Electronic Journal of Statistics
Vol. 7 (2013) 2656–2682
ISSN: 1935-7524
DOI: 10.1214/13-EJS859

Penalized profiled semiparametric estimating functions

Lan Wang

School of Statistics, University of Minnesota, Minneapolis, MN 55455
e-mail: wangx346@umn.edu

Bo Kai

Department of Mathematics, College of Charleston, Charleston, SC 29424
e-mail: KaiB@cofc.edu

Cédric Heuchenne

HEC-Management School of University of Liège, Rue Louvrex 14, B-4000 Liège, Belgium
e-mail: C.Heuchenne@ulg.ac.be

and

Chih-Ling Tsai

Graduate School of Management, University of California at Davis, Davis, CA 95616
College of Management, National Taiwan University, Taiwan
e-mail: cltsai@ucdavis.edu

Abstract: In this paper, we propose a general class of penalized profiled semiparametric estimating functions which is applicable to a wide range of statistical models, including quantile regression, survival analysis, and missing data, among others. It is noteworthy that the estimating function can be non-smooth in the parametric and/or nonparametric components. Without imposing a specific functional structure on the nonparametric component or assuming a conditional distribution of the response variable for the given covariates, we establish a unified theory which demonstrates that the resulting estimator for the parametric component possesses the oracle property. Monte Carlo studies indicate that the proposed estimator performs well. An empirical example is also presented to illustrate the usefulness of the new method.

AMS 2000 subject classifications: Primary 62G05, 62G08; secondary 62G20.
Keywords and phrases: Profiled semiparametric estimating functions, nonconvex penalty, non-smooth estimating functions.

Received March 2013.

Contents

1. Introduction

In statistical estimation, regularization or penalization has flourished during the last twenty years or so as an effective approach for controlling model complexity and avoiding overfitting, see for example Bickel and Li (2006) for a general survey. To estimate an unknown p-dimensional vector of parameters $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T$, the regularized estimator is defined as

$$\underset{\boldsymbol{\beta}}{\operatorname{argmin}} \Big\{ L_n(\boldsymbol{\beta}) + \sum_{i=1}^{p} p_{\lambda_n}(|\beta_j|) \Big\},$$

where L_n is a loss function that measures the goodness-of-fit of the model, and $p_{\lambda_n}(\cdot)$ is a penalty function that depends on a positive tuning parameter λ_n. Despite the large amount of work on regularized estimation, most existing studies were restricted to linear regression and likelihood based models. Recent statistical literature has witnessed increasingly growing interest on regularized semiparametric models, due to their balance between flexibility and parsimony. However, current results usually focus on a specific type of semiparametric regression model. For example, Bunea (2004), Xie and Huang (2009) and Liang and Li (2009) studied the partially linear regression model; Wang and Xia (2009) investigated shrinkage estimation of the varying coefficient model; Li and Liang (2008) proposed the nonconcave penalized quasilikelihood method for variable selection in semiparametric varying-coefficient models; Liang et al. (2010) considered partially linear single-index models; Kai, Li and Zou (2011) investigated the varying-coefficient partially linear models; Wang et al. (2011) studied estimation and variable selection for generalized additive partial linear models. Although the aforementioned work convincingly demonstrate the merits of regularization in a semiparametric setting, a general theory is still lacking. Furthermore, most of the existing theory assumes a smooth loss function which excludes many interesting applications, such as those arising from quantile regression, survival analysis and missing data analysis.

Instead of penalizing the loss function, Fu (2003) proposed to directly penalize the estimating function for generalized linear models. Later, Johnson, Lin and Zeng (2008) derived impressive results on the asymptotic theory for a broad class of penalized estimating functions when the regression model is linear but the error distribution is unspecified. It is noteworthy that their approach allows the estimating function to be discrete. In addition, Chen, Linton and Keilegom (2003) introduced a non-smooth estimating function for the semiparametric models, but they only focused on non-penalized estimation. Since the non-parametric component in their estimating function has been profiled, we refer to it as the profiled semiparametric estimating equation for the purpose of simplicity. Both of the above two innovative approaches motivate us to propose a general class of penalized profiled estimating functions that substantially expands the scope of applicability of the regularization approach for semiparametric models.

In this paper, we provide a unified approach for obtaining penalized semiparametric estimation that is applicable for many commonly used likelihood based models as well as non-likelihood based semiparametric models. This broad class of models has three appealing features:

- First, the models incorporate nonparametric components for nonlinearity without imposing any assumptions on the conditional distribution of the response variable for the given covariates.
- Second, the profiled estimating function allows the preliminary estimators of the nonparametric functions to depend on the unknown parametric component. Furthermore, the estimator for the nonparametric component is only assumed to satisfy mild conditions. Thus, the widely used nonparametric estimation methods, such as kernel smoothing or spline approximation, can be applied.
- Third, the profiled semiparametric estimation function can be non-smooth in both the parametric and/or nonparametric components, which is particularly useful for models arising from quantile regression, survival analysis, and missing data analysis, among others.

Based on the profiled semiparametric estimating function with an appropriate nonconvex penalty function, we demonstrate that the penalized estimator of the parametric component possesses the oracle property under suitable conditions. That is, the zero coefficients in the parametric component are estimated to be exactly zero with probability approaching one and the nonzero coefficients have the same asymptotic normal distribution as if it is known *a priori*. It is noteworthy that asymptotic results are established under a set of mild conditions and without assuming a parametric likelihood function. In addition, the proposed estimator can be computed via an efficient algorithm.

The rest of the paper is organized as follows. Section 2 introduces the methodology of the penalized profiled semiparametric estimating function and then illustrates its applicability via four analytical examples. Section 3 presents a set of sufficient conditions and provides asymptotic theories of the penalized estimator. Monte Carlo studies and an empirical example are reported in Section 4

to demonstrate the finite sample performance and the usefulness of proposed method, respectively. Section 5 concludes the article with a brief discussion. All detailed proofs are relegated to the Appendix.

2. Penalized profiled semiparametric estimating function

2.1. Estimating function

Let $m(\mathbf{z}, \boldsymbol{\beta}, h)$ be a p-dimensional vector that is a function of a p-dimensional parameter vector $\boldsymbol{\beta}$ and an infinite dimensional parameter h. The nonparametric component h can depend on both $\boldsymbol{\beta}$ and $\mathbf{z}$, and thus is written as $h(\mathbf{z}, \boldsymbol{\beta})$ when clarity is needed. We assume that $\boldsymbol{\beta} \in \mathbb{B}$, a compact subset of $\mathbb{R}^p$; and that $h \in \mathbb{H}$ which is a vector space of functions endowed with the sup-norm metric $||h||_\infty = \sup_{\boldsymbol{\beta}} \sup_{\mathbf{z}} |h(\mathbf{z}, \boldsymbol{\beta})|$. We further assume that the data $\{\mathbf{Z}_i = (\tilde{\mathbf{X}}_i^T, Y_i)^T, i = 1, \ldots, n\}$ are randomly generated from a distribution which satisfies

$$E[m(\mathbf{Z}_i, \boldsymbol{\beta}_0, h_0(\mathbf{Z}_i, \boldsymbol{\beta}_0))|\tilde{\mathbf{X}}_i] = 0, \tag{1}$$

for some $\boldsymbol{\beta}_0 \in \mathbb{B}$ and $h_0 \in \mathbb{H}$, where $\tilde{\mathbf{X}}_i$ is a generic notation for the covariate vector and Y_i is a response variable. In this paper, we consider semiparametric models satisfying the above moment condition, and denote the true values of the finite and infinite dimensional parameters as $\boldsymbol{\beta}_0$ and $h_0(\cdot, \boldsymbol{\beta}_0)$, respectively.

In many real applications, the researchers are interested in estimating the parametric component $\boldsymbol{\beta}_0$ and treat the nonparametric component h_0 as a nuisance function. To this end, for a given $\boldsymbol{\beta}$, we consider the "profiled" estimator $\widehat{h}(\cdot, \boldsymbol{\beta})$ (abbreviated as $\widehat{h}$), which serves as a nonparametric estimator for $h(\cdot, \boldsymbol{\beta})$ in the semiparametric setting. To estimate $\boldsymbol{\beta}_0$, we subsequently define the p-dimensional profiled semiparametric estimating function

$$\mathbf{M}_n(\boldsymbol{\beta}, \widehat{h}) = n^{-1} \sum_{i=1}^{n} m(\mathbf{Z}_i, \boldsymbol{\beta}, \widehat{h}). \tag{2}$$

Let $\mathbf{M}(\boldsymbol{\beta}, h) = E[m(\mathbf{Z}_i, \boldsymbol{\beta}, h(\mathbf{Z}_i, \boldsymbol{\beta}))]$ be the population version of the estimating function. In this paper, we assume that $\mathbf{M}(\boldsymbol{\beta}, h)$ is smooth in $\boldsymbol{\beta}$, while its sample version $\mathbf{M}_n(\boldsymbol{\beta}, \widehat{h})$ may be non-smooth in $\boldsymbol{\beta}$. Based on the above estimating function, Chen, Linton and Keilegom (2003) considered the problem of estimating $\boldsymbol{\beta}_0$ by emphasizing that m is a non-smooth function in $\boldsymbol{\beta}$ and/or h. Although $\mathbf{M}_n(\boldsymbol{\beta}, \widehat{h})$ only contains a profile estimator, $\hat{h}$, it may implicitly depend on the additional estimators induced by the model setting. For the sake of explicitness, we sometimes include those augmented components in the estimating functions (e.g., see Examples 2 and 3 in the next subsection).

In this paper, we study a related but different problem of variable selection and estimation for the parametric component. We assume that some of the components in $\boldsymbol{\beta}_0 = (\beta_{01} \ldots, \beta_{0p})^T$ are zero, corresponding to redundant covariates.

To estimate $\boldsymbol{\beta}_0$ and identify its nonzero components, we propose the following penalized profiled (PP) semiparametric estimating function:

$$\mathbf{U}_n(\boldsymbol{\beta}, \widehat{h}) = \mathbf{M}_n(\boldsymbol{\beta}, \widehat{h}) + q_{\lambda_n}(|\boldsymbol{\beta}|)\text{sgn}(\boldsymbol{\beta}), \tag{3}$$

where the notation $\mathbf{q}_{\lambda_n}(|\boldsymbol{\beta}|)\text{sgn}(\boldsymbol{\beta})$ denotes the component-wise product of $\mathbf{q}_{\lambda_n}(|\boldsymbol{\beta}|) = (\mathbf{q}_{\lambda_n}(|\beta_1|), \ldots, \mathbf{q}_{\lambda_n}(|\beta_p|))^T$ with $\text{sgn}(\boldsymbol{\beta}) = (\text{sgn}(\beta_1), \ldots, \text{sgn}(\beta_p))^T$ and $\text{sgn}(t) = I(t > 0) - I(t < 0)$. The function $q_{\lambda_n}(\cdot)$ is the gradient of some penalty function. Based on the penalty function setting in Section 3, $q_{\lambda_n}(|\beta_j|)$ is zero for large values of $|\beta_j|$, whereas it is relatively large for small values of $|\beta_j|$. Accordingly, $M_{nj}(\boldsymbol{\beta}, \widehat{h})$ (the jth component of $\mathbf{M}_n(\boldsymbol{\beta}, \widehat{h})$) is not penalized when $|\beta_j|$ is large. In contrast, if $|\beta_j|$ is close (but not equal) to zero, $M_{nj}(\boldsymbol{\beta}, \widehat{h})$ is heavily penalized, which forces the estimator of β_{0j} to shrink to zero. Once an estimated coefficient shrinks towards zero, its associated covariate is excluded from the final selected model.

It is known that the convex L_1 penalty or Lasso Tibshirani (1996) is computationally attractive and demonstrates excellent predictive ability. However, it requires stringent assumptions to yield consistent variable selection (Greenshtein and Ritov, 2004; Meinshausen and Bühlmann, 2006; Zhao and Yu, 2006, among others). A useful alternative to the L_1 penalty function is the noncovex penalty function SCAD (Fan and Li, 2001) or MCP (Zhang, 2010), which alleviates the bias of Lasso and achieves model selection consistency under more relaxed conditions on the design matrix. Hence, we focus on nonconvex penalty functions that satisfy the general conditions given in Section 3.1.

When $\mathbf{U}_n(\boldsymbol{\beta}, \widehat{h})$ is a non-smooth function, an exact solution to $\mathbf{U}_n(\boldsymbol{\beta}, \widehat{h}) = 0$ may not exist. Hence, we estimate $\boldsymbol{\beta}_0$ by any $\widehat{\boldsymbol{\beta}}$ that satisfies $||\mathbf{U}_n(\widehat{\boldsymbol{\beta}}, \widehat{h})|| = O_p(n^{-1/2})$, where $||\cdot||$ denotes the L_2 or Euclidean norm. For the sake of simplicity, we name it an approximate estimator. In Section 3, we demonstrate that the oracle estimator is an approximate solution of the penalized profiling estimating equations; and any root-n consistent approximate estimator of the penalized profiling estimating equations possesses the oracle property with probability tending to one.

2.2. Analytical examples

The proposed PP semiparametric estimating function can be applied to a wide range of statistical models. To illustrate its broadness, we consider the four motivating examples given below, some of which will be further discussed later to demonstrate the theory and applications. Since the penalty function in (3) does not depend on the model structure, we only present the profiled semiparametric estimating function.

Example 1 (Partially linear quantile regression)**.** We consider a random sample $(\mathbf{X}_i, W_i, Y_i)$, $i = 1, \ldots, n$, from the partially linear quantile regression model

$$Y_i = \mathbf{X}_i^T \boldsymbol{\beta}_0 + h_0(W_i) + \epsilon_i,$$

where $\boldsymbol{\beta}_0$ and the function $h_0(\cdot)$ are unknown, and the random error ϵ_i satisfies $P(\epsilon_i \le 0|\mathbf{X}_i, W_i) = \tau$. Thus, $\mathbf{X}_i^T\boldsymbol{\beta}_0 + h_0(W_i)$ is the τth conditional quantile of Y_i. Define $\rho_\tau(u) = \tau u - uI(u < 0)$ to be the quantile loss function. For a given $\boldsymbol{\beta}$, let $h(w, \boldsymbol{\beta}) = \text{argmin}_f\, E\big[\rho_\tau(Y_i - \mathbf{X}_i^T\boldsymbol{\beta} - f(W_i)|W_i = w)\big]$, where f is any function such that $f : W \to \mathbf{R}^1$; that is, $h(w, \boldsymbol{\beta})$ is the conditional quantile of $Y_i - \mathbf{X}_i^T\boldsymbol{\beta}$ given $W_i = w$. Then $h_0(w) = h(w, \boldsymbol{\beta}_0)$. Accordingly, the profiled semiparametric estimating function is

$$\mathbf{M}_n(\boldsymbol{\beta}, \widehat{h}) = n^{-1}\sum_{i=1}^{n}\mathbf{X}_i\left[I(Y_i \le \mathbf{X}_i^T\boldsymbol{\beta} + \widehat{h}(W_i, \boldsymbol{\beta})) - \tau\right], \tag{4}$$

where $\widehat{h}(w, \boldsymbol{\beta})$ is a nonparametric estimator of $h(w, \boldsymbol{\beta})$. In Section 4, $\widehat{h}(w, \boldsymbol{\beta})$ is obtained by the local linear smoothing of quantile regression. Specifically, for a given $\boldsymbol{\beta}$, we have that

$$(\widehat{a}_1, \widehat{a}_2) = \text{argmin}\sum_{i=1}^{n}\rho_\tau\left(Y_i - \mathbf{X}_i^T\boldsymbol{\beta} - a_1 - a_2(W_i - w)\right)K_t(W_i - w), \tag{5}$$

where $K_t(\cdot) = t^{-1}K(\cdot/t)$, $K(\cdot)$ is a kernel function, and $t > 0$ is the bandwidth. Accordingly, the local linear estimator is $\widehat{h}(w, \boldsymbol{\beta}) = \widehat{a}_1$.

Example 2 (Single-index mean regression)**.** We observe a random sample $(\mathbf{X}_i, Y_i)$, $i = 1, \ldots, n$, from the model

$$Y_i = h_0(\mathbf{X}_i^T\boldsymbol{\beta}_0) + \epsilon_i, \tag{6}$$

where $\boldsymbol{\beta}_0$ and the function $h_0(\cdot)$ are unknown, and the random error ϵ_i satisfies $E(\epsilon_i|\mathbf{X}_i) = 0$. For a given $\boldsymbol{\beta}$, let $h(\mathbf{X}^T\boldsymbol{\beta}) = E(Y|\mathbf{X}^T\boldsymbol{\beta})$, where $(\mathbf{X}, Y)$ has the same distribution as $(\mathbf{X}_i, Y_i)$. Then $h_0(\mathbf{X}^T\boldsymbol{\beta}_0) = h(\mathbf{X}^T\boldsymbol{\beta}_0)$. There are various approaches to estimate h, for example, the leave-one-out Nadaraya-Watson kernel estimator $\frac{\sum_{j\ne i}K_t((\mathbf{X}_j - \mathbf{X}_i)^T\boldsymbol{\beta})Y_j}{\sum_{j\ne i}K_t((\mathbf{X}_j - \mathbf{X}_i)^T\boldsymbol{\beta})}$, where $K_t(\cdot)$ is defined as in Example 1. Furthermore, adopting Ichimura (1993)'s suggestion, the profiled semiparametric estimating function is

$$\mathbf{M}_n(\boldsymbol{\beta}, \widehat{h}, \widehat{s}) = n^{-1}\sum_{i=1}^{n}\widehat{s}(\mathbf{X}_i, \boldsymbol{\beta})\left[Y_i - \widehat{h}(\mathbf{X}_i^T\boldsymbol{\beta})\right],$$

where $\widehat{s}(\mathbf{X}_i, \boldsymbol{\beta})$ is a nonparametric estimator of the gradient $s(\mathbf{X}_i, \boldsymbol{\beta}) = \frac{\partial h(\mathbf{X}_i^T\boldsymbol{\beta})}{\partial\boldsymbol{\beta}}$, for example, the derivative of the Nadaraya-Watson kernel estimator.

Example 3 (Partially linear mean regression with missing covariates)**.** Consider the partially linear regression model $Y_i = \mathbf{X}_i^T\boldsymbol{\beta}_0 + h_0(W_i) + \epsilon_i$, where $\boldsymbol{\beta}_0$ and $h_0(\cdot)$ are unknown, and $E(\epsilon_i|\mathbf{X}_i, W_i) = 0$. For a given $\boldsymbol{\beta}$, let $h(w, \boldsymbol{\beta}) = E(Y_i - \mathbf{X}_i^T\boldsymbol{\beta}|W_i = w)$, then $h_0(w) = h(w, \boldsymbol{\beta}_0)$. Liang et al. (2004) studied this model when the data on $\mathbf{X}_i$ may not be completely observed. Let δ_i be the observing data indicator: $\delta_i = 1$ if $\mathbf{X}_i$ is observed and $\delta_i = 0$ otherwise. Assume

that $\mathbf{X}_i$ is missing at random in the sense that $P(\delta_i = 1|\mathbf{X}_i, W_i, Y_i) = P(\delta_i = 1|W_i, Y_i)$, and denote the probability of $\mathbf{X}_i$ being observed by $\pi(Y_i, W_i) = P(\delta_i = 1|W_i, Y_i)$. In addition, let $m_1(w) = E(\mathbf{X}|W = w)$, $m_2(w) = E(Y|W = w)$, $m_3(y, w) = E(\mathbf{X}|Y = y, W = w)$, and $m_4(y, w) = E(\mathbf{X}\mathbf{X}^T|Y = y, W = w)$. Then $h(w, \boldsymbol{\beta}) = m_2(w) - m_1(w)^T\boldsymbol{\beta}$. Moreover, let $\widehat{m}_j$ be a nonparametric estimator of m_j for $j = 1, \ldots, 4$. As a result, $\widehat{h}(w, \boldsymbol{\beta}) = \widehat{m}_2(w) - \widehat{m}_1(w)^T\boldsymbol{\beta}$. Finally, let $\widehat{\pi}$ be an estimator of π based on a parametric (e.g., logistic regression) model or a nonparametric regression approach. Adapting Liang et al. (2004)'s method, we obtain the following estimating function

$$\mathbf{M}_n(\boldsymbol{\beta}, \widehat{h}, \widehat{A}) = n^{-1}\sum_{i=1}^{n}\Phi(\boldsymbol{\beta}, \widehat{h}, \widehat{A}, Y_i, \mathbf{X}_i, W_i, \delta_i), \tag{7}$$

where $\widehat{A} = (\widehat{m}_1, \widehat{m}_2, \widehat{m}_3, \widehat{m}_4, \widehat{\pi})$ and

$$\begin{aligned}\Phi(\boldsymbol{\beta}, \widehat{h}, \widehat{A}, Y_i, \mathbf{X}_i, W_i, \delta_i) &= \{\mathbf{X}_i - \widehat{m}_1(W_i)\}\{Y - \mathbf{X}_i^T\boldsymbol{\beta} - \widehat{h}(W_i, \boldsymbol{\beta})\}\frac{\delta_i}{\widehat{\pi}_i}\\ &- \big[\{Y_i - \widehat{m}_2(W_i)\}\{\widehat{m}_3(Y_i, W_i) - \widehat{m}_1(W_i)\} - \{\widehat{m}_4(Y_i, W_i) - \widehat{m}_3(Y_i, W_i)\widehat{m}_1^T(W_i)\\ &- \widehat{m}_1(W_i)\widehat{m}_3^T(Y_i, W_i) + \widehat{m}_1(W_i)\widehat{m}_1^T(W_i)\}\boldsymbol{\beta}\big]\frac{\delta_i - \widehat{\pi}_i}{\widehat{\pi}_i}.\end{aligned}$$

In Section 4.1, the Horvitz–Thompson (HT) weighted local linear kernel estimators (Wang et al., 1998; Liang et al., 2004) are used for estimating $m_j(w)$ $(j = 1, \ldots, 4)$, which collectively yield the estimate of $h(w, \beta)$.

Example 4 (Locally weighted censored quantile regression)**.** Censored quantile regression has been recognized as a useful alternative to the classical proportional hazards model for analyzing survival data. It accommodates heterogeneity in the data and relaxes the proportional hazards assumption. The survival time (or a transformation of it) T_i is subject to random right censoring and may not be completely observed. However, we observe the i.i.d. triples $(\mathbf{X}_i, Y_i, \delta_i^*)$, where $Y_i = \min(T_i, C_i)$, $\delta_i^* = I(T_i \le C_i)$ is the indicator for censoring and C_i is the censoring variable. we further assume that

$$T_i = \mathbf{X}_i^T\boldsymbol{\beta}_0 + \epsilon_i,$$

where $P(\epsilon_i \le 0|\mathbf{X}_i) = \tau$, $0 < \tau < 1$. Therefore, $\mathbf{X}_i^T\boldsymbol{\beta}_0$ is the τth conditional quantile of the survival time. Following Wang and Wang (2009) approach, we obtain the profiled semiparametric estimating function

$$\mathbf{M}_n(\boldsymbol{\beta}, \widehat{h}) = n^{-1}\sum_{i=1}^{n}\mathbf{X}_i\big[\tau - w_i^*(\widehat{h})I(Y_i - \mathbf{X}_i^T\boldsymbol{\beta} < 0)\big],$$

where $\widehat{h}(\cdot|\mathbf{X}_i)$ is the local Kaplan-Meier estimator of $h_0(\cdot|\mathbf{X}_i)$, which is the conditional distribution function of T_i given $\mathbf{X}_i$, and the weight function is

$$w_i^*(h_0) = \begin{cases} 1, & \delta_i^* = 1 \text{ or } h_0(C_i|\mathbf{X}_i) > \tau,\\ \frac{\tau - h_0(C_i|\mathbf{X}_i)}{1 - h_0(C_i|\mathbf{X}_i)}, & \delta_i^* = 0 \text{ and } h_0(C_i|\mathbf{X}_i) < \tau,\end{cases}$$

$i = 1, \ldots, n$. Wang and Wang (2009) showed that the estimator obtained by solving the above estimating function is consistent for $\boldsymbol{\beta}_0$, and it is also asymptotically normal under weaker conditions than those in the literature.

3. Theoretical properties and estimation algorithm

3.1. Asymptotic properties

In this paper, we assume that $\mathbf{U}_n(\boldsymbol{\beta}, \widehat{h})$ can be a non-smooth function due to either $\mathbf{M}_n(\boldsymbol{\beta}, \widehat{h})$ or $q_{\lambda_n}(|\boldsymbol{\beta}|)$. For example, the popular SCAD penalty function (Fan and Li, 2001) has

$$q_{\lambda_n}(\theta) = \lambda_n \left\{ I(\theta \leq \lambda_n) + \frac{(a\lambda_n - \theta)_+}{(a-1)\lambda_n} I(\theta > \lambda_n) \right\} \tag{8}$$

for $\theta \geq 0$ and some $a > 2$, where the notation $\tilde{b}_+$ stands for the positive part of $\tilde{b}$, i.e., $\tilde{b}_+ = \tilde{b} I(\tilde{b} > 0)$. Hence, the $q_{\lambda_n}(\theta)$ function is not differentiable at $\theta = \lambda_n$ and $\theta = a\lambda_n$. It is not surprising that an exact solution to $\mathbf{U}_n(\boldsymbol{\beta}, \widehat{h}) = 0$ may not exist. Hence, we consider an approximate estimator for $\boldsymbol{\beta}_0$ that satisfies $||\mathbf{U}_n(\widehat{\boldsymbol{\beta}}, \widehat{h})|| = O_p(n^{-1/2})$, where $||\cdot||$ denotes the L_2 or Euclidean norm, see also the non-penalized approximate estimator in Chen, Linton and Keilegom (2003). Alternatively, we may consider the estimator as an approximate zero-crossing of $\mathbf{U}_n(\boldsymbol{\beta}, \widehat{h})$; see Johnson, Lin and Zeng (2008).

Without loss of generality, we assume that $\boldsymbol{\beta}_0 = (\boldsymbol{\beta}_{10}^T, \boldsymbol{\beta}_{20}^T)^T$, where $\boldsymbol{\beta}_{10}$ consists of the nonzero components and $\boldsymbol{\beta}_{20} = \mathbf{0}$ contains the zero components. Let $A = \{1 \leq j \leq p : \beta_{0j} \neq 0\}$ be the index set of the nonzero components and denote the dimension of $\boldsymbol{\beta}_{10}$ by s, where $1 \leq s \leq p$. Our goal is to simultaneously estimate $\boldsymbol{\beta}_0$ and identify its nonzero components.

Under the moment condition (1), the population version of the estimating function $\mathbf{M}(\boldsymbol{\beta}, h)$ satisfies $\mathbf{M}(\boldsymbol{\beta}_0, h_0) = \mathbf{M}(\boldsymbol{\beta}_0, h_0(\mathbf{Z}_i, \boldsymbol{\beta}_0)) = 0$. To characterize the influence of the parametric component and nonparametric component on estimation, we adopt the approach of Chen, Linton and Keilegom (2003) and define the ordinary derivative and the path-wise functional derivative of $\mathbf{M}(\boldsymbol{\beta}, h)$. Specifically, the *ordinary derivative* of $\mathbf{M}(\boldsymbol{\beta}, h)$ with respect to $\boldsymbol{\beta}$ is the $p \times p$ matrix $\boldsymbol{\Gamma}_1(\boldsymbol{\beta}, h)$, which satisfies

$$\boldsymbol{\Gamma}_1(\boldsymbol{\beta}, h)(\overline{\boldsymbol{\beta}} - \boldsymbol{\beta}) = \lim_{\tau \to 0} \frac{\left[\mathbf{M}(\boldsymbol{\beta} + \tau(\overline{\boldsymbol{\beta}} - \boldsymbol{\beta}), h(\cdot, \boldsymbol{\beta} + \tau(\overline{\boldsymbol{\beta}} - \boldsymbol{\beta}))) - \mathbf{M}(\boldsymbol{\beta}, h(\cdot, \boldsymbol{\beta}))\right]}{\tau},$$

for all $\overline{\boldsymbol{\beta}} \in \mathbb{B}$. In addition, the *path-wise derivative* of $\mathbf{M}(\boldsymbol{\beta}, h)$ at $h \in \mathbb{H}$ in the direction $[\overline{h} - h]$ is the $p \times 1$ vector $\boldsymbol{\Gamma}_2(\boldsymbol{\beta}, h)$, which satisfies

$$\boldsymbol{\Gamma}_2(\boldsymbol{\beta}, h)[\overline{h} - h] = \lim_{\tau \to 0} \frac{\left[\mathbf{M}(\boldsymbol{\beta}, h(\cdot, \boldsymbol{\beta}) + \tau(\overline{h}(\cdot, \boldsymbol{\beta}) - h(\cdot, \boldsymbol{\beta}))) - \mathbf{M}(\boldsymbol{\beta}, h(\cdot, \boldsymbol{\beta}))\right]}{\tau},$$

where $\{h + \tau(\overline{h} - h) : \tau \in [0, 1]\} \subset \mathbb{H}$.

To facilitate the presentation of the large-sample theory for the penalized profiling semiparametric estimating equations, we consider the following three sets of conditions.

(I) *Conditions on the PP estimating equation*

Let $\mathbb{B}_n = \{\boldsymbol{\beta} : ||\boldsymbol{\beta} - \boldsymbol{\beta}_0|| \le \delta_n\}$ and $\mathbb{H}_n = \{h : ||h - h_0||_\infty \le \delta_n\}$, where the sequence $\delta_n = o(1)$.

(C1) The ordinary derivative $\boldsymbol{\Gamma}_1(\boldsymbol{\beta}, h_0)$ exists for $\boldsymbol{\beta}$ in a small neighborhood of $\boldsymbol{\beta}_0$ and is continuous at $\boldsymbol{\beta} = \boldsymbol{\beta}_0$.

(C2) For $\boldsymbol{\beta} \in \mathbb{B}$, $\boldsymbol{\Gamma}_2(\boldsymbol{\beta}, h_0)[h - h_0]$ exists in all directions $[h - h_0] \in \mathbb{H}$. Furthermore, for all $(\boldsymbol{\beta}, h) \in \mathbb{B}_n \times \mathbb{H}_n$, $||\mathbf{M}(\boldsymbol{\beta}, h) - \mathbf{M}(\boldsymbol{\beta}, h_0) - \boldsymbol{\Gamma}_2(\boldsymbol{\beta}, h_0)[h - h_0]|| \le c||h - h_0||_\infty^2$ for a constant $c > 0$ and $||\boldsymbol{\Gamma}_2(\boldsymbol{\beta}, h_0)[h - h_0] - \boldsymbol{\Gamma}_2(\boldsymbol{\beta}_0, h_0)[h - h_0]|| \le o(1)||\boldsymbol{\beta} - \boldsymbol{\beta}_0||$.

(C3) $P(\widehat{h} \in \mathbb{H}) \to 1$ and $||\widehat{h} - h_0||_\infty = o_p(n^{-1/4})$.

(C4) $\sup_{\boldsymbol{\beta} \in \mathbb{B}_n, h \in \mathbb{H}_n} ||\mathbf{M}_n(\boldsymbol{\beta}, h) - \mathbf{M}(\boldsymbol{\beta}, h) - \mathbf{M}_n(\boldsymbol{\beta}_0, h_0)|| = o_p(n^{-1/2})$.

(C5) $\sqrt{n}\Big[\mathbf{M}_n(\boldsymbol{\beta}_0, h_0) + \boldsymbol{\Gamma}_2(\boldsymbol{\beta}_0, h_0)[\widehat{h} - h_0]\Big] \to N(0, \mathbf{V})$ in distribution for a $p \times p$ positive definite matrix $\mathbf{V}$.

(II) *Conditions on the penalty function and the tuning parameter*

(P1) The penalty function $q_{\lambda_n}(\cdot)$ is nonnegative and non-increasing on the interval $(0, +\infty)$. There exist positive constants $b_1 < b_2$ such that $q_{\lambda_n}(\theta)$ is differentiable on $(0, b_1\lambda_n)$ and $(b_2\lambda_n, \infty)$. For any $|\beta| > b_1\lambda_n$, $\lim_{n\to\infty} n^{1/2} q_{\lambda_n}(|\beta|) = 0$ and $\lim_{n\to\infty} q'_{\lambda_n}(|\beta|) = 0$. In addition, we assume that $\lim_{n\to\infty} \sqrt{n} \inf_{|\beta| \le dn^{-1/2}} q_{\lambda_n}(|\beta|) = \infty$ and $\lim_{n\to\infty} \sup_{|\beta| \le dn^{-1/2}} q'_{\lambda_n}(|\beta|) = 0$, $\forall\ d > 0$.

(P2) $\lim_{n\to\infty} \lambda_n = 0$ and $\lim_{n\to\infty} \sqrt{n}\lambda_n = \infty$.

(III) *Conditions on the true parameters and the unpenalized estimating equation*

(T1) $\boldsymbol{\beta}_0 \in \mathbb{B}$ satisfies $\mathbf{M}(\boldsymbol{\beta}_0, h_0(\cdot, \boldsymbol{\beta}_0)) = 0$.

(T2) For all $\xi > 0$, there exists $\epsilon(\xi) > 0$ such that

$$\inf_{||\boldsymbol{\beta_1} - \boldsymbol{\beta_{10}}|| > \xi} ||\mathbf{M}_1(\boldsymbol{\beta}_1, h_0)|| \ge \epsilon(\xi),$$

where $\mathbf{M}_1(\boldsymbol{\beta}_1, h_0)$ denotes the subvector that consists of the first s components of $\mathbf{M}(\boldsymbol{\beta}, h_0)$ evaluated at $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^T, \mathbf{0}^T)^T$.

(T3) Let $\boldsymbol{\Gamma}_{11}$ denote the $s \times s$ submatrix in the upper-left corner of $\boldsymbol{\Gamma}_1(\boldsymbol{\beta}, h_0)$, which is assumed to be positive definite.

(T4) $\min_{1 \le j \le s} |\beta_{0j}|/\lambda_n \to \infty$ as $n \to \infty$.

Remark 1. Conditions (C1)–(C5) and (T1)–(T3) are similar to those in Chen, Linton and Keilegom (2003) to ensure good performance of the profiled estimating equations, and they are general enough to allow the estimating equations to be non-smooth. In addition, Condition (T4) imposes a constraint on the

magnitude of the smallest signal, which is common for the theory of penalized estimators. It is noteworthy that Condition (P1) is satisfied by popular nonconvex penalty functions, such as SCAD and MCP. Condition (P2) is a standard requirement on the rate of the tuning parameter in achieving the oracle property (Fan and Li, 2001).

Theorem 1. *Assume Conditions (C1)–(C5), (P1)–(P2) and (T1)–(T4) hold.*

1. *Let* $\widehat{\boldsymbol{\beta}}_1$ *be the estimator obtained by solving the unpenalized profiled estimating equations when the true model is known in advance. Then, with probability approaching one, the oracle estimator* $\widehat{\boldsymbol{\beta}} = (\widehat{\boldsymbol{\beta}}_1^T, \mathbf{0}^T)^T$ *is an approximate solution of the penalized profiling estimating equation in the sense that* $||\mathbf{U}_n(\widehat{\boldsymbol{\beta}}, \widehat{h})|| = O_p(n^{-1/2})$.
2. *For any root-n consistent approximate solution* $\widehat{\boldsymbol{\beta}} = (\widehat{\boldsymbol{\beta}}_1^T, \widehat{\boldsymbol{\beta}}_2^T)^T$, *we have that* $P(\widehat{\boldsymbol{\beta}}_2 = \mathbf{0}) \to 1$. *Furthermore, if* $||\mathbf{U}_n(\widehat{\boldsymbol{\beta}}, \widehat{h})|| = o_p(n^{-1/2})$, *then* $\widehat{\boldsymbol{\beta}}_1$ *has the oracle asymptotic distribution*

$$\sqrt{n}(\boldsymbol{\Gamma}_{11} + \boldsymbol{\Sigma}_{11})\big[\widehat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_{10} + (\boldsymbol{\Gamma}_{11} + \boldsymbol{\Sigma}_{11})^{-1}\mathbf{b}_n\big] \to N(0, \mathbf{V}_{11})$$

as $n \to \infty$, *where* $\boldsymbol{\Gamma}_{11}$ *is defined in Condition (T3),* $\mathbf{V}_{11}$ *denotes the* $s \times s$ *submatrix in the upper-left corner of* $\mathbf{V}$, $\boldsymbol{\Sigma}_{11} = diag(q'_{\lambda_n}(|\beta_{01}|), \ldots, q'_{\lambda_n}(|\beta_{0s}|))$, *and*

$$\mathbf{b}_n = (q_{\lambda_n}(|\boldsymbol{\beta}_{01}|)sgn(\boldsymbol{\beta}_{01}), \ldots, q_{\lambda_n}(|\boldsymbol{\beta}_{0s}|)sgn(\boldsymbol{\beta}_{0s}))^T.$$

Remark 2. The property described in this theorem is often referred to as the *oracle property* of parameter estimators in the variable selection context. In addition, for a nonconvex penalty function such as SCAD, we have that $\boldsymbol{\Sigma}_{11} = \mathbf{b}_n = \mathbf{0}$ as $\lambda_n \to 0$. Hence, when $\sqrt{n}\lambda_n \to \infty$, we have $\sqrt{n}\big[\widehat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_{10}\big] \to N(0, \boldsymbol{\Gamma}_{11}^{-1}\mathbf{V}_{11}\boldsymbol{\Gamma}_{11}^{-1})$. This is the asymptotic normal distribution that would be obtained if the true model is known *a priori*.

Theorem 1 establishes the asymptotic property of the approximate estimator for a possibly non-smooth estimating function. If the unpenalized estimating function is continuous in the true parameter space, then an exact solution can be found. This leads us to investigate the property of its resulting estimator given below. Before presenting the result, let us define $\mathbf{U}_{n1}(\boldsymbol{\beta}, \widehat{h})$ and $\mathbf{M}_{n1}(\boldsymbol{\beta}, \widehat{h})$ be the subvectors that contain the first s components of $\mathbf{U}_n(\boldsymbol{\beta}, \widehat{h})$ and $\mathbf{M}_n(\boldsymbol{\beta}, \widehat{h})$, respectively.

Theorem 2. *Assume Conditions (C1)–(C5), (P1)–(P2) and (T1)–(T4) are satisfied. If* $\mathbf{M}_{n1}((\boldsymbol{\beta}_1^T, \mathbf{0}^T)^T, \widehat{h})$ *is continuous in* $\boldsymbol{\beta}_1$, *then with probability approaching one, there exists* $\widehat{\boldsymbol{\beta}}_1$ *that is root-n consistent for* $\boldsymbol{\beta}_{10}$ *and satisfies*

$$\mathbf{U}_{n1}((\widehat{\boldsymbol{\beta}}_1^T, \mathbf{0}^T)^T, \widehat{h}) = 0.$$

Furthermore, $\widehat{\boldsymbol{\beta}}_1$ *has the same asymptotic normal distribution as stated in Theorem 1(2).*

To apply the above two theorems, the main efforts lie in checking Conditions (C2)–(C5). Condition (C2) usually can be verified based on the smoothness of the population version of the objective function $M(\boldsymbol{\beta}, h)$. Condition (C3) is often satisfied for frequently used nonparametric estimators. Condition (C4) holds if we can show that the function class $\{m(\mathbf{Z}, \boldsymbol{\beta}, h) : \boldsymbol{\beta} \in \mathbb{B}, h \in \mathbb{H}\}$ is a Donsker class (e.g. van der Vaart and Wellner, 1996). In addition, the three sufficient conditions for (C4) are provided in Theorem 3 of Chen, Linton and Keilegom (2003). Condition (C5) can usually be established by applying a uniform Bahadur representation of $\widehat{h} - h_0$, which is available for commonly used nonparametric smoothers. We have checked four analytical examples, which satisfy all conditions. It is noteworthy that Chen, Linton and Keilegom (2003) examined Conditions (C4) and (C5) for a partially linear median regression model that is a special case of Example 1. For the sake of illustration, we briefly demonstrate the examination of Conditions (C4) and (C5) for Example 2 in Appendix B.

3.2. Parameter estimation

To allow for the PP semiparametric estimating function to be non-smooth, we apply the idea of the MM (majorization-minimization) algorithm to both the profiled semiparametric estimating function and the penalty function. We refer to Hunter and Lange (2004) for a general tutorial on the MM algorithm. Specifically, we first obtain the nonparametric estimate $\widehat{h}(W_i, \boldsymbol{\beta})$ for the given $\boldsymbol{\beta}$. Then, we adopt Hunter and Lange (2000)'s MM algorithm to the unpenalized profiled estimating function and Hunter and Li (2005)'s MM algorithm to the penalty function, which yields their corresponding MM functions: $\mathbf{M}_n^{\epsilon}(\boldsymbol{\beta}, \widehat{h})$ and $nq_{\lambda_n}(|\boldsymbol{\beta}|)\frac{\boldsymbol{\beta}}{\epsilon+|\boldsymbol{\beta}|}$, respectively, where the explicit form of $\mathbf{M}_n^{\epsilon}(\boldsymbol{\beta}, \widehat{h})$ depends on the specific model form under study and the constant ϵ stands for a small perturbation, which we take to be 10^{-6} in our simulation studies, see (12) below for an example. Accordingly, the penalized estimator $\widehat{\boldsymbol{\beta}} = (\widehat{\beta}_1, \ldots, \widehat{\beta}_p)^T$ approximately satisfies:

$$U_n^{\epsilon}(\widehat{\boldsymbol{\beta}}, \widehat{h}) = \mathbf{M}_n^{\epsilon}(\widehat{\boldsymbol{\beta}}, \widehat{h}) + nq_{\lambda_n}(|\widehat{\boldsymbol{\beta}}|)\frac{\widehat{\boldsymbol{\beta}}}{\epsilon + |\widehat{\boldsymbol{\beta}}|} = 0, \tag{9}$$

where the product in the last term of (9) denotes the component-wise product. It is noteworthy that $\mathbf{M}_n^{\epsilon}(\widehat{\boldsymbol{\beta}}, \widehat{h}) = \mathbf{M}_n(\widehat{\boldsymbol{\beta}}, \widehat{h})$ when $\mathbf{M}_n$ is a smooth function. To obtain $\widehat{\boldsymbol{\beta}}$, we employ the concept of the Newton-Raphson algorithm to the function $U_n^{\epsilon}(\boldsymbol{\beta}, \widehat{h})$, which yields the following iterative algorithm:

$$\widehat{\boldsymbol{\beta}}^{(k+1)} = \widehat{\boldsymbol{\beta}}^{(k)} - \left[\mathbf{H}(\widehat{\boldsymbol{\beta}}^{(k)}) + n\mathbf{E}(\widehat{\boldsymbol{\beta}}^{(k)})\right]^{-1}\left[\mathbf{M}_n^{\epsilon}(\widehat{\boldsymbol{\beta}}^{(k)}, \widehat{h}) + n\mathbf{E}(\widehat{\boldsymbol{\beta}}^{(k)})\widehat{\boldsymbol{\beta}}^{(k)}\right], \tag{10}$$

where $\mathbf{E}(\widehat{\boldsymbol{\beta}}^{(k)}) = \mathrm{diag}(q_{\lambda_n}(|\widehat{\beta}_1^{(k)}|)/(\epsilon + |\widehat{\beta}_1^{(k)}|), \ldots, q_{\lambda_n}(|\widehat{\beta}_p^{(k)}|)/(\epsilon + |\widehat{\beta}_p^{(k)}|))$, and $\mathbf{H}(\widehat{\boldsymbol{\beta}}^{(k)})$ is the derivative matrix of $\mathbf{M}_n^{\epsilon}(\boldsymbol{\beta}, \widehat{h})$ with respect to $\boldsymbol{\beta}$ evaluated at $\widehat{\boldsymbol{\beta}}^{(k)}$.

The above algorithm is iterated until certain stopping criterion is met, for example, $||\widehat{\boldsymbol{\beta}}^{(k+1)} - \widehat{\boldsymbol{\beta}}^{(k)}|| \leq 10^{-4}$. In addition, any coefficient sufficiently small is suppressed to zero, i.e., if $|\widehat{\beta}_j| \leq 10^{-4}$ upon convergence, then the estimator of this coefficient is set to be exactly zero. It is noteworthy that $\widehat{h}$ in the iterative algorithm is updated along with $\widehat{\boldsymbol{\beta}}^{(k)}$. Finally, we select the tuning parameter λ_n by minimizing a Bayesian Information Criterion (Schwarz, 1978),

$$\text{BIC}(\lambda_n) = \log\Big(L_n(\widehat{\boldsymbol{\beta}}, \widehat{h})\Big) + \text{df}_{\lambda_n}\frac{\log(n)}{n}, \tag{11}$$

where $L_n(\widehat{\boldsymbol{\beta}}, \widehat{h})$ is the loss function that leads to $\mathbf{M}_n(\widehat{\boldsymbol{\beta}}, \widehat{h})$ and the effective number of parameters is

$$\text{df}_{\lambda_n} = \text{trace}\big\{(\mathbf{H}(\widehat{\boldsymbol{\beta}}) + n\mathbf{E}(\widehat{\boldsymbol{\beta}}))^{-1}\mathbf{H}(\widehat{\boldsymbol{\beta}})\big\}.$$

For the sake of illustration, we revisit Example 1 by briefly presenting the estimating equation and its relevant quantities. Based on equations (3) and (4), the penalized estimator of partially linear quantile regression satisfies the following equation,

$$n^{-1}\sum_{i=1}^{n}\mathbf{X}_i\Big[I(Y_i \leq \mathbf{X}_i^T\boldsymbol{\beta} + \widehat{h}(W_i, \boldsymbol{\beta})) - \tau\Big] + q_{\lambda_n}(|\boldsymbol{\beta}|)\text{sgn}(\boldsymbol{\beta}) = 0.$$

Note that to estimate $h(W, \boldsymbol{\beta})$, the minimization of the objective function in (5) can be solved using existing software packages, for example, the quantile regression package in R. Furthermore, the non-penalized MM function is

$$\mathbf{M}_n^{\epsilon}(\widehat{\boldsymbol{\beta}}, \widehat{h}) = -\frac{1}{2}\sum_{i=1}^{n}\mathbf{X}_i\frac{Y_i - \mathbf{X}_i^T\widehat{\boldsymbol{\beta}} - \widehat{h}(W_i, \widehat{\boldsymbol{\beta}})}{\epsilon + |Y_i - \mathbf{X}_i^T\widehat{\boldsymbol{\beta}} - \widehat{h}(W_i, \widehat{\boldsymbol{\beta}})|} - \Big(\tau - \frac{1}{2}\Big)\sum_{i=1}^{n}\mathbf{X}_i. \tag{12}$$

4. Numerical results

In this section, we use the SCAD penalty function defined in (8) with $a = 3.7$ for both simulations and real data analyses.

4.1. Monte Carlo simulated examples

To evaluate the finite sample performance of the proposed method, we first consider the partially linear quantile regression model given in Example 1,

$$Y_i = \mathbf{X}_i^T\boldsymbol{\beta}_0 + h_0(W_i) + \sigma\epsilon_i, \quad i = 1, \ldots, n, \tag{13}$$

where $\boldsymbol{\beta}_0 = (3, 1.5, 0, 0, 2, 0, 0, 0)^T$ and ϵ_i is the random error. As a result, the number of nonzero coefficients is 3. Furthermore, the vectors $\mathbf{X}_i$ are generated from a multivariate normal distribution with mean $\mathbf{0}$ and an AR-1 correlation matrix with the auto-correlation coefficient 0.5. The covariates W_i are simulated

from a uniform (0,1) distribution, and they are independent of $\mathbf{X}_i$ and ϵ_i. Moreover, we consider two nonparametric functions: $h_0(w) = 2\sin(4\pi w)$ adapted from Fan and Huang (2005) and $h_0(w) = 16w(1-w) - 2$ adapted from Li and Liang (2008); two values for σ: 1 and 3; two sample sizes: $n = 200$ and 400; three different quantile levels: $\tau = 0.25$, 0.5, 0.75, and four error distributions of ϵ_i: (1) the standard normal distribution, (2) the t distribution with 3 degrees of freedom, (3) the mixture normal distribution with heavy tails: $0.9N(0,1) + 0.1N(0,10^2)$, and (4) the Gamma(2,2) distribution. We standardize ϵ_i such that it satisfies $P(\epsilon_i \le 0|\mathbf{X}_i, W_i) = \tau$ for a given quantile level τ of interest. For each of the above settings, a total of 500 realizations are conducted.

To assess the model selection properties, we report the average number of nonzero coefficients that are correctly estimated to be nonzero (labeled 'C'), the average number of zero coefficients that are incorrectly estimated to be nonzero (labeled 'I'), and the proportion of the selected model being underfitted (missing any significant variables, labeled 'UF'), correctly fitted (being the exact subset model, labeled 'CF') and overfitted (including all significant variables and some noise variables, labeled 'OF'). To examine the estimation accuracy, we report the mean squared error (MSE), $500^{-1}\sum_{m=1}^{500}||\widehat{\boldsymbol{\beta}}_{(m)} - \boldsymbol{\beta}||^2$, where $\widehat{\boldsymbol{\beta}}_{(m)}$ is the estimate from the mth realization. As a benchmark, we also compute the mean squared error of the oracle estimate (in parentheses), which is the un-penalized quantile estimate of the true model.

When $n = 200$ and 400, Tables 1 and 2, respectively, present the results for the partially linear quantile regression model with the nonparametric function $h_0(w) = 2\sin(4\pi w)$. We observe the following important findings. (i) As the sample size gets larger, MSE becomes smaller and approaches that of the oracle estimate, which is consistent with the theoretical finding. When the signal gets stronger (i.e., σ decreases from 3 to 1), the measurements of MSE, I, UF and OF decrease and those of C and CF increase as expected. (ii) In the symmetric distributions, which are standard normal, t_3, and mixture, it is not surprising that $\tau = 0.5$ yields better performance than $\tau = 0.25$ and $\tau = 0.75$ in terms of all measurements. In the positively skewed Gamma(2,2) distribution, it is also sensible that $\tau = 0.25$ outperforms $\tau = 0.5$ and $\tau = 0.75$. It is noteworthy that the proportion of underfitted models is high for the Gamma(2,2) distribution with $\sigma = 3$ and $\tau = 0.75$. This is because the signal is too weak in this case, due to a large variance and skewness. Because the simulations with the nonparametric function $h_0(w) = 16w(1-w) - 2$ exhibit similar results, we do not present them here to save space.

To further illustrate the proposed method, we next generate random data from a partially linear mean regression model with missing covariates given in Example 3,

$$Y_i = \mathbf{X}_i^T\boldsymbol{\beta}_0 + h_0(W_i) + \sigma\epsilon_i, \quad i = 1, \ldots, n, \tag{14}$$

where the ϵ_i are independently generated from a $N(0,1)$ distribution and $\boldsymbol{\beta}_0$ is the same as that in (13). The variables $\mathbf{X}_i$ and W_i are also generated from the same distributions as in the previous example. Moreover, $h_0(w)$, n, and σ are defined as above. Let $\delta_i = 1$ if $\mathbf{X}_i$ is observed; and $\delta_i = 0$ otherwise. Then,

TABLE 1
Simulation results of quantile regression with $h(w) = 2\sin(4\pi w)$ *and* $n = 200$

			No. of nonzeros		**Proportion**		
σ	τ	**MSE(Oracle)**	**C**	**I**	**UF**	**CF**	**OF**
			Standard Normal				
	0.25	0.0562(0.0372)	3.000	0.360	0.000	0.760	0.240
1	0.5	0.0409(0.0321)	3.000	0.216	0.000	0.834	0.166
	0.75	0.0555(0.0392)	3.000	0.312	0.000	0.778	0.222
			t-distribution with df $= 3$				
	0.25	0.0872(0.0588)	3.000	0.286	0.000	0.802	0.198
1	0.5	0.0505(0.0405)	3.000	0.160	0.000	0.868	0.132
	0.75	0.0893(0.0580)	3.000	0.290	0.000	0.790	0.210
			Mixture $0.9N(0,1) + 0.1N(0,10^2)$				
	0.25	0.0545(0.0461)	3.000	0.114	0.000	0.904	0.096
1	0.5	0.0381(0.0369)	3.000	0.086	0.000	0.926	0.074
	0.75	0.0594(0.0523)	3.000	0.110	0.000	0.900	0.100
			Gamma(2,2)				
	0.25	0.1526(0.1105)	3.000	0.224	0.000	0.820	0.180
1	0.5	0.3093(0.2082)	3.000	0.286	0.000	0.768	0.232
	0.75	0.8415(0.4255)	2.986	0.580	0.014	0.602	0.384
			Standard Normal				
	0.25	0.5474(0.3150)	2.994	0.382	0.006	0.714	0.280
3	0.5	0.4450(0.2790)	2.992	0.292	0.008	0.758	0.234
	0.75	0.5701(0.3387)	2.990	0.416	0.010	0.692	0.298
			t-distribution with df $= 3$				
	0.25	0.9514(0.5131)	2.928	0.300	0.072	0.712	0.216
3	0.5	0.5739(0.3542)	2.970	0.206	0.030	0.800	0.170
	0.75	1.1048(0.5054)	2.910	0.356	0.090	0.656	0.254
			Mixture $0.9N(0,1) + 0.1N(0,10^2)$				
	0.25	0.8230(0.4049)	2.896	0.136	0.104	0.798	0.098
3	0.5	0.4645(0.3170)	2.960	0.086	0.040	0.890	0.070
	0.75	0.8021(0.4401)	2.896	0.120	0.104	0.804	0.092
			Gamma(2,2)				
	0.25	2.5188(0.9616)	2.618	0.350	0.362	0.486	0.152
3	0.5	4.9104(1.8305)	2.264	0.384	0.626	0.258	0.116
	0.75	10.8207(3.7964)	1.816	0.692	0.846	0.064	0.090

consider the case where the covariates $\mathbf{X}_i$ are missing at random in the sense that $\pi(Y_i, W_i) = P(\delta_i = 1|\mathbf{X}_i, Y_i, W_i) = P(\delta_i = 1|Y_i, W_i)$. Subsequently, we employ logistic regression to generate the missing data indicators:

$$P(\delta_i = 1|Y_i, W_i) = \frac{\exp(\gamma_0 + \gamma_1 Y_i + \gamma_2 W_i)}{1 + \exp(\gamma_0 + \gamma_1 Y_i + \gamma_2 W_i)}.$$

To assess the sensitivity of parameter estimates against the missing rate, we study the following four cases: Case 1: $(\gamma_0, \gamma_1, \gamma_2) = (1, 1, 2)$; Case 2: $(\gamma_0, \gamma_1, \gamma_2) = (3, 1, 2)$; Case 3: $(\gamma_0, \gamma_1, \gamma_2) = (6, 1, 2)$; and Case 4: $(\gamma_0, \gamma_1, \gamma_2) = (8, 1, 2)$. The average missing rates are approximately 0.35, 0.25, 0.10 and 0.05, respectively. Based on the simulated data from each of the four cases, we are able to estimate

TABLE 2
Simulation results of quantile regression with $h(w) = 2\sin(4\pi w)$ *and* $n = 400$

			No. of nonzeros		Proportion		
σ	τ	MSE(Oracle)	C	I	UF	CF	OF
			Standard Normal				
	0.25	0.0247(0.0183)	3.000	0.262	0.000	0.822	0.178
1	0.5	0.0188(0.0150)	3.000	0.186	0.000	0.860	0.140
	0.75	0.0245(0.0187)	3.000	0.248	0.000	0.816	0.184
			t-distribution with df $= 3$				
	0.25	0.0400(0.0277)	3.000	0.232	0.000	0.810	0.190
1	0.5	0.0225(0.0191)	3.000	0.142	0.000	0.878	0.122
	0.75	0.0386(0.0249)	3.000	0.272	0.000	0.812	0.188
			Mixture $0.9N(0,1) + 0.1N(0,10^2)$				
	0.25	0.0270(0.0252)	3.000	0.082	0.000	0.924	0.076
1	0.5	0.0180(0.0169)	3.000	0.100	0.000	0.912	0.088
	0.75	0.0250(0.0229)	3.000	0.066	0.000	0.936	0.064
			Gamma(2,2)				
	0.25	0.0756(0.0558)	3.000	0.216	0.000	0.820	0.180
1	0.5	0.1305(0.0959)	3.000	0.188	0.000	0.854	0.146
	0.75	0.3250(0.1985)	3.000	0.338	0.000	0.734	0.266
			Standard Normal				
	0.25	0.2537(0.1611)	3.000	0.302	0.000	0.762	0.238
3	0.5	0.1923(0.1301)	3.000	0.206	0.000	0.826	0.174
	0.75	0.2484(0.1609)	2.998	0.282	0.002	0.776	0.222
			t-distribution with df $= 3$				
	0.25	0.3795(0.2484)	2.996	0.212	0.004	0.820	0.176
3	0.5	0.2233(0.1680)	2.998	0.124	0.002	0.888	0.110
	0.75	0.3628(0.2279)	3.000	0.264	0.000	0.794	0.206
			Mixture $0.9N(0,1) + 0.1N(0,10^2)$				
	0.25	0.2628(0.2220)	2.998	0.074	0.002	0.936	0.062
3	0.5	0.1685(0.1491)	3.000	0.040	0.000	0.960	0.040
	0.75	0.2419(0.1962)	2.998	0.068	0.002	0.932	0.066
			Gamma(2,2)				
	0.25	0.8225(0.4912)	2.928	0.198	0.072	0.764	0.164
3	0.5	1.8962(0.8534)	2.732	0.242	0.264	0.594	0.142
	0.75	5.1416(1.8160)	2.332	0.492	0.562	0.292	0.146

$(\gamma_0, \gamma_1, \gamma_2)$ from the above logistic regression model and then get $\widehat{\pi}_i$. Since simulation settings lead to $E((X - E(X))\epsilon|Y, W) = 0$, we could follow Liang et al. (2004)'s comment and use the first part of function Φ defined after equation (7), together with the estimation process of Section 3.2, to obtain the penalized estimates. Finally, the tuning parameter is selected by minimizing BIC(λ_n) in equation (11).

When $h_0(w) = 2\sin(4\pi w)$, Table 3 indicates that MSE decreases and approaches that of the oracle estimate when the sample size becomes large, which confirms the theoretical result. It is also not surprising that the measurements of MSE, I, UF, and OF decrease and C and CF increase as σ decreases from 3 to 1. Since the missing rate decreases from Case 1 to Case 4, it is sensible that Case 4

TABLE 3
Simulation results of missing covariates with $h(w) = 2\sin(4\pi w)$

Case	MSE(Oracle)	No. of nonzeros		Proportion		
		C	I	UF	CF	OF
		$n = 200, \sigma = 1$				
1	0.0773(0.0593)	3.000	0.206	0.000	0.814	0.186
2	0.0492(0.0408)	3.000	0.130	0.000	0.876	0.124
3	0.0312(0.0266)	3.000	0.068	0.000	0.934	0.066
4	0.0253(0.0218)	3.000	0.060	0.000	0.942	0.058
		$n = 200, \sigma = 3$				
1	1.1672(0.8464)	2.928	0.268	0.070	0.712	0.218
2	0.7538(0.5308)	2.966	0.234	0.034	0.758	0.208
3	0.3862(0.2992)	2.984	0.104	0.016	0.890	0.094
4	0.3074(0.2449)	2.996	0.108	0.004	0.904	0.092
		$n = 400, \sigma = 1$				
1	0.0459(0.0372)	3.000	0.196	0.000	0.824	0.176
2	0.0280(0.0224)	3.000	0.146	0.000	0.866	0.134
3	0.0160(0.0139)	3.000	0.066	0.000	0.938	0.062
4	0.0127(0.0117)	3.000	0.040	0.000	0.960	0.040
		$n = 400, \sigma = 3$				
1	0.7508(0.5802)	2.988	0.278	0.010	0.750	0.240
2	0.4681(0.3525)	2.992	0.248	0.008	0.768	0.224
3	0.2249(0.1870)	2.998	0.114	0.002	0.892	0.106
4	0.1556(0.1310)	3.000	0.086	0.000	0.920	0.080

performs the best while Case 1 performs the worst in terms of all assessing measures. Moreover, the nonparametric function, $h_0(w) = 16w(1-w) - 2$, yields similar findings, which we omit here to save space. In summary, our proposed estimates perform well for simultaneous estimation and variable selection.

4.2. A real example

To demonstrate the practical usefulness of the proposed method, we consider the Female Labor Supply data collected in East Germany that has been analyzed by Fan, Härdle and Mammen (1998). The data set consists of 607 observations, and the response variable y is the 'wage per hour'. There are eight explanatory variables: x_1 is the number of working hours in a week (HRS); x_2 is the 'Treiman prestige index' of the woman's job (PRTG); x_3 is the monthly net income of the woman's husband (HUS); x_4 and x_5 are dummy variables for the woman's education (EDU): $x_4 = 1$ if the woman received between 13 and 16 years of education, and $x_4 = 0$ otherwise (EDU_1); $x_5 = 1$ if the woman received at least 17 years of education, and $x_5 = 0$ otherwise (EDU_2); x_6 is a dummy variable for children (CLD): $x_6 = 1$ if the woman has children less than 16 years old, and $x_6 = 0$ otherwise; x_7 is the unemployment rate in the place where she lives (UNEM); and w is her age.

Recently, Wang, Li and Tsai (2007) employed the penalized partially linear mean regression model to fit the data by including a nonparametric component

TABLE 4
Estimated coefficients and their standard errors for Female Labor Supply data

Variable	$\tau = 0.25$	$\tau = 0.5$	$\tau = 0.75$
HRS(x_1)	−0.556(0.036)	−0.592(0.035)	−0.727(0.072)
PRTG(x_2)	1.209(0.038)	1.494(0.043)	1.527(0.048)
HUS(x_3)	0	0	0
$\text{EDU}_1(x_4)$	1.281(0.073)	1.016(0.061)	0.949(0.091)
$\text{EDU}_2(x_5)$	2.307(0.137)	3.040(0.150)	3.434(0.216)
CLD(x_6)	−0.438(0.052)	0	0.607(0.075)
UNEM(x_7)	0	0	0
x_1^2	−0.344(0.018)	−0.330(0.015)	0
x_2^2	0.267(0.027)	0	0.275(0.026)
x_3^2	0	0	0
x_1x_2	0	0	−0.365(0.056)
x_1x_3	0	0	0
x_2x_3	0	0	0

w and seven linear main effects together with some of the first-order interaction effects among x_1, x_2 and x_3. The covariates x_1, x_2, x_3 and x_7 were standardized. To further understand the relationship between the wage and other variables, we adopt the quantile regression model given in analytic Example 1, which could provide more comprehensive and insightful findings. To this end, we consider $\tau = 0.25, 0.5$, and 0.75, which correspond to the responses of lower-paid females, middle-paid females, and well-paid females, respectively. After preliminary analyses, one observation with Age=60 is deleted because it is an outlier that has high leverage and low response. Then, we apply the five-fold cross validation method to choose smoothing bandwidths for $\hat{h}(w)$, which are $t_{\tau=0.25} = 7.63$, $t_{\tau=0.5} = 4.13$, $t_{\tau=0.75} = 4.56$. Subsequently, we employ the BIC criterion to select the tuning parameters λ_n, which are $\lambda_{n,\tau=0.25} = 0.061$, $\lambda_{n,\tau=0.5} = 0.093$, and $\lambda_{n,\tau=0.75} = 0.073$. Accordingly, the penalized profile estimates are obtained. In addition, we adapt equation (4.1) of Hunter and Li (2005) to compute the standard errors of parameter estimates.

Table 4 reports the penalized regression estimates and their standard errors that yield the following interesting results. (a) The associated coefficient estimates of x_1, x_2, x_1^2, x_2^2, and x_1x_2 indicate that a unit increase in HRS has a larger negative impact on middle-paid females than on lower-paid females. In addition, it leads to a stronger negative effect on well-paid females when PRTG is at a higher level than when PRTG is at a lower level. In contrast, a unit increase in PRTG has a larger positive impact on middle-paid females than on lower-paid females. Moreover, it yields a smaller positive effect on well-paid females when HRS is at a higher level than when HRS is at a lower level. (b) The associated coefficient estimates of x_4 (EDU_1) and x_5 (EDU_2) indicate that higher education usually yields a larger positive effect on well-paid females than on middle-paid and lower-paid females. (c) It is not surprising that variable x_6 (CLD) is not selected into the median regression, since it has not been included in the mean regression (see Wang, Li and Tsai, 2007). However, it is chosen into the quantile regression models with $\tau = 0.25$ and $\tau = 0.75$. The

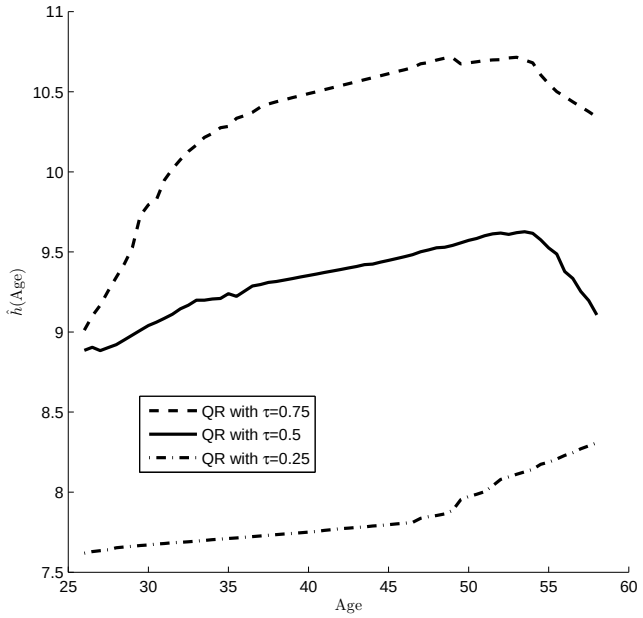

FIG 1. *The plot of* $\hat{h}(Age)$ *versus Age for Female Labor Study data.*

associated coefficient estimates indicate that, for well-paid females with young children, they are better motivated and have the ability to earn more; while, for lower-paid females with young children, their salaries are negatively affected possibly due to limited skills and time spent on child care. This result demonstrates that quantile regressions could provide more comprehensive findings than mean regression alone. (d) Two variables, x_3 (HUS) and x_7 (UNEM), are not selected in any of the quantile regression models. Hence, they do not appear to affect the hourly wage.

Figure 1 depicts the estimated nonparametric functions $\hat{h}(w)$ for all three quantile models. It indicates that the difference between starting wage at age 26 for well-paid versus middle-paid females is much smaller than that between middle-paid and lower-paid females. In addition, between ages 26 and 33, the rate of growth in wage of well-paid females increases much faster than that of middle-paid females. Afterward, these two groups exhibit similar rates of growth and decrease. Moreover, the rate of growth in wage of lower-paid females increases faster after age 48. This is because they have more time and experience to earn higher wages. In sum, the starting wage and the strong rate of growth in wage at earlier age play a significant role in females' lifetime earnings.

5. Conclusion and discussions

In this paper, we study a class of penalized profiled semiparametric estimating functions that are flexible enough to incorporate nonlinearity and non-

smoothness. Hence, they cover various regression models, such as quantile regression, survival regression, and regression with missing data. Under very general conditions, we establish the oracle property of the resulting estimator for parametric components.

The oracle property implies that the regularized estimator for the subvector of nonzero coefficients has the asymptotic variance as that of the estimator based on the unpenalized estimating equation when the true model is known *a priori*. Hence, when the moment condition in (1) comes from a semiparametric efficient score function, it is expected that the corresponding regularized estimator achieves the semiparametric efficiency bound for estimating the subvector of nonzero coefficients. For instance, consider the mean single-index regression model in Example 2, and let $(\mathbf{x}_{1i}^T, \mathbf{x}_{2i}^T)$ be a partition of $\mathbf{X}_i$ corresponding to $(\boldsymbol{\beta}_{10}^T, \boldsymbol{\beta}_{20}^T)$. A direct calculation reveals that the regularized estimator with the SCAD penalty for $\boldsymbol{\beta}_{10}$ has an asymptotic covariance matrix $\boldsymbol{\Gamma}_{11}^{-1}\mathbf{V}_{11}\boldsymbol{\Gamma}_{11}^{-1}$, where

$$\boldsymbol{\Gamma}_{11} = -E\big[h_0'(\mathbf{x}_{1i}^T\boldsymbol{\beta}_{10})^2\{\mathbf{x}_{1i} - E(\mathbf{x}_{1i}|\mathbf{x}_{1i}^T\boldsymbol{\beta}_{10})\}\{\mathbf{x}_{1i} - E(\mathbf{x}_{1i}|\mathbf{x}_{1i}^T\boldsymbol{\beta}_{10})\}^T\big],$$

$$\mathbf{V}_{11} = E\big[h_0'(\mathbf{x}_{1i}^T\boldsymbol{\beta}_{10})^2\{\mathbf{x}_{1i} - E(\mathbf{x}_{1i}|\mathbf{x}_{1i}^T\boldsymbol{\beta}_{10})\}\{\mathbf{x}_{1i} - E(\mathbf{x}_{1i}|\mathbf{x}_{1i}^T\boldsymbol{\beta}_{10})\}^T\sigma^2(\mathbf{x}_{1i})\big],$$

and $\sigma^2(\mathbf{x}_{1i}) = E(\epsilon^2|\mathbf{x}_{1i})$. When the error is homoscedastic, one then can apply Carroll et al. (1997) result and show that the proposed regularized estimator asymptotically achieves the semiparametric efficiency bound for estimating $\boldsymbol{\beta}_{10}$. For the partially linear quantile regression discussed in Example 1, one can use the semiparametric efficiency score derived in Section 5 of Lee (2003), which requires estimating the conditional error density function. In general, obtaining a semiparametric efficient estimator can be computationally cumbersome. For example, for the missing data problem discussed in Example 3, Liang et al. (2004) in their Section 4.1 pointed out that one needs to solve a complex integral equation to obtain the optimal weight for the semiparametric efficient score function.

To further explore the proposed function, one could link the current work to ultrahigh dimensional analysis by incorporating the screening methods from Fan and Lv (2008), Wang (2009), Fan, Feng and Song (2011), and Liang, Wang and Tsai (2012). It is also of interest to extend the estimation function to nonlinear time series models (see Fan and Yao, 2003) and financial time series models (see Tsay, 2005). We believe that these efforts would broaden the usefulness of the penalized profiled semiparametric estimating function.

Appendix A: Technical proofs

Proof of Theorem 1

(1) Assume that the non-zero and zero components of $\boldsymbol{\beta}_0$ are known *a prior*. Then, we can estimate the vector of nonzero coefficients $\boldsymbol{\beta}_{10}$ by solving the s-dimensional profiled estimation function $\mathbf{M}_{n1}(\boldsymbol{\beta}_1, \widehat{h}) = 0$, where $\mathbf{M}_{n1}(\boldsymbol{\beta}_1, \widehat{h})$ denotes the subvector that consists of the first s components of $\mathbf{M}_n(\boldsymbol{\beta}, \widehat{h})$ evaluated at $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^T, \mathbf{0}^T)^T$. Applying the result of Chen, Linton and Keilegom

(2003), under Conditions (C1)–(C5) and (T1)–(T3), there exists an approximate solution $\widehat{\boldsymbol{\beta}}_1$ satisfying $\widehat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_{10} = O_p(n^{-1/2})$. We next consider the oracle estimator $\widehat{\boldsymbol{\beta}} = (\widehat{\boldsymbol{\beta}}_1^T, \mathbf{0}^T)^T$ of $\boldsymbol{\beta}_0$, and show that it is an approximate solution to the minimization problem of $\min_{\beta\in\mathbb{B}} ||\mathbf{U}_n(\boldsymbol{\beta}, \widehat{h})||$.

By Conditions (T4) and (P2) and the observation $\widehat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_{10} = O_p(n^{-1/2})$, we have that $\min_{1\le j\le s} |\beta_{0j}|/\lambda_n \to \infty$ and $||\widehat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_{10}|| = o_p(\lambda_n)$, respectively. Hence, $\forall \nu > 0$,

$$\begin{aligned} & P(\min_{1\le j\le s} |\beta_{0j}| - ||\boldsymbol{\beta}_{10} - \widehat{\boldsymbol{\beta}}_{10}|| \ge \nu\lambda_n) \\ = \;& P(||\boldsymbol{\beta}_{10} - \widehat{\boldsymbol{\beta}}_{10}|| \le \min_{1\le j\le s} |\beta_{0j}| - \nu\lambda_n) \to 1. \end{aligned}$$

This, together with the fact that $\min_{1\le j\le s} |\widehat{\beta}_j| \ge \min_{1\le j\le s} |\beta_{0j}| - \max_{1\le j\le s} |\beta_{0j} - \widehat{\beta}_j| \ge \min_{1\le j\le s} |\beta_{0j}| - ||\boldsymbol{\beta}_{10} - \widehat{\boldsymbol{\beta}}_{10}||$, leads to $P(\min_{1\le j\le s} |\widehat{\beta}_j| \ge \nu\lambda_n) \to 1$ as $n \to \infty$. By Condition (P1), we then have $q_{\lambda_n}(|\widehat{\beta}_j|) = o_p(n^{-1/2})$ for $j = 1, \ldots, s$. Consequently, $\mathbf{U}_n(\widehat{\boldsymbol{\beta}}, \widehat{h}) = \mathbf{M}_n(\widehat{\boldsymbol{\beta}}, \widehat{h}) + o_p(n^{-1/2})$.

Under Conditions (C3) and (C4), the oracle estimator $\widehat{\boldsymbol{\beta}}$ satisfies

$$||\mathbf{M}_n(\widehat{\boldsymbol{\beta}}, \widehat{h}) - \mathbf{M}(\widehat{\boldsymbol{\beta}}, \widehat{h}) - \mathbf{M}_n(\boldsymbol{\beta}_0, h_0)|| = o_p(n^{-1/2}).$$

Applying the triangle inequality and using $\mathbf{M}(\boldsymbol{\beta}_0, h_0) = 0$, we have

$$\begin{aligned} ||\mathbf{M}_n(\widehat{\boldsymbol{\beta}}, \widehat{h})|| \;\le\; & ||\mathbf{M}(\widehat{\boldsymbol{\beta}}, \widehat{h}) + \mathbf{M}_n(\boldsymbol{\beta}_0, h_0)|| + o_p(n^{-1/2}) \\ = \;& ||\mathbf{M}(\widehat{\boldsymbol{\beta}}, \widehat{h}) - \mathbf{M}(\widehat{\boldsymbol{\beta}}, h_0) - \boldsymbol{\Gamma}_2(\widehat{\boldsymbol{\beta}}, h_0)[\widehat{h} - h_0]|| \\ & + ||\boldsymbol{\Gamma}_2(\widehat{\boldsymbol{\beta}}, h_0)[\widehat{h} - h_0] - \boldsymbol{\Gamma}_2(\boldsymbol{\beta}_0, h_0)[\widehat{h} - h_0]|| \\ & + ||\mathbf{M}(\widehat{\boldsymbol{\beta}}, h_0) - \mathbf{M}(\boldsymbol{\beta}_0, h_0)|| \\ & + ||\mathbf{M}_n(\boldsymbol{\beta}_0, h_0) + \boldsymbol{\Gamma}_2(\boldsymbol{\beta}_0, h_0)[\widehat{h} - h_0]|| + o_p(n^{-1/2}) \\ = \;& o_p(n^{-1/2}) + o_p(n^{-1/2}) + O_p(n^{-1/2}) + O_p(n^{-1/2}) + o_p(n^{-1/2}) \\ = \;& O_p(n^{-1/2}), \end{aligned}$$

where the last equality follows from the fact $\widehat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_{10} = O_p(n^{-1/2})$ and Conditions (C2), (C1) and (C4). Hence, $||\mathbf{U}_n(\widehat{\boldsymbol{\beta}}, \widehat{h})|| = O_p(n^{-1/2})$.

(2) We first demonstrate that for any root-n consistent approximate estimator $\widehat{\boldsymbol{\beta}} = (\widehat{\boldsymbol{\beta}}_1^T, \widehat{\boldsymbol{\beta}}_2^T)^T$, $P(\widehat{\boldsymbol{\beta}}_2 = \mathbf{0}) \to 1$ as $n \to \infty$. The proof follows similar ideas to those used for proving Theorem 1(b) in Johnson, Lin and Zeng (2008). We first note that $\widehat{\beta}_j = O_p(n^{-1/2})$ for $j = s+1, \ldots, p$. Hence, $\forall\ \kappa > 0$, there exists $d_1 > 0$ such that, for sufficiently large n,

$$P(\widehat{\beta}_j \ne 0) \le \frac{\kappa}{2} + P(\widehat{\beta}_j \ne 0, |\widehat{\beta}_j| < d_1 n^{-1/2}). \tag{15}$$

By the definition of an approximate estimator and Conditions (C3) and (C4), we obtain that

$$M_j(\widehat{\boldsymbol{\beta}}, \widehat{h}) + M_{nj}(\boldsymbol{\beta}_0, h_0) + q_{\lambda_n}(|\widehat{\beta}_j|)\mathrm{sgn}(\widehat{\beta}_j) = O_p(n^{-1/2}),$$

where M_j and M_{nj} are the jth component of $\mathbf{M}$ and $\mathbf{M}_n$, respectively. This implies

$$\begin{aligned}&[M_j(\widehat{\boldsymbol{\beta}},\widehat{h}) - M_j(\widehat{\boldsymbol{\beta}},h_0) - \mathbf{e}_j^T\boldsymbol{\Gamma}_2(\widehat{\boldsymbol{\beta}},h_0)[\widehat{h}-h_0]] + [\mathbf{e}_j^T\boldsymbol{\Gamma}_2(\widehat{\boldsymbol{\beta}},h_0)[\widehat{h}-h_0]\\ &- \mathbf{e}_j^T\boldsymbol{\Gamma}_2(\boldsymbol{\beta}_0,h_0)[\widehat{h}-h_0]] + [M_j(\widehat{\boldsymbol{\beta}},h_0) - M_j(\boldsymbol{\beta}_0,h_0)]\\ &+ [M_{nj}(\boldsymbol{\beta}_0,h_0) + \mathbf{e}_j^T\boldsymbol{\Gamma}_2(\boldsymbol{\beta}_0,h_0)[\widehat{h}-h_0]] + q_{\lambda_n}(|\widehat{\beta}_j|)\mathrm{sgn}(\widehat{\beta}_j) = O_p(n^{-1/2}),\end{aligned}$$

where $\mathbf{e}_j$ is a unit vector with the jth component being one and all the other components being zero. By the root-n consistency of $\widehat{\boldsymbol{\beta}}$ and Conditions (C1)–(C4), it is straightforward to see that the first four terms on the left-hand side are of order $O_p(n^{-1/2})$. Thus, $q_{\lambda_n}(|\widehat{\beta}_j|)\mathrm{sgn}(\widehat{\beta}_j) = O_p(n^{-1/2})$. Accordingly, there exists $d_2 > 0$ such that, for sufficiently large n,

$$P(\widehat{\beta}_j \neq 0, |\widehat{\beta}_j| < d_1 n^{-1/2}, n^{1/2} q_{\lambda_n}(|\widehat{\beta}_j|) > d_2) < \kappa/2. \tag{16}$$

By the root-n consistency and Condition (P1), $n^{1/2}q_{\lambda_n}(|\widehat{\beta}_j|) > d_2$ for sufficiently large n. This, together with equations (15) and (16), leads to, for sufficiently large n,

$$P(\widehat{\beta}_j \neq 0) \leq \epsilon/2 + P(\widehat{\beta}_j \neq 0, |\widehat{\beta}_j| < d_1 n^{-1/2}, n^{1/2} q_{\lambda_n}(|\widehat{\beta}_j|) > d_2) < \kappa.$$

It follows that $P(\widehat{\beta}_j = 0, j = s+1, \ldots, p) \to 0$.

We next show the asymptotic normality of $\widehat{\boldsymbol{\beta}}_1$ when the order of $||\mathbf{U}_n(\widehat{\boldsymbol{\beta}})||$ is $o_p(n^{-1/2})$. Applying Conditions (C1)–(C4), we have

$$\begin{aligned}&[\mathbf{M}_1(\widehat{\boldsymbol{\beta}},h_0) - \mathbf{M}_1(\boldsymbol{\beta}_0,h_0)] + [\mathbf{M}_{n1}(\boldsymbol{\beta}_0,h_0) + (\boldsymbol{\Gamma}_2(\boldsymbol{\beta}_0,h_0)[\widehat{h}-h_0])_1]\\ &+ q_{\lambda_n}(|\widehat{\boldsymbol{\beta}}_1|)\mathrm{sgn}(\widehat{\boldsymbol{\beta}}_1) = o_p(n^{-1/2}),\end{aligned}$$

where $(\boldsymbol{\Gamma}_2(\boldsymbol{\beta}_0,h_0)[\widehat{h}-h_0])_1$ denotes the subvector that contains the first s components of $\boldsymbol{\Gamma}_2(\boldsymbol{\beta}_0,h_0)[\widehat{h}-h_0]$. Employing Taylor expansions of $\mathbf{M}_1(\widehat{\boldsymbol{\beta}},h_0)$ and $q_{\lambda_n}(|\widehat{\boldsymbol{\beta}}_1|)\mathrm{sgn}(\widehat{\boldsymbol{\beta}}_1)$ at $\boldsymbol{\beta}_0$ yields

$$\begin{aligned}&(\boldsymbol{\Gamma}_{11} + o_p(1))(\widehat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_{10}) + [\mathbf{M}_{n1}(\boldsymbol{\beta}_0,h_0) + (\boldsymbol{\Gamma}_2(\boldsymbol{\beta}_0,h_0)[\widehat{h}-h_0])_1]\\ &+ q_{\lambda_n}(|\boldsymbol{\beta}_{10}|)\mathrm{sgn}(\boldsymbol{\beta}_{10}) + q'_{\lambda_n}(|\boldsymbol{\beta}_{10}|)(\widehat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_{10})(1+o_p(1)) = o_p(n^{-1/2}),\end{aligned}$$

where $\boldsymbol{\Gamma}_{11}$ is the $s \times s$ submatrix in the upper-left corner of $\boldsymbol{\Gamma}_1$. As a result,

$$\begin{aligned}&\sqrt{n}(\widehat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_{10})\\ =\ &-\sqrt{n}\big[\boldsymbol{\Gamma}_{11} + \boldsymbol{\Sigma}_{11}\big]^{-1}\big[\mathbf{M}_{n1}(\boldsymbol{\beta}_0,h_0) + (\boldsymbol{\Gamma}_2(\boldsymbol{\beta}_0,h_0)[\widehat{h}-h_0])_1 + \mathbf{b}_n\big] + o_p(1),\end{aligned}$$

where

$$\boldsymbol{\Sigma}_{11} = \mathrm{diag}(q'_{\lambda_n}(|\beta_{01}|), \ldots, q'_{\lambda_n}(|\beta_{0s}|))$$

and

$$\mathbf{b}_n = (q_{\lambda_n}(|\boldsymbol{\beta}_{01}|)\mathrm{sgn}(\boldsymbol{\beta}_{01}), \ldots, q_{\lambda_n}(|\boldsymbol{\beta}_{0s}|)\mathrm{sgn}(\boldsymbol{\beta}_{0s}))^T.$$

By Condition (C5), we then obtain that

$$\sqrt{n}(\boldsymbol{\Gamma}_{11} + \boldsymbol{\Sigma}_{11})\big[(\widehat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_{10}) + (\boldsymbol{\Gamma}_{11} + \boldsymbol{\Sigma}_{11})^{-1}\mathbf{b}_n\big] \to N(0, \mathbf{V}_{11}),$$

where $\mathbf{V}_{11}$ is the $s \times s$ submatrix in the upper-left corner of $\mathbf{V}$. This completes the proof.

Proof of Theorem 2

We first prove the existence of a root-n consistent estimator of $\boldsymbol{\beta}_{10}$. By the result 6.3.4 of Ortega and Rheinboldt (1970), it suffices to show that, $\forall\ \xi > 0$, there exists a constant $\Delta > 0$ such that, for sufficiently large n,

$$P\left(\sup_{\boldsymbol{\beta}_1 \in \mathbf{B}_{n1}} (\boldsymbol{\beta}_1 - \boldsymbol{\beta}_{10})^T \mathbf{U}_{n1}((\boldsymbol{\beta}_1^T, \mathbf{0}^T)^T, \widehat{h}) > 0\right) \geq 1 - \xi, \tag{17}$$

where $\mathbf{B}_{n1} = \{\boldsymbol{\beta}_1 :\ ||\boldsymbol{\beta}_1 - \boldsymbol{\beta}_{10}|| = \Delta/\sqrt{n}\}$.

For any $\boldsymbol{\beta}_1 \in \mathbf{B}_{n1}$, employing similar techniques as those used in the proof of Theorem 1 and Conditions (C1)–(C4), we have

$$\begin{aligned}
& (\boldsymbol{\beta}_1 - \boldsymbol{\beta}_{10})^T \mathbf{U}_{n1}((\boldsymbol{\beta}_1^T, \mathbf{0}^T)^T, \widehat{h}) \\
= \ & (\boldsymbol{\beta}_1 - \boldsymbol{\beta}_{10})^T \left[\mathbf{M}_{n1}((\boldsymbol{\beta}_1^T, \mathbf{0}^T)^T, \widehat{h}) + q_{\lambda_n}(|\boldsymbol{\beta}_1|)\mathrm{sgn}(\boldsymbol{\beta}_1)\right] \\
= \ & (\boldsymbol{\beta}_1 - \boldsymbol{\beta}_{10})^T \left[\mathbf{M}_{n1}((\boldsymbol{\beta}_1^T, \mathbf{0}^T)^T, h_0) + \boldsymbol{\Gamma}_{11}(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_{10}) + (\boldsymbol{\Gamma}_2(\boldsymbol{\beta}_0, h_0)[\widehat{h} - h_0])_1\right. \\
& \left. + o_p(n^{-1/2})\right] + (\boldsymbol{\beta}_1 - \boldsymbol{\beta}_{10})^T q_{\lambda_n}(|\boldsymbol{\beta}_1|)\mathrm{sgn}(\boldsymbol{\beta}_1) \\
= \ & (\boldsymbol{\beta}_1 - \boldsymbol{\beta}_{10})^T \left[\mathbf{M}_{n1}(\boldsymbol{\beta}_0, h_0) + (\boldsymbol{\Gamma}_2(\boldsymbol{\beta}_0, h_0)[\widehat{h} - h_0])_1\right] \\
& + (\boldsymbol{\beta}_1 - \boldsymbol{\beta}_{10})^T \boldsymbol{\Gamma}_{11}(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_{10}) \\
& + o_p(n^{-1}) + \sum_{j=1}^{s} (\beta_j - \beta_{0j}) q_{\lambda_n}(|\beta_j|)\mathrm{sgn}(\beta_j).
\end{aligned} \tag{18}$$

Applying the Cauchy-Schwarz inequality and Condition (C5), we obtain that

$$\begin{aligned}
& \left|(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_{10})^T \left[\mathbf{M}_{n1}(\boldsymbol{\beta}_0, h_0) + (\boldsymbol{\Gamma}_2(\boldsymbol{\beta}_0, h_0)[\widehat{h} - h_0])_1\right]\right| \\
\leq \ & ||\boldsymbol{\beta}_1 - \boldsymbol{\beta}_{10}|| \cdot ||\mathbf{M}_{n1}(\boldsymbol{\beta}_0, h_0) + (\boldsymbol{\Gamma}_2(\boldsymbol{\beta}_0, h_0)[\widehat{h} - h_0])_1|| \leq d_3 \Delta/n,
\end{aligned} \tag{19}$$

with large probability, for some positive constant d_3 and sufficiently large n. In addition, Condition (T3) implies that

$$(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_{10})^T \boldsymbol{\Gamma}_{11}(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_{10}) \geq \lambda_{\min}(\boldsymbol{\Gamma}_{11})||\boldsymbol{\beta}_1 - \boldsymbol{\beta}_{10}||^2 \geq d_4 \Delta^2/n \tag{20}$$

for some $d_4 > 0$, where $\lambda_{\min}(\boldsymbol{\Gamma}_{11})$ denotes the smallest eigenvalue of $\boldsymbol{\Gamma}_{11}$. Moreover, employing the Cauchy-Schwarz inequality and Condition (C5) again, we have

$$\begin{aligned}
\left|\sum_{j=1}^{s} (\beta_j - \beta_{0j}) q_{\lambda_n}(|\beta_j|)\mathrm{sign}(\beta_j)\right| & \leq ||\boldsymbol{\beta}_1 - \boldsymbol{\beta}_{10}|| \sqrt{s} q_{\lambda_n}(\min_{1 \leq j \leq s} |\boldsymbol{\beta}_{nj}|) \\
& \leq \frac{\sqrt{s}\Delta}{n} \sqrt{n} q_{\lambda_n}(\min_{1 \leq j \leq s} |\boldsymbol{\beta}_j|).
\end{aligned}$$

By Conditions (P2) and (T4), $\min_{1 \leq j \leq s} |\beta_j| \geq \min_{1 \leq j \leq s} |\beta_{0j}| - \Delta/\sqrt{n} \geq \frac{1}{2} \min_{1 \leq j \leq s} |\beta_{0j}|$ for $\boldsymbol{\beta}_1 \in \mathbf{B}_{n1}$. Then, under Condition (P1), we have $\sqrt{n} q_{\lambda_n}(\min_{1 \leq j \leq s} |\boldsymbol{\beta}_j|) \to 0$. Hence, $\left|\sum_{j=1}^{s} (\beta_j - \beta_{0j}) q_{\lambda_n}(|\beta_j|)\mathrm{sign}(\beta_j)\right| =$

$o(\Delta n^{-1})$. This, together with equations (18), (19), and (20), leads to the fact that, for a large Δ, $(\boldsymbol{\beta}-\boldsymbol{\beta}_0)^T\mathbf{U}_{n1}((\boldsymbol{\beta}_1^T,\mathbf{0}^T)^T,\widehat{h})$ is asymptotically dominated by $(\boldsymbol{\beta}_1-\boldsymbol{\beta}_{10})^T\boldsymbol{\Gamma}_{11}(\boldsymbol{\beta}_1-\boldsymbol{\beta}_{10})$, which is nonnegative. As a result, (17) holds and with probability approaching one there exists $\widehat{\boldsymbol{\beta}}_1$ that is a root-n consistent estimator for $\boldsymbol{\beta}_{10}$ and satisfies $\mathbf{U}_{n1}((\widehat{\boldsymbol{\beta}}_1^T,\mathbf{0}^T)^T,\widehat{h})=0$. Finally, the asymptotic normality of $\widehat{\boldsymbol{\beta}}_1$ can be obtained via similar techniques as those used in the proof of Theorem 1(2). This completes the proof.

Appendix B: Examination of Conditions (C4) & (C5) for Example 2

We consider the single index mean regression model defined in (6). Assume that $\mathbf{X}\in\mathbf{R}_{\mathbf{X}}$, $\boldsymbol{\beta}\in\mathbb{B}$, and that both $\mathbf{R}_{\mathbf{X}}$ and $\mathbb{B}$ are compact subsets of $\mathbf{R}^p$. The true parameter value $\boldsymbol{\beta}_0$ is assumed to be in the interior of $\mathbb{B}$. Let $\mathbb{T}=\{t: t=\mathbf{X}^T\boldsymbol{\beta},\ \mathbf{X}\in\mathbf{R}_{\mathbf{X}},\boldsymbol{\beta}\in\mathbb{B}\}$, then $\mathbb{T}$ is a compact subset of $\mathbf{R}$. We consider the following two classes of smooth functions: $\mathbb{H}=\{h(t): h(t)$ is twice continuously differentiable on $\mathbb{T}\}$ and $\mathbb{S}=\{S(\mathbf{X},\boldsymbol{\beta}): S(\mathbf{X},\boldsymbol{\beta})$ has continuous partial derivatives w.r.t $\mathbf{X}\in\mathbb{R}_{\mathbf{X}}$ and $\boldsymbol{\beta}\in\mathbb{B}\}$.

To verify (C4), we check three sufficient conditions in Theorem 3 of Chen, Linton and Keilegom (2003). Based on $\mathbb{H}$ and $\mathbb{S}$ defined above, their Conditions (3.2) and (3.3) are satisfied. To check their Condition (3.1), we further assume that $E(Y^2)$ is bounded. Then,

$$\begin{aligned}
& |m_j(\mathbf{Z},\boldsymbol{\beta}_1,h_1,s_1)-m_j(\mathbf{Z},\boldsymbol{\beta}_2,h_2,s_2)| \\
= \ & \left|s_{1j}(\mathbf{X},\boldsymbol{\beta}_1)[Y-h_1(\mathbf{X}^T\boldsymbol{\beta}_1)]-s_{2j}(\mathbf{X},\boldsymbol{\beta}_2)[Y-h_2(\mathbf{X}^T\boldsymbol{\beta}_2)]\right| \\
\le \ & |s_{1j}(\mathbf{X},\boldsymbol{\beta}_1)|\cdot\left|h_1(\mathbf{X}^T\boldsymbol{\beta}_1)-h_2(\mathbf{X}^T\boldsymbol{\beta}_2)\right| \\
& +|s_{1j}(\mathbf{X},\boldsymbol{\beta}_1)-s_{2j}(\mathbf{X},\boldsymbol{\beta}_2)|\cdot\left|Y-h_2(\mathbf{X}^T\boldsymbol{\beta}_2)\right| \\
\le \ & C_1(|Y|+C_2)(||\boldsymbol{\beta}_1-\boldsymbol{\beta}_2||+||h_1-h_2||_\infty+||s_1-s_2||_\infty),
\end{aligned}$$

where C_1 and C_2 are positive constants, and s_{1j} and s_{2j} denote the j components of s_1 and s_2, respectively. Accordingly, Condition (3.1) is satisfied and (C4) holds.

In this example, the population version of the estimating equation is

$$\mathbf{M}(\boldsymbol{\beta},h,s)=E[s(\mathbf{X},\boldsymbol{\beta})(Y-h(\mathbf{X}^T\boldsymbol{\beta}))]=E\Big[\frac{\partial h(\mathbf{X}^T\boldsymbol{\beta})}{\partial\boldsymbol{\beta}}\{h_0(\mathbf{X}^T\boldsymbol{\beta}_0)-h(\mathbf{X}^T\boldsymbol{\beta})\}\Big].$$

After algebraic simplification, we obtain the path-wise derivative of $\mathbf{M}(\boldsymbol{\beta},h,s)$ in the direction $[\overline{h}-h,\overline{s}]$:

$$\begin{aligned}
& \boldsymbol{\Gamma}_2(\boldsymbol{\beta},h,s)[\overline{h}-h,\overline{s}-s] \\
= \ & \lim_{\tau\to0}\frac{1}{\tau}E\Big[\big\{s(\mathbf{X},\boldsymbol{\beta})+\tau(\overline{s}(\mathbf{X},\boldsymbol{\beta})-s(\mathbf{X},\boldsymbol{\beta}))\big\} \\
& \cdot\big\{h_0(\mathbf{X}^T\boldsymbol{\beta}_0)-h(\mathbf{X}^T\boldsymbol{\beta})-\tau(\overline{h}(\mathbf{X}^T\boldsymbol{\beta})-h(\mathbf{X}^T\boldsymbol{\beta}))\big\} \\
& -s(\mathbf{X},\boldsymbol{\beta})\big\{h_0(\mathbf{X}^T\boldsymbol{\beta}_0)-h(\mathbf{X}^T\boldsymbol{\beta})\big\}\Big]
\end{aligned}$$

$$= E\Big[\{\overline{s}(\mathbf{X},\beta) - \frac{\partial h(\mathbf{X}^T\beta)}{\partial\beta}\}\{h_0(\mathbf{X}^T\beta_0) - h(\mathbf{X}^T\beta)\} - \frac{\partial h(\mathbf{X}^T\beta)}{\partial\beta}\{\overline{h}(\mathbf{X}^T\beta) - h(\mathbf{X}^T\beta)\}\Big].$$

Using the fact that $h(\mathbf{x}^T\boldsymbol{\beta}_0) = h_0(\mathbf{x}^T\boldsymbol{\beta}_0)$ and $\frac{\partial h(\mathbf{X}^T\boldsymbol{\beta})}{\partial\boldsymbol{\beta}}\big|_{\boldsymbol{\beta}=\boldsymbol{\beta}_0} = h'(\mathbf{X}^T\boldsymbol{\beta}_0)[\mathbf{X} - E(\mathbf{X}|\mathbf{X}^T\boldsymbol{\beta}_0)]$ and then applying Lemma 7.1 of Pagan and Ullah (1999), we obtain that

$$\begin{aligned} &\boldsymbol{\Gamma}_2(\boldsymbol{\beta}_0, h_0, s_0)[\overline{h} - h_0, \overline{s} - s_0] \\ = \;& -E\Big[h'(\mathbf{X}^T\boldsymbol{\beta}_0)\{\mathbf{X} - E(\mathbf{X}|\mathbf{X}^T\boldsymbol{\beta}_0)\}\{\overline{h}(\mathbf{X}^T\boldsymbol{\beta}) - h(\mathbf{X}^T\boldsymbol{\beta})\}\Big] = \mathbf{0}, \end{aligned}$$

where $h'(t) = \frac{d}{dt}E(Y|\mathbf{X}^T\boldsymbol{\beta} = t)$. This, together with the classical multivariate central limit theorem, leads to

$$\begin{aligned} &\sqrt{n}\Big[\mathbf{M}_n(\boldsymbol{\beta}_0, h_0, s_0) + \boldsymbol{\Gamma}_2(\boldsymbol{\beta}_0, h_0, s_0)[\widehat{h} - h_0, \widehat{s} - s_0]\Big] \\ = \;& n^{-1/2}\sum_{i=1}^{n}[Y_i - h_0(\mathbf{X}_i^T\boldsymbol{\beta}_0)]h_0'(\mathbf{X}^T\boldsymbol{\beta}_0)[\mathbf{X} - E(\mathbf{X}|\mathbf{X}^T\boldsymbol{\beta}_0)] \to N(\mathbf{0}, \mathbf{V}), \end{aligned}$$

where $\mathbf{V} = E\big[h_0'(\mathbf{X}^T\boldsymbol{\beta}_0)^2\{\mathbf{X} - E(\mathbf{X}|\mathbf{X}^T\boldsymbol{\beta}_0)\}\{\mathbf{X} - E(\mathbf{X}|\mathbf{X}^T\boldsymbol{\beta}_0)\}^T\sigma^2(\mathbf{X})\big]$ and $\sigma^2(\mathbf{X}) = E(\epsilon_i^2|\mathbf{X})$. Hence, Condition (C5) is satisfied.

Acknowledgements

The authors are very grateful to the editor, an associate editor and two referees for their constructive comments and suggestions. Wang's research was supported by NSF Grant DMS-1308960. Heuchenne's research was supported by IAP research network grant nr. P7/06 of the Belgian government (Belgian Science Policy).

References

Bickel, P. J. and Li, B. (2006). Regularization in statistics (with discussion). *Test* **15** 271–344. MR2273731

Bunea, F. (2004). Consistent covariate selection and post model selection inference in semiparametric regression. *The Annals of Statistics* **32** 898–927. MR2065193

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. MR1467842

Chen, X., Linton, O. B. and Keilegom, I. V. (2003). Estimation of semiparametric models when the criterion function is not smooth. *Econometrica* **71** 1591–1608. MR2000259

Fan, J., Feng, Y. and Song, R. (2011). Nonparametric independence screening in sparse ultra-high dimensional additive models. *Journal of the American Statistical Association* **106** 544–557. MR2847969

Fan, J., Härdle, W. and Mammen, E. (1998). Direct estimation of low-dimensional components in additive models. *The Annals of Statistics* **26** 943–971. MR1635422

Fan, J. and Huang, T. (2005). Profile likelihood inferences on semiparametric varying-coefficient partially linear models. *Bernoulli* **11** 1031–1057. MR2189080

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. MR1946581

Fan, J. and Lv, J. (2008). Sure independence screening for ultrahigh dimensional feature space (with discussion). *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **70** 849–911. MR2530322

Fan, J. and Yao, Q. (2003). *Nonlinear Time Series: Nonparametric and Parametric Methods*. Springer-Verlag, New York. MR1964455

Fu, W. J. (2003). Penalized estimating equations. *Biometrics* **59** 126–132. MR1978479

Greenshtein, E. and Ritov, Y. (2004). Persistence in high-dimensional linear predictor selection and the virtue of overparametrization. *Bernoulli* **10** 971–988. MR2108039

Hunter, D. R. and Lange, K. (2000). Quantile regression via an MM algorithm. *Journal of Computational and Graphical Statistics* **9** 60–77. MR1819866

Hunter, D. R. and Lange, K. (2004). A tutorial on MM algorithms. *The American Statistician* **58** 30–37. MR2055509

Hunter, D. R. and Li, R. (2005). Variable selection using MM algorithms. *The Annals of Statistics* **33** 1617–1642. MR2166557

Ichimura, H. (1993). Semiparametric least squares (SLS) and weighted SLS estimation of single-index models. *Journal of Econometrics* **58** 71–120. MR1230981

Johnson, B. A., Lin, D. Y. and Zeng, D. (2008). Penalized estimating functions and variable selection in semiparametric regression models. *Journal of the American Statistical Association* **103** 672–680. MR2435469

Kai, B., Li, R. and Zou, H. (2011). New efficient estimation and variable selection methods for semiparametric varying-coefficient partially linear models. *The Annals of Statistics* **39** 305–332. MR2797848

Lee, S. (2003). Efficient semiparametric estimation of a partially linear quantile regression model. *Econometric Theory* **19** 1–31. MR1965840

Li, R. and Liang, H. (2008). Variable selection in semiparametric regression modeling. *The Annals of Statistics* **36** 261–286. MR2387971

Liang, H. and Li, R. (2009). Variable selection for partially linear models with measurement errors. *Journal of the American Statistical Association* **104** 234–248. MR2504375

Liang, H., Wang, H. and Tsai, C.-L. (2012). Profiled forward regression for ultrahigh dimensional variable screening in semiparametric partially linear models. *Statistica Sinica* **22** 531–554. MR2954351

Liang, H., Wang, S., Robins, J. M. and Carroll, R. J. (2004). Estimation in partially linear models with missing covariates. *Journal of the American Statistical Association* **99** 357–367. MR2062822

Liang, H., Liu, X., Li, R. and Tsai, C.-L. (2010). Estimation and testing for partially linear single-index models. *The Annals of Statistics* **38** 3811–3836. MR2766869

Meinshausen, N. and Bühlmann, P. (2006). High-dimensional graphs and variable selection with the Lasso. *The Annals of Statistics* **34** 1436–1462. MR2278363

Ortega, J. M. and Rheinboldt, W. C. (1970). *Iterative Solution of Nonlinear Equations in Several Variables*. Academic Press, San Diego. MR0273810

Pagan, A. and Ullah, A. (1999). *Nonparametric Econometrics*. Cambridge University Press, New York. MR1699703

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics* **6** 461–464. MR0468014

Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society, Series B (Methodological)* **58** 267–288. MR1379242

Tsay, R. S. (2005). *Analysis of Financial Time Series, 2nd Edition*. Wiley-Interscience, New York. MR2162112

van der Vaart, A. W. and Wellner, J. A. (1996). *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer-Verlag, New York. MR1385671

Wang, H. (2009). Forward regression for ultra-high dimensional variable screening. *Journal of the American Statistical Association* **104** 1512–1524. MR2750576

Wang, H., Li, R. and Tsai, C.-L. (2007). Tuning parameter selectors for the smoothly clipped absolute deviation method. *Biometrika* **94** 553–568. MR2410008

Wang, H. J. and Wang, L. (2009). Locally weighted censored quantile regression. *Journal of the American Statistical Association* **104** 1117–1128. MR2562007

Wang, H. and Xia, Y. (2009). Shrinkage estimation of the varying coefficient model. *Journal of the American Statistical Association* **104** 747–757. MR2541592

Wang, C.-Y., Wang, S., Gutierrez, R. G. and Carroll, R. J. (1998). Local linear regression for generalized linear models with missing data. *The Annals of Statistics* **26** 1028–1050. MR1635438

Wang, L., Liu, X., Liang, H. and Carroll, R. J. (2011). Estimation and variable selection for generalized additive partial linear models. *The Annals of Statistics* **39** 1827–1851. MR2893854

Xie, H. and Huang, J. (2009). SCAD-penalized regression in high-dimensional partially linear models. *The Annals of Statistics* **37** 673–696. MR2502647

Zhang, C.-H. (2010). Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics* **38** 894–942. MR2604701

Zhao, P. and Yu, B. (2006). On model selection consistency of Lasso. *Journal of Machine Learning Research* **7** 2541–2563. MR2274449