Accepted Manuscript

Estimation and variable selection for quantile partially linear single-index models

Yuankun Zhang, Heng Lian, Yan Yu

PII: S0047-259X(17)30562-6
DOI: https://doi.org/10.1016/j.jmva.2017.09.006
Reference: YJMVA 4289

To appear in: *Journal of Multivariate Analysis*

Received date : 20 July 2016

Please cite this article as: Y. Zhang, H. Lian, Y. Yu, Estimation and variable selection for quantile partially linear single-index models, *Journal of Multivariate Analysis* (2017), https://doi.org/10.1016/j.jmva.2017.09.006

Estimation and variable selection for quantile partially linear single-index models

Yuankun Zhang[a], Heng Lian[b], Yan Yu[c,]

[a]Department of Mathematical Sciences, University of Cincinnati, Ohio, USA
[b]Department of Mathematics, City University of Hong Kong, Hong Kong
[c]Carl H. Lindner College of Business, University of Cincinnati, Ohio, USA

Abstract

Partially linear single-index models are flexible dimension reduction semiparametric tools yet still retain ease of interpretability as linear models. This paper is concerned with the estimation and variable selection for partially linear single-index quantile regression models. Polynomial splines are used to estimate the unknown link function. We first establish the asymptotic properties of the quantile regression estimators. For feature selection, we adopt the smoothly clipped absolute deviation penalty (SCAD) approach to select simultaneously single-index variables and partially linear variables. We show that the regularized variable selection estimators are consistent and possess oracle properties. The consistency and oracle properties are also established under the proposed linear approximation of the nonparametric link function that facilitates fast computation. Furthermore, we show that the proposed SCAD tuning parameter selectors via the Schwarz information criterion can consistently identify the true model. Monte Carlo studies and an application to Boston Housing price data are presented to illustrate the proposed approach.

Keywords: Asymptotics, check loss minimization, oracle properties, polynomial splines, quantile regression, SCAD.

1. Introduction

This paper is partly motivated by the famous Boston Housing Price data used in Harrison and Rubinfeld [10]. One important issue in that study was the identification of the key variables that affect housing prices in Boston. The response variable of interest, median housing price, is skewed and truncated due to possible outliers. Quantile regression introduced by Koenker and Bassett [17] is a robust alternative to least squares to analyze these data. Nonparametric quantile regression, in contrast, can unveil important nonlinear features; see, e.g., [18, 34, 35]. Cai and Xu [1] also discovered that some interaction effect exists besides nonlinearity and suggest that a further investigation of a semiparametric model and significant variables to include is warranted.

In this paper, we focus on estimation and variable selection for quantile partially linear single-index models (Q-PLSIM), where the conditional quantiles take the form $g(\mathbf{X}_i^\top \boldsymbol{\gamma}) + \mathbf{Z}_i^\top \boldsymbol{\beta}$. By projecting the q-dimensional covariates $\mathbf{X}$ to a univariate single index $\mathbf{X}_i^\top \boldsymbol{\gamma}$ for the nonparametric part and allowing partially linear terms $\mathbf{Z}_i^\top \boldsymbol{\beta}$, the partially linear single-index models can overcome the curse of dimensionality yet preserve model flexibility.

Partially linear single-index models have been widely studied in the mean regression context. For example, Carroll et al. [2] worked on generalized PLSIM using local methods. Yu and Ruppert [36] proposed penalized spline estimation for PLSIM, and Yu et al. [37] extended it to generalized exponential families with a computationally expedient approach. Liang et al. [22] proposed a profile least squares procedure with kernel estimation for PLSIM and studied estimation, variable selection, and testing. Ma et al. [25] investigated PLSIM for repeated measurements. Unlike additive models, interactions among the single-index variables can be modeled due to the nonparametric link function on the index [2, 36].

Partially linear single-index quantile regression models also include various models as special cases. For example, with no partially linear terms, the Q-PLSIM reduces to the single-index quantile models. Wu et al. [34] first proposed single-index quantile regression models. Kong and Xia [18] investigated the Bahadur representation. Most recently, Ma and He [24] studied inference with profile optimization. When the single-index parameter is known or $\mathbf{X}$ is univariate, the estimation reduces to a partially linear quantile problem. For example, He and Liang [11] investigated errors-in-variables partially linear quantile models. Chen and Khan [4] examined censored partially linear quantile models.

We estimate the unknown link function g by polynomial splines nonparametrically. As noted in Wang et al. [30], polynomial splines, as a global smoothing method without involving large systems of equations, are advantageous in computational expedience compared with local methods. Penalized splines may also offer great computational advantages. However, their theory is still the object of ongoing research and a fixed knot assumption often has to be used [32, 36]. In computation, we adopt linear approximation to the nonparametric link function g to further facilitate fast computation. Specifically, we propose an efficient iterative algorithm for estimation, where in each iteration a linear quantile regression algorithm can be readily applied. This can avoid solving a high-dimensional nonlinear optimization, unlike in a direct one-step estimation.

Variable selection is often of vital interest in applications, where many predictors are often collected but insignificant variables should be left out for a final model. In this paper, we select significant single-index variables in $\mathbf{X}$ and partially linear variables in $\mathbf{Z}$ simultaneously using a regularization method with the popular smoothly clipped absolute deviation (SCAD) penalty proposed by Fan and Li [8]. In this paper, the variable dimension is considered fixed as in the seminal papers of Tibshirani [28] (1996, Lasso) and Fan and Li [8] (2001, SCAD). There is a vast literature on penalized variable selection with fixed-dimensional predictors following these seminal papers, such as Zou and Hastie [41] (2005, Elastic Net), Zou [40] (2006, Adaptive Lasso), Yuan and Lin [38] (2006, grouped lasso), etc. Please refer to these for further reading on penalized/shrinkage variable selection.

Schwarz's information criterion is used to choose SCAD tuning parameters. We adopt the difference convex algorithm from Wu and Liu [33] to transform the non-convex SCAD penalty into two convex functions. We discuss detailed algorithms in Section 4. We then conduct various simulation studies for the proposed Q-PLSIM estimation and variable selection and examine the Boston housing price data in Section 5. Our findings shed some new light on the important variables for Boston housing prices.

Finally, we make the following contributions to the theoretical literature: (i) we establish the existence, convergence rate, and asymptotic normality for the estimation of Q-PLSIM parameters; (ii) for variable selection using the penalized check-loss function, we establish the existence, convergence rate, and asymptotic normality of the penalized estimators; more importantly, we prove their oracle properties; (iii) with the linear approximation of the nonparametric link function g, we again prove their oracle properties for variable selection; (iv) we formally establish the variable selection consistency when the tuning parameters for regularization are chosen by the Schwarz information criterion.

We note that even though partially linear single-index quantile regression models are appealing in practice, theoretically dealing with single-index models is more challenging than some other semiparametric models, including partially linear varying coefficient models (see Fan and Zhang [9] for a survey on mean regression; Kim [16], Wang et al. [29] on quantile regression) and partially linear additive models [7, 14, 35]. Since the single-index parameters $\boldsymbol{\gamma}$ are nested within the unknown link function g, this makes higher order terms in the quadratic approximation of the check-loss functions more complicated. Techniques used in the proof of oracle properties for quantile semiparametric regression are also different from least squares models due to the fact that the check-loss function is not differentiable. Instead, we rely on the convexity of the check-loss function and use the subgradient of the check-loss function in our proofs. This gives a linear lower bound for differences in the loss function, instead of the usual quadratic approximation, which is nevertheless sufficient for the proof of the oracle properties. In addition, even though spline smoothing is computationally expedient, theorems for spline-based smoothing are generally more challenging to establish than those for local methods. We hope that some of our proof techniques can be useful for establishing asymptotic properties for other model estimation and variable selection problems.

The rest of the paper is organized as follows. In Section 2, we introduce the partially linear single-index quantile regression model. We propose polynomial spline estimators minimizing a quantile check-loss function. We establish the rate of convergence and asymptotic properties of the proposed estimators. In Section 3, we describe the variable selection procedures for the single-index and partially linear terms via a penalized check-loss function. We establish their asymptotic properties and prove that they possess oracle properties along with theoretical results under Schwarz's Information Criterion. In Section 4, we elaborate the iterative algorithm for estimation and penalized variable selection in detail. Numerical studies and an application to the Boston Housing data are presented in Section 5. Lemmas and proofs of the main theorems are relegated to the Appendix. Detailed proofs of all lemmas are given in the Online Supplement.

2. Quantile partially linear single-index models

2.1. Estimation

Suppose a random sample $(Y_1, \mathbf{X}_1, \mathbf{Z}_1), \ldots, (Y_n, \mathbf{X}_n, \mathbf{Z}_n)$ is observed. We consider the quantile partially linear single-index model (Q-PLSIM) defined, for each $i \in \{1, \ldots, n\}$, by

$$Y_i = g(\mathbf{X}_i^\top \boldsymbol{\gamma}_0) + \mathbf{Z}_i^\top \boldsymbol{\beta}_0 + e_i,$$

where g is an unknown link function and $(\boldsymbol{\gamma}_0^\top, \boldsymbol{\beta}_0^\top)^\top$ is the unknown parameter vector, such that $\Pr(e_i \le 0 \mid \mathbf{X}_i, \mathbf{Z}_i) = \tau$, and $\mathbf{X}_i$ and $\mathbf{Z}_i$ are q-dimensional and p-dimensional covariates, respectively. For identifiability, according to Yu and Ruppert [36] and Lin and Kulasekera [23], we impose $\|\boldsymbol{\gamma}\| = 1$ with its first nonzero element positive.

Note that when the data are actually generated from a linear model in which all the coefficients associated with the predictors $\mathbf{X}$ are trivial, the function g would then be constant in a partially linear single-index model but then this model, and in particular the single-index coefficients $\boldsymbol{\gamma}$, would no longer be identifiable. In such a case, the partially linear single-index model reduces to a linear model.

In the even more extreme case where the data are actually generated from a constant where all of the coefficients associated with all the predictors are trivial, the function g would be constant and all the partially linear coefficients $\boldsymbol{\beta}$ would be zero in a partially linear single-index model. In such a case, the partially linear single-index model reduces to a constant. In other words, if no variable is actually needed in the model, the quantiles would be trivially constant. The situation would be very similar to the usual linear models when all coefficients reduce to zero. The corresponding estimates of g would be expected close to a constant and estimates of the partially linear coefficients $\boldsymbol{\beta}$ close to zero in a large sample. Again the single-index coefficients $\boldsymbol{\gamma}$ would be unidentifiable. In practice we usually first fit a linear model to the data and would further resort to PLSIM if some standard diagnostic procedure suggests nonlinearity is present.

Here we model the τth conditional quantile of Y by $Q_\tau(Y \mid \boldsymbol{x}, \boldsymbol{z}) = m(\boldsymbol{x}, \boldsymbol{z}) = g(\boldsymbol{x}^\top \boldsymbol{\gamma}) + \boldsymbol{z}^\top \boldsymbol{\beta}$, where $(\boldsymbol{x}, \boldsymbol{z}) \in S$. Note that we can rewrite our model as $Q_\tau(Y \mid \boldsymbol{v}) = m(\boldsymbol{v}) = g(\boldsymbol{v}^\top \boldsymbol{\phi}) + \boldsymbol{v}^\top \boldsymbol{\psi}$, where $\boldsymbol{v} = (\boldsymbol{x}, \boldsymbol{z}) \in S \subset \mathbb{R}^{q+p}$, $\boldsymbol{\phi} = (\boldsymbol{\gamma}, \mathbf{0}_p)$ and $\boldsymbol{\psi} = (\mathbf{0}_q, \boldsymbol{\beta})$. Since we use different variables for the single-index part and the partially linear part, the covariates $\boldsymbol{v}$ are naturally divided into distinct groups $\boldsymbol{x}$ and $\boldsymbol{z}$. Consequently, parameter vectors $\boldsymbol{\phi}$ and $\boldsymbol{\psi}$ are perpendicular to each other. Therefore, following the argument in the proof of Case 4 of Theorem 2 in Lin and Kulasekera (2007), we can prove that for all $(\boldsymbol{x}, \boldsymbol{z}) \in S$, $m(\boldsymbol{x}, \boldsymbol{z}) = g_1(\boldsymbol{x}^\top \boldsymbol{\gamma}_1) + \boldsymbol{z}^\top \boldsymbol{\beta}_1 = g_2(\boldsymbol{x}^\top \boldsymbol{\gamma}_2) + \boldsymbol{z}^\top \boldsymbol{\beta}_2$ for some continuous nonlinear functions g_1, g_2, and for $k \in \{1, 2\}$, $\boldsymbol{\beta}_k \in \mathbb{R}^p$ and $\boldsymbol{\gamma}_k \in D = \{\boldsymbol{\gamma} \in \mathbb{R}^q : \|\boldsymbol{\gamma}\| = 1$ with first nonzero element positive$\}$. Then $\boldsymbol{\gamma}_1 = \boldsymbol{\gamma}_2$, $\boldsymbol{\beta}_1 = \boldsymbol{\beta}_2$ and $g_1 = g_2$. Hence, the conditional quantile model $Q_\tau(Y \mid \boldsymbol{x}, \boldsymbol{z}) = g(\boldsymbol{x}^\top \boldsymbol{\gamma}) + \boldsymbol{z}^\top \boldsymbol{\beta}$ is identifiable under these assumptions. See Lin and Kulasekera [23] for more details.

To take into account the unit norm constraint, we use the popular "delete-one-component" method [5, 36]. We can write $\boldsymbol{\gamma} = ((1 - \|\boldsymbol{\gamma}^{(-1)}\|^2)^{1/2}, \gamma_2, \ldots, \gamma_q)^\top$ where $\boldsymbol{\gamma}^{(-1)} = (\gamma_2, \ldots, \gamma_q)^\top$ is $\boldsymbol{\gamma}$ without the first component. Thus $\boldsymbol{\gamma}$ is a function of $\boldsymbol{\gamma}^{(-1)}$. The $q \times (q-1)$ Jacobian matrix is

$$\widetilde{\mathbf{J}} = \frac{\partial \boldsymbol{\gamma}}{\partial \boldsymbol{\gamma}^{(-1)}} = \begin{pmatrix} -\frac{\boldsymbol{\gamma}^{(-1)}}{(1-\|\boldsymbol{\gamma}^{(-1)}\|^2)^{1/2}} \\ \mathbf{I}_{(q-1)\times(q-1)} \end{pmatrix},$$

where $\mathbf{I}_{(q-1)\times(q-1)}$ is the $(q-1) \times (q-1)$ identity matrix. In what follows, we use $\mathbf{J} = \mathrm{diag}(\widetilde{\mathbf{J}}, \mathbf{I}_{p\times p})$. We also use polynomial splines to approximate the nonparametric function g.

Let $t_0 = a < t_1 < \cdots < t_{K'} < b = t_{K'+1}$ be a partition of $[a, b]$ into subintervals $[t_k, t_{k+1})$ with $k \in \{0, \ldots, K'\}$ with K' internal knots. We only restrict our attention to equally spaced knots although data-driven choices can be considered such as putting knots at certain sample quantiles of the single-index values u. A polynomial spline of order s is a function whose restriction to each subinterval is a polynomial of degree $s-1$ and globally $s-2$ times continuously differentiable on $[a, b]$. The collection of splines with a fixed sequence of knots has a B-spline basis $B_1(u), \ldots, B_K(u)$ with $K = K' + s$. In the empirical implementations, we use the minimal and maximal values of $u_i = \mathbf{X}_i^\top \boldsymbol{\gamma}$ as a and b to generate B-spline basis functions for a given $\boldsymbol{\gamma}$.

We assume that the B-spline basis is normalized to have $B_1(u) + \cdots + B_K(u) = \sqrt{K}$. Such a normalization is not essential and is just imposed to simplify some expressions in subsequent theoretical derivations. Let $\mathbf{B} = (B_1, \ldots, B_K)^\top$. With the given single index $u_i = \mathbf{X}_i^\top \boldsymbol{\gamma}$, g can be estimated by B-spline expansion, viz. $g(u_i) \approx \mathbf{B}^\top(u_i)\boldsymbol{\delta}$, where $\boldsymbol{\delta}$ is the spline coefficient vector of dimension K. Throughout this paper, we use cubic B-spline basis functions, i.e., the order s of B-splines is 4. Using the spline smoothing, we will minimize

$$L_n(\boldsymbol{\gamma}, \boldsymbol{\beta}, \boldsymbol{\delta}) = \sum_{i=1}^{n} \rho_\tau\{Y_i - \mathbf{B}^\top(\mathbf{X}_i^\top \boldsymbol{\gamma})\boldsymbol{\delta} - \mathbf{Z}_i^\top \boldsymbol{\beta}\} \tag{1}$$

over $(\boldsymbol{\gamma}, \boldsymbol{\beta}, \boldsymbol{\delta})$ with the constraint $\|\boldsymbol{\gamma}\| = 1$ and $\gamma_1 > 0$, or equivalently regarding $\boldsymbol{\gamma} = \boldsymbol{\gamma}(\boldsymbol{\gamma}^{(-1)})$ as a function of $\boldsymbol{\gamma}^{(-1)}$ and optimize over $(\boldsymbol{\gamma}^{(-1)}, \boldsymbol{\beta}, \boldsymbol{\delta})$. $\rho_\tau(s) = \tau s - s\mathbf{1}(s < 0)$ is the check loss function for quantile regression. The resulting estimators are denoted by $(\widehat{\boldsymbol{\gamma}}, \widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\delta}})$.

2.2. Large-sample properties

To establish asymptotic normality and the convergence rate, we need to "orthogonalize" the parametric part with respect to the nonparametric part using the following projection. Let $\mathcal{M} = \{m : m(\mathbf{x}) = f(\mathbf{x}^\top \boldsymbol{\gamma}), \mathrm{E}\{m^2(\mathbf{X})\} < \infty\}$ be the space of single-index functions. In this paper, the projection of any random variable W onto $\mathcal{M}$, denoted by $\mathrm{E}_{\mathcal{M}}(W)$, is defined as $m(\mathbf{X})$, m being the minimizer of

$$\mathrm{E}[f(0 \mid \mathbf{X}, \mathbf{Z})\{W - m(\mathbf{X})\}^2],$$

with $m \in \mathcal{M}$. This definition can be extended trivially to the case where $\mathbf{W} = (W_1, \ldots, W_q)^\top$ is a random vector by $\mathrm{E}_{\mathcal{M}}(\mathbf{W}) = (\mathrm{E}_{\mathcal{M}}(W_1), \ldots, \mathrm{E}_{\mathcal{M}}(W_q))^\top$. We impose the following assumptions.

(A1) The covariates $\mathbf{Z}$, $\mathbf{X}$ are bounded.

(A2) Let $f(\cdot \mid \mathbf{X}_i, \mathbf{Z}_i)$ be the conditional density of e_i. We assume that $f(\cdot \mid \mathbf{X}_i, \mathbf{Z}_i)$ is bounded and bounded away from zero in a neighborhood of 0, uniformly over the support of $\mathbf{X}_i, \mathbf{Z}_i$. The derivative of $f(\cdot \mid \mathbf{X}_i, \mathbf{Z}_i)$ is uniformly bounded in a neighborhood of 0 over the support of $\mathbf{X}_i, \mathbf{Z}_i$.

(A3) The function g is in the Hölder space of order $d \geq 2$, i.e., $|g^{(m)}(x) - g^{(m)}(y)| \leq C|x - y|^r$ for $d = m + r$ and m is the largest integer strictly smaller than d, where $g^{(m)}$ is the mth derivative of g.

(A4) Suppose $\mathrm{E}_{\mathcal{M}}\{X_j g^{(1)}(\mathbf{X}^\top \boldsymbol{\gamma}_0)\} = f_j(\mathbf{X}^\top \boldsymbol{\gamma}_0)$ for all $j \in \{1, \ldots, q\}$. The functions f_j are in the Hölder space of order $d' \geq 1$. The order of the B-spline used satisfies $s \geq \max(d, d') + 1$. The same smoothness condition is satisfied by the component functions of $\mathrm{E}_{\mathcal{M}}(\mathbf{Z})$.

(A5) The matrix

$$\mathrm{E}\left\{ f(0 \mid \mathbf{X}, \mathbf{Z}) \begin{pmatrix} \widetilde{\mathbf{J}}^\top \mathbf{X} g^{(1)}(\mathbf{X}^\top \boldsymbol{\gamma}_0) - \mathrm{E}_{\mathcal{M}}\{\widetilde{\mathbf{J}}^\top \mathbf{X} g^{(1)}(\mathbf{X}^\top \boldsymbol{\gamma}_0)\} \\ \mathbf{Z} - E_{\mathcal{M}}(\mathbf{Z}) \end{pmatrix}^{\otimes 2} \right\}$$

is positive definite, where for any matrix $\mathbf{A}$, $\mathbf{A}^{\otimes 2} = \mathbf{A}\mathbf{A}^\top$.

Boundedness of $\mathbf{Z}$ is assumed mainly for convenience of proof; it could possibly be replaced by moment conditions with lengthier arguments. Boundedness of $\mathbf{X}$ is tied to our estimation approach, typically assumed when using regression splines. Assumption (A2) on conditional density is commonly used in quantile regression [12, 29]. Smoothness of g is required to establish the convergence rate. Smoothness of functions in the representation of $\mathrm{E}_{\mathcal{M}}\{X_j g^{(1)}(\mathbf{X}^\top \boldsymbol{\gamma}_0)\}$ is typically used in semiparametric models to show the asymptotic normality of the parametric part. Finally (A5) can be regarded as an identifiability assumption for semiparametric models [2, 19, 30, 31].

Theorem 1. *If conditions (A1)–(A5) hold, $K \to \infty$, and $K^{d+3/2} \ln n/n \to 0$, then there exists a local minimizer of* (1) *with*

$$\|\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0\| + \|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\| + \|\widehat{\boldsymbol{\delta}} - \boldsymbol{\delta}_0\| = O_p(\sqrt{K/n} + K^{-d}).$$

In particular, $\|\widehat{\boldsymbol{\delta}} - \boldsymbol{\delta}_0\| = O_p(\sqrt{K/n} + K^{-d})$ implies that $\|\widehat{g} - g\| = O_p(\sqrt{K/n} + K^{-d})$, with $\widehat{g} = \mathbf{B}^\top \widehat{\boldsymbol{\delta}}$.

The convergence rate above takes a familiar form as in nonparametric regression with the two terms corresponding to bias and variance, respectively. The optimal choice of K is obviously $K \sim n^{1/(2d+1)}$. In Theorem 2, stronger assumptions on the choice of K and smoothness of nonparametric functions allow us to establish the asymptotic normality of the single-index parameter estimates $\widehat{\boldsymbol{\gamma}}$ and partially linear parameter estimates $\widehat{\boldsymbol{\beta}}$. Note that when d' is large enough (e.g., $d' = d$), $K \sim n^{1/(2d+1)}$ is still contained in the permissible range.

Theorem 2. *If conditions (A1)–(A5) hold, $K \to \infty$, $K^{\max(4,d+3/2)} \ln n/n \to 0$, $\sqrt{n}\, K^{-2d+3/2} \to 0$, and $\sqrt{n}\, K^{-d-d'} \to 0$, then*

$$\sqrt{n}\left\{ \begin{pmatrix} \widehat{\boldsymbol{\gamma}} \\ \widehat{\boldsymbol{\beta}} \end{pmatrix} - \begin{pmatrix} \boldsymbol{\gamma}_0 \\ \boldsymbol{\beta}_0 \end{pmatrix} \right\} \rightsquigarrow \mathcal{N}[0, \mathbf{J}(\mathbf{J}^\top \boldsymbol{\Phi} \mathbf{J})^{-1} \mathbf{J}^\top \boldsymbol{\Sigma} \mathbf{J} (\mathbf{J}^\top \boldsymbol{\Phi} \mathbf{J})^{-1}) \mathbf{J}^\top],$$

where

$$\boldsymbol{\Phi} = \mathrm{E}\left\{ f(0 \mid \mathbf{X}, \mathbf{Z}) \begin{pmatrix} g^{(1)}(\mathbf{X}_i^\top \boldsymbol{\gamma}_0)\mathbf{X} - \mathrm{E}_{\mathcal{M}}\{g^{(1)}(\mathbf{X}_i^\top \boldsymbol{\gamma}_0)\mathbf{X}\} \\ \mathbf{Z}_i - E_{\mathcal{M}}(\mathbf{Z}_i) \end{pmatrix}^{\otimes 2} \right\},$$

$$\boldsymbol{\Sigma} = \tau(1-\tau)\mathrm{E}\left\{ \begin{pmatrix} g^{(1)}(\mathbf{X}_i^\top \boldsymbol{\gamma}_0)\mathbf{X} - \mathrm{E}_{\mathcal{M}}\{g^{(1)}(\mathbf{X}_i^\top \boldsymbol{\gamma}_0)\mathbf{X}\} \\ \mathbf{Z}_i - E_{\mathcal{M}}(\mathbf{Z}_i) \end{pmatrix}^{\otimes 2} \right\},$$

and the Jacobian matrix $\mathbf{J}$ is evaluated at the true single-index parameter $\boldsymbol{\gamma}_0$.

3. Penalized variable selection

When the number of predictors is relatively large, it is desirable to select the relevant predictors both in the parametric and nonparametric part. We use a penalization method for variable selection. More specifically, given the estimated spline coefficients $\widehat{\boldsymbol{\delta}}$, we find the penalized estimators $\widetilde{\boldsymbol{\gamma}}, \widetilde{\boldsymbol{\beta}}$ through minimization of

$$\begin{aligned} Q_n(\boldsymbol{\gamma},\boldsymbol{\beta}) &= L_n(\boldsymbol{\gamma},\boldsymbol{\beta}) + n\sum_{j=1}^{q} p_{\lambda_1}(|\gamma_j|) + n\sum_{j=1}^{p} p_{\lambda_2}(|\beta_j|) \\ &= \sum_{i=1}^{n} \rho_\tau\{Y_i - \mathbf{B}^\top(\mathbf{X}_i^\top\boldsymbol{\gamma})\widehat{\boldsymbol{\delta}} - \mathbf{Z}_i^\top\boldsymbol{\beta}) + n\sum_{j=1}^{q} p_{\lambda_1}(|\gamma_j|) + n\sum_{j=1}^{p} p_{\lambda_2}(|\beta_j|), \end{aligned} \tag{2}$$

where λ_1 and λ_2 are two regularization tuning parameters. There are more than one way to specify the penalty function and here we only focus on the SCAD penalty function [8], defined by its first derivative, viz.

$$p'_\lambda(v) = \lambda\left\{\mathbf{1}(v \le \lambda) + \frac{(a\lambda - v)_+}{(a-1)\lambda}\mathbf{1}(v > \lambda)\right\},$$

with $a > 2$ and $p_\lambda(0) = 0$. We will use $a = 3.7$ as suggested in [8]. Other choices of penalty, such as an adaptive lasso [40] or minimax concave penalty [39], are expected to produce similar results in both theory and practice.

For theoretical purposes, we assume that only the first q_1 components of $\boldsymbol{\gamma}$ and the first p_1 components of $\boldsymbol{\beta}$ are nonzero. Let $\boldsymbol{\gamma}^{(1)} = (\gamma_1, \ldots, \gamma_{q_1})^\top$ and $\boldsymbol{\beta}^{(1)} = (\beta_1, \ldots, \beta_{p_1})^\top$. The following theorem presents the oracle properties [8] of the penalized estimator. That is, the asymptotic normality property is the same as when the nonzero components in $\boldsymbol{\gamma}$ and $\boldsymbol{\beta}$ are known.

Theorem 3. *Under the conditions of Theorem 2, and* $\lambda_1, \lambda_2 = o(1)$, $\sqrt{K}(\sqrt{K/n} + K^{-d})/\lambda_1 = o(1)$, $\sqrt{K}(\sqrt{K/n} + K^{-d})/\lambda_2 = o(1)$, *there exists a local minimizer* $(\widetilde{\boldsymbol{\gamma}}, \widetilde{\boldsymbol{\beta}})$ *of* (2) *such that*

$$\sqrt{n}\left\{\begin{pmatrix} \widetilde{\boldsymbol{\gamma}}^{(1)} \\ \widetilde{\boldsymbol{\beta}}^{(1)} \end{pmatrix} - \begin{pmatrix} \boldsymbol{\gamma}_0^{(1)} \\ \boldsymbol{\beta}_0^{(1)} \end{pmatrix}\right\} \rightsquigarrow \mathcal{N}[0, \mathbf{J}^{(1)}(\mathbf{J}^{(1)\top}\boldsymbol{\Phi}^{(1)}\mathbf{J}^{(1)})^{-1}\mathbf{J}^{(1)\top}\boldsymbol{\Sigma}^{(1)}\mathbf{J}^{(1)}(\mathbf{J}^{(1)\top}\boldsymbol{\Phi}^{(1)}\mathbf{J}^{(1)})^{-1}\mathbf{J}^{(1)\top}],$$

where $\mathbf{J}^{(1)}$, $\boldsymbol{\Phi}^{(1)}$ *and* $\boldsymbol{\Sigma}^{(1)}$ *are defined in a similar way as in Theorem 2, using only the first* q_1 *components of* $\boldsymbol{\gamma}_0$, *as well as the first* q_1 *components of* $\mathbf{X}$ *and the first* p_1 *components of* $\mathbf{Z}$. *In addition,* $\widetilde{\gamma}_{q_1+1} = \cdots = \widetilde{\gamma}_q = \widetilde{\beta}_{p_1+1} = \cdots = \widetilde{\beta}_p = 0$ *with probability approaching* 1.

For fast computation, we employ a linear approximation of g in the loss function, so that $g(\mathbf{X}_i^\top\boldsymbol{\gamma})$ is replaced by its linear approximation, viz.

$$g(\mathbf{X}_i^\top\boldsymbol{\gamma}) \cong g(\mathbf{X}_i^\top\boldsymbol{\gamma}_0) + g'(\mathbf{X}_i^\top\boldsymbol{\gamma}_0)\mathbf{X}_i^\top(\boldsymbol{\gamma} - \boldsymbol{\gamma}_0).$$

By the spline smoothing, $g(\mathbf{X}_i^\top\boldsymbol{\gamma}_0)$ can be represented by $\mathbf{B}^\top(\mathbf{X}_i^\top\boldsymbol{\gamma}_0)\boldsymbol{\delta}$ and $g'(\mathbf{X}_i^\top\boldsymbol{\gamma}_0)$ is analogously represented by $\mathbf{B}^{(1)\top}(\mathbf{X}_i^\top\boldsymbol{\gamma}_0)\boldsymbol{\delta}$, in which $\mathbf{B}^{(1)}$ is the first derivative of B-spline basis functions. Given the initial unpenalized estimates $(\widehat{\boldsymbol{\gamma}}, \widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\delta}})$, we can solve the following optimization problem:

$$\begin{aligned} Q_n^*(\boldsymbol{\gamma},\boldsymbol{\beta}) &= L_n^*(\boldsymbol{\gamma},\boldsymbol{\beta}) + n\sum_{j=1}^{q} p_{\lambda_1}(|\gamma_j|) + n\sum_{j=1}^{p} p_{\lambda_2}(|\beta_j|) \\ &= \sum_{i=1}^{n} \rho_\tau\{Y_i - \mathbf{B}^\top(\mathbf{X}_i^\top\widehat{\boldsymbol{\gamma}})\widehat{\boldsymbol{\delta}} - \mathbf{B}^{(1)\top}(\mathbf{X}_i^\top\widehat{\boldsymbol{\gamma}})\widehat{\boldsymbol{\delta}}\mathbf{X}_i^\top(\boldsymbol{\gamma} - \widehat{\boldsymbol{\gamma}}) - \mathbf{Z}_i^\top\boldsymbol{\beta}\} + n\sum_{j=1}^{q} p_{\lambda_1}(|\gamma_j|) + n\sum_{j=1}^{p} p_{\lambda_2}(|\beta_j|). \end{aligned} \tag{3}$$

Next we show that the minimizer of (3), still denoted by $(\widetilde{\boldsymbol{\gamma}}, \widetilde{\boldsymbol{\beta}})$, again satisfies the oracle properties.

Theorem 4. *Under the same conditions assumed for Theorem 3, the minimizer of* (3)*, still denoted by* $(\widetilde{\boldsymbol{\gamma}}, \widetilde{\boldsymbol{\beta}})$*, satisfies the same oracle properties as those stated in Theorem 3.*

Finally, we can choose the penalty parameters λ_1, λ_2 using the Schwarz Information Criterion (SIC), viz.

$$\mathrm{SIC}(\lambda_1, \lambda_2) = \ln\{L_n^*(\widetilde{\boldsymbol{\gamma}}, \widetilde{\boldsymbol{\beta}})\} + \frac{1}{2n}(\ln n)(\widetilde{p} + \widetilde{q})$$

$$= \ln\left[\sum_{i=1}^{n}\rho_\tau\{Y_i - \mathbf{B}^\top(\mathbf{X}_i^\top\widetilde{\boldsymbol{\gamma}})\widehat{\boldsymbol{\delta}} - \mathbf{B}^{(1)\top}(\mathbf{X}_i^\top\widetilde{\boldsymbol{\gamma}})\widehat{\boldsymbol{\delta}}\mathbf{X}_i^\top(\widetilde{\boldsymbol{\gamma}} - \widehat{\boldsymbol{\gamma}}) - \mathbf{Z}_i^\top\widetilde{\boldsymbol{\beta}}\}\right] + \frac{1}{2n}(\ln n)(\widetilde{p} + \widetilde{q}), \quad (4)$$

where $(\widetilde{\boldsymbol{\gamma}}, \widetilde{\boldsymbol{\beta}})$ are the estimates based on a given pair of values of (λ_1, λ_2) and $\widetilde{q}$ and $\widetilde{p}$ are the number of nonzero coefficients in $\widetilde{\boldsymbol{\gamma}}$ and $\widetilde{\boldsymbol{\beta}}$, respectively [14, 20]. The Schwarz information criterion is consistent in this context.

Theorem 5. *The SIC defined in* (4) *is consistent in the sense that using the* (λ_1, λ_2) *selected by SIC, with probability approaching* 1*, we have* $\widetilde{\gamma}_{q_1+1} = \cdots = \widetilde{\gamma}_q = \widetilde{\beta}_{p_1+1} = \cdots = \widetilde{\beta}_p = 0$ *while other components are not estimated as zero.*

In the above, we use the loss function from (3) with linear approximation of g. We can also similarly define SIC based on the loss function $L_n(\widetilde{\boldsymbol{\gamma}}, \widetilde{\boldsymbol{\beta}})$ in (2)

$$\mathrm{SIC}(\lambda_1, \lambda_2) = \ln\left[\sum_{i=1}^{n}\rho_\tau\{Y_i - \mathbf{B}^\top(\mathbf{X}_i^\top\widetilde{\boldsymbol{\gamma}})\widehat{\boldsymbol{\delta}} - \mathbf{Z}_i^\top\widetilde{\boldsymbol{\beta}}\}\right] + \frac{1}{2n}(\ln n)(\widetilde{p} + \widetilde{q}).$$

We show in the Appendix that Theorem 5 also holds.

4. Algorithms and computation

There are a number of approaches to minimizing the objective function (1). One could use a direct "one-step" estimation by minimizing $L_n(\boldsymbol{\gamma}, \boldsymbol{\beta}, \boldsymbol{\delta})$ over all parameters $(\boldsymbol{\gamma}, \boldsymbol{\beta}, \boldsymbol{\delta})$. This might be challenging due to the nonlinear optimization over a possibly high-dimensional space. A natural alternative is through an iterative approach, where the estimation for the spline coefficients $\boldsymbol{\delta}$ versus the estimation for the single-index parameters $\boldsymbol{\gamma}$ and partially linear $\boldsymbol{\beta}$ are iteratively updated until convergence. This is especially appealing because under the linear approximation of g, each iteration is essentially a linear quantile problem. Furthermore, given the spline coefficient estimates $\widehat{\boldsymbol{\delta}}$, the penalized estimators of $\widetilde{\boldsymbol{\gamma}}$ and $\widetilde{\boldsymbol{\beta}}$ can then be obtained via a penalized linear quantile regression using some existing algorithms, such as the Difference Convex Algorithm (DCA).

In particular, given the single-index parameters $\boldsymbol{\gamma}$ and partially linear parameters $\boldsymbol{\beta}$, the estimates of the spline coefficients $\boldsymbol{\delta}$ can be obtained by minimizing

$$\sum_{i=1}^{n}\rho_\tau\{Y_i - \mathbf{Z}_i^\top\boldsymbol{\beta} - \mathbf{B}^\top(\mathbf{X}_i^\top\boldsymbol{\gamma})\boldsymbol{\delta}\},$$

which is a univariate quantile smoothing problem. If we view $Y_i - \mathbf{Z}_i^\top\boldsymbol{\beta}$ as a response and the B-spline basis as a design matrix at given knots, this may even be regarded as a linear quantile problem in computation.

Given the estimated spline coefficients $\widehat{\boldsymbol{\delta}}$, under the linear approximation to the nonparametric function g, the loss function $L_n^*(\boldsymbol{\gamma}, \boldsymbol{\beta})$ for $\boldsymbol{\gamma}$ and $\boldsymbol{\beta}$ is

$$\sum_{i=1}^{n}\rho_\tau\{Y_i - \mathbf{B}^\top(\mathbf{X}_i^\top\widehat{\boldsymbol{\gamma}})\widehat{\boldsymbol{\delta}} - \mathbf{B}^{(1)\top}(\mathbf{X}_i^\top\widehat{\boldsymbol{\gamma}})\widehat{\boldsymbol{\delta}}\mathbf{X}_i^\top(\boldsymbol{\gamma} - \widehat{\boldsymbol{\gamma}}) - \mathbf{Z}_i^\top\boldsymbol{\beta}\}.$$

Let $Y_i^* = Y_i - \mathbf{B}^\top(\mathbf{X}_i^\top\widehat{\boldsymbol{\gamma}})\widehat{\boldsymbol{\delta}} + \mathbf{B}^{(1)\top}(\mathbf{X}_i^\top\widehat{\boldsymbol{\gamma}})\widehat{\boldsymbol{\delta}}\mathbf{X}_i^\top\widehat{\boldsymbol{\gamma}}$ and $\mathbf{X}_i^* = (\mathbf{B}^{(1)\top}(\mathbf{X}_i^\top\widehat{\boldsymbol{\gamma}})\widehat{\boldsymbol{\delta}}\mathbf{X}_i, \mathbf{Z}_i)$. Thus, the unpenalized estimator of $\boldsymbol{\xi} = (\boldsymbol{\gamma}^\top, \boldsymbol{\beta}^\top)^\top$ will be

$$\widehat{\boldsymbol{\xi}} = \arg\min\sum_{i=1}^{n}\rho_\tau(Y_i^* - \mathbf{X}_i^{*\top}\boldsymbol{\xi}),$$

again a linear quantile problem. Consequently, we propose the following iterative algorithm.

Algorithm. Initialize $\widehat{\boldsymbol{\xi}}^{(0)} = (\widehat{\boldsymbol{\gamma}}^{(0)}, \widehat{\boldsymbol{\beta}}^{(0)})^\top$.
Step 1. Given $\widehat{\boldsymbol{\xi}}^{(m-1)} = (\widehat{\boldsymbol{\gamma}}^{(m-1)}, \widehat{\boldsymbol{\beta}}^{(m-1)})^\top$, the spline coefficients are estimated by

$$\widehat{\boldsymbol{\delta}}^{(m)} = \arg\min\sum_{i=1}^{n}\rho_\tau\{Y_i - \mathbf{Z}_i^\top\widehat{\boldsymbol{\beta}}^{(m-1)} - \mathbf{B}^\top(\mathbf{X}_i^\top\widehat{\boldsymbol{\gamma}}^{(m-1)})\boldsymbol{\delta}\}.$$

Step 2. Given the estimated spline coefficients $\widehat{\boldsymbol{\delta}}^{(m)}$, compute

$$Y_i^* = Y_i - \mathbf{B}^\top(\mathbf{X}_i^\top\widehat{\boldsymbol{\gamma}}^{(m-1)})\widehat{\boldsymbol{\delta}}^{(m)} + \mathbf{B}^{(1)\top}(\mathbf{X}_i^\top\widehat{\boldsymbol{\gamma}}^{(m-1)})\widehat{\boldsymbol{\delta}}^{(m)}\mathbf{X}_i^\top\widehat{\boldsymbol{\gamma}}^{(m-1)} \quad \text{and} \quad \mathbf{X}_i^* = (\mathbf{B}^{(1)\top}(\mathbf{X}_i^\top\widehat{\boldsymbol{\gamma}}^{(m-1)})\widehat{\boldsymbol{\delta}}^{(m)}\mathbf{X}_i, \mathbf{Z}_i),$$

and the estimated $\widehat{\boldsymbol{\xi}}^{(m)}$ can be obtained by minimizing $L_n^*(\boldsymbol{\xi}) = \sum_{i=1}^n \rho_\tau(Y_i^* - \mathbf{X}_i^{*\top}\boldsymbol{\xi})$.

Repeat Steps 1 and 2 until convergence.

For initializing the iterative algorithms, we may simply use mean linear regression or quantile linear regression estimates from the model $Q_\tau(Y) = \mathbf{X}^\top\boldsymbol{\gamma} + \mathbf{Z}^\top\boldsymbol{\beta}$. Normalize $\boldsymbol{\gamma}$ such that $\|\boldsymbol{\gamma}\| = 1$ and its first nonzero element is positive for identifiability. This type of initial values works well in both our simulation studies and in the application to the Boston Housing Price data. Estimates from partially linear single-index mean regression models can also be used. In general, we further recommend trying different random starting values as discussed in Yu and Ruppert [36]. In particular, we recommend using some random starting values and choosing the initial estimate that gives the minimum value of the objective function

$$\sum_{i=1}^n \rho_\tau\{Y_i - \mathbf{Z}_i^\top\widehat{\boldsymbol{\beta}}^{(0)} - \mathbf{B}^\top(\mathbf{X}_i^\top\widehat{\boldsymbol{\gamma}}^{(0)})\boldsymbol{\delta}\}.$$

Note that we may consider extending the profile approach in Ma and He [24] to Q-PLSIM. Here we would treat $\widehat{\boldsymbol{\delta}}$ as a function of $\boldsymbol{\gamma}$ and $\boldsymbol{\beta}$ and then optimize one objective function $L_n(\boldsymbol{\gamma},\boldsymbol{\beta})$ over $\boldsymbol{\xi} = (\boldsymbol{\gamma}^\top,\boldsymbol{\beta}^\top)^\top$. However, it is easy to see that this will involve solving the nonlinear optimization problem over a possibly high-dimensional space $\boldsymbol{\xi} = (\boldsymbol{\gamma}^\top,\boldsymbol{\beta}^\top)^\top$. Another difficulty lies in the absence of an explicit solution for $\widehat{\boldsymbol{\delta}}$ as a function of parameters $\boldsymbol{\gamma}$ and $\boldsymbol{\beta}$, unlike that in the mean regression context as in Liang et al. [22]. In fact, our iterative algorithm combined with the linear approximation of nonparametric function g can be viewed essentially as iterating over two linear quantile estimations and thus can further accelerate the variable selection procedure via penalized estimation.

With the unpenalized estimates $\widehat{\boldsymbol{\gamma}}$, $\widehat{\boldsymbol{\beta}}$, and $\widehat{\boldsymbol{\delta}}$ from the above iterative algorithm, the penalized estimators $\widetilde{\boldsymbol{\gamma}}$, $\widetilde{\boldsymbol{\beta}}$ can be obtained by minimizing the penalized objective function (3). In the empirical studies, we choose the tuning parameters λ_1, λ_2 using the SIC as defined in (4).

One of the most appealing features of the proposed iterative algorithm, along with the linear approximation of g, is that given the spline coefficient estimates $\boldsymbol{\delta}$, estimation and variable selection by minimizing the penalized objective function (3) can reduce to estimating and selecting relevant variables in a linear quantile model. We therefore can directly adopt the Difference Convex Algorithm (DCA) as in Wu and Liu [33] for variable selection in linear quantile models. DCA transforms the non-convex minimization of the penalized objective function $Q_n^*(\boldsymbol{\gamma},\boldsymbol{\beta})$ to a linear programming problem by decomposing the SCAD penalty function into two convex functions. In fact, the SCAD penalty $p_\lambda(v)$ with $v > 0$ can be represented as $p_\lambda(v) = p_{1\lambda}(v) - p_{2\lambda}(v)$, in which $p_{1\lambda}(v)$ and $p_{2\lambda}(v)$ are both convex with derivatives $p'_{1\lambda}(v) = \lambda$ and $p'_{2\lambda}(v) = \lambda\{1 - (a\lambda - v)_+/\lambda(a-1)\}\mathbf{1}(v > \lambda)$, respectively. With the second function $p_{2\lambda}(|\boldsymbol{\xi}|)$ approximated by a linear function, DCA solves the optimization problem

$$\min \sum_{i=1}^n \rho_\tau(Y_i^* - \mathbf{X}_i^{*\top}\boldsymbol{\xi}) + n\sum_{j=1}^{q+p} p_{1\lambda}(|\xi_j|) - n\sum_{j=1}^{q+p} p'_{2\lambda}(|\hat{\xi}_j^{(m)}|)\text{sign}(\hat{\xi}_j^{(m)})(\xi_j - \hat{\xi}_j^{(m)})$$

at each $(m+1)$th step when minimizing $Q_n^*(\boldsymbol{\gamma},\boldsymbol{\beta})$. Further, by introducing some slack variables, the above minimization can be formulated as a linear programming problem, viz.

$$\min \sum_{i=1}^n \{\tau\phi_i + (1-\tau)\psi_i\} + n\lambda\sum_{j=1}^{q+p} t_j - n\sum_{j=1}^{q+p} p'_{2\lambda}(|\hat{\xi}_j^{(m)}|)\text{sign}(\hat{\xi}_j^{(m)})(\xi_j - \hat{\xi}_j^{(m)}),$$

subject to $\phi_i \geq 0$, $\psi_i \geq 0$, $\phi_i - \psi_i = Y_i^* - \mathbf{X}_i^{*\top}\boldsymbol{\xi}$ for $i \in \{1,\ldots,n\}$ and $t_j \geq |\xi_j|$ for $j \in \{1,\ldots,q+p\}$. For notational simplicity, here λ refers to the penalty parameters λ_1, λ_2 with respect to $\boldsymbol{\gamma}$ and $\boldsymbol{\beta}$. The unpenalized estimates $\widehat{\boldsymbol{\xi}}$ can be used for initialization. The resulting minimizer is the penalized estimator $\widetilde{\boldsymbol{\xi}} = (\widetilde{\boldsymbol{\gamma}}^\top,\widetilde{\boldsymbol{\beta}}^\top)^\top$.

Besides the difference convex algorithm, approximation to the SCAD penalty is commonly used to handle its non-convexity. There are several approaches available in the literature. For example, Fan and Li [2001] proposed a quadratic approximation and Zou and Li [42] developed a linear approximation for the penalty function. Here the difference convex algorithm is readily available through a penalized linear quantile estimation and thus it is implemented in our numerical studies.

5. Numerical illustrations

In this section, we conduct Monte Carlo studies to assess the finite-sample performance of the proposed estimation and variable selection approaches. The first example shows the effectiveness of the proposed iterative

estimation procedure and the second example illustrates how the variable selection performs with SCAD penalty. As an illustration, we then look for the most important variables affecting the Boston housing prices using the proposed Q-PLSIMs.

5.1. Simulation

Example 1. (Estimation for Q-PLSIMs with polynomial splines.) We generate the sample with $n = 200$ observations from the sine-bump model defined, for all $i \in \{1, \ldots, n\}$, by

$$Y_i = \sin\{(\mathbf{X}_i^\top \boldsymbol{\gamma} - a)\pi/(b-a)\} + Z_i\boldsymbol{\beta} + 0.1\, e_i,$$

where the true parameters are $\boldsymbol{\gamma}_0 = (1/\sqrt{3}, 1/\sqrt{3}, 1/\sqrt{3})^\top$ and $\boldsymbol{\beta}_0 = 0.3$. a and b are taken as $\sqrt{3}/2 - 1.645/\sqrt{12}$ and $\sqrt{3}/2 + 1.645/\sqrt{12}$, respectively. This model is widely used in the semiparametric modeling literature, e.g., [2, 22]. $\mathbf{X}$ consists of three covariates X_1, X_2 and X_3 that are independent and uniformly distributed as $\mathcal{U}(0, 1)$, while Z takes 0 for the odd observations and 1 for the even observations. We consider different distributions for the error terms, namely, the standard Gaussian distribution, the Student t distribution with 3 degrees of freedom, and the Laplace distribution with location 0 and scale 1. For each scenario, 500 data sets are generated. Estimation results at different quantile levels, i.e., median ($\tau = 0.5$), first quartile ($\tau = 0.25$) and third quartile ($\tau = 0.75$) are reported.

We discuss in Theorem 1 the order and the optimal choice of spline dimension $K = K' + s$, where K' is the number of interior knots and s is the order of spline basis function. In practice given $\widehat{\boldsymbol{\gamma}}$ and $\widehat{\boldsymbol{\beta}}$, we choose the number of spline basis functions using the smallest Schwarz-type information criterion [13, 29], viz.

$$SIC_1(K) = \ln\left[\sum_{i=1}^{n} \rho_\tau\{Y_i - \mathbf{B}^\top(\mathbf{X}_i^\top\widehat{\boldsymbol{\gamma}})\boldsymbol{\delta} - \mathbf{Z}_i^\top\widehat{\boldsymbol{\beta}}\}\right] + \frac{\ln n}{2n} K.$$

For the spline smoothing, one needs to determine the degree of the splines, the number of knots, and the location of these knots. Cubic splines are most commonly used to estimate a smooth function [6, 26]. Usually the spline knots are taken to be equally spaced or at the equally-spaced quantiles; see, e.g., [15, 36]. Throughout our empirical studies, we follow the common practice and use cubic splines ($s = 4$) with equally spaced knots. This choice works well in our numerical study. To determine the number of knots, we implement the following. For each random sample, 0 to 10 possible interior knots were used to estimate the nonparametric function and the number of knots with the smallest $SIC_1(K)$ considered as the optimal. Over different quantile levels and different error distributions, the number of interior knots is frequently selected to be 1, in which case the number of spline basis functions K is 5.

For each generated sample with sample size $n = 200$, we can obtain the single-index estimates $\widehat{\boldsymbol{\gamma}}$ and partially linear estimates $\widehat{\boldsymbol{\beta}}$. In Monte Carlo studies with 500 replicates, we evaluate the performance of our estimation approach by computing the Monte Carlo mean, standard error, and bias of the parameter estimates over 500 Monte Carlo samples. Table 1 summarizes the results. Overall, the estimates are close to the true parameters, while the standard errors are slightly larger for the models with t-distributed errors and Laplace-distributed errors than those with Gaussian errors. Boxplots for parameter estimates with Gaussian errors are displayed in Figure 1 at quantile levels 0.5, 0.25 and 0.75. The fitted curves for median regression with Gaussian errors are presented in Figure 2.

In addition, we find that the proposed iterative approach based on essentially two linear quantile regressions yields fast computation compared with a profile approach. In this simulation study, we observe that the profile approach produces similar parameter estimates. However, the profile approach takes noticeably longer computation time than the proposed iterative algorithm even in this moderate 4-dimensional nonlinear optimization problem. For the profile approach, R function `optim` is used for nonlinear optimization over the parameter space $\boldsymbol{\xi} = (\boldsymbol{\gamma}^\top, \boldsymbol{\beta}^\top)^\top$. All numerical studies were completed using Macbook Pro with 2.4 GHz Intel Core 2 Duo under OS X Version 10.9.5.

Example 2. (Penalized Variable selection via SCAD for Q-PLSIMs.) We further investigate the performance of variable selection using SCAD penalty for quantile partially linear single-index models, simultaneously selecting the single-index variables and partially linear variables. We simulate the data from the sine-bump design as in Example 1, viz. $Y_i = \sin\{(\mathbf{X}_i^\top\boldsymbol{\gamma} - a)\pi/(b-a)\} + \mathbf{Z}_i^\top\boldsymbol{\beta} + 0.1\, e_i$ for all $i \in \{1, \ldots, n\}$, where $\mathbf{X}_i \in \mathbb{R}^8$ and $\mathbf{Z}_i \in \mathbb{R}^{12}$ are generated from the uniform distribution $\mathcal{U}(0, 1)$, and three distributions are used for random errors e: $e \sim \mathcal{N}(0, 1)$, $e \sim t_3$, and $e \sim Laplace(0, 1)$. Again, 500 realizations are sampled for each case and each realization consists of

Table 1: Summary of parameter estimates from Q-PLSIMs in Example 1. True $\boldsymbol{\gamma}_0$ is $(0.5774, 0.5774, 0.5774)^\top$ and true $\boldsymbol{\beta}_0$ is 0.3. The sample mean ("mean"), standard error ("se", in parentheses), and bias ("bias") of the parameter estimates are calculated over 500 simulations across different quantile levels ($\tau \in \{0.25, 0.5, 0.75\}$) and different error terms.

	$\tau = 0.25$		$\tau = 0.5$		$\tau = 0.75$	
	Mean(se)	bias	Mean(se)	Bias	Mean(se)	Bias
			$e \sim \mathcal{N}(0,1)$			
γ_1	0.5777 (0.0139)	0.0003	0.5774 (0.0117)	0.0000	0.5772 (0.0126)	−0.0001
γ_1	0.5765 (0.0149)	−0.0008	0.5766 (0.0135)	−0.0007	0.5766 (0.0146)	−0.0007
γ_1	0.5773 (0.0141)	0.0000	0.5777 (0.0123)	0.0003	0.5777 (0.0135)	0.0004
γ_1	0.3028 (0.0199)	0.0028	0.3016 (0.0172)	0.0016	0.3038 (0.0184)	0.0038
			$e \sim t_3$			
γ_1	0.5776 (0.0099)	0.0002	0.5775 (0.0083)	0.0001	0.5773 (0.0096)	−0.0001
γ_2	0.5769 (0.0104)	−0.0005	0.5767 (0.0086)	−0.0007	0.5764 (0.0105)	−0.0009
γ_3	0.5774 (0.0097)	0.0000	0.5777 (0.0078)	0.0003	0.5781 (0.0094)	0.0008
β	0.3032 (0.0156)	0.0032	0.3015 (0.0122)	0.0015	0.3021 (0.0137)	0.0021
			$e \sim Laplace(0,1)$			
γ_1	0.5778 (0.0126)	0.0004	0.5770 (0.0085)	−0.0004	0.5765 (0.0107)	−0.0009
γ_2	0.5774 (0.0130)	0.0000	0.5773 (0.0095)	−0.0001	0.5775 (0.0131)	0.0002
γ_3	0.5765 (0.0127)	−0.0009	0.5776 (0.0093)	0.0002	0.5777 (0.0128)	0.0003
β	0.3029 (0.0174)	0.0029	0.3012 (0.0115)	0.0012	0.3023 (0.0167)	0.0023

Table 2: Summary of variable selection results from Q-PLSIMs for Example 2. SCAD penalty is used to select relevant variables in "SCAD", while only true variables are used in "Oracle". "C" is the average number of correctly detected zero components in the single-index part ($\boldsymbol{\gamma}$) and partially linear part ($\boldsymbol{\beta}$) over 500 simulations. "I" is the average number of nonzero components that are incorrectly set to zero. "MSE" is the mean squared error of the parameter estimates. "MRME" is the median relative model error.

		$\boldsymbol{\gamma}$				$\boldsymbol{\beta}$			
Quantile	Method	C	I	MSE	MRME	C	I	MSE	MRME
					$e \sim \mathcal{N}(0,1)$				
0.5	SCAD	3.89	0	0.0008	0.2989	5.78	0.068	0.0122	0.4645
	Oracle	4	0	0.0006	0.1909	6	0	0.0077	0.2983
0.75	SCAD	3.83	0	0.0010	0.4162	5.77	0.002	0.0112	0.3976
	Oracle	4	0	0.0007	0.2812	6	0	0.0081	0.1992
0.25	SCAD	3.80	0	0.0013	0.2607	5.73	0.004	0.0141	0.4498
	Oracle	4	0	0.0007	0.1255	6	0	0.0150	0.6501
					$e \sim t_3$				
0.5	SCAD	3.94	0	0.0003	0.2958	5.90	0	0.0039	0.4768
	Oracle	4	0	0.0003	0.2377	6	0	0.0044	0.7672
0.75	SCAD	3.85	0	0.0005	0.3556	5.72	0	0.0073	0.4390
	Oracle	4	0	0.0004	0.2307	6	0	0.0046	0.1453
0.25	SCAD	3.84	0	0.0007	0.2448	5.57	0	0.0139	0.5965
	Oracle	4	0	0.0004	0.1137	6	0	0.0121	0.5859
					$e \sim Laplace(0,1)$				
0.5	SCAD	3.94	0	0.0004	0.2943	5.92	0.002	0.0045	0.4445
	Oracle	4	0	0.0003	0.2137	6	0	0.0052	0.7004
0.75	SCAD	3.86	0	0.0007	0.3851	5.67	0.002	0.0100	0.4805
	Oracle	4	0	0.0005	0.2257	6	0	0.0063	0.1359
0.25	SCAD	3.84	0	0.0010	0.2272	5.58	0.002	0.0157	0.5616
	Oracle	4	0	0.0006	0.1076	6	0	0.0136	0.6270

$n = 200$ observations. The true values for the single-index parameter $\boldsymbol{\gamma}$ are $(1, 3, 1.5, 0.5, 0, 0, 0, 0)^\top / \sqrt{12.5}$ and $(3, 2, 0, 0, 0, 1.5, 0, 0.2, 0.3, 0.15, 0, 0)^\top$ for the partially linear parameter $\boldsymbol{\beta}$. There are four zero components in the single-index part and six zero components in the partially linear part.

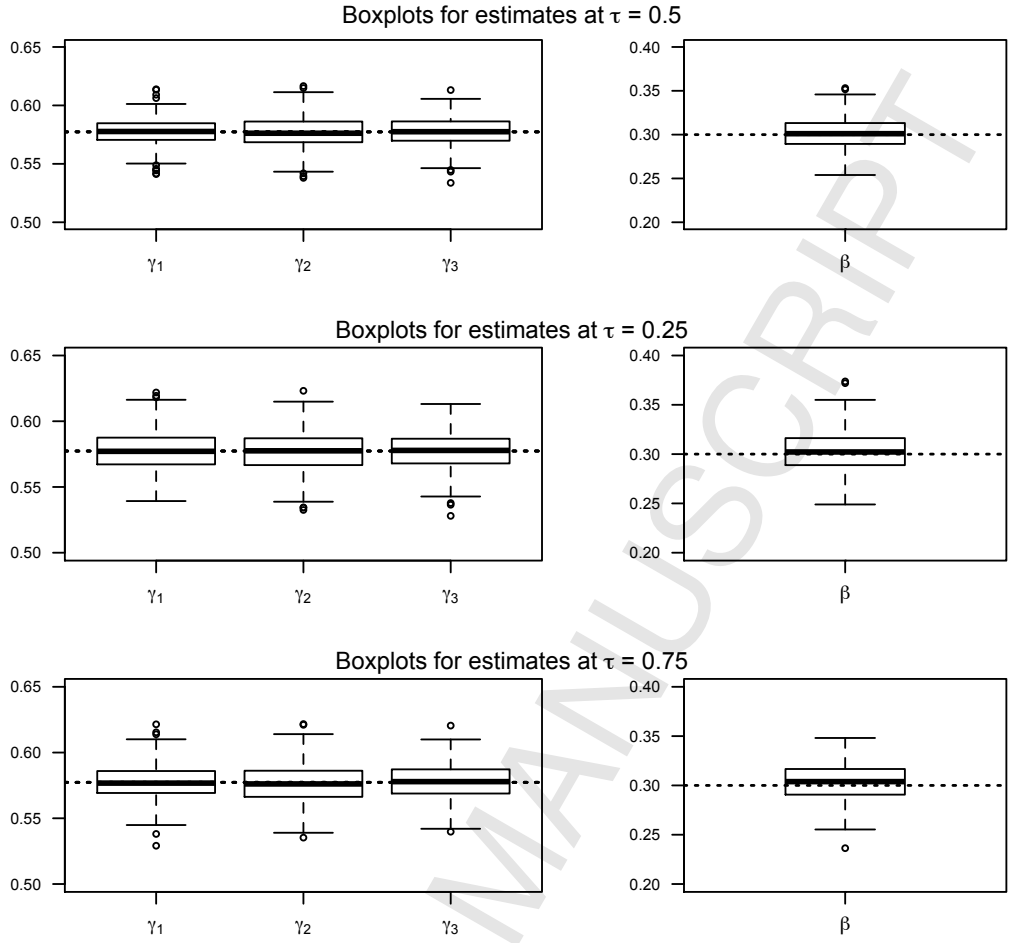

Figure 1: Boxplots of estimates from Q-PLSIM in Example 1 for the Gaussian error case. The left panels are for estimates of single-index parameters $\boldsymbol{\gamma}$ and the right panels are for the estimates of partially linear parameter $\boldsymbol{\beta}$ at quantile level 0.5, 0.25 and 0.75.

Table 2 reports the penalized variable selection results. "SCAD" gives the estimation and variable selection results for our proposed Q-PLSIMs using SCAD penalty. The "Oracle" only uses the exact nonzero components in the single-index part as well as the partially linear part. Overall, Q-PLSIMs with "SCAD" penalty can identify the irrelevant variables by setting zero parameters to zeros but still keep relevant variables. For median regression when $\tau = 0.5$, we observe correctly identified zeros (C) are more than 3.8 out of 4 in the single-index part and more than 5.7 out of 6 in the partially linear part, i.e., more than 95% of the zero components are correctly identified. For quantile levels at $\tau = 0.25$ and $\tau = 0.75$, correctly identified zeros are slightly fewer than those in the median regression but at least 92% of the zeros are correctly detected. In the single-index part, almost all those nonzero components are kept, namely, the average number of nonzero components that are incorrectly set to zero (I) is almost uniformly 0 for different quantiles and different error terms. Nonzero components in the partially linear part are occasionally set to zero, but on average no more than one nonzero component is incorrectly set to zero.

To assess the estimation accuracy, we compute the mean squared errors (MSE) defined as

$$\frac{1}{500}\sum_{j=1}^{500}\|\widetilde{\boldsymbol{\gamma}}_{(j)}-\boldsymbol{\gamma}_0\|^2$$

for the penalized single-index parameter estimates $\widetilde{\boldsymbol{\gamma}}$ and

$$\frac{1}{500}\sum_{j=1}^{500}\|\widetilde{\boldsymbol{\beta}}_{(j)}-\boldsymbol{\beta}_0\|^2$$

for the penalized partially linear parameter estimates $\widetilde{\boldsymbol{\beta}}$ over 500 simulations. Similar measures are used in Ma et al. [25], and by Liang and Li [21]. Furthermore, in order to show the effectiveness of penalized variable selection procedure, we also compute the median relative model errors [MRME, Fan and Li [8]; Liang et al. [22]] over 500 repetitions. The relative model error (RME) is defined as $\mathrm{ME}/\mathrm{ME}_{\mathrm{full}}$. ME for the single-index part can be calculated by $(\widetilde{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0)^\top \mathbf{X}\mathbf{X}^\top (\widetilde{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0)$, while the model error for the partially linear part is calculated by $(\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)^\top \mathbf{Z}\mathbf{Z}^\top (\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)$. ME_{full} stands for the model error under the full model without variable selection. In Table 2, MSEs and MRMEs are also computed under Oracle models ("Oracle"). We observe that MSEs and MRMEs from the proposed penalized variable selection procedure for Q-PLSIMs using SCAD penalty ("SCAD") are close to those from oracle models.

5.2. *An application to the Boston Housing Price data*

We apply the quantile partially linear single-index models to the Boston Housing Price data and examine important variables that affect housing prices. The data set can be downloaded from R package `MASS` (`https://cran.r-project.org/web/packages/MASS`). It consists of 506 observations and 14 variables, among which *medv* (median value of owner occupied homes in $1,000s in Boston suburb area in 1970s) is the dependent variable of interest, and 13 independent variables are *crim* (per capita crime rate by town), *zn* (proportion of residential land zoned for lots over 25,000 squared feet), *indus* (proportion of non-retail business acres per town), *chas* (Charles River dummy variable taking 1 if tract bounds river, 0 otherwise), *nox* (nitric oxides concentration, parts per 10 million), *rm* (average number of rooms per dwelling), *age* (proportion of owner-occupied units built prior to 1940), *dis* (weighted distances to five Boston employment centers), *rad* (index of accessibility to radial highways), *tax* (full-value property-tax rate per USD 10,000), *ptratio* (pupil-teacher ratio by town), *black* ($= 1thou(B - 0.63)^2$ where B is the proportion of blacks by town), and *lstat* (percentage of lower status of the population).

The Boston Housing Price data have been widely studied using nonparametric quantile regressions because of important nonlinear features and the fact that the median housing price has a ceiling at $50,000 and the distributions of *medv* and the major covariates such as *lstat* are skewed. Pre-specified important covariates are usually adopted. For example, Yu and Lu [35] and Wu et al. [34] included four covariates *(rm,* $\ln(tax)$*, ptratio,* $\ln(lstat)$*)* using an additive and a single-index quantile model, respectively; Kong and Xia [18] considered all 13 covariates in a single-index quantile model. Cai and Xu [1] examined a smooth coefficient quantile regression where the smooth coefficient is a function of *lstat* along with two variables (rm, $\ln(crim)$). They stressed the need for a more careful study of significant variables to include using a semiparametric model incorporating some interactions.

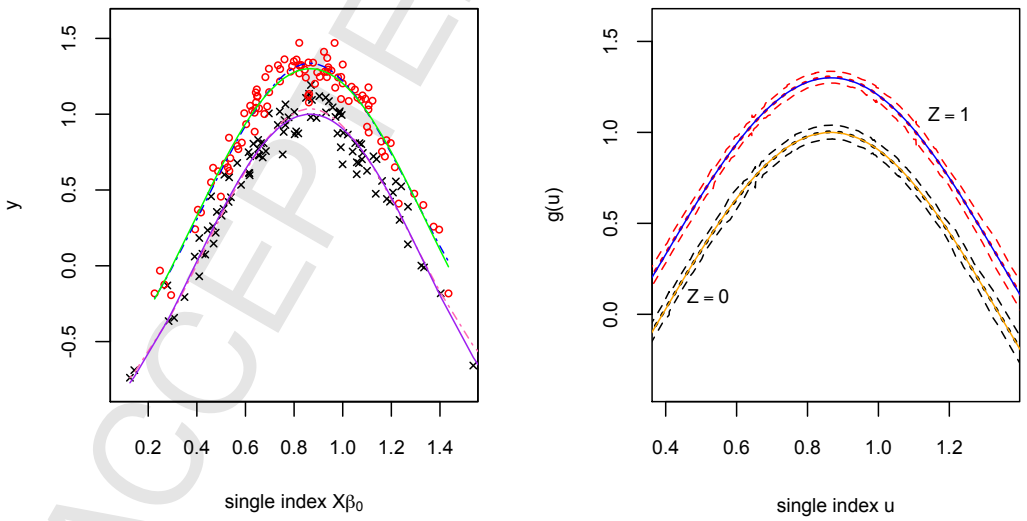

Figure 2: Fitted curves from Q-PLSIM in Example 1 for median and Gaussian error. The solid lines are true sinusoidal function when $Z = 1$ and $Z = 0$. The dot-dash lines in the left panel are the fitted curve from one simulation sample. In the right panel the dotted lines are the average fitted curves over 500 simulations and the dashed curves are the corresponding 2.5% and 97.5% confidence bands.

The proposed estimation and penalized variable selection for quantile partially linear single-index models can enable simultaneously estimating and selecting important variables in both the single-index part (SI) and the partially linear part (PL). It is natural to keep the two categorical variables *chas* and *rad* in the partially linear part. Harrison and Rubinfeld [10] claimed that the proportion of blacks by town (B) should have a quadratic relationship with the response variable, hence we also keep *black* in the PL part. Next, we use the proposed penalized variable selection for Q-PLSIMs with the remaining 10 continuous covariates into SI part and the three partially linear covariates to screen whether other continuous variables are significant in the single-index part. As in the previous studies, we take the logarithm of *tax* and *lstat*. Since *crim* is severely skewed to the left, we use ln(*crim*). All continuous covariates are standardized after the aforementioned transformations.

We consider the simultaneous estimation and variable selection for Q-PLSIMs over different quantile levels at 0.1, 0.25, 0.5, 0.75, and 0.9. First we run the penalized variable selection procedure and find that ln(*crim*), *zn*, *indus*, and *age* are not selected in the SI part over all quantile levels. We therefore include these four continuous covariates into the PL part rather than stopping at this point. Then we will focus on the following Q-PLSIM:

$$Q_\tau(medv) = g\{\gamma_1 rm + \gamma_2 \ln(tax) + \gamma_3 ptratio + \gamma_4 \ln(lstat) + \gamma_5 nox + \gamma_6 dis\} + \beta_1 \ln(crim) + \beta_2 zn + \beta_3 indus + \beta_4 age + \beta_5 black + \beta_6 chas + \beta_7 rad.$$

Table 3 summarizes the estimation and penalized variable selection results of Q-PLSIMs across different quantile levels. The selected number of interior knots is 2 for $\tau \in \{0.1, 0.25, 0.5\}$, 1 for $\tau = 0.75$, and 0 for $\tau = 0.9$ based on the smallest SIC_1. Equivalently, the number of spline basis functions K is 6, 5, and 4, respectively. Not surprisingly, the remaining variables in the single-index part show significant effect on housing prices. However, there is some variation at lower or upper quantiles; for example, the air pollution factor *nox* appears negligible at more extreme quantiles $\tau \in \{0.1, 0.75, 0.9\}$; also, ln(*tax*) has little effect on upper quantiles of housing prices.

Only a few variables in the partially linear part are selected by the penalized variable selection procedure. It is interesting to find that ln(*crim*) considered in the smooth-coefficient quantile model of Cai and Xu [1] is not selected across all quantiles here. The partially linear coefficients for *age* are negative and significant at high quantiles $\tau \in \{0.75, 0.9\}$. Older houses seem not attractive to high-income households that can afford more expensive properties. The partially linear coefficients for *black* are mostly positive and significant. This is consistent with the previous findings that the lower the proportion of blacks in a town, the higher house prices are in that town.

Figure 3 displays the single-index parameter estimates at nine quantile levels $\tau \in \{0.1, 0.2, 0.25, 0.4, 0.5, 0.6, 0.75, 0.8, 0.9\}$. The dots are estimated values of singe-index parameters, and the blank squares are corresponding 5% and 95% confidence limits over 200 bootstrap samples. To obtain the standard errors, one might use the asymptotic variance formulas derived in Theorems 2–3. However, they may be quite complicated to compute. Instead, we rely on the bootstrap, as Wu et al. [34]. The asymptotic variances of Theorems 2–3 involve several unknown quantities, such as the conditional density function $f(0 \mid \mathbf{X}, \mathbf{Z})$ and conditional expectation $\mathrm{E}_{\mathcal{M}}\{g^{(1)}(\mathbf{X}_i^\top \boldsymbol{\gamma}_0)\mathbf{X}\}$, etc. We may use estimates and proceed with the estimated conditional density and sample averages. However, estimating the unknown multivariate conditional density may be quite challenging; see Stone [27].

We did not include zero estimates in the estimation of bootstrapping samples. Therefore, no confidence limits are shown for those. Based on Figure 3, we found that the number of rooms per house (*rm*) has a positive effect uniformly across different quantiles. This matches the intuition that people usually like more space and multi-functional rooms. The effect of *rm* on housing prices seems less stronger at lower quantiles. Relatively lower-income households may need to balance their budge and size of the house. One would expect the property tax rate ln(*tax*) to have a negative impact on housing prices. It is worth noticing that the tax rate is not significant at higher quantile levels $\tau \in \{0.75, 0.8, 0.9\}$. Higher-income households may be less concerned with the tax rate when they can afford better houses. Both the pupil-teacher ratio (*ptratio*) and the percentage of lower (educational) status of the population ln(*lstat*) show negative influence on housing values. People care about the educational environment for their children and also have safety concerns related to adults without higher education degrees. It appears that air pollution (*nox*) tends to affect moderate-income households decision of buying a house more, while those luxury houses purchased by higher-income households are usually well located with better natural environment. Figure 3 also shows that the weighted distance to five industrial areas (*dis*) is negatively related to housing values. As mentioned in Harrison and Rubinfeld [10], housing prices tend to be higher near employment centers according to traditional theories of urban land rent gradients.

To compare our selected quantile partially linear single-index models with Wu et al. [34], Chaudhuri et al. [3], and Yu and Lu [35], we compute the average sum of check loss function defined as

$$R_\tau = \frac{1}{n}\sum_{i=1}^{n} \rho_\tau\{y_i - \widehat{g}(\boldsymbol{X}_i\widehat{\boldsymbol{\gamma}}) - \boldsymbol{Z}_i\widehat{\boldsymbol{\beta}}\}$$

Table 3: Estimation and variable selection results for Boston Housing Price Data. Estimates ("est") are from quantile partially linear single-index models with SCAD penalty. Bootstrap standard errors ("se") are only evaluated for selected variables.

	$\tau = 0.1$	$\tau = 0.25$	$\tau = 0.5$	$\tau = 0.75$	$\tau = 0.9$
Variable	est (se)	est (se)	est (se)	est (se)	est (se)
			Single-index terms		
rm	0.4054 (0.04)	0.5399 (0.03)	0.6622 (0.02)	0.7496 (0.03)	0.6890 (0.10)
ln(tax)	−0.4595 (0.04)	−0.3387 (0.02)	−0.1857 (0.02)	0 (–)	0 (–)
ptratio	−0.3507 (0.02)	−0.3256 (0.02)	−0.2834 (0.01)	−0.1774 (0.01)	0 (–)
ln(lstat)	−0.6320 (0.04)	−0.6322 (0.03)	−0.5750 (0.02)	−0.5385 (0.02)	−0.6604 (0.10)
nox	0 (–)	-0.2181 (0.02)	−0.2070 (0.02)	0 (–)	0 (–)
dis	−0.2523 (0.02)	−0.2012 (0.02)	−0.2706 (0.02)	−0.3180 (0.02)	−0.2261 (0.04)
			Partially linear terms		
ln(crim)	0 (–)	0 (–)	0 (–)	0 (–)	0 (–)
zn	0 (–)	0 (–)	0 (–)	0 (–)	0 (–)
indus	0 (–)	0 (–)	0 (–)	0 (–)	0 (–)
age	0 (–)	0 (–)	0 (–)	−1.1371 (0.11)	−0.8378 (0.25)
black	0.5805 (0.08)	0.8549 (0.07)	1.2718 (0.08)	1.3216 (0.06)	0 (–)
chas	0 (–)	0 (–)	0 (–)	0 (–)	0 (–)
rad	0.0422 (0.03)	0 (–)	0 (–)	0 (–)	0 (–)

in Table 4 as in Wu et al. [34] (WYY). WYY-1 in Table 4 refers to the first single-index model of WYY using four pre-specified variables rm, ln(tax), ptratio and ln(lstat). WYY-2 in Table 4 refers to the second single-index model of WYY using three variables *rm*, *lstat*, and *dist*. The last row ADE refers to the average derivative methods (ADE) of Chaudhuri et al. [3] using variables *rm*, *lstat*, and *dist*. Additive refers to Yu and Lu [35] using local linear additive models $Q_\tau(medv) = g_1(rm) + g_2\{\ln(tax)\} + g_3(ptratio) + g_4\{\ln(lstat)\}$. Note that all these papers use pre-specified variables without formal variable selection. We see from Table 4 that the Q-PLSIM models perform consistently best across the five quantile levels in terms of model errors measured by the check loss function.

It is a bit surprising that the Q-PLSIM even gives smaller check loss than those from additive model of Yu and Lu [35] at most quantile levels except for quantile level 0.9. Unlike in the mean regression context, Carroll et al. [2] found mixed results in their applications where generalized additive models sometimes have better performance but with larger "effective" model degrees freedom, i.e., more complex model. Note that a model complexity measure is not considered in the check loss function R_τ. As pointed out in Wu et al. [34], the theory on model complexity measure for quantile regression, especially for nonparametric/semiparametric quantile regression, is very limited. This may deserve more research.

In summary, quantile partially linear single-index models reduce dimensionality and are flexible. We find that the proposed estimation and penalized variable selection for Q-PLSIMs using polynomial splines are effective. They can also give some natural guidance in partitioning variables into the single-index (nonparametric) part versus partially linear (parametric) part. Our findings also shed some new light in the Boston Housing Price application. Besides commonly considered variables such as *rm,* ln(*tax*)*, ptratio,* ln(*lstat*) in the literature, we find in the single-index part air pollution *nox* are negative and significant in the middle quantile levels; and the weighted distances to five Boston employment centers *dis* are negative and significant across all quantile levels. However, ln(*crim*) is not selected across all quantiles.

Table 4: Model comparison for Boston Housing Price Data. The averages of check loss function are calculated for different models at the quantile level 0.1, 0.25, 0.5, 0.75 and 0.9.

τ	0.1	0.25	0.5	0.75	0.9
PLSIM	0.504	0.954	1.249	1.147	0.769
WYY-1	1.102	2.105	2.845	2.577	1.749
WYY-2	1.228	2.229	2.874	2.490	3.320
ADE	1.559	2.696	3.042	2.430	3.126
Additive	0.544	1.011	1.336	1.185	0.744

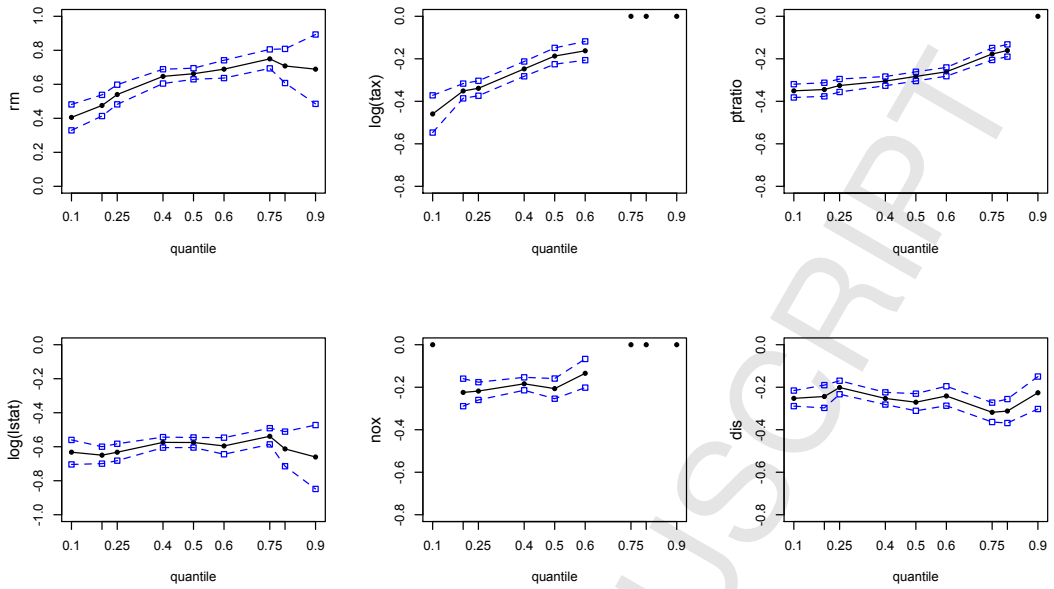

Figure 3: Single-index parameter estimates for Boston Housing Price Data over Different Quantile Levels from Q-PLSIM with SCAD Penalty. Quantile levels considered are 0.1, 0.2, 0.25, 0.4, 0.5, 0.6, 0.75, 0.8, and 0.9. The solid dots are the estimates at each quantile and the blank squares are point-wise 2.5% and 97.5% confidence limits for non-zero components.

Appendix: Technical proofs

In this appendix, we give proofs of the five main theorems. First, the following seven lemmas need to be established. To save space, detailed proofs of all lemmas are relegated to the online supplemental resources.

Let $\boldsymbol{\delta}_0$ be spline coefficients in the best spline approximation of g with $\sup_t |g(t) - \mathbf{B}^\top(t)\boldsymbol{\delta}_0| \le CK^{-d}$, which is possible by (A3). Let $F(\cdot \mid \mathbf{X}, \mathbf{Z})$ be the conditional cdf of e given the covariates. We also write $g(\mathbf{X}_i^\top \boldsymbol{\gamma}_0)$ as g_i and let $m_i = g_i + \mathbf{Z}_i^\top \boldsymbol{\beta}_0$. In the proofs C denotes a generic positive constant which may assume different values even on the same line.

Lemma 1. *Let* $r_n = \sqrt{K/n} + K^{-d}$.

$$\sup_{\|\boldsymbol{\gamma}-\boldsymbol{\gamma}_0\|+\|\boldsymbol{\beta}-\boldsymbol{\beta}_0\|+\|\boldsymbol{\delta}-\boldsymbol{\delta}_0\| \le Cr_n} \sum_{i=1}^n \rho_\tau\{Y_i - \mathbf{B}^\top(\mathbf{X}_i^\top\boldsymbol{\gamma})\boldsymbol{\delta} - \mathbf{Z}_i^\top\boldsymbol{\beta}\} - \sum_{i=1}^n \rho_\tau\{Y_i - \mathbf{B}^\top(\mathbf{X}_i^\top\boldsymbol{\gamma}_0)\boldsymbol{\delta}_0 - \mathbf{Z}_i^\top\boldsymbol{\beta}_0\}$$
$$+ \sum_{i=1}^n \{\mathbf{B}^\top(\mathbf{X}_i^\top\boldsymbol{\gamma})\boldsymbol{\delta} - \mathbf{B}^\top(\mathbf{X}_i^\top\boldsymbol{\gamma}_0)\boldsymbol{\delta}_0 + \mathbf{Z}_i^\top(\boldsymbol{\beta}-\boldsymbol{\beta}_0)\}\{\tau - \mathbf{1}(e_i \le 0)\}$$
$$-\mathrm{E}\sum_{i=1}^n \rho_\tau\{Y_i - \mathbf{B}^\top(\mathbf{X}_i^\top\boldsymbol{\gamma})\boldsymbol{\delta} - \mathbf{Z}_i^\top\boldsymbol{\beta}\} + \mathrm{E}\sum_{i=1}^n \rho_\tau\{Y_i - \mathbf{B}^\top(\mathbf{X}_i^\top\boldsymbol{\gamma}_0)\boldsymbol{\delta}_0 - \mathbf{Z}_i^\top\boldsymbol{\beta}_0\} = o_p(nr_n^2),$$

where the expectations are over Y_i *conditional on* $\mathbf{X}_i, \mathbf{Z}_i$ *(all expectations below are also such conditional expectations).*

Now to show the convergence rate of the estimator, suppose $\|\boldsymbol{\gamma}-\boldsymbol{\gamma}_0\| + \|\boldsymbol{\beta}-\boldsymbol{\beta}_0\| + \|\boldsymbol{\delta}-\boldsymbol{\delta}_0\| = Lr_n$ for sufficiently large $L > 0$.

Lemma 2.

$$\inf_{\|\boldsymbol{\gamma}-\boldsymbol{\gamma}_0\|+\|\boldsymbol{\beta}-\boldsymbol{\beta}_0\|+\|\boldsymbol{\delta}-\boldsymbol{\delta}_0\|=Lr_n} \sum_{i=1}^n \mathrm{E}\rho_\tau\{e_i + m_i - \mathbf{B}^\top(\mathbf{X}_i^\top\boldsymbol{\gamma})\boldsymbol{\delta} - \mathbf{Z}_i^\top\boldsymbol{\beta}\} - \sum_{i=1}^n \mathrm{E}\rho_\tau\{e_i + m_i - \mathbf{B}^\top(\mathbf{X}_i^\top\boldsymbol{\gamma}_0)\boldsymbol{\delta}_0 - \mathbf{Z}_i^\top\boldsymbol{\beta}_0\} \ge L^2 Cnr_n^2.$$

Lemma 3. *The eigenvalues of*

$$\frac{1}{n}\sum_{i=1}^{n}\begin{pmatrix}\widetilde{\mathbf{J}}^\top\mathbf{B}^{(1)}(\mathbf{X}_i^\top\boldsymbol{\gamma}_0)^\top\boldsymbol{\delta}_0\mathbf{X}_i\\ \mathbf{Z}_i\\ \mathbf{B}(\mathbf{X}_i^\top\boldsymbol{\gamma}_0)\end{pmatrix}(\mathbf{B}^{(1)}(\mathbf{X}_i^\top\boldsymbol{\gamma}_0)^\top\boldsymbol{\delta}_0\mathbf{X}_i^\top\widetilde{\mathbf{J}},\mathbf{Z}_i^\top,\mathbf{B}^\top(\mathbf{X}_i^\top\boldsymbol{\gamma}_0))$$

are bounded and bounded away from zero with probability approaching 1.

The following lemma deals with one of the terms in the statement of Lemma 1. For models with single-index structure, its proof is more complicated than for additive or varying coefficient models (due to the fact that for the latter models the parametric part and nonparametric part are added up to give the regression function).

Lemma 4.

$$\sup_{\|\boldsymbol{\gamma}-\boldsymbol{\gamma}_0\|+\|\boldsymbol{\beta}-\boldsymbol{\beta}_0\|+\|\boldsymbol{\delta}-\boldsymbol{\delta}_0\|=Lr_n}\sum_{i=1}^{n}\{\mathbf{B}^\top(\mathbf{X}_i^\top\boldsymbol{\gamma})\boldsymbol{\delta}-\mathbf{B}^\top(\mathbf{X}_i^\top\boldsymbol{\gamma}_0)\boldsymbol{\delta}_0+\mathbf{Z}_i^\top(\boldsymbol{\beta}-\boldsymbol{\beta}_0)\}\{\tau-\mathbf{1}(e_i\le 0)\}=LO_p(nr_n^2).$$

Proof of Theorem 1. Combining Lemmas 1, 2, 4, we get, for sufficiently large $L>0$,

$$\Pr\Big[\inf_{\|\boldsymbol{\gamma}-\boldsymbol{\gamma}_0\|+\|\boldsymbol{\beta}-\boldsymbol{\beta}_0\|+\|\boldsymbol{\delta}-\boldsymbol{\delta}_0\|=Lr_n}\sum_{i=1}^{n}\rho_\tau\{Y_i-\mathbf{B}^\top(\mathbf{X}_i^\top\boldsymbol{\gamma})\boldsymbol{\delta}-\mathbf{Z}_i^\top\boldsymbol{\beta}\}>\sum_{i=1}^{n}\rho_\tau\{Y_i-\mathbf{B}^\top(\mathbf{X}_i^\top\boldsymbol{\gamma}_0)\boldsymbol{\delta}_0-\mathbf{Z}_i^\top\boldsymbol{\beta}_0\}\Big]\to 1,$$

and thus there is a local minimizer of $(\widehat{\boldsymbol{\gamma}},\widehat{\boldsymbol{\delta}})$ with $\|\widehat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}_0\|+\|\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0\|+\|\widehat{\boldsymbol{\delta}}-\boldsymbol{\delta}_0\|=O_p(r_n)$. □

Now we consider asymptotic normality. The challenge here is the need to perform orthogonalization appropriately. Since the parametric part is nested within the spline basis, orthogonalization is more complicated than partially linear models as studied in [29, 30].

Let $\boldsymbol{\Pi}_i=\mathbf{B}(\mathbf{X}_i^\top\boldsymbol{\gamma}_0)$ and $\boldsymbol{\Pi}$ be the $n\times K$ matrix with rows $\boldsymbol{\Pi}_i^\top$. The empirical version of the minimization problem corresponding to the projection is

$$\min_{\boldsymbol{\delta}}\sum_{i=1}^{n}f(0\mid\mathbf{X}_i,\mathbf{Z}_i)(W_i-\boldsymbol{\Pi}_i^\top\boldsymbol{\delta})^2,$$

with the minimizer $(\boldsymbol{\Pi}^\top\boldsymbol{\beta}\boldsymbol{\Pi})^{-1}\boldsymbol{\Pi}^\top\boldsymbol{\beta}\mathbf{W}$ where $\boldsymbol{\beta}$ is the diagonal matrix with diagonal elements $f(0\mid\mathbf{X}_i,\mathbf{Z}_i)$ and $\mathbf{W}=(W_1,\ldots,W_n)^\top$. Define $\mathbf{P}=\boldsymbol{\Pi}(\boldsymbol{\Pi}^\top\boldsymbol{\beta}\boldsymbol{\Pi})^{-1}\boldsymbol{\Pi}^\top\boldsymbol{\beta}$. We write

$$\begin{aligned}&\rho_\tau\{e_i+m_i-\mathbf{B}^\top(\mathbf{X}_i^\top\boldsymbol{\gamma})\boldsymbol{\delta}-\mathbf{Z}_i^\top\boldsymbol{\beta}\}\\&\qquad=\rho_\tau\{e_i-\mathbf{B}^\top(\mathbf{X}_i^\top\boldsymbol{\gamma}_0)(\boldsymbol{\delta}-\boldsymbol{\delta}_0)-\mathbf{B}^{(1)\top}(\mathbf{X}_i^\top\boldsymbol{\gamma}_0)\boldsymbol{\delta}_0\mathbf{X}_i^\top(\boldsymbol{\gamma}-\boldsymbol{\gamma}_0)-\mathbf{Z}_i^\top(\boldsymbol{\beta}-\boldsymbol{\beta}_0)-R_i(\boldsymbol{\gamma},\boldsymbol{\delta})\}\\&\qquad=\rho_\tau\{e_i-\boldsymbol{\Pi}_i^\top(\boldsymbol{\delta}-\boldsymbol{\delta}_0)-\mathbf{U}_i^\top(\boldsymbol{\gamma}-\boldsymbol{\gamma}_0)-\mathbf{Z}_i^\top(\boldsymbol{\beta}-\boldsymbol{\beta}_0)-R_i(\boldsymbol{\gamma},\boldsymbol{\delta})\},\end{aligned}$$

where we define $\mathbf{U}_i=\mathbf{B}^{(1)\top}(\mathbf{X}_i^\top\boldsymbol{\gamma}_0)\boldsymbol{\delta}_0\mathbf{X}_i$ and

$$R_i(\boldsymbol{\gamma},\boldsymbol{\delta})=\{\mathbf{B}(\mathbf{X}_i^\top\boldsymbol{\gamma})-\mathbf{B}(\mathbf{X}_i^\top\boldsymbol{\gamma}_0)\}^\top\boldsymbol{\delta}-\mathbf{B}^{(1)\top}(\mathbf{X}_i^\top\boldsymbol{\gamma}_0)\boldsymbol{\delta}_0\mathbf{X}_i^\top(\boldsymbol{\gamma}-\boldsymbol{\gamma}_0)+\{\mathbf{B}^\top(\mathbf{X}_i^\top\boldsymbol{\gamma}_0)\boldsymbol{\delta}_0-g_i\}=R_{i1}(\boldsymbol{\gamma},\boldsymbol{\delta})+R_{i2}(\boldsymbol{\gamma},\boldsymbol{\delta}).$$

Let $\mathbf{V}=\mathbf{U}-\mathbf{P}\mathbf{U}$ with the ith row of $\mathbf{V}$ denoted by $\mathbf{V}_i^\top=\mathbf{U}_i^\top-\mathbf{P}_i^\top\mathbf{U}$. Let $\mathbf{W}=\mathbf{Z}-\mathbf{P}\mathbf{Z}$ with the ith row of $\mathbf{W}$ denoted by $\mathbf{W}_i^\top=\mathbf{Z}_i^\top-\mathbf{P}_i^\top\mathbf{Z}$. Let $\mathbf{Q}=(\mathbf{V},\mathbf{W})$ with rows $\mathbf{Q}_i^\top=(\mathbf{V}_i^\top,\mathbf{W}_i^\top)$ and denote $\boldsymbol{\xi}=(\boldsymbol{\gamma}^\top,\boldsymbol{\beta}^\top)^\top$, $\boldsymbol{\xi}_0=(\boldsymbol{\gamma}_0^\top,\boldsymbol{\beta}_0^\top)^\top$. To carry out orthogonalization, we further write

$$\begin{aligned}&\rho_\tau\{e_i-\boldsymbol{\Pi}_i^\top(\boldsymbol{\delta}-\boldsymbol{\delta}_0)-\mathbf{U}_i^\top(\boldsymbol{\gamma}-\boldsymbol{\gamma}_0)-\mathbf{Z}_i^\top(\boldsymbol{\beta}-\boldsymbol{\beta}_0)-R_i(\boldsymbol{\gamma},\boldsymbol{\delta})\}\\&\qquad=\rho_\tau\{e_i-\boldsymbol{\Pi}_i^\top((\boldsymbol{\delta}-\boldsymbol{\delta}_0)+(\boldsymbol{\Pi}^\top\boldsymbol{\beta}\boldsymbol{\Pi})^{-1}\boldsymbol{\Pi}^\top\boldsymbol{\beta}\mathbf{Q}(\boldsymbol{\xi}-\boldsymbol{\xi}_0))-\mathbf{Q}_i^\top(\boldsymbol{\xi}-\boldsymbol{\xi}_0)-R_i(\boldsymbol{\gamma},\boldsymbol{\delta})\}\\&\qquad=\rho_\tau\{e_i-\boldsymbol{\Pi}_i^\top\boldsymbol{\eta}-\mathbf{Q}_i^\top(\boldsymbol{\xi}-\boldsymbol{\xi}_0)-R_i(\boldsymbol{\gamma},\boldsymbol{\delta})),\end{aligned}$$

with $\boldsymbol{\eta}=\boldsymbol{\delta}-\boldsymbol{\delta}_0+(\boldsymbol{\Pi}^\top\boldsymbol{\beta}\boldsymbol{\Pi})^{-1}\boldsymbol{\Pi}^\top\boldsymbol{\beta}\mathbf{Q}(\boldsymbol{\xi}-\boldsymbol{\xi}_0)$. The proof of the following lemma is similar to Lemma 1, but requires finer analysis of $R_i(\boldsymbol{\gamma},\boldsymbol{\delta})$.

Lemma 5.

$$\sup_{\|\boldsymbol{\xi}-\boldsymbol{\xi}_0\|\le C/\sqrt{n},\|\boldsymbol{\eta}\|\le Cr_n}\Bigg|\sum_{i=1}^n\rho_\tau\{e_i-\boldsymbol{\Pi}_i^\top\boldsymbol{\eta}-\mathbf{Q}_i^\top(\boldsymbol{\xi}-\boldsymbol{\xi}_0)-R_i(\boldsymbol{\gamma},\boldsymbol{\delta})\}$$
$$-\sum_{i=1}^n\rho_\tau\{e_i-\boldsymbol{\Pi}_i^\top\boldsymbol{\eta}-R_i(\boldsymbol{\gamma}_0,\boldsymbol{\delta})\}+\sum_{i=1}^n\mathbf{Q}_i^\top(\boldsymbol{\xi}-\boldsymbol{\xi}_0)e_i$$
$$-\mathrm{E}\sum_{i=1}^n\rho_\tau\{e_i-\boldsymbol{\Pi}_i^\top\boldsymbol{\eta}-\mathbf{Q}_i^\top(\boldsymbol{\xi}-\boldsymbol{\xi}_0)-R_i(\boldsymbol{\gamma},\boldsymbol{\delta})\}+\mathrm{E}\sum_{i=1}^n\rho_\tau\{e_i-\boldsymbol{\Pi}_i^\top\boldsymbol{\eta}-R_i(\boldsymbol{\gamma}_0,\boldsymbol{\delta})\}\Bigg|=o_p(1).$$

Lemma 6.

$$\sup_{\|\boldsymbol{\eta}\|\le Cr_n,\|\boldsymbol{\xi}-\boldsymbol{\xi}_0\|\le C/\sqrt{n}}\Bigg|\sum_{i=1}^n\mathrm{E}\rho_\tau\{e_i-\boldsymbol{\Pi}_i^\top\boldsymbol{\eta}-\mathbf{Q}_i(\boldsymbol{\xi}-\boldsymbol{\xi}_0)-R_i(\boldsymbol{\gamma},\boldsymbol{\delta})\}$$
$$-\sum_{i=1}^n\mathrm{E}\rho_\tau\{e_i-\boldsymbol{\Pi}_i^\top\boldsymbol{\eta}-R_i(\boldsymbol{\gamma}_0,\boldsymbol{\delta})\}-\sum_{i=1}^n\frac{f(0\mid\mathbf{X}_i,\mathbf{Z}_i)}{2}(\boldsymbol{\xi}-\boldsymbol{\xi}_0)\mathbf{Q}_i\mathbf{Q}_i^\top(\boldsymbol{\xi}-\boldsymbol{\xi}_0)\Bigg|=o_p(1).$$

Lemma 7.

$$\frac{1}{n}\sum_{i=1}^n f(0\mid\mathbf{X}_i,\mathbf{Z}_i)\mathbf{Q}_i\mathbf{Q}_i^\top\to\mathrm{E}\left[f(0\mid\mathbf{X},\mathbf{Z})\left[\begin{array}{c}g^{(1)}(\mathbf{X}^\top\boldsymbol{\gamma}_0)\mathbf{X}-\mathrm{E}_{\mathcal{M}}\{g^{(1)}(\mathbf{X}^\top\boldsymbol{\gamma}_0)\mathbf{X}\}\\ \mathbf{Z}-E_{\mathcal{M}}(\mathbf{Z})\end{array}\right]^{\otimes2}\right]\ \textit{in probability},$$

$$\frac{1}{n}\sum_{i=1}^n\mathbf{Q}_i\mathbf{Q}_i^\top\to\mathrm{E}\left[\left[\begin{array}{c}g^{(1)}(\mathbf{X}^\top\boldsymbol{\gamma}_0)\mathbf{X}-\mathrm{E}_{\mathcal{M}}\{g^{(1)}(\mathbf{X}^\top\boldsymbol{\gamma}_0)\mathbf{X}\}\\ \mathbf{Z}-E_{\mathcal{M}}(\mathbf{Z})\end{array}\right]^{\otimes2}\right]\ \textit{in probability}.$$

Proof of Theorem 2. Let $\widehat{\boldsymbol{\eta}}=\widehat{\boldsymbol{\delta}}-\boldsymbol{\delta}_0+(\boldsymbol{\Pi}^\top\boldsymbol{\beta}\boldsymbol{\Pi})^{-1}\boldsymbol{\Pi}^\top\boldsymbol{\beta}\mathbf{Q}(\boldsymbol{\xi}-\boldsymbol{\xi}_0)$. By Lemmas 5, 6, and 7,

$$\sup_{\|\boldsymbol{\xi}-\boldsymbol{\xi}_0\|\le C/\sqrt{n}}\Bigg|\sum_{i=1}^n\rho_\tau\{e_i-\boldsymbol{\Pi}_i^\top\widehat{\boldsymbol{\eta}}-\mathbf{Q}_i^\top(\boldsymbol{\xi}-\boldsymbol{\xi}_0)-R_i(\boldsymbol{\gamma},\widehat{\boldsymbol{\delta}})\}$$
$$-\sum_{i=1}^n\rho_\tau\{e_i-\boldsymbol{\Pi}_i^\top\widehat{\boldsymbol{\eta}}-R_i(\boldsymbol{\gamma}_0,\widehat{\boldsymbol{\delta}})\}+\sum_{i=1}^n\mathbf{Q}_i^\top(\boldsymbol{\xi}-\boldsymbol{\xi}_0)e_i-\frac{n}{2}(\boldsymbol{\xi}-\boldsymbol{\xi}_0)^\top\boldsymbol{\Phi}(\boldsymbol{\xi}-\boldsymbol{\xi}_0)\Bigg|=o_p(1).$$

Let $\boldsymbol{\xi}^{(-1)}=(\boldsymbol{\gamma}^{(-1)T},\boldsymbol{\beta}^\top)^\top$ (as before we regard $\boldsymbol{\xi}$ as a function of $\boldsymbol{\xi}^{(-1)}$), the above easily implies

$$\sup_{\|\boldsymbol{\xi}-\boldsymbol{\xi}_0\|\le C/\sqrt{n}}\Bigg|\sum_{i=1}^n\rho_\tau\{e_i-\boldsymbol{\Pi}_i^\top\widehat{\boldsymbol{\eta}}-\mathbf{Q}_i^\top(\boldsymbol{\xi}-\boldsymbol{\xi}_0)-R_i(\boldsymbol{\gamma},\widehat{\boldsymbol{\delta}})\}$$
$$-\sum_{i=1}^n\rho_\tau\{e_i-\boldsymbol{\Pi}_i^\top\widehat{\boldsymbol{\eta}}-R_i(\boldsymbol{\gamma}_0,\widehat{\boldsymbol{\delta}})\}+\sum_{i=1}^n\mathbf{Q}_i^\top\mathbf{J}(\boldsymbol{\xi}^{(-1)}-\boldsymbol{\xi}_0^{(-1)})e_i$$
$$-\frac{n}{2}(\boldsymbol{\xi}^{(-1)}-\boldsymbol{\xi}_0^{(-1)})^\top\mathbf{J}^\top\boldsymbol{\Phi}\mathbf{J}(\boldsymbol{\xi}^{(-1)}-\boldsymbol{\xi}_0^{(-1)})\Bigg|=o_p(1)\quad(A.1)$$

Denote

$$Q(\boldsymbol{\xi})=\frac{n}{2}(\boldsymbol{\xi}^{(-1)}-\boldsymbol{\xi}_0^{(-1)})^\top\mathbf{J}^\top\boldsymbol{\Phi}\mathbf{J}(\boldsymbol{\xi}^{(-1)}-\boldsymbol{\xi}_0^{(-1)})-\sum_{i=1}^n\mathbf{Q}_i^\top\mathbf{J}(\boldsymbol{\xi}^{(-1)}-\boldsymbol{\xi}_0^{(-1)})e_i$$

and define

$$\tilde{\boldsymbol{\xi}}^{(-1)}=\boldsymbol{\xi}_0^{(-1)}+\frac{1}{n}(\mathbf{J}^\top\boldsymbol{\Phi}\mathbf{J})^{-1}\sum_{i=1}^n\mathbf{J}^\top\mathbf{Q}_ie_i.$$

It follows from the Central Limit Theorem that

$$\sqrt{n}\,(\tilde{\boldsymbol{\xi}}^{(-1)}-\boldsymbol{\xi}_0^{(-1)})\rightsquigarrow\mathcal{N}[0,(\mathbf{J}^\top\boldsymbol{\Phi}\mathbf{J})^{-1}\mathbf{J}^\top\boldsymbol{\Sigma}\mathbf{J}(\mathbf{J}^\top\boldsymbol{\Phi}\mathbf{J})^{-1}],$$

and thus

$$\sqrt{n}\,(\tilde{\boldsymbol{\xi}} - \boldsymbol{\xi}_0) \rightsquigarrow \mathcal{N}[0, \mathbf{J}(\mathbf{J}^\top \boldsymbol{\Phi} \mathbf{J})^{-1} \mathbf{J}^\top \boldsymbol{\Sigma} \mathbf{J} (\mathbf{J}^\top \boldsymbol{\Phi} \mathbf{J})^{-1} \mathbf{J}^\top].$$

Note that $\tilde{\boldsymbol{\xi}} = (\tilde{\boldsymbol{\gamma}}^\top, \tilde{\boldsymbol{\beta}}^\top)^\top$ is the minimizer of $Q(\boldsymbol{\xi})$ (we think of $\boldsymbol{\xi}$ as a function of $\boldsymbol{\xi}^{(-1)}$ when appropriate) and $Q(\boldsymbol{\xi})$ is equal to $(\boldsymbol{\xi}^{(-1)} - \tilde{\boldsymbol{\xi}}^{(-1)})^\top \mathbf{J}^\top \boldsymbol{\Phi} \mathbf{J} (\boldsymbol{\xi}^{(-1)} - \tilde{\boldsymbol{\xi}}^{(-1)})$ plus a term that does not involve $\boldsymbol{\xi}$.

For any $\boldsymbol{\xi} = (\boldsymbol{\gamma}^\top, \boldsymbol{\beta}^\top)^\top$ with $\|\boldsymbol{\gamma}\| = 1$ and $\|\boldsymbol{\xi} - \tilde{\boldsymbol{\xi}}\| = \nu/\sqrt{n}$ with some small $\nu > 0$. By the quadratic form of Q, we have $Q(\boldsymbol{\xi}) - Q(\tilde{\boldsymbol{\xi}}) \geq C\nu^2$ and thus by (A.1),

$$\Pr\left[\sum_{i=1}^n \rho_\tau\{e_i - \boldsymbol{\Pi}_i^\top \widehat{\boldsymbol{\eta}} - \mathbf{Q}_i^\top (\boldsymbol{\xi} - \boldsymbol{\xi}_0) - R_i(\boldsymbol{\gamma}, \widehat{\boldsymbol{\delta}})\} > \sum_{i=1}^n \rho_\tau\{e_i - \boldsymbol{\Pi}_i^\top \widehat{\boldsymbol{\eta}} - \mathbf{Q}_i^\top (\tilde{\boldsymbol{\xi}} - \boldsymbol{\xi}_0) - R_i(\tilde{\boldsymbol{\gamma}}, \widehat{\boldsymbol{\delta}})\}\right] \to 1.$$

Since ν is arbitrarily small, we get $\|\widehat{\boldsymbol{\xi}} - \tilde{\boldsymbol{\xi}}\| = o_p(1/\sqrt{n})$ and the proof is complete. □

Proof of Theorem 3. We first show that there is a root-n consistent local minimizer of (2). As in the proof of Theorem 2, for $\widehat{\boldsymbol{\eta}} = \widehat{\boldsymbol{\delta}} - \boldsymbol{\delta}_0 + (\boldsymbol{\Pi}^\top \boldsymbol{\beta} \boldsymbol{\Pi})^{-1} \boldsymbol{\Pi}^\top \boldsymbol{\beta} \mathbf{Q} (\boldsymbol{\xi} - \boldsymbol{\xi}_0)$, we have

$$\sup_{\|\boldsymbol{\xi} - \boldsymbol{\xi}_0\| \leq C/\sqrt{n}} \Bigg| \sum_{i=1}^n \rho_\tau\{e_i - \boldsymbol{\Pi}_i^\top \widehat{\boldsymbol{\eta}} - \mathbf{Q}_i^\top (\boldsymbol{\xi} - \boldsymbol{\xi}_0) - R_i(\boldsymbol{\gamma}, \widehat{\boldsymbol{\delta}})\} - \sum_{i=1}^n \rho_\tau\{e_i - \boldsymbol{\Pi}_i^\top \widehat{\boldsymbol{\eta}} - R_i(\boldsymbol{\gamma}_0, \widehat{\boldsymbol{\delta}})\} + \sum_{i=1}^n \mathbf{Q}_i^\top (\boldsymbol{\xi} - \boldsymbol{\xi}_0) e_i - \frac{n}{2} (\boldsymbol{\xi} - \boldsymbol{\xi}_0)^\top \boldsymbol{\Phi} (\boldsymbol{\xi} - \boldsymbol{\xi}_0) \Bigg| = o_p(1).$$

For $\|\boldsymbol{\xi} - \boldsymbol{\xi}_0\| = L/\sqrt{n}$ with $L > 0$ sufficiently large, the quantity

$$-\sum_{i=1}^n \mathbf{Q}_i^\top (\boldsymbol{\xi} - \boldsymbol{\xi}_0) e_i + \frac{n}{2} (\boldsymbol{\xi} - \boldsymbol{\xi}_0)^\top \boldsymbol{\Phi} (\boldsymbol{\xi} - \boldsymbol{\xi}_0)$$

is bounded away from zero with probability approaching 1, and thus

$$\sum_{i=1}^n \rho_\tau\{e_i - \boldsymbol{\Pi}_i^\top \widehat{\boldsymbol{\eta}} - \mathbf{Q}_i^\top (\boldsymbol{\xi} - \boldsymbol{\xi}_0) - R_i(\boldsymbol{\gamma}, \widehat{\boldsymbol{\delta}})\} - \sum_{i=1}^n \rho_\tau\{e_i - \boldsymbol{\Pi}_i^\top \widehat{\boldsymbol{\eta}} - R_i(\boldsymbol{\gamma}_0, \widehat{\boldsymbol{\delta}})\}$$

is bounded away from zero with probability approaching 1. For the penalty terms, when $j \leq q_1$, with $|\gamma_j - \gamma_{0j}| \leq L/\sqrt{n}$, by the property of the SCAD penalty, we have $p_{\lambda_1}(|\gamma_j|) = p_{\lambda_1}(|\gamma_{0j}|)$ since $\lambda_1 = o(1)$ and both $|\gamma_j|$ and $|\gamma_{0j}|$ are bounded away from zero. For $j > q_1$, we have $p_{\lambda_1}(|\gamma_j|) \geq p_{\lambda_1}(|\gamma_{0j}|) = 0$. Similarly, $p_{\lambda_2}(|\beta_j|) = p_{\lambda_2}(|\beta_{0j}|)$ for $j \leq p_1$ and $p_{\lambda_2}(|\beta_j|) \geq p_{\lambda_2}(|\beta_{0j}|) = 0$ when $j > p_1$. Summarizing, we get

$$n \sum_{j=1}^q p_{\lambda_1}(|\gamma_j|) + n \sum_{j=1}^p p_{\lambda_2}(|\beta_j|) \geq n \sum_{j=1}^q p_{\lambda_1}(|\gamma_{0j}|) + n \sum_{j=1}^p p_{\lambda_2}(|\beta_{0j}|).$$

Combining the two displayed equations above, we get uniformly for $\|\boldsymbol{\xi} - \boldsymbol{\xi}_0\| = L/\sqrt{n}$, for L sufficiently large,

$$\sum_{i=1}^n \rho_\tau\{Y_i - \mathbf{B}^\top (\mathbf{X}_i^\top \boldsymbol{\gamma}) \widehat{\boldsymbol{\delta}} - \mathbf{Z}_i^\top \boldsymbol{\beta}\} + n \sum_{j=1}^q p_{\lambda_1}(|\gamma_j|) + n \sum_{j=1}^p p_{\lambda_2}(|\beta_j|) > \sum_{i=1}^n \rho_\tau\{Y_i - \mathbf{B}^\top (\mathbf{X}_i^\top \boldsymbol{\gamma}_0) \widehat{\boldsymbol{\delta}} - \mathbf{Z}_i^\top \boldsymbol{\beta}\} + n \sum_{j=1}^q p_{\lambda_1}(|\gamma_{0j}|) + n \sum_{j=1}^p p_{\lambda_2}(|\beta_{0j}|).$$

This implies there is a root-n consistent local minimizer of (2).

Next, we show the root-n consistent local minimizer in the proof above, denoted by $(\widetilde{\boldsymbol{\gamma}}, \widetilde{\boldsymbol{\beta}})$, satisfies part (ii) of Theorem 3. By way of contradiction, suppose $\widetilde{\beta}_{j^*} \neq 0$ for some $j^* > p_1$. Let $\widetilde{\boldsymbol{\beta}}^*$ be the same as $\widetilde{\boldsymbol{\beta}}$ except that we replace $\widetilde{\beta}_{j^*}$ by $\widetilde{\beta}^*_{j^*} = 0$. Using the convexity of the check loss function, we have $\rho_\tau(x) - \rho_\tau(y) \geq \{\tau - \mathbf{1}(y \leq 0)\}(x - y)$, which leads to

$$\sum_{i=1}^n \rho_\tau\{Y_i - \mathbf{B}^\top (\mathbf{X}_i^\top \widetilde{\boldsymbol{\gamma}}) \widehat{\boldsymbol{\delta}} - \mathbf{Z}_i^\top \widetilde{\boldsymbol{\beta}}\} - \sum_{i=1}^n \rho_\tau\{Y_i - \mathbf{B}^\top (\mathbf{X}_i^\top \widetilde{\boldsymbol{\gamma}}) \widehat{\boldsymbol{\delta}} - \mathbf{Z}_i^\top \widetilde{\boldsymbol{\beta}}^*\}$$

$$\begin{aligned} &\geq -\sum_{i=1}^{n}[\tau - \mathbf{1}\{Y_i \leq \mathbf{B}^\top(\mathbf{X}_i^\top\widetilde{\boldsymbol{\gamma}})\widehat{\boldsymbol{\delta}} + \mathbf{Z}_i^\top\widetilde{\boldsymbol{\beta}}^*\}]Z_{ij^*}\widetilde{\beta}_{j^*} \\ &= -\sum_{i=1}^{n}\{\tau - \mathbf{1}(e_i \leq 0)\}Z_{ij^*}\widetilde{\beta}_{j^*} - \sum_{i=1}^{n}[\mathbf{1}(e_i \leq 0) - \mathbf{1}\{e_i \leq \mathbf{B}^\top(\mathbf{X}_i^\top\widetilde{\boldsymbol{\gamma}})\widehat{\boldsymbol{\delta}} + \mathbf{Z}_i^\top\widetilde{\boldsymbol{\beta}}^* - m_i\}]Z_{ij^*}\widetilde{\beta}_{j^*}. \end{aligned} \tag{A.2}$$

For the first term above, we easily find that

$$\sum_{i=1}^{n}\{\tau - \mathbf{1}(e_i \leq 0)\}Z_{ij^*}\widetilde{\beta}_{j^*} = O_p(\sqrt{n})|\widetilde{\beta}_{j^*}|.$$

For the second term above, we have, using $|\mathbf{B}^\top(\mathbf{X}_i^\top\widetilde{\boldsymbol{\gamma}})\widehat{\boldsymbol{\delta}} + \mathbf{Z}_i^\top\widetilde{\boldsymbol{\beta}}^* - m_i| = O_p(\sqrt{K}r_n)$,

$$\begin{aligned} &\mathrm{E}\left[\left[\sum_{i=1}^{n}[\mathbf{1}(e_i \leq 0) - \mathbf{1}\{e_i \leq \mathbf{B}^\top(\mathbf{X}_i^\top\widetilde{\boldsymbol{\gamma}})\widehat{\boldsymbol{\delta}} + \mathbf{Z}_i^\top\widetilde{\boldsymbol{\beta}}^* - m_i\}]Z_{ij^*}\widetilde{\beta}_{j^*}\right]^2\right] \\ &\quad\leq \mathrm{E}\left[\left(\sum_{i=1}^{n}|\mathbf{1}(e_i \leq C\sqrt{K}r_n) - \mathbf{1}(e_i \leq -C\sqrt{K}r_n)| \times |Z_{ij^*}\widetilde{\beta}_{j^*}|\right)^2\right] \\ &\quad= \mathrm{E}\left[\sum_{i=1}^{n}\mathbf{1}(-C\sqrt{K}r_n \leq e_i \leq C\sqrt{K}r_n) \times |Z_{ij^*}\widetilde{\beta}_{j^*}|^2\right] \\ &\qquad+ \sum_{i\neq i'}\mathrm{E}\left[\mathbf{1}(-C\sqrt{K}r_n \leq e_i \leq C\sqrt{K}r_n)\mathbf{1}(-C\sqrt{K}r_n \leq e_{i'} \leq C\sqrt{K}r_n)|Z_{ij^*}Z_{i'j^*}| \times |\widetilde{\beta}_{j^*}|^2\right] \\ &\quad\leq C(n\sqrt{K}r_n + n^2Kr_n^2)\widetilde{\beta}_{j^*}^2, \end{aligned}$$

and thus the second term of (A.2) is $O_p(n\sqrt{K}r_n)|\widetilde{\beta}_{j^*}|$. Furthermore, the difference of the penalty terms is

$$n\sum_{j=1}^{q}p_{\lambda_1}(|\widetilde{\gamma}_j|) + n\sum_{j=1}^{p}p_{\lambda_2}(|\widetilde{\beta}_j|) - n\sum_{j=1}^{q}p_{\lambda_1}(|\widetilde{\gamma}_j|) - n\sum_{j=1}^{p}p_{\lambda_2}(|\widetilde{\beta}_j^*|) = np_{\lambda_2}(|\widetilde{\beta}_{j^*}|) = n\lambda_2|\widetilde{\beta}_{j^*}|, \tag{A.3}$$

where the last equality is due to that $p_\lambda(|x|) = \lambda|x|$ when $|x| < \lambda$. Combining the bound for (A.2) and (A.3), we have

$$\begin{aligned} &\sum_{i=1}^{n}\rho_\tau\{Y_i - \mathbf{B}^\top(\mathbf{X}_i^\top\widetilde{\boldsymbol{\gamma}})\widehat{\boldsymbol{\delta}} - \mathbf{Z}_i^\top\widetilde{\boldsymbol{\beta}}\} + n\sum_{j=1}^{q}p_{\lambda_1}(|\widetilde{\gamma}_j|) + n\sum_{j=1}^{p}p_{\lambda_2}(|\widetilde{\beta}_j|) \\ &\qquad - \sum_{i=1}^{n}\rho_\tau\{Y_i - \mathbf{B}^\top(\mathbf{X}_i^\top\widetilde{\boldsymbol{\gamma}})\widehat{\boldsymbol{\delta}} - \mathbf{Z}_i^\top\widetilde{\boldsymbol{\beta}}^*\} - n\sum_{j=1}^{q}p_{\lambda_1}(|\widetilde{\gamma}_j|) - n\sum_{j=1}^{p}p_{\lambda_2}(|\widetilde{\beta}_j^*|) > 0 \end{aligned}$$

with probability approaching 1, if $\sqrt{K}r_n = o(\lambda_2)$. This leads to a contradiction. Similarly we can show that $\widetilde{\gamma}_j = 0$ when $j > q_1$.

Finally, to show part (i) of the theorem, we only need to note that given part (ii) and within a root-n neighborhood, the penalty

$$n\sum_{j=1}^{q_1}p_{\lambda_1}(|\widetilde{\gamma}_j|) + n\sum_{j=1}^{p_1}p_{\lambda_2}(|\widetilde{\beta}_j|)$$

remains a constant, and thus local minimizer of

$$\sum_{i=1}^{n}\rho_\tau\{Y_i - \mathbf{B}^\top(\mathbf{X}_i^\top\boldsymbol{\gamma})\widehat{\boldsymbol{\delta}} - \mathbf{Z}_i^\top\boldsymbol{\beta}\}$$

without penalty is also a local minimizer of the objective function with penalty. Then the asymptotic normality directly follows from Theorem 2. □

Proof of Theorem 4. The arguments are similar to that used in the proof of Theorem 3 with some modifications, dealing with approximated loss function as well as approximated penalties. We have

We first show that results similar to Lemmas 5 and 6 hold with the approximated loss function.

$$\begin{aligned}
&\rho_\tau\{Y_i - \mathbf{B}^\top(\mathbf{X}_i^\top\widehat{\boldsymbol{\gamma}})\widehat{\boldsymbol{\delta}} - \mathbf{B}^{(1)\top}(\mathbf{X}_i^\top\widehat{\boldsymbol{\gamma}})\widehat{\boldsymbol{\delta}}\mathbf{X}_i^\top(\boldsymbol{\gamma}-\widehat{\boldsymbol{\gamma}}) - \mathbf{Z}_i^\top\boldsymbol{\beta}\} \\
&\qquad = \rho_\tau\{e_i + m_i - \mathbf{B}^\top(\mathbf{X}_i^\top\widehat{\boldsymbol{\gamma}})\widehat{\boldsymbol{\delta}} - \mathbf{B}^{(1)\top}(\mathbf{X}_i^\top\widehat{\boldsymbol{\gamma}})\widehat{\boldsymbol{\delta}}\mathbf{X}_i^\top(\boldsymbol{\gamma}-\widehat{\boldsymbol{\gamma}}) - \mathbf{Z}_i^\top\boldsymbol{\beta}\} \\
&\qquad = \rho_\tau\{e_i - \mathbf{B}^\top(\mathbf{X}_i^\top\boldsymbol{\gamma}_0)(\widehat{\boldsymbol{\delta}}-\boldsymbol{\delta}_0) - \mathbf{B}^{(1)\top}(\mathbf{X}_i^\top\boldsymbol{\gamma}_0)\boldsymbol{\delta}_0\mathbf{X}_i^\top(\boldsymbol{\gamma}-\boldsymbol{\gamma}_0) - \mathbf{Z}_i^\top(\boldsymbol{\beta}-\boldsymbol{\beta}_0) - \widetilde{R}_i(\boldsymbol{\gamma})\} \\
&\qquad = \rho_\tau\{e_i - \boldsymbol{\Pi}_i^\top(\widehat{\boldsymbol{\delta}}-\boldsymbol{\delta}_0) - \mathbf{U}_i^\top(\boldsymbol{\gamma}-\boldsymbol{\gamma}_0) - \mathbf{Z}_i^\top(\boldsymbol{\beta}-\boldsymbol{\beta}_0) - \widetilde{R}_i(\boldsymbol{\gamma})\},
\end{aligned}$$

where $\mathbf{U}_i = \mathbf{B}^{(1)\top}(\mathbf{X}_i^\top\boldsymbol{\gamma}_0)\boldsymbol{\delta}_0\mathbf{X}_i$ as before and

$$\begin{aligned}
\widetilde{R}_i(\boldsymbol{\gamma}) = \{\mathbf{B}(\mathbf{X}_i^\top\widehat{\boldsymbol{\gamma}}) - \mathbf{B}(\mathbf{X}_i^\top\boldsymbol{\gamma}_0)\}^\top\widehat{\boldsymbol{\delta}} + \{\mathbf{B}^{(1)\top}(\mathbf{X}_i^\top\widehat{\boldsymbol{\gamma}})\widehat{\boldsymbol{\delta}} - \mathbf{B}^{(1)\top}(\mathbf{X}_i^\top\boldsymbol{\gamma}_0)\boldsymbol{\delta}_0\}\mathbf{X}_i^\top(\boldsymbol{\gamma}-\widehat{\boldsymbol{\gamma}}) \\
- \mathbf{B}^{(1)\top}(\mathbf{X}_i^\top\boldsymbol{\gamma}_0)\boldsymbol{\delta}_0\mathbf{X}_i^\top(\widehat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}_0) + (\mathbf{B}^\top(\mathbf{X}_i^\top\boldsymbol{\gamma}_0)\boldsymbol{\delta}_0 - g_i) \equiv \widetilde{R}_{i1}(\boldsymbol{\gamma}) + \widetilde{R}_{i2},
\end{aligned}$$

where $\widetilde{R}_{i2} = \mathbf{B}^\top(\mathbf{X}_i^\top\boldsymbol{\gamma}_0)\boldsymbol{\delta}_0 - g_i$ (this is the same as what we have previously denoted $R_{i2}(\boldsymbol{\gamma},\boldsymbol{\delta}) = R_i(\boldsymbol{\gamma}_0,\boldsymbol{\delta})$) and $\widetilde{R}_{i1}(\boldsymbol{\gamma})$ contains the remaining three terms above. We have

$$\begin{aligned}
\widetilde{R}_{i1}(\boldsymbol{\gamma}) &= \{\mathbf{B}(\mathbf{X}_i^\top\widehat{\boldsymbol{\gamma}}) - \mathbf{B}(\mathbf{X}_i^\top\boldsymbol{\gamma}_0)\}^\top(\widehat{\boldsymbol{\delta}}-\boldsymbol{\delta}_0) + \{\mathbf{B}^{(1)\top}(\mathbf{X}_i^\top\widehat{\boldsymbol{\gamma}})\widehat{\boldsymbol{\delta}} - \mathbf{B}^{(1)\top}(\mathbf{X}_i^\top\boldsymbol{\gamma}_0)\boldsymbol{\delta}_0\}\mathbf{X}_i^\top(\boldsymbol{\gamma}-\widehat{\boldsymbol{\gamma}}) \\
&\quad + \{\mathbf{B}^\top(\mathbf{X}_i^\top\widehat{\boldsymbol{\gamma}})\boldsymbol{\delta}_0 - \mathbf{B}^\top(\mathbf{X}_i^\top\boldsymbol{\gamma}_0)\boldsymbol{\delta}_0 - \mathbf{B}^{(1)\top}(\mathbf{X}_i^\top\boldsymbol{\gamma}_0)\boldsymbol{\delta}_0\mathbf{X}_i^\top(\widehat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}_0)\} \\
&= \{\mathbf{B}(\mathbf{X}_i^\top\widehat{\boldsymbol{\gamma}}) - \mathbf{B}(\mathbf{X}_i^\top\boldsymbol{\gamma}_0)\}^\top(\widehat{\boldsymbol{\delta}}-\boldsymbol{\delta}_0) + \{\mathbf{B}^{(1)\top}(\mathbf{X}_i^\top\widehat{\boldsymbol{\gamma}})\widehat{\boldsymbol{\delta}} - \mathbf{B}^{(1)\top}(\mathbf{X}_i^\top\boldsymbol{\gamma}_0)\boldsymbol{\delta}_0)\mathbf{X}_i^\top(\boldsymbol{\gamma}-\widehat{\boldsymbol{\gamma}}) \\
&\quad + \mathbf{B}^{(2)T}(\mathbf{X}_i^\top\boldsymbol{\gamma}^*)\boldsymbol{\delta}_0(\mathbf{X}_i^\top(\widehat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}_0)\}^2.
\end{aligned}$$

Comparing $\widetilde{R}_{i1}(\boldsymbol{\gamma})$ with $R_{i1}(\boldsymbol{\gamma},\boldsymbol{\delta})$ used in the proof of Theorem 1 and Lemma 5, although the two expressions are different, it is easy to see that we still have $|\widetilde{R}_{i1}| \le C\sqrt{K^3/nr_n}$ and $\sum_{i=1}^n \widetilde{R}_{i1}^2 = O_p(r_n^2K^2)$, which are the only bounds required in the proof of Lemma 5. Thus the latter holds with $R_i(\boldsymbol{\gamma},\boldsymbol{\delta})$ replaced by $\widetilde{R}_i(\boldsymbol{\gamma})$. That is, we have

$$\begin{aligned}
\sup_{\|\boldsymbol{\xi}-\boldsymbol{\xi}_0\|\le C/\sqrt{n}} \Bigg| &\sum_{i=1}^n \rho_\tau\{e_i - \boldsymbol{\Pi}_i^\top\widehat{\boldsymbol{\eta}} - \mathbf{Q}_i^\top(\boldsymbol{\xi}-\boldsymbol{\xi}_0) - \widetilde{R}_i(\boldsymbol{\gamma})\} - \sum_{i=1}^n \rho_\tau\{e_i - \boldsymbol{\Pi}_i^\top\widehat{\boldsymbol{\eta}} - \widetilde{R}_i(\boldsymbol{\gamma}_0)\} + \sum_{i=1}^n \mathbf{Q}_i^\top(\boldsymbol{\xi}-\boldsymbol{\xi}_0)e_i \\
&- \mathrm{E}\sum_{i=1}^n \rho_\tau\{e_i - \boldsymbol{\Pi}_i^\top\widehat{\boldsymbol{\eta}} - \mathbf{Q}_i^\top(\boldsymbol{\xi}-\boldsymbol{\xi}_0) - \widetilde{R}_i(\boldsymbol{\gamma})\} + \mathrm{E}\sum_{i=1}^n \rho_\tau\{e_i - \boldsymbol{\Pi}_i^\top\widehat{\boldsymbol{\eta}} - \widetilde{R}_i(\boldsymbol{\gamma}_0)\}\Bigg| = o_p(1),
\end{aligned}$$

where $\widehat{\boldsymbol{\eta}} = \widehat{\boldsymbol{\delta}} - \boldsymbol{\delta}_0 + (\boldsymbol{\Pi}^\top\boldsymbol{\beta}\boldsymbol{\Pi})^{-1}\boldsymbol{\Pi}^\top\boldsymbol{\beta}\mathbf{Q}(\boldsymbol{\xi}-\boldsymbol{\xi}_0)$. It is also straightforward to see that a result similar to Lemma 6 holds, following exactly the same arguments:

$$\begin{aligned}
\sup_{\|\boldsymbol{\xi}-\boldsymbol{\xi}_0\|\le C/\sqrt{n}} \Bigg| &\sum_{i=1}^n \mathrm{E}\rho_\tau\{e_i - \boldsymbol{\Pi}_i^\top\widehat{\boldsymbol{\eta}} - \mathbf{Q}_i(\boldsymbol{\xi}-\boldsymbol{\xi}_0) - \widetilde{R}_i(\boldsymbol{\gamma})\} \\
&- \sum_{i=1}^n \mathrm{E}\rho_\tau\{e_i - \boldsymbol{\Pi}_i^\top\widehat{\boldsymbol{\eta}} - \widetilde{R}_i(\boldsymbol{\gamma}_0)\} - \sum_{i=1}^n \frac{f(0 \mid \mathbf{X}_i, \mathbf{Z}_i)}{2}(\boldsymbol{\xi}-\boldsymbol{\xi}_0)\mathbf{Q}_i\mathbf{Q}_i^\top(\boldsymbol{\xi}-\boldsymbol{\xi}_0)\Bigg| = o_p(1).
\end{aligned}$$

Thus as in the proof of Theorem 3, we see that for $\|\boldsymbol{\xi}-\boldsymbol{\xi}_0\| = L/\sqrt{n}$ with $L > 0$ sufficiently large,

$$\sum_{i=1}^n \rho_\tau\{e_i - \boldsymbol{\Pi}_i^\top\widehat{\boldsymbol{\eta}} - \mathbf{Q}_i^\top(\boldsymbol{\xi}-\boldsymbol{\xi}_0) - \widetilde{R}_i(\boldsymbol{\gamma})\} - \sum_{i=1}^n \rho_\tau\{e_i - \boldsymbol{\Pi}_i^\top\widehat{\boldsymbol{\eta}} - \widetilde{R}_i(\boldsymbol{\gamma}_0)\}$$

is bounded away from zero with probability approaching 1.

For the penalty terms in the proof of Theorem 3, we had

$$n\sum_{j=1}^q p_{\lambda_1}(|\gamma_j|) + n\sum_{j=1}^p p_{\lambda_2}(|\beta_j|) \ge n\sum_{j=1}^q p_{\lambda_1}(|\gamma_{0j}|) + n\sum_{j=1}^p p_{\lambda_2}(|\beta_{0j}|).$$

Here, when $\lambda_1, \lambda_2 = o(1)$, we still have

$$n\sum_{j=1}^q p'_{\lambda_1}(|\widehat{\gamma}_j|)|\gamma_j| + n\sum_{j=1}^p p'_{\lambda_2}(|\widehat{\beta}_j|)|\beta_j| \ge n\sum_{j=1}^q p'_{\lambda_1}(|\widehat{\gamma}_j|)|\gamma_{0j}| + n\sum_{j=1}^p p'_{\lambda_2}(|\widehat{\beta}_{0j}|)|\beta_{0j}|,$$

since $p'_{\lambda_1}(|\widehat{\gamma}_j|) = 0$ for $j \le q_1$, $p'_{\lambda_2}(|\widehat{\beta}_j|) = 0$ for $j \le p_1$, while $|\gamma_{0j}| = 0$ for $j > q_1$, $|\beta_{0j}| = 0$ for $j > p_1$. These results imply the minimizer of (3) is root-n consistent.

Next, we show the root-n consistent minimizer $(\widetilde{\boldsymbol{\gamma}}, \widetilde{\boldsymbol{\beta}})$ satisfies part (ii) of Theorem 4. By way of contradiction, suppose $\widetilde{\beta}_{j^*} \neq 0$ for some $j^* > p_1$. Let $\widetilde{\boldsymbol{\beta}}^*$ be the same as $\widetilde{\boldsymbol{\beta}}$ except that we replace $\widetilde{\beta}_{j^*}$ by $\widetilde{\beta}^*_{j^*} = 0$. Using the convexity of the check loss function, we have $\rho_\tau(x) - \rho_\tau(y) \ge \{\tau - \mathbf{1}(y \le 0)\}(x - y)$, which leads to

$$\sum_{i=1}^n \rho_\tau\{Y_i - \mathbf{B}^\top(\mathbf{X}_i^\top\widehat{\boldsymbol{\gamma}})\widehat{\boldsymbol{\delta}} - \mathbf{B}^{(1)\top}(\mathbf{X}_i^\top\widehat{\boldsymbol{\gamma}})\widehat{\boldsymbol{\delta}}\mathbf{X}_i^\top(\widetilde{\boldsymbol{\gamma}} - \widehat{\boldsymbol{\gamma}}) - \mathbf{Z}_i^\top\widetilde{\boldsymbol{\beta}}\} - \sum_{i=1}^n \rho_\tau\{Y_i - \mathbf{B}^\top(\mathbf{X}_i^\top\widehat{\boldsymbol{\gamma}})\widehat{\boldsymbol{\delta}} - \mathbf{B}^{(1)\top}(\mathbf{X}_i^\top\widehat{\boldsymbol{\gamma}})\widehat{\boldsymbol{\delta}}\mathbf{X}_i^\top(\widetilde{\boldsymbol{\gamma}} - \widehat{\boldsymbol{\gamma}}) - \mathbf{Z}_i^\top\widetilde{\boldsymbol{\beta}}^*\}$$
$$\ge -\sum_{i=1}^n [\tau - \mathbf{1}\{Y_i \le \mathbf{B}^\top(\mathbf{X}_i^\top\widehat{\boldsymbol{\gamma}})\widehat{\boldsymbol{\delta}} + \mathbf{B}^{(1)\top}(\mathbf{X}_i^\top\widehat{\boldsymbol{\gamma}})\widehat{\boldsymbol{\delta}}\mathbf{X}_i^\top(\widetilde{\boldsymbol{\gamma}} - \widehat{\boldsymbol{\gamma}}) + \mathbf{Z}_i^\top\widetilde{\boldsymbol{\beta}}^*\}]Z_{ij^*}\widetilde{\beta}_{j^*},$$

and the right-hand side can be rewritten as

$$-\sum_{i=1}^n \{\tau - \mathbf{1}(e_i \le 0)\}Z_{ij^*}\widetilde{\beta}_{j^*} - \sum_{i=1}^n [\mathbf{1}(e_i \le 0) - \mathbf{1}\{e_i \le \mathbf{B}^\top(\mathbf{X}_i^\top\widehat{\boldsymbol{\gamma}})\widehat{\boldsymbol{\delta}} + \mathbf{B}^{(1)\top}(\mathbf{X}_i^\top\widehat{\boldsymbol{\gamma}})\widehat{\boldsymbol{\delta}}\mathbf{X}_i^\top(\widetilde{\boldsymbol{\gamma}} - \widehat{\boldsymbol{\gamma}}) + \mathbf{Z}_i^\top\widetilde{\boldsymbol{\beta}}^* - m_i\}]Z_{ij^*}\widetilde{\beta}_{j^*}.$$

The rest of the proof follows exactly the same lines of the proof of Theorem 3, except that (A.3) is replaced by

$$n\sum_{j=1}^q p'_{\lambda_1}(|\widehat{\gamma}_j|)|\widetilde{\gamma}_j| + n\sum_{j=1}^p p'_{\lambda_2}(|\widehat{\beta}_j|)|\widetilde{\beta}_j| - n\sum_{j=1}^q p'_{\lambda_1}(|\widehat{\gamma}_j|)|\widetilde{\gamma}_j| - n\sum_{j=1}^p p'_{\lambda_2}(|\widehat{\beta}_j|)|\widetilde{\beta}^*_j| = np'_{\lambda_2}(|\widehat{\beta}_{j^*}|)|\widetilde{\beta}_{j^*}| = n\lambda_2|\widetilde{\beta}_{j^*}|.$$

This completes the proof of Theorem 4. □

Proof of Theorem 5. We consider the approximated penalized problem (3), and (2) can also be dealt with similarly. Let S_0 be the set of indices of the nonzero components of $(\boldsymbol{\gamma}_0, \boldsymbol{\beta}_0)$ in the true model. Let S be the set of indices of the nonzero components of $(\widetilde{\boldsymbol{\gamma}}, \widetilde{\boldsymbol{\beta}})$ using λ_1, λ_2 selected by SIC. As usual, the proof is split into two cases. First we introduce the following notations. Let

$$\Omega_n(\boldsymbol{\gamma}, \boldsymbol{\beta}) = \sum_{i=1}^n \rho_\tau\{Y_i - \mathbf{B}^\top(\mathbf{X}_i^\top\widehat{\boldsymbol{\gamma}})\widehat{\boldsymbol{\delta}} - \mathbf{B}^{(1)\top}(\mathbf{X}_i^\top\widehat{\boldsymbol{\gamma}})\widehat{\boldsymbol{\delta}}\mathbf{X}_i^\top(\boldsymbol{\gamma} - \widehat{\boldsymbol{\gamma}}) - \mathbf{Z}_i^\top\boldsymbol{\beta}\}.$$

Define $\Omega_{nS}(\boldsymbol{\gamma}, \boldsymbol{\beta})$ similarly with $\mathbf{X}_i$, $\mathbf{Z}_i$ replaced by their subvectors containing only components in S. In the following, such subvectors of $\mathbf{X}_i$, $\mathbf{Z}_i$ are still denoted by $\mathbf{X}_i$, $\mathbf{Z}_i$ for simplicity of notation. Let $(\widetilde{\boldsymbol{\gamma}}_S, \widetilde{\boldsymbol{\beta}}_S)$ be the minimizer of $\Omega_{nS}(\boldsymbol{\gamma}, \boldsymbol{\beta})$, $(\boldsymbol{\gamma}^*_S, \boldsymbol{\beta}^*_S)$ be the minimizer of $\mathrm{E}\Omega_{nS}(\boldsymbol{\gamma}, \boldsymbol{\beta})$.

Case 1. $S_0 \nsubseteq S$ (underfitted). Obviously, by definition, we have $\Omega_n(\widetilde{\boldsymbol{\gamma}}, \widetilde{\boldsymbol{\beta}}) \ge \Omega_{nS}(\widetilde{\boldsymbol{\gamma}}_S, \widetilde{\boldsymbol{\beta}}_S)$. Proceeding as in the proof of our theorems above, it can be shown that $(\widetilde{\boldsymbol{\gamma}}_S, \widetilde{\boldsymbol{\beta}}_S)$ is a root-n consistent estimator of $(\boldsymbol{\gamma}^*_S, \boldsymbol{\beta}^*_S)$. Similar to the proof of Lemma 5, we have

$$|\Omega_{nS}(\widetilde{\boldsymbol{\gamma}}_S, \widetilde{\boldsymbol{\beta}}_S) - \Omega_{nS}(\boldsymbol{\gamma}^*_S, \boldsymbol{\beta}^*_S) - \mathrm{E}\Omega_{nS}(\widetilde{\boldsymbol{\gamma}}_S, \widetilde{\boldsymbol{\beta}}_S) + \mathrm{E}\Omega_{nS}(\boldsymbol{\gamma}^*_S, \boldsymbol{\beta}^*_S)| = O_p(1).$$

The above is $O_p(1)$ instead of $o_p(1)$ since we now do not need the linear term $\boldsymbol{\xi} - \boldsymbol{\xi}_0$ as in Lemma 5. Similar to Lemma 6, we have

$$|\mathrm{E}\Omega_{nS}(\widetilde{\boldsymbol{\gamma}}_S, \widetilde{\boldsymbol{\beta}}_S) - \mathrm{E}\Omega_{nS}(\boldsymbol{\gamma}^*_S, \boldsymbol{\beta}^*_S)| = O_p(1).$$

Furthermore, by the Law of Large Numbers, we can also show that $|\Omega_{nS}(\boldsymbol{\gamma}^*_S, \boldsymbol{\beta}^*_S) - \mathrm{E}\Omega_{nS}(\boldsymbol{\gamma}^*_S, \boldsymbol{\beta}^*_S)| = o_p(n)$. Thus

$$\Omega_{nS}(\widetilde{\boldsymbol{\gamma}}_S, \widetilde{\boldsymbol{\beta}}_S) = \mathrm{E}\Omega_{nS}(\boldsymbol{\gamma}^*_S, \boldsymbol{\beta}^*_S) + o_p(n). \tag{A.4}$$

Similar bounds hold when S is replaced by S_0 above, i.e.,

$$\Omega_{nS_0}(\widetilde{\boldsymbol{\gamma}}_{S_0}, \widetilde{\boldsymbol{\beta}}_{S_0}) = \mathrm{E}\Omega_{nS_0}(\boldsymbol{\gamma}^*_{S_0}, \boldsymbol{\beta}^*_{S_0}) + o_p(n). \tag{A.5}$$

Next, comparing $\mathrm{E}\Omega_{nS}(\boldsymbol{\gamma}^*_S, \boldsymbol{\beta}^*_S)$ with $\mathrm{E}\Omega_{nS}(\boldsymbol{\gamma}^*_{S_0}, \boldsymbol{\beta}^*_{S_0})$, we have

$$\frac{1}{n}\mathrm{E}\Omega_{nS}(\boldsymbol{\gamma}_S^*,\boldsymbol{\beta}_S^*)-\frac{1}{n}\mathrm{E}\Omega_{nS}(\boldsymbol{\gamma}_{S_0}^*,\boldsymbol{\beta}_{S_0}^*)$$
$$=\frac{1}{n}\sum_{i=1}^{n}\int_{\mathbf{B}^\top(\mathbf{X}_i^\top\widehat{\boldsymbol{\gamma}})\widehat{\boldsymbol{\delta}}+\mathbf{B}^{(1)\top}(\mathbf{X}_i^\top\widehat{\boldsymbol{\gamma}})\widehat{\boldsymbol{\delta}}\mathbf{X}_i^\top(\boldsymbol{\gamma}_S^*-\widehat{\boldsymbol{\gamma}})+\mathbf{Z}_i^\top\boldsymbol{\beta}_S^*}^{\mathbf{B}^\top(\mathbf{X}_i^\top\widehat{\boldsymbol{\gamma}})\widehat{\boldsymbol{\delta}}+\mathbf{B}^{(1)\top}(\mathbf{X}_i^\top\widehat{\boldsymbol{\gamma}})\widehat{\boldsymbol{\delta}}\mathbf{X}_i^\top(\boldsymbol{\gamma}_{S_0}^*-\widehat{\boldsymbol{\gamma}})+\mathbf{Z}_i^\top\boldsymbol{\beta}_{S_0}^*}F(t\mid\mathbf{X}_i,\mathbf{Z}_i)-f(0\mid\mathbf{X}_i,\mathbf{Z}_i)dt$$
$$=\frac{1}{n}\sum_{i=1}^{n}\frac{f(0\mid\mathbf{X}_i,\mathbf{Z}_i)}{2}[\mathbf{B}^{(1)\top}(\mathbf{X}_i^\top\widehat{\boldsymbol{\gamma}})\widehat{\boldsymbol{\delta}}\mathbf{X}_i^\top(\boldsymbol{\gamma}_S^*-\boldsymbol{\gamma}_{S_0}^*)+\mathbf{Z}_i^\top(\boldsymbol{\beta}_S^*-\boldsymbol{\beta}_{S_0}^*)]^2\{1+o_p(1)\}.$$

Using Lemma 3, the above is bounded away from zero. Combining this fact with (A.4) and (A.5), we see that $\Omega_{nS}(\widetilde{\boldsymbol{\gamma}}_S,\widetilde{\boldsymbol{\beta}}_S)/n-\Omega_{nS_0}(\widetilde{\boldsymbol{\gamma}}_{S_0},\widetilde{\boldsymbol{\beta}}_{S_0})/n$ is bounded away from zero. Let $(\lambda_{01},\lambda_{02})$ be the theoretically optimal (satisfying the conditions of Theorem 3) tuning parameters with the corresponding estimator $(\widetilde{\boldsymbol{\gamma}}_{S_0},\widetilde{\boldsymbol{\beta}}_{S_0})$. Finally, we get

$$\mathrm{SIC}(\lambda_1,\lambda_2)-\mathrm{SIC}(\lambda_{01},\lambda_{02})\geq\ln\Omega_{nS}(\widetilde{\boldsymbol{\gamma}}_S,\widetilde{\boldsymbol{\beta}}_S)-\ln\Omega_{nS_0}(\widetilde{\boldsymbol{\gamma}}_{S_0},\widetilde{\boldsymbol{\beta}}_{S_0})+O_p(\ln n/n)$$
$$=\ln\left\{1+\frac{\Omega_{nS}(\widetilde{\boldsymbol{\gamma}}_S,\widetilde{\boldsymbol{\beta}}_S)/n-\Omega_{nS_0}(\widetilde{\boldsymbol{\gamma}}_{S_0},\widetilde{\boldsymbol{\beta}}_{S_0})/n}{\Omega_{nS_0}(\widetilde{\boldsymbol{\gamma}}_{S_0},\widetilde{\boldsymbol{\beta}}_{S_0})/n}\right\}+O_p(\ln n/n).$$

Case 2. $S_0\subsetneq S$ (overfitted). In this case, both $(\widetilde{\boldsymbol{\gamma}}_S,\widetilde{\boldsymbol{\beta}}_S)$ and $(\widetilde{\boldsymbol{\gamma}}_{S_0},\widetilde{\boldsymbol{\beta}}_{S_0})$ are root-n consistent estimators of $(\boldsymbol{\gamma}_0,\boldsymbol{\beta}_0)$. Using results similar to Lemma 6 again, we get $\Omega_{nS}(\widetilde{\boldsymbol{\gamma}}_S,\widetilde{\boldsymbol{\beta}}_S)/n-\Omega_{nS_0}(\widetilde{\boldsymbol{\gamma}}_{S_0},\widetilde{\boldsymbol{\beta}}_{S_0})/n=O_p(1/n)$. Thus

$$\mathrm{SIC}(\lambda_1,\lambda_2)-\mathrm{SIC}(\lambda_{01},\lambda_{02})\geq\ln\left\{1+\frac{\Omega_{nS}(\widetilde{\boldsymbol{\gamma}}_S,\widetilde{\boldsymbol{\beta}}_S)/n-\Omega_{nS_0}(\widetilde{\boldsymbol{\gamma}}_{S_0},\widetilde{\boldsymbol{\beta}}_{S_0})/n}{\Omega_{nS_0}(\widetilde{\boldsymbol{\gamma}}_{S_0},\widetilde{\boldsymbol{\beta}}_{S_0})/n}\right\}+\ln n/(2n)$$
$$=O_p(1/n)+\ln n/(2n)>0,$$

which is again in contradiction to the fact that (λ_1,λ_2) minimizes SIC. □

References

[1] Z. Cai, X. Xu, Nonparametric quantile estimations for dynamic smooth coefficient models, J. Amer. Statist. Assoc. 103 (2008) 1595–1608.
[2] R.J. Carroll, J. Fan, I. Gijbels, M.P. Wand, Generalized partially linear single-index models, J. Amer. Statist. Assoc. 92 (1997) 477–489.
[3] P. Chaudhuri, K. Doksum, A. Samarov, On average derivative quantile regression, Ann. Statist. 25 (1997) 715–744.
[4] S. Chen, S., Khan, Semiparametric estimation of a partially linear censored regression model, Econom. Theory 17 (2001) 567–590.
[5] X. Cui, W.K. Härdle, L. Zhu, The EFM approach for single-index models, Ann. Statist. 39 (2011) 1658–1688.
[6] C. de Boor, A Practical Guide to Splines, Revised Ed., Springer, New York, 2001.
[7] J.G. De Gooijer, D. Zerom, On additive conditional quantiles with high-dimensional covariates, J. Amer. Statist. Assoc. 98 (2003) 135–146.
[8] J. Fan, R. Li, Variable selection via nonconcave penalized likelihood and its oracle properties, J. Amer. Statist. Assoc. 96 (2001) 1348–1360.
[9] J. Fan, W. Zhang, Statistical methods with varying coefficient models, Statist. Interface 1 (2008) 179–195.
[10] D. Harrison, D.L. Rubinfeld, Hedonic prices and the demand for clean air, J. Environ. Econom. Manag. 5 (1978) 81–102.
[11] X. He, H. Liang, Quantile regression estimates for a class of linear and partially linear errors-in-variables models, Statist. Sinica 10 (2000) 129–140.
[12] X. He, P. Shi, Convergence rate of B-spline estimators of nonparametric conditional quantile functions, J. Nonparam. Statist. 3 (1994) 299–308.
[13] X. He, Z.-Y. Zhu, W.-K. Fung, Estimation in a semiparametric model for longitudinal data with unspecified dependence structure, Biometrika 89 (2002) 579–590.
[14] J.L. Horowitz, S. Lee, Nonparametric estimation of an additive quantile regression model, J. Amer. Statist. Assoc. 100 (2005) 1238–1249.
[15] J. Huang, H. Shen, Functional coefficient regression models for nonlinear time series: A polynomial spline approach, Scand. J. Statist. 31 (2004) 515–534.
[16] M. Kim, Quantile regression with varying coefficients, Ann. Statist. 35 (2007) 92–108.
[17] R. Koenker, G. Bassett, Regression quantile, Econometrica 1 (1978) 33–50.
[18] E. Kong, Y. Xia, A single-index quantile regression model and its estimation, Econom. Theory 28 (2012) 730–768.
[19] Q. Li, Efficient estimation of additive partially linear models, Internat. Econom. Rev. 41 (2000) 1073–1092.
[20] H. Lian, A note on the consistency of Schwarz's criterion in linear quantile regression with the SCAD penalty, Statist. Probab. Lett. 82 (2012) 1224–1228.
[21] H. Liang, R.Z. Li, Variable selection for partially linear models with measurement errors, J. Amer. Statist. Assoc. 104 (2009) 234–248.
[22] H. Liang, X. Liu, R. Li, C.L. Tsai, Estimation and testing for partially linear single-index models, Ann. Statist. 38 (2010) 3811–3836.
[23] W. Lin, K.B. Kulasekera, Identifiability of single-index models and additive-index models, Biometrika 94 (2007) 496–501.
[24] S. Ma, X. He, Inference for single-index quantile regression models with profile optimization, Ann. Statist. 44 (2016) 1234–1268.
[25] S. Ma, H. Liang, C.-L. Tsai, Partially linear single index models for repeated measurements, J. Multivariate Anal. 130 (2014) 354–375.
[26] D. Ruppert, W.P. Wand, R.J. Carroll, Semiparametric Regression, Cambridge University Press, 2003.

[27] C.J. Stone, The use of polynomial splines and their tensor products in multivariate function estimation, Ann. Statist. 22 (1994) 118–184.
[28] R.J. Tibshirani, Regression shrinkage and selection via the lasso, J. Roy. Statist. Soc. Ser. B 58 (1996) 267–288.
[29] H.J. Wang, Z. Zhu, J. Zhou, Quantile regression in partially linear varying coefficient models, Ann. Statist. 37 (2009) 3841–3866.
[30] L. Wang, X. Liu, H. Liang, R.J. Carroll, Estimation and variable selection for generalized additive partial linear models, Ann. Statist. 39 (2011) 1827–1851.
[31] Y. Wei, X. He, Conditional growth charts, Ann. Statist. 34 (2006) 2069–2097.
[32] C. Wu, Y. Yu, Partially linear modeling of conditional quantiles using penalized splines, Comput. Statist. Data Anal. 77 (2014) 170–187.
[33] Y. Wu, Y. Liu, Variable selection in quantile regression, Statist. Sinica 19 (2009) 801–817.
[34] Z. Wu, K. Yu, Y. Yu, Single-index quantile regression. J. Multivariate Anal. 101 (2010) 1607–1621.
[35] K. Yu, Z. Lu, Local linear additive quantile regression, Scand. J. Statist. 31 (2004) 333–346.
[36] Y. Yu, D. Ruppert, Penalized spline estimation for partially linear single-index models, J. Amer. Statist. Assoc. 97 (2002) 1042–1054.
[37] Y. Yu, C. Wu, Y. Zhang, Penalized spline estimation for generalized partially linear single-index models, Statist. Comput. 27 (2017) 571–582.
[38] M. Yuan, Y. Lin, Model selection and estimation in regression with grouped variables, J. Roy. Statist. Soc. Ser. B 68 (2006) 49–67.
[39] C. Zhang, Nearly unbiased variable selection under minimax concave penalty, Ann. Statist. 38 (2010) 894–942.
[40] H. Zou, The adaptive lasso and its oracle properties, J. Amer. Statist. Assoc. 101 (2006) 1418–1429.
[41] H. Zou, T. Hastie, Regularization and variable selection via the elastic net, J. Roy. Statist. Soc. Ser. B 67 (2005) 301–320.
[42] H. Zou, R.Z. Li, One-step sparse estimates in nonconcave penalized likelihood models, Ann. Statist. 36 (2008) 1509–1533.