

# ESTIMATION AND TESTING FOR PARTIALLY LINEAR SINGLE-INDEX MODELS (TECHNICAL REPORT)

BY HUA LIANG<sup>\*</sup> XIANG LIU RUNZE LI<sup>†</sup> AND CHIH-LING TSAI

*University of Rochester, The Pennsylvania State University, and University of California*

In partially linear single-index models, we obtain the semiparametrically efficient profile least-squares estimators of regression coefficients. We also employ the smoothly clipped absolute deviation penalty (SCAD) approach to simultaneously select variables and estimate regression coefficients. We show that the resulting SCAD estimators are consistent and possess the oracle property. Subsequently, we demonstrate that a proposed tuning parameter selector, BIC, identifies the true model consistently. Finally, we develop a linear hypothesis test of parametric coefficients and a goodness-of-fit test of the nonparametric component, respectively. Monte Carlo studies and an empirical example are presented.

## 1. Introduction

Regression analysis is commonly used to explore the relationship between a response variable  $Y$  and its covariates  $Z$ . For the sake of convenience, one often employs a linear regression model  $E(Y|Z) = Z^T \alpha$  to assess the impact of the covariates on the response, where  $\alpha$  is an unknown vector and  $E$  stands for expectation. In practice, however, the linear assumption may not be valid. Hence, it is natural to consider a single-index model  $E(Y|Z) = \eta(Z^T \alpha)$ , in which the link function  $\eta$  is unknown. Accordingly, various parameter estimators of single-index models have been proposed (e.g., Powell, Stock and Stocker, 1989; Duan and Li, 1991; Härdle, Hall and Ichimura, 1993; Ichimura, 1993; Horowitz and Härdle, 1996; Liang and Wang, 2005). Detailed

---

<sup>\*</sup>Corresponding author. Liang's research was partially supported by NIH/NIAID grant AI59773 and NSF grant DMS-0806097. Liu's Research was partially supported by Merck Quantitative Sciences Fellowship Program.

<sup>†</sup> Li's research is supported by National Science Foundation grants DMS 0348869 and NIDA, NIH grants R21 DA024260 and P50 DA10075. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIDA or the NIH.

*AMS 2000 subject classifications:* Primary 62G08; secondary 62G10, 62G20, 62J02 62F12

*Keywords and phrases:* Efficiency, Hypothesis testing, Local linear regression, Nonparametric regression, Profile likelihood, SCAD

discussion and illustration of the usefulness of this model can be found in Horowitz (1998). Although the single-index model plays an important role in data analysis, it may not be sufficient to explain the variation of responses via covariates  $Z$ . Therefore, Carroll, Fan, Gijbels and Wand (1997) augmented the single-index model in a linear form with additional covariates  $X$ , which yields a partially linear single-index model (PLSIM),  $E(Y|Z, X) = \eta(Z^T \boldsymbol{\alpha}) + X^T \boldsymbol{\beta}$ . When  $Z$  is scalar and  $\boldsymbol{\alpha} = 1$ , the PLSIM reduces to the partially linear model,  $E(Y|Z, X) = \eta(Z) + X^T \boldsymbol{\beta}$  (see Speckman, 1988). A comprehensive review on partially linear models can be found in Härdle, Liang and Gao (2000).

To estimate parameters in partially linear single-index models, Carroll et al. (1997) proposed the backfitting algorithm. However, the resulting estimators may be unstable (see Yu and Ruppert, 2002) and undersmoothing the nonparametric function is necessary to reduce the bias of the parametric estimators. Accordingly, Yu and Ruppert (2002) proposed the penalized spline estimation procedure, while Xia and Härdle (2006) applied the minimum average variance estimation (MAVE) method, which was originally introduced by Xia, Tong, Li and Zhu (2002) for dimension reduction. Although Yu and Ruppert's procedure is useful, it may not yield efficient estimators. In addition, these estimators need to be solved via an iterative procedure; that is, iteratively estimating the nonparametric component and the parametric component. We therefore propose the profile least squares approach, which obtains efficient estimators and provides the efficient bound. Moreover, the resulting estimators can be found without using the iterative procedure mentioned above, and hence the computational gain is substantial.

In data analysis, the true model is often unknown; this allows the possibility of selecting an underfitted (or overfitted) model, leading to biased (or inefficient) estimators and predictions. To address this problem, Tibshirani (1996) introduced the least absolute shrinkage and selection operator (LASSO) to shrink the estimated coefficients of superfluous variables to zero in linear regression models. Subsequently, Fan and Li (2001) proposed the smoothly clipped absolute deviation (SCAD) approach that not only selects important variables consistently, but also produces parameter estimators as efficient as if the true model were known, a property not possessed by the LASSO. Because it is not a simple matter to formulate the penalized function via the MAVE's procedure, we employ the profile least squares approach to obtain the SCAD estimators for both parameter vectors  $\boldsymbol{\alpha}$  and  $\boldsymbol{\beta}$ . Furthermore, we establish the asymptotical

results of the SCAD estimators, which include the consistency and oracle properties (see Fan and Li, 2001). Simulation results are consistent with theoretical findings.

After estimating unknown parameters, it is natural to construct hypothesis tests to assess the appropriateness of the linear hypothesis on parametric components as well as the linearity of the nonparametric function. We demonstrate that the resulting test statistics are asymptotically chi-square distributed under the null hypothesis. In addition, simulation studies indicate that the test statistics perform well. The rest of this paper is organized as follows. Section 2 introduces the profile least squares estimators and the penalized SCAD estimators. The asymptotic properties of these estimators are obtained. Section 3 presents hypothesis tests and their large sample properties. Monte Carlo studies and an empirical example are presented in Section 4. Section 5 concludes the article with a brief discussion. All detailed proofs are relegated to the Appendix.

## 2. Profile Least Squares Procedure

Suppose that  $\{(Y_i, Z_i, X_i), i = 1, \dots, n\}$  is a random sample generated from the PLSIM

$$(2.1) \quad Y = \eta(Z^T \boldsymbol{\alpha}) + X^T \boldsymbol{\beta} + \varepsilon,$$

where  $Z$  and  $X$  are  $p$ -dimensional and  $q$ -dimensional covariate vectors, respectively,  $\boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_p)^T$ ,  $\boldsymbol{\beta} = (\beta_1, \dots, \beta_q)^T$ ,  $\eta(\cdot)$  is an unknown differentiable function,  $\varepsilon$  is the random error with mean zero and finite variance  $\sigma^2$ , and  $(Z^T, X^T)^T$  and  $\varepsilon$  are independent. Furthermore, we assume that  $\|\boldsymbol{\alpha}\| = 1$ , and the first element of  $\boldsymbol{\alpha}$  is positive to ensure identifiability. We then employ the profile least squares procedure to obtain efficient estimators and SCAD estimators in the following two subsections, respectively.

### 2.1. Profile least squares estimator

In semiparametric models, Severini and Wong (1992) applied the profile likelihood approach to estimate the parametric component. We adapt this approach to estimate unknown parameters of partially linear single-index models. To begin, we re-express model (2.1) as

$$Y_i^* = \eta(\Lambda_i) + \varepsilon_i,$$

where  $Y_i^* = Y_i - X_i^T \boldsymbol{\beta}$  and  $\Lambda_i = Z_i^T \boldsymbol{\alpha}$ . Then, for a given  $\boldsymbol{\zeta} = (\boldsymbol{\alpha}^T, \boldsymbol{\beta}^T)^T$ , we employ the local linear regression technique (Fan and Gijbels, 1996) to estimate  $\eta$ , that is to minimize

$$(2.2) \quad \sum_{i=1}^n \{a + b(\Lambda_i - u) + X_i^T \boldsymbol{\beta} - Y_i\}^2 K_h(\Lambda_i - u),$$

with respect to  $a, b$ , where  $K_h(\cdot) = 1/hK(\cdot/h)$ ,  $K(\cdot)$  is a kernel function, and  $h$  is a bandwidth.

Let  $(\hat{a}, \hat{b})$  be the minimizer of (2.2). Then,

$$(2.3) \quad \hat{\eta}(u, \boldsymbol{\zeta}) = \hat{a} = \frac{K_{20}(u, \boldsymbol{\zeta})K_{01}(u, \boldsymbol{\zeta}) - K_{10}(u, \boldsymbol{\zeta})K_{11}(u, \boldsymbol{\zeta})}{K_{00}(u, \boldsymbol{\zeta})K_{20}(u, \boldsymbol{\zeta}) - K_{10}^2(u, \boldsymbol{\zeta})},$$

where  $K_{jl}(u, \boldsymbol{\zeta}) = \sum_{i=1}^n K_h(Z_i^T \boldsymbol{\alpha} - u)(Z_i^T \boldsymbol{\alpha} - u)^j (X_i^T \boldsymbol{\beta} - Y_i)^l$  for  $j = 0, 1, 2$  and  $l = 0, 1$ . Subsequently, following Jennrich's (1969) assumption (a), there exists a profile least squares estimator  $\hat{\boldsymbol{\zeta}} = (\hat{\boldsymbol{\alpha}}^T, \hat{\boldsymbol{\beta}}^T)^T$  obtained by minimizing the following profile least squares function

$$(2.4) \quad Q(\boldsymbol{\zeta}) = \sum_{i=1}^n \{Y_i - \hat{\eta}(Z_i^T \boldsymbol{\alpha}, \boldsymbol{\zeta}) - X_i^T \boldsymbol{\beta}\}^2.$$

It is noteworthy that we apply the Newton-Naphson iterative method to find the estimator  $\hat{\boldsymbol{\zeta}}$ ; this technique is distinct from the more commonly used iterative procedure in partially single-index models which iteratively updates estimations of the nonparametric component and the parametric component obtained from their corresponding objective functions. In addition, our proposed profile least squares approach allows us to directly introduce the penalized function as given in the next subsection.

To study the large sample properties of parameter estimators, we consider the true model with an unknown parameter vector  $\boldsymbol{\zeta}_0 = (\boldsymbol{\alpha}_0^T, \boldsymbol{\beta}_0^T)^T$ . In addition, we assume that  $\boldsymbol{\alpha}_0^T$  and  $\boldsymbol{\beta}_0^T$  have the same dimensions as their corresponding parameter vectors  $\boldsymbol{\alpha}^T$  and  $\boldsymbol{\beta}^T$  of the candidate model in this subsection. Moreover, we introduce the following notations:  $A^{\otimes 2} = AA^T$  for a matrix  $A$ ,  $\Lambda = Z^T \boldsymbol{\alpha}$ ,  $\hat{\Lambda} = Z^T \hat{\boldsymbol{\alpha}}$ ,  $\Lambda_0 = Z^T \boldsymbol{\alpha}_0$ ,  $\tilde{\xi} = \xi - E(\xi|\Lambda)$ ,  $\tilde{\xi}_0 = \xi - E(\xi|\Lambda_0)$ , and  $\hat{\xi} = \xi - \hat{E}(\xi|\Lambda)$  for any random variable (or vector)  $\xi$ , where  $\hat{E}(\xi|\Lambda)$  is the local linear estimator of  $E(\xi|\Lambda)$ . For example,  $\tilde{Z} = Z - E(Z|\Lambda)$  and  $\hat{X} = X - \hat{E}(X|\Lambda)$ . We next present six conditions and then obtain the weak consistency and asymptotic normality of the profile least squares estimators.

- (i) The function  $\eta(\cdot)$  is differentiable and not constant on the support  $\mathcal{U}$  of  $Z^T \boldsymbol{\alpha}$ .
- (ii) The function  $\eta(z^T \boldsymbol{\alpha})$  and the density function of  $Z^T \boldsymbol{\alpha}$ ,  $f_{\boldsymbol{\alpha}}(z)$ , are both three times continuously differentiable with respect to  $z$ . The third derivatives are uniformly Lipschitz continuous over  $\mathcal{A} \subset \mathcal{R}^p$  for all  $u \in \{u = z^T \boldsymbol{\alpha} : \boldsymbol{\alpha} \in \mathcal{A}, z \in \mathcal{Z} \subset \mathcal{R}^p\}$ .

- (iii)  $E(|Y|^{m_1}) < \infty$  for some  $m_1 \geq 3$ . The conditional variance of  $Y$  given  $(X, Z)$  is bounded and bounded away from 0.
- (iv) The kernel function  $K(\cdot)$  is twice continuously differentiable with the support  $(-1, 1)$ . In addition, its second derivative is Lipschitz continuous. Moreover,  $\int u^j K(u) du = 1$  if  $j = 0$ , 0 if  $j = 1$ .
- (v) The bandwidth  $h$  satisfies  $\{nh^{3+3/(m_1-1)}\} \log^{-1} n \rightarrow \infty$  and  $nh^8 \rightarrow 0$  as  $n \rightarrow \infty$ .
- (vi)  $E(\tilde{X}^{\otimes 2})$  and  $E\{\tilde{Z}\eta'(\Lambda)\}^{\otimes 2}$  are positive finite, where  $\eta'$  is the first derivative of  $\eta$ .

**Theorem 1** *Under the regularity conditions (i)-(vi), with probability tending to one,  $\hat{\zeta}$  is a consistent estimator of  $\zeta$ . Furthermore,  $\sqrt{n}(\hat{\zeta} - \zeta_0) \rightarrow N(0, \sigma^2 \mathbf{D}^{-1})$  in distribution, where  $\mathbf{D} = E\left[\{\eta'(\Lambda_0)\tilde{Z}_0^T, \tilde{X}_0^T\}^T\right]^{\otimes 2}$ . Moreover,  $\hat{\zeta}$  is a semiparametrically efficient estimator.*

Because  $(Z^T, X^T)^T$  is independent of  $\varepsilon$ , we are able to demonstrate that the asymptotic variance of Xia and Härdle's (2006) minimum average variance estimator is  $\sigma^2 \mathbf{D}^{-1}$ , which indicates that MAVE is also an efficient estimator.

After having estimated  $\alpha$  and  $\beta$ , we obtain the following estimator of  $\eta(u)$ ,

$$\hat{\eta}(u) = \hat{\eta}(u, \hat{\zeta}) = \frac{K_{20}(u, \hat{\zeta})K_{01}(u, \hat{\zeta}) - K_{10}(u, \hat{\zeta})K_{11}(u, \hat{\zeta})}{K_{00}(u, \hat{\zeta})K_{20}(u, \hat{\zeta}) - K_{10}^2(u, \hat{\zeta})}.$$

If the density function  $f_{\Lambda_0}$  of  $\Lambda_0$  is positive and the derivative of  $E(\varepsilon^2 | \Lambda_0 = u)$  exists, then we can further demonstrate that  $(nh)^{1/2}\{\hat{\eta}(u) - \eta(u) - 1/2k_2\eta''(u)h^2\}$  converges to a normal distribution  $N(0, \sigma_{\eta,u}^2)$ , where  $\sigma_{\eta,u}^2 = f_{\Lambda_0}^{-1}(u) \int K^2(t) dt E(\varepsilon^2 | \Lambda_0 = u)$ ,  $k_2 = \int K(t)t^2 dt$ , and  $\eta''$  is the second derivative of  $\eta$ . It is also noteworthy that one usually either introduces a trimming function or adds a ridge parameter (see Seifert and Gassert, 1996) when the division of  $\hat{\eta}$  closes to zero.

**Remark 1** *Condition (v) indicates that Theorem 1 is applicable for a reasonable range of bandwidths. Numerical studies confirm it; our results remain stable by employing various bandwidths around the optimal bandwidth selected by cross-validation, in particular when the sample size becomes large.*

## 2.2. Penalized profile least squares estimator

In practice, the true model is often unknown a priori. An underfitted model can yield biased estimates and predicted values, while an overfitted model can degrade the efficiency of the

parameter estimates and predictions. This motivates us to apply the penalized least squares approach to simultaneously estimate parameters and select important variables. To this end, we consider a penalized profile least squares function

$$(2.5) \quad \mathcal{L}_P(\boldsymbol{\zeta}) = \frac{1}{2}Q(\boldsymbol{\zeta}) + n \sum_{j=1}^p p_{\lambda_{1j}}(|\alpha_j|) + n \sum_{k=1}^q p_{\lambda_{2k}}(|\beta_k|),$$

where  $p_\lambda(\cdot)$  is a penalty function with a regularization parameter  $\lambda$ . Throughout this paper, we allow different elements of  $\boldsymbol{\alpha}$  and  $\boldsymbol{\beta}$  to have different penalty functions with different regularization parameters. For the purpose of selecting  $X$ -variables only, we simply set  $p_{\lambda_{1j}}(\cdot) = 0$  and the resulting penalized profile least squares function becomes

$$(2.6) \quad \mathcal{L}_P(\boldsymbol{\zeta}) = \frac{1}{2}Q(\boldsymbol{\zeta}) + n \sum_{k=1}^q p_{\lambda_{2k}}(|\beta_k|).$$

Similarly, if we are only interested in selecting  $Z$ -variables, then set  $p_{\lambda_{2k}}(\cdot) = 0$  so that

$$(2.7) \quad \mathcal{L}_P(\boldsymbol{\zeta}) = \frac{1}{2}Q(\boldsymbol{\zeta}) + n \sum_{j=1}^p p_{\lambda_{1j}}(|\alpha_j|).$$

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

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

and  $p_\lambda(0) = 0$ , where  $a = 3.7$  and  $(t)_+ = tI\{t > 0\}$  is the hinge loss function. For the given tuning parameters, we obtain the penalized estimators by minimizing  $\mathcal{L}_P(\boldsymbol{\alpha}, \boldsymbol{\beta})$  with respect to  $\boldsymbol{\alpha}$  and  $\boldsymbol{\beta}$ . For the sake of simplicity, we denote the resulting estimators by  $\widehat{\boldsymbol{\alpha}}_{\lambda_1}$  and  $\widehat{\boldsymbol{\beta}}_{\lambda_2}$ .

In what follows, we study the theoretical properties of the penalized profile least squares estimators with the SCAD penalty. Without loss of generality, it is assumed that the correct model has regression coefficients  $\boldsymbol{\alpha}_0 = (\boldsymbol{\alpha}_{10}^\top, \boldsymbol{\alpha}_{20}^\top)^\top$  and  $\boldsymbol{\beta}_0 = (\boldsymbol{\beta}_{10}^\top, \boldsymbol{\beta}_{20}^\top)^\top$ , where  $\boldsymbol{\alpha}_{10}$  and  $\boldsymbol{\beta}_{10}$  are  $p_0 \times 1$  and  $q_0 \times 1$  nonzero components of  $\boldsymbol{\alpha}_0$  and  $\boldsymbol{\beta}_0$ , respectively, and  $\boldsymbol{\alpha}_{20}$  and  $\boldsymbol{\beta}_{20}$  are  $(p-p_0) \times 1$  and  $(q-q_0) \times 1$  vectors with zeros. In addition, we denote that  $Z_1$  and  $X_1$  consist of the first  $p_0$  and  $q_0$  elements of  $Z$  and  $X$ , respectively. We define  $\widetilde{Z}_1$  and  $\widetilde{X}_1$  analogously. Finally, we use the following notation for simplicity:  $\Gamma_{\widetilde{X}_1 \widetilde{X}_1} = E(\widetilde{X}_1^{\otimes 2})$ ,  $\Gamma_{\widetilde{X}_1 \widetilde{Z}_1} = E\{\widetilde{Z}_1 \widetilde{X}_1^\top \eta'(\Lambda)\}$ , and  $\Gamma_{\widetilde{Z}_1 \widetilde{Z}_1} = E\{\widetilde{Z}_1 \eta'(\Lambda)\}^{\otimes 2}$ .

**Theorem 2** *Under the regularity conditions (i)-(vi), if for all  $k$  and  $j$ ,  $\lambda_{1j} \rightarrow 0$ ,  $\sqrt{n}\lambda_{1j} \rightarrow \infty$ ,  $\lambda_{2k} \rightarrow 0$  and  $\sqrt{n}\lambda_{2k} \rightarrow \infty$  as  $n \rightarrow \infty$ , with probability tending to one, then the penalized estimators  $\hat{\alpha}_{\lambda_1} = (\hat{\alpha}_{1\lambda_1}^T, \hat{\alpha}_{2\lambda_1}^T)^T$  and  $\hat{\beta}_{\lambda_2} = (\hat{\beta}_{1\lambda_2}^T, \hat{\beta}_{2\lambda_2}^T)^T$  satisfy*

- (a)  $\hat{\alpha}_{2\lambda_1} = 0$  and  $\hat{\beta}_{2\lambda_2} = 0$ ;
- (b)  $\sqrt{n}(\hat{\alpha}_{1\lambda_1} - \alpha_{10}) \rightarrow N \left\{ 0, \sigma^2 (\Gamma_{\tilde{Z}_1 \tilde{Z}_1} - \Gamma_{\tilde{X}_1 \tilde{Z}_1} \Gamma_{\tilde{X}_1 \tilde{X}_1}^{-1} \Gamma_{\tilde{X}_1 \tilde{Z}_1}^T)^{-1} \right\}$  and  $\sqrt{n}(\hat{\beta}_{1\lambda_2} - \beta_{10}) \rightarrow N \left\{ 0, \sigma^2 (\Gamma_{\tilde{X}_1 \tilde{X}_1} - \Gamma_{\tilde{X}_1 \tilde{Z}_1}^T \Gamma_{\tilde{Z}_1 \tilde{Z}_1}^{-1} \Gamma_{\tilde{X}_1 \tilde{Z}_1})^{-1} \right\}$ .

Analogous results can be established for those parameter estimators obtained via the penalized functions (2.6) or (2.7).

Theorem 2 indicates that the proposed variable selection procedure possesses the oracle property. However, this attractive feature relies on the tuning parameters. To this end, we adopt Wang, Li and Tsai's (2007) BIC selector to choose the regularization parameters  $\lambda_{1j}$  and  $\lambda_{2k}$ . Because it is computationally expensive to minimize BIC, defined below, with respect to the  $(p+q)$ -dimensional regularization parameters, we follow the approach of Fan and Li (2004) to set  $\lambda_{1j} = \lambda \text{SE}(\hat{\alpha}_j^u)$  and  $\lambda_{2k} = \lambda \text{SE}(\hat{\beta}_k^u)$ , where  $\lambda$  is the tuning parameter, and  $\text{SE}(\hat{\alpha}_j^u)$  and  $\text{SE}(\hat{\beta}_k^u)$  are the standard errors of the unpenalized profile least squares estimators of  $\alpha_j$  and  $\beta_k$ , respectively, for  $j = 1, \dots, p$  and  $k = 1, \dots, q$ . Let the resulting SCAD estimators be  $\hat{\alpha}_\lambda$  and  $\hat{\beta}_\lambda$ . Then, we select  $\lambda$  by minimizing the following objective function,

$$(2.8) \quad \text{BIC}(\lambda) = \log\{\text{MSE}(\lambda)\} + (\log(n)/n)\text{DF}_\lambda,$$

where  $\text{MSE}(\lambda) = n^{-1} \sum_{i=1}^n \{Y_i - \hat{\eta}(Z_i^T \hat{\alpha}_\lambda) - X_i^T \hat{\beta}_\lambda\}^2$  and  $\text{DF}_\lambda$  is the number of nonzero coefficients of both  $\hat{\alpha}_\lambda$  and  $\hat{\beta}_\lambda$ . More specifically, we choose  $\lambda$  to be the minimizer among a set of grid points over bounded interval  $[0, \lambda_{max}]$ , where  $\lambda_{max}/\sqrt{n} \rightarrow 0$  as  $n \rightarrow \infty$ . The resulting optimal tuning parameter is denoted by  $\hat{\lambda}$ . In practice, a plot of the  $\text{BIC}(\lambda)$  against  $\lambda$  can be used to determine an appropriate  $\lambda_{max}$  to ensure that the  $\text{BIC}(\lambda)$  reaches its minimum around the middle of the range of  $\lambda$ . Then, the grid points for  $\lambda$  are taken to be evenly distributed over  $[0, \lambda_{max}]$  so that they are chosen to be fine enough to avoid multiple minimizers of  $\text{BIC}(\lambda)$ . In our numerical studies, the range for  $\lambda$  and the number of grid points are set in the same manner as those in Wang, Li and Tsai (2007) and Zhang, Li and Tsai (2010). Based on our limited experience, the resulting estimate of  $\lambda$  is quite stable with respect to the number of grid points when they are sufficiently fine.

To investigate the theoretical properties of the BIC selector, we denote as  $S = \{j_1, \dots, j_d\}$  the set of the indices of the covariates in the given candidate model, which contains indices of both  $X$  and  $Z$ . In addition, let  $S_T$  be the true model,  $S_F$  be the full model, and  $S_\lambda$  be the set of the indices of the covariates selected by the SCAD procedure with tuning parameter  $\lambda$ . For a given candidate model  $S$  with parameter vectors  $\alpha_s$  and  $\beta_s$ , let  $\hat{\alpha}_s$  and  $\hat{\beta}_s$  be the corresponding profile least squares estimators. Then, define  $\sigma_n^2(S) = n^{-1} \sum_{i=1}^n \{Y_i - \hat{\eta}(Z_i^T \hat{\alpha}_s) - X_i^T \hat{\beta}_s\}^2$ , and further assume that

(A) for any  $S \subset S_F$ ,  $\sigma_n^2(S) \rightarrow \sigma^2(S)$  in probability for some  $\sigma^2(S) > 0$ ;

(B) for any  $S \not\subset S_T$ , we have  $\sigma^2(S) > \sigma^2(S_T)$ .

It is noteworthy that (A) and (B) are the standard conditions for investigating parameter estimation under model misspecification (e.g., see Wang, Li and Tsai, 2007).

We next present the asymptotic property of the BIC-type tuning parameter selector.

**Theorem 3** *Under conditions (A) and (B) and the regularity conditions (i)-(vi), we have*

$$(2.9) \quad P(S_\lambda = S_T) \rightarrow 1.$$

Theorem 3 demonstrates that the BIC tuning parameter selector enables us to select the true model consistently.

### 3. Hypothesis Tests

Applying the estimation method described in the previous section, we propose two hypothesis tests. The first one is for general hypothesis testing of regression parameters, and the second one is for testing the nonparametric function.

#### 3.1. Testing parametric components

Consider the general linear hypothesis

$$(3.1) \quad H_0 : \mathbf{A}\boldsymbol{\zeta} = \boldsymbol{\delta} \text{ versus } H_1 : \mathbf{A}\boldsymbol{\zeta} \neq \boldsymbol{\delta},$$

where  $\mathbf{A}$  is a known  $m \times (p + q)$  full-rank matrix and  $\boldsymbol{\delta}$  is an  $m \times 1$  vector. A simple example of (3.1) is to test whether some elements of  $\alpha$  and  $\beta$  are zero; that is,

$$H_0 : \alpha_{i_1} = \dots = \alpha_{i_k} = 0 \text{ and } \beta_{j_1} = \dots = \beta_{j_l} = 0$$

versus

$$H_1 : \text{not all } \alpha_{i_1}, \dots, \alpha_{i_k} \text{ and } \beta_{j_1}, \dots, \beta_{j_l} \text{ are equal to } 0.$$

Under  $H_0$  and  $H_1$ , let  $\zeta_0 = (\boldsymbol{\alpha}_0^\top, \boldsymbol{\beta}_0^\top)^\top$  and  $\zeta_1 = (\boldsymbol{\alpha}_1^\top, \boldsymbol{\beta}_1^\top)^\top$  be the corresponding parameter vectors, and let  $\Omega_0$  and  $\Omega_1$  be the parameter spaces of  $\zeta_0$  and  $\zeta_1$ , respectively. It is noteworthy that this is a slight abuse of notation because  $\zeta_0$  has previously been used to denote the true value of  $\zeta$  in Section 2.1. Furthermore, define

$$Q(H_0) = \inf_{\Omega_0} Q(\zeta_0) = \sum_{i=1}^n \{Y_i - \mathbf{X}_i^\top \hat{\boldsymbol{\beta}}_0 - \hat{\eta}(Z_i^\top \hat{\boldsymbol{\alpha}}_0, \hat{\zeta}_0)\}^2,$$

and

$$Q(H_1) = \inf_{\Omega_1} Q(\zeta_1) = \sum_{i=1}^n \{Y_i - \mathbf{X}_i^\top \hat{\boldsymbol{\beta}}_1 - \hat{\eta}(Z_i^\top \hat{\boldsymbol{\alpha}}_1, \hat{\zeta}_1)\}^2,$$

where  $\{\hat{\boldsymbol{\beta}}_0, \hat{\zeta}_0\}$  and  $\{\hat{\boldsymbol{\beta}}_1, \hat{\zeta}_1\}$  are the profile least squares estimators of  $\{\boldsymbol{\beta}_0, \zeta_0\}$  and  $\{\boldsymbol{\beta}_1, \zeta_1\}$ , respectively, and  $\hat{\eta}$  is the nonparametric estimator of  $\eta$  obtained via (2.3). Subsequently, we propose a test statistic

$$T_1 = \frac{n\{Q(H_0) - Q(H_1)\}}{Q(H_1)},$$

and give its theoretical property below.

**Theorem 4** *Assume that the regularity conditions (i)-(vi) hold. Then,*

- (a) *under  $H_0$  in (3.1),  $T_1 \rightarrow \chi_m^2$ ;*
- (b) *under  $H_1$  in (3.1),  $T_1$  converges to a noncentral chi-squared distribution with  $m$  degrees of freedom and noncentrality parameter  $\phi = \lim_{n \rightarrow \infty} n\sigma^{-2}(\mathbf{A}\boldsymbol{\zeta} - \boldsymbol{\delta})^\top (\mathbf{A}\mathbf{D}^{-1}\mathbf{A}^\top)^{-1}(\mathbf{A}\boldsymbol{\zeta} - \boldsymbol{\delta})$ , where  $\mathbf{D}$  is defined in Theorem 1.*

Analogously, we are able to construct the Wald test,  $W_n = (\mathbf{A}\hat{\boldsymbol{\zeta}} - \boldsymbol{\delta})^\top (\mathbf{A}\hat{\mathbf{D}}^{-1}\mathbf{A}^\top)^{-1}(\mathbf{A}\hat{\boldsymbol{\zeta}} - \boldsymbol{\delta})$ , and demonstrate that  $W_n$  and  $T_1$  have the same asymptotic distribution.

### 3.2. Testing the nonparametric component

The nonparametric estimate of  $\eta(\cdot)$  provides us with descriptive and graphical information for exploratory data analysis. Using this information, it is possible to formulate a parametric model that takes into account the features that emerged from the preliminary analysis. To this

end, we introduce a goodness-of-fit test to assess the appropriateness of a proposed parametric model. Without loss of generality, we consider a simple linear model under the null hypothesis. Accordingly, the null and alternative hypotheses are given as follows:

$$(3.2) \quad H_0 : \eta(u) = \theta_0 + \theta_1 u, \quad \text{versus} \quad H_1 : \eta(u) \neq \theta_0 + \theta_1 u \text{ for some } u,$$

where  $\theta_0$  and  $\theta_1$  are unknown constant parameters.

Under  $H_1$ , let  $\hat{\alpha}$ ,  $\hat{\beta}$ , and  $\hat{\eta}$  be their corresponding profile least squares and nonparametric estimators of  $\alpha$ ,  $\beta$ , and  $\eta$ . Under  $H_0$ , we use the same parametric estimators  $\hat{\alpha}$  and  $\hat{\beta}$  as those obtained under  $H_1$ , while the estimator of  $\eta$  is  $\tilde{\eta}(u) = \hat{\theta}_0 + \hat{\theta}_1 u$ , where  $\hat{\theta}_0$  and  $\hat{\theta}_1$  are the ordinary least squares estimator of  $\theta_0$  and  $\theta_1$ , respectively, by fitting  $Y_i - X_i^T \hat{\beta}$  versus  $Z_i^T \hat{\alpha}$ . Then, the resulting residual sum of squares under the null and alternative hypotheses are

$$\text{RSS}(H_0) = \sum_{i=1}^n \{Y_i - \tilde{\eta}(Z_i^T \hat{\alpha}) - X_i^T \hat{\beta}\}^2$$

and

$$\text{RSS}(H_1) = \sum_{i=1}^n \{Y_i - \hat{\eta}(Z_i^T \hat{\alpha}) - X_i^T \hat{\beta}\}^2.$$

To test the null hypothesis, we consider the following generalized  $F$ -test,

$$T_2 = \frac{r_K}{2} \frac{n \{\text{RSS}(H_0) - \text{RSS}(H_1)\}}{\text{RSS}(H_1)},$$

where  $r_K = \{K(0) - 0.5 \int K^2(u) du\} \{\int \{K(u) - 0.5K * K(u)\} du\}^{-1}$  and  $K * K$  denotes the convolution of  $K$ . The theoretical property of  $T_2$  is given below.

**Theorem 5** *Assume that the regularity conditions (i)-(vi) hold. Under  $H_0$  in (3.2),  $T_2$  has an asymptotic  $\chi^2$  distribution with  $df_n$  degrees of freedom, where  $df_n = r_K |\mathcal{U}| \{K(0) - 0.5 \int K^2(u) du\} / h$ ,  $|\mathcal{U}|$  stands for the length of  $\mathcal{U}$ , and  $\mathcal{U}$  is defined in regularity condition (i).*

The above theorem unveils the Wilks phenomenon for the partially linear single index model. Furthermore, we can obtain an analogous result when the simple linear model under  $H_0$  is replaced by a multiple regression model.

## 4. Simulation Studies and an Empirical Example

In this section, we present four Monte Carlo studies which evaluate the finite sample performance of the proposed estimation and testing methods. The first two examples illustrate

the performance of the profile least squares estimator and the SCAD-based variable selection procedure proposed in Sections 2.1 and 2.2, respectively. The next two examples explore the performance of the test statistics developed in Sections 3.1 and 3.2. Finally, we demonstrate the usefulness of those approaches via a real application.

**Example 4.1** We generated 500 realizations, each consisting of  $n = 50, 100$ , and 200 observations, from each one of the following models:

$$(4.1) \quad y = 4\{(x_1 + x_2 - 1)/\sqrt{2}\}^2 + 4 + 0.2\varepsilon,$$

$$(4.2) \quad y = \sin\{(x_1 + x_2 + x_3)/\sqrt{3} - a\}\pi/(b - a) + \beta Z + 0.1\varepsilon,$$

where  $x_1, x_2, x_3$  are independent and uniformly distributed on  $[0, 1]$ ,  $Z = 0$  for the odd numbered observations and  $Z = 1$  for the even numbered observations,  $\varepsilon$  has the standard normal distribution,  $a = 0.3912$  and  $b = 1.3409$ . The resulting parameters of models (4.1) and (4.2) are  $\boldsymbol{\alpha} = (\alpha_1, \alpha_2)^T = (0.7071, 0.7071)^T$  and  $(\boldsymbol{\alpha}, \beta) = (\alpha_1, \alpha_2, \alpha_3, \beta)^T = (0.5774, 0.5774, 0.5774, 0.3)^T$  with  $\phi_0 = \arccos(\alpha_1) = 0.7854$  as in Xia and Härdle (2006) (or  $\pi/4$  in Härdle, Hall and Ichimura, 1993, p. 165).

Model (4.1) was analyzed by Härdle et al. (1993) and Xia and Härdle (2006), while model (4.2) was investigated by Carroll et al. (1997) and Xia and Härdle (2006). For both models, Xia and Härdle claimed that their MAVE approach outperforms those of Härdle et al. (1993) and Carroll et al. (1997), respectively. It is therefore of interest to compare the profile least squares (abbreviated to PrLS) method with MAVE. Tables 1 and 2 present the results for models (4.1) and (4.2), respectively. Both tables indicate that the profile least squares method yields accurate estimates, and that the mean squared error becomes smaller as the sample size gets larger, which is consistent with the theoretical finding. Furthermore, Table 1 shows that the mean squared errors of  $\hat{\boldsymbol{\alpha}}$  and  $\hat{\boldsymbol{\beta}}$  are smaller than those computed via the MAVE method (see Table 1 of Xia and Härdle, 2006). Table 2 suggests that the biases and their associated mean squared errors of  $\hat{\boldsymbol{\alpha}}$  and  $\hat{\boldsymbol{\beta}}$  are comparable to those calculated via MAVE. In sum, the Monte Carlo studies indicate that PrLS performs well. Because the penalized MAVE is not easy to obtain, we study only the penalized PrLS estimates in the following example.

**Example 4.2** We simulated 500 realizations, each consisting of  $n = 100$  and 200 random samples, from model (4.2) with  $\sigma = 0.1$  and 0.25, respectively. The mean function has coefficients  $\boldsymbol{\alpha} = (1, 3, 1.5, 0.5, 0, 0, 0, 0)^T / \sqrt{12.5}$  and  $\boldsymbol{\beta} = (3, 2, 0, 0, 0, 1.5, 0, 0.2, 0.3, 0.15, 0, 0)^T$ . To assess the robustness of estimates, we further generate the linear and nonlinear covariates from the following three scenarios: (i) The covariate vectors  $\mathbf{X}$  and  $\mathbf{Z}$  have 12 and 8 elements, respectively, which are independent and uniformly distributed on  $[0, 1]$ ; (ii) The covariate vector  $\mathbf{X}$  has 12 elements. The first 5 and last 5 elements are independent and standard normally distributed, while the 6th and 7th elements are independently Bernoulli distributed with success probability 0.5. The covariate vector  $\mathbf{Z}$  has 8 elements, which are independent and standard normally distributed; (iii) A covariate vector  $W$  was generated from a 12-dimensional normal distribution with mean 0 and variance 0.25. The correlation between  $w_i$  and  $w_j$  is  $\rho^{|i-j|}$  with  $\rho = 0.4$ . Then, the covariate vector  $\mathbf{X} = W + \{1.5 \exp(1.5z_1), 5z_1, 5\sqrt{z_2}, 3z_1 + z_2^2, 0, 0, 0, 0, 0, 0, 0, 0\}^T$ . Moreover, the covariate vector  $\mathbf{Z}$  has 8 elements, which are independent and uniformly distributed on  $[0, 1]$ .

Based on the above model settings, we next explore the performance of the penalized profile least squares approach via SCAD-BIC. Because the Akaike information criterion (Akaike, 1973) has been commonly used for classical variable selections, we also study SCAD-AIC by replacing  $\log(n)$  in (2.8) with  $2(p+q)$ . To assess the performance, we consider Fan and Li's (2001) median of relative model error (MRME), where the relative model errors is defined as  $\text{RME} = \text{ME} / \text{ME}_{\text{SF}}$ ,  $\text{ME} = E[\{\hat{\eta}(Z^T \hat{\boldsymbol{\alpha}}_{\hat{\lambda}}) + X^T \hat{\boldsymbol{\beta}}_{\hat{\lambda}}\} - \{\eta(Z^T \boldsymbol{\alpha}) + X^T \boldsymbol{\beta}\}]^2$ , and  $\text{ME}_{\text{SF}}$  is the model error calculated by fitting the data with the full model via the unpenalized estimates. In addition, we calculate the average number of the eight true zero coefficients that were correctly set to zero, and the average number of the four truly nonzero coefficients that were incorrectly set to zero. Table 3 shows that the SCAD-BIC outperforms SCAD-AIC in terms of model error measures. Moreover, SCAD-BIC has a much better rate of correctly identifying the true submodel than that of SCAD-AIC, although it sometimes shrinks slightly nonzero coefficients to zero. Unsurprisingly, SCAD-BIC improves as the signal gets stronger and the sample size becomes larger, which corroborates our theoretical findings.

**Example 4.3** To study the finite sample performance of the test statistic  $T_1$  in Section 3.1, we

consider the same model as that of scenario (i) in Example 4.2. Due to the model's parameter setting, we naturally consider the following null and alternative hypotheses:

$$H_0 : \beta_3 = \beta_4 = \beta_5 = \beta_7 = 0 \quad \text{versus} \quad H_1 : \beta_3 = \beta_4 = \beta_5 = \beta_7 = c_1,$$

where  $c_1$  ranges from 0 to 0.1 with increment 0.01 for  $\sigma = 0.1$ , whereas  $c_1$  is the value from a set of  $\{0, 0.01, 0.02, \dots, 0.09, 0.15, 0.2\}$  for  $\sigma = 0.25$ . In addition, 500 realizations were generated with  $n = 200$  to calculate the size and power of  $T_1$ . Figure 1 depicts the power function versus  $c_1$ . It shows that the empirical size at  $c_1 = 0$  is very close to the nominal level 0.05. Furthermore, the power of the test is greater than 0.95 as  $c_1$  increases to 0.05 and 0.15, respectively, when  $\sigma = 0.1$  and  $\sigma = 0.25$ . It is not surprising that the power increases as the signal gets stronger. In sum,  $T_1$  not only controls the size well, but is also a powerful test.

**Example 4.4** To examine the performance of the test statistic  $T_2$  in Section 3.2, we generated 500 realizations from the model given below with  $n = 200$ .

$$(4.3) \quad y = \eta\{(z_1 + z_2 + z_3)/\sqrt{3}\} - 0.5x_1 + 0.3x_2 + \sigma\varepsilon,$$

where  $\sigma = 0.1$  and  $\sigma = 0.25$ , respectively. Then, we consider the following hypotheses,

$$H_0 : \eta(u) = u, \quad \text{versus} \quad H_1 : \eta(u) = c_2 \sin\{\pi(u - a)/(b - a)\} + u,$$

where  $c_2$  ranges from 0 to 0.1 with increment 0.025 for  $\sigma = 0.1$ , while  $c_2$  ranges from 0 to 0.2 with increment 0.05 for  $\sigma = 0.25$ . Figure 2 demonstrates that the empirical size at  $c_2 = 0$  is very close to the nominal level 0.05. Furthermore, the power of the test is greater than 0.95 as  $c_2$  increases to 0.075 and 0.2, respectively. As expected, the power increases when the signal becomes stronger. Although  $T_2$  is slightly less powerful than  $T_1$ , it controls the size well and is a reliable test.

**Example 4.5** To illustrate the usefulness of the proposed methods in real applications, we analyze an automobile data set (see Johnson, 2003) to discover the factors that influence the price. The suggested retail price (the manufacturer's assessment of the vehicle's worth, including adequate profit for the automaker and the dealer) serves as the response variable,  $y$ . The type of vehicle is measured by 4 explanatory variables: sports car,  $x_1$  (1=yes, 0=no), sport utility vehicle,  $x_2$  (1=yes, 0=no), wagon,  $x_3$  (1=yes, 0=no), and minivan,  $x_4$  (1=yes, 0=no).

In addition, two binary variables  $x_5$  (1=yes, 0=no) and  $x_6$  (1=yes, 0=no) are used to indicate whether the car/truck is all-wheel drive and rear-wheel drive, respectively. Since the number of cylinders takes the values 3, 4, 5, 6, 8 and 12, we created 5 binary variables  $x_7, x_8, x_9, x_{10}$  and  $x_{11}$  to indicate whether the number of cylinders is 3, 5, 6, 8 and 12, respectively. Moreover, there are 7 additional measurements that include engine size,  $z_1$ , horsepower,  $z_2$ , the average of city miles per gallon and highway miles per gallon,  $z_3$ , weight in pounds,  $z_4$ , wheel base in inches,  $z_5$ , length in inches,  $z_6$ , and width in inches,  $z_7$ . We have a total of 18 explanatory variables and 386 observations by removing 41 missing values and one large outlier from the original data set with 428 observations.

We fit the data via the model (2.1) with SCAD-BIC. The resulting model is

$$\hat{y} = \hat{\eta} \left( \sum_{j=1 \sim 7} z_j \hat{\alpha}_{\hat{\lambda}, j} \right) + \sum_{k=1, 6, 8 \sim 11} x_k \hat{\beta}_{\hat{\lambda}, k},$$

where  $(\hat{\alpha}_{\hat{\lambda}, 1}, \hat{\alpha}_{\hat{\lambda}, 2}, \hat{\alpha}_{\hat{\lambda}, 3}, \hat{\alpha}_{\hat{\lambda}, 4}, \hat{\alpha}_{\hat{\lambda}, 5}, \hat{\alpha}_{\hat{\lambda}, 6}, \hat{\alpha}_{\hat{\lambda}, 7})^T = (-0.168, 0.634, 0.189, 0.565, -0.224, 0.216, -0.343)^T$ ,

$(\hat{\beta}_{\hat{\lambda}, 1}, \hat{\beta}_{\hat{\lambda}, 6}, \hat{\beta}_{\hat{\lambda}, 8}, \hat{\beta}_{\hat{\lambda}, 9}, \hat{\beta}_{\hat{\lambda}, 10}, \hat{\beta}_{\hat{\lambda}, 11})^T = (0.386, 0.397, 0.448, 0.329, 0.626, 1.728)^T$ , and  $\hat{\lambda} = 0.044$ .

The standard errors of the parameter estimates are  $(0.069, 0.159, 0.074, 0.162, 0.088, 0.086, 0.098)^T$

and  $(0.096, 0.066, 0.172, 0.040, 0.076, 0.276)^T$ , respectively. It shows that all continuous variables

$z_1$  to  $z_7$  are important, while the significantly discrete variables are  $x_1, x_6, x_8$  to  $x_{11}$ . Hence,

the type of vehicle (except for the sports car) does not play an important role for retail price

when the information of seven attributes is available. This discovery is not surprising because

the sports car is intrinsically different from the other types of cars and its contribution cannot

be substituted by the seven continuous measurements. In addition, the rear-wheel provides

more useful information than all-wheel drive. Moreover, the 3 or 4 cylinder makes no significant

difference for deciding the price.

It is of interest to employ the test statistics  $T_1$  and  $T_2$  to examine the above model. As a result,

the  $p$  value of  $T_1$  for testing  $H_0 : \beta_2 = \beta_3 = \beta_4 = \beta_5 = \beta_7 = 0$  is 0.0556, which is consistent

with SCAD-BIC. Because it is slightly larger than 0.05, we conduct further analysis and find

that variable  $x_2$  (sport utility vehicle) may provide some useful information even though it is a

generic marketing description for a vehicle similar to a station wagon but built on a light-truck

chassis. Furthermore, the  $p$  value of  $T_2$  for testing  $H_0 : \eta(u) = u$  is 0.001. This indicates the

linear fitting is not appropriate, which can be seen in Figure 3. This figure also shows that a

larger single index composed from the car's efficiency and power yields a higher retail price. In conclusion, the above model is sensible and useful for the manufacturer's suggested retail price.

## 5. Discussion

In partially linear single-index models, we propose using the SCAD approach to shrink parameters contained in both parametric and nonparametric components. The resulting estimators enjoy the oracle property when the regularization parameters satisfy the proper conditions. To further exploit SCAD, one could extend the current results to partially linear multiple-index models by allowing  $\varepsilon$  to be dependent on  $(Z^T, X^T)^T$ . In addition, one could obtain the SCAD estimator for generalized partially linear single-index models. Finally, an investigation of partially linear single-index model selection with error-prone covariates could also be of interest. We believe that these efforts would enhance the usefulness of SCAD in data analysis.

## Appendix: Proofs of Theorems

**Proof of Theorem 1.** Under the conditions of Theorem 1, we follow similar arguments to those used by Ichimura (1993) and show that  $\widehat{\zeta} = (\widehat{\alpha}^T, \widehat{\beta}^T)^T$  is a root- $n$  consistent estimator of  $\zeta$ . Because the proof is straightforward, we do not present it here. We next demonstrate the asymptotic normality of  $\widehat{\zeta}$  by using a general result of Newey (1994).

Let  $m_x(\Lambda) = E(X|\Lambda)$ ,  $m_z(\Lambda) = E(Z|\Lambda)$ , and  $\kappa = \eta'(\Lambda)\{Z - m_z(\Lambda)\}$ . In addition, let

$$(A.1) \quad \Psi(m_x, \eta, \kappa, \alpha, \beta, Y, Z, X) = (Y - \eta - X^T \beta) \begin{Bmatrix} \kappa \\ X - m_x(\Lambda) \end{Bmatrix}.$$

For any given  $m_x^*$ ,  $\eta^*$ , and  $\kappa^*$ , define

$$\begin{aligned} D(m_x^* - m_x, \eta^* - \eta, \kappa^* - \kappa, \alpha, \beta, Y, Z, X) \\ = \frac{\partial \Psi}{\partial m_x}(m_x^* - m_x) + \frac{\partial \Psi}{\partial \eta}(\eta^* - \eta) + \frac{\partial \Psi}{\partial \kappa}(\kappa^* - \kappa), \end{aligned}$$

where the partial derivatives are the Frechet partial derivatives. After algebraic simplification, we have

$$\frac{\partial \Psi}{\partial m_x} = (Y - \eta - X^T \beta) \begin{pmatrix} 0 \\ -1 \end{pmatrix},$$

$$\begin{aligned}\frac{\partial \Psi}{\partial \eta} &= - \left\{ \begin{array}{c} \kappa \\ X - m_x(\Lambda) \end{array} \right\}, \\ \frac{\partial \Psi}{\partial \kappa} &= (Y - \eta - X^T \boldsymbol{\beta}) \begin{pmatrix} 1 \\ 0 \end{pmatrix},\end{aligned}$$

where the partial derivatives are zero. Accordingly,

$$\begin{aligned}(\text{A.2}) \quad & \|\Psi(m_x^*, \eta^*, \kappa^*, \boldsymbol{\alpha}, \boldsymbol{\beta}, Y, Z, X) - \Psi(m_x, \eta, \kappa, \boldsymbol{\alpha}, \boldsymbol{\beta}, Y, Z, X) \\ & - D(m_x^* - m_x, \eta^* - \eta, \kappa^* - \kappa, \boldsymbol{\alpha}, \boldsymbol{\beta}, Y, Z, X)\| \\ & = O\left(\|m_x^* - m_x\|^2 + \|\eta^* - \eta\|^2 + \|\kappa^* - \kappa\|^2\right),\end{aligned}$$

where  $\|\cdot\|$  denotes the Sobolev norm, i.e, the supremum norm of the function itself as well as its derivatives. Equation (A.2) is Newey's assumption 5.1(i). It is also noteworthy that his assumption 5.2 holds by the expression of  $D(\cdot, \cdot, \cdot, \boldsymbol{\alpha}, \boldsymbol{\beta}, Y, Z, X)$ . Moreover, the result

$$E \{D(m_x^* - m_x, \eta^* - \eta, \kappa^* - \kappa, \boldsymbol{\alpha}, \boldsymbol{\beta}, Y, Z, X)\} = 0$$

leads to Newey's assumption 5.3.

In addition to Newey's assumptions mentioned above, we need to verify one more assumption before employing his result. To this end, we re-express the solution of (2.2) as

$$\hat{\eta}(u) = \frac{1}{n} \sum_{i=1}^n \frac{\{\hat{s}_2(u, h) - \hat{s}_1(u, h)(\Lambda_i - u)\} K_h(\Lambda_i - u)(Y_i - X_i^T \boldsymbol{\beta})}{\hat{s}_2(u, h)\hat{s}_0(u, h) - \hat{s}_1^2(u, h)},$$

where  $\Lambda_i$  is the  $i$ -th row of  $\Lambda$  and

$$\hat{s}_r(u, h) = \frac{1}{n} \sum_{i=1}^n (\Lambda_i - u)^r K_h(\Lambda_i - u), \text{ for } r = 0, 1, 2.$$

Then, let

$$\begin{aligned}\hat{m}_x(u) &= \frac{1}{n} \sum_{i=1}^n \frac{\{\hat{s}_2(u, h) - \hat{s}_1(u, h)(\Lambda_i - u)\} K_h(\Lambda_i - u) X_i^T}{\hat{s}_2(u, h)\hat{s}_0(u, h) - \hat{s}_1^2(u, h)}, \\ \hat{m}_z(u) &= \frac{1}{n} \sum_{i=1}^n \frac{\{\hat{s}_2(u, h) - \hat{s}_1(u, h)(\Lambda_i - u)\} K_h(\Lambda_i - u) Z_i^T}{\hat{s}_2(u, h)\hat{s}_0(u, h) - \hat{s}_1^2(u, h)}.\end{aligned}$$

Applying similar techniques to those used in Mack and Silverman (1982), we obtain the following equations, which hold uniformly in  $u \in \mathcal{U}$ .

$$\begin{aligned}(\text{A.3}) \quad & \hat{\eta}(u) - \eta(u) = o_p(n^{-1/4}), \quad \hat{\eta}'(u) - \eta'(u) = o_p(n^{-1/4}), \\ & \hat{m}_x(u) - m_x(u) = o_p(n^{-1/4}), \quad \text{and} \quad \hat{m}_z(u) - m_z(u) = o_p(n^{-1/4}).\end{aligned}$$

These results imply  $\widehat{\kappa} - \kappa = o_p(n^{-1/4})$ . Thus, Newey's assumption 5.1(ii) holds.

After examining Newey's assumptions 5.1 to 5.3, we apply his Lemma 5.1 and find that  $\widehat{\zeta}$  has the same limit distribution as the solution to the equation

$$(A.4) \quad 0 = \sum_{i=1}^n \Psi(m_x, \eta, \kappa, \boldsymbol{\alpha}, \boldsymbol{\beta}, Y_i, Z_i, X_i).$$

Furthermore, it is easy to show that the solution to (A.4) has the same limit distribution as described in the statement of Theorem 1. Hence, we complete the proof for asymptotic normality.

Finally, we show the efficiency of  $\widehat{\zeta}$ . Let  $p_\varepsilon(\varepsilon)$  be the probability density function of  $\varepsilon$  and  $p'_\varepsilon(\varepsilon)$  be its first-order derivative with respect to  $\varepsilon$ . Then, the score function of  $(\boldsymbol{\alpha}, \boldsymbol{\beta})$  is

$$S_{(\boldsymbol{\alpha}, \boldsymbol{\beta})} = -\sigma^2 \begin{Bmatrix} \eta'(\Lambda)Z \\ X \end{Bmatrix} \frac{p'_\varepsilon(\varepsilon)}{p_\varepsilon(\varepsilon)}.$$

For any given function  $g$  of  $(Z, X)$ , it can be shown that the nuisance tangent space  $\mathcal{P}$ , for the three nuisance parameters,  $g_{(Z, X)}(z, x)$ ,  $p_\varepsilon(\varepsilon)$ , and  $\eta(\Lambda)$ , is  $\{g(Z, X) : E(g) = 0, E(\varepsilon g)$  is a function of  $(Z, X)$  only $\}$ . Furthermore, the orthogonal component of  $\mathcal{P}$  is

$$\mathcal{P}^\perp = \{\varepsilon g(Z, X) : E(g|\Lambda) = 0\}.$$

Subsequently, we apply Bickel et al.'s (1993) approach and obtain the following semiparametric efficient score function via equation (2.4),

$$(A.5) \quad S_{eff} = \varepsilon \begin{Bmatrix} \eta'(\Lambda)\tilde{Z} \\ \tilde{X} \end{Bmatrix}.$$

It can be seen that  $S_{eff} \in \mathcal{P}^\perp$ .

For any  $\varepsilon g \in \mathcal{P}^\perp$ , we have  $E(g|\Lambda) = 0$ . Accordingly,

$$\begin{aligned} E\{(S_{(\boldsymbol{\alpha}, \boldsymbol{\beta})} - S_{eff})^\top \varepsilon g(Z, X)\} &= E\left(-\sigma^2 \begin{Bmatrix} \eta'(\Lambda)Z \\ X \end{Bmatrix}^\top g E\left\{\frac{\varepsilon p'_\varepsilon(\varepsilon)}{p_\varepsilon(\varepsilon)}\right\}\right. \\ &\quad \left.- E(\varepsilon^2) \left[ \begin{Bmatrix} \eta'(\Lambda)Z \\ X \end{Bmatrix}^\top g - E\left\{ \begin{Bmatrix} \eta'(\Lambda)E(Z|\Lambda) \\ E(X|\Lambda) \end{Bmatrix}^\top g \right\} \right] \right). \end{aligned}$$

Because  $E\{\varepsilon p'_\varepsilon(\varepsilon)/p_\varepsilon(\varepsilon)\} = -1$ , it follows that

$$E\{(S_{(\boldsymbol{\alpha}, \boldsymbol{\beta})} - S_{eff})^\top \varepsilon g(Z, X)\} = E[\{\eta'(\Lambda)E(Z^\top|\Lambda), E(X^\top|\Lambda)\}E(g|\Lambda)] = 0.$$

That is,  $S_{eff}$  is the projection of  $S_{(\boldsymbol{\alpha}, \boldsymbol{\beta})}$  to  $\mathcal{P}^\perp$ , and the estimator  $\hat{\boldsymbol{\zeta}}$  is therefore efficient (see Bickel et al., 1993). We complete the proof.

**Proof of Theorem 2.** To prove this theorem, we consider the following three steps: Step I establishes the order of the minimizer  $(\hat{\boldsymbol{\alpha}}_{\lambda_1}^\top, \hat{\boldsymbol{\beta}}_{\lambda_2}^\top)^\top$  of  $\mathcal{L}_P(\boldsymbol{\alpha}, \boldsymbol{\beta})$ ; Step II shows that  $(\hat{\boldsymbol{\alpha}}_{\lambda_1}^\top, \hat{\boldsymbol{\beta}}_{\lambda_2}^\top)^\top$  attains sparsity; Step III derives the asymptotic distribution of the penalized estimators.

**Step I.** Let  $\gamma_n = n^{-1/2} + a_n + c_n$ ,  $\mathbf{v}_1 = (v_{11}, \dots, v_{1p})^\top$ ,  $\mathbf{v}_2 = (v_{21}, \dots, v_{2q})^\top$ , and  $\|\mathbf{v}_1\| = \|\mathbf{v}_2\| = C$  for some positive constant  $C$ , where  $a_n = \max_{1 \leq j \leq p} \{ |p'_{\lambda_{1j}}(|\alpha_{0j}|)|, \alpha_{0j} \neq 0 \}$ ,  $c_n = \max_{1 \leq k \leq q} \{ |p'_{\lambda_{2k}}(|\beta_{0k}|)|, \beta_{0k} \neq 0 \}$ , and  $\alpha_{0j}$  and  $\beta_{0k}$  are the  $j$ -th and  $k$ -th elements of  $\boldsymbol{\alpha}_0$  and  $\boldsymbol{\beta}_0$ , respectively, for  $j = 1, \dots, p$  and  $k = 1, \dots, q$ . Furthermore, define

$$\begin{aligned} D_{n,1} &= \sum_{i=1}^n \{ Y_i - \hat{\eta}(Z_i^\top(\boldsymbol{\alpha}_0 + \gamma_n \mathbf{v}_1), \boldsymbol{\alpha}_0 + \gamma_n \mathbf{v}_1, \boldsymbol{\beta}_0 + \gamma_n \mathbf{v}_2) - X_i^\top(\boldsymbol{\beta}_0 + \gamma_n \mathbf{v}_2) \}^2 \\ &\quad - \sum_{i=1}^n \{ Y_i - \hat{\eta}(X_i^\top \boldsymbol{\alpha}_0, \boldsymbol{\alpha}_0, \boldsymbol{\beta}_0) - X_i^\top \boldsymbol{\beta}_0 \}^2 \end{aligned}$$

and

$$D_{n,2} = -n \sum_{j=1}^{p_0} \{ p_{\lambda_{1j}}(|\alpha_{0j} + \gamma_n v_{1j}|) - p_{\lambda_{1j}}(|\alpha_{0j}|) \} - n \sum_{k=1}^{q_0} \{ p_{\lambda_{2k}}(|\beta_{0k} + \gamma_n v_{2k}|) - p_{\lambda_{2k}}(|\beta_{0k}|) \}.$$

After algebraic simplification, we have

$$\begin{aligned} D_{n,1} &= \sum_{i=1}^n \{ \hat{\eta}(Z_i^\top(\boldsymbol{\alpha}_0 + \gamma_n \mathbf{v}_1), \boldsymbol{\alpha}_0 + \gamma_n \mathbf{v}_1, \boldsymbol{\beta}_0 + \gamma_n \mathbf{v}_2) - \hat{\eta}(Z_i^\top \boldsymbol{\alpha}_0, \boldsymbol{\alpha}_0, \boldsymbol{\beta}_0) + X_i^\top \gamma_n \mathbf{v}_2 \} \\ &\quad \{ \hat{\eta}(Z_i^\top(\boldsymbol{\alpha}_0 + \gamma_n \mathbf{v}_1), \boldsymbol{\alpha}_0 + \gamma_n \mathbf{v}_1, \boldsymbol{\beta}_0 + \gamma_n \mathbf{v}_2) \\ &\quad + \hat{\eta}(Z_i^\top \boldsymbol{\alpha}_0, \boldsymbol{\alpha}_0, \boldsymbol{\beta}_0) + X_i^\top(\boldsymbol{\beta}_0 + \gamma_n \mathbf{v}_2) + X_i^\top \boldsymbol{\beta}_0 - 2Y_i \} \\ \text{(A.6)} \quad &= \sum_{i=1}^n \{ \tilde{Z}_i^\top \eta'(\Lambda_i) \mathbf{v}_1 \gamma_n + \tilde{X}_i^\top \mathbf{v}_2 \gamma_n \}^2 - \sum_{i=1}^n \{ \tilde{Z}_i^\top \eta'(\Lambda_i) \mathbf{v}_1 \gamma_n + \tilde{X}_i^\top \mathbf{v}_2 \gamma_n \} \varepsilon_i + o_p(1). \end{aligned}$$

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

$$\sqrt{p_0} \gamma_n a_n \|\mathbf{v}_1\| + \gamma_n^2 b_n \|\mathbf{v}_1\|^2 + \sqrt{q_0} \gamma_n c_n \|\mathbf{v}_2\| + \gamma_n^2 d_n \|\mathbf{v}_2\|^2 \leq C \gamma_n^2 (\sqrt{p_0} + b_n C + \sqrt{q_0} + d_n C),$$

where  $b_n = \max_{1 \leq j \leq p} \{ |p''_{\lambda_{1j}}(|\alpha_{0j}|)|, \alpha_{0j} \neq 0 \}$  and  $d_n = \max_{1 \leq k \leq q} \{ |p''_{\lambda_{2k}}(|\beta_{0k}|)|, \beta_{0k} \neq 0 \}$ . When  $b_n$  and  $d_n$  tend to 0 and  $C$  is sufficiently large, the first term on the right-hand side of (A.6) dominates the second term on the right-hand side of (A.6) and  $D_{n,2}$ . As a result, for any given  $\nu > 0$ , there exists a large constant  $C$  such that

$$P \left\{ \inf_{\mathcal{V}_{12}} \mathcal{L}_P(\boldsymbol{\alpha}_0 + \gamma_n \mathbf{v}_1, \boldsymbol{\beta}_0 + \gamma_n \mathbf{v}_2) > \mathcal{L}_P(\boldsymbol{\alpha}_0, \boldsymbol{\beta}_0) \right\} \geq 1 - \nu,$$

where  $\mathcal{V}_{12} = \{(\mathbf{v}_1, \mathbf{v}_2) : \|\mathbf{v}_1\| = C, \|\mathbf{v}_2\| = C\}$ . We therefore conclude that the rate of convergence of  $(\widehat{\boldsymbol{\alpha}}_{\lambda_1}^\top, \widehat{\boldsymbol{\beta}}_{\lambda_2}^\top)^\top$  is  $O_P(n^{-1/2} + a_n + c_n)$ .

**Step II.** Let  $\boldsymbol{\alpha}_1$  and  $\boldsymbol{\beta}_1$  satisfy  $\|\boldsymbol{\alpha}_1 - \boldsymbol{\alpha}_{10}\| = O_P(n^{-1/2})$  and  $\|\boldsymbol{\beta}_1 - \boldsymbol{\beta}_{10}\| = O_P(n^{-1/2})$ , respectively. We next show that

$$(A.7) \quad \mathcal{L}_P \left\{ \begin{pmatrix} \boldsymbol{\alpha}_1 \\ \mathbf{0} \end{pmatrix}, \begin{pmatrix} \boldsymbol{\beta}_1 \\ \mathbf{0} \end{pmatrix} \right\} = \min_{\mathcal{C}} \mathcal{L}_P \left\{ \begin{pmatrix} \boldsymbol{\alpha}_1 \\ \boldsymbol{\alpha}_2 \end{pmatrix}, \begin{pmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \end{pmatrix} \right\},$$

where  $\mathcal{C} = \{\|\boldsymbol{\alpha}_2\| \leq C^*n^{-1/2}, \|\boldsymbol{\beta}_2\| \leq C^*n^{-1/2}\}$  and  $C^*$  is a positive constant.

Consider  $\beta_k \in (-C^*n^{-1/2}, C^*n^{-1/2})$  for  $k = q_0 + 1, \dots, q$ . When  $\beta_k \neq 0$ , we have  $\partial \mathcal{L}_P(\boldsymbol{\alpha}, \boldsymbol{\beta}) / \partial \beta_k = \ell_k(\boldsymbol{\alpha}, \boldsymbol{\beta}) + np'_{\lambda_{2k}}(|\beta_k|) \text{sgn}(\beta_k)$ , where

$$\begin{aligned} \ell_k(\boldsymbol{\alpha}, \boldsymbol{\beta}) &= \frac{1}{n} \sum_{i=1}^n \{Y_i - \widehat{\eta}(Z_i^\top \boldsymbol{\alpha}, \boldsymbol{\alpha}, \boldsymbol{\beta}) + X_i^\top \boldsymbol{\beta}\} \left\{ X_{ik} + \frac{\partial \widehat{\eta}(Z_i^\top \boldsymbol{\alpha}, \boldsymbol{\alpha}, \boldsymbol{\beta})}{\partial \beta_k} \right\} \\ &= \frac{1}{n} \sum_{i=1}^n \{ \widehat{\eta}(Z_i^\top \boldsymbol{\alpha}, \boldsymbol{\alpha}, \boldsymbol{\beta}) - \eta(Z_i^\top \boldsymbol{\alpha}_0) + X_i^\top (\boldsymbol{\beta} - \boldsymbol{\beta}_0) - \varepsilon_i \} \left\{ X_{ik} + \frac{\partial \widehat{\eta}(Z_i^\top \boldsymbol{\alpha}, \boldsymbol{\alpha}, \boldsymbol{\beta})}{\partial \beta_k} \right\}. \end{aligned}$$

Applying similar arguments to those used in the proof of Theorem 5.2 of Ichimura (1993) together with algebraic simplifications, the above term can be expressed as

$$\frac{1}{n} \sum_{i=1}^n (\boldsymbol{\alpha} - \boldsymbol{\alpha}_0)^\top \widetilde{Z}_i \eta'(\Lambda_i) \widetilde{X}_{ik} + \frac{1}{n} \sum_{i=1}^n (\boldsymbol{\beta} - \boldsymbol{\beta}_0)^\top \widetilde{X}_i \widetilde{X}_{ik} + o_p(n^{-1/2}).$$

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

$$\frac{\partial \mathcal{L}_P(\boldsymbol{\alpha}, \boldsymbol{\beta})}{\partial \beta_k} = -n \lambda_{2k} \{ \lambda_{2k}^{-1} p'_{\lambda_{2k}}(|\beta_k|) \text{sgn}(\beta_k) + O_P(n^{-1/2} \lambda_{2k}^{-1}) \}.$$

Because  $\liminf_{n \rightarrow \infty} \liminf_{\beta_j \rightarrow 0^+} \lambda_{2k}^{-1} p'_{\lambda_{2k}}(|\beta_k|) > 0$  and  $n^{-1/2} / \lambda_{2k} \rightarrow 0$ ,  $\partial \mathcal{L}_P(\boldsymbol{\alpha}, \boldsymbol{\beta}) / \partial \beta_k$  and  $\beta_k$  have different signs for  $\beta_k \in (-C^*n^{-1/2}, C^*n^{-1/2})$ . Analogously, we can show that  $\partial \mathcal{L}_P(\boldsymbol{\alpha}, \boldsymbol{\beta}) / \partial \alpha_j$  and  $\alpha_j$  have different signs when  $\alpha_j \in (-C^*n^{-1/2}, C^*n^{-1/2})$  for  $j = p_0 + 1, \dots, p$ . Consequently, the minimum is attained at  $\boldsymbol{\alpha}_2 = \mathbf{0}$  and  $\boldsymbol{\beta}_2 = \mathbf{0}$ . This completes the proof of (A.7).

**Step III.** Finally, we demonstrate the asymptotic normality of  $\widehat{\boldsymbol{\alpha}}_{\lambda_{11}}$  and  $\widehat{\boldsymbol{\beta}}_{\lambda_{21}}$ . For the sake of simplicity, we define

$$\mathbf{R}_{\lambda_1} = \{p'_{\lambda_{11}}(|\alpha_{01}|) \text{sgn}(\alpha_{01}), \dots, p'_{\lambda_{1p_0}}(|\alpha_{0p_0}|) \text{sgn}(\alpha_{0p_0})\}^\top,$$

$$\Sigma_{\lambda_1} = \text{diag}\{p''_{\lambda_{11}}(|\alpha_{01}|), \dots, p''_{\lambda_{1p_0}}(|\alpha_{0p_0}|)\},$$

$$\mathbf{R}_{\lambda_2} = \{p'_{\lambda_{21}}(|\beta_{01}|) \text{sgn}(\beta_{01}), \dots, p'_{\lambda_{2q_0}}(|\beta_{0q_0}|) \text{sgn}(\beta_{0q_0})\}^\top,$$

$$\Sigma_{\lambda_2} = \text{diag}\{p''_{\lambda_{21}}(|\beta_{01}|), \dots, p''_{\lambda_{2q_0}}(|\beta_{0q_0}|)\},$$

where  $\alpha_{0j}$  and  $\beta_{0k}$  are the  $j$ -th and  $k$ -th elements of  $\boldsymbol{\alpha}_{10}$  and  $\boldsymbol{\beta}_{10}$ , respectively, for  $j = 1, \dots, p_0$  and  $k = 1, \dots, q_0$ . It follows from (2.5) that  $\widehat{\boldsymbol{\alpha}}_{\lambda_{11}}$  and  $\widehat{\boldsymbol{\beta}}_{\lambda_{21}}$  satisfy

$$(A.8) \quad \mathbf{0} = \left\{ \begin{array}{c} \frac{\partial \mathcal{L}_P(\widehat{\boldsymbol{\alpha}}_{\lambda_{11}}, \widehat{\boldsymbol{\beta}}_{\lambda_{21}})}{\partial \boldsymbol{\alpha}_1} \\ \frac{\partial \mathcal{L}_P(\widehat{\boldsymbol{\alpha}}_{\lambda_{11}}, \widehat{\boldsymbol{\beta}}_{\lambda_{21}})}{\partial \boldsymbol{\beta}_1} \end{array} \right\} = l(\widehat{\boldsymbol{\alpha}}_{\lambda_{11}}, \widehat{\boldsymbol{\beta}}_{\lambda_{21}}) + \left\{ \begin{array}{c} \mathbf{R}_{\lambda_1} - \Sigma_{\lambda_1}(\widehat{\boldsymbol{\alpha}}_{\lambda_{11}} - \boldsymbol{\alpha}_{10}) \\ \mathbf{R}_{\lambda_2} - \Sigma_{\lambda_2}(\widehat{\boldsymbol{\beta}}_{\lambda_{21}} - \boldsymbol{\beta}_{10}) \end{array} \right\},$$

where

$$l(\widehat{\boldsymbol{\alpha}}_{\lambda_{11}}, \widehat{\boldsymbol{\beta}}_{\lambda_{21}}) = \frac{1}{n} \sum_{i=1}^n \{Y_i - \widehat{\eta}(\Lambda_i, \widehat{\boldsymbol{\zeta}}_{\lambda_{11}}) - X_{i,1}^T \widehat{\boldsymbol{\beta}}_{\lambda_{21}}\} \left\{ \begin{array}{c} \frac{\partial \widehat{\eta}(\widehat{\Lambda}_i, \widehat{\boldsymbol{\zeta}}_{\lambda_{11}})}{\partial \boldsymbol{\alpha}} \\ \frac{\partial \widehat{\eta}(\widehat{\Lambda}_i, \widehat{\boldsymbol{\zeta}}_{\lambda_{11}})}{\partial \boldsymbol{\beta}} + X_{i,1} \end{array} \right\},$$

where  $\widehat{\Lambda}_i$ , and  $X_{i,1}$  are the  $i$ -th rows of  $\widehat{\Lambda}$ , and  $X_1$ , respectively, and  $\widehat{\boldsymbol{\zeta}}_{\lambda_{11}} = (\widehat{\boldsymbol{\alpha}}_{\lambda_{11}}^T, \widehat{\boldsymbol{\beta}}_{\lambda_{21}}^T)^T$  is the penalized least squares estimator of  $\boldsymbol{\zeta}_1 = (\boldsymbol{\alpha}_1^T, \boldsymbol{\beta}_1^T)^T$ .

Applying the Taylor expansion, we obtain

$$\begin{aligned} l(\widehat{\boldsymbol{\alpha}}_{\lambda_{11}}, \widehat{\boldsymbol{\beta}}_{\lambda_{21}}) &= \frac{1}{n} \sum_{i=1}^n \{Y_i - \widehat{\eta}(\Lambda_i, \boldsymbol{\zeta}_1) - X_{i,1}^T \boldsymbol{\beta}_1\} \left\{ \begin{array}{c} \frac{\partial \widehat{\eta}(\Lambda_i, \boldsymbol{\zeta}_1)}{\partial \boldsymbol{\alpha}_1} \\ \frac{\partial \widehat{\eta}(\Lambda_i, \boldsymbol{\zeta}_1)}{\partial \boldsymbol{\beta}_1} + X_{i,1} \end{array} \right\} \\ &\quad - \frac{1}{n} \sum_{i=1}^n \left\{ \begin{array}{c} \frac{\partial \widehat{\eta}(\bar{\Lambda}_i, \bar{\boldsymbol{\zeta}}_1)}{\partial \boldsymbol{\alpha}_1} \\ \frac{\partial \widehat{\eta}(\bar{\Lambda}_i, \bar{\boldsymbol{\zeta}}_1)}{\partial \boldsymbol{\beta}_1} + X_{i,1} \end{array} \right\}^{\otimes 2} (\widehat{\boldsymbol{\zeta}}_{\lambda_{11}} - \boldsymbol{\zeta}_1) \\ &\quad - \frac{1}{n} \sum_{i=1}^n \{Y_i - \widehat{\eta}(\bar{\Lambda}_i, \bar{\boldsymbol{\zeta}}_1) - X_{i,1}^T \bar{\boldsymbol{\beta}}_1\} \frac{\partial^2 \widehat{\eta}(\bar{\Lambda}_i, \bar{\boldsymbol{\zeta}}_1)}{\partial \boldsymbol{\zeta}_1 \partial \boldsymbol{\zeta}_1^T} (\widehat{\boldsymbol{\zeta}}_{\lambda_{11}} - \boldsymbol{\zeta}_1), \end{aligned}$$

where  $\bar{\boldsymbol{\beta}}_1$ ,  $\bar{\Lambda}_i$  and  $\bar{\boldsymbol{\zeta}}_1$  are the interior points between  $\boldsymbol{\beta}_1$  and  $\widehat{\boldsymbol{\beta}}_{\lambda_{21}}$ ,  $\Lambda_i$  and  $\widehat{\Lambda}_i$ , and  $\boldsymbol{\zeta}_1$  and  $\widehat{\boldsymbol{\zeta}}_{\lambda_{11}}$ , respectively. Furthermore, using the similar arguments to the proof of Theorem 1, we have that

$$\begin{aligned} l(\widehat{\boldsymbol{\alpha}}_{\lambda_{11}}, \widehat{\boldsymbol{\beta}}_{\lambda_{21}}) &= \frac{1}{\sqrt{n}} \sum_{i=1}^n \left\{ \begin{array}{c} \tilde{Z}_{i,1} \eta'(\Lambda_i) \\ \tilde{X}_{i,1} \end{array} \right\} \varepsilon_i \\ &\quad - \frac{1}{n} \sum_{i=1}^n \left\{ \begin{array}{c} \tilde{Z}_{i,1} \eta'(\Lambda_i) \\ \tilde{X}_{i,1} \end{array} \right\}^{\otimes 2} \sqrt{n} \begin{pmatrix} \widehat{\boldsymbol{\alpha}}_{\lambda_{11}} - \boldsymbol{\alpha}_{10} \\ \widehat{\boldsymbol{\beta}}_{\lambda_{21}} - \boldsymbol{\beta}_{10} \end{pmatrix} + o_p(1). \end{aligned}$$

Moreover, the summand of the matrix over  $n$  in the second term of above equation converges

to  $\begin{pmatrix} \Gamma_{\tilde{Z}_1 \tilde{Z}_1} & \Gamma_{\tilde{X}_1 \tilde{Z}_1} \\ \Gamma_{\tilde{X}_1 \tilde{Z}_1}^T & \Gamma_{\tilde{X}_1 \tilde{X}_1} \end{pmatrix}$ . These results together with (A.8) lead to

$$\begin{aligned} &n^{1/2} \begin{pmatrix} \Gamma_{\tilde{Z}_1 \tilde{Z}_1} + \Sigma_{\lambda_1} & \Gamma_{\tilde{X}_1 \tilde{Z}_1} \\ \Gamma_{\tilde{X}_1 \tilde{Z}_1}^T & \Gamma_{\tilde{X}_1 \tilde{X}_1} + \Sigma_{\lambda_2} \end{pmatrix} \begin{pmatrix} \widehat{\boldsymbol{\alpha}}_{\lambda_{11}} - \boldsymbol{\alpha}_{10} \\ \widehat{\boldsymbol{\beta}}_{\lambda_{21}} - \boldsymbol{\beta}_{10} \end{pmatrix} - n^{1/2} \begin{pmatrix} \mathbf{R}_{\lambda_1} \\ \mathbf{R}_{\lambda_2} \end{pmatrix} \\ &= \frac{1}{\sqrt{n}} \sum_{i=1}^n \left\{ \begin{array}{c} \tilde{Z}_{i,1} \eta'(\Lambda_i) \varepsilon_i \\ \tilde{X}_{i,1} \varepsilon_i \end{array} \right\} + o_p(1). \end{aligned}$$

It follows that

$$\begin{aligned}
& n^{1/2}(\Gamma_{\tilde{Z}_1\tilde{Z}_1} + \Sigma_{\lambda_1})(\hat{\boldsymbol{\alpha}}_{\lambda_1 1} - \boldsymbol{\alpha}_{10}) + n^{1/2}\Gamma_{\tilde{X}_1\tilde{Z}_1}(\hat{\boldsymbol{\beta}}_{\lambda_2 1} - \boldsymbol{\beta}_{10}) - n^{1/2}\mathbf{R}_{\lambda_1} \\
& \quad = \frac{1}{\sqrt{n}} \sum_{i=1}^n \tilde{Z}_{i,1} \eta'(\Lambda_i) \varepsilon_i + o_p(1) \\
\text{and} \quad & n^{1/2}\Gamma_{\tilde{X}_1\tilde{Z}_1}^T(\hat{\boldsymbol{\alpha}}_{\lambda_1 1} - \boldsymbol{\alpha}_{10}) + n^{1/2}(\Gamma_{\tilde{X}_1\tilde{X}_1} + \Sigma_{\lambda_2})(\hat{\boldsymbol{\beta}}_{\lambda_2 1} - \boldsymbol{\beta}_{10}) - n^{1/2}\mathbf{R}_{\lambda_2} \\
& \quad = \frac{1}{\sqrt{n}} \sum_{i=1}^n \tilde{X}_{i,1} \varepsilon_i + o_p(1).
\end{aligned}$$

After simplification, we have

$$\begin{aligned}
& \sqrt{n}\{(\Gamma_{\tilde{Z}_1\tilde{Z}_1} + \Sigma_{\lambda_1}) - \Gamma_{\tilde{X}_1\tilde{Z}_1}(\Gamma_{\tilde{X}_1\tilde{X}_1} + \Sigma_{\lambda_2})^{-1}\Gamma_{\tilde{X}_1\tilde{Z}_1}^T\}(\hat{\boldsymbol{\alpha}}_{\lambda_1 1} - \boldsymbol{\alpha}_{10}) \\
& \quad + n^{1/2}\{\Gamma_{\tilde{X}_1\tilde{Z}_1}(\Gamma_{\tilde{X}_1\tilde{X}_1} + \Sigma_{\lambda_1})^{-1}\mathbf{R}_{\lambda_1} - \mathbf{R}_{\lambda_2}\} \\
\text{(A.9)} \quad & = \frac{1}{\sqrt{n}} \sum_{i=1}^n \{\tilde{Z}_{i,1} \eta'(\Lambda_i) - \Gamma_{\tilde{X}_1\tilde{Z}_1}(\Gamma_{\tilde{X}_1\tilde{X}_1} + \Sigma_{\lambda_2})^{-1}\tilde{X}_{i,1}\} \varepsilon_i + o_p(1)
\end{aligned}$$

and

$$\begin{aligned}
& \sqrt{n}\{(\Gamma_{\tilde{X}_1\tilde{X}_1} + \Sigma_{\lambda_2}) - \Gamma_{\tilde{X}_1\tilde{Z}_1}^T(\Gamma_{\tilde{Z}_1\tilde{Z}_1} + \Sigma_{\lambda_1})^{-1}\Gamma_{\tilde{X}_1\tilde{Z}_1}\}(\hat{\boldsymbol{\beta}}_{\lambda_2 1} - \boldsymbol{\beta}_{10}) \\
& \quad + n^{1/2}\{\Gamma_{\tilde{X}_1\tilde{Z}_1}^T(\Gamma_{\tilde{Z}_1\tilde{Z}_1} + \Sigma_{\lambda_1})^{-1}\mathbf{R}_{\lambda_2} - \mathbf{R}_{\lambda_1}\} \\
\text{(A.10)} \quad & = \frac{1}{\sqrt{n}} \sum_{i=1}^n \{\tilde{X}_{i,1} - \Gamma_{\tilde{X}_1\tilde{Z}_1}^T(\Gamma_{\tilde{Z}_1\tilde{Z}_1} + \Sigma_{\lambda_1})^{-1}\tilde{Z}_{i,1}\eta'(\Lambda_i)\} \varepsilon_i + o_p(1).
\end{aligned}$$

Equations (A.9) and (A.10) together with the central limit theorem yield that

$$\begin{aligned}
& \sqrt{n}\{\Gamma_{\tilde{Z}_1\tilde{Z}_1} + \Sigma_{\lambda_1} - \Gamma_{\tilde{X}_1\tilde{Z}_1}(\Gamma_{\tilde{X}_1\tilde{X}_1} + \Sigma_{\lambda_2})^{-1}\Gamma_{\tilde{X}_1\tilde{Z}_1}^T\}(\hat{\boldsymbol{\alpha}}_{\lambda_1 1} - \boldsymbol{\alpha}_{10}) \\
& \quad + n^{1/2}\{\Gamma_{\tilde{X}_1\tilde{Z}_1}(\Gamma_{\tilde{X}_1\tilde{X}_1} + \Sigma_{\lambda_2})^{-1}\mathbf{R}_{\lambda_1} - \mathbf{R}_{\lambda_2}\} \rightarrow N(0, \Sigma_{\boldsymbol{\alpha}_1}) \\
\text{and} \quad & \sqrt{n}\{\Gamma_{\tilde{X}_1\tilde{X}_1} + \Sigma_{\lambda_2} - \Gamma_{\tilde{X}_1\tilde{Z}_1}^T(\Gamma_{\tilde{Z}_1\tilde{Z}_1} + \Sigma_{\lambda_1})^{-1}\Gamma_{\tilde{X}_1\tilde{Z}_1}\}(\hat{\boldsymbol{\beta}}_{\lambda_2 1} - \boldsymbol{\beta}_{10}) \\
& \quad + n^{1/2}\{\Gamma_{\tilde{X}_1\tilde{Z}_1}^T(\Gamma_{\tilde{Z}_1\tilde{Z}_1} + \Sigma_{\lambda_1})^{-1}\mathbf{R}_{\lambda_2} - \mathbf{R}_{\lambda_1}\} \rightarrow N(0, \Sigma_{\boldsymbol{\beta}_1}),
\end{aligned}$$

where

$$\begin{aligned}
\Sigma_{\boldsymbol{\alpha}_1} & = \{\Gamma_{\tilde{Z}_1\tilde{Z}_1} - 2\Gamma_{\tilde{X}_1\tilde{Z}_1}(\Gamma_{\tilde{X}_1\tilde{X}_1} + \Sigma_{\lambda_2})^{-1}\Gamma_{\tilde{X}_1\tilde{Z}_1}^T \\
& \quad + \Gamma_{\tilde{X}_1\tilde{Z}_1}(\Gamma_{\tilde{X}_1\tilde{X}_1} + \Sigma_{\lambda_2})^{-1}\Gamma_{\tilde{X}_1\tilde{Z}_1}(\Gamma_{\tilde{X}_1\tilde{X}_1} + \Sigma_{\lambda_2})^{-1}\Gamma_{\tilde{X}_1\tilde{Z}_1}^T\} \sigma^2
\end{aligned}$$

and

$$\begin{aligned}
\Sigma_{\boldsymbol{\beta}_1} & = \{\Gamma_{\tilde{X}_1\tilde{X}_1} - 2\Gamma_{\tilde{X}_1\tilde{Z}_1}^T(\Gamma_{\tilde{Z}_1\tilde{Z}_1} + \Sigma_{\lambda_1})^{-1}\Gamma_{\tilde{X}_1\tilde{Z}_1} \\
& \quad + \Gamma_{\tilde{X}_1\tilde{Z}_1}^T(\Gamma_{\tilde{Z}_1\tilde{Z}_1} + \Sigma_{\lambda_1})^{-1}\Gamma_{\tilde{X}_1\tilde{Z}_1}(\Gamma_{\tilde{Z}_1\tilde{Z}_1} + \Sigma_{\lambda_1})^{-1}\Gamma_{\tilde{X}_1\tilde{Z}_1}^T\} \sigma^2.
\end{aligned}$$

Because each element of  $n^{1/2}\Sigma_{\lambda_1}$ ,  $n^{1/2}\Sigma_{\lambda_2}$ ,  $n^{1/2}\mathbf{R}_{\lambda_1}$  and  $n^{1/2}\mathbf{R}_{\lambda_2}$  tends to zero, we complete the proof.

**Proof of Theorem 3.** Let  $\tau_n = \log(n)$ ,  $\lambda_{n1j} = \tau_n \text{SE}(\hat{\alpha}_j^u)$ ,  $\lambda_{n2k} = \tau_n \text{SE}(\hat{\beta}_k^u)$ , and

$$\text{BIC}(S_T) = \log\{\sigma_n^2(S_T)\} + \{\log(n)/n\}\text{DF}(S_T),$$

where  $\text{DF}(S_T)$  stands for the degrees of freedom of the true model  $S_T$ .  $\text{SE}(\hat{\alpha}_k^u) = O(1/\sqrt{n})$  and  $\text{SE}(\hat{\beta}_j^u) = O(1/\sqrt{n})$ . Thus,  $\lambda_{n1j} = O_P\{\log(n)/\sqrt{n}\}$  and  $\lambda_{n2k} = O_P\{\log(n)/\sqrt{n}\}$ . Then, employing the similar techniques to those in Wang, Li and Tsai (2007), we obtain that

$$(A.11) \quad P\{\text{BIC}(\tau_n) = \text{BIC}(S_T)\} = 1.$$

Therefore, to prove the theorem, it suffices to show that

$$(A.12) \quad P\left\{\inf_{\lambda \in \Omega_- \cup \Omega_+} \text{BIC}(\lambda) > \text{BIC}(\tau_n)\right\} \rightarrow 1,$$

where

$$\Omega_- = \{\lambda : S_\lambda \not\supset S_T\} \quad \text{and} \quad \Omega_+ = \{\lambda : S_\lambda \supset S_T\}$$

represent the underfitted and overfitted models, respectively.

To demonstrate (A.12), we consider two separate cases given below.

*Case 1: Underfitted model* (i.e., the model misses at least one covariate from the true model).

For any  $\lambda \in \Omega_-$ , (A.11) together with assumptions (A) and (B), implies that, with probability tending to 1,

$$\begin{aligned} \text{BIC}(\lambda) - \text{BIC}(\tau_n) &= \log\{\text{MSE}(\lambda)\} + \{\log(n)/n\}\text{DF}_\lambda - \text{BIC}(S_T) \\ &\geq \log\{\text{MSE}(\lambda)\} - \text{BIC}(S_T) \\ &\geq \log\{\sigma_n^2(S_\lambda)\} - \text{BIC}(S_T) \\ &\geq \inf_{\lambda \in \Omega_-} \log\{\sigma_n^2(S_\lambda)\} - \text{BIC}(S_T) \\ &\geq \min_{S \not\supset S_T} \log\{\sigma_n^2(S)\} - \log\{\sigma_n^2(S_T)\} - \{\log(n)/n\}\text{DF}(S_T) \\ &\rightarrow \min_{S \not\supset S_T} \log\{\sigma^2(S)/\sigma^2(S_T)\} \geq 0. \end{aligned}$$

*Case 2: Overfitted model*, (i.e., the model contains all the covariates in the true model and includes at least one covariate that does not belong to the true model). For any  $\lambda \in \Omega_+$ , it

follows by (A.11) that, with probability tending to one,

$$\begin{aligned}
n\{\text{BIC}(\lambda) - \text{BIC}(\tau_n)\} &= n\{\text{BIC}(\lambda) - \text{BIC}(S_T)\} \\
&= n \log\{\text{MSE}(\lambda)/\sigma_n^2(S_T)\} + (\text{DF}_\lambda - \text{DF}_{S_T}) \log(n) \\
&\geq n \log\{\sigma_n^2(S_\lambda)/\sigma_n^2(S_T)\} + (\text{DF}_\lambda - \text{DF}_{S_T}) \log(n) \\
&= \frac{n\{\sigma_n^2(S_\lambda) - \sigma_n^2(S_T)\}}{\sigma_n^2(S_T)} \{1 + o_P(1)\} + (\text{DF}_\lambda - \text{DF}_{S_T}) \log(n).
\end{aligned}$$

Applying the result of Theorem 4, we know that  $n\{\sigma_n^2(S_\lambda) - \sigma_n^2(S_T)\}/\sigma_n^2(S_T)$  is an asymptotically chi-squared distribution with  $\text{DF}_\lambda - \text{DF}_{S_T}$  degrees of freedom. Accordingly, we obtain that  $n\{\sigma_n^2(S_\lambda) - \sigma_n^2(S_T)\}/\sigma_n^2(S_T) = O_P(1)$ . Moreover, for any  $\lambda \in \Omega_+$ ,  $\text{DF}_\lambda - \text{DF}_{S_T} \geq 1$ , and hence  $(\text{DF}_\lambda - \text{DF}_{S_T}) \log(n)$  diverges to  $+\infty$  as  $n \rightarrow \infty$ . Consequently,

$$P \left\{ \inf_{\lambda \in \Omega_+} n\{\text{BIC}(\lambda) - \text{BIC}(\tau_n)\} > 0 \right\} = P \left\{ \inf_{\lambda \in \Omega_+} \text{BIC}(\lambda) - \text{BIC}(\tau_n) > 0 \right\} \rightarrow 1.$$

The results of Cases 1 and 2 complete the proof.

**Proof of Theorem 4.** We apply similar techniques to those in the proofs of Theorems 3.1 and 3.2 in Fan and Huang (2005) to show this theorem. Accordingly, we only provide a sketch of a proof here, and detailed derivations can be obtained from the authors upon request.

Let

$$B_n = \left[ \sum_{i=1}^n \begin{Bmatrix} \eta'(\Lambda_i) \widehat{Z}_i \\ \widehat{X}_i \end{Bmatrix}^{\otimes 2} \right]^{-1} \mathbf{A}^\top \left( \mathbf{A} \left[ \sum_{i=1}^n \begin{Bmatrix} \eta'(\Lambda_i) \widehat{Z}_i \\ \widehat{X}_i \end{Bmatrix}^{\otimes 2} \right]^{-1} \mathbf{A}^\top \right)^{-1} \mathbf{A}.$$

The difference  $Q(H_0) - Q(H_1)$  can be expressed as

$$\begin{aligned}
&\sum_{i=1}^n \{Y_i - X_i^\top \widehat{\beta}_0 - \widehat{\eta}(Z_i^\top \widehat{\alpha}_0, \widehat{\zeta}_0)\}^2 - \sum_{i=1}^n \{Y_i - X_i^\top \widehat{\beta}_1 - \widehat{\eta}(Z_i^\top \widehat{\alpha}_1, \widehat{\zeta}_1)\}^2 \\
&= \sum_{i=1}^n \{\widehat{\eta}(Z_i^\top \widehat{\alpha}_1, \widehat{\zeta}_1) - \widehat{\eta}(Z_i^\top \widehat{\alpha}_0, \widehat{\zeta}_0) + X_i^\top (\widehat{\beta}_1 - \widehat{\beta}_0)\}^2 \\
&\quad + 2 \sum_{i=1}^n \{Y_i - X_i^\top \widehat{\beta}_1 - \widehat{\eta}(Z_i^\top \widehat{\alpha}_1, \widehat{\zeta}_1)\} \{\widehat{\eta}(Z_i^\top \widehat{\alpha}_1, \widehat{\zeta}_1) - \widehat{\eta}(Z_i^\top \widehat{\alpha}_0, \widehat{\zeta}_0) + X_i^\top (\widehat{\beta}_1 - \widehat{\beta}_0)\} \\
\text{(A.13)} \quad &\stackrel{\text{def}}{=} Q_1 + Q_2.
\end{aligned}$$

It can be shown that  $Q_2$  is asymptotically negligible in probability. Furthermore,  $Q_1$  can be simplified as

$$Q_1 = (\widehat{\zeta}_1 - \widehat{\zeta}_0)^\top \sum_{i=1}^n \begin{Bmatrix} \eta'(Z_i^\top \widehat{\alpha}_0, \widehat{\zeta}_0) \widehat{Z}_i \\ \widehat{X}_i \end{Bmatrix}^{\otimes 2} (\widehat{\zeta}_1 - \widehat{\zeta}_0) + o_p(1).$$

A direct calculation yields that  $\widehat{\zeta}_0 = \widehat{\zeta}_1 - B_n \widehat{\zeta}_1$ . This, together with the asymptotic normality and consistency of  $\widehat{\zeta}_1$  obtained from Theorem 1, implies  $\sigma^{-2}Q_1 \rightarrow \chi_m^2$  in distribution under  $H_0$ . Moreover, under  $H_1$ ,  $\sigma^{-2}Q_1$  asymptotically follows a noncentral chi-squared distribution with  $m$  degrees of freedom and noncentrality parameter  $\phi$ . This completes the proof.

**Proof of Theorem 5.** It is noteworthy that

$$\begin{aligned} RSS(H_0) &= \sum_{i=1}^n \{Y_i - X_i^T \boldsymbol{\beta} - \widetilde{\eta}(Z_i^T \widehat{\boldsymbol{\alpha}})\}^2 \\ &\quad + \sum_{i=1}^n [\{Y_i - X_i^T \widehat{\boldsymbol{\beta}} - \widehat{\eta}(Z_i^T \widehat{\boldsymbol{\alpha}})\}^2 - \{Y_i - X_i^T \boldsymbol{\beta} - \widetilde{\eta}(Z_i^T \widehat{\boldsymbol{\alpha}})\}^2] \\ &\stackrel{\text{def}}{=} RSS^*(H_0) + I_{n0} \end{aligned}$$

and

$$\begin{aligned} RSS(H_1) &= \sum_{i=1}^n \{Y_i - X_i^T \boldsymbol{\beta} - \widehat{\eta}(Z_i^T \widehat{\boldsymbol{\alpha}})\}^2 \\ &\quad + \sum_{i=1}^n [\{Y_i - X_i^T \widehat{\boldsymbol{\beta}} - \widehat{\eta}(Z_i^T \widehat{\boldsymbol{\alpha}})\}^2 - \{Y_i - X_i^T \boldsymbol{\beta} - \widehat{\eta}(Z_i^T \widehat{\boldsymbol{\alpha}})\}^2] \\ &\stackrel{\text{def}}{=} RSS^*(H_1) + I_{n1}. \end{aligned}$$

Applying similar arguments to those in the proof of Theorem 5 in Fan, Zhang, and Zhang (2001), under  $H_0$ , we have

$$\frac{nr_K}{2} \frac{RSS^*(H_0) - RSS^*(H_1)}{RSS^*(H_1)} \sim \chi_{df_n}^2,$$

where  $df_n$  is defined in Theorem 5 and it approaches infinity as  $n \rightarrow \infty$ . Furthermore, it can be shown straightforwardly that  $n^{-1}I_{n0} = \sigma^2\{1 + o_P(1)\}$  and  $n^{-1}I_{n1} = \sigma^2\{1 + o_P(1)\}$ . These results complete the proof.

## References

- AKAIKE, H. (1973). Information theory as an extension of the maximum likelihood principle. In *2nd International Symposium on Information Theory*, Ed. B. N. Petrov and F. Csaki, pp. 267-281, Budapest: Akademia Kiado.
- BICKEL, P. J., KLAASEN, C. A. J., RITOV, YA'ACOV and WELLNER, J. A. (1993). *Efficient and Adaptive Estimation for Semiparametric Models*. The Johns Hopkins University Press.

- CARROLL, R. J., FAN, J., GIJBELS, I. and WAND, M. P. (1997). Generalized partially linear single-index models. *Journal of the American Statistical Association*, **92**, 477-489.
- DUAN, N. H. and LI, K. C. (1991). Slicing regression: a link-free regression method. *The Annals of Statistics*, **19**, 505-530.
- FAN, J. and GIJBELS, I. (1996). *Local Polynomial Modelling and Its Applications*. New York: Chapman and Hall.
- FAN, J. and HUANG, T. (2005). Profile likelihood inferences on semiparametric varying-coefficient partially linear models. *Bernoulli*, **11**, 1031-1057.
- FAN, J. and LI, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, **96**, 1348-1360.
- FAN, J. and LI, R. (2004). New estimation and model selection procedures for semiparametric modeling in longitudinal data analysis. *Journal of the American Statistical Association*, **99**, 710-723.
- FAN, J., ZHANG, C., and ZHANG, J. (2001). Generalized likelihood ratio statistical and wilks phenomenon. *The Annals of Statistics*, **29**, 153-193.
- HÄRDLE, W., HALL, P. and ICHIMURA, H. (1993). Optimal smoothing in single-index models. *The Annals of Statistics*, **21**, 157-178.
- HÄRDLE, W., LIANG, H. and GAO, J. (2000). *Partially Linear Models*. Heidelberg: Springer Physica-Verlag.
- HOROWITZ, J. (1998). *Semiparametric Methods in Econometrics* (Lecture Notes in Statistics). New York: Springer.
- HOROWITZ, J. L. and HÄRDLE, W. (1996). Direct semiparametric estimation of single-index Models with discrete covariates. *Journal of the American Statistical Association*, **91**, 1632-1639.
- ICHIMURA, H. (1993). Semiparametric least squares (SLS) and weighted SLS estimation of single-index models. *Journal of Econometrics*, **58**, 71-120.
- JENNRICH, R. I. (1969). Asymptotic properties of nonlinear least squares estimators. *Annals of Mathematical Statistics*, **40**, 633-643.
- JOHNSON, R. W. (2003). Kiplinger's personal finance. *Journal of Statistics Education*, **57**, 104-123.
- LIANG, H. and WANG, N. (2005). Partially linear single-index measurement error models. *Statistica Sinica*, **15**, 99-116.

- MACK, Y. and SILVERMAN, B. (1982). Weak and strong uniform consistency of kernel regression estimates. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, **60**, 405–415.
- NEWNEY, W. K. (1994). The asymptotic variance of semiparametric estimators. *Econometrica*, **62**, 1349–1382.
- POWELL, J. L., STOCK, J. H., and STOKER, T. M. (1989). Semiparametric estimation of index coefficient. *Econometrica*, **51**, 1403–1430.
- SEIFERT, B. and GASSER, T. (1996). Finite-sample variance of local polynomials: analysis and solutions. *Journal of the American Statistical Association*, **91**, 267–275.
- SEVERINI, T. A. and WONG, W. H. (1992). Profile likelihood and conditionally parametric models. *The Annals of Statistics*, **20**, 1768–1802.
- SPECKMAN, P. (1988). Kernel smoothing in partial linear models. *Journal of the Royal Statistical Society, Series B*, **50**, 413–436.
- TIBSHIRANI, R. (1996). Regression shrinkage and selection via the LASSO. *Journal of the Royal Statistical Society, Series B*, **58**, 267–288.
- WANG, H. S., LI, R., and TSAI, C. L. (2007). Tuning parameter selectors for the smoothly clipped absolute deviation method. *Biometrika*, **94**, 553–568.
- XIA, Y. C. and HÄRDLE, W. (2006). Semi-parametric estimation of partially linear single-index models. *Journal of Multivariate Analysis*, **97**, 1162–1184.
- XIA, Y. C., TONG, H., LI, W. K. and ZHU, L. (2002). An adaptive estimation of dimension reduction space (with discussion). *Journal of the Royal Statistical Society, Series B*, **64**, 363–410.
- YU, Y. and RUPPERT, D. (2002). Penalized spline estimation for partially linear single-index models. *Journal of the American Statistical Association*, **97**, 1042–1054.
- ZHANG, Y. LI, R. and TSAI, C.-L. (2010). Regularization parameter selections via generalized information criterion. *Journal of American Statistical Association*. In press.

TABLE 1

*Simulation Results for Example 4.1. The profile least squares estimates (PrLS) and their corresponding mean squared errors ( $\times 10^{-4}$ ) for model (4.1)*

$n$	$\alpha_1(= 0.7071)$	$\alpha_2(= 0.7071)$	$\phi_0(= 0.7854)$
50	0.7053(21.5274)	0.7059(21.3398)	0.7859(42.9158)
100	0.7054(8.5874)	0.7076(8.4175)	0.7869(17.0116)
200	0.7067(4.4636)	0.7069(4.4287)	0.7856(8.8942)

TABLE 2

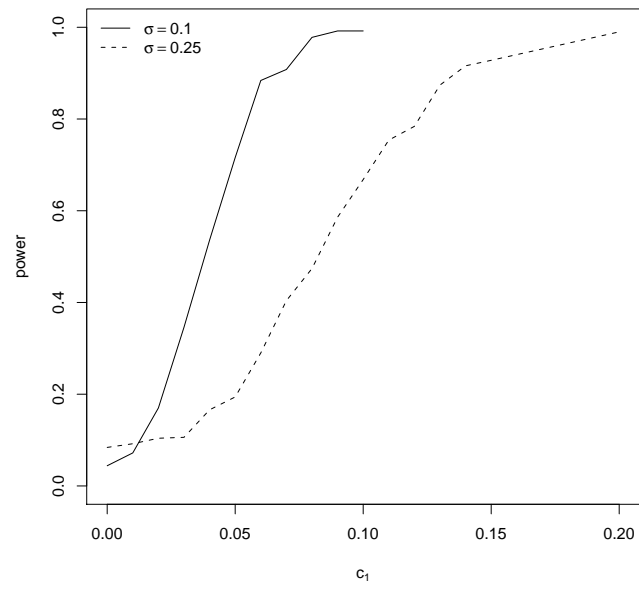
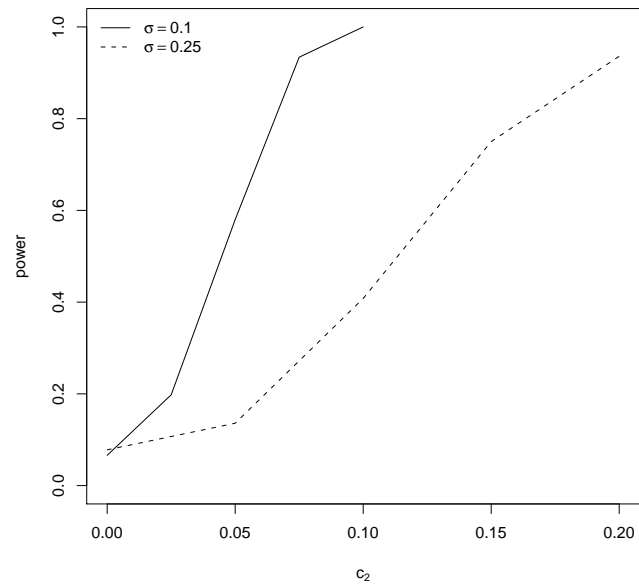
*Simulation Results for Example 4.1. The profile least squares estimates (PrLS) and their corresponding mean squared errors ( $\times 10^{-4}$ ) for model (4.2)*

$n$	$\alpha_1(= 0.5774)$	$\alpha_{22}(= 0.5774)$	$\alpha_3(= 0.5774)$	$\beta(= 0.3)$
50	0.5753(5.8336)	0.5771(5.5685)	0.5782(5.9245)	0.2923(11.4582)
100	0.5776(2.5009)	0.5774(2.4606)	0.5764(2.4770)	0.3000(4.7030)
200	0.5782(1.1533)	0.5771(1.0852)	0.5764(1.2483)	0.3004(2.2026)

TABLE 3

Simulation Results for Example 4.2. S-AIC: SCAD(AIC); S-BIC: SCAD(BIC). MRME: median of relative model error; C: the average number of the eight true zero coefficients that were correctly set to zero; I: the average number of the four truly nonzero coefficients incorrectly set to zero.

n	$\sigma = 0.1$						$\sigma = 0.25$					
	$\alpha$			$\beta$			$\alpha$			$\beta$		
	MRME	C	I	MRME	C	I	MRME	C	I	MRME	C	I
	scenario (i)											
100 Oracle	0.26	4	0	0.28	6	0	0.2	4	0	0.27	6	0
S-BIC	0.37	3.60	0.08	0.91	5.32	0.29	0.73	3.29	0.30	0.86	4.91	1.02
S-AIC	0.66	3.08	0.05	0.97	4.12	0.15	0.75	2.70	0.11	0.91	4.02	0.68
200 Oracle	0.27	4	0	0.34	6	0	0.31	4	0	0.39	6	0
S-BIC	0.33	3.89	0.02	0.85	5.55	0.02	0.36	3.86	0.03	0.94	5.50	0.57
S-AIC	0.60	3.39	0.01	0.92	4.49	0.01	0.62	3.29	0.01	0.93	4.43	0.23
	scenario (ii)											
100 Oracle	0.29	4	0	0.35	6	0	0.24	4	0	0.24	6	0
S-BIC	0.36	3.75	0.05	0.88	5.44	0.19	0.66	3.47	0.27	0.94	5.11	1.07
S-AIC	0.65	3.26	0.02	0.91	4.35	0.07	0.70	2.86	0.09	0.96	4.04	0.67
200 Oracle	0.31	4	0	0.36	6	0	0.32	4	0	0.3	6	0
S-BIC	0.36	3.91	0.01	0.79	5.64	0.01	0.40	3.87	0.03	0.85	5.53	0.50
S-AIC	0.62	3.32	0	0.93	4.51	0.01	0.72	3.29	0.01	0.97	4.45	0.19
	scenario (iii)											
100 Oracle	0.28	4	0	0.15	6	0	0.19	4	0	0.18	6	0
S-BIC	0.48	3.67	0.03	0.82	5.24	0.05	0.50	3.35	0.21	0.85	4.99	0.56
S-AIC	0.74	3.09	0.02	0.94	4.35	0.03	0.71	2.70	0.12	0.98	4.17	0.32
200 Oracle	0.32	4	0	0.18	6	0	0.29	4	0	0.17	6	0
S-BIC	0.39	3.89	0	0.73	5.52	0.01	0.39	3.80	0.04	0.83	5.29	0.12
S-AIC	0.68	3.30	0	0.84	4.54	0	0.66	3.13	0.02	0.85	4.48	0.03

FIG 1. Power function of the test statistic  $T_1$ .FIG 2. Power function of the test statistic  $T_2$ .

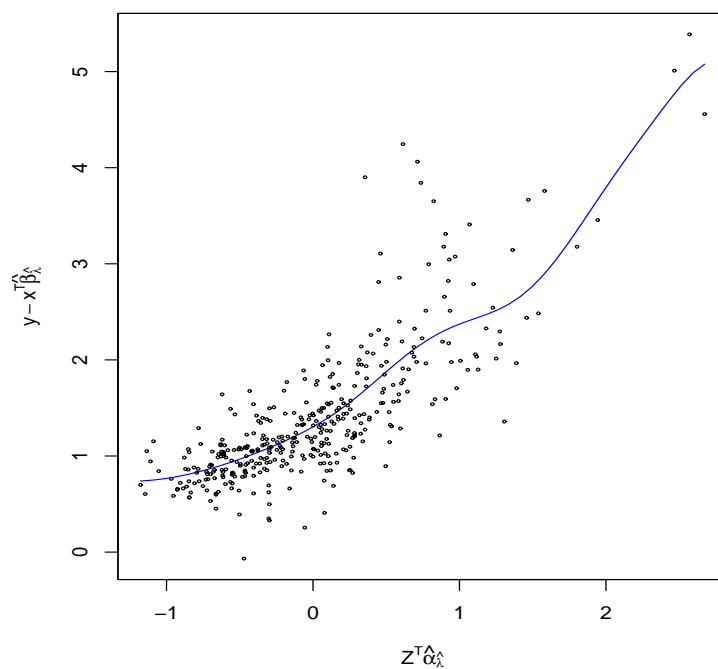


FIG 3. Scatter plot of  $y - x^T \hat{\beta}_\lambda$  versus  $z^T \hat{\alpha}_\lambda$  with local linear estimate superimposed.

HUA LIANG and XIANG LIU  
 DEPARTMENT OF BIostatISTICS  
 AND COMPUTATIONAL BIOLOGY  
 UNIVERSITY OF ROCHESTER  
 ROCHESTER, NY 14642, USA  
 E-MAIL: [hliang@bst.rochester.edu](mailto:hliang@bst.rochester.edu)  
 E-MAIL: [xliu@bst.rochester.edu](mailto:xliu@bst.rochester.edu)

RUNZE LI  
 DEPARTMENT OF STATISTICS  
 THE PENNSYLVANIA STATE UNIVERSITY  
 UNIVERSITY PARK, PA 16802, USA  
 E-MAIL: [rli@stat.psu.edu](mailto:rli@stat.psu.edu)

CHIH-LING TSAI  
 GRADUATE SCHOOL OF MANAGEMENT  
 UNIVERSITY OF CALIFORNIA  
 DAVIS, CA 95616, USA  
 E-MAIL: [cltsai@ucdavis.edu](mailto:cltsai@ucdavis.edu)