Ann Inst Stat Math (2017) 69:761–789
DOI 10.1007/s10463-016-0558-9

Quantile regression and variable selection of single-index coefficient model

Weihua Zhao[1,2] · Riquan Zhang[1] · Yazhao Lv[1] · Jicai Liu[1,3]

Received: 6 November 2014 / Revised: 17 February 2016 / Published online: 31 March 2016

Abstract In this paper, a minimizing average check loss estimation (MACLE) procedure is proposed for the single-index coefficient model (SICM) in the framework of quantile regression (QR). The resulting estimators have the asymptotic normality and achieve the best convergence rate. Furthermore, a variable selection method is investigated for the QRSICM by combining MACLE method with the adaptive LASSO penalty, and we also established the oracle property of the proposed variable selection method. Extensive simulations are conducted to assess the finite sample performance of the proposed estimation and variable selection procedure under various error settings. Finally, we present a real-data application of the proposed approach.

Keywords Single index coefficient model · Quantile regression · Asymptotic normality · Variable selection · Adaptive LASSO · Oracle property

1 Introduction

Consider the varying-coefficient model (VCM), whose standard form can be written as

The research was supported in part by National Natural Science Foundation of China (11501372, 11571112), Project of National Social Science Fund (15BTJ027), Doctoral Fund of Ministry of Education of China (20130076110004), Program of Shanghai Subject Chief Scientist (14XD1401600) and the 111 Project of China (B14019).

✉ Riquan Zhang
zhangriquan@163.com

[1] School of Statistics, East China Normal University, Shanghai 200241, China

[2] School of Science, NanTong University, Nantong 226007, China

[3] College of Mathematics and Sciences, Shanghai Normal University, Shanghai 200234, China

$$Y = \mathbf{g}(\mathbf{X})^T \mathbf{Z} + \varepsilon, \tag{1}$$

where Y is the response variable, $\mathbf{X} = (X_1, \ldots, X_p)^T \in \mathbb{R}^p$ and $\mathbf{Z} = (Z_0, Z_1, \ldots, Z_{d-1})^T \in \mathbb{R}^d$ are two covariates vectors, ε is the model error, $\mathbf{g}(\cdot) = (g_0(\cdot), g_1(\cdot), \ldots, g_{d-1}(\cdot))^T$ is an unknown coefficient function vector. Without loss generality, we assume $Z_0 \equiv 1$, i.e., the corresponding nonparametric function $g_0(\cdot)$ can be seen as the baseline function.

Though much research has been done on the VCM (1), the voluminous literature mostly focused on the case when the variate $\mathbf{X}$ is scalar. When the dimension of $\mathbf{X}$ is high, how to effectively estimate the multivariate nonparametric $\mathbf{g}(\mathbf{X})$ is challenging in practice because of model (1) still facing the problem of "curse of dimensionality". To this end, Xia et al. (1999) proposed an elegant solution for multivariate $\mathbf{X}$ by introducing a single-index structure for the index vector, resulting in

$$Y = \mathbf{g}(\mathbf{X}^T \boldsymbol{\theta})^T \mathbf{Z} + \varepsilon, \tag{2}$$

with $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)^T \in \mathbb{R}^p$, which was termed the single-index coefficient model (SICM). For identifiability, we assume $\|\boldsymbol{\theta}\| = 1$ and $\theta_1 > 0$. A related model, termed the adaptive varying-coefficient linear models (Fan et al. 2003), has the same form as SICM with $\mathbf{X} = \mathbf{Z}$. For the simplicity, in this paper, we assume that $\mathbf{X} \neq \mathbf{Z}$ in model (2), otherwise we need the additional identifiability condition like in Fan et al. (2003).

On the other hand, SICM can be also viewed as the useful extension of the single-index model proposed by Härdle et al. (1993). Xia et al. (1999) investigated the least-squares cross-validation estimation method for the index parameter $\boldsymbol{\theta}$, and its estimator can achieve the best convergence rate without "undersmoothing" the nonparametric coefficient function. However, the least-squares cross-validation method is computationally expensive and not practical in reality. Lu et al. (2007) established the asymptotic theory of the profile likelihood estimation of SICM, but they did not provide any simulation studies and real data analysis. Recently, Xue and Pang (2013) proposed an estimation method based on estimating equation and obtained the confidence region of the nonparametric coefficient function. Huang and Zhang (2012) derived a confidence interval of the index parameter $\boldsymbol{\theta}$ in SICM by profile empirical likelihood method. Furthermore, Feng and Xue (2013) proposed an estimation procedure of SICM based on spline approximation and further considered the variable selection issue of the parameter and the nonparametric coefficient functions.

However, the estimation methods aforementioned all focused on the mean regression for SICM. It is well known that when the error deviates far from the normal distribution and/or the data include some outliers, the least square-based method or likelihood estimation approach may loss efficiency and lead to incorrect inference. In this case, quantile regression proposed by Koenker and Basset (1978) can be chosen as an alternative approach to investigate the underlying relationship of the response and the multidimensional covariates, and it can provide the full description of the conditional distribution for response variable at different quantile level. There have been some researches on the quantile regression of the two simplified form of SICM, varying-coefficient model (VCM) (see Honda 2004; Kim 2007; Cai and Xu 2008) and single-index model (SIM) (see Wu et al. 2010; Jiang et al. 2012). However, there is no work for the SICM based on the quantile method.

In this paper, an estimation procedure, called as minimizing average check loss estimation(MACLE) method, is proposed for SICM based on the quantile regression framework. We describe the implementation details of the proposed algorithm and establish the theoretical properties of the estimators. In special, the estimator of the index parameter can achieve the best convergence rate without "undersmoothing" the nonparametric coefficient vector function. Meanwhile, we address the variable selection method for quantile regression SICM by combining the MACLE method and the adaptive LASSO method (Zou 2006), and the corresponding oracle properties are also established.

The paper is organized as follows. In Sect. 2, we outline the estimation procedure and the algorithm for the quantile regression of SICM. In Sect. 3, the asymptotic properties of the estimators are established. To select the important index variables, we investigated the variable selection method in Sect. 4, and the corresponding oracle properties are also established. In Sect. 5, we conduct two simulations with different error settings to assess the finite sample performance of our proposed method. We further illustrate the method by the analysis of the Boston Housing data in Sect. 6. The technical proof and the regularity conditions are relegated in the Appendix.

2 Estimation methodology

To apply the quantile method, we assume that the τ-th quantile of ε in model (2) is zero, i.e., $P\{\varepsilon < 0|\mathbf{X} = \mathbf{x}, \mathbf{Z} = \mathbf{z}\} = \tau$. Let $\rho_\tau(u) = u[\tau - I(u < 0)]$ be the check loss function for $\tau \in (0, 1)$. Quantile regression is used to estimate the conditional quantile of the response variable Y, which is defined as

$$q_\tau(\mathbf{x}, \mathbf{z}) = \underset{a}{\operatorname{argmin}}\ \mathrm{E}\left\{\rho_\tau(Y - a)|\mathbf{X} = \mathbf{x}, \mathbf{Z} = \mathbf{z}\right\}.$$

Suppose $\{\mathbf{X}_i,\ \mathbf{Z}_i,\ Y_i\}_{i=1}^n$ is an independent and identically distributed (i.i.d.) samples from (2). Theoretically, the estimate $\hat{\boldsymbol{\theta}}$ satisfies

$$\hat{\boldsymbol{\theta}} = \underset{\|\boldsymbol{\theta}\|=1, \theta_1>0}{\operatorname{argmin}}\ \mathrm{E}[\rho_\tau(Y - \mathbf{g}(\mathbf{X}^T\boldsymbol{\theta})^T\mathbf{Z})]. \tag{3}$$

By the property of conditional expectation, the right side of (3) can be re-expressed as

$$\mathrm{E}\left[\rho_\tau(Y - \mathbf{g}(\mathbf{X}^T\boldsymbol{\theta})^T\mathbf{Z})\right] = \mathrm{E}\left\{\mathrm{E}\left[\rho_\tau(Y - \mathbf{g}(\mathbf{X}^T\boldsymbol{\theta})^T\mathbf{Z})\middle|\mathbf{X}^T\boldsymbol{\theta}\right]\right\}, \tag{4}$$

where $\mathrm{E}[\rho_\tau(Y - \mathbf{g}(\mathbf{X}^T\boldsymbol{\theta})^T\mathbf{Z})|\mathbf{X}^T\boldsymbol{\theta}]$ is the conditional expected check loss function given $\mathbf{X}^T\boldsymbol{\theta}$. In the following, we will construct an empirical form of the theoretical loss (4). By minimizing the empirical loss function, we can derive the estimation of the index parameter.

Given $\boldsymbol{\theta}$, when $\mathbf{X}_i^T\boldsymbol{\theta}$ in the neighborhood of u, for $0 \le j \le d-1$, the jth element of $\mathbf{g}(\mathbf{X}_i^T\boldsymbol{\theta})$, $g_j(\mathbf{X}_i^T\boldsymbol{\theta})$ can be approximated local linearly as

$$g_j\left(\mathbf{X}_i^T\boldsymbol{\theta}\right) \simeq g_j(u) + g_j'(u)\left(\mathbf{X}_i^T\boldsymbol{\theta} - u\right).$$

Then the local linear approximation of $\mathrm{E}[\rho_\tau(Y - \mathbf{g}(\mathbf{X}^T\boldsymbol{\theta})^T\mathbf{Z})|\mathbf{X}^T\boldsymbol{\theta} = u]$ will be

$$\sum_{i=1}^n \rho_\tau\left(Y_i - \left[\mathbf{g}(u) + \mathbf{g}'(u)\left(\mathbf{X}_i^T\boldsymbol{\theta} - u\right)\right]^T\mathbf{Z}_i\right)\omega_{i0},$$

where $\omega_{i0} = K_h(\mathbf{X}_i^T\boldsymbol{\theta} - u)/\sum_{l=1}^n K_h(\mathbf{X}_l^T\boldsymbol{\theta} - u)$ satisfy $\sum_{i=1}^n \omega_{i0} = 1$, $K(\cdot)$ is kernel function, $K_h(\cdot) = K(\cdot/h)/h$, and h is the bandwidth. By averaging on $u_j = \mathbf{X}_j^T\boldsymbol{\theta}$, $j = 1, \ldots, n$, we can get the empirical form of (4) as

$$\frac{1}{n}\sum_{j=1}^n\sum_{i=1}^n \rho_\tau\left(Y_i - \left[\mathbf{g}(u_j) + \mathbf{g}'(u_j)\mathbf{X}_{ij}^T\boldsymbol{\theta}\right]^T\mathbf{Z}_i\right)\omega_{ij}, \tag{5}$$

where $\mathbf{X}_{ij} = \mathbf{X}_i - \mathbf{X}_j$, $\omega_{ij} = K_h(\mathbf{X}_{ij}^T\boldsymbol{\theta})/\sum_{l=1}^n K_h(\mathbf{X}_{lj}^T\boldsymbol{\theta})$ satisfy $\sum_{i=1}^n \omega_{ij} = 1$, $\forall j = 1, \ldots, n$.

Now the parameter $\boldsymbol{\theta}$ can be estimated by

$$\hat{\boldsymbol{\theta}} = \underset{\|\boldsymbol{\theta}\|=1,\ \theta_1>0}{\operatorname{argmin}} \sum_{j=1}^n\sum_{i=1}^n \rho_\tau\left(Y_i - \left[\mathbf{g}(u_j) + \mathbf{g}'(u_j)\mathbf{X}_{ij}^T\boldsymbol{\theta}\right]^T\mathbf{Z}_i\right)\omega_{ij}. \tag{6}$$

We call the estimation of $\boldsymbol{\theta}$ as the minimizing average check loss estimation(MACLE). Since both $\mathbf{g}(\cdot)$ and $\mathbf{g}'(\cdot)$ are unknown vector functions in (6), the direct minimization of (6) is impossible. To obtain the estimator, the unknown functions can be firstly replaced by their estimates, and then we can obtain the MACLE estimator for index parameter $\boldsymbol{\theta}$. The details of the minimization algorithm are given as follows.

- **Step 1.** Given initial value of $\boldsymbol{\theta}$ by $\tilde{\boldsymbol{\theta}}$, standardize $\tilde{\boldsymbol{\theta}}$ s.t. $\|\tilde{\boldsymbol{\theta}}\| = 1$, $\tilde{\theta}_1 > 0$. Denote $\alpha_j = \mathbf{g}(\mathbf{X}_j^T\tilde{\boldsymbol{\theta}})$, $\beta_j = \mathbf{g}'(\mathbf{X}_j^T\tilde{\boldsymbol{\theta}})$, $j = 1, \ldots, n$, which can be estimated by

$$(\tilde{\alpha}_j,\ \tilde{\beta}_j) = \underset{\alpha_j,\beta_j}{\operatorname{argmin}} \sum_{i=1}^n \rho_\tau\left[Y_i - \left(\alpha_j + \beta_j\mathbf{X}_{ij}^T\tilde{\boldsymbol{\theta}}\right)^T\mathbf{Z}_i\right]\omega_{ij} \quad \text{for } j = 1, \ldots, n.$$

- **Step 2.** Given $\tilde{\alpha}_j$, $\tilde{\beta}_j$, $j = 1, \ldots, n$, the estimation value of $\boldsymbol{\theta}$ can be updated by

$$\tilde{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}}{\operatorname{argmin}} \sum_{j=1}^n\sum_{i=1}^n \rho_\tau\left[Y_i - \left(\tilde{\alpha}_j + \tilde{\beta}_j\mathbf{X}_{ij}^T\boldsymbol{\theta}\right)^T\mathbf{Z}_i\right]\omega_{ij}, \tag{7}$$

where the values of ω_{ij} are calculated based on the value of $\tilde{\boldsymbol{\theta}}$ and h in **Step 1.**

- **Step 3.** Repeat **Step 1** and **Step 2** until convergence, then we obtain the final estimate of $\boldsymbol{\theta}$ denoted by $\hat{\boldsymbol{\theta}}$.
- **Step 4.** After obtaining the estimate $\hat{\boldsymbol{\theta}}$, for any inner point u on the tight support of $\mathbf{X}^T\hat{\boldsymbol{\theta}}$, $\mathbf{g}(u)$ can be estimated by $\hat{\mathbf{g}}(u; h, \hat{\boldsymbol{\theta}}) = \hat{\alpha}$, where

$$(\hat{\alpha}, \hat{\beta}) = \underset{\alpha,\beta}{\operatorname{argmin}} \sum_{i=1}^{n} \rho_\tau \left\{ Y_i - \left[\alpha + \beta \left(\mathbf{X}_i^T\hat{\boldsymbol{\theta}} - u\right)\right]^T \mathbf{Z}_i \right\} K_h \left(\mathbf{X}_i^T\hat{\boldsymbol{\theta}} - u\right). \quad (8)$$

Remark 1 After implementing **Step 2** in the above algorithm, $\tilde{\boldsymbol{\theta}}$ needs standardization as: $\tilde{\boldsymbol{\theta}} = \operatorname{sign}(\tilde{\theta}_1)\tilde{\boldsymbol{\theta}}/\|\tilde{\boldsymbol{\theta}}\|$, where $\operatorname{sign}(\tilde{\theta}_1)$ is the first component of $\tilde{\boldsymbol{\theta}}$. In addition, the initial estimate $\tilde{\boldsymbol{\theta}}$ in **Step 1** can be obtained using the single-index quantile regression method proposed in Wu et al. (2010) based on data $\{\mathbf{X}_i, Y_i\}_{i=1}^n$. Our simulations show that our proposed estimation procedure works well.

Remark 2 The optimal bandwidth h used in above algorithm can be selected by cross-validation method. To reduce the computation task, we can use the K-fold cross-validation method as following. Denote $F_1, \ldots, F_K$ as a partition of $\{1, \ldots, n\}$, and each F_i being roughly the same size, we may define a cross-validation score for the given bandwidth h

$$\mathrm{CV}(h) = \sum_{k=1}^{K} \rho_\tau \left(Y^{(F_k)} - \left(\hat{\mathbf{g}}^{(-F_k)} \left(\mathbf{X}^{(F_k)T}\hat{\boldsymbol{\theta}}^{(-F_k)}\right)\right)^T \mathbf{Z}^{(F_k)} \right),$$

where $Y^{(F_k)}$, $\mathbf{X}^{(F_k)}$ and $\mathbf{Z}^{(F_k)}$ denote using the observations from F_k only, and $\hat{\mathbf{g}}^{(-F_k)}$ and $\hat{\boldsymbol{\theta}}^{(-F_k)}$ are the estimates based on observations in $\{1, \ldots, n\}\backslash F_k$ with the given bandwidth h. Then the optimal bandwidth is selected by

$$h_{\mathrm{opt}} = \operatorname{argmin}_h \mathrm{CV}(h).$$

In the simulations and real data analysis, we use the fivefold cross-validation method.

3 Asymptotic properties

In this section, we present the asymptotic properties of the resulting estimators $\hat{\boldsymbol{\theta}}$ and $\hat{\mathbf{g}}(\cdot;\ h, \hat{\boldsymbol{\theta}})$. We first give some notations.

Let $f_Y(\cdot|\mathbf{X}^T\boldsymbol{\theta})$ and $F_Y(\cdot|\mathbf{X}^T\boldsymbol{\theta})$ be the density and cumulative distribution function of Y when given $\mathbf{X}^T\boldsymbol{\theta}$, respectively. Choose $K(\cdot)$ as a symmetric density function, and denote $\mu_j = \int u^j K(u)\mathrm{d}u$ and $\nu_j = \int u^j K^2(u)\mathrm{d}u$, $j = 0,\ 1,\ 2, \ldots$. Then, for the estimator $\hat{\boldsymbol{\theta}}$ obtained in (6), we have the following results.

Theorem 1 *Suppose the condition A.1–A.8 in the Appendix hold, then*

$$\sqrt{n}\left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\right) \xrightarrow{\mathcal{L}} N(0, \tau(1-\tau)\mathcal{G}^{-1}\mathcal{G}_0\mathcal{G}^{-1}), \quad (9)$$

where $\xrightarrow{\mathcal{L}}$ *denote convergence in distribution,* $\mathcal{G}_0 = \mathrm{E}(\mathcal{D})$, $\mathcal{G} = \mathrm{E}\left(f_Y(q_\tau(\mathbf{X}, \mathbf{Z})|\mathbf{X}^T\boldsymbol{\theta})\mathcal{D}\right)$, $\mathcal{D} = \mathbf{g}'(\mathbf{X}^T\boldsymbol{\theta})^T\pi_{\boldsymbol{\theta}}(\mathbf{X})\mathbf{g}'(\mathbf{X}^T\boldsymbol{\theta})\tilde{\mathbf{X}}\tilde{\mathbf{X}}^T$, $\tilde{\mathbf{X}} = \mathbf{X} - \mathrm{E}(\mathbf{X}|\mathbf{X}^T\boldsymbol{\theta})$, $\pi_{\boldsymbol{\theta}}(\mathbf{X}) = \mathrm{E}(\mathbf{Z}\mathbf{Z}^T|\mathbf{X}^T\boldsymbol{\theta})$.

By Theorem 1, we can get $\sqrt{n}$-consistence estimation of $\boldsymbol{\theta}$, then based on $\hat{\boldsymbol{\theta}}$, we can derive the estimation of $\mathbf{g}(\cdot)$ by (8). In the following, we present the asymptotic property of the nonparametric estimation of $\mathbf{g}(\cdot)$.

Theorem 2 *Suppose* $\mathbf{x}$ *be the inner point of the tight support of* $\mathbf{X}$*, and the conditions A.1–A.7 in appendix hold, then we have*

$$\sqrt{nh}\left\{\hat{\mathbf{g}}(\mathbf{x}^T\hat{\boldsymbol{\theta}}; h, \hat{\boldsymbol{\theta}}) - \mathbf{g}(\mathbf{x}^T\boldsymbol{\theta}) - \frac{1}{2}\mathbf{g}''(\mathbf{x}^T\boldsymbol{\theta})\mu_2 h^2\right\} \xrightarrow{\mathcal{L}} N(0, \Gamma_\tau(\mathbf{x}^T\boldsymbol{\theta})), \tag{10}$$

where $\Gamma_\tau(\mathbf{x}^T\boldsymbol{\theta}) = \tau(1-\tau)\nu_0\left[f_{\mathcal{U}}(\mathbf{x}^T\boldsymbol{\theta})f_Y(q_\tau(\mathbf{X}, \mathbf{Z})|\mathbf{x}^T\boldsymbol{\theta})^2\mathrm{E}(\mathbf{Z}\mathbf{Z}^T|\mathbf{X}^T\boldsymbol{\theta})\right]^{-1}$, $f_Y(q_\tau(\mathbf{X}, \mathbf{Z})|\mathbf{x}^T\boldsymbol{\theta})$ *is the conditional density value of* Y *at* $q_\tau(\mathbf{X}, \mathbf{Z})$ *given* $\mathbf{X}^T\boldsymbol{\theta}$*,* $f_{\mathcal{U}}(\cdot)$ *is the marginal density function of* $\mathbf{X}^T\boldsymbol{\theta}$*.*

4 Variable selection

In practice, the true model is unknown previously. An underfitted model will yield biased estimates and large prediction deviation, while an overfitted model will increase the complexity of the model and difficult to interpret. This motivates us to apply the penalized approach to simultaneously estimate the parameter $\boldsymbol{\theta}$ and select important variables of $\mathbf{X}$.

To conduct the variable selection, we firstly fit the model with all the index predictors. According to Theorem 1, the MACLE estimator, denoted as $\hat{\boldsymbol{\theta}}^{QR}$, is $\sqrt{n}$-consistent to the true parameter $\boldsymbol{\theta}$. Then, based on $\hat{\boldsymbol{\theta}}^{QR}$, we can obtain the penalized estimator $\hat{\boldsymbol{\theta}}^{\lambda}$ by minimizing the adaptive LASSO penalized average check loss function defined as

$$G_n(\boldsymbol{\theta}) = \sum_{j=1}^{n}\sum_{i=1}^{n}\rho_\tau\left(Y_i - \hat{\mathbf{g}}\left(\mathbf{X}_i^T\boldsymbol{\theta}\right)^T\mathbf{Z}_i\right)\omega_{ij} + \lambda\sum_{k=1}^{p}\frac{|\theta_k|}{|\hat{\theta}_k^{QR}|^2}, \tag{11}$$

where θ_k and $\hat{\theta}_k^{QR}$ are the kth element of $\boldsymbol{\theta}$ and $\hat{\boldsymbol{\theta}}^{QR}$ respectively, and $\hat{\mathbf{g}}(\cdot)$ is obtained from (8).

Without loss of generality, we assume that the first component of $\mathbf{X}$ is a relevant variable. For given tuning parameter λ, by minimizing $G_n(\boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$ under the constrains that $\|\boldsymbol{\theta}\| = 1$ and the first nonzero element of $\boldsymbol{\theta}$ is positive, we can obtain the sparse estimator $\hat{\boldsymbol{\theta}}^{\lambda}$, which is called as the adaptive LASSO penalized MACLE of $\boldsymbol{\theta}$.

Remark 3 Other variable selection methods such as SCAD proposed by Fan and Li (2001) or MCP proposed by Zhang (2010) can be also used here and the oracle property can be derived similarly. For the sake of computation, we adopt the adaptive LASSO method, which can be solved conveniently by nonlinear programming.

To obtain the sparse estimator, the optimal tuning parameter λ can be chosen through the Bayesian Information Criterion. Following Wang and Leng (2007), denote

$$\mathrm{BIC}(\lambda) = \log P_\tau(\lambda) + \frac{\log n}{n}\mathrm{DF}_\lambda, \tag{12}$$

where

$$P_\tau(\lambda) = \sum_{j=1}^{n}\sum_{i=1}^{n}\rho_\tau\left(Y_i - \hat{\mathbf{g}}\left(\mathbf{X}_i^T\hat{\boldsymbol{\theta}}^{\lambda}\right)^T\mathbf{Z}_i\right)\hat{\omega}_{ij}^{\lambda}$$
$$\text{with } \hat{\omega}_{ij}^{\lambda} = K_h\left(\mathbf{X}_{ij}^T\hat{\boldsymbol{\theta}}^{\lambda}\right)/\sum_{l=1}^{n}K_h\left(\mathbf{X}_{lj}^T\hat{\boldsymbol{\theta}}^{\lambda}\right),$$

and DF_λ is the number of non-zeros elements of $\hat{\boldsymbol{\theta}}^{\lambda}$. Then, the optimal tuning parameter is selected by

$$\hat{\lambda}_{\mathrm{opt}} = \underset{\lambda}{\mathrm{argmin}}\ \mathrm{BIC}(\lambda).$$

According to our simulation experience, this tuning parameter selection strategy works very well.

In the following, we will show that the adaptive LASSO penalized MACLE $\hat{\boldsymbol{\theta}}^{\lambda}$ enjoys the oracle property. Denote $\mathcal{A}_{\boldsymbol{\theta}} = \left\{j : \theta_j \neq 0\right\}$. Without loss of generality, suppose the true index parameter $\boldsymbol{\theta} = \binom{\boldsymbol{\theta}^1}{\boldsymbol{\theta}^2}$, where $\boldsymbol{\theta}^1$ is the sub-vector composed by the first p_0 nonzero elements of $\boldsymbol{\theta}$, $\boldsymbol{\theta}^2$ is composed by the remaining $p - p_0$ zero elements of $\boldsymbol{\theta}$. Thus, we have $\mathcal{A}_{\boldsymbol{\theta}} = \{1, \ldots, p_0\}$. Similarly, we define $\mathbf{X}_1$ be the sub-vector composed by the first p_0 elements of $\mathbf{X}$ and define $\tilde{\mathbf{X}}_1 = \mathbf{X}_1 - \mathrm{E}(\mathbf{X}_1|\mathbf{X}_1^T\boldsymbol{\theta}^1)$.

Theorem 3 (Oracle Property) *Suppose the conditions A.1–A.8 in the Appendix hold and $\lambda \to \infty$, $\lambda/\sqrt{n} \to 0$ as $n \to \infty$. Then for the adaptive LASSO penalized MACLE $\hat{\boldsymbol{\theta}}^{\lambda}$, we have*

(1) *Model selection consistency:* $\Pr(\{j : \hat{\theta}_j^{\lambda} \neq 0\} = \mathcal{A}_{\boldsymbol{\theta}}) = 1$,
(2) *Asymptotic normality:*

$$\sqrt{n}\left(\hat{\boldsymbol{\theta}}^{1\lambda} - \boldsymbol{\theta}^1\right) \xrightarrow{\mathcal{L}} N\left(0, \tau(1-\tau)\left(\mathcal{G}^*\right)^{-1}\mathcal{G}_0^*\left(\mathcal{G}^*\right)^{-1}\right), \tag{13}$$

where $\hat{\boldsymbol{\theta}}^{1\lambda}$ is the sub-vector composed by the first p_0 elements of $\hat{\boldsymbol{\theta}}^{\lambda}$, $\mathcal{G}_0^* = \mathrm{E}(\mathcal{D}^*)$, $\mathcal{G}^* = \mathrm{E}\left(f_Y(q_\tau(\mathbf{X}_1, \mathbf{Z})|\mathbf{X}_1^T\boldsymbol{\theta}^1)\mathcal{D}^*\right)$, $\mathcal{D}^* = \mathbf{g}'(\mathbf{X}_1^T\boldsymbol{\theta}^1)^T\pi_{\boldsymbol{\theta}^1}^*(\mathbf{X}_1)\mathbf{g}'(\mathbf{X}_1^T\boldsymbol{\theta}^1)\tilde{\mathbf{X}}_1\tilde{\mathbf{X}}_1^T$, $\tilde{\mathbf{X}}_1 = \mathbf{X}_1 - \mathrm{E}(\mathbf{X}_1|\mathbf{X}_1^T\boldsymbol{\theta}^1)$, $\pi_{\boldsymbol{\theta}}^*(\mathbf{X}) = \mathrm{E}(\mathbf{Z}\mathbf{Z}^T|\mathbf{X}_1^T\boldsymbol{\theta}^1)$.

Remark 4 Based on the asymptotic result of Theorem 1 or 3, it can be used to obtain the standard deviation estimate for index parameter by replacing some unknown quantities by their estimates. However, due to the ignoring estimation bias of these unknown quantities in small sample size, we propose to use the bootstrap method to obtain the standard deviation in practice when the sample size is small. Our limited experience shows that the bootstrap variance method works well in practice. However, its theoretical property of consistency issue for bootstrap variance remains an open question.

5 Monte Carlo simulation

In this subsection, we conduct two simulation studies with different error settings to examine the performance of the MACLE method and the proposed variable selection procedure.

Example 1 We conduct a simulation with the data generating from the following model

$$Y = g_1\left(\pi(\mathbf{X}^T\boldsymbol{\theta} - a)/(b-a)\right) Z_1 + g_2\left(\pi(\mathbf{X}^T\boldsymbol{\theta} - a)/(b-a)\right) Z_2 + \sigma\varepsilon, \quad (14)$$

where $\mathbf{X} = (X_1, X_2, X_3)^T$, $X_i \sim U[0, 1]$, and the correlation $\text{corr}(X_i, X_j) = 0.5^{|i-j|}$, $1 \leq i, j \leq 3$; (Z_1, Z_2) follows bivariate normal distribution with marginal distribution $N(0, 1)$ and correlation coefficient 0.5; $g_1(u) = \sin(u)$, $g_2(u) = \cos(u)$, $\boldsymbol{\theta} = (1, 1, 1)^T/\sqrt{3}$, $a = 0.3912$, $b = 1.3409$, $\sigma = 0.1$ or 0.25 denotes the low or high noise level. The sample size is set to be $n = 100, 200$ and 300. In our simulation, we consider the following four different error distributions as $N(0, 1)$, t distribution with degree of freedom 3 ($t(3)$), standard Cauchy and mixture normal $0.9N(0, 1) + 0.1N(0, 10^2)$, and $\mathbf{X}$, $\mathbf{Z}$ and ε are generated independently. For each type of error, $\boldsymbol{\theta}$ and $a_1(\cdot)$, $a_2(\cdot)$ are estimated by MACLE method under three quantile levels $\tau = 0.25$, 0.5 and 0.75. All the simulations are conducted 200 replications.

To assess the performance of our proposed method, we report the bias of the estimate for the index parameter and the mean integrated squared errors (MISE) of the estimate for nonparametric function, that is, $\text{MISE} = \frac{1}{2}\sum_{j=1}^{2}\text{ISE}_j$, where

$$\text{ISE}_j = \frac{1}{n_{\text{grid}}}\sum_{k=1}^{n_{\text{grid}}}(\hat{g}_j(u_k) - g_j(u_k))^2,$$

and $\{u_k : k = 1, \ldots, n_{\text{grid}}\}$ are regular grid points with $n_{\text{grid}} = 100$.

The results of 200 times simulation are summarized in Tables 1–3, where we present the bias and the standard deviation of $\boldsymbol{\theta}$ and MISE of nonparametric function $g(\cdot)$. From Tables 1, 2 and 3, we can see that all the biases of $\boldsymbol{\theta}$ are close to zero and the corresponding estimates of the standard deviation decrease as the sample size increases for all the case. Meanwhile, the MISE of the nonparametric function $g(\cdot)$ becomes more smaller when the sample size is large. Even for the large noise level $\sigma = 0.25$, the performance of our proposed MACLE method is still satisfactory. Therefore, our

Table 1 Summary of the bias and MISE for $\sigma = 0.1$, where the std denotes the sample standard deviation calculated over 200 replications

n	Error type	τ	$\hat{\theta}_1$ (bias)	$\hat{\theta}_1$ (std)	$\hat{\theta}_2$ (bias)	$\hat{\theta}_2$ (std)	$\hat{\theta}_3$ (bias)	$\hat{\theta}_3$ (std)	MISE	Std
100	Normal	0.25	0.0010	0.0177	−0.0039	0.0179	0.0021	0.0165	0.0084	0.0021
		0.5	−0.0270	0.0764	0.0174	0.0923	−0.0088	0.0765	0.0031	0.0014
		0.75	0.0006	0.0216	0.0005	0.0256	−0.0025	0.0230	0.0045	0.0020
	$t(3)$	0.25	0.0014	0.0215	−0.0055	0.0262	0.0027	0.0239	0.0125	0.0099
		0.5	−0.0015	0.0201	−0.0012	0.0233	0.0016	0.0188	0.0036	0.0020
		0.75	0.0016	0.0212	−0.0050	0.0258	0.0019	0.0251	0.0077	0.0124
	Cauchy	0.25	−0.0017	0.0444	−0.0039	0.0589	−0.0012	0.0498	0.4560	0.5194
		0.5	−0.0017	0.0262	−0.0002	0.0317	−0.0002	0.0271	0.0119	0.0414
		0.75	0.0009	0.0351	−0.0007	0.0401	−0.0036	0.0322	0.0882	0.4378
	Mixture normal	0.25	0.0007	0.0180	−0.0016	0.0214	0.0001	0.0183	0.0148	0.0151
		0.5	0.0006	0.0166	0.0009	0.0211	−0.0023	0.0173	0.0054	0.0124
		0.75	0.0009	0.0216	−0.0013	0.0226	−0.0008	0.0198	0.0137	0.0310
200	Normal	0.25	0.0004	0.0076	−0.0002	0.0112	−0.0008	0.0101	0.0073	0.0011
		0.5	−0.0016	0.0071	0.0026	0.0105	0.0001	0.0103	0.0019	0.0004
		0.75	0.0009	0.0067	−0.0007	0.0107	−0.0011	0.0101	0.0032	0.0008
	$t(3)$	0.25	−0.0002	0.0074	0.0004	0.0114	−0.0002	0.0102	0.0092	0.0019
		0.5	0.0000	0.0062	−0.0006	0.0110	0.0002	0.0091	0.0022	0.0007
		0.75	−0.0002	0.0071	−0.0003	0.0114	0.0002	0.0108	0.0045	0.0018
	Cauchy	0.25	0.0002	0.0088	−0.0014	0.0145	0.0003	0.0117	0.0274	0.0931
		0.5	0.0002	0.0073	−0.0005	0.0124	−0.0002	0.0100	0.0032	0.0024
		0.75	0.0005	0.0090	−0.0002	0.0145	−0.0011	0.0121	0.0122	0.0099
	Mixture normal	0.25	0.0000	0.0076	0.0011	0.0116	−0.0011	0.0105	0.0098	0.0028
		0.5	−0.0003	0.0069	−0.0000	0.0104	0.0002	0.0095	0.0024	0.0016
		0.75	0.0003	0.0075	0.0009	0.0110	−0.0014	0.0098	0.0049	0.0026
300	Normal	0.25	−0.0006	0.0064	0.0003	0.0104	0.0005	0.0084	0.0072	0.0008
		0.5	0.0003	0.0058	−0.0003	0.0097	−0.0003	0.0077	0.0017	0.0004
		0.75	0.0001	0.0062	0.0005	0.0097	−0.0007	0.0083	0.0031	0.0006
	$t(3)$	0.25	−0.0003	0.0073	−0.0011	0.0114	0.0010	0.0097	0.0090	0.0016
		0.5	−0.0003	0.0068	−0.0006	0.0099	0.0006	0.0090	0.0018	0.0004
		0.75	−0.0005	0.0066	−0.0002	0.0104	0.0007	0.0090	0.0040	0.0010
	Cauchy	0.25	−0.0000	0.0092	0.0009	0.0172	−0.0011	0.0110	0.0160	0.0042
		0.5	−0.0004	0.0068	0.0001	0.0109	0.0002	0.0091	0.0022	0.0009
		0.75	−0.0009	0.0085	0.0011	0.0130	0.0002	0.0109	0.0090	0.0051
	Mixture normal	0.25	0.0005	0.0073	0.0004	0.0111	−0.0013	0.0092	0.0088	0.0019
		0.5	−0.0008	0.0057	−0.0000	0.0099	0.0010	0.0087	0.0018	0.0005
		0.75	0.0005	0.0071	−0.0001	0.0109	−0.0008	0.0094	0.0042	0.0027

Table 2 Summary of the bias and MISE for $\sigma = 0.25$, where the std denotes the sample standard deviation calculated over 200 replications

n	Error type	τ	$\hat{\theta}_1$ (bias)	$\hat{\theta}_1$ (std)	$\hat{\theta}_2$ (bias)	$\hat{\theta}_2$ (std)	$\hat{\theta}_3$ (bias)	$\hat{\theta}_3$ (std)	MISE	Std
100	Normal	0.25	0.0012	0.0100	−0.0004	0.0154	−0.0020	0.0140	0.0308	0.0117
		0.5	0.0002	0.0100	−0.0004	0.0167	−0.0005	0.0131	0.0096	0.0050
		0.75	−0.0002	0.0109	−0.0008	0.0172	0.0004	0.0133	0.0213	0.0131
	$t(3)$	0.25	−0.0017	0.0101	0.0004	0.0178	0.0016	0.0147	0.0470	0.0243
		0.5	−0.0018	0.0106	0.0008	0.0162	0.0015	0.0155	0.0198	0.0587
		0.75	−0.0011	0.0129	0.0009	0.0190	0.0003	0.0149	0.0417	0.0554
	Cauchy	0.25	−0.0021	0.0155	0.0029	0.0284	−0.0005	0.0233	0.8090	2.9809
		0.5	−0.0009	0.0134	−0.0002	0.0234	0.0004	0.0195	0.4864	2.8403
		0.75	−0.0060	0.0556	0.0052	0.0364	0.0001	0.0248	0.8596	2.7364
	Mixture normal	0.25	−0.0004	0.0098	0.0001	0.0180	−0.0001	0.0136	0.1317	0.4017
		0.5	0.0000	0.0124	0.0004	0.0180	−0.0010	0.0177	0.0157	0.0297
		0.75	−0.0001	0.0095	0.0008	0.0140	−0.0009	0.0121	0.0614	0.1373
200	Normal	0.25	−0.0015	0.0096	0.0002	0.0150	0.0016	0.0135	0.0284	0.0106
		0.5	0.0004	0.0099	−0.0007	0.0173	−0.0007	0.0142	0.0106	0.0103
		0.75	−0.0003	0.0107	0.0020	0.0153	−0.0016	0.0140	0.0205	0.0081
	$t(3)$	0.25	−0.0014	0.0110	−0.0013	0.0186	0.0023	0.0148	0.0336	0.0102
		0.5	−0.0007	0.0108	−0.0005	0.0173	0.0007	0.0146	0.0063	0.0034
		0.75	−0.0010	0.0107	0.0021	0.0145	−0.0005	0.0135	0.0232	0.0083
	Cauchy	0.25	−0.0027	0.0240	0.0017	0.0354	0.0001	0.0309	0.1150	0.2715
		0.5	−0.0012	0.0120	0.0014	0.0182	−0.0000	0.0169	0.0149	0.0320
		0.75	−0.0002	0.0128	−0.0017	0.0208	0.0006	0.0181	0.0821	0.1537
	Mixture normal	0.25	−0.0006	0.0107	−0.0012	0.0159	0.0012	0.0143	0.0341	0.0139
		0.5	0.0014	0.0102	−0.0030	0.0168	−0.0004	0.0140	0.0070	0.0108
		0.75	−0.0001	0.0103	0.0005	0.0162	−0.0007	0.0134	0.0235	0.0151
300	Normal	0.25	−0.0001	0.0089	0.0010	0.0160	−0.0011	0.0124	0.0236	0.0042
		0.5	−0.0014	0.0092	0.0002	0.0164	0.0015	0.0130	0.0037	0.0012
		0.75	0.0004	0.0088	−0.0006	0.0135	−0.0006	0.0116	0.0143	0.0033
	$t(3)$	0.25	0.0001	0.0106	0.0003	0.0170	−0.0009	0.0135	0.0318	0.0071
		0.5	−0.0005	0.0093	0.0018	0.0160	−0.0012	0.0126	0.0043	0.0019
		0.75	−0.0013	0.0099	0.0003	0.0167	0.0013	0.0118	0.0207	0.0063
	Cauchy	0.25	0.0009	0.0155	−0.0013	0.0259	−0.0016	0.0200	0.0655	0.0301
		0.5	0.0003	0.0113	−0.0005	0.0209	−0.0009	0.0155	0.0060	0.0037
		0.75	−0.0015	0.0155	−0.0002	0.0249	0.0012	0.0203	0.0501	0.0392
	Mixture normal	0.25	0.0009	0.0102	−0.0009	0.0176	−0.0013	0.0126	0.0304	0.0075
		0.5	0.0004	0.0092	0.0003	0.0158	−0.0013	0.0120	0.0045	0.0018
		0.75	−0.0005	0.0096	0.0006	0.0148	−0.0001	0.0128	0.0205	0.0073

Table 3 Summary of the bias and MISE for Example 2

n	Error type	τ	$\hat{\theta}_1$ (bias)	$\hat{\theta}_1$ (std)	$\hat{\theta}_2$ (bias)	$\hat{\theta}_2$ (std)	$\hat{\theta}_3$ (bias)	$\hat{\theta}_3$ (std)	MISE	Std
100	Normal	0.25	0.0006	0.0125	−0.0002	0.0127	−0.0004	0.0080	0.1449	0.2438
		0.5	−0.0003	0.0110	−0.0000	0.0111	0.0030	0.0140	0.0547	0.0602
		0.75	0.0019	0.0115	−0.0012	0.0109	0.0016	0.0130	0.0534	0.0526
	$t(3)$	0.25	−0.0003	0.0135	−0.0013	0.0134	0.0005	0.0086	0.2976	0.4530
		0.5	−0.0002	0.0107	−0.0001	0.0107	−0.0012	0.0144	0.1393	0.4218
		0.75	0.0010	0.0114	−0.0006	0.0119	0.0011	0.0130	0.1119	0.1300
	Cauchy	0.25	−0.0009	0.0130	0.0009	0.0135	−0.0003	0.0080	3.2268	4.4820
		0.5	−0.0018	0.0128	0.0014	0.0123	0.0006	0.0158	2.0617	2.3269
		0.75	0.0006	0.0130	−0.0004	0.0129	0.0010	0.0148	3.6395	3.2962
	Mixture normal	0.25	−0.0003	0.0125	0.0003	0.0128	−0.0002	0.0078	0.2583	0.4760
		0.5	−0.0004	0.0102	0.0001	0.0102	0.0013	0.0146	0.1535	0.3437
		0.75	0.0005	0.0108	−0.0014	0.0117	−0.0009	0.0126	0.3456	0.8059
200	Normal	0.25	0.0018	0.0140	−0.0004	0.0139	−0.0010	0.0079	0.0550	0.0344
		0.5	−0.0002	0.0107	−0.0001	0.0107	−0.0003	0.0135	0.0183	0.0119
		0.75	0.0023	0.0141	−0.0018	0.0134	0.0013	0.0127	0.0847	0.1792
	$t(3)$	0.25	0.0016	0.0147	−0.0011	0.0146	−0.0006	0.0095	0.1265	0.1944
		0.5	−0.0002	0.0110	−0.0001	0.0110	0.0011	0.0150	0.0289	0.0297
		0.75	0.0042	0.0147	−0.0033	0.0138	0.0020	0.0130	0.1596	0.3562
	Cauchy	0.25	0.0031	0.0166	0.0005	0.0168	−0.0022	0.0108	0.7402	1.1607
		0.5	0.0005	0.0137	−0.0009	0.0138	−0.0001	0.0141	0.1444	0.5194
		0.75	0.0125	0.0357	−0.0089	0.0758	−0.0101	0.0641	0.4324	0.7214
	Mixture normal	0.25	0.0008	0.0153	0.0002	0.0146	−0.0008	0.0088	0.2244	0.2363
		0.5	0.0000	0.0112	−0.0004	0.0112	−0.0010	0.0147	0.0280	0.0313
		0.75	0.0002	0.0120	−0.0006	0.0125	−0.0000	0.0129	0.2535	0.5345
300	Normal	0.25	0.0044	0.0122	−0.0001	0.0137	−0.0024	0.0080	0.0417	0.0189
		0.5	−0.0001	0.0105	−0.0002	0.0105	0.0011	0.0131	0.0108	0.0082
		0.75	0.0031	0.0123	−0.0023	0.0126	0.0017	0.0120	0.0803	0.1981
	$t(3)$	0.25	0.0038	0.0126	0.0018	0.0156	−0.0031	0.0094	0.0650	0.0378
		0.5	0.0004	0.0101	−0.0007	0.0101	0.0008	0.0134	0.0165	0.0121
		0.75	0.0030	0.0135	−0.0026	0.0124	0.0012	0.0128	0.1210	0.2187
	Cauchy	0.25	0.0065	0.0214	−0.0009	0.0210	−0.0035	0.0134	0.8551	1.0081
		0.5	0.0001	0.0134	−0.0005	0.0133	−0.0011	0.0142	0.2521	0.1360
		0.75	0.0025	0.0167	−0.0031	0.0167	0.0001	0.0147	0.7850	1.2013
	Mixture normal	0.25	0.0050	0.0140	−0.0021	0.0148	−0.0018	0.0089	0.1186	0.2197
		0.5	−0.0002	0.0103	−0.0001	0.0103	0.0001	0.0139	0.0183	0.0284
		0.75	0.0021	0.0129	−0.0020	0.0124	0.0007	0.0122	0.2681	0.2786

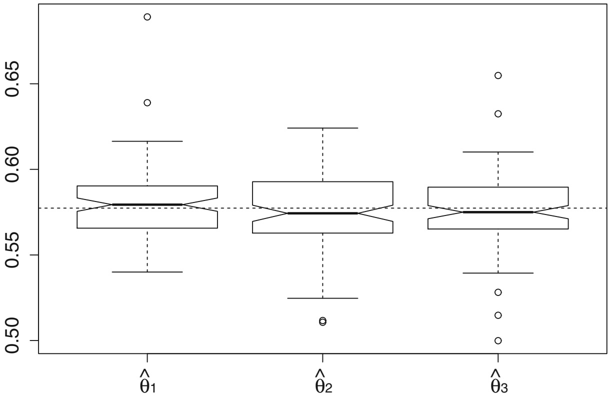

Fig. 1 *Boxplot* of 200 times estimates of $\boldsymbol{\theta}$ at $\tau = 0.5$ for Cauchy error

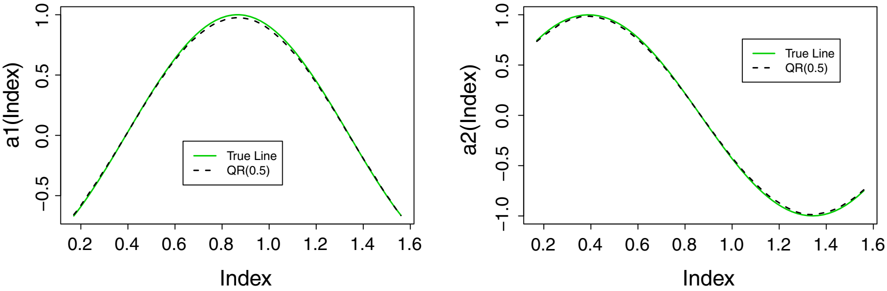

Fig. 2 Nonparametric estimate when error follows standard Cauchy and $\tau = 0.5$

proposed MACLE procedure is robust to different error distribution, especially for the Cauchy distributed error. Particularly, Fig. 1 shows the boxplot of the 200 times estimates of $\boldsymbol{\theta}$ with sample size $n = 200$, noise level $\sigma = 0.1$ and quantile level $\tau = 0.5$ when the error follows standard Cauchy. In addition, we present the median of 200 times estimation of the nonparametric coefficient functions in Fig. 2. It is clear that the estimation curve (dashed line) is very close to the true curve (solid line).

Example 2 In this example, the data are generated from the following heteroscedastic model

$$\begin{aligned} Y = g_0(\mathbf{X}^T\boldsymbol{\theta}) + g_1\left(\pi(\mathbf{X}^T\boldsymbol{\theta} - a)/(b-a)\right) Z_1 \\ + g_2\left(\pi(\mathbf{X}^T\boldsymbol{\theta} - a)/(b-a)\right) Z_2 + 0.25 \cdot (1 + |X_1|)\varepsilon, \end{aligned}$$

where the true value $\boldsymbol{\theta} = (\tau, \tau, 1 - 2\tau)^T/\sqrt{6\tau^2 - 4\tau + 1}$, which depends on the quantile level τ; $g_0(u) = 2\exp(-(u-\tau)^2)$, $g_1(u)$ and $g_2(u)$ are the same in Example 1. The covariate $\mathbf{X} = (X_1, X_2, X_3)^T$, $X_i \sim U[0, 1]$, and the correlation $\text{corr}(X_i, X_j) = 0.5$, $1 \le i, j \le 3$; $\mathbf{Z} = (Z_1, Z_2)$ is generated as following two steps: we first generate $\mathbf{U} = (U_1, U_2)$, which follows bivariate normal distribution with marginal distribution

$N(0, 1)$ and correlation coefficient 0.5, then we get $Z_j = U_j + \mathbf{X}^T\boldsymbol{\theta}$, $j = 1, 2$. The error settings are the same as Example 1. In this example, we note that two covariates $\mathbf{X}$ and $\mathbf{Z}$ are independent for given $\mathbf{X}^T\boldsymbol{\theta}$, but they are not mutually independent. The aim of this example is to examine whether our proposed MACLE still works well for heteroscedastic model with correlation between covariates. The results over 200 replications are shown in Table 3.

As we can see from Table 3, all the biases of the estimate for index parameter are close to zero, and the MISE of the nonparametric functions become smaller as the sample size increases for each error distribution. To conclude, queryKindly check and confirm the edit in the sentence “To conclude, our....”our estimation procedure still performs well for heteroscedastic model.

Example 3 Reconsider model (14), where $\mathbf{X} = (X_1, \ldots, X_8)^T$, $X_i \sim U[0, 1], i = 1, \ldots, 8$ with $\text{corr}(X_i, X_j) = \frac{1}{2}^{|i-j|}$, $\boldsymbol{\theta} = (3,\ 1.5, 0, 0,\ 2, 0, 0, 0, 0)^T/\sqrt{15.25}$ and $\sigma = 0.1$. Other settings are the same as in Example 1. In each simulation, we get 100 i.i.d. sample and consider adaptive LASSO penalized MACLE variable selection methods for $\tau = 0.25,\ 0.5,\ 0.75$, respectively. For each case, we conduct 200 times simulation.

The results of the variable selection are summarized in Table 4, where column “C” shows the average number of the zero elements in $\boldsymbol{\theta}$ correctly identified to be zero and column “IC” presents the average number of the non-zero elements of $\boldsymbol{\theta}$ incorrectly estimated to be zero. The column “U-fit” shows the proportion of trials excluding any nonzero coefficients in 200 replications, i.e. at least one important variable not been selected in the final model. Additionally, we report the proportion of trials selecting the exact sub-model by “C-fit” and the proportion of trials selecting all three significant variables and at least including one noise variables by “O-fit”, respectively. Several observations can be seen from Table 4. Firstly, the adaptive LASSO penalized MACLE variable selection method is robust to various error distributions. Secondly,

Table 4 Summarize of 200 times variable selection of SICM

Error type	τ	C	IC	U-fit	O-fit	C-fit
Standard normal	0.5	4.985	0	0	0.005	0.995
	0.25	5	0	0	0	1
	0.75	5	0	0	0	1
$t(3)$	0.5	4.995	0	0	0.005	0.995
	0.25	4.995	0	0	0.005	0.995
	0.75	4.990	0.010	0.010	0.010	0.980
Standard cauchy	0.5	5	0.140	0.085	0	0.915
	0.25	4.970	0.180	0.110	0.015	0.875
	0.75	4.990	0.160	0.140	0.010	0.850
Mixture normal	0.5	5	0	0	0	1
	0.25	4.995	0	0	0.005	0.995
	0.75	4.995	0.067	0.050	0.005	0.945

the performance of the variable selection method is satisfactory at different quantile levels. Thirdly, we can see that the BIC tuning parameter selection strategy performs well. These findings further demonstrate our theoretical results in Sect. 4.

6 Real data analysis

In this section, we consider the Boston housing data, which can be get from http://lib.stat.cmu.edu/datasets/bostoncorrected.txt, with some corrections and augmentation by the latitude and longitude of each observation, called the Corrected Boston House Price Data. There are 506 observations, 15 non-constant predictor variables and one response variable, corrected median value of owner-occupied homes (CMEDV). Predictors include longitude (LON), Latitude (LAT), crime rate (CRIM), proportion of area zoned with large lots (ZN), proportion of non-retail business acres per town (INDUS), Charles River as a dummy variable (= 1 if tract bounds river; 0 otherwise) (CHAS), nitric oxides concentration (NOX), average number of rooms per dwelling (RM), proportion of owner-occupied units built prior to 1940 (AGE), weighted distances to five Boston employment centers (DIS), index of accessibility to radial highways (RAD), property tax rate (TAX), pupil–teacher ratio by town (PTRATIO), black population proportion town (B), and lower status population proportion (LSTAT). Following previous studies we take logarithmic transformation on TAX and LSTAT. For simplicity, we exclude the categorical variable RAD and standardize the other covariates aside from CHAS. We construct SICM as follows

$$\begin{aligned} q_\tau(\text{CMDEV}) &= g_0(\text{Index}) + g_1(\text{Index})\text{DIS} + g_2(\text{Index})\text{LON};\\ \text{Index} &= \text{RM}\theta_1 + \text{Log(TAX)}\theta_2 + \text{PTRATIO}\theta_3 + \text{Log(LSTAT)}\theta_4 + \text{CRIM}\theta_5\\ &\quad + \text{B}\theta_6 + \text{NOX}\theta_7 + \text{LAT}\theta_8 + \text{ZN}\theta_9. \end{aligned} \tag{15}$$

The adaptive LASSO penalized MACLE estimates of $\boldsymbol{\theta}$ are presented in Table 5. From which we can see that there is difference in the influence of the covariates on the different conditional quantile of the CMDEV. For $\tau = 0.5$, we plot the estimate of baseline function $g_0(\cdot)$ in Fig. 3 and the estimates of the coefficient functions in Fig. 4. We found that two coefficient functions have significant nonlinear effects, which indicate that the relationships between index variable and covariates DIS and LON have important interaction effects for the response variable. On the other hand, by analyzing the normality of the residuals by Shapiro Wilk test (Shapiro and Wilk 1965), the p value is small than 2.2×10^{-16}, which means that the error can not be

Table 5 The sparse estimate of the $\boldsymbol{\theta}$ in Boston Housing data

τ	RM	PTRATIO	Log(LSTAT)	CRIM	B	Log (TAX)	NOX	ZN	LAT
0.25	0.534	−0.213	−0.714	−0.220	0.204	−0.264	0	0	0
0.5	0.550	−0.244	−0.722	0	0.247	−0.237	0	0	0
0.75	0.537	−0.262	−0.802	0	0	0	0	0	0

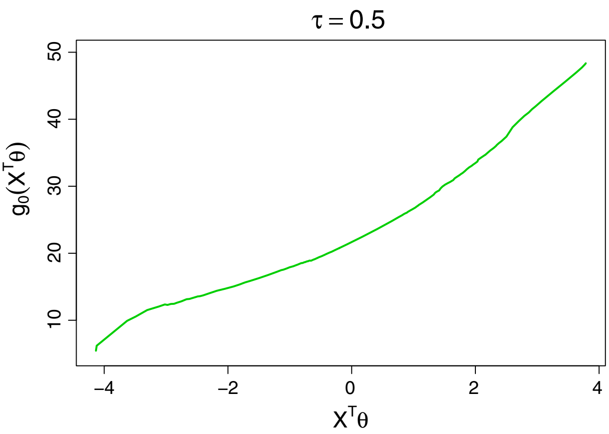

Fig. 3 Estimate of $g_0(\cdot)$ in Boston Housing data

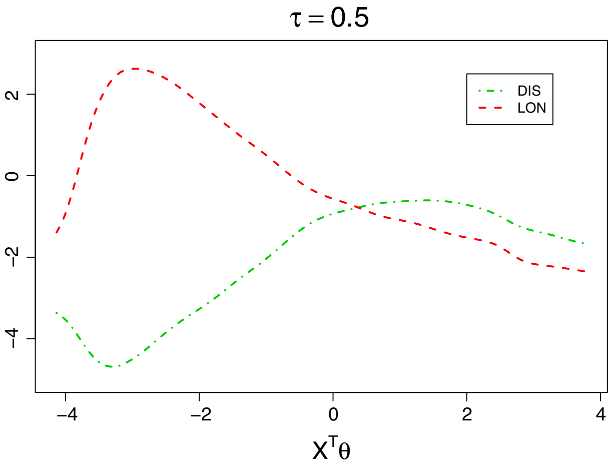

Fig. 4 Estimate of $g_1(\cdot)$ and $g_2(\cdot)$ in Boston Housing data

normal. The normal Q–Q plot of the residuals when $\tau = 0.5$ is presented in Fig. 5. This phenomenon may throw some light on the usage and robustness of the quantile regression of semiparametric models. In practice, the error's distribution can not be available previously; hence, the quantile regression methods will be useful to provide the underlying relationships between the response and the covariates.

To further illustrate the usefulness of SICM, we also fit the data by the single-index model (SIM)

$$q_\tau(\text{CMDEV}) = g_0(\text{Index})$$

and partially linear single-index model (PLSIM)

$$q_\tau(\text{CMDEV}) = g_0(\text{Index}) + \gamma_1\text{DIS} + \gamma_2\text{LON}$$

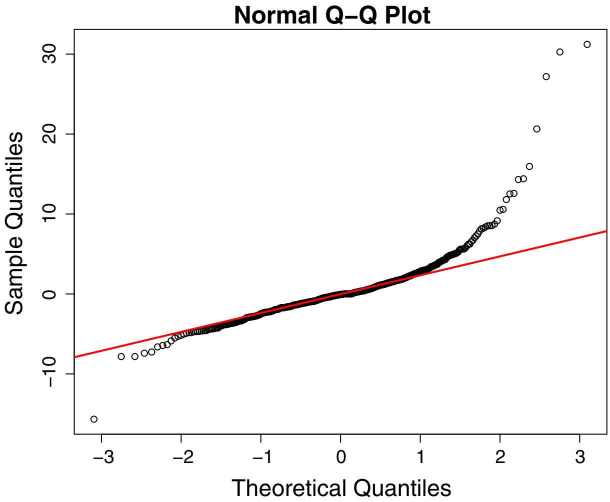

Fig. 5 Normal Q–Q plot of the residuals in QR($\tau = 0.5$) of Boston Housing data by SICM

Table 6 FAD and PAD of the three models for Boston housing data

Model	SICM	SIM	PLSIM
FAD	2.4098	2.6320	2.5822
PAD	12.0221	12.2338	12.0890

at the quantile level $\tau = 0.5$. The results of the fitted absolute deviation (FAD)

$$\text{FAD} = \frac{1}{n}\sum_{i=1}^{n}|Y_i - \hat{Y}_i|$$

are shown in Table 6. Moreover, we also reported the prediction absolute deviation (PAD) of three different models in Table 6, where

$$\text{PAD} = \frac{1}{n}\sum_{i=1}^{n}|Y_i - \hat{Y}_i^{(-i)}|,$$

and $\hat{Y}_i^{(-i)}$, $i = 1, \ldots, n$, denote the fitted value based on the $n - 1$ observations after deleting the ith sample.

From Table 6, we can see that the values of both FAD and PAD for SICM are the smallest among three candidate models.

Appendix

To establish the asymptotic properties and the Oracle property of the proposed methods, we need the following regularity conditions:

- A.1 The kernel function $K(\cdot)$ is a symmetric Lipschitz continues density function with a compact support and it satisfies $\int_{-\infty}^{\infty} z^2 K(z)dz < \infty$, $\int_{-\infty}^{\infty} z^j K^2(z)dz < \infty$, $j = 0, 1, 2$;
- A.2 Denote $\boldsymbol{\Theta}$ as the local neighborhood of $\boldsymbol{\theta}$ and Ξ as the compact support of the covariate $\mathbf{X}$. Let $\mathcal{U} = \left\{u = \mathbf{x}^T\boldsymbol{\theta}; \mathbf{x} \in \Xi, \boldsymbol{\theta} \in \boldsymbol{\Theta}\right\}$ be the compact support of $\mathbf{X}^T\boldsymbol{\theta}$ with marginal density $f_{\mathcal{U}}(u)$. Furthermore, $f_{\mathcal{U}}(u)$ is first-order Lipschitz continuous and its lower bound is positive;
- A.3 Denote $u_{\boldsymbol{\theta}} = \mathbf{x}^T\boldsymbol{\theta}$, the index function $\alpha(u_{\boldsymbol{\theta}})$ is second order differentiable with respect to $u_{\boldsymbol{\theta}}$ and it is Lipschitz continues with respect to $\boldsymbol{\theta}$;
- A.4 Given $\mathrm{X}^T\boldsymbol{\theta} = u$, the conditional density $f(y|u)$ is Lipschitz continues with respect to y and u;
- A.5 The matrix functions $\mathrm{E}(\mathbf{X}|\mathbf{X}^T\boldsymbol{\theta} = u)$, $\mathrm{E}(\mathbf{Z}|\mathbf{X}^T\boldsymbol{\theta} = u)$, $\mathrm{E}(\mathbf{X}^{\otimes 2}|\mathbf{X}^T\boldsymbol{\theta} = u)$, $\mathrm{E}(\mathbf{Z}^{\otimes 2}|\mathbf{X}^T\boldsymbol{\theta} = u)$ and $\mathrm{E}(\mathbf{X}\mathbf{Z}^T|\mathbf{X}^T\boldsymbol{\theta} = u)$ are consistently Lipschitz continuous with respect to $u \in \mathcal{U}$ and $\boldsymbol{\theta} \in \Theta$, where $A^{\otimes 2} = AA^T$, A is matrix or vector;
- A.6 The bandwidth h satisfies $h \sim n^{-\delta}$, where $1/6 < \delta < 1/4$;
- A.7 $\forall u \in \mathcal{U}$ and $\boldsymbol{\theta} \in \Theta$, the matrix $\mathrm{E}(\mathbf{Z}^{\otimes 2}|\mathbf{X}^T\boldsymbol{\theta} = u)$ is invertible;
- A.8 $\forall \boldsymbol{\theta} \in \Theta$, the matrix $\mathcal{G}$ defined in Theorem 1 is positive definite.

Remark 5 The above conditions are commonly used in the semi-parametric literature and they can be easily satisfied in many applications. Condition A.1 simply requires that the kernel function is a proper density with finite second moment, which is required to derive the asymptotic variance of estimators. Condition A.2 guarantees the existence of any ratio terms with the density appearing as part of the denominator. Conditions A.3 and A.4 are commonly used in single-index model and quantile regression literature, see Wu et al. (2010), Kai et al. (2011) and Xue and Pang (2013). Condition A.5 list some common assumptions in semi-parametric model, see for example Huang and Zhang (2012), Kai et al. (2011) and Xue and Pang (2013). Condition A.6 admits the optimal bandwidth in nonparametric estimation. Condition A.7 comes from Lu et al. (2007) and Kai et al. (2011). Condition A.8 is used to derive the consistence of the variable selection method.

The following two lemmas will be frequently used in our proof.

Lemma 1 *Suppose $A_n(s)$ is convex and can be represented as $\frac{1}{2}s^T Vs + U_n^T s + C_n + r_n(s)$, where V is symmetric and positive definite, U_n is stochastically bounded, C_n is arbitrary, and $r_n(s)$ goes to zero in probability for each s. Then the argmin of A_n is only $o_p(1)$ away from $\beta_n = -V^{-1}U_n$, the argmin of $\frac{1}{2}s^T Vs + U_n^T s + C_n$.*

Proof This lemma comes from the Basic proposition in Hjort and Pollard (1993). □

Lemma 2 *Let $(U_1, Y_1), \ldots, (U_n, Y_n)$ be independent and identically distributed random vectors, where Y_i and U_i are scalar random variable. Assume further that $\mathrm{E}|Y|^s < \infty$ and $\sup\limits_u \int |y|^s f(u, y)\mathrm{d}y < \infty$, where $f(\cdot,\cdot)$ denotes the joint density of (U, Y). Let $K(\cdot)$ be a bounded positive function with a bounded support and satisfying a Lipschitz condition. Then*

$$\sup_{u\in\mathcal{U}}\left|\frac{1}{n}\sum_{i=1}^{n}[K_h(U_i-u)Y_i - \mathrm{E}(K_h(U_i-u)Y_i)]\right| = O_p\left[\left(\frac{\ln(1/h)}{nh}\right)^{1/2}\right],$$

provided that $n^{2\varepsilon-1}h \to \infty$ *for some* $\varepsilon < 1 - s^{-1}$*, where* $\mathscr{U}$ *is the compact support of* U*.*

Proof This follows from the result by Mack and Silverman (1982). □

Let $\tilde{\boldsymbol{\theta}}$ be the initial consistency estimate of parameter $\boldsymbol{\theta}$, which can be obtained using existing methods, see Remark 1. In the following, we assume $\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta} = o_p(1)$. Denote $\delta_n = [\ln(1/h)/nh]^{1/2}$, $\tau_n = h^2 + \delta_n$, $\delta_{\boldsymbol{\theta}} = \|\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}\|$ and $K_{ih}^{\boldsymbol{\theta}} = K_{i,h}^{\boldsymbol{\theta}}(\mathbf{x}) = K_h(\mathbf{X}_{i0}^T\boldsymbol{\theta})$, where $\mathbf{X}_{i0} = \mathbf{X}_i - \mathbf{x}$. Then we have the following Lemma 3.

Lemma 3 *Assume* $\mathbf{x}$ *as the interior point of* Ξ*, denote*

$$S_l(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n} K_{ih}^{\tilde{\boldsymbol{\theta}}}\mathbf{Z}_i\mathbf{Z}_i^T\left(\frac{\mathbf{X}_{i0}^T\tilde{\boldsymbol{\theta}}}{h}\right)^l, \quad l = 0, 1, 2,$$

$$E_l(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n} K_{ih}^{\tilde{\boldsymbol{\theta}}}\mathbf{Z}_i\mathbf{Z}_i^T\left(\frac{\mathbf{X}_i - \mathbf{x}}{h}\right)^{\otimes l}, \quad l = 1, 2,$$

then we have

$$\begin{aligned}
S_0(\mathbf{x}) &= \pi_{\tilde{\boldsymbol{\theta}}}(\mathbf{x})f_{\mathcal{U}}(\mathbf{x}^T\tilde{\boldsymbol{\theta}}) + O(h^2 + \delta_n),\\
&= \pi_{\boldsymbol{\theta}}(\mathbf{x})f_{\mathcal{U}}(\mathbf{x}^T\tilde{\boldsymbol{\theta}}) + O(h^2 + \delta_{\boldsymbol{\theta}} + \delta_n),\\
S_1(\mathbf{x}) &= O(h + h\delta_{\boldsymbol{\theta}} + \delta_n),\\
S_2(\mathbf{x}) &= \mu_2\pi_{\boldsymbol{\theta}}(\mathbf{x})f_{\mathcal{U}}(\mathbf{x}^T\tilde{\boldsymbol{\theta}}) + O(h^2 + \delta_{\boldsymbol{\theta}} + \delta_n),\\
E_1(\mathbf{x}) &= f_{\mathcal{U}}(\mathbf{x}^T\tilde{\boldsymbol{\theta}})\pi_{\boldsymbol{\theta}}(\mathbf{x})(\mu_{\boldsymbol{\theta}}(\mathbf{x}) - \mathbf{x}) + O(h^2 + \delta_{\boldsymbol{\theta}} + \delta_n),\\
E_2(\mathbf{x}) &= 2f_{\mathcal{U}}(\mathbf{x}^T\tilde{\boldsymbol{\theta}})\pi_{\boldsymbol{\theta}}(\mathbf{x})\Sigma_{\boldsymbol{\theta}}(\mathbf{x}) + O(h^2 + \delta_{\boldsymbol{\theta}} + \delta_n),
\end{aligned}$$

where $\mu_{\boldsymbol{\theta}}(\mathbf{x}) = \mathrm{E}(X|\mathbf{X}^T\boldsymbol{\theta} = \mathbf{x}^T\boldsymbol{\theta})$*,* $\nu_{\boldsymbol{\theta}}(\mathbf{x}) = \mathrm{E}(Z|\mathbf{X}^T\boldsymbol{\theta} = \mathbf{x}^T\boldsymbol{\theta})$*,* $\pi_{\boldsymbol{\theta}}(\mathbf{x}) = \mathrm{E}(\mathbf{Z}\mathbf{Z}^T|\mathbf{X}^T\boldsymbol{\theta} = \mathbf{x}^T\boldsymbol{\theta})$*,* $\Sigma_{\boldsymbol{\theta}}(\mathbf{x}) = \mathrm{E}\left((\mathbf{X} - \mu_{\boldsymbol{\theta}}(\mathbf{x}))(\mathbf{X} - \mu_{\boldsymbol{\theta}}(\mathbf{X}))^T|\mathbf{X}^T\boldsymbol{\theta} = \mathbf{x}^T\boldsymbol{\theta}\right)$*.*

Proof By the Condition 2, after some direct calculations, we can easily obtain the above conclusions. □

Lemma 4 *For the given interior point* $\mathbf{x}$ *of* $\mathbf{X}$*, then the estimates of* $\mathbf{g}(\mathbf{x}^T\tilde{\boldsymbol{\theta}})$ *and* $\mathbf{g}'(\cdot)$ *are*

$$(\hat{\mathbf{g}}(\mathbf{x}^T\tilde{\boldsymbol{\theta}}), \hat{\mathbf{g}}'(\mathbf{x}^T\tilde{\boldsymbol{\theta}})) = \underset{\mathbf{a},\mathbf{b}}{\operatorname{argmin}}\sum_{i=1}^{n}\rho_\tau\left(Y_i - \left(\mathbf{a} + \mathbf{b}\mathbf{X}_{i0}^T\tilde{\boldsymbol{\theta}}\right)^T\mathbf{Z}_i)K(\mathbf{X}_{i0}^T\tilde{\boldsymbol{\theta}}/h\right).$$

Under the conditions A.1–A.7, we have

$$\begin{aligned}
\hat{\mathbf{g}}(\mathbf{x}^T\tilde{\boldsymbol{\theta}}) &= \mathbf{g}(\mathbf{x}^T\tilde{\boldsymbol{\theta}}) + \frac{1}{2}\mathbf{g}''(\mathbf{x}^T\tilde{\boldsymbol{\theta}})\mu_2h^2 - \mathbf{g}'(\mathbf{x}^T\tilde{\boldsymbol{\theta}})\mu_{\boldsymbol{\theta}}(\mathbf{x})^T\boldsymbol{\theta}_d\\
&\quad + R_{n1}^{\tilde{\boldsymbol{\theta}}}\left(\mathbf{x}\right) + O(h^2(h^2 + \delta_{\boldsymbol{\theta}} + \delta_n) + \delta_{\boldsymbol{\theta}}^2\right),\\
\hat{\mathbf{g}}'(\mathbf{x}^T\tilde{\boldsymbol{\theta}}) &= \mathbf{g}'(\mathbf{x}^T\tilde{\boldsymbol{\theta}}) + \frac{1}{h}R_{n2}^{\tilde{\boldsymbol{\theta}}}(\mathbf{x}) + O(h^2 + \delta_n + \delta_{\boldsymbol{\theta}}),
\end{aligned}$$

where $\boldsymbol{\theta}_d = \tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}$, $\mathbf{X}_{i0} = \mathbf{X}_i - \mathbf{x}$, $\psi_\tau(u) = \tau - I(u < 0)$,

$$R_{n1}^{\boldsymbol{\theta}}(\mathbf{x}) = [nf_Y(q_\tau(\mathbf{x}, \mathbf{z})|\mathbf{x}^T\boldsymbol{\theta}) f_{\mathcal{U}}(\mathbf{x}^T\boldsymbol{\theta})]^{-1}\pi_{\boldsymbol{\theta}}(\mathbf{x})^{-1}\sum_{i=1}^{n} K_{i,h}^{\boldsymbol{\theta}}\psi_\tau(\varepsilon_i),$$

$$R_{n2}^{\boldsymbol{\theta}}(\mathbf{x}) = [nh\mu_2 f_Y(q_\tau(\mathbf{x}, \mathbf{x})|\mathbf{x}^T\boldsymbol{\theta}) f_{\mathcal{U}}(u)]^{-1}\pi_{\boldsymbol{\theta}}(\mathbf{x})^{-1}\sum_{i=1}^{n} K_{i,h}^{\boldsymbol{\theta}}\psi_\tau(\varepsilon_i)\mathbf{X}_{i0}^T\boldsymbol{\theta}.$$

In particular, $\sup_{\mathbf{x}\in\Xi}\|\hat{\mathbf{g}}'(\mathbf{x}^T\tilde{\boldsymbol{\theta}}) - \mathbf{g}'(\mathbf{x}^T\tilde{\boldsymbol{\theta}})\| = O(h^2 + h^{-1}\delta_n + \delta_{\boldsymbol{\theta}})$ *holds.*

Proof For notation simplicity, let $\mathbf{x}^T\tilde{\boldsymbol{\theta}} = u$, denote

$$\eta = \sqrt{nh}\begin{pmatrix}\mathbf{a}-\mathbf{g}(u)\\ h(\mathbf{b}-\mathbf{g}'(u))\end{pmatrix}, \hat{\eta}_n = \sqrt{nh}\begin{pmatrix}\hat{\mathbf{g}}(u)-\mathbf{g}(u)\\ h(\hat{\mathbf{g}}'(u)-\mathbf{g}'(u))\end{pmatrix}, \ M_i = \begin{pmatrix}\mathbf{Z}_i\\ \mathbf{Z}_i\mathbf{X}_{i0}^T\tilde{\boldsymbol{\theta}}/h\end{pmatrix}$$

and

$$r_i(u) = \left[-\mathbf{g}(\mathbf{X}_i^T\boldsymbol{\theta}) + \mathbf{g}(u) + \mathbf{g}'(u)\mathbf{X}_{i0}^T\tilde{\boldsymbol{\theta}}\right]^T\mathbf{Z}_i, \ K_i = K\left(\mathbf{X}_{i0}^T\tilde{\boldsymbol{\theta}}/h\right).$$

Then $\hat{\eta}_n$ is the minimizer of the following object function

$$Q_n(\eta) = \sum_{i=1}^{n}\left[\rho_\tau\left(\varepsilon_i - r_i(u) - \eta^T M_i/\sqrt{nh}\right) - \rho_\tau(\varepsilon_i - r_i(u))\right]K_i.$$

By the identify equation in Knight (1998),

$$\rho_\tau(u - v) - \rho_\tau(u) = -v\psi_\tau(u) + \int_0^v (I(u \le s) - I(u \le 0)\mathrm{d}s, \tag{16}$$

it follows that $Q_n(\eta)$ can be restated as

$$\begin{aligned} Q_n(\eta) &= \frac{1}{\sqrt{nh}}\sum_{i=1}^{n} K_i M_i\psi_\tau(\varepsilon_i) + \sum_{i=1}^{n} K_i\int_{r_i(u)}^{r_i(u)+M_i^T\eta/\sqrt{nh}} (I(\varepsilon_i \le s) - I(\varepsilon_i) \le 0))\mathrm{d}s,\\ &\equiv -\eta^T W_n + B_n(\eta), \end{aligned} \tag{17}$$

where $W_n = \frac{1}{\sqrt{nh}}\sum\limits_{i=1}^{n} K_i M_i\psi_\tau(\varepsilon_i)$,

$$B_n(\eta) = \sum_{i=1}^{n} K_i\int_{r_i(u)}^{r_i(u)+M_i^T\eta/\sqrt{nh}} [I(\varepsilon_i \le s) - I(\varepsilon_i \le 0)]\,\mathrm{d}s.$$

We next consider $B_n(\eta)$. Denote $\tilde{\mathcal{X}}$ as the σ field generated by $\{\mathbf{X}_1^T\tilde{\boldsymbol{\theta}}, \mathbf{X}_2^T\tilde{\boldsymbol{\theta}}, \ldots, \mathbf{X}_n^T\tilde{\boldsymbol{\theta}}\}$. Take the conditional expectation of $B_n(\eta)$, we have

$$\begin{aligned}
\mathrm{E}\left(B_n(\eta)|\tilde{\mathcal{X}}\right) &= \sum_{i=1}^n K_i \int_{r_i(u)}^{r_i(u)+M_i^T\eta/\sqrt{nh}} \mathrm{E}\left(I(\varepsilon_i \le s) - I(\varepsilon_i \le 0)|\mathbf{X}_i^T\tilde{\boldsymbol{\theta}}\right) \mathrm{d}s \\
&= \frac{1}{2} f_Y(q_\tau(\mathbf{x}, \mathbf{z})|u)\eta^T \left(\frac{1}{nh}\sum_{i=1}^n M_i {M_i}^T K_i\right)\eta \\
&\quad + \left(\frac{f_Y(q_\tau(\mathbf{x}, \mathbf{z})|u)}{\sqrt{nh}} \sum_{i=1}^n K_i r_i(u) M_i\right)^T \eta + o_p(1) \\
&\equiv B_{n1}(\eta) + B_{n2}(\eta) + o_p(1),
\end{aligned}$$

where $B_{n1}(\eta) = \frac{1}{2} f_Y(q_\tau(\mathbf{x}, \mathbf{z})|u)\eta^T \left(\frac{1}{nh}\sum_{i=1}^n M_i {M_i}^T K_i\right)\eta$,

$$B_{n2}(\eta) = \left(\frac{f_Y(q_\tau(\mathbf{x}, \mathbf{z})|u)}{\sqrt{nh}} \sum_{i=1}^n K_i r_i(u) M_i\right)^T \eta + o_p(1).$$

We next calculate $\mathrm{Var}(B_n(\eta)|\tilde{\mathcal{X}})$. Denote

$$\Delta_i = M_i^T\eta/\sqrt{nh} = \left[\mathbf{a} - \mathbf{g}(u) + h(\mathbf{b} - \mathbf{g}'(u))(\mathbf{X}_i^T\tilde{\theta} - u)\right]^T \mathbf{Z}_i.$$

Since

$$\begin{aligned}
\mathrm{Var}\left[B_n(\eta)|\tilde{\chi}\right] &= \sum_{i=1}^n \mathrm{Var}\left\{\left(K_i \int_{r_i(u)}^{r_i(u)+\Delta_i} \left[I\left\{\varepsilon_i \le s\right\} - I\left\{\varepsilon \le 0\right\}\right] \mathrm{d}s\right)|\tilde{\chi}\right\} \\
&= \sum_{i=1}^n \mathrm{Var}\left\{\left(K_i \int_0^{\Delta_i} \left[I\left\{\varepsilon_i \le r_i(u) + t\right\} - I\left\{\varepsilon \le r_i(u)\right\}\right] \mathrm{d}t\right)|\tilde{\chi}\right\} \\
&\le \sum_{i=1}^n \mathrm{E}\left[\left(K_i \int_0^{\Delta_i} \left[I\left\{\varepsilon_i \le r_i(u) + t\right\} - I\left\{\varepsilon \le r_i(u)\right\}\right] \mathrm{d}t\right)^2 |\tilde{\chi}\right] \\
&\le \sum_{i=1}^n K_i^2 \int_0^{|\Delta_i|} \int_0^{|\Delta_i|} \left[F(r_i(u) + |\Delta_i|) - F(r_i(u))\right] \mathrm{d}v_1 \mathrm{d}v_2 \\
&= o\left(\sum_{i=1}^n K_i^2 \Delta_i^2\right) = o_p(1).
\end{aligned}$$

Therefore, we have $\mathrm{Var}(B_n(\eta)|\tilde{\mathcal{X}}) = o(1)$, and it follows that

$$B_n(\eta) = B_{n1}(\eta) + B_{n2}(\eta) + o_p(1). \tag{18}$$

Denote $\mathbb{S}_n = \frac{1}{nh} f_Y(q_\tau(\mathbf{x}, \mathbf{z})|u) \sum_{i=1}^{n} M_i {M_i}^T K_i$. By the above Lemma 3, it is easy to prove $\mathbb{S}_n = \mathbb{S} + O_p(\tau_n + \delta_{\boldsymbol{\theta}})$, where

$$\mathbb{S} = f_Y(q_\tau(\mathbf{x}, \mathbf{z})|u) f_{\mathcal{U}}(u) \mathrm{E}(\mathbf{Z}\mathbf{Z}^T|\mathbf{X}^T\boldsymbol{\theta}) \otimes dia\mathbf{g}(1, \mu_2),$$

and $A \otimes B$ denotes the Kronecker product of two matrixes.

Combining the above results, we have

$$Q_{n1}(\eta) = \frac{1}{2}\eta^T \mathbb{S} \eta + o_p(1). \tag{19}$$

Now we begin to consider $B_{n2}(\eta)$. Note that

$$r_i(u) = \left(\mathbf{X}_i^T \boldsymbol{\theta}_d \mathbf{g}\left(\mathbf{X}_i^T \tilde{\boldsymbol{\theta}}\right) - \frac{1}{2}\mathbf{g}''(u) \left(\mathbf{X}_{i0}^T \tilde{\boldsymbol{\theta}}\right)^2 + O\left(\boldsymbol{\theta}_d^2 + \left(\mathbf{X}_{i0}^T \tilde{\boldsymbol{\theta}}\right)^3\right)\right)^T \mathbf{Z}_i,$$

hence it follows that

$$\begin{aligned} \frac{1}{\sqrt{nh}} \sum_{i=1}^{n} f_Y(q_\tau(\mathbf{x}, \mathbf{z})|u) \mathbf{Z}_i K_i r_i(u) &= \sqrt{nh} \mathrm{E}(\mathbf{Z}\mathbf{Z}^T|\mathbf{X}^T\boldsymbol{\theta}) f_Y(q_\tau(\mathbf{x}, \mathbf{z})|u) f_{\mathcal{U}}(u) \\ &\quad \times \bigg(\mathbf{g}'(u)\mu_{\boldsymbol{\theta}}(\mathbf{x})^T \boldsymbol{\theta}_d - \frac{1}{2}\mathbf{g}''(u)\mu_2 h^2 \\ &\quad + O(h^4 + \delta_{\boldsymbol{\theta}}^2 + h^2\delta_{\boldsymbol{\theta}}) \bigg), \end{aligned}$$

$$\frac{1}{\sqrt{nh}} \sum_{i=1}^{n} f_Y(q_\tau(\mathbf{x}, \mathbf{z})|u) K_i \frac{\mathbf{X}_{i0}^T \tilde{\boldsymbol{\theta}}}{h} \mathbf{Z}_i r_i(u) = \sqrt{nh}[O(h^3 + h\delta_{\boldsymbol{\theta}})]. \tag{20}$$

Combining the results from (17), (18), (19) and (20), we have

$$\begin{aligned} Q_n(\eta) = \frac{1}{2}\eta^T \mathbb{S} \eta - W_n^T \eta + \sqrt{nh} f_Y(q_\tau(\mathbf{x}, \mathbf{z})|u) f_{\mathcal{U}}(u) \\ \times \begin{pmatrix} \mathrm{E}(\mathbf{Z}\mathbf{Z}^T|\mathbf{X}^T\boldsymbol{\theta})\left[\mathbf{g}'(u)\mu_{\boldsymbol{\theta}}(\mathbf{x})^T\boldsymbol{\theta}_d - \frac{1}{2}\mathbf{g}''(u)\mu_2 h^2 + O(h^4 + \delta_{\boldsymbol{\theta}}^2 + h^2\delta_{\boldsymbol{\theta}})\right] \\ O(h^3 + h\delta_{\boldsymbol{\theta}}) \end{pmatrix}^T \eta + o_p(1). \end{aligned}$$

By the result of (1), the minimizer of $Q_n(\eta)$ can be expressed as

$$\hat{\eta}_n = \mathbb{S}^{-1} W_n - \sqrt{nh} \begin{pmatrix} \mathbf{g}'(u)\mu_{\boldsymbol{\theta}}(\mathbf{x})^T\boldsymbol{\theta}_d - \frac{1}{2}\mathbf{g}''(u)\mu_2 h^2 + O(\delta_{\boldsymbol{\theta}}^2 + (h^2 + \delta_{\boldsymbol{\theta}})\tau_n) \\ O(h^3 + h\delta_{\boldsymbol{\theta}}) \end{pmatrix} + o_p(1).$$

According to the definition of $\hat{\eta}_n$ and W_n, the result of the first part follows. Meanwhile, by the Lemma 2, the second part also follows. □

Proof of Theorem 1 Given the estimates $\hat{\mathbf{g}}(\mathbf{X}_j^T\tilde{\boldsymbol{\theta}}),\ \hat{\mathbf{g}}'(\mathbf{X}_j^T\tilde{\boldsymbol{\theta}})$ of $\mathbf{g}(\mathbf{X}_j^T\tilde{\boldsymbol{\theta}})$ and $\mathbf{g}'(\mathbf{X}_j^T\tilde{\boldsymbol{\theta}})$, $j = 1, \ldots, n$, by (6), the estimate $\boldsymbol{\theta}$ can be obtained as

$$\hat{\boldsymbol{\theta}} = \underset{\|\boldsymbol{\theta}\|=1,\boldsymbol{\theta}_1>0}{\operatorname{argmin}} \sum_{j=1}^n \sum_{i=1}^n \rho_\tau \left(Y_i - [\hat{\mathbf{g}}(\mathbf{X}_j^T\tilde{\boldsymbol{\theta}}) + \hat{\mathbf{g}}'(\mathbf{X}_j^T\tilde{\boldsymbol{\theta}})\mathbf{X}_{ij}^T\boldsymbol{\theta}]^T\mathbf{Z}_i\right)\omega_{ij}.$$

Denote $\tilde{U}_i = \mathbf{X}_i^T\tilde{\boldsymbol{\theta}},\ \tilde{U}_j = \mathbf{X}_j^T\tilde{\boldsymbol{\theta}}$. Let

$$\hat{\boldsymbol{\theta}}^* = \sqrt{n}\left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\right),\ M_{ij} = \mathbf{Z}_i^T\hat{\mathbf{g}}'(\tilde{U}_j)\mathbf{X}_{ij},$$
$$r_{ij} = \left(-\mathbf{g}(\mathbf{X}_i^T\boldsymbol{\theta}) + \hat{\mathbf{g}}(\tilde{U}_j) + \hat{\mathbf{g}}'(\tilde{U}_j)\mathbf{X}_{ij}\boldsymbol{\theta}\right)^T\mathbf{Z}_i,$$

then $\hat{\boldsymbol{\theta}}^*$ is the minimizer of

$$\mathcal{Q}_n(\boldsymbol{\theta}^*) = \sum_{j=1}^n\sum_{i=1}^n \omega_{ij}\left[\rho_\tau\left(\varepsilon_i - r_{ij} - M_{ij}^T\boldsymbol{\theta}^*/\sqrt{n}\right) - \rho_\tau(\varepsilon_i - r_{ij})\right].$$

By Knight (1998) identify Eq. (16), we can rewritten $\mathcal{Q}_n(\boldsymbol{\theta}^*)$ as

$$\begin{aligned}\mathcal{Q}_n(\boldsymbol{\theta}^*) &= -\frac{1}{\sqrt{n}}\sum_{j=1}^n\sum_{i=1}^n \omega_{ij}\psi_\tau(\varepsilon_i)M_{ij}^T\boldsymbol{\theta}^* \\ &\quad + \sum_{j=1}^n\sum_{i=1}^n \omega_{ij}\int_{r_{ij}}^{r_{ij}+M_{ij}^T\boldsymbol{\theta}^*/\sqrt{n}} [I(\varepsilon_i \le s) - I(\varepsilon_i \le 0)]\mathrm{d}s \\ &\equiv \mathcal{Q}_{1n}(\boldsymbol{\theta}^*) + \mathcal{Q}_{2n}(\boldsymbol{\theta}^*),\end{aligned}$$

where $\mathcal{Q}_{1n}(\boldsymbol{\theta}^*) = -\frac{1}{\sqrt{n}}\sum_{j=1}^n\sum_{i=1}^n \omega_{ij}\psi_\tau(\varepsilon_i)M_{ij}^T\boldsymbol{\theta}^*$,

$$\mathcal{Q}_{2n}(\boldsymbol{\theta}^*) = \sum_{j=1}^n\sum_{i=1}^n \omega_{ij}\int_{r_{ij}}^{r_{ij}+M_{ij}^T\boldsymbol{\theta}^*/\sqrt{n}}(I(\varepsilon_i \le s) - I(\varepsilon_i \le 0))\mathrm{d}s.$$

Firstly, we consider the conditional expectation of $\mathcal{Q}_{2n}(\boldsymbol{\theta}^*)$ on $\tilde{\mathcal{X}}$. By directly calculating, we have

$$\begin{aligned}\mathrm{E}\left(\mathcal{Q}_{2n}(\boldsymbol{\theta}^*)|\tilde{\mathcal{X}}\right) &= \sum_{j=1}^n\sum_{i=1}^n\int_{r_{ij}}^{r_{ij}+M_{ij}^T\boldsymbol{\theta}^*/\sqrt{n}} \omega_{ij}\left[sf_Y(q_\tau(\mathbf{X}_i,\mathbf{Z}_i)|\tilde{U}_i)(1+o(1))\right]\mathrm{d}s \\ &= \frac{1}{2}\boldsymbol{\theta}^{*T}\left(\frac{1}{n}\sum_{j=1}^n\sum_{i=1}^n f_Y(q_\tau(\mathbf{X}_i,\mathbf{Z}_i)|\tilde{U}_i)M_{ij}M_{ij}^T\omega_{ij}\right)\boldsymbol{\theta}^*\end{aligned}$$

$$+\left(\frac{1}{\sqrt{n}}\sum_{j=1}^{n}\sum_{i=1}^{n}\omega_{ij}f_Y(q_\tau(\mathbf{X}_i,\mathbf{Z}_i)|\tilde{U}_i)r_{ij}M_{ij}\right)^T\boldsymbol{\theta}^*+o_p(1)$$
$$\equiv\mathcal{Q}_{2n1}(\boldsymbol{\theta}^*)+\mathcal{Q}_{2n2}(\boldsymbol{\theta}^*)+o_p(1),$$

where $\mathcal{Q}_{2n1}(\boldsymbol{\theta}^*)=\frac{1}{2}\boldsymbol{\theta}^{*T}\left(\frac{1}{n}\sum_{j=1}^{n}\sum_{i=1}^{n}\omega_{ij}f_Y(q_\tau(\mathbf{X}_i,\mathbf{Z}_i)|\tilde{U}_i)M_{ij}M_{ij}^T\right)\boldsymbol{\theta}^*,$

$$\mathcal{Q}_{2n2}(\boldsymbol{\theta}^*)=\left(\frac{1}{\sqrt{n}}\sum_{j=1}^{n}\sum_{i=1}^{n}\omega_{ij}f_Y(q_\tau(\mathbf{X}_i,\mathbf{Z}_i)|\tilde{U}_i)M_{ij}r_{ij}\right)^T\boldsymbol{\theta}^*+o_p(1).$$

Denote $\mathcal{R}_n(\boldsymbol{\theta}^*)=\mathcal{Q}_{2n}(\boldsymbol{\theta}^*)-\mathrm{E}(\mathcal{Q}_{2n}(\boldsymbol{\theta}^*)|\tilde{\mathcal{X}})$. It is easy to obtain $\mathcal{R}_n(\boldsymbol{\theta}^*)=o_p(1)$, then we have $\mathcal{Q}_{2n}(\boldsymbol{\theta}^*)=\mathcal{Q}_{2n1}(\boldsymbol{\theta}^*)+\mathcal{Q}_{2n2}(\boldsymbol{\theta}^*)+o_p(1)$.

Next, we consider $\mathcal{Q}_{2n1}(\boldsymbol{\theta}^*)$ and $\mathcal{Q}_{2n2}(\boldsymbol{\theta}^*)$, respectively. For $\mathcal{Q}_{2n1}(\boldsymbol{\theta}^*)$, let

$$\mathcal{G}_n^{\tilde{\boldsymbol{\theta}}}=\frac{1}{n}\sum_{j=1}^{n}\sum_{i=1}^{n}f_Y(q_\tau(\mathbf{X}_i,\mathbf{Z}_i)|\tilde{U}_i)M_{ij}M_{ij}^T\omega_{ij}.$$

By the Lemma 2, it is easy to have $\mathcal{G}_n^{\tilde{\boldsymbol{\theta}}}=2\mathcal{G}+O(h^2+\delta_n+\delta_{\boldsymbol{\theta}})$, where the definition of $\mathcal{G}$ can be seen in Theorem 1.

Denote $W_{\boldsymbol{\theta}}(\mathbf{x})=\mathrm{E}(f_Y(q_\tau(\mathbf{X},\mathbf{Z})|\mathbf{X}^T\boldsymbol{\theta})\mathbf{Z}\mathbf{Z}^T|\mathbf{X}^T\boldsymbol{\theta}=\mathbf{x}^T\boldsymbol{\theta})$, then

$$\mathcal{Q}_{2n1}(\boldsymbol{\theta}^*)=\frac{1}{2}\boldsymbol{\theta}^{*T}\mathcal{G}\boldsymbol{\theta}^*+o_p(1). \tag{21}$$

For $\mathcal{Q}_{2n2}(\boldsymbol{\theta}^*)$, note that

$$r_{ij}=\mathbf{Z}_i^T\left(\mathbf{g}'(\tilde{U}_i)\mathbf{X}_i^T\boldsymbol{\theta}_d-\frac{1}{2}\mathbf{g}''(\tilde{U}_j)\left(\mathbf{X}_{ij}^T\tilde{\boldsymbol{\theta}}\right)^2-\hat{\mathbf{g}}'(\tilde{U}_j)\mathbf{X}_{ij}^T\boldsymbol{\theta}_d\right)$$
$$+\mathbf{Z}_i^T\left(\hat{\mathbf{g}}(\tilde{U}_j)-\mathbf{g}(\tilde{U}_j)+(\hat{\mathbf{g}}'(\tilde{U}_j)-\mathbf{g}'(\tilde{U}_j))\mathbf{X}_{ij}^T\tilde{\boldsymbol{\theta}}+O\left(\boldsymbol{\theta}_d^2+\left(\mathbf{X}_{ij}^T\tilde{\boldsymbol{\theta}}\right)^3\right)\right).$$

Hence, we obtain

$$\mathcal{Q}_{2n2}(\boldsymbol{\theta}^*)=\frac{1}{\sqrt{n}}\sum_{j=1}^{n}\sum_{i=1}^{n}f_Y(q_\tau(\mathbf{X}_i,\mathbf{Z}_i)|\tilde{U}_i)\omega_{ij}M_{ij}^T\boldsymbol{\theta}^*(\mathbf{Z}_i^T,\mathbf{Z}_i^T\mathbf{X}_{ij}^T\tilde{\boldsymbol{\theta}}/h)\begin{pmatrix}\hat{\mathbf{g}}(\tilde{U}_j)-\mathbf{g}(\tilde{U}_j)\\ h(\hat{\mathbf{g}}'(\tilde{U}_j)-\mathbf{g}'(\tilde{U}_j))\end{pmatrix}$$
$$+\frac{1}{\sqrt{n}}\sum_{j=1}^{n}\sum_{i=1}^{n}f_Y(q_\tau(\mathbf{X}_i,\mathbf{Z}_i)|\tilde{U}_i)\omega_{ij}M_{ij}^T\boldsymbol{\theta}^*\mathbf{Z}_i^T$$
$$\times\left(\mathbf{g}'(\tilde{U}_i)\mathbf{X}_i^T\boldsymbol{\theta}_d-\hat{\mathbf{g}}'(\tilde{U}_j)\mathbf{X}_{ij}^T\boldsymbol{\theta}_d-\frac{1}{2}\mathbf{g}''(\tilde{U}_j)(\mathbf{X}_{ij}^T\tilde{\boldsymbol{\theta}})^2\right)$$
$$\equiv(\mathcal{Q}_{2n21}+\mathcal{Q}_{2n22})^T\boldsymbol{\theta}^*+O(\delta_{\boldsymbol{\theta}}^2+h^3),$$

where

$$
\begin{aligned}
\mathcal{Q}_{2n21} &= \frac{1}{\sqrt{n}} \sum_{j=1}^{n} \sum_{i=1}^{n} f_Y(q_\tau(\mathbf{X}_i, \mathbf{Z}_i)|\tilde{U}_i)\omega_{ij} M_{ij} \left(\mathbf{Z}_i^T, \mathbf{Z}_i^T \mathbf{X}_{ij}^T \tilde{\boldsymbol{\theta}}/h\right) \left(\begin{smallmatrix} \hat{\mathbf{g}}(\tilde{U}_j)-\mathbf{g}(\tilde{U}_j) \\ h(\hat{\mathbf{g}}'(\tilde{U}_j)-\mathbf{g}'(\tilde{U}_j)) \end{smallmatrix}\right), \\
\mathcal{Q}_{2n22} &= \frac{1}{\sqrt{n}} \sum_{j=1}^{n} \sum_{i=1}^{n} f_Y(q_\tau(\mathbf{X}_i, \mathbf{Z}_i)|\tilde{U}_i)\omega_{ij} M_{ij} \mathbf{Z}_i^T, \\
&\quad \times \left(\mathbf{g}'(\tilde{U}_i)\mathbf{X}_i^T \boldsymbol{\theta}_d - \hat{\mathbf{g}}'(\tilde{U}_j)\mathbf{X}_{ij}^T \boldsymbol{\theta}_d - \frac{1}{2}\mathbf{g}''(\tilde{U}_j)(\mathbf{X}_{ij}^T \tilde{\boldsymbol{\theta}})^2\right).
\end{aligned}
$$

Now, we begin to consider $\mathcal{Q}_{2n21}$ and $\mathcal{Q}_{2nn2}$. By the asymptotic expressions $\hat{\mathbf{g}}(\mathbf{x}^T \tilde{\boldsymbol{\theta}})$ and $\hat{\mathbf{g}}'(\mathbf{X}^T \tilde{\boldsymbol{\theta}})$ obtained in Lemma 4, we have

$$
\begin{aligned}
\mathcal{Q}_{2n21} &= \frac{1}{\sqrt{n}} \sum_{j=1}^{n} \sum_{i=1}^{n} \omega_{ij} f_Y(q_\tau(\mathbf{X}_i, \mathbf{Z}_i)|\tilde{U}_i) M_{ij} \left(\mathbf{Z}_i^T, \mathbf{X}_{ij}^T \tilde{\boldsymbol{\theta}} \mathbf{Z}_i^T\right) \left(\begin{smallmatrix} R_{n1}^{\tilde{\boldsymbol{\theta}}}(\mathbf{X}_j) \\ R_{n2}^{\tilde{\boldsymbol{\theta}}}(\mathbf{X}_j) \end{smallmatrix}\right) \\
&\quad + \frac{1}{\sqrt{n}} \sum_{j=1}^{n} \sum_{i=1}^{n} \omega_{ij} f_Y(q_\tau(\mathbf{X}_i, \mathbf{Z}_i)|\tilde{U}_i) M_{ij} \mathbf{Z}_i^T \\
&\quad \times \left(\frac{1}{2}\mathbf{g}''(\tilde{U}_j)\mu_2 h^2 - \mathbf{g}'(\tilde{U}_j)\mu_{\boldsymbol{\theta}}(\mathbf{X}_j)^T \boldsymbol{\theta}_d\right) \\
&\quad + O_p((h^2 + \delta_{\boldsymbol{\theta}})\tau_n + \delta_{\boldsymbol{\theta}}^2 + h^3 + h\delta_{\boldsymbol{\theta}}) \\
&\equiv T_1 + T_2 + o_p(1),
\end{aligned}
$$

where

$$
\begin{aligned}
T_1 &= \frac{1}{\sqrt{n}} \sum_{j=1}^{n} \sum_{i=1}^{n} \omega_{ij} f_Y(q_\tau(\mathbf{X}_i, \mathbf{Z}_i)|\tilde{U}_i) M_{ij} \left(\mathbf{Z}_i^T, \mathbf{X}_{ij}^T \tilde{\boldsymbol{\theta}} \mathbf{Z}_i^T\right) \left(\begin{smallmatrix} R_{n1}^{\tilde{\boldsymbol{\theta}}}(\mathbf{X}_j) \\ R_{n2}^{\tilde{\boldsymbol{\theta}}}(\mathbf{X}_j) \end{smallmatrix}\right), \\
T_2 &= \frac{1}{\sqrt{n}} \sum_{j=1}^{n} \sum_{i=1}^{n} \omega_{ij} f_Y(q_\tau(\mathbf{X}_i, \mathbf{Z}_i)|\tilde{U}_i) M_{ij} \mathbf{Z}_i^T \\
&\quad \times \left(\frac{1}{2}\mathbf{g}''(\tilde{U}_j)\mu_2 h^2 - \mathbf{g}'(\tilde{U}_j)\mu_{\boldsymbol{\theta}}(\mathbf{X}_j)^T \boldsymbol{\theta}_d\right).
\end{aligned}
$$

By directly calculating, it follows that

$$
\begin{aligned}
T_1 &= \frac{1}{\sqrt{n}} \sum_{j=1}^{n} \sum_{i=1}^{n} \frac{\omega_{ij} f_Y(q_\tau(\mathbf{X}_i, \mathbf{Z}_i)|\tilde{U}_i)}{n f_{\mathcal{U}}(\tilde{U}_j) f_Y(q_\tau(\mathbf{X}_j, \mathbf{Z}_j)|\tilde{U}_j)} M_{ij} \left(\mathbf{Z}_i^T,\ \mathbf{Z}_i^T \frac{\mathbf{X}_{ij}^T \tilde{\boldsymbol{\theta}}}{h}\right) W_{\tilde{\boldsymbol{\theta}}}(\mathbf{X}_j)^{-1} \\
&\quad \times \sum_{k=1}^{n} \left(\begin{matrix} \mathbf{Z}_k \\ \mathbf{Z}_k \frac{\mathbf{X}_{kj}^T \tilde{\boldsymbol{\theta}}}{h} \end{matrix}\right) K_h\left(\mathbf{X}_{kj}^T \tilde{\boldsymbol{\theta}}\right) \psi_\tau(\varepsilon_k) \\
&= \frac{1}{\sqrt{n}} \sum_{k=1}^{n} \sum_{j=1}^{n} \psi_\tau(\varepsilon_k)\omega_{kj}[\mu_{\boldsymbol{\theta}}(\mathbf{X}_j) - \mathbf{X}_j]\mathbf{g}'(\tilde{U}_j)^T \mathbf{Z}_k + o_p(1).
\end{aligned}
$$

Combining T_1 and $\mathcal{Q}_{1n}(\boldsymbol{\theta}^*)$, we have

$$\mathcal{Q}_{1n}(\boldsymbol{\theta}^*)+T_1^T\boldsymbol{\theta}^*=\left[-\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\sum_{j=1}^{n}\psi_\tau(\varepsilon_i)\omega_{ij}\hat{\mathbf{g}}'(\tilde{U}_j)^T\mathbf{Z}_i\left(\mathbf{X}_i-\mu_{\boldsymbol{\theta}}(\mathbf{X}_j)\right)\right]^T\boldsymbol{\theta}^*+o_p(1)$$
$$=-\sqrt{n}\mathcal{W}_n^T\boldsymbol{\theta}^*+o_p(1), \qquad (22)$$

where $\mathcal{W}_n=\frac{1}{\sqrt{n}}\sum\limits_{i=1}^{n}\sum\limits_{j=1}^{n}\psi_\tau(\varepsilon_i)\omega_{ij}\hat{\mathbf{g}}'(\tilde{U}_j)^T\mathbf{Z}_i\left[\mathbf{X}_i-\mu_{\boldsymbol{\theta}}(\mathbf{X}_j)\right]$. By the Lemma 2, we obtain

$$\mathcal{W}_n=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\psi_\tau(\varepsilon_i)\mathbf{g}'(\tilde{U}_i)^T\mathbf{Z}_i(\mathbf{X}_i-\mu_{\boldsymbol{\theta}}(\mathbf{X}_i)). \qquad (23)$$

According to the Cramér–Wald device and the central limit theorem, we have

$$\mathcal{W}_n\xrightarrow{\mathcal{L}}N(0,\tau(1-\tau)\mathcal{G}_0), \qquad (24)$$

where the definition of $\mathcal{G}_0$ is given in Theorem 1.

Merging T_2 and $\mathcal{Q}_{2n22}$, we obtain

$$\mathcal{Q}_{2n22}+T_2=\frac{1}{\sqrt{n}}\sum_{j=1}^{n}\sum_{i=1}^{n}f_Y(q_\tau(\mathbf{X}_j,\mathbf{Z}_j)|\tilde{U}_j)\omega_{ij}M_{ij}\mathbf{Z}_i^T\Big[\mathbf{g}'(\tilde{U}_i)\mathbf{X}_i^T\boldsymbol{\theta}_d-\hat{\mathbf{g}}'(\tilde{U}_j)\mathbf{X}_{ij}^T\boldsymbol{\theta}_d$$
$$-\frac{1}{2}\mathbf{g}''(\tilde{U}_j)\left(\mathbf{X}_{ij}^T\tilde{\boldsymbol{\theta}}\right)^2+\frac{1}{2}\mathbf{g}''(\tilde{U}_j)\mu_2h^2-\mathbf{g}'(\tilde{U}_j)\mu_{\boldsymbol{\theta}}(\mathbf{X}_j)^T\boldsymbol{\theta}_d\Big]+o_p(1)$$
$$=\frac{1}{\sqrt{n}}\sum_{j=1}^{n}\mathbf{g}'(\tilde{U}_j)^TW_{\tilde{\boldsymbol{\theta}}}(\mathbf{X}_j)\mathbf{g}'(\tilde{U}_j)\left(\mu_{\boldsymbol{\theta}}(\mathbf{X}_j)-\mathbf{X}_j^T\right)\left(\mu_{\boldsymbol{\theta}}(\mathbf{X}_j)-\mathbf{X}_j^T\right)^T\boldsymbol{\theta}_d$$
$$+o_p(1).$$

By Lemmas 2 and 3, it is easy to obtain

$$\mathcal{Q}_{2n22}+T_2=-\sqrt{n}\mathcal{G}\boldsymbol{\theta}_d+o_p(1). \qquad (25)$$

Therefore, by (21), (22) and (25), we have

$$\mathcal{Q}_n(\boldsymbol{\theta}^*)=\boldsymbol{\theta}^{*T}\mathcal{G}\boldsymbol{\theta}^*-\left[\mathcal{W}_n+\sqrt{n}\mathcal{G}\boldsymbol{\theta}_d\right]^T\boldsymbol{\theta}^*+o_p(1).$$

By the Lemma 1, the minimizer $\hat{\boldsymbol{\theta}}^*$ of $\mathcal{Q}_n(\boldsymbol{\theta}^*)$ can be written as $\hat{\boldsymbol{\theta}}^*=\frac{1}{2}\mathcal{G}^{-1}\mathcal{W}_n+\frac{1}{2}\sqrt{n}\boldsymbol{\theta}_d+o_p(1)$. Note that $\hat{\boldsymbol{\theta}}^*=\sqrt{n}\left(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}\right)$, then we have

$$\left(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}\right)=\frac{1}{2}\mathcal{G}^{-1}\frac{1}{\sqrt{n}}\mathcal{W}_n+\frac{1}{2}\left(\tilde{\boldsymbol{\theta}}-\boldsymbol{\theta}\right)+o_p(1/\sqrt{n}). \qquad (26)$$

The convergence of the estimate algorithm can be followed by the above equation.

Define $\tilde{\theta}_k$ as the kth estimate, $\forall k$, the Eq. (26) still satisfies if we replace $\tilde{\boldsymbol{\theta}}$ and $\hat{\boldsymbol{\theta}}$ as $\tilde{\theta}_k$ and $\tilde{\theta}_{k+1}$, respectively. Therefore, for the sufficiently large k, we have $\hat{\boldsymbol{\theta}} - \boldsymbol{\theta} = \mathcal{G}^{-1}\frac{1}{\sqrt{n}}\mathcal{W}_n + \frac{1}{2}\left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\right) + o(1/\sqrt{n})$. Then

$$\hat{\boldsymbol{\theta}} - \boldsymbol{\theta} = \mathcal{G}^{-1}\frac{1}{\sqrt{n}}\mathcal{W}_n + o(1/\sqrt{n}).$$

Combining the above result in (24), we complete the proof of Theorem 1. □

Lemma 5 *Suppose u is an inner point of the tight support of $f_{\mathcal{U}}(\cdot)$, and the conditions A.1–A.7 in appendix hold, then we have*

$$\sqrt{nh}\left\{\hat{\mathbf{g}}(u; h, \hat{\boldsymbol{\theta}}) - \mathbf{g}(u) - \frac{1}{2}\mathbf{g}''(u)\mu_2 h^2\right\} \xrightarrow{\mathcal{L}} N(0, \Gamma_\tau(u)), \tag{27}$$

where $\Gamma_\tau(\cdot)$ is defined in Theorem 2.

Proof of Lemma 5 When the parameter $\boldsymbol{\theta}$ is known, for the given interior point $u = \mathbf{x}^T\boldsymbol{\theta}$ of $\mathcal{U}$, denote $R_{n1}^{\boldsymbol{\theta}}(\mathbf{x})$ as $R_{n1}^{\boldsymbol{\theta}}$. By the similar proof as Theorem 4, the estimate of $g(u)$ can be written as

$$\hat{\mathbf{g}}(u; h, \boldsymbol{\theta}) = \mathbf{g}(u) + \frac{1}{2}\mathbf{g}''(u)\mu_2 h^2 + R_{n1}^{\boldsymbol{\theta}} + O(h^3).$$

By the central limit theorem, it is easy to prove

$$\sqrt{nh}\left(\hat{\mathbf{g}}(u; h, \boldsymbol{\theta}) - \mathbf{g}(u) - \frac{1}{2}\mathbf{g}''(u)\mu_2 h^2\right) \xrightarrow{\mathcal{L}} N(0, \Gamma(u)).$$

By the Lemma 4, we consider the difference between the two estimate

$$\begin{aligned}\hat{\mathbf{g}}(u; h, \tilde{\boldsymbol{\theta}}) - \hat{\mathbf{g}}(u; h, \boldsymbol{\theta}) = &-\mathrm{E}(X|\mathbf{X}^T\boldsymbol{\theta} = u)^T\boldsymbol{\theta}_d - \mathrm{E}(Z|\mathbf{X}^T\boldsymbol{\theta} = u)^T \\ &+ R_{n1}^{\tilde{\boldsymbol{\theta}}} - R_{n1}^{\boldsymbol{\theta}} + O(\delta_{\boldsymbol{\theta}} + h\delta_n + h^3).\end{aligned}$$

Since $\boldsymbol{\theta}_d = O_p(1/\sqrt{n})$, we only need to prove

$$\sqrt{nh}\left(R_{n1}^{\tilde{\boldsymbol{\theta}}} - R_{n1}^{\boldsymbol{\theta}}\right) = o_p(1). \tag{28}$$

When the bandwidth h satisfies $nh^4 \to \infty$, since $\boldsymbol{\theta}_d = O_p(1/\sqrt{n})$, by directly calculating, we have

$$\begin{aligned}\mathrm{Var}\left[\sqrt{nh}\left(R_{n,1}^{\hat{\boldsymbol{\theta}}} - R_{n,1}^{\boldsymbol{\theta}}\right)\right] &\leq (\tau - \tau^2)\mathrm{E}\left[K_h(\mathbf{X}^T\boldsymbol{\theta} - u) - K_h(\mathbf{X}^T\hat{\boldsymbol{\theta}} - u)\right]^2 \\ &= (\tau - \tau^2)\int\left(K(t) - K(t + \mathbf{X}^T\boldsymbol{\theta}_d/h)\right)^2 f(u + ht)\mathrm{d}t \\ &\leq \int\frac{1}{4}K'(t^*)^2(\mathbf{X}^T\boldsymbol{\theta}_d/h)^2 f(u + ht)\mathrm{d}t = O\left(\frac{1}{nh^2}\right) = o(1).\end{aligned}$$

Therefore (28) holds and the proof of the Lemma 5 is completed. □

Proof of Theorem 2 Given the interior $\mathbf{x}$ of Ξ, we have

$$\begin{aligned}&(nh)^{1/2}[\hat{\mathbf{g}}(\mathbf{x}^T\hat{\boldsymbol{\theta}}; h, \hat{\boldsymbol{\theta}}) - \mathbf{g}(\mathbf{x}^T\boldsymbol{\theta})]\\ &\quad = (nh)^{1/2}[\hat{\mathbf{g}}(\mathbf{x}^T\hat{\boldsymbol{\theta}}; h, \hat{\boldsymbol{\theta}}) - \hat{\mathbf{g}}(\mathbf{x}^T\boldsymbol{\theta}; h, \hat{\boldsymbol{\theta}}) + \hat{\mathbf{g}}(\mathbf{x}^T\boldsymbol{\theta}; h, \hat{\boldsymbol{\theta}}) - \mathbf{g}(\mathbf{x}^T\boldsymbol{\theta})]\\ &\quad = E + (nh)^{1/2}[\hat{\mathbf{g}}(\mathbf{x}^T\boldsymbol{\theta}; h, \hat{\boldsymbol{\theta}}) - \mathbf{g}(\mathbf{x}^T\boldsymbol{\theta})].\end{aligned}$$

By Taylor expansion,

$$E = \sqrt{nh}[\hat{\mathbf{g}}(\mathbf{x}^T\hat{\boldsymbol{\theta}}; h, \hat{\boldsymbol{\theta}}) - \hat{\mathbf{g}}(\mathbf{x}^T\boldsymbol{\theta}; h, \hat{\boldsymbol{\theta}})] = \sqrt{nh}\hat{\mathbf{g}}'(\mathbf{x}^T\boldsymbol{\theta})O_p(\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\|) = o_p(1).$$

By the result of Lemma 5, we can conclude that Theorem 2 holds. □

Proof of Theorem 3 For convenience, redefine $\mathbf{u} = \sqrt{n}(\hat{\boldsymbol{\theta}}^{\lambda} - \boldsymbol{\theta})$, $\hat{\boldsymbol{\theta}}_d = \hat{\boldsymbol{\theta}}^{QR} - \boldsymbol{\theta}$, where $\hat{\boldsymbol{\theta}}^{QR}$ is the estimate of $\boldsymbol{\theta}$ in Theorem 1. Then, $\mathbf{u}$ is the minimizer of the following object function:

$$\begin{aligned}G_n(\mathbf{u}) = &\sum_{j=1}^{n}\sum_{i=1}^{n}\omega_{ij}\left(\rho_\tau\left(\varepsilon_i + r_{ij} + M_{ij}^T\mathbf{u}/\sqrt{n}\right) - \rho_\tau(\varepsilon_i + r_{ij})\right)\\ &+ \sum_{k=1}^{p}\frac{\lambda}{\sqrt{n}|\hat{\theta}_k^{QR}|^2}\sqrt{n}\left[\left|\theta_k + \frac{u_k}{\sqrt{n}}\right| - |\theta_k|\right].\end{aligned}$$

Similar to the proof of Theorem 1, we can write $G_n(\mathbf{u})$ as:

$$\begin{aligned}G_n(\mathbf{u}) = &\frac{1}{2}\mathbf{u}^T\mathcal{G}\mathbf{u} - \mathcal{W}_n^T\mathbf{u} + \sqrt{n}\hat{\boldsymbol{\theta}}_d^T C_0^T\mathbf{u} + o_p(1)\\ &+ \sum_{k=1}^{p}\frac{\lambda}{\sqrt{n}|\hat{\theta}_k^{QR}|^2}\sqrt{n}\left[\left|\theta_k + \frac{u_k}{\sqrt{n}}\right| - |\theta_k|\right].\end{aligned}$$

For $1 \le k \le p_0$, $\theta_k \neq 0$, we have $|\hat{\theta}_k^{QR}|^2 \to_p |\theta_k|^2$, and $\sqrt{n}(|\theta_k + u_k/\sqrt{n}| - |\theta_k|) \to u_k\text{sgn}(\theta_k)$. By the Slutsky's Theorem, $\frac{\lambda}{\sqrt{n}|\hat{\theta}_k^{QR}|^2}\sqrt{n}(|\theta_k + u_k/\sqrt{n}| - |\theta_k|) \to_p 0$.

For $p_0 < k \le p$, $\theta_k = 0$, then we have $\sqrt{n}(|\theta_k + u_k/\sqrt{n}| - |\theta_k|) \to_p \infty$. Therefore, we have

$$\frac{\lambda}{\sqrt{n}|\hat{\theta}_k^{QR}|^2}\sqrt{n}\left[\left|\theta_k + \frac{u_k}{\sqrt{n}}\right| - |\theta_k|\right] \to_p W(\theta_k, u_k) = \begin{cases}0, & \text{if } \theta_k \neq 0,\\ 0, & \text{if } \theta_k = 0 \text{ and } u_k = 0,\\ \infty, & \text{if } \theta_k = 0 \text{ and } u_k \neq 0.\end{cases}$$

For $\boldsymbol{\theta} = \binom{\boldsymbol{\theta}^1}{\boldsymbol{\theta}^2}$, denote $\mathbf{u} = \binom{\mathbf{u}_1}{\mathbf{u}_2}$, we have

$$\begin{aligned} G_n(\mathbf{u}) &\rightarrow \frac{1}{2}\mathbf{u}^T \mathcal{G}\mathbf{u} - W_n^T \mathbf{u} + \left(\hat{\boldsymbol{\theta}}_d^T, \hat{\boldsymbol{\beta}}_d^T\right) C_0^T \mathbf{u} + \sum_{j=1}^{p} W(\theta_j, u_j) + o_p(1) \\ &\rightarrow L(\mathbf{u}) = \begin{cases} \frac{1}{2}\mathbf{u}^T \mathcal{G}\mathbf{u} - W_n^T \mathbf{u} + \hat{\boldsymbol{\theta}}_d^T C_0^T \mathbf{u}, & \text{if } \mathbf{u}_2 = 0 \\ \infty, & \text{otherwise.} \end{cases} \end{aligned}$$

Note that $G_n(\mathbf{u})$ is convex about u, and $L(\mathbf{u})$ has unique minimal solution. By the epi-convergence result Geyer (1994), we can obtain the asymptotic normality by following the proof of Theorem 1.

Next, we consider the convergence of the model selection. Note that the form of two formulas $G_n(\mathbf{u})$ and $L(\mathbf{u})$ are similar to Zou (2006), and by the condition A.8, $\mathcal{G}$ is positive definite; hence, we can easily obtain the model consistency by following the idea of Zou (2006). □

References

Cai, Z., Xu, X. (2008). Nonparametric quantile estimations for dynamic smooth coefficient models. *Journal of the American Statistical Association*, *103*, 1595–1608.

Fan, J., Li, R. (2001). Variable selection via non-concave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, *96*, 1348–1360.

Fan, J., Yao, Q., Cai, Z. (2003). Adaptive varying-coefficient linear models. *Journal of the Royal Statistical Society Series B (Statistical Methodology)*, *65*, 57–80.

Feng, S., Xue, L. (2013). Variable selection for single-index varying-coefficient model. *Frontiers of Mathematics in China*, *8*, 541–565.

Geyer, C. J. (1994). On the asymptotics of constrained m-estimation. *The Annals of Statistics*, *22*, 1993–2010.

Härdle, W., Hall, P., Ichimura, H. (1993). Optimal smoothing in single-index models. *The Annals of Statistics*, *21*, 157–178.

Hjort, N., Pollard, D. (1993). *Asymptotics for minimizers of convex processes* (preprint).

Honda, T. (2004). Quantile regression in varying coefficient models. *Journal of Statistical Planning and Inference*, *121*, 113–125.

Huang, Z., Zhang, R. (2013). Profile empirical-likelihood inferences for the single-index-coefficient regression model. *Statistics and Computing*, *23*, 455–465.

Jiang, R., Zhou, Z., Qian, W., Shao, W. (2012). Single-index composite quantile regression. *Journal of the Korean Statistical Society*, *41*, 323–332.

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

Kim, M.-O. (2007). Quantile regression with varying coefficients. *The Annals of Statistics*, *35*, 92–108.

Knight, K. (1998). Limiting distributions for l_1 regression estimators under general conditions. *The Annals of Statistics*, *26*, 755–770.

Koenker, R., Basset, G. S. (1978). Regression quantiles. *Econometrica*, *46*, 33–50.

Lu, Z., Tjøstheim, D., Yao, Q. (2007). Adaptive varying-coefficient linear models for stochastic processes: Asymptotic theory. *Statistica Sinica*, *17*, 177–197.

Mack, Y. P., Silverman, B. W. (1982). Weak and strong uniform consistency of kernel regression estimates. *Probability Theory and Related Fields*, *61*, 405–415.

Shapiro, S. S., Wilk, M. B. (1965). An analysis of variance test for normality (complete samples). *Biometrika*, *52*, 591–611.

Wang, H., Leng, C. (2007). Unified lasso estimation via least squares approximation. *Journal of the American Statistical Association*, *102*, 1039–1048.

Wu, T., Yu, K., Yu, Y. (2010). Single-index quantile regression. *Journal of Multivariate Analysis*, *101*, 1607–1621.

Xia, Y., Tong, H., Li, W. K. (1999). On extended partially linear single-index models. *Biometrika*, *86*, 831–842.

Xue, L., Pang, Z. (2013). Statistical inference for a single-index varying-coefficient model. *Statistics and Computing*, *23*, 589–599.

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

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