Notation: The gradient and hessian matrix of a function $f(x)$ are denoted by $\nabla f(x)$ and $\nabla^2 f(x)$ respectively. We write $\|f\|_2$ and $\|f\|_\infty$ for the L_2 and sup norm of a function f, respectively. The L_q norm of a p-vector v is defined as $\|v\|_q = (\sum_{j=1}^p |v_j|^q)^{1/q}$ for $q \geq 1$ with $\|v\|_\infty = \max_{1\leq j\leq p} |v_j|$, and $\|v\|_0 = |\mathrm{supp}(v)|$ where $\mathrm{supp}(v) = \{j : v_j \neq 0\}$ and $|S|$ is the cardinality of a set S. Let $M_{i\cdot}$, $M_{\cdot j}$ and M_{ij} be the ith row, jth column and (i,j) entry of the matrix M, respectively. Let $\|M\|_q = \sup_{\|v\|_q=1} \|Mv\|_q$ be the matrix L_q operator norm. We use $\|\cdot\|$ as a shorthand for $\|\cdot\|_2$. We use c and C to denote generic positive constants that may vary from place to place. Moreover, the operator $\xrightarrow{P}$ denotes convergence in probability, and $\xrightarrow{D}$ denotes convergence in distribution.

2. Methods and results

2.1. Penalized bias-corrected profile least squares estimator. As in [3], if X_i is observable we can apply the profile least squares estimation to estimate the parametric component and apply the local polynomial estimation to estimate the nonparametric component. Profile least squares is a useful approach and will be showed to be semiparametrically efficient for model (1.1). When $\varepsilon_i \sim N(0, \sigma^2)$, the approach becomes profile likelihood estimation. For the paper to be self-contained, we summarize the main ingredients as follows. If β is known, (1.1) can be written as

$$Y_i - X_i^\top \beta = Z_i^\top \alpha(T_i) + \varepsilon_i, \tag{2.1}$$

which can be treated as a varying coefficient model. Thus, we may apply a local linear regression technique to estimate the varying coefficient functions $\{\alpha_j(\cdot),\ j = 1, \ldots, q\}$. For T in a small neighbourhood of t, approximate each $\alpha_j(t)$ by $\alpha_j(T) \approx \alpha_j(t) + \alpha_j'(t)(T - t)$, $j = 1, \ldots, q$. This leads to the following weighted local least-squares problem: find

$\alpha_j(t), \alpha_j'(t)$ to minimize

$$\sum_{i=1}^{n}\left[Y_i - X_i^\top\beta - \sum_{j=1}^{q} Z_{ij}\left\{\alpha_j(t) + \alpha_j'(t)(T_i - t)\right\}\right]^2 K_h(T_i - t), \tag{2.2}$$

where $K_h(\cdot) = K(\cdot/h)/h$, $K(\cdot)$ is a kernel function and h is a bandwidth.

For the sake of descriptive convenience, we denote $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^\top$, write $\boldsymbol{X}$, $\boldsymbol{Z}$, $\boldsymbol{\varepsilon}$ in a similar fashion. Let $\omega_t = \mathrm{diag}\{K_h(T_1 - t), \ldots, K_h(T_n - t)\}$ and

$$D_t = \begin{pmatrix} Z_1 & \cdots & Z_n \\ \frac{T_1 - t}{h} Z_1 & \cdots & \frac{T_n - t}{h} Z_n \end{pmatrix}^\top.$$

It is easy to show that the minimizers of (2.2) are given by

$$(\tilde{\alpha}(t)^\top, h\tilde{\alpha}'(t)^\top)^\top = \{D_t^\top \omega_t D_t\}^{-1} D_t^\top \omega_t(\boldsymbol{Y} - \boldsymbol{X}\beta).$$

This solutions depend on β implicitly. Then we can estimate $\alpha(t)$, when β is given, by

$$\tilde{\alpha}(t;\beta) = (I_{q\times q}, 0_{q\times q})\{D_t^\top \omega_t D_t\}^{-1} D_t^\top \omega_t(\boldsymbol{Y} - \boldsymbol{X}\beta), \tag{2.3}$$

where $I_{q\times q}$ denote the q by q identity matrix, and $0_{q\times q}$ denote a q by q matrix of zeros. Substituting $\tilde{\alpha}(t;\beta)$ into model (2.1), we can obtain the profile least square estimator of β by the following regression problem

$$\tilde{\beta} = \arg\min_{\beta} \frac{1}{2}\sum_{i=1}^{n}(Y_i - X_i^\top\beta - Z_i^\top\tilde{\alpha}(T_i;\beta))^2 \tag{2.4}$$

Moreover, plug $\tilde{\beta}$ into (2.3), the estimators of $\alpha(t)$ can be obtained, see [3] for details.

However, in our case, X_i cannot be exactly observed. If one ignores the measurement error and replaces X_i by W_i in (2.4), one can show that the resulting estimator is inconsistent. By the correction for attenuation technique as in [21], the bias-corrected profile least squares estimator of β can be defined by minimizing

$$\widehat{L}_n(\beta) = \frac{1}{2}\sum_{i=1}^{n}(Y_i - W_i^\top\beta - Z_i^\top\hat{\alpha}(T_i,\beta))^2 - \frac{n}{2}\beta^\top\Sigma_U\beta, \tag{2.5}$$

where $\hat{\alpha}(T_i, \beta)$ is obtained by replace $\boldsymbol{X}$ with $\boldsymbol{W}$ in the right hand side of (2.3). The second term is included to correct the bias in the squared loss function due to measurement error.

In high dimensional data analysis, to perform variable selection and estimation simultaneously, based on (2.5) we propose the penalized bias-corrected profile least squares function defined as

$$\widehat{Q}_n(\beta) = \widehat{L}_n(\beta) + n\sum_{j=1}^{p_n} p_\lambda(|\beta_j|), \tag{2.6}$$

where $p_\lambda(\cdot)$ is a prespecified penalty function with a tuning parameter λ, which may be chosen by a data-driven method. It is worth noting that the penalty functions and the tuning parameters are not necessarily the same for all coefficients. For instance, we want to keep important variables in the final model, and therefore we should not penalize their coefficients. For ease of presentation, we assume that the penalty functions and the regularization parameters are the same for all coefficients in this paper.

The choice of the penalty functions has been studied in [4] in depth. A property of (2.6) is that with a proper choice of penalty functions, such as the SCAD and Lasso penalty, the resulting estimate contains some exact zero coefficients. This is equivalent to excluding the corresponding variables from the final selected model, thus variable selection is achieved at the same time as parameter estimation. Solving for $\hat{\beta}$ from (2.6) gives the estimate of β. Moreover, the fact that $\mathrm{E}(Y_i - X_i^\top \beta | T_i) = \mathrm{E}(Y_i - W_i^\top \beta | T_i)$ suggests us to estimate $\alpha(\cdot)$ by

$$\hat{\alpha}(t) = (I_{q\times q}, 0_{q\times q})\{D_t^\top \omega_t D_t\}^{-1} D_t^\top \omega_t (\boldsymbol{Y} - \boldsymbol{W}\hat{\beta}). \tag{2.7}$$

2.2. Asymptotic properties. In this subsection we consider the large sampling properties of the proposed estimator. For convenience of notation, we assume the true value $\beta^* = (\beta_I^{*\top}, \beta_{II}^{*\top})^\top$, where β_I^* consists of all nonzero components of β^* and $\beta_{II}^* = 0$. Let s_n denote the dimension of β_I^*. Furthermore, denote

$$B = (p'_{\lambda_n}(|\beta_1^*|)\mathrm{sign}(\beta_1^*), \ldots, p'_{\lambda_n}(|\beta_{s_n}^*|)\mathrm{sign}(\beta_{s_n}^*))^\top \quad \text{and}$$
$$\Sigma_{\lambda_n} = \mathrm{diag}\{p''_{\lambda_n}(|\beta_1^*|), \ldots, p''_{\lambda_n}(|\beta_{s_n}^*|)\},$$

where we write λ as λ_n to emphasize its dependence on the sample size n. To give the asymptotic results, here are regularity conditions required.

(C1) The random variable T has a bounded support $\mathcal{T}$. Its density function $f_T(t)$ is Lipschitz continuous and bounded away from 0 on $\mathcal{T}$.

(C2) The $q \times q$ matrix $\mathrm{E}(ZZ^\top|T)$ is nonsingular for each $T \in \mathcal{T}$. $\mathrm{E}(XX^\top|T)$, $\mathrm{E}(ZZ^\top|T)$ and $\mathrm{E}(XZ^\top|T)$ are all Lipschitz continuous.

(C3) There is an $\kappa > 2$ such that $\mathrm{E}\|X\|^{2\kappa} < \infty$, $\mathrm{E}\|Z\|^{2\kappa} < \infty$, $\mathrm{E}\|\varepsilon\|^{2\kappa} < \infty$ and $\mathrm{E}\|U\|^{2\kappa} < \infty$, and for some $\delta < 2 - \kappa^{-1}$ there is $n^{2\delta-1}h \to \infty$ as $n \to \infty$.

(C4) All of the coefficient functions $\{\alpha_j(\cdot), j = 1, \ldots, q\}$ are Lipschitz continuous and have continuous second order derivatives on $\mathcal{T}$.

(C5) The function $K(\cdot)$ is a symmetric density function with compact support and the bandwidth h satisfies $nh^8 \to 0$ and $nh^2/(\log n)^2 \to \infty$ as $n \to \infty$.

(C6) $\min\{|\beta_j^*|, \; j = 1, \ldots, s_n\}/\lambda_n \to \infty$ as $n \to \infty$.

(C7) There exist constant c and C such that $0 < c < \Lambda_{\min}(\Sigma_1) < \Lambda_{\max}(\Sigma_1) < C < \infty$ for all n, where $\Lambda_{\min}(M)$ and $\Lambda_{\max}(M)$ denote respectively the smallest and largest eigenvalues of symmetric matrix M.

(P1) Let $a_n = \max_{1\le j\le p_n}\{p'_{\lambda_n}(|\beta_j^*|), \beta_j^* \neq 0\}$ and $b_n = \max_{1\le j\le p_n}\{p''_{\lambda_n}(|\beta_j^*|), \; \beta_j^* \neq 0\}$. Assume that $a_n = O(n^{-1/2})$ and $b_n \to 0$ as $n \to \infty$. In addition, there exist constants c and C such that, when $\theta_1, \theta_2 \ge c\lambda_n$, $|p''_{\lambda_n}(\theta_1) - p''_{\lambda_n}(\theta_2)| \le C|\theta_1 - \theta_2|$.

(P2) $\liminf_{n\to\infty} \liminf_{\theta\to 0+} p'_{\lambda_n}(\theta)/\lambda_n > 0$.

These conditions, while a little bit lengthy at first look, are actually quite mild and may be further relaxed. Conditions (C1)–(C5) are also used by [3]. Conditions (C6)–(C7) and (P1)–(P2) are adopted from [5], see [5] for details. Condition (C6) gives the rate at which the penalized estimator can distinguish nonvanishing parameters from 0, which is necessary for obtaining the oracle property. In the finite-parameter situation this condition is implicitly assumed, and is in fact stronger than that imposed here. Condition (C7) assumes that the Σ_1 is positive definite and its eigenvalues are uniformly bounded. Conditions (P1)–(P2) are regularity conditions on penalty function.

The following theorem demonstrates that the convergence rate for the penalized bias-corrected estimator depends on the penalty function and the regularization parameter λ_n through a_n.

2.1. Theorem. *(Existence) Suppose the penalty function satisfies condition (P1). Under regularity conditions (C1)–(C5), if $\lambda_n \to 0$ and $p_n^4/n \to 0$ as $n \to \infty$, then with probability tending to 1, there is a local minimizer $\hat{\beta}$ of* (2.6) *such that $\|\hat{\beta} - \beta^*\| = O_P\{\sqrt{p_n}(n^{-1/2} + a_n)\}$.*

The proof of this theorem is given in Section 6. As it can be seen from the statement of Theorem 2.1, it requires that λ_n and the penalty function must be chosen such that $a_n = O(n^{-1/2})$ to achieve $\sqrt{n/p_n}$ convergence rate (or $\sqrt{n}$ convergence rate for finite and fixed p). For the L_1 penalty, $a_n = \lambda_n$. Thus, the $\sqrt{n/p_n}$ convergence rate requires that $\lambda_n = O(n^{-1/2})$. This requirement will make it difficult to choose λ_n in practice. However, if condition (C6) is satisfied, it is clear that $a_n = 0$ as when n is large enough for the SCAD penalty. Thus, the resulting estimator is $\sqrt{n/p_n}$ consistent, and no requirements are imposed on the convergence rate of λ_n. Note that the optimal bandwidth $h = O(n^{-1/5})$ is included in Theorem 2.1. Hence $\sqrt{n/p_n}$-consistency is achieved without the need of undersmoothing of the nonparametric component.

2.2. Theorem. *(Oracle property). Suppose the penalty function satisfies conditions (P1)–(P2). Under regularity conditions (C1)–(C7), if $\lambda_n \to 0$, $p_n^5/n \to 0$ and $\sqrt{n/p_n}\lambda_n \to \infty$ as $n \to \infty$, then with probability tending to 1, the $\sqrt{n/p_n}$-consistent local minimizer $\hat\beta = (\hat\beta_I^\top, \hat\beta_{II}^\top)^\top$ in Theorem 2.1 must satisfy: (i) (Sparsity) $\hat\beta_{II} = 0$; (ii)(Asymptotic normality) Let A_n be a determinstic $l \times s_n$ matrix with l fixed and $A_nA_n^\top \to G$, a positive definite matrix. Then*

$$\sqrt{n}A_n\Sigma_{2I}^{-1/2}\{\Sigma_{1I} + \Sigma_{\lambda_n}\}[\hat\beta_I - \beta_I^* + \{\Sigma_{1I} + \Sigma_{\lambda_n}\}^{-1}B] \xrightarrow{D} N(0_l, G),$$

where Σ_{1I} and Σ_{2I} are the top left-hand $s_n \times s_n$ submatrix of Σ_1 and Σ_2, respectively.

Theorem 2.2 is proved in Section 6. It is easy to see that sparsity and asymptotic normality are still valid when the number of parameter diverges in PLVCE models. For some penalty functions, including the SCAD penalty, B and Σ_{λ_n} are zero when n is large enough. Hence the results in Theorem 2.2 imply that the proposed procedure has the celebrated oracle property, i.e., $\hat\beta_{II} = 0$ and $\sqrt{n}A_n\Sigma_{2I}^{-1/2}\Sigma_{1I}(\hat\beta_I - \beta_I^*) \xrightarrow{D} N(0_l, G)$. On the other hand, the L_q penalty, $q \geq 1$, cannot simultaneously satisfy the conditions $\lambda_n = O_P(n^{-1/2})$ and $\sqrt{n/p_n}\lambda_n \to \infty$ as $n \to \infty$. These penalty functions cannot produce estimators with the oracle property. The L_q penalty, $q < 1$, may satisfy these two conditions at same time, but the bias term in Theorem 2.2(ii) cannot be ignored.

To make statistical inference on β_I^*, we need to estimate the standard error of the estimator of $\hat\beta_I$. The standard errors for estimated parameters can be obtained directly because we are estimating parameters and selecting variables at the same time. From Theorem 2.2, we can further approximate the estimation variance of the resulting estimator by the sandwich formula. Namely

$$\frac{1}{n}\{\widehat\Sigma_{1I} + \Sigma_{\lambda_n}(\hat\beta_I)\}^{-1}\widehat\Sigma_{2I}\{\widehat\Sigma_{1I} + \Sigma_{\lambda_n}(\hat\beta_I)\}^{-1}, \tag{2.8}$$

where $\hat\Sigma_{1I}$, a consistent estimate of Σ_{1I}, is defined as

$$\hat\Sigma_{1I} = \frac{1}{n}\nabla^2\widehat{L}_{nI}(\hat\beta_I) = \frac{1}{n}\sum_{i=1}^{n}\left(W_{Ii} + \frac{\partial\hat\alpha(T_i;\hat\beta_I)}{\partial\beta_I}Z_i\right)^{\otimes 2} - \Sigma_{UI},$$

and $\hat\Sigma_{2I} = \mathrm{Cov}(\nabla\widehat{L}_{nI}(\hat\beta_I))$ is given by

$$\frac{1}{n}\sum_{i=1}^{n}\left\{(Y_i - W_{Ii}^\top\hat\beta_I - Z_i^\top\hat\alpha(T_i;\hat\beta_I))(W_{Ii} + \frac{\partial\hat\alpha(T_i;\hat\beta_I)}{\partial\beta_I}Z_i) + \Sigma_{UI}\hat\beta_I\right\}^{\otimes 2},$$

furthermore, $\Sigma_{\lambda_n}(\hat\beta_I)$ is obtained by replacing β_I^* by $\hat\beta_I$ in Σ_{λ_n}.

The consistency of the proposed sandwich formula can be shown by using similar techniques as in [5]. The accuracy of this sandwich formula will be tested in our simulation studies.

3. Issues in practical implementation

In this section, we present a computational algorithm for obtaining the estimator and selection methods for the tuning parameters.

3.1. Computational algorithm. Since some penalty functions such as the SCAD penalty and $L_q, 0 \leq q \leq 1$ penalty are singular at the origin, it is challenging to minimize the penalized bias-corrected least squares function of (2.6). Following the idea of [4], we apply iterative algorithm based on the local quadratic approximation (LQA) of the penalty function. More specifically, suppose that at the kth step of the iteration, we obtain the value $\hat{\beta}^{(k)}$ that is close to the true value β^*. If $\hat{\beta}_j^{(k)}$ is very close to 0, then set $\hat{\beta}_j^{(k+1)} = 0$, and exclude the corresponding covariate from the model. Otherwise, an approximation of the penalty function at value $\hat{\beta}_j^{(k)}$ can be given by

$$p_\lambda(|\beta_j|) \approx p_\lambda(|\hat{\beta}_j^{(k)}|) + \frac{1}{2}\frac{p'_\lambda(|\hat{\beta}_j^{(k)}|)}{|\hat{\beta}_j^{(k)}|}(\beta_j^2 - \hat{\beta}_j^{(k)2}),$$

Consequently, with a slight abuse of notation, removing irrelevant terms we undate the estimate of β repeatedly until convergence with

$$\beta^{(k+1)} = \arg\min_{\beta}\left\{\widehat{L}_n(\beta) + \frac{n}{2}\beta^\top \Sigma_{\lambda_n}^{\mathrm{LQA}}(\hat{\beta}^{(k)})\beta\right\}, \tag{3.1}$$

where $\Sigma_{\lambda_n}^{\mathrm{LQA}}(\hat{\beta}^{(k)}) = \mathrm{diag}\{p'_{\lambda_n}(|\hat{\beta}_1^{(k)}|)/|\hat{\beta}_1^{(k)}|, \ldots, p'_{\lambda_n}(|\hat{\beta}_{p_n}^{(k)}|)/|\hat{\beta}_{p_n}^{(k)}|\}$. Hence, the foregoing discussion leads to the following iterating algorithm:

Step 1. Given an initial estimate $\hat{\beta}^{(0)}$.

Step 2. Update $\hat{\beta}^{(1)}$ by (3.1).

Step 3. Set $\hat{\beta}^{(0)} = \hat{\beta}^{(1)}$. Iterate Step 1 and 2 until convergence, and denote the final estimator $\hat{\beta}$.

In the initialization step, the initial estimators do not affect the degree of sparsity of the solution and the accuracy of the final estimator, but they will affect the speed of convergence of our iterative algorithm. In the following simulations, we obtain an initial estimator using a bias-corrected ordinary least-squares method based on (2.5). The simulation results show that such a choice is workable. During the iterations, to avoid numerical instability we need to keep track of zero coefficients and modify the penalty terms accordingly once $|\hat{\beta}_j^{(0)}|$ drops below a certain threshold ϵ ($\epsilon = 10^{-4}$ in our implementation). Specifically, in Step 2, if $|\hat{\beta}_j^{(0)}| < \epsilon$, then set $\hat{\beta}_j^{(1)} = 0$, delete the jth component of the covariates from the iteration.

3.2. Tuning parameters selection. To implement the proposed method, the bandwidth h and the tuning parameters λ_n in the penalty functions should be chosen. It is desirable to have automatic, data-driven methods to select h and λ_n.

Bandwidth selection. Condition (C5) reveal the rate of h. Any bandwidths with this rate lead to the same limiting distribution for $\hat{\beta}$. Therefore, the bandwidth selection can be done in a standard routine. For simple calculation, the bandwidth h is taken to be $h = 0.5n^{-1/5}$ in this paper, which we find to work satisfactorily in a variety of setting. We also conduct a sensitivity analysis by shifting bandwidths around the selected values, and found that the results are stable. Thus, the simulation results are not sensitive to the choice of h within certain range.

Regularization parameters selection. Here, given h, we use the "leave one sample out" method to select the tuning parameter λ_n. This method has been widely applied in practice. The cross-validation score for λ_n is defined as

$$\mathrm{CV}(\lambda_n) = \sum_{i=1}^{n} (Y_i - W_i^\top \hat{\beta}^{-i} - Z_i^\top \hat{\alpha}^{-i}(T_i))^2 - \sum_{i=1}^{n} (\hat{\beta}^{-i})^\top \Sigma_U \hat{\beta}^{-i} \tag{3.2}$$

where $\hat{\beta}^{-i}$ is the solution based on (2.6) after deleting the ith observation, and $\hat{\alpha}^{-i}(T_i)$ is the estimator defined in (2.7) with $\hat{\beta}$ replaced by $\hat{\beta}^{-i}$. The CV tuning parameter λ_n^{CV} is selected to minimize (3.2), that is, $\lambda_n^{\mathrm{CV}} = \arg\min_{\lambda_n} \mathrm{CV}(\lambda_n)$.

We also can use any other appropriate selection method to select the tuning parameters such as GCV, AIC and BIC. However, the definition of the degrees of freedom for the effective parameters in our variable selection procedure poses great challenges. Then, it is inconvenient to use such selection criteria for our variable selection procedure. In addition, from our simulation experience, we found that the CV method used in this paper works well. Further study of the asymptotic property of the proposed tuning parameter selection is needed, but it is outside the scope of this paper.

4. Simulation studies

In this section we corroborate our theoretical results with numerical experiments on synthetic data examples. That is, we conduct simulations to evaluate the finite sample performance of the proposed methods. We focus on only the SCAD penalty and referred to the proposed procedure as $\mathrm{C}_{\mathrm{SCAD}}$. The $\mathrm{C}_{\mathrm{SCAD}}$ is compared with four alternative procedures as follows. The first is the naive penalized procedures with a direct replacement of X by W ignoring measurement error ($\mathrm{N}_{\mathrm{SCAD}}$). The second is the estimators with considering measurement errors, but not penalized for complexity (Full). As a benchmark, two oracle methods in which the nonzero subset of slope β were known are implemented. In particular, the first (Oracle_1) serves as the gold standard, in which X can be observed. The second (Oracle_2) is another type, in which using W based on bias-corrected due to measurement errors.

We simulate data from model (1.1) and (1.2) with $q = 2$ and $p_n = \lfloor 1.8n^{1/3} \rfloor$ where $\lfloor k \rfloor$ denote the largest integer not greater than k, in which $\alpha_1(t) = 2\sin(2\pi t)$ and $\alpha_2(t) = 16t(1-t) - 2$, and $\beta = (2, -1.5, 4, 0, \ldots, 0)^\top$. Thus the first $s_n = 3$ regression variables were significant, but the remaining were not. The rate $p_n = \lfloor 1.8n^{1/3} \rfloor$ is not the same as presented in the theorems in Section 2, but we use this to show the capability of handling a higher rate of parameters growth for proposed method. The index variable T is sampled uniformly on $[0, 1]$. The covariates (X, Z) are taken from multivariate normal distribution $N_{p_n+q}(0, \Sigma)$. We consider Toeplitz convariance matrices $\Sigma_{ij} = \varrho^{|i-j|}$, in which both independent ($\varrho = 0$) and correlated cases ($\varrho = 0.5$) are taken into account. Y is generated according to the model, where noise term $\varepsilon \sim N(0, \sigma^2)$, and two different value $\sigma^2 = 0.5$ and 1, which represent strong and weak signal-to-noise ratios, were considered. Moreover, we assume that measurement error $U \sim N(0, \sigma_U^2 I_{p_n})$, where we take $\sigma_U = 0.2$ and 0.4 to represent different level of measurement errors. We perform 1000 simulations for all configurations with sample size $n = 100$ and $n = 400$ respectively. In all simulations, as a commonly adopted strategy we use the Epanechnikov kernel function $K(t) = 0.75(1 - t^2)_+$.

To assess the performance of different methods, we adopt the following criteria. For model error, the performance of estimator $\hat{\beta}$ will be assessed by using the generalized mean square error (GMSE), defined as

$$\mathrm{GMSE} = (\hat{\beta} - \beta^*)^\top (\mathrm{E}WW^\top - \Sigma_U)(\hat{\beta} - \beta^*).$$

Table 1. Simulation results with different methods for $\sigma^2 = 1$ over 1000 repetitions

		$\sigma_U = 0.2$					$\sigma_U = 0.5$				
ϱ	Method	True	C	IC	GMSE	RASE	True	C	IC	GMSE	RASE
						$(n, p_n) = (100, 8)$					
0	$\mathrm{C}_{\mathrm{SCAD}}$	99.9	5.000	0.001	0.072	0.556	91.0	4.970	0.063	0.546	1.030
	$\mathrm{N}_{\mathrm{SCAD}}$	99.6	5.000	0.004	0.100	0.550	80.0	5.000	0.205	1.377	0.897
	Full	0.00	0.020	0.000	0.188	0.571	0.00	0.016	0.000	1.060	1.107
	Oracle$_1$	100	5	0	0.033	0.551	100	5	0	0.033	0.950
	Oracle$_2$	100	5	0	0.071	0.556	100	5	0	0.399	1.018
0.5	$\mathrm{C}_{\mathrm{SCAD}}$	99.6	5.000	0.004	0.076	0.621	69.3	4.447	0.078	2.051	1.448
	$\mathrm{N}_{\mathrm{SCAD}}$	88.7	5.000	0.114	0.244	0.622	1.70	4.999	1.112	2.071	0.985
	Full	0.00	0.008	0.000	0.202	0.647	0.00	0.006	0.000	2.801	1.638
	Oracle$_1$	100	5	0	0.033	0.614	100	5	0	0.035	1.075
	Oracle$_2$	100	5	0	0.071	0.620	100	5	0	0.607	1.220
						$(n, p_n) = (400, 13)$					
0	$\mathrm{C}_{\mathrm{SCAD}}$	100	10.00	0.000	0.015	0.274	99.9	10.00	0.001	0.072	0.469
	$\mathrm{N}_{\mathrm{SCAD}}$	100	10.00	0.000	0.046	0.271	97.7	10.00	0.023	0.982	0.430
	Full	0.0	0.12	0.000	0.067	0.277	0.0	0.055	0.000	0.300	0.481
	Oracle$_1$	100	10	0	0.008	0.273	100	10	0	0.007	0.463
	Oracle$_2$	100	10	0	0.015	0.274	100	10	0	0.071	0.469
0.5	$\mathrm{C}_{\mathrm{SCAD}}$	99.9	9.999	0.000	0.016	0.314	99.2	10.00	0.009	0.099	0.548
	$\mathrm{N}_{\mathrm{SCAD}}$	98.7	10.00	0.013	0.085	0.310	0.0	10.00	1.008	1.873	0.475
	Full	0.0	0.090	0.000	0.073	0.321	0.0	0.032	0.000	0.384	0.589
	Oracle$_1$	100	10	0	0.007	0.313	100	10	0	0.007	0.534
	Oracle$_2$	100	10	0	0.016	0.314	100	10	0	0.087	0.548

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

$$\mathrm{RASE} = \left\{ N_{\mathrm{grid}}^{-1} \sum_{k=1}^{N_{\mathrm{grid}}} \|\hat{\alpha}(t_k) - \alpha(t_k)\|^2 \right\}^{1/2},$$

over $N_{\mathrm{grid}} = 200$ grid points $\{t_k\}$. Table 1 presents the mean of GMSE and RASE over the 1000 simulations. For the selected model, the model complexity is summarized in terms of the number of zero coefficients for the parametric components, as also reported in Table 1. In Table 1, the column labeled "C" is the average numbers of zero coefficients correctly estimated to be zero, and the column labeled "IC" depicts the average numbers of nonzero coefficients erroneously set to zero. Furthermore, the column labeled "True" is the proportion of times the true model is exactly identified.

From Table 1, we can make the following observations: (i) The performances of both $\mathrm{C}_{\mathrm{SCAD}}$ and $\mathrm{N}_{\mathrm{SCAD}}$ procedures become better in terms of model error and model complexity as the level of measurement error decreases. (ii) Both variable selection procedures perform very similarly when the level of measurement error is small. However, when the level of measurement error is large, the performance of $\mathrm{C}_{\mathrm{SCAD}}$ is significantly better than that of $\mathrm{N}_{\mathrm{SCAD}}$. The latter cannot eliminate some unimportant variables and gives larger model errors. This implies that the estimators based on the $\mathrm{N}_{\mathrm{SCAD}}$ procedure are biased. (iii) In addition, as expected, the performance of the Oracle$_1$ procedure is best in all cases in terms of model error. Furthermore, the performance of $\mathrm{C}_{\mathrm{SCAD}}$ becomes increasingly closer to that based on the Oracle$_2$ procedure as the level of measurement error decreases or ϱ decreases. (iv) As the sample size increases, the performance of all methods becomes better. To save space the simulation results, for others settings with $\sigma^2 = 0.5$, are not showed here. The above conclusions can also be drawn similarly except now all approaches perform better than they done when $\sigma^2 = 1$ as presented in Table 1. These findings imply that the model selection result based on the $\mathrm{C}_{\mathrm{SCAD}}$ approach is satisfactory and the selected model is very close to the true model in terms of nonzero coefficients.

Table 2. Bias and standard deviations of estimators for $\sigma^2 = 1$, $\sigma_U = 0.5$ and $\varrho = 0.5$

	$\hat{\beta}_1$			$\hat{\beta}_2$			$\hat{\beta}_3$		
Method	Bias	SD	SDE(sd(SDE))	Bias	SD	SDE(sd(SDE))	Bias	SD	SDE(sd(SDE))
				$(n, p_n) = (100, 8)$					
$\mathrm{C}_{\mathrm{SCAD}}$	0.769	3.689	0.687(2.701)	1.378	9.414	0.973(4.749)	1.532	13.419	0.823(3.894)
$\mathrm{N}_{\mathrm{SCAD}}$	0.875	0.492	0.251(0.063)	1.471	0.178	0.263(0.049)	1.215	0.252	0.228(0.033)
Oracle_1	0.101	0.127	0.119(0.049)	1.111	0.139	0.139(0.146)	0.100	0.126	0.121(0.084)
Oracle_2	0.430	0.532	0.785(6.753)	0.597	0.678	1.364(4.364)	0.521	0.585	0.837(8.114)
Full	0.847	3.695	1.091(7.934)	1.534	9.477	1.784(8.565)	1.806	13.476	1.211(9.255)
				$(n, p_n) = (400, 13)$					
$\mathrm{C}_{\mathrm{SCAD}}$	0.162	0.200	0.193(0.023)	0.212	0.282	0.241(0.039)	0.180	0.217	0.206(0.026)
$\mathrm{N}_{\mathrm{SCAD}}$	0.758	0.145	0.117(0.021)	1.500	0.000	0.125(0.010)	1.253	0.101	0.111(0.008)
Oracle_1	0.046	0.058	0.058(0.004)	0.055	0.068	0.064(0.004)	0.050	0.063	0.058(0.004)
Oracle_2	0.159	0.193	0.194(0.023)	0.204	0.249	0.243(0.045)	0.177	0.211	0.207(0.026)
Full	0.181	0.208	0.200(0.025)	0.249	0.277	0.249(0.044)	0.263	0.280	0.215(0.030)

We now verify the consistency of the estimators and test the accuracy the standard error formula. Table 2 displays the bias (columns labeled Bias) and sample standard deviation (columns labeled SD) of the estimates for three nonzero coefficients, over 1000 simulations. These can be regard as the true standard errors and compared with 1000 estimated standard errors. The 1000 estimated standard errors by using the sandwich formula are summarized by their mean (columns labeled SDE) and the sample standard deviations (sd(SDE)). The accuracy gets better when n increases. We omit here the results for other configurations, only for case $\sigma^2 = 1$, $\sigma_U = 0.5$ and $\varrho = 0.5$. Overall, the estimators are consistent and the sandwich formula works well.

5. Discussion

In this paper, we have proposed a variable selection procedure for the high dimensional PLVCE models. Our method extends the variable selection procedure to the setting, in which high dimension, measurement error, semiparametric models are considered at the same time. We have shown that the proposed method is consistent in variable selections, and the estimators of the regression coefficients have oracle property. Simulation studies indicate that the proposed method seems rather encouraging. To conclude this article, we would like to discuss some interesting topics for future study. Firstly, in this paper, we assume that the covariance matrix of measurement errors is known. However, it is usually unknown in many applications. If the covariance matrix is unknown, the variable selection procedure proposed by this paper will not work any more unless repeated measurements of the data are available. As a future research topic, it is interest to consider the variable selection for the high dimensional PLVCE models when the covariance matrix of measurement errors is unknown. Secondly, it is interesting to perform variable selection for $p_n \gg n$. Variable selection for large p_n, small n setting is a very active research topic. However, it is challenging to extend the existing procedures for large p_n, small n problems to measurement error data. The details will also be further investigated in the future.

6. Proofs

In order to prove the main results, we first introduce several lemmas. Let $\mu_k = \int t^k K(t)dt$, $\nu_k = \int t^k K^2(t)dt$, $c_n = h^2 + [\log(1/h)/nh]^{1/2}$. Set $\Psi(T_1) = \mathrm{E}(X_1 Z_1^\top | T_1)$, $\Upsilon(T_1) = \mathrm{E}(Z_1 Z_1^\top | T_1)$ and $\Xi(T_1;\beta) = \mathrm{E}[Z_1(Y_1 - X_1^\top \beta)|T_1]$. Furthermore, denote by $\alpha(t;\beta)$ the 'least favorable curve' of the nonparametric function $\alpha(t)$, which is defined as

$$\alpha(t;\beta) = \arg\min_{\eta} \mathrm{E}[(Y_i - W_i^\top \beta - Z_i^\top \eta)^2 | T_i = t] = \Upsilon^{-1}(t)\Xi(t;\beta), \tag{6.1}$$

and let $Q_n(\beta) = L_n(\beta) + +n\sum_{j=1}^{p_n} p_\lambda(|\beta_j|)$, where $2L_n(\beta) = \sum_{i=1}^n (Y_i - W_i^\top \beta - Z_i^\top \alpha(T_i;\beta))^2$ $-n\beta^\top \Sigma_U \beta$. Apparently, $\alpha(t;\beta^*) = \alpha^*(t)$ and $\frac{\partial \alpha(t;\beta)}{\partial \beta} = \Psi(t)\Upsilon^{-1}(t)$ is a p_n by q matrix. The following Lemma 6.1 can be found in [3].

6.1. Lemma. *Let $(X_i, Y_i), i = 1, \ldots, n$ be be i.i.d. random vectors, where the Y_i are scale random variables. Further assume that $E|y|^\kappa < \infty$ and $\sup_x \int |y|^\kappa f(x,y)dy < \infty$, where f denotes the joint density of (X, Y). Let K be a bounded positive function with a bounded support, satisfying a Lipschitz condition. Given that $n^{2\delta-1}h \to \infty$ for some $\delta < 1 - \kappa^{-1}$, then*

$$\sup_x \left| \frac{1}{n}\sum_{i=1}^n \Big\{ K_h(X_i - x)Y_i - \mathrm{E}[K_h(X_i - x)Y_i] \Big\} \right| = O_P\left(\left\{ \frac{\log(1/h)}{nh} \right\}^{1/2} \right).$$

6.2. Lemma. *Under regularity conditions (C1)–(C5), the following holds uniformly in $t \in \mathfrak{T}$,*

$$\hat{\alpha}(t;\beta) - \alpha(t;\beta) = O_P(c_n),$$

$$\frac{\partial \hat{\alpha}(t;\beta)}{\partial \beta_k} - \frac{\partial \alpha(t;\beta)}{\partial \beta_k} = O_P(c_n), \text{ for } k = 1, \ldots, p_n.$$

Proof. From Lemma 6.1, we have that

$$\begin{aligned} \frac{1}{n} D_t^\top \omega_t D_t &= \frac{1}{n}\sum_{i=1}^n \begin{pmatrix} Z_i Z_i^\top & Z_i Z_i^\top \frac{T_i - t}{h} \\ Z_i Z_i^\top \frac{T_i - t}{h} & Z_i Z_i^\top (\frac{T_i - t}{h})^2 \end{pmatrix} K_h(T_i - t) \\ &= \begin{pmatrix} 1 & 0 \\ 0 & \mu_2 \end{pmatrix} \otimes \Upsilon(t) f_T(t)\{1 + O_P(c_n)\} \quad \text{and} \end{aligned}$$

$$\begin{aligned} \frac{1}{n} D_t^\top \omega_t (\boldsymbol{Y} - \boldsymbol{W}\beta) &= \frac{1}{n}\sum_{i=1}^n \begin{pmatrix} Z_i(Y_i - W_i^\top \beta) \\ Z_i(Y_i - W_i^\top \beta)\frac{T_i - t}{h} \end{pmatrix} K_h(T_i - t) \\ &= \begin{pmatrix} 1 \\ 0 \end{pmatrix} \otimes \Xi(t;\beta) f_T(t)\{1 + O_P(c_n)\} \end{aligned}$$

hold uniformly in $t \in \mathfrak{T}$. Here the symbol $\otimes$ represent the Kronecker product between matrices. Hence, invoking equation (6.1) and $\hat{\alpha}(t;\beta)$ in Section 2, the first conclusion follows. The second assertion can get similarly. □

6.3. Lemma. *Under regularity conditions (C1)-(C5), if $p_n^\kappa / n \to 0$ for $\kappa > 5/4$, $h = O(n^{-\varsigma})$ with $(4\kappa)^{-1} < \varsigma < 1 - \kappa^{-1}$, then for any β,*

$$n^{-1/2} \|\nabla \widehat{L}_n(\beta) - \nabla L_n(\beta)\| = o_P(1).$$

Proof. Invoking Lemma 6.2, the column vector $n^{-1/2}(\nabla\widehat{L}_n(\beta)-\nabla L_n(\beta))$ has the kth component equals

$$\begin{aligned}&\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\Big\{(Y_i-W_i^\top\beta-Z_i^\top\hat{\alpha}(T_i;\beta))(-W_{ik}-\frac{\partial\hat{\alpha}(T_i;\beta)}{\partial\beta_k}Z_i)\\&\qquad-(Y_i-W_i^\top\beta-Z_i^\top\alpha(T_i;\beta))(-W_{ik}-\frac{\partial\alpha(T_i;\beta)}{\partial\beta_k}Z_i)\Big\}\\&=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\Big\{(Y_i-W_i^\top\beta-Z_i^\top\alpha(T_i;\beta))\frac{\partial\alpha(T_i;\beta)}{\partial\beta_k}Z_i+Z_i^\top\alpha(T_i;\beta)W_{ik}\Big\}O_P(c_n)\\&=O_P(c_n).\end{aligned}$$

Hence we have shown

$$n^{-1/2}\|\nabla\widehat{L}_n(\beta)-\nabla L_n(\beta)\|=O_P(\sqrt{p_n}c_n)=o_P(1),$$

and the proof is complete. □

6.4. Lemma. *Under the conditions of Theorem 1, we have*

$$\frac{[n^{-1/2}\nabla^\top L_n(\beta^*)](\Sigma_2)^{-1}[n^{-1/2}\nabla L_n(\beta^*)]-p_n}{\sqrt{2p_n}}\xrightarrow{D}N(0,1),$$

where $\Sigma_2=\mathrm{E}[(\varepsilon_i-U_i^\top\beta^*)(\Psi(T_i)\Upsilon^{-1}(T_i)Z_i-X_i)-\Sigma_U\beta^*]^{\otimes 2}$. *In addition,* $\nabla L_n(\beta^*)=O_P(\sqrt{np_n})$. *Likewise, the results above hold also by* $L_n(\beta^*)$ *replaced with* $\widehat{L}_n(\beta^*)$.

Proof. From (6.1), we get the following formulas $\mathrm{E}[Z_i(Y_i-X_i^\top\beta-Z_i^\top\alpha(T_i;\beta))|T_i=t]=0$ and $\mathrm{E}[X_iZ_i^\top+\frac{\partial\alpha(T_i;\beta)}{\partial\beta}Z_iZ_i^\top|T_i=t]=0$. Then $\mathrm{E}[\nabla L_n(\beta)]=0$ follows. Direct calculation yields

$$\nabla L_n(\beta)=\sum_{i=1}^{n}(Y_i-W_i^\top\beta-Z_i^\top\alpha(T_i;\beta))(\Psi(T_i)\Upsilon^{-1}(T_i)Z_i-W_i)-n\Sigma_U\beta.$$

Thus,

$$\frac{1}{\sqrt{n}}\nabla L_n(\beta^*)=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\Big\{(\varepsilon_i-U_i^\top\beta^*)(\Psi(T_i)\Upsilon^{-1}(T_i)Z_i-W_i)-\Sigma_U\beta^*\Big\}.$$

By applying the martingale central limit theorem as given in [9], we can easily obtain the first part. The second part follows from Lemma 6.3. □

6.5. Lemma. *Under regularity conditions C1–C5, and* $p^4/n=o(1)$,

$$\begin{aligned}&\|\frac{1}{n}\nabla^2L_n(\beta)-\Sigma_1\|=o_P(p_n^{-1}),\\&\|\frac{1}{n}\nabla^2\widehat{L}_n(\beta)-\Sigma_1\|=o_P(p_n^{-1})+O_P(p_nc_n),\end{aligned}$$

where $\Sigma_1=\mathrm{E}(X_1X_1^\top)-\mathrm{E}\{\Psi(T_1)\Upsilon^{-1}(T_1)\Psi^\top(T_1)\}$.

Proof. Direct calculation yields $n^{-1}\nabla^2L_n(\beta)=n^{-1}\sum_{i=1}^n(W_i-\Psi(T_i)\Upsilon^{-1}(T_i)Z_i)\ (W_i-\Psi(T_i)\Upsilon^{-1}(T_i)Z_i)^\top-\Sigma_U$. Then $\mathrm{E}[n^{-1}\nabla^2L_n(\beta)]=\mathrm{E}\{\mathrm{E}[(W_i-\Psi(T_i)\Upsilon^{-1}(T_i)Z_i)\ (W_i-\Psi(T_i)\Upsilon^{-1}(T_i)Z_i)^\top|T_i]\}-\Sigma_U=\Sigma_1$. The first conclusion follows from

$$\begin{aligned}\mathrm{E}p_n^2\|\frac{1}{n}\nabla^2L_n(\beta)-\Sigma_1\|^2&=p_n^2\mathrm{E}\sum_{j,k=1}^{p_n}\left\{\frac{1}{n}\nabla^2L_n(\beta)-\Sigma_1\right\}_{jk}^2\\&=O\left(\frac{p_n^4}{n}\right)=o(1).\end{aligned}$$

From this, triangle inequality immediately gives the second conclusion if we can show that

$$\|\frac{1}{n}\nabla^2\widehat{L}_n(\beta) - \frac{1}{n}\nabla^2 L_n(\beta)\| = O_P(p_nc_n). \tag{6.2}$$

To this end, for $k = 1, \dots, p_n$,

$$\begin{aligned}
&n^{-1}\frac{\partial}{\partial\beta_k}(\nabla\widehat{L}_n(\beta) - \nabla L_n(\beta)) \\
=&n^{-1}\frac{\partial}{\partial\beta_k}\sum_{i=1}^{n}\Big\{(Y_i - W_i^\top\beta - Z_i^\top\hat{\alpha}(T_i;\beta))(-W_i - \frac{\partial\hat{\alpha}(T_i;\beta)}{\partial\beta}Z_i) \\
&\qquad\qquad -(Y_i - W_i^\top\beta - Z_i^\top\alpha(T_i;\beta))(-W_i - \frac{\partial\alpha(T_i;\beta)}{\partial\beta}Z_i)\Big\} \\
=&n^{-1}\sum_{i=1}^{n}\Big\{(W_{ik} + \frac{\partial\hat{\alpha}(T_i;\beta)}{\partial\beta_k}Z_i)(W_i + \frac{\partial\hat{\alpha}(T_i;\beta)}{\partial\beta}Z_i) \\
&\qquad -(W_{ik} + \frac{\partial\alpha(T_i;\beta)}{\partial\beta_k}Z_i)(W_i + \frac{\partial\alpha(T_i;\beta)}{\partial\beta}Z_i)\Big\} \\
=&O_P(\sqrt{p_n}c_n)
\end{aligned}$$

where the last line follows from Lemma 6.2. Hence (6.2) follows and the proof completes. □

Proof of Theorem 2.1. Let $\vartheta_n = \sqrt{p_n}(n^{-1/2} + a_n)$ and set $\|v\| = C$, where C is a large enough constant. Our aim is to show that for any given $\epsilon > 0$ there is a large constant C such that, for large n we have

$$\Pr\Big\{\inf_{\|v\|=C}\widehat{Q}_n(\beta^* + \vartheta_n v) > \widehat{Q}_n(\beta^*)\Big\} \geq 1 - \epsilon. \tag{6.3}$$

This implies that with probability tending to 1 there is a local minimizer $\hat{\beta}$ in the ball $\{\beta^* + \vartheta_n v : \|v\| \leq C\}$ such that $\|\hat{\beta} - \beta^*\| = O_P(\vartheta_n)$.

Let $\Delta_n(v) = \widehat{Q}_n(\beta^* + \vartheta_n v) - \widehat{Q}_n(\beta^*)$. Recall that the first s_n components of β^* are nonzero, and $p_\lambda(\cdot)$ is nonnegative and $p_\lambda(0) = 0$. By the Taylor expansion and the fact that $\widehat{L}_n(\beta)$ is quadratic, we have

$$\begin{aligned}
\Delta_n(v) &\geq \widehat{L}_n(\beta^* + \vartheta_n v) - \widehat{L}_n(\beta^*) + n\sum_{j=1}^{s_n}\{p_\lambda(|\beta_j^* + \vartheta_n v_j|) - p_\lambda(|\beta_j^*|)\} \\
&\geq \vartheta_n v^\top\nabla\widehat{L}_n(\beta^*) + \frac{1}{2}\vartheta_n^2 v^\top\nabla^2\widehat{L}_n(\beta^*)v \\
&\quad + \sum_{j=1}^{s_n} n\vartheta_n p_\lambda'(|\beta_j^*|)\mathrm{sign}(\beta_j^*)v_j + \frac{1}{2}\sum_{j=1}^{s_n} n\vartheta_n^2 p_\lambda''(|\beta_j^*|)v_j^2\{1 + o(1)\} \\
&\triangleq D_1 + D_2 + D_3 + D_4.
\end{aligned}$$

By Lemma 6.4 and $\sqrt{p_n} \leq \sqrt{n}\vartheta_n$, we get

$$\begin{aligned}
|D_1| &= |\vartheta_n v^\top\nabla\widehat{L}_n(\beta^*)| \leq \vartheta_n\|\nabla\widehat{L}_n(\beta^*)\|\|v\| \\
&\leq O_P(\vartheta_n\sqrt{np_n})\|v\| \leq O_P(n\vartheta_n^2)\|v\|
\end{aligned}$$

Next we consider D_2, An application of Lemma 6.5 yields that

$$\begin{aligned} D_2 &= \frac{1}{2}\vartheta_n^2 v^\top \nabla^2 \widehat{L}_n(\beta^*) v = \frac{1}{2} n\vartheta_n^2 v^\top [\frac{1}{n}\nabla^2 \widehat{L}_n(\beta^*) - \Sigma_1] v + \frac{1}{2} n\vartheta_n^2 v^\top \Sigma_1 v \\ &= \frac{1}{2} n\vartheta_n^2 v^\top \Sigma_1 v + o_P(1) n\vartheta_n^2 \|v\|^2. \end{aligned}$$

With regard to D_3 and D_4, for $\sqrt{s_n} a_n \leq \sqrt{s_n}(n^{-1/2} + a_n) \leq \vartheta_n$, we have

$$\begin{aligned} |D_3| &\leq \sum_{j=1}^{s_n} \left| n\vartheta_n p'_\lambda(|\beta_j^*|) \text{sign}(\beta_j^*) v_j \right| \\ &\leq n\vartheta_n a_n \sum_{j=1}^{s_n} |v_j| \leq n\vartheta_n a_n \sqrt{s_n} \|v\| \leq n\vartheta_n^2 \|v\|, \quad \text{and} \\ |D_4| &= \frac{1}{2} \sum_{j=1}^{s_n} n\vartheta_n^2 p''_\lambda(|\beta_j^*|) v_j^2 \{1 + o(1)\} \leq b_n n\vartheta_n^2 \|v\|^2. \end{aligned}$$

Therefore, under the condition (P1), by allowing C to be large enough, all terms D_1, D_3, D_4 are dominated by D_2, which is positive. This proves (6.3) and completes the proof. □

Proof of Theorem 2.2. Let $\zeta_n = C\sqrt{p_n/n}$. It is sufficient to show that with probability tending to 1 as $n \to \infty$, for any β satisfying $\|\beta - \beta^*\| = O_P(\sqrt{p_n/n})$ we have, for $j = s_n + 1, \ldots, p_n$,

$$\text{(6.4)} \qquad \frac{\partial \widehat{Q}_n(\beta)}{\partial \beta_j} < 0 \text{ for } \beta_j \in (-\zeta_n, 0) \quad \text{and} \quad \frac{\partial \widehat{Q}_n(\beta)}{\partial \beta_j} > 0 \text{ for } \beta_j \in (0, \zeta_n).$$

By Taylor expansion and the fact that $\widehat{L}_n(\beta)$ is quadratic in β, we get

$$\begin{aligned} \frac{\partial \widehat{Q}_n(\beta)}{\partial \beta_j} &= \frac{\partial \widehat{L}_n(\beta)}{\partial \beta_j} + np'_\lambda(|\beta_j|)\text{sign}(\beta_j) \\ &= \frac{\partial \widehat{L}_n(\beta^*)}{\partial \beta_j} + \sum_{k=1}^{p_n} \frac{\partial^2 \widehat{L}_n(\beta^*)}{\partial \beta_j \partial \beta_k}(\beta_k - \beta_k^*) + np'_\lambda(|\beta_j|)\text{sign}(\beta_j) \\ &\triangleq J_1 + J_2 + J_3. \end{aligned}$$

Next, we consider J_1, J_2. Invoking Lemma 6.4, we have

$$J_1 = O_P(\sqrt{n}) = O_P(\sqrt{np_n}).$$

The term J_2 can be written as $J_2 = \sum_{k=1}^{p_n} \left\{ \frac{\partial^2 \widehat{L}_n(\beta^*)}{\partial \beta_j \partial \beta_k} - n\Sigma_{1,jk} \right\} (\beta_k - \beta_k^*) + n\sum_{k=1}^{p_n} \Sigma_{1,jk}(\beta_k - \beta_k^*) \triangleq J_{21} + J_{22}$. Using the Cauchy-Schwarz inequality and $\|\beta - \beta^*\| = O_P(\sqrt{p_n/n})$, we have

$$|J_{22}| \leq n \sum_{k=1}^{p_n} |\Sigma_{1,jk}(\beta_k - \beta_k^*)| \leq nO_P(\sqrt{p_n/n}) \left[\sum_{k=1}^{p_n} (\Sigma_{1,jk})^2 \right]^{1/2}.$$

As the eigenvalues of Σ_1 are bounded according to condition (C7), we have $\sum_{k=1}^{p_n} (\Sigma_1)_{jk}^2 = O(1)$. This entails that $J_{22} = O_P(\sqrt{np_n})$. For J_{21}, applying the Cauchy-Schwarz inequality,

$$|J_{21}| \leq \|\beta - \beta^*\| \left[\sum_{k=1}^{p_n} \left\{ \frac{\partial^2 \widehat{L}_n(\beta^*)}{\partial \beta_j \partial \beta_k} - n\Sigma_{1,jk} \right\}^2 \right]^{1/2}.$$

By a standard argument from condition (C7), we have

$$\left[\sum_{k=1}^{p_n}\left\{\frac{\partial^2 L_n(\beta^*)}{\partial\beta_j\partial\beta_k} - n\Sigma_{1,jk}\right\}^2\right]^{1/2} = O_P(n).$$

Then $J_{21} = O_P(\sqrt{np_n})$ follows form $\|\hat\beta - \beta^*\| = O_P(\sqrt{p_n/n})$. Now we have

$$J_2 = O_P(\sqrt{np_n}).$$

Hence we have

$$\frac{\partial\widehat{Q}_n(\beta)}{\partial\beta_j} = n\lambda\left\{\frac{p'_\lambda(|\beta_j|)}{\lambda}\mathrm{sign}(\beta_j) + O_P\left(\frac{\sqrt{p_n/n}}{\lambda}\right)\right\}.$$

Because of $\sqrt{p_n/n}/\lambda \to 0$ and (P2), the sign of β_j completely determines the sign of $\partial\widehat{Q}_n(\beta)/\partial\beta_j$. Then (6.4) follows from the continuity of $\partial\widehat{Q}_n(\beta)/\partial\beta_j$. Combining with the result of Theorem 2.1, there is a $\sqrt{n/p_n}$-consistent local minimizer $\hat\beta$ of $\widehat{Q}_n(\beta)$ and $\hat\beta$ has the form $(\hat\beta_I^\top, 0^\top)^\top$, i.e. part (i) holds.

Now we prove part (ii). As shown in Theorem 2.1, we let λ_n be sufficiently small so that $a_n = o(n^{-1/2})$, then $\hat\beta$ is $\sqrt{n/p_n}$ consistent. By part (i), each component of $\hat\beta_I$ stays away from zero for a sufficiently large sample size n because β_I^* is away from zero. At the same time, $\hat\beta_{II} = 0$ with probability tending to 1. As a consequence, the estimate $\hat\beta_I$ based on the penalized estimation are necessarily the solution of the following estimation equation

$$\nabla\widehat{L}_{nI}(\hat\beta_I) + nP'_\lambda(|\hat\beta_I|) = 0 \tag{6.5}$$

where $P'_\lambda(|\hat\beta_I|)$ is a s_n-vector whose jth element is $p'_\lambda(|\hat\beta_j|)\mathrm{sign}(\hat\beta_j)$. Applying a Taylor expansion to (6.5) and re-arranging the resulting terms, we have

$$(\Sigma_{1I} + \Sigma_{\lambda_n})(\hat\beta_I - \beta_I^*) + P'_\lambda(|\beta_I^*|) = -\frac{1}{n}\nabla L_{nI}(\beta_I^*) + R_1 + R_2$$

where $R_1 = -\left[\frac{1}{n}\nabla^2\widehat{L}_{nI}(\beta_I^*) + P''_\lambda(|\tilde\beta_I|) - \Sigma_{1I} - \Sigma_{\lambda_n}\right](\hat\beta_I - \beta_I^*)$ and $R_2 = \frac{1}{n}\nabla L_{nI}\ (\beta_I^*) - \frac{1}{n}\nabla\widehat{L}_{nI}(\beta_I^*)$. By Lemma 6.5 and Cauchy-Schwarz inequality, $\|R_1\| = o_P((np_n)^{-1/2}) + O_P(\sqrt{p_n^3/nc_n}) = o_P(n^{-1/2})$. By Lemma 6.3, we have $R_2 = o_P(n^{-1/2})$. Hence, we have

$$\begin{aligned}&\sqrt{n}A_n\Sigma_{2I}^{-1/2}\{\Sigma_{1I} + \Sigma_\lambda\}\{(\hat\beta_I - \beta_I^*) + \{\Sigma_{1I} + \Sigma_\lambda\}^{-1}B\}\\ &= -\frac{1}{\sqrt{n}}A_n\Sigma_{2I}^{-1/2}\nabla L_{nI}(\beta_I^*) + o_P(1),\end{aligned}$$

Since $\|A_n\Sigma_{2I}^{-1/2}\| = O(1)$ by conditions of this theorem.

Next, we verify the Lindeberg-Feller Central Limit Theorem for the last term above. Let

$$\psi_{ni} = \frac{1}{\sqrt{n}}A_n\Sigma_{2I}^{-1/2}\nabla L_{nIi}(\beta_I^*), \quad i = 1, \ldots, n,$$

where $\nabla L_{nIi}(\beta_I^*) = \Big\{(Y_i - W_{Ii}^\top\beta_I^* - Z_i^\top\alpha(T_i;\beta_I^*))(W_{Ii} + \frac{\partial\alpha(T_i;\hat\beta_I)}{\partial\beta_I}) + \Sigma_{UI}\beta_I^*\Big\}$. For any $\epsilon > 0$,

$$\begin{aligned}\sum_{i=1}^n \mathrm{E}\|\psi_{ni}\|^2 I\{\|\psi_{ni}\| > \epsilon\} &= n\mathrm{E}\|\psi_{n1}\|^2 I\{\|\psi_{n1}\| > \epsilon\}\\ &\le n\{\mathrm{E}\|\psi_{n1}\|^4\}^{1/2}\{\Pr(\|\psi_{n1}\| > \epsilon)\}^{1/2}.\end{aligned}$$

Using Chebyshev's inequality, we have $\Pr(\|\psi_{n1}\| > \epsilon) \leq \frac{\mathrm{E}\|\psi_{n1}\|^2}{\epsilon^2}$ $= \frac{\mathrm{E}\|A_n\Sigma_{2I}^{-1/2}\nabla L_{nIi}(\beta_I^*)\|^2}{n\epsilon^2} = O(n^{-1})$ and $\mathrm{E}\|\psi_{n1}\|^4 = \mathrm{E}(\psi_{ni}^\top\psi_{ni})^2 \leq \frac{1}{n^2}\Lambda_{\max}^2(A_nA_n^\top)$ $\Lambda_{\max}^2(\Sigma_{2I}^{-1})\mathrm{E}\|\nabla L_{nIi}^\top(\beta_I^*)\ \nabla L_{nIi}(\beta_I^*)\|^2 = O(\frac{s_n^2}{n^2})$, by condition (C7). Hence, we get

$$\sum_{i=1}^n \mathrm{E}\|\psi_{ni}\|^2 I\{\|\psi_{ni}\| > \epsilon\} = O(n\frac{s_n}{n}\frac{1}{\sqrt{n}}) = o(1).$$

Also, note that $\mathrm{E}\psi_{ni} = 0$ and

$$\sum_{i=1}^n \mathrm{Cov}(\psi_{ni}) = n\mathrm{Cov}(\psi_{n1}) = \mathrm{Cov}(A_n\Sigma_{2I}^{-1/2}\nabla L_{nIi}(\beta_I^*)) = A_nA_n^\top \to G.$$

From the foregoing argument, ψ_{ni} satisfies the conditions of the Lindeberg-Feller central limit theorem, then we complete the proof of part (ii). □

Acknowledgments

This work was supported by the National Natural Science Foundation of China (11571025, Key grant: 11331011), and the Beijing Natural Science Foundation (1182002).

References

[1] Ahmad, I., Leelahanon, S. and Li, Q. *Efficient estimation of a semiparametric partially linear varying coefficient model*, The Annals of Statistics **33** (1), 258–283, 2005.

[2] Carroll, R. J., Ruppert, D., Stefanski, L. A. and Crainiceanu, C. M. *Measurement Error in Nonlinear Models (2nd ed)*, New York: Chapman and Hall, 2006.

[3] Fan, J.Q. and Huang, T. *Profile likelihood inferences on semiparametric varying coefficient partially linear models*, Bernoulli **11** (6), 1031–1057, 2005.

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

[5] Fan, J.Q. and Peng, H. *Nonconcave penalized likelihood with a diverging number of parameters*, The Annals of Statistics **32** (3), 928–961, 2004.

[6] Feng, S.Y. and Xue, L.G. *Bias-corrected statistical inference for partially linear varying coefficient errors-in-variables models with restricted condition*, Annals of the Institute of Statistical Mathematics **66** (1), 121–140, 2014.

[7] Fuller, W. A. *Measurement Error Models.* New York: John Wiley, 1987.

[8] Hu, X.M., Wang, Z.Z. and Zhao, Z.Z. *Empirical likelihood for semiparametric varying-coefficient partially linear errors-in-variables models*, Statistics and Probability Letters **79** (8), 1044–1052, 2009.

[9] Hall, P. and Heyde, C. (1980). *Martingale limit theory and its application.* Academic Press.

[10] Huang, Z. and Zhang, R.Q. *Empirical likelihood for nonparametric parts in semiparametric varying-coefficient partially linear models*, Statistics and Probability Letters **79** (16), 1798–1808, 2009.

[11] Kai, B.,Li, R. Z. and Zou, H. *New efficient estimation and variable selection methods for semiparametric varying-coefficient partially linear models*, The Annals of Statistics **39** (1), 305–332, 2011.

[12] Li, G.R., Feng, S.Y. and Peng, H. *Profile-type smoothed score function for a varying coefficient partially linear model*, Journal of Multivariate Analysis **102** (2), 372–385, 2011.

[13] Li, G.R., Lin, L. and Zhu, L.X. *Empirical likelihood for varying coefficient partially linear model with diverging number of parameters*, Journal of Multivariate Analysis **105** (1), 85–111, 2012.

[14] Li, R.Z. and Liang, H. *Variable selection in semiparametric regression modeling*, The Annals of Statistics **36** (1), 261–286, 2008.

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

[16] Ruppert, D., Wand, M. and Carroll, R. *Semiparametric Regression.* Cambridge University Press, 2003.

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

[18] Wang, K. and Lin, L. *Simultaneous structure estimation and variable selection in partial linear varying coefficient models for longitudinal data*, Journal of Statistical Computation and Simulation **85** (7), 1459–1473, 2015.

[19] Wang, X. L., Li, G. R. and Lin, L. *Empirical likelihood inference for semiparametric varying-coefficient partially linear EV models*, Metrika **73** (2), 171–185, 2011.

[20] Wei, C. H. *Statistical inference for restricted partially linear varying coefficient errors-in-variables models*, Journal of Statistical Planning and Inference **142** (8), 2464–2472, 2012.

[21] You, J. H. and Chen, G. M. *Estimation of a semiparametric varying-coefficient partially linear errors-in-variables model*, Journal of Multivariate Analysis **97** (2), 324–341, 2006.

[22] Zhang, C.H. *Nearly unbiased variable selection under minimax concave penalty*, The Annals of Statistics **38** (2), 894–942, 2010.

[23] Zhang, W., Lee, S., Y. and Song, X. *Local polynomial fitting in semivarying coefficient models*, Journal of Multivariate Analysis **82** (1), 166–188, 2002.

[24] Zhao, P.X. and Xue, L.G. *Variable selection for semiparametric varying coefficient partially linear models*, Statistics and Probability Letters **79** (20), 2148–2157, 2009.

[25] Zhao, P.X. and Xue, L.G. *Variable selection for semiparametric varying coefficient partially linear errors-in-variables models*, Journal of Multivariate Analysis **101** (8), 1872–1883, 2010.

[26] Zhou, Y. and Liang, H. *Statistical inference for semiparametric varying-coefficient partially linear models with generated regressors*, The Annals of Statistics **37** (1), 427–458, 2009.

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

[28] Zou, H. and Hastie, T. *Regularization and variable selection via the elastic net*, Journal of the Royal Statistical Society: Series B (Statistical Methodology) **67** (2), 301–320, 2005.