Chin. Ann. Math.
29B(2), 2008, 207–222
DOI: 10.1007/s11401-006-0210-8

Chinese Annals of Mathematics, Series B

Statistical Inference for Partially Linear Regression Models with Measurement Errors

Jinhong YOU[*] Qinfeng XU[**] Bin ZHOU[***]

Abstract In this paper, the authors investigate three aspects of statistical inference for the partially linear regression models where some covariates are measured with errors. Firstly, a bandwidth selection procedure is proposed, which is a combination of the difference-based technique and GCV method. Secondly, a goodness-of-fit test procedure is proposed, which is an extension of the generalized likelihood technique. Thirdly, a variable selection procedure for the parametric part is provided based on the nonconcave penalization and corrected profile least squares. Same as "Variable selection via nonconcave penalized likelihood and its oracle properties" (J. Amer. Statist. Assoc., **96**, 2001, 1348–1360), it is shown that the resulting estimator has an oracle property with a proper choice of regularization parameters and penalty function. Simulation studies are conducted to illustrate the finite sample performances of the proposed procedures.

Keywords Partially linear model, Measurement error, Bandwidth selection, Goodness-of-fit test, Oracle property
2000 MR Subject Classification 62G08, 62J12

1 Introduction

Parametric regression provides powerful tools for analyzing practical data when the models are correctly specified, but may suffer from large modeling biases if the structures of models are misspecified. As an alternative, nonparametric smoothing eases the concerns on modeling biases. However, the nonparametric method is hampered by the so-called "curse of dimensionality" in multivariate settings (see [14, 22] among others). One of the methods for attenuating this difficulty is to model covariate effects via a partially linear structure, a combination of linear and nonparametric parts. This results in the partially linear regression models (see [12]). In general, a partially linear regression model can be written as

$$Y = X^\tau \boldsymbol{\beta} + g(U) + \varepsilon, \tag{1.1}$$

where Y is the response, both X and U are possibly vector-valued covariates, ε is a random error independent of (X, U) with $E(\varepsilon) = 0$ and $\mathrm{Var}(\varepsilon) = \sigma^2$, $\boldsymbol{\beta}$ is an unknown parameter vector having the same dimension as X, $g(\,\cdot\,)$ is an unknown smooth function, and the superscript $(^\tau)$ denotes the transpose of a vector or a matrix.

Manuscript received May 19, 2006. Revised May 27, 2007. Published online January 28, 2008.

[*] Department of Biostatistics, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599-7400, USA.
[**] Corresponding author. School of Mathematical Sciences, Fudan University, Shanghai 200433, China. E-mail: qfxu@fudan.edu.cn
[***] Department of Statistics, East China Normal University, Shanghai 200062, China.

The partially linear regression model has broad applicability in the fields of biology, economics, education and social sciences. This model and various associated estimators, test statistics, and generalizations have generated a substantial body of literature, which include [6–8, 11, 13, 21, 23, 28–31]. To name just a few.

In many practical situations, however, there often exist covariate measurement errors. For example, it has been well documented in the literature that covariates such as blood pressure, urinary sodium chloride level and exposure to pollutants are often subject to measure error. Some work has been done in the estimation of regression coefficients of the partially linear model (1.1) in the presence of additive measurement errors in the predictors. For example, Cui and Li [10] and Liang, Härdle and Carroll [27] discussed the estimation problem when the covariates are measured with additive errors by the nearest neighbor estimation and general kernel smoothing for the nonparametric component, respectively. Liang [26] discussed estimation of the partially linear model with measurement errors in the nonparametric part.

However, according to the knowledge of the authors, there is no article investigating statistical inference beyond point estimation for partially linear regression models with measurement errors, whereas the statistical inference theory has been well developed for partially linear regression models without measurement errors. Actually, even for classical regression models with measurement errors, there have not yet been many articles investigating statistical inference except recent papers by Cheng and Tsai [9] and You and Xu [33]. Cheng and Tsai [9] investigated the invariance property of score tests for assessing heteroscedasticity, first-order autoregressive disturbance, and the need for a Box-Cox per transformation in the context of linear regression models with additive measurement errors. They showed that the score tests for measurement error models are identical to the corresponding well-established tests derived from the standard linear regression models. You and Xu [33] provided a procedure to select the significant covariates of the linear regression models in which some or all covariates are measured with errors. The proposed method is based on the combination of a nonconcave penalization and a corrected least squares, and it simultaneously selects significant covariates and estimates the unknown regression coefficients.

The objective of the present paper is to fill this gap. In this paper, same as [27], we consider the case that the covariate vector X is measured with additive errors, and U is error free, i.e., we can not observe X but W where

$$W = X + \zeta, \tag{1.2}$$

and ζ is the measurement error vector. We assume that ζ is independent of (X, U, ε), $E(\zeta) = \mathbf{0}$ and $\mathrm{Cov}(\zeta) = \mathbf{\Sigma}_\zeta$ where $\mathbf{\Sigma}_\zeta$ is assumed known, same as in [24, 34] among others. However, we will take up the case that it is estimated in Section 6.

Due to the curse of dimensionality, for simplicity, we assume that U is univariate throughout this paper. Suppose that the dimension of X is p, and $\{(Y_i, W_i, U_i)\}_{i=1}^n$ is a random sample from model (1.1) with measurement errors (1.2).

The contribution of this paper is three fold. We first propose a simple bandwidth selection procedure which is based on the combination of the difference technique and GCV method.

After fitting model (1.1), one often asks if there exists a parametric structure for $g(\cdot)$. This amounts to testing if $g(\cdot)$ is in a certain parametric form. However, for such a frequently-asked question, there are limited tools available for model (1.1) with measurement errors (1.2). We propose a goodness-of-fit test procedure, which is an extension of the generalized likelihood technique proposed by Fan, Zhang and Zhang [19] to the setting of partially linear measurement error model. The bootstrap method is used to evaluate the p-value of the test.

Like traditional parametric regression, covariate selection is also important in the semiparametric model (1.1) with measurement errors (1.2). To reduce possible modeling biases, the nonlinear terms and interactions between covariates are often introduced. This makes the number of covariates in the parametric part of semiparametric model (1.1) easily be large. It is common in practice to include only important variables in the model to enhance predictability and to give a parsimonious description between the response and the covariates. Due to the complexity caused by measurement errors, the well-developed stepwise deletion and best variable selection can not be extended to semiparametric model (1.1). Recently, Fan and Li [15] proposed a covariate selection method via nonconcave penalized likelihood. This method deletes insignificant covariates by estimating their coefficients as 0 and simultaneously selects significant covariates and estimates regression coefficients. From their simulations, Fan and Li [15] showed that the penalized likelihood estimator with smoothly clipped absolute deviation (SCAD) penalty outperforms the best subset variable selection in terms of computational cost and stability using the terminology of [3]. In addition, they have demonstrated that with a proper choice of regularization parameters and penalty functions (such as SCAD), the penalized likelihood estimator possesses an oracle property. Namely, the true regression coefficients that are zero are automatically estimated as zero, and the remaining coefficients are estimated as well as if the correct submodel is known in advance. Hence, the SCAD and its siblings are ideal for variable selection, at least from the theoretical point of view. Fan and Li [16, 17], Cai et. al. [4] and Fan and Peng [18] extended their nonconcave penalized likelihood approach to the Cox model, frailty model, multivariate Cox model, longitudinal partially linear model and regression model with infinite parameters. These nice properties encourage us to extend the technique to model (1.1) with measurement errors (1.2).

The layout of the remainder of this paper is as follows. In Section 2, we present the corrected profile least squares estimation proposed by Liang, Härdle and Carroll [27]. A bandwidth selection is described in Section 3. A bootstrap based test for the goodness of fit of models is developed in Section 4. A model selection procedure is shown in Section 5. Section 6 states the corresponding results when the measurement error variance $\mathbf{\Sigma}_\zeta$ is estimated. Simulations are conducted in Section 7. Section 8 concludes. The proofs of the main results are collected in Appendix.

2 Corrected Profile Least Squares Estimation

A corrected profile least squares estimation proposed by Liang, Härdle and Carroll [27] has the following form

$$\widehat{\boldsymbol{\beta}}_n=\Big(\sum_{i=1}^n\widehat{W}_i\widehat{W}_i^\tau-n\mathbf{\Sigma}_\zeta\Big)^{-1}\sum_{i=1}^n\widehat{W}_i\widehat{Y}_i,$$

where $\widehat{\mathbf{Y}}=(\widehat{Y}_1,\cdots,\widehat{Y}_n)^\tau=(\mathbf{I}_n-\mathbf{S})\mathbf{Y}$, $\widehat{\mathbf{W}}=(\widehat{W}_1,\cdots,\widehat{W}_n)^\tau=(\mathbf{I}_n-\mathbf{S})\mathbf{W}$,

$$\mathbf{S}=\begin{pmatrix}(1\quad 0)(\mathbf{D}_{U_1}^\tau\boldsymbol{\omega}_{U_1}\mathbf{D}_{U_1})^{-1}\mathbf{D}_{U_1}^\tau\boldsymbol{\omega}_{U_1}\\ \vdots\\ (1\quad 0)(\mathbf{D}_{U_n}^\tau\boldsymbol{\omega}_{U_n}\mathbf{D}_{U_n})^{-1}\mathbf{D}_{U_n}^\tau\boldsymbol{\omega}_{U_n}\end{pmatrix},\quad \mathbf{D}_u=\begin{pmatrix}1 & \frac{U_1-u}{h}\\ \vdots & \vdots\\ 1 & \frac{U_n-u}{h}\end{pmatrix},$$

$\boldsymbol{\omega}_u=\mathrm{diag}(K_h(U_1-u),\cdots,K_h(U_n-u))$, $K(\cdot)$ is a kernel function, h is a bandwidth, $K_h(\cdot)=\frac{K(\frac{\cdot}{h})}{h}$, $\mathbf{Y}=(Y_1,\cdots,Y_n)^\tau$, and $\mathbf{W}=(W_1,\cdots,W_n)^\tau$. Moreover, the fact $g(U_i)=E(Y_i-$

$X_i^\tau\boldsymbol{\beta}\,|\,U_i) = E(Y_i - W_i^\tau\boldsymbol{\beta}\,|\,U_i)$ suggests one to estimate the nonparametric component $g(\,\cdot\,)$ by

$$\widehat{g}_n(u) = (1\ \ 0)(\mathbf{D}_u^\tau\boldsymbol{\omega}_u\mathbf{D}_u)^{-1}\mathbf{D}_u^\tau\boldsymbol{\omega}_u(\mathbf{Y} - \mathbf{W}\widehat{\boldsymbol{\beta}}_n).$$

The following assumptions are needed to present the asymptotic properties of $\widehat{\boldsymbol{\beta}}_n$ and $\widehat{g}_n(u)$, and other results developed in the subsequent sections.

Assumption 2.1 *The random variable U has a bounded support Ω. Its density function $f(\,\cdot\,)$ is Lipschitz continuous and bounded away from* 0 *on its support.*

Assumption 2.2 *There is an $s > 2$ such that $E\|X\|^{2s} < \infty$ and $E\|\zeta\|^{2s} < \infty$ and for some $\delta < 2 - s^{-1}$, such that $n^{2\delta-1}h \to \infty$ as $n \to \infty$.*

Assumption 2.3 *$g(\,\cdot\,)$ has the continuous second derivative in Ω.*

Assumption 2.4 *The function $K(\,\cdot\,)$ is a symmetric density function with compact support and the bandwidth satisfies $\frac{nh^4}{(\log\log n)^{1/2}} \to 0$ and $\frac{nh^2}{(\log n)^2\to\infty}$ as $n \to \infty$.*

The following theorem gives the asymptotic normality of $\widehat{\boldsymbol{\beta}}_n$.

Theorem 2.1 *Suppose that Assumptions* 2.1–2.4 *hold. Then the corrected profile least squares estimator $\widehat{\boldsymbol{\beta}}_n$ of $\boldsymbol{\beta}$ is asymptotically normal, namely*

$$\sqrt{n}(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}) \to_D N(0, \boldsymbol{\Sigma}_1^{-1}\boldsymbol{\Sigma}_2\boldsymbol{\Sigma}_1^{-1}), \quad \textit{as } n \to \infty,$$

where $\boldsymbol{\Sigma}_1 = E\{X - E(X\,|\,U)\}\{X - E(X\,|\,U)\}^\tau$ and

$$\boldsymbol{\Sigma}_2 = E(\varepsilon - \zeta^\tau\boldsymbol{\beta})^2\boldsymbol{\Sigma}_1 + \sigma^2\boldsymbol{\Sigma}_\zeta + E\{(\zeta\zeta^\tau - \boldsymbol{\Sigma}_\zeta)\boldsymbol{\beta}\boldsymbol{\beta}^\tau(\zeta\zeta^\tau - \boldsymbol{\Sigma}_\zeta)\}.$$

The following theorem gives the asymptotic normality of the estimator $\widehat{g}_n(\,\cdot\,)$.

Theorem 2.2 *Suppose that Assumptions* 2.1–2.4 *hold. Then the local linear estimator $\widehat{g}_n(\,\cdot\,)$ of $g(\,\cdot\,)$ is asymptotically normal, namely*

$$\sqrt{nh}\Big\{\widehat{g}_n(u_0) - g(u_0) - \frac{h^2}{2}\frac{\mu_2^2 - \mu_1\mu_3}{\mu_2 - \mu_1^2}g''(u_0)\Big\} \to_D N(0, \boldsymbol{\Sigma}_3), \quad \textit{as } n \to \infty,$$

where

$$\boldsymbol{\Sigma}_3 = \frac{(c_0^2\nu_0 + 2c_0c_1\nu_1 + c_1^2\nu_2)}{f(u_0)}(\sigma^2 + \boldsymbol{\beta}^\tau\boldsymbol{\Sigma}_\zeta\boldsymbol{\beta})$$

with $c_0 = \frac{\mu_2}{(\mu_2-\mu_1^2)}$, $c_1 = -\frac{\mu_1}{(\mu_2-\mu_1^2)}$, $\mu_j = \int_{-\infty}^{\infty} u^jK(u)\mathrm{d}u$, $\nu_j = \int_{-\infty}^{\infty} u^jK^2(u)\mathrm{d}u$.

It should be noted that Liang, Härdle and Carroll [27] just presented the convergence rate of $\widehat{g}_n(\,\cdot\,)$, no asymptotic normality.

3 Bandwidth Selection Procedure

The corrected profile least square estimators $\widehat{\boldsymbol{\beta}}_n$ and local linear estimator $\widehat{g}_n(\,\cdot\,)$ depend on the choice of bandwidth. Furthermore, the issue of bandwidth selection arises naturally in practice. Selecting bandwidths for semiparametric models, particularly for estimating the parametric component, was posed by Bickel and Kwon [2] as an important and unsolved problem. As discussed by Fan in [2], the estimation of the parametric component does not very sensitively depend on the choice of bandwidth, as long as the selected bandwidth does not create excessive biases in the estimation of the nonparametric components. The reason is that the biases in the estimation of nonparametric components can not be averaged out in the process of estimating the parametric component, yet the variance in nonparametric estimation can be averaged out. Hence, we choose a bandwidth that is suitable for estimating function $g(\cdot)$.

We propose to use the following technique to determine an appropriate value of the smoothing parameter h. We first construct a root-n consistent estimator $\widehat{\boldsymbol{\beta}}_n^\star$ of $\boldsymbol{\beta}$ which does not involve h, and then, based on the modified data set $\{Y_i - W_i^\tau \widehat{\boldsymbol{\beta}}_n^\star, U_i\}_{i=1}^n$, select an appropriate value of h by CV method.

Let the sample $\{(Y_i, W_i, U_i)\}_{i=1}^n$ be ordered according to variable U. Under some mild conditions, the spacing between $U_{i+1} - U_i$ is $O_p(\frac{1}{n})$ so that $g(U_{i+1}) - g(U_i) = O_p(\frac{1}{n})$. Then by model (1.1),

$$\begin{aligned} Y_{i+1} - Y_i &= (X_{i+1,1} - X_{i,1})\beta_1 + \cdots + (X_{i+1,p} - X_{i,p})\beta_p + g(U_{i+1}) - g(U_i) + \varepsilon_{i+1} - \varepsilon_i \\ &= (X_{i+1,1} - X_{i,1})\beta_1 + \cdots + (X_{i+1,p} - X_{i,p})\beta_p + \varepsilon_{i+1} - \varepsilon_i + O_p\Big(\frac{1}{n}\Big), \end{aligned}$$

where $X_i = (X_{i,1}, \cdots, X_{i,p})^\tau$, and $\boldsymbol{\beta} = (\beta_1, \cdots, \beta_p)^\tau$. Let $\varepsilon_i^\star$'s be corrected stochastic errors with $\varepsilon_i^\star = \varepsilon_{i+1} - \varepsilon_i$. Thus, the nonparametric function $g(\cdot)$ in model (1.1) is eliminated. The coefficient $\boldsymbol{\beta}$ can be estimated by ordinary least-squares from the above approximation model. That is

$$\widetilde{\boldsymbol{\beta}}_n^\star = \Big\{ \sum_{i=1}^{n-1} (X_{i+1} - X_i)(X_{i+1} - X_i)^\tau \Big\}^{-1} \sum_{i=1}^{n-1} (X_{i+1} - X_i)(Y_{i+1} - Y_i).$$

However, in our case X_i can not be observed. Therefore, we propose the following corrected estimator

$$\widehat{\boldsymbol{\beta}}_n^\star = \Big\{ \sum_{i=1}^{n-1} (W_{i+1} - W_i)(W_{i+1} - W_i)^\tau - 2(n-1)\boldsymbol{\Sigma}_\zeta \Big\}^{-1} \sum_{i=1}^{n-1} (W_{i+1} - W_i)(Y_{i+1} - Y_i).$$

Under some regularity conditions, we can show $\widehat{\boldsymbol{\beta}}_n^\star$ is root-n consistent.

Based on $\widehat{\boldsymbol{\beta}}_n^\star$, we can get the modified data set $\{Y_i - W_i^\tau \widehat{\boldsymbol{\beta}}_n^\star, U_i\}_{i=1}^n$. Then the usual cross-validation method can be used. Define the squares cross-validation function by

$$\mathrm{CV}(h) = n^{-1} \sum_{i=1}^{n} \{Y_i - W_i^\tau \widehat{\boldsymbol{\beta}}_n^\star - \widehat{g}_{h,-i}(U_i)\}^2, \tag{3.1}$$

where $\widehat{g}_{h,-i}(\cdot)$ is the local linear estimate from the data $\{Y_i - W_i^\tau \widehat{\boldsymbol{\beta}}_n^\star, U_i\}_{i=1}^n$ omitting the ith point $(Y_i - W_i^\tau \widehat{\boldsymbol{\beta}}_n^\star, U_i)$. Depending on the smoothing parameter h, formula (3.1) is used as an overall measure of effectiveness of the estimation scheme. The cross-validation bandwidth selector is the one that minimizes (3.1), namely $\widehat{h}_{\mathrm{CV}} = \arg\min_h \mathrm{CV}(h)$.

4 Goodness of Fit Test Based on Bootstrap

To test whether model (1.1) holds with a specified parametric form such as a linear model, we propose a goodness-of-fit test by comparing the pseudo residual sums of squares (PRSS) between parametric and semiparametric fittings. This method is an extension of the generalized likelihood technique developed by Fan, Zhang and Zhang [19] to the partially linear regression models with measurement errors.

Consider the null hypothesis

$$H_0 : g(u) = a(u, \boldsymbol{\theta}), \tag{4.1}$$

where $a(\,\cdot\,, \boldsymbol{\theta})$ is a given family of function indexed by unknown parameter vector $\boldsymbol{\theta}$. Let $\widehat{\boldsymbol{\theta}}_n$ be an estimator of $\boldsymbol{\theta}$. The pseudo residual sum of squares under the null hypothesis is

$$\mathrm{PRSS}_0 = n^{-1} \sum_{i=1}^{n} \{Y_i - W_i^{\tau} \widehat{\boldsymbol{\beta}}_n - a(U_i, \widehat{\boldsymbol{\theta}}_n)\}^2.$$

Analogously, the pseudo residual sum of squares corresponding to model (1.1) is

$$\mathrm{PRSS}_1 = n^{-1} \sum_{i=1}^{n} \{Y_i - W_i^{\tau} \widehat{\boldsymbol{\beta}}_n - \widehat{g}_n(U_i)\}^2.$$

The test statistic is defined as

$$T_n = \frac{\mathrm{PRSS}_0 - \mathrm{PRSS}_1}{\mathrm{PRSS}_1} = \frac{\mathrm{PRSS}_0}{\mathrm{PRSS}_1} - 1.$$

We have the following theorem for T_n.

Theorem 4.1 *Under the null hypothesis* (4.1) *and Assumptions* 2.1–2.4*, if $h \to 0$ in such a way that $nh^{3/2} \to \infty$, then the test statistic*

$$r_K T_n \to_D \chi^2_{\delta_n}, \quad \textit{as } n \to \infty,$$

where

$$r_K = \Big(K(0) - \frac{1}{2}\int K^2(u)\mathrm{d}u\Big)\Big\{\int \Big(K(u) - \frac{1}{2}K * K(u)\Big)^2 \mathrm{d}u\Big\}^{-1},$$
$$\delta_n = r_K \frac{|\mathcal{U}|}{h}\Big(K(0) - \frac{1}{2}\int K^2(u)\mathrm{d}u\Big),$$

$K * K$ denotes the convolution of K and $|\mathcal{U}|$ is the length of the support Ω of U.

We reject the null hypothesis (4.1) for large values of T_n. The following bootstrap approach is used to evaluate p-value of the test.

(1) By fitting the model, we estimate the pseudo residuals by

$$\widehat{\varepsilon}_i = Y_i - W_i^{\tau} \widehat{\boldsymbol{\beta}}_n - \widehat{g}_n(U_i), \quad i = 1, \cdots, n.$$

(2) Generate the bootstrap residuals $\{\varepsilon_i^{\star}\}_{i=1}^n$ from the empirical distribution of the centralized residuals $\{\widehat{\varepsilon}_i - \overline{\varepsilon}\}_{i=1}^n$ where $\overline{\varepsilon} = n^{-1} \sum\limits_{i=1}^{n} \widehat{\varepsilon}_i$. Define

$$Y_i^{\star} = W_i^{\tau} \widehat{\boldsymbol{\beta}}_n + \widehat{g}_n(U_i) + \varepsilon_i^{\star} \quad \text{for } i = 1, \cdots, n.$$

(3) Calculate the bootstrap test statistic $T_n^{\star}$ based on the sample $\{U_i, W_i, Y_i^{\star}\}_{i=1}^n$.

(4) Reject the null hypothesis H_0 when T_n is greater than the upper-α point of the conditional distribution of $T_n^{\star}$ given by $\{U_i, W_i, Y_i\}_{i=1}^n$.

The p-value of the test is simply the relative frequency of the event $\{T_n^\star \geq T_n\}$ in the replications of the bootstrap sampling. For the sake of simplicity, we use the same bandwidth in calculating $T_n^\star$ as that in T_n. Note that we bootstrap the centralized residuals from the semiparametric fit instead of the parametric fit, because the semiparametric estimator of the residuals is always consistent, no matter the null or the alternative hypothesis is correct. The method should provide a consistent estimator of the null distribution even when the null hypothesis does not hold.

5 Covariate Selection

Model selection is an indispensable tool for statistical data analysis. However, the problem has not been studied in the semiparametric regression model with measurement errors. Fan and Li [15] proposed a variable selection method via nonconcave penalized likelihood and found some oracle properties. These nice properties encourage us to extend the technique to the partially linear regression model (1.1) with measured errors (1.2). It gives us a quick and effective method for eliminating unimportant variables. We here propose a nonconcave penalized corrected profile least squares procedure, which is described as follows.

5.1 Penalized corrected profile least squares

Suppose that $\boldsymbol{\beta}$ consists of p components, and some of these are not significant. A penalized corrected profile least squares takes the form

$$\mathcal{L}(\boldsymbol{\beta}) \equiv \ell(\boldsymbol{\beta}) + n\sum_{s=1}^{p} \lambda_s p_s(|\beta_s|),$$

where the $p_s(\,\cdot\,)$'s are penalty functions, λ_s's are tuning parameters which control the model complexity and can be selected by some data-driven methods, such as cross-validation or generalized cross validation, and

$$\ell(\beta) = \frac{1}{2}\left(\mathbf{Y} - \mathbf{W}\boldsymbol{\beta} - G\right)^\tau \left(\mathbf{Y} - \mathbf{W}\boldsymbol{\beta} - G\right) - \frac{n}{2}\boldsymbol{\beta}^\tau \boldsymbol{\Sigma}_\zeta \boldsymbol{\beta},$$

where $G = (g(U_1), \cdots, g(U_n))^\tau$, $\mathbf{Y}$ and $\mathbf{W}$ are defined in Section 2. Here the penalty functions $p_s(\,\cdot\,)$ and the regularization parameters λ_s are not necessarily the same for all $s = 1, \cdots, p$. This allows us to incorporate prior information for the unknown coefficients by using different penalty functions or taking different values of λ_s. For instance, we may wish to keep important predictors in the parametric part of model (1.1) and hence do not want to penalize their coefficients. For ease of presentation, we denote $\lambda_s p_s(\,\cdot\,)$ by $p_{\lambda_s}(\,\cdot\,)$.

After eliminating the nuisance function $g(\,\cdot\,)$ by the profile techniques, we obtain the following penalized corrected profile least squares:

$$\mathcal{L}(\boldsymbol{\beta}) = \frac{1}{2}(\widehat{\mathbf{Y}} - \widehat{\mathbf{W}}\boldsymbol{\beta})^\tau (\widehat{\mathbf{Y}} - \widehat{\mathbf{W}}\boldsymbol{\beta}) - \frac{n}{2}\boldsymbol{\beta}^\tau \boldsymbol{\Sigma}_\zeta \boldsymbol{\beta} + n\sum_{s=1}^{p} \lambda_s p_s(|\beta_{js}|), \tag{5.1}$$

where $\widehat{\mathbf{Y}}$ and $\widehat{\mathbf{W}}$ are defined in Section 2.

Many penalty functions, such as the family of L_q-penalty ($q \geq 0$), have been used for penalized least squares and penalized likelihood in various parametric models. For instance, $q = 0$ corresponds to the entropy penalty, L_1 penalty results in the LASSO, proposed by

Tibshirani [32], and bridge regression (see [20]) corresponds to $0 < q < 1$. Antoniadis and Fan [1] and Fan and Li [15] provided various insights into how a penalty function should be chosen. They advocate that a good penalty function should yield an estimator with the following three properties: unbiasedness for a large true coefficient to avoid unnecessary estimation bias, sparsity (estimating a small coefficient as zero) to reduce model complexity, and continuity to avoid unnecessary variation in model prediction. Necessary conditions are given in [1]. None of the L_q penalties produce any estimator satisfying simultaneously the above three properties. According to [15], a simple penalty function, which results in an estimator with the three desired properties, is the smoothly clipped absolute deviation (SCAD) penalty. Its first derivative is defined by

$$p'_\lambda(\beta) = \lambda\Big\{I(\beta \le \lambda) + \frac{(a\lambda - \beta)_+}{(a-1)\lambda} I(\beta > \lambda)\Big\} \quad \text{for some } a > 2 \text{ and } \beta > 0,$$

and $p_\lambda(0) = 0$. The SCAD involves two unknown parameters, λ and a. Fan and Li [15] suggested using $a = 3.7$ from a Bayesian point of view. Hence, this value will be used throughout the rest of the paper.

5.2 Asymptotic properties

Now, we study the asymptotic properties of the resulting estimator of the penalized corrected profile least squares (5.1). First, we establish the convergence rate of the penalized corrected profile least squares estimator. Assume that all penalty functions $p_{\lambda_s}(\cdot)$ are negative, nondecreasing with $p_{\lambda_s}(0) = 0$. Denote by $\boldsymbol{\beta}_0$ the true value of $\boldsymbol{\beta}$, and

$$a_n = \max_s\{|p'_{\lambda_s}(|\beta_{0s}|)| : \beta_{0s} \neq 0\}, \quad b_n = \max_s\{|p''_{\lambda_s}(|\beta_{0s}|)| : \beta_{0s} \neq 0\}.$$

Then, we have the following theorem.

Theorem 5.1 *Suppose that Assumptions 2.1–2.4 hold. If a_n and b_n tend to zero as $n \to \infty$, then with probability tending to one, there exists a local minimizer $\widetilde{\boldsymbol{\beta}}_n$ of $\mathcal{L}(\boldsymbol{\beta})$ such that $\|\widetilde{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0\| = O_p(n^{-1/2} + a_n)$.*

Theorem 5.1 demonstrates how the rate of convergence of the penalized corrected profile least squares estimator $\widetilde{\boldsymbol{\beta}}_n$ depends on λ_s. To achieve the root-n convergence rate, we have to take λ_s small enough so that $a_n = O(n^{-1/2})$. Next, we establish the oracle property for the penalized corrected profile least squares estimator $\widetilde{\boldsymbol{\beta}}_n$. For ease of presentation, we assume, without loss of generality, that all of the first q components of $\boldsymbol{\beta}_0$ are not equal to 0, and all other $p - q$ components are equal to 0.

Let

$$\mathbf{B} = \mathrm{diag}\{p''_{\lambda_1}(|\beta_{01}|), \cdots, p''_{\lambda_q}(|\beta_{0q}|)\} \text{ and } \mathbf{b} = (p'_{\lambda_1}(|\beta_{01}|)\mathrm{sgn}(\beta_{01}), \cdots, p'_{\lambda_q}(|\beta_{0q}|)\mathrm{sgn}(\beta_{0q}))^\tau.$$

Further, let $\widetilde{\boldsymbol{\beta}}_{n1}$ consist of the first q components of $\widetilde{\boldsymbol{\beta}}_n$ and $\widetilde{\boldsymbol{\beta}}_{n2}$ consist of the last $p - q$ ones.

Theorem 5.2 (Oracle Property) *Assume that $\lambda_s \to 0$ and $\sqrt{n}\,\lambda_s \to \infty$ as $n \to \infty$ for $s = 1, \cdots, p$, and the penalty function $p_{\lambda_s}(|\beta_s|)$ satisfies that*

$$\liminf_{n\to\infty} \liminf_{\beta_s \to 0+} \frac{p_{\lambda_s}(\beta_s)}{\lambda_s} > 0. \tag{5.2}$$

If $a_n = O(n^{-1/2})$, then under the conditions of Theorem 5.1, *with probability tending to* 1, *the root-n consistent local minimizer $\widetilde{\boldsymbol{\beta}}_n = (\widetilde{\boldsymbol{\beta}}_{n1}^{\tau}, \widetilde{\boldsymbol{\beta}}_{n2}^{\tau})^{\tau}$ in Theorem* 5.1 *must satisfy*

(i) (Sparsity) $\widetilde{\boldsymbol{\beta}}_{n2} = 0$;

(ii) (Asymptotic Normality)

$$\sqrt{n}\,(\boldsymbol{\Sigma}_1^{(1)} + \mathbf{B})\{\widetilde{\boldsymbol{\beta}}_{n1} - \boldsymbol{\beta}_{01} + (\boldsymbol{\Sigma}_1^{(1)} + \mathbf{B})^{-1}\mathbf{b}\} \rightarrow_D N_q(0, \boldsymbol{\Sigma}_2^{(1)}),$$

where $\boldsymbol{\Sigma}_1^{(1)}$ and $\boldsymbol{\Sigma}_2^{(1)}$ consist of the first q rows and columns of $\boldsymbol{\Sigma}_1$ and $\boldsymbol{\Sigma}_2$, respectively, and $\boldsymbol{\Sigma}_1$ and $\boldsymbol{\Sigma}_2$ are defined in Theorem 2.1.

From Theorem 5.2, if $\lambda_s \to 0$, $\sqrt{n}\lambda_s \to \infty$ for $s = 1, \cdots, p$, $a_n = O(n^{-1/2})$, and condition (5.2) is satisfied, then the resulting estimator possesses an oracle property. This implies that the procedure correctly specifies the true model and estimates the unknown regression coefficients as efficiently as we knew the submodel. If all the penalty functions are SCAD, then a_n tends to 0 as $n \to \infty$, and hence the resulting estimator possesses the oracle property.

5.3 Choice of regularization parameters

It is challenging to find the solution of the penalized corrected profile least squares of (5.1) because the penalty function $p_{\lambda_s}(|\beta_s|)$, such as the L_q penalty $(0 < q \le 1)$ and the SCAD penalty, is irregular at the origin and may not have a second derivative at some points. Following [15], we can locally approximate the penalty functions by quadratic functions as follows. Given an initial value $\boldsymbol{\beta}^{(0)}$ that is close to the minimizer of (5.1). If $\beta_s^{(0)}$ is very close to 0 (for instance, $|\beta_s^{(0)}|$ less than a prescribed value η), then set $\widehat{\beta}_s = 0$. Otherwise, the penalty $p_{\lambda_s}(|\beta_s|)$ can be locally approximated by a quadratic function as

$$[p_{\lambda_s}(|\beta_s|)]' = p'_{\lambda_s}(|\beta_s|)\mathrm{sgn}(\beta_s) \approx \left\{\frac{p'_{\lambda_s}(|\beta_s^{(0)}|)}{|\beta_s^{(0)}|}\right\}\beta_s.$$

The further details can be found in [15]. With the local quadratic approximation, the Newton-Raphson algorithm can be implemented directly for minimizing $\mathcal{L}(\boldsymbol{\beta})$.

To implement the method described in the previous sections, it is desirable to have an automatic data-driven method for estimating the tuning parameters $\lambda_1, \cdots, \lambda_p$. Similarly to [15], we can estimate $(\lambda_1, \cdots, \lambda_p)$ by minimizing an approximate generalized cross-validation score.

6 Estimated Errors Variance

Although in some cases the measurement error covariance matrix $\boldsymbol{\Sigma}_\zeta$ has been established by independent experiments, in others it is unknown and must be estimated. According to [5, Chapter 3], the usual method for doing so is by partial replication, so that we observe $W_{ij} = X_i + \zeta_{ij}$, $j = 1, \cdots, m_i$.

For notational convenience, same as [27], we consider here only the case that $m_i \le 2$ and assume that a fraction δ of the data has such replicates. Let $\overline{W}_i$ and $\overline{Y}_i$ be the corresponding sample means of the replicates. Then a consistent, unbiased moment estimator for $\boldsymbol{\Sigma}_\zeta$, is

$$\widehat{\boldsymbol{\Sigma}}_\zeta = \frac{\sum\limits_{i=1}^{n}\sum\limits_{j=1}^{m_i}(W_{ij} - \overline{W}_i)(W_{ij} - \overline{W}_i)^{\tau}}{\sum\limits_{i=1}^{n}(m_i - 1)}.$$

Correspondingly, the corrected profile least squares estimator is

$$\check{\boldsymbol{\beta}}_n = \Big\{ \sum_{i=1}^n \widehat{\overline{W}}_i \widehat{\overline{W}}_i^\tau - n(1-\delta/2)\widehat{\boldsymbol{\Sigma}}_\zeta \Big\}^{-1} \sum_{i=1}^n \widehat{\overline{W}}_i \widehat{\overline{Y}}_i,$$

where $(\widehat{\overline{Y}}_1, \cdots, \widehat{\overline{Y}}_n)^\tau = (\mathbf{I}_n - \mathbf{S})(\overline{Y}_1, \cdots, \overline{Y}_n)^\tau$, and $(\widehat{\overline{W}}_1, \cdots, \widehat{\overline{W}}_n)^\tau = (\mathbf{I}_n - \mathbf{S})(\overline{W}_1, \cdots, \overline{W}_n)^\tau$. According to [27], $\check{\boldsymbol{\beta}}_n$ is root-n consistent and asymptotically normal with asymptotic covariance matrix $\boldsymbol{\Sigma}_1^{-1}\boldsymbol{\Gamma}\boldsymbol{\Sigma}_1^{-1}$, where

$$\begin{aligned}\boldsymbol{\Gamma} = {}& (1-\delta)E\{(\varepsilon - \zeta^\tau\boldsymbol{\beta})(X - E(X\,|\,U))\}^{\otimes 2} + \delta E\{(\varepsilon - \overline{\zeta}^\tau\boldsymbol{\beta})(X - E(X\,|\,U))\}^{\otimes 2} \\ & + (1-\delta)E\Big[\Big\{\Big(\zeta\zeta^\tau - \Big(1 - \frac{\delta}{2}\Big)\boldsymbol{\Sigma}_\zeta\Big)\boldsymbol{\beta}\Big\}^{\otimes 2} + \zeta\zeta^\tau\varepsilon^2\Big] \\ & + \delta E\Big[\Big\{\Big(\overline{\zeta\zeta}^\tau - \Big(1 - \frac{\delta}{2}\Big)\boldsymbol{\Sigma}_\zeta\Big)\boldsymbol{\beta}\Big\}^{\otimes 2} + \overline{\zeta\zeta}^\tau\varepsilon^2\Big],\end{aligned}$$

$A^{\otimes 2} = AA^\tau$ and $\overline{\zeta}$ refers to the mean of two ζ's.

If we replace Y_i by $\overline{Y}_i$ and W_i by $\overline{W}_i$ in Sections 3–5, the bandwidth selection method still work, Theorems 4.1, 5.1 and 5.2 still hold except that $\boldsymbol{\Sigma}_2^{(1)}$ consists of the first q rows and columns of $\boldsymbol{\Gamma}$.

7 Some Simulation Studies

In this section, we carry out some simulation studies to demonstrate the finite sample performances of the proposed procedures.

Example 7.1 ($\boldsymbol{\Sigma}_\zeta$ Known) The data are generated from the following semiparametric regression model

$$\begin{aligned} y_i &= x_{i1}\beta_1 + x_{i2}\beta_2 + \cdots + x_{i8}\beta_8 + g(u_i) + \varepsilon_i, \quad i = 1, \cdots, n, \\ w_{is} &= x_{is} + \zeta_{is}, \end{aligned}$$

where $x_i = (x_{i1}, \cdots, x_{i8})^\tau \sim N(0, 4\mathbf{I}_8)$, $\beta_1 = \beta_2 = \beta_3 = \beta_7 = \beta_8 = 0$, $\beta_4 = 0.2$, $\beta_5 = 1.5$, $\beta_6 = 2$, $u_i \sim U(0,1)$, $g(u) = \sin(2\pi u)$, $\varepsilon_i \sim N(0,1)$ and $\zeta_i = (\zeta_{i1}, \cdots, \zeta_{i8})^\tau \sim N(0, \boldsymbol{\Sigma}_\zeta)$. We take $\Sigma_\zeta = \mathbf{I}_8$, $0.5\mathbf{I}_8$ and $0.3\mathbf{I}_8$.

The means and standard deviations of RGMSEs over 1,000 simulated data are summarized in the rows labeled "SM" and "STD" of Table 1, respectively. Here, the RGMSE means the Relative GMSE, the ratio of GMSE of an underlying procedure to that of the corrected profile least squares estimator without penalization. And for estimator $\overline{\boldsymbol{\beta}}_n$, the GMSE is defined as

$$\text{GMSE} = (\overline{\boldsymbol{\beta}}_n - \boldsymbol{\beta})^\tau\Big(\frac{1}{n}\widehat{\mathbf{W}}^\tau\widehat{\mathbf{W}} - \boldsymbol{\Sigma}_\zeta\Big)(\overline{\boldsymbol{\beta}}_n - \boldsymbol{\beta}).$$

In addition, the average number of zero coefficients is also reported in Table 1, where the row labeled "C" presents the average number, restricted only to the true zero coefficients, while the row label "I" depicts the average number of coefficients erroneously set to 0.

Moreover, we also study the case that $\boldsymbol{\Sigma}_\zeta$ is unknown.

Example 7.2 ($\boldsymbol{\Sigma}_\zeta$ Unknown) For w, we have replicated measurements

$$w_{ijs} = x_{is} + \zeta_{ijs}, \quad i = 1, \cdots, n, \; j = 1, 2, \; s = 1, \cdots, 8,$$

Table 1 Relative approximate model error for Example 7.1 with known $\mathbf{\Sigma}_\zeta$

		$n = 200$	$n = 300$	$n = 400$	$n = 500$
$\mathbf{\Sigma}_\zeta = \mathbf{I}_8$	SM	0.5002	0.4651	0.4668	0.4376
	STD	0.2580	0.2526	0.2587	0.2537
	C	4.6720	4.7200	4.7500	4.7780
	I	0.5540	0.4320	0.2960	0.1760
$\mathbf{\Sigma}_\zeta = 0.5\mathbf{I}_8$	SM	0.4862	0.4921	0.4985	0.4950
	STD	0.2465	0.2507	0.2431	0.2431
	C	4.7000	4.7200	4.7240	4.7100
	I	0.0980	0.0160	0.0020	0
$\mathbf{\Sigma}_\zeta = 0.3\mathbf{I}_8$	SM	0.5054	0.5147	0.4866	0.4970
	STD	0.2638	0.2452	0.2379	0.2514
	C	4.6800	4.7300	4.7680	4.7060
	I	0.0080	0.0040	0	0

Table 2 Relative approximate model error for Example 7.2 with unknown $\mathbf{\Sigma}_\zeta$

		$n = 200$	$n = 300$	$n = 400$	$n = 500$
$\mathbf{\Sigma}_\zeta = \mathbf{I}_8$	SM	0.4887	0.4971	0.4894	0.4762
	STD	0.2625	0.2517	0.2512	0.2508
	C	4.7080	4.7020	4.7040	4.7560
	I	0.5920	0.4140	0.2840	0.1980
$\mathbf{\Sigma}_\zeta = 0.5\mathbf{I}_8$	SM	0.4779	0.4971	0.5060	0.4979
	STD	0.2518	0.2508	0.2455	0.2422
	C	4.7260	4.7640	4.7580	4.7380
	I	0.0900	0.0200	0.0060	0
$\mathbf{\Sigma}_\zeta = 0.3\mathbf{I}_8$	SM	0.5027	0.5270	0.5100	0.4713
	STD	0.2470	0.2330	0.2453	0.2332
	C	4.7140	4.7160	4.7300	4.7840
	I	0.0160	0	0	0

and other symbols are the same as those in Example 7.1. The results are summarized in Table 2. Here, for estimator $\overline{\boldsymbol{\beta}}_n$, the GMSE is defined as

$$\text{GMSE} = (\overline{\boldsymbol{\beta}}_n - \boldsymbol{\beta})^\tau \Big(\frac{1}{n}\widehat{\mathbf{W}}^\tau \widehat{\mathbf{W}} - \widehat{\mathbf{\Sigma}}_\zeta\Big)(\overline{\boldsymbol{\beta}}_n - \boldsymbol{\beta}).$$

From Tables 1 and 2, we can see that the proposed covariate selection procedure performs very well. Whether $\mathbf{\Sigma}_\zeta$ is known or unknown almost has no influence on the results.

The following example is used to demonstrate the level and power of the proposed bootstrap based goodness-of-fit test.

Example 7.3 The data are generated from the following semiparametric regression model

$$y_i = x_{i1}\beta_1 + x_{i2}\beta_2 + g(u_i) + \varepsilon_i, \quad i = 1, \cdots, n,$$

$$w_{is} = x_{is} + \zeta_{is},$$

where $(x_{i1}, x_{i2})^\tau \sim N(0, 4\mathbf{I}_2)$, $\beta_1 = 1, \beta_2 = 1.5$, $u_i \sim U(0,1)$, $\varepsilon_i \sim N(0,1)$ and $\zeta_i = (\zeta_{i1}, \zeta_{i2})^\tau \sim N(0, \mathbf{\Sigma}_\zeta)$. We take $\Sigma_\zeta = 0.5\mathbf{I}_2$ and $n = 200$. We consider the following null hypothesis:

$$H_0 : g(U_i) = \theta U_i \quad \text{for all } i = 1, \cdots, n \text{ (a linear regression model)}$$

against the alternative

$$H_1 : g(U_i) \neq \theta U_i \quad \text{for at least one } i.$$

The power function is evaluated under the following alternatives indexed by c:

$$H_1 : g(U_i) = \theta U_i + \sin(c\pi U_i), \quad i = 1, \cdots, n.$$

The goodness-of-fit test described in Section 4 is applied to simulations with 500 replicates. Figure 1 plots the simulated power curve against c. When the null hypothesis is true, the power is very close to the significance level 5%. This demonstrates that bootstrap estimate of the null distribution is accurate. The power curve also shows that our test is quite powerful.

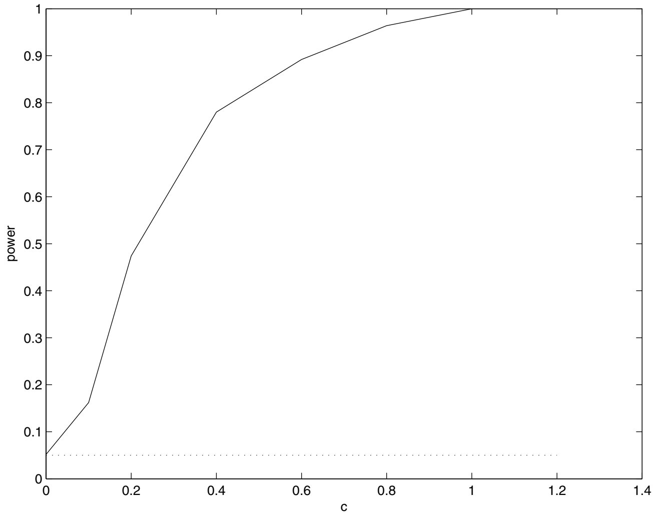

Figure 1 The power curve of testing the goodness of fit of model with $n = 200$ and $\mathbf{\Sigma}_\zeta = 0.5\mathbf{I}_2$. Power curve (slide) and 5% level (dotted)

8 Concluding Remarks

Covariate with measurement error is very common in practical application. In this paper, we have placed the emphasis on the partially linear regression models in which the covariates of the parametric part are measured with additive errors. We studied the statistical inference of this type of models. We proposed a simple bandwidth selection procedure which is a combination of the difference-based technique and cross validation method. Moreover, we proposed a goodness-of-fit test, which is an extension of the generalized likelihood technique to the setting of such models. In addition, we proposed a covariate selection procedure to select the significant covariates in the parametric part of such a model. The procedure is based on the nonconcave penalization and corrected profile least squares, and the resulting estimator owns an oracle property.

In some situations, the additivity of the measurement errors may be not true (cf. [24, 25] and so on). Thus, new procedures are needed to develop. Moreover, to extend our results to more general semiparametric measurement error regression models such as varying-coefficient partially linear measurement error regression models is also an interesting topic.

Appendix Proofs of Main Results

The proof of Theorem 2.1 is similar to [27, Theorem 3.1]. Combining the root-n consistency, Theorem 2.2 is a traditional result of nonparametric regression. The proof of Theorem 4.1 is similar to [19, Theorem 5]. Therefore, we focus our proof on Theorems 5.1 and 5.2.

In order to prove Theorems 5.1 and 5.2, we first present a lemma.

Lemma A.1 *Under the conditions of Theorem* 5.2*, with probability tending to* 1*, for any given* $\boldsymbol{\beta}_1^*$ *satisfying* $\|\boldsymbol{\beta}_1^* - \boldsymbol{\beta}_{01}\| = O_p(n^{-1/2})$ *and any constant* c*,*

$$\mathcal{L}\{(\boldsymbol{\beta}_1^{*\tau}, \mathbf{0}^\tau)^\tau\} = \min_{\|\boldsymbol{\beta}_2^*\| \le cn^{-1/2}} \mathcal{L}\{(\boldsymbol{\beta}_1^{*\tau}, \boldsymbol{\beta}_2^{*\tau})^\tau\}.$$

Proof The proof is the same as that of [17, Lemma A.1]. We are going to show that with probability tending to 1 as $n \to \infty$, for any $\boldsymbol{\beta}_1^*$ satisfying $\|\boldsymbol{\beta}_1^* - \boldsymbol{\beta}_{01}\| = O_p(n^{-1/2})$, and $\|\boldsymbol{\beta}_2^*\| \le cn^{-1/2}$, $\frac{\partial \mathcal{L}(\boldsymbol{\beta}^*)}{\partial \beta_s}$ and β_s^* have the same signs for $\beta_s^* \in (-cn^{-1/2}, cn^{-1/2})$ for $s = q+1, \cdots, p$. Thus, the minimizer attains at $\boldsymbol{\beta}_2 = \mathbf{0}$.

For $\beta_s^* \neq 0$ and $s = q+1, \cdots, p$,

$$\frac{\partial \mathcal{L}(\boldsymbol{\beta}^*)}{\partial \beta_s} = \ell_s'(\boldsymbol{\beta}^*) + np_{\lambda_s}'(|\beta_s^*|)\operatorname{sgn}(\beta_s^*),$$

where $\ell_s'(\boldsymbol{\beta}^*) = \frac{\partial \ell(\boldsymbol{\beta}^*)}{\partial \beta_s}$. It is easy to see

$$\ell_s'(\boldsymbol{\beta}^*) = -\sum_{i=1}^n \widehat{W}_{is}(\widehat{Y}_i - \widehat{W}_i^\tau \boldsymbol{\beta}_0) - n\sum_{j=1}^p \beta_{0j}\Sigma_{\zeta js} + \sum_{i=1}^n \widehat{W}_{is}\widehat{W}_i^\tau(\boldsymbol{\beta}^* - \boldsymbol{\beta}_0) - n\sum_{j=1}^p (\beta_j^* - \beta_{0j})\Sigma_{\zeta js},$$

where $\Sigma_{\zeta js}$ is the (j,s)th element of $\boldsymbol{\Sigma}_\zeta$. By the same argument as for the proof of Theorem 2.1, we can show that

$$n^{-1}\sum_{i=1}^n \widehat{W}_{is}(\widehat{Y}_i - \widehat{W}_i^\tau \boldsymbol{\beta}_0) - \sum_{j=1}^p \beta_{0j}\Sigma_{\zeta js} = O_p(n^{-1/2}).$$

Further, noting that $\|\boldsymbol{\beta}^* - \boldsymbol{\beta}_0\| = O_p(n^{-1/2})$ by the assumption, we can show that $n^{-1}\ell_s'(\boldsymbol{\beta}^*)$ is of the order $O_p(n^{-1/2})$. Therefore

$$\frac{\partial \mathcal{L}(\beta^*)}{\partial \beta_s} = n\lambda_s\{\lambda_s^{-1}p_{\lambda_s}'(|\beta_s^*|)\operatorname{sgn}(\beta_s^*) + O_p(n^{-\frac{1}{2}}\lambda_s^{-1})\}.$$

Since

$$\liminf_{n\to\infty}\liminf_{|\beta_s^*|\to 0+} \lambda_s^{-1}p_{\lambda_s}'(|\beta_s^*|) > 0 \quad \text{and} \quad (n^{\frac{1}{2}}\lambda_s)^{-1} \to 0,$$

the sign of the derivative is completely determined by that of β_s^*. This completes the proof.

Proof of Theorem 5.1 Let

$$\mathcal{L}(\boldsymbol{\beta}) = \frac{1}{2}(\widehat{\mathbf{Y}} - \widehat{\mathbf{W}}\boldsymbol{\beta})^\tau(\widehat{\mathbf{Y}} - \widehat{\mathbf{W}}\boldsymbol{\beta}) - \frac{n}{2}\boldsymbol{\beta}^\tau\boldsymbol{\Sigma}_{\boldsymbol{\zeta}}\boldsymbol{\beta} + n\sum_{s=1}^p p_{\lambda_s}(|\beta_s|).$$

Denote $\alpha_n = n^{-1/2} + a_n$. It is sufficient to show that for any given $d > 0$, there exists a large constant c such that

$$P\Big\{\inf_{\|\mathbf{u}\|=c}\mathcal{L}(\boldsymbol{\beta} + \alpha_n\mathbf{u}) \ge \mathcal{L}(\boldsymbol{\beta})\Big\} \ge 1 - d.$$

This implies, with probability at least $1-d$, there exists a local minimizer in the ball $\{\boldsymbol{\beta}+\alpha_n\mathbf{u} : \|\mathbf{u}\| \le c\}$. Define

$$D_n(\mathbf{u}) = \mathcal{L}(\boldsymbol{\beta}+\alpha_n\mathbf{u}) - \mathcal{L}(\boldsymbol{\beta}).$$

Note that $p_{\lambda_s}(0)=0$ and $p_{\lambda_s}(|\beta_s|)$ is nonnegative. Therefore, it holds that

$$\begin{aligned} n^{-1}D_n(\mathbf{u}) \ge{}& \frac{1}{2n}\{(\widehat{\mathbf{Y}} - \widehat{\mathbf{W}}(\boldsymbol{\beta}+\alpha_n\mathbf{u}))^\tau(\widehat{\mathbf{Y}} - \widehat{\mathbf{W}}(\boldsymbol{\beta}+\alpha_n\mathbf{u})) - (\widehat{\mathbf{Y}} - \widehat{\mathbf{W}}\boldsymbol{\beta})^\tau(\widehat{\mathbf{Y}} - \widehat{\mathbf{W}}\boldsymbol{\beta})\} \\ &- \frac{1}{2}\{(\boldsymbol{\beta}+\alpha_n\mathbf{u})^\tau\boldsymbol{\Sigma}_\zeta(\boldsymbol{\beta}+\alpha_n\mathbf{u}) - \boldsymbol{\beta}^\tau\boldsymbol{\Sigma}_\zeta\boldsymbol{\beta}\} + \sum_{s=1}^{q}\{p_{\lambda_s}(|\beta_s+\alpha_n u_s|) - p_{\lambda_s}(|\beta_s|)\}. \end{aligned}$$

Clearly

$$\begin{aligned} &\frac{1}{2n}\{(\widehat{\mathbf{Y}} - \widehat{\mathbf{W}}(\boldsymbol{\beta}+\alpha_n\mathbf{u}))^\tau(\widehat{\mathbf{Y}} - \widehat{\mathbf{W}}(\boldsymbol{\beta}+\alpha_n\mathbf{u})) - (\widehat{\mathbf{Y}} - \widehat{\mathbf{W}}\boldsymbol{\beta})^\tau(\widehat{\mathbf{Y}} - \widehat{\mathbf{W}}\boldsymbol{\beta})\} \\ &- \frac{1}{2}\{(\boldsymbol{\beta}+\alpha_n\mathbf{u})^\tau\boldsymbol{\Sigma}_\zeta(\boldsymbol{\beta}+\alpha_n\mathbf{u}) - \boldsymbol{\beta}^\tau\boldsymbol{\Sigma}_\zeta\boldsymbol{\beta}\} \\ ={}& \frac{\alpha_n^2}{2n}\mathbf{u}^\tau\widehat{\mathbf{W}}^\tau\widehat{\mathbf{W}}\mathbf{u} - \frac{1}{n}(\alpha_n\widehat{\mathbf{W}}\mathbf{u})^\tau(\widehat{\mathbf{Y}} - \widehat{\mathbf{W}}\boldsymbol{\beta}) - \alpha_n\boldsymbol{\beta}^\tau\boldsymbol{\Sigma}_\zeta\mathbf{u} - \frac{\alpha_n^2}{2}\mathbf{u}^\tau\boldsymbol{\Sigma}_\zeta\mathbf{u} \\ ={}& \frac{\alpha_n^2}{2n}\mathbf{u}^\tau\widehat{\mathbf{W}}^\tau\widehat{\mathbf{W}}\mathbf{u} - \frac{1}{n}(\alpha_n\widehat{\mathbf{W}}\mathbf{u})^\tau(\widehat{\boldsymbol{\varepsilon}} - \widehat{\boldsymbol{\zeta}}\boldsymbol{\beta}) - \alpha_n\boldsymbol{\beta}^\tau\boldsymbol{\Sigma}_\zeta\mathbf{u} - \frac{\alpha_n^2}{2}\mathbf{u}^\tau\boldsymbol{\Sigma}_\zeta\mathbf{u} \\ ={}& \frac{\alpha_n^2}{2}\mathbf{u}^\tau\Big(\frac{1}{n}\widehat{\mathbf{W}}^\tau\widehat{\mathbf{W}} - \boldsymbol{\Sigma}_\zeta\Big)\mathbf{u} - \frac{\alpha_n}{n}\mathbf{u}^\tau\widehat{\mathbf{W}}(\boldsymbol{\varepsilon} - \boldsymbol{\zeta}\boldsymbol{\beta}) - \alpha_n\boldsymbol{\beta}^\tau\boldsymbol{\Sigma}_\zeta\mathbf{u} + o_p(n^{-\frac{1}{2}}\alpha_n\|\mathbf{u}\|) \\ ={}& \frac{\alpha_n^2}{2}\mathbf{u}^\tau\Big(\frac{1}{n}\widehat{\mathbf{W}}^\tau\widehat{\mathbf{W}} - \boldsymbol{\Sigma}_\zeta\Big)\mathbf{u} - \frac{\alpha_n}{n}\mathbf{u}^\tau(\widehat{\mathbf{X}} + \widehat{\boldsymbol{\zeta}})(\boldsymbol{\varepsilon} - \boldsymbol{\zeta}\boldsymbol{\beta}) - \alpha_n\boldsymbol{\beta}^\tau\boldsymbol{\Sigma}_\zeta\mathbf{u} + o_p(n^{-\frac{1}{2}}\alpha_n\|\mathbf{u}\|) \\ ={}& \frac{\alpha_n^2}{2}\mathbf{u}^\tau\Big(\frac{1}{n}\widehat{\mathbf{W}}^\tau\widehat{\mathbf{W}} - \boldsymbol{\Sigma}_\zeta\Big)\mathbf{u} - \frac{\alpha_n}{n}\mathbf{u}^\tau\mathbf{X}^*(\mathbf{U})(\boldsymbol{\varepsilon} - \boldsymbol{\zeta}\boldsymbol{\beta}) - \frac{\alpha_n}{n}\mathbf{u}^\tau\boldsymbol{\zeta}\boldsymbol{\varepsilon} + \alpha_n\boldsymbol{\beta}^\tau\Big(\frac{1}{n}\boldsymbol{\zeta}^\tau\boldsymbol{\zeta} - \boldsymbol{\Sigma}_\zeta\Big)\mathbf{u} \\ &+ O_p(n^{-\frac{1}{2}}\alpha_n\|\mathbf{u}\|) \\ ={}& \frac{\alpha_n^2}{2}\mathbf{u}^\tau\Big(\frac{1}{n}\widehat{\mathbf{W}}^\tau\widehat{\mathbf{W}} - \boldsymbol{\Sigma}_\zeta\Big)\mathbf{u} - \alpha_n\Big\{\frac{1}{n}\mathbf{u}^\tau\mathbf{X}^*(\mathbf{U})(\boldsymbol{\varepsilon} - \boldsymbol{\zeta}\boldsymbol{\beta}) + \frac{1}{n}\mathbf{u}^\tau\boldsymbol{\zeta}\boldsymbol{\varepsilon} - \mathbf{u}^\tau\Big(\frac{1}{n}\boldsymbol{\zeta}^\tau\boldsymbol{\zeta} - \boldsymbol{\Sigma}_\zeta\Big)\boldsymbol{\beta}\Big\} \\ &+ O_p(n^{-\frac{1}{2}}\alpha_n\|\mathbf{u}\|) \\ ={}& J_1 + J_2 + O_p(n^{-\frac{1}{2}}\alpha_n\|\mathbf{u}\|), \end{aligned}$$

say, where $\boldsymbol{\zeta} = (\zeta_1,\cdots,\zeta_n)^\tau$, $\boldsymbol{\varepsilon} = (\varepsilon_1,\cdots,\varepsilon_n)^\tau$, $\widehat{\boldsymbol{\varepsilon}} = (\mathbf{I}_n - \mathbf{S})\boldsymbol{\varepsilon}$, $\widehat{\boldsymbol{\zeta}} = (\mathbf{I}_n - \mathbf{S})\boldsymbol{\zeta}$, and $\widehat{\mathbf{X}}^*(\mathbf{U}) = (E(X_1\,|\,U_1),\cdots,E(X_n\,|\,U_n))^\tau$. Since when n is large enough

$$\frac{1}{n}\widehat{\mathbf{W}}^\tau\widehat{\mathbf{W}} - \boldsymbol{\Sigma}_\zeta = \mathrm{Cov}(X - E(X|U)) + O_p(n^{-\frac{1}{2}}) > 0,$$

J_1 is of the order $c^2\alpha_n^2$. Note that $n^{-1/2}\alpha_n = O_p(\alpha_n^2)$. By choosing a sufficiently large c, J_1 will dominate the second term, uniformly in $\|\mathbf{u}\| = c$. Furthermore,

$$\sum_{s=1}^{p}\{p_{\lambda_s}(|\beta_s+\alpha_n u_s|) - p_{\lambda_s}(|\beta_s|)\}$$

is bounded by

$$\sqrt{q}\alpha_n a_n\|\mathbf{u}\| + \alpha_n^2 b_n\|\mathbf{u}\|^2 = c\alpha_n^2(\sqrt{q} + b_n c)$$

by the Taylor expansion and Cauchy-Schwarz inequality, where q is the number of components of $\boldsymbol{\beta}_1$. $c\alpha_n^2(\sqrt{q}+b_n c)$ is dominated by J_1 as $b_n \to 0$, by taking c sufficiently large. This completes the proof of the theorem.

Proof of Theorem 5.2 Part (i) directly follows by Lemma A.1. Now we prove Part (ii). Using argument similar to the proof of Theorem 5.1, it can be shown that there exists a $\widetilde{\boldsymbol{\beta}}_{n1}$ in Theorem 5.1 that is a root-n consistent minimizer of $\mathcal{L}\{(\boldsymbol{\beta}_1^\tau, \mathbf{0}^\tau)^\tau\}$, satisfying the penalized corrected profile least squares equations:

$$\frac{\partial \mathcal{L}\{(\widetilde{\boldsymbol{\beta}}_{n1}^\tau, \mathbf{0}^\tau)^\tau\}}{\partial \boldsymbol{\beta}_1} = 0.$$

Further, we have

$$\begin{aligned}\frac{\partial \mathcal{L}\{(\widetilde{\boldsymbol{\beta}}_{n1}^\tau, \mathbf{0}^\tau)^\tau\}}{\partial \boldsymbol{\beta}_1} &= -\widehat{\mathbf{W}}^{(1)}(\widehat{\mathbf{Y}} - \widehat{\mathbf{W}}^{(1)}\boldsymbol{\beta}_1) - n\boldsymbol{\Sigma}_\zeta^{(1)}\boldsymbol{\beta}_1 + (\widehat{\mathbf{W}}^{(1)}\widehat{\mathbf{W}}^{(1)\tau} - n\boldsymbol{\Sigma}_\zeta^{(1)})(\widetilde{\boldsymbol{\beta}}_{n1} - \boldsymbol{\beta}_1)\\ &\quad - n[\mathbf{b} + \{\mathbf{B} + o_p(1)\}(\widetilde{\boldsymbol{\beta}}_{n1} - \boldsymbol{\beta}_1)],\end{aligned}$$

where $\widehat{\mathbf{W}}^{(1)}$ consists of the first q columns of $\widehat{\mathbf{W}}$, and $\boldsymbol{\Sigma}_\zeta^{(1)}$ consists of the first q rows and columns of $\boldsymbol{\Sigma}_\zeta$.

Similarly to Theorem 2.1, we can show that

$$-\frac{1}{\sqrt{n}}\{\widehat{\mathbf{W}}^{(1)}(\widehat{\mathbf{Y}} - \widehat{\mathbf{W}}^{(1)}\boldsymbol{\beta}_1) + n\boldsymbol{\Sigma}_\zeta^{(1)}\boldsymbol{\beta}_1\} \to_D N(\mathbf{0}, \boldsymbol{\Sigma}_2^{(1)}), \quad \text{as } n \to \infty,$$

where $\boldsymbol{\Sigma}_2^{(1)}$ consists of the first q rows and columns of $\boldsymbol{\Sigma}_2$. Thus, by Slutsky's Theorem, it follows that

$$\sqrt{n}\{\boldsymbol{\Sigma}_1^{(1)} + \mathbf{B}\}\{\widetilde{\boldsymbol{\beta}}_{n1} - \boldsymbol{\beta}_1 + (\boldsymbol{\Sigma}_1^{(1)} + \mathbf{B})^{-1}\mathbf{b}\} \to_D N(\mathbf{0}, \boldsymbol{\Sigma}_2^{(1)}),$$

where $\boldsymbol{\Sigma}_1^{(1)}$ consists of the first q rows and columns of $\boldsymbol{\Sigma}_1$. This completes the proof of Theorem 5.2.

Acknowledgement The authors are grateful to the editors and referees whose comments improved this paper.

References

[1] Antoniadis, A. and Fan, J., Regularization of wavelet approximations (with discussion and a rejoinder by the authors), *J. Amer. Statist. Assoc.*, **96**, 2001, 939–967.

[2] Bickel, P. J. and Kwon, J., Inference for semiparametric models: some questions and an answer (with comments and a rejoinder by the authors), *Statist. Sinica*, **11**, 2001, 863–960.

[3] Breiman, L., Better subset regression using the nonnegative garrote, *Technometrics*, **37**, 1995, 373–384.

[4] Cai, J., Fan, J., Li, R. and Zhou, H., Model selection for multivariate failure time data, *Biometrika*, **92**, 2005, 303–316.

[5] Carroll, R. J., Ruppert, D. and Stefanski, L. A., Measurement Error in Nonlinear Models, Chapman and Hall, London, 1995.

[6] Chen, H., Convergence rates for parametric components in a partly linear model, *Ann. Statist.*, **16**, 1988, 136–146.

[7] Chen, H. and Shiau, J. H., A two-stage spline smoothing method for partially linear models, *J. Statist. Plann. Infer.*, **27**, 1991, 187–202.

[8] Chen, H. and Shiau, J. H., Data-driven efficient estimators for a partially linear model, *Ann. Statist.*, **22**, 1994, 211–237.

[9] Cheng, C. L. and Tsai, C. L., The invariance of some score tests in the linear model with classical measurement error, *J. Amer. Statist. Assoc.*, **99**, 2004, 805–809.

[10] Cui, H. and Li, R., On parameter estimation for semi-linear errors-in-variables models, *J. Multivar. Anal.*, **64**, 1998, 1–24.

[11] Donald, G. and Newey, K., Series estimation of semilinear models, *J. Multivar. Anal.*, **50**, 1994, 30–40.

[12] Engle, R. F., Granger, C. W. J., Rice, J. and Weiss, A., Semiparametric estimates of the relation between weather and electricity sales, *J. Amer. Statist. Assoc.*, **81**, 1986, 310–320.

[13] Eubank, R., Hart, J. D. and Speckman, P., Trigonometric series regression estimators with an application to partially linear models, *J. Multivar. Anal.*, **32**, 1990, 70–83.

[14] Fan, J. and Gijbels, I., Local Polynomial Modeling and Its Applications, Chapman and Hall, London, 1996.

[15] Fan, J. and Li, R., Variable selection via nonconcave penalized likelihood and its oracle properties, *J. Amer. Statist. Assoc.*, **96**, 2001, 1348–1360.

[16] Fan, J. and Li, R., Variable selection for Cox's proportional Hazards model and frailty model, *Ann. Statist.*, **30**, 2002, 74–99.

[17] Fan, J. and Li, R., New estimation and model selection procedures for semiparametric modeling in longitudinal data analysis, *J. Amer. Statist. Assoc.*, **99**, 2004, 710–723.

[18] Fan, J. and Peng, H., Non-concave penalized likelihood with diverging number of parameters, *Ann. Statist.*, **32**, 2004, 928–961.

[19] Fan, J., Zhang, C. and Zhang, J., Generalized likelihood ratio statistics and Wilks phenomenon, *Ann. Statist.*, **29**, 2001, 153–193.

[20] Frank, I. E. and Friedman, J. H., A statistical view of some chemometrics regression tools (with discussion), *Technometrics*, **35**, 1993, 109–148.

[21] Hamilton, S. A. and Truong, Y. K., Local linear estimation in partly linear models, *J. Multivar. Anal.*, **60**, 1997, 1–19.

[22] Härdle, W., Applied Nonparametric Regression, Cambridge University Press, Cambridge, New York, 1990.

[23] Härdle, W., Liang, H. and Gao, J. T., Partially Linear Models, Physica-Verlag, Heidelberg, 2000.

[24] Hwang, J. T., Multiplicative errors-in-variables models with applications to recent data released by the U.S. department of energy, *J. Amer. Statist. Assoc.*, **81**, 1986, 680–688.

[25] Iturria, S. J., Carroll, R. J. and Firth, D., Polynomial regression and estimating functions in the presence of multiplicative measurement error, *J. Roy. Statist. Soc. Ser. B*, **61**, 1999, 547–561.

[26] Liang, H., Asymptotic normality of parametric part in partially linear models with measurement error in the nonparametric part, *J. Statist. Plann. Infer.*, **86**, 2000, 51–62.

[27] Liang, H., Härdle, W. and Carroll, R. J., Estimation in a semiparametric partially linear errors-in-variables model, *Ann. Statist.*, **27**, 1999, 1519–1535.

[28] Rice, J., Convergence rates for partially splined models, *Statist. Probab. Lett.*, **4**, 1986, 203–208.

[29] Robinson, P. M., Root-N-consistent semiparametric regression, *Econometrica*, **56**, 1988, 931–954.

[30] Shi, P. D. and Li, G. Y., A note on the convergence rates of M-estimates for partly linear model, *Statistics*, **26**, 1995, 27–47.

[31] Speckman, P., Kernel smoothing in partial linear models, *J. Roy. Statist. Soc., Ser. B*, **50**, 1988, 413–436.

[32] Tibshirani, R., Regression shrinkage and selection via the lasso, *J. Roy. Statist. Soc. Ser. B*, **58**, 1996, 267–288.

[33] You, J. H. and Xu, Q., Covariate selection for linear errors-in-variables regression models, *Comm. Statist. Theory Methods*, **36**, 2007, 375–386.

[34] Zhu, L. and Cui, H., A semiparametric regression model with errors in variables, *Scand. J. Statist.*, **30**, 2003, 429–442.