Ann Inst Stat Math (2018) 70:553–582
https://doi.org/10.1007/s10463-017-0599-8

Adaptive varying-coefficient linear quantile model: a profiled estimating equations approach

Weihua Zhao[1] · **Jianbo Li**[2] · **Heng Lian**[3]

Received: 22 January 2016 / Revised: 30 December 2016 / Published online: 20 February 2017
© The Institute of Statistical Mathematics, Tokyo 2017

Abstract We consider an estimating equations approach to parameter estimation in adaptive varying-coefficient linear quantile model. We propose estimating equations for the index vector of the model in which the unknown nonparametric functions are estimated by minimizing the check loss function, resulting in a profiled approach. The estimating equations have a bias-corrected form that makes undersmoothing of the nonparametric part unnecessary. The estimating equations approach makes it possible to obtain the estimates using a simple fixed-point algorithm. We establish asymptotic properties of the estimator using empirical process theory, with additional complication due to the nuisance nonparametric part. The finite sample performance of the new model is illustrated using simulation studies and a forest fire dataset.

Keywords Asymptotic normality · Bias-corrected estimating equations · Check loss · Empirical processes · Single-index model

1 Introduction

The varying-coefficient model (VCM) has gained much attention in the literature. Hastie and Tibshirani (1993) and Chen and Tsay (1993) are the two seminal works on the VCM for cross-sectional data and time series data, respectively. Fan and Zhang (1999) proposed a two-step local linear estimator in the VCM which achieves uni-

✉ Weihua Zhao
weihuazhao66@hotmail.com

1 School of Science, Nantong University, Nantong 226019, P. R. China

2 School of Mathematics and Statistics, Jiangsu Normal University, Xuzhou 221116, P. R. China

3 Department of Mathematics, City University of Hong Kong, Kowloon Tong, Hong Kong

variate optimal convergence rate. The VCM also has wide applications in longitudinal studies; see Hoover et al. (1998), Fan and Zhang (2000) and Huang et al. (2002). Recently, the VCM has been applied to high-dimensional data settings with large p; see Wei et al. (2011), Lian (2012b) and Xue and Qu (2012) for the penalization method and Fan et al. (2014b); Liu et al. (2014) for the independence screening method.

We consider the following varying-coefficient model,

$$Y_i = g_0(\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}) + \sum_{j=1}^{p} g_j(\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta})X_{ij} + e_i, \tag{1}$$

where $\mathbf{X}_i = (X_{i1}, \ldots, X_{ip})^{\mathrm{T}}$ are the covariates, Y_i is the response, $g_j, 0 \leq j \leq p$ are $p+1$ unknown coefficient functions and $\boldsymbol{\beta}$ is an unknown p-vector, henceforth referred to as the index vector. Fan et al. (2003) considered mean regression for the model (1) with $E[e_i|\mathbf{X}_i] = 0$ and termed it the adaptive varying-coefficient linear model. The adaptiveness comes from that the coefficients are functions of $\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}$, with $\boldsymbol{\beta}$ estimated from data, in contrast to the standard varying-coefficient model where the coefficient functions vary with some directly observed variable(s). The adaptive varying-coefficient linear model avoids the curse of dimensionality for multivariate data since only one-dimensional functions are being estimated. This model is also similar to that proposed in Xia and Li (1999). Fan et al. (2003) proposed a new computationally efficient procedure based on the profile least-squares local linear regression. Large sample properties of the model are later established by Lu et al. (2007).

Model (1) was previously considered for estimating the conditional mean of the response. In this paper, we are instead interested in the conditional quantile, by assuming $P(e_i \leq 0|\mathbf{X}_i) = \tau$ for some $\tau \in (0, 1)$. Quantile regression, since its introduction by Koenker and Bassett Jr (1978) in a celebrated Econometrica paper, has been intensively studied in both the econometrics and the statistics literature. Both parametric and nonparametric modeling in quantile regression have been investigated (Bondell et al. 2010; Koenker et al. 1994; Reich et al. 2010; Wu and Liu 2009; Yu and Jones 1998). Semiparametric quantile models have also attracted much attention. For example, Kim (2007) considered varying-coefficient quantile regression and Wang et al. (2009) extended this to partially linear varying coefficient models, both using polynomial spline. Cai and Xiao (2012) investigated the partially linear varying-coefficient models for time series data. Lee (2003) proposed efficient estimation method for partially linear quantile regression. Horowitz and Lee (2005) used a two-stage estimation procedure to achieve oracle efficiency for additive quantile regression models, using polynomial splines in the first stage and kernel estimator in the second. Lian (2012a) also considered additive models with a focus on model selection. Bayesian methods are also popular, including Hu et al. (2013); Kottas and Krnjajic (2009); Tokdar and Kadane (2011); Yang and He (2012); Yu and Moyeed (2001). Quantile regression in high-dimensional situations is studied in Belloni and Chernozhukov (2011); Fan et al. (2014a); Wang et al. (2012); Zhu et al. (2012).

In this paper, we consider the estimation problem for (1) assuming the conditional τ-quantile of Y_i is $g_0(\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}) + \sum_{j=1}^{p} g_j(\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta})X_{ij}$. This semiparametric model can be regarded as an extension of quantile varying-coefficient regression model with the coefficient varying with an unknown linear combination of the covariates. It can also be regarded as an extension of quantile single-index models, for which Wu et al. (2010) and Kong and Xia (2012) were the first to study the associated estimation problem using kernel-based methods. However, to our knowledge quantile version of (1) has not been studied. As shown in Lu et al. (2007), establishing asymptotic properties of adaptive varying-coefficient linear models is nontrivial even for mean regression. For the coefficient functions, here we use polynomial splines as our estimation method. The advantage of using regression splines largely resides in its computational expediency, with all coefficient functions estimated simultaneously, while for local polynomial-based method, one needs to choose a grid for the support of $\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}$ and the coefficient function values at each point on the grid are estimated separately. However, our main intent is not to promote the splines. Estimation procedure for quantile varying-coefficient linear model could also be developed using local polynomial regression. Both approaches have their advantages and disadvantages.

Conceptually, we treat the nonparametric coefficient functions as the nuisance and propose estimating equations for the estimation of $\boldsymbol{\beta}$, with the unknown coefficient functions replaced by their estimates for a given $\boldsymbol{\beta}$. In this respect, we are using a profiled procedure for constructing the estimating equations. This is the same as the approach used in Cui et al. (2011) for generalized single-index models. Due to profiling, the estimating equations have a bias-corrected form that resembles estimating equations used in Cui et al. (2011); Lai et al. (2012); Wang et al. (2010); Zhu et al. (2010). The bias-corrected form results from the profile approach and is the reason that no undersmoothing for the nonparametric functions is necessary, as discussed in Carroll et al. (1997). However, unlike those estimating equations mentioned in those papers, the quantile version here is nonsmooth, and thus, we heavily rely on empirical process theory to derive its asymptotic properties.

The rest of the article is organized as follows. In Sect. 2, we formally derive the estimating equations and define the estimation as an approximate solution to the estimating equations. We also establish the asymptotic normality of the estimator for the index vector. Section 3 contains numerical experiments including simulation and a real data application. We conclude with some discussions in Sect. 4. The technical details for establishing the asymptotic properties are relegated to Appendix.

2 Methods

2.1 Derivation of estimating equations

Given independent and identically distributed data $(\mathbf{X}_i, Y_i)$, $i = 1, \ldots, n$, we write (1) more succinctly as

$$Y_i = \mathbf{g}^{\mathrm{T}}(\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta})\mathbf{Z}_i + e_i,$$

with $\mathbf{g} = (g_0, \ldots, g_p)^{\mathrm{T}}$ and $\mathbf{Z}_i = (1, \mathbf{X}_i^{\mathrm{T}})^{\mathrm{T}}$. e_i is the error whose conditional τth quantile is 0, that is, $P(e_i \leq 0|\mathbf{X}_i) = \tau$. For identifiability, we assume $\|\boldsymbol{\beta}\| = 1$ with $\beta_1 > 0$.

We first motivate the estimating equations based on calculations in the population. Given $\boldsymbol{\beta}$, let $\mathbf{g}(u; \boldsymbol{\beta})$ be the minimizer of

$$\min_{\mathbf{g}\in R^p} E\left[\rho_\tau(Y - \mathbf{g}^{\mathrm{T}}\mathbf{Z})|\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta} = u\right],$$

where $\rho_\tau(x) = x(\tau - I\{x \le 0\})$ is the check loss function. Since $g(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}; \boldsymbol{\beta})$ and its derivatives appear often below, it is important to distinguish between the different partial derivatives. We use $g'(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}; \boldsymbol{\beta})$ to denote $\partial g(u; \boldsymbol{\beta})/\partial u|_{u=\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}}$, use $\partial g(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}; \boldsymbol{\beta})/\partial\boldsymbol{\beta}$ to denote $\partial g(u; \boldsymbol{\beta})/\partial\boldsymbol{\beta}|_{u=\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}}$ and use $dg(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}; \boldsymbol{\beta})/d\boldsymbol{\beta}$ to denote $g'(\mathbf{X}^{T}\boldsymbol{\beta}; \boldsymbol{\beta})\mathbf{X} + \partial g(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}; \boldsymbol{\beta})/\partial\boldsymbol{\beta}$ which results from the chain rule.

We now derive the estimating equations. For now, we ignore the constraint $\|\boldsymbol{\beta}\| = 1$. In the population, the first-order condition for minimizing the profiled functional over $\boldsymbol{\beta}$, $\min_{\boldsymbol{\beta}} E\left[\rho_\tau(Y - \mathbf{g}^{\mathrm{T}}(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}; \boldsymbol{\beta})\mathbf{Z})\right]$, is given by

$$E\left[\left(\mathbf{X}(\mathbf{g}'(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}; \boldsymbol{\beta}))^{\mathrm{T}}\mathbf{Z} + \frac{\partial\mathbf{g}^{\mathrm{T}}(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}; \boldsymbol{\beta})}{\partial\boldsymbol{\beta}}\mathbf{Z}\right)\psi_\tau(Y - \mathbf{g}^{\mathrm{T}}(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}; \boldsymbol{\beta})\mathbf{Z})\right] = \mathbf{0}, \quad (2)$$

where $\psi_\tau(x) = \tau - I\{x \le 0\}$. To derive the estimating equations, we need a more concrete expression for $\frac{\partial\mathbf{g}^{\mathrm{T}}(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}; \boldsymbol{\beta})}{\partial\boldsymbol{\beta}}$. Since $\mathbf{g}(u; \boldsymbol{\beta})$ is the minimizer of

$$\min_{\mathbf{g}\in R^{p+1}} E\left[\rho_\tau(Y - \mathbf{g}^{\mathrm{T}}\mathbf{Z})|\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta} = u\right], \quad (3)$$

we have

$$E\left[\mathbf{Z}\psi_\tau(Y - \mathbf{g}(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}; \boldsymbol{\beta})^{\mathrm{T}}\mathbf{Z})|\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}\right] \equiv \mathbf{0} \quad (4)$$

for all $\boldsymbol{\beta}$. We take the derivative of (4) with respect to $\boldsymbol{\beta}$ at $\boldsymbol{\beta}_0$. Using that $\partial E[\psi_\tau(Y - g)|\mathbf{X}]/\partial g|_{g=\mathbf{g}^{\mathrm{T}}(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta};\boldsymbol{\beta})\mathbf{Z}} = -f(\mathbf{g}^{\mathrm{T}}(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}; \boldsymbol{\beta})\mathbf{Z}|\mathbf{X})$, where $f(.|\mathbf{X})$ is the conditional density of Y given $\mathbf{X}$, we get

$$E\left[\mathbf{Z}f(\mathbf{g}^{\mathrm{T}}(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}; \boldsymbol{\beta})\mathbf{Z}|\mathbf{X})\mathbf{Z}^{\mathrm{T}}\mathbf{g}'(\mathbf{X}\boldsymbol{\beta}; \boldsymbol{\beta})\mathbf{X}^{\mathrm{T}} + \mathbf{Z}f(\mathbf{g}^{\mathrm{T}}(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}; \boldsymbol{\beta})\mathbf{Z}|\mathbf{X})\mathbf{Z}^{\mathrm{T}}\frac{\partial\mathbf{g}}{\partial\boldsymbol{\beta}^{\mathrm{T}}}|\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}\right] = \mathbf{0} \quad (5)$$

at $\boldsymbol{\beta} = \boldsymbol{\beta}_0$. Equation (5) is derived using the chain rule, and the fact that at $\boldsymbol{\beta}_0$, we have

$$\begin{aligned}
&\frac{\partial E\left[\mathbf{Z}\psi_\tau(Y - \mathbf{g}(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}_0; \boldsymbol{\beta}_0)^{\mathrm{T}}\mathbf{Z})|\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}\right]}{\partial\boldsymbol{\beta}} \\
&= \frac{\partial E\left[E\left[\mathbf{Z}\psi_\tau(e)|\mathbf{X}\right]|\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}\right]}{\partial\boldsymbol{\beta}} \\
&= \mathbf{0},
\end{aligned}$$

since $E[\mathbf{Z}\psi_\tau(e)|\mathbf{X}] = 0$. Thus by (5) we have

$$-\frac{\partial \mathbf{g}^{\mathrm{T}}(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta};\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = E\left[\mathbf{X}\mathbf{Z}^{\mathrm{T}}\mathbf{g}'(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta};\boldsymbol{\beta}) f(\mathbf{g}^{\mathrm{T}}(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta};\boldsymbol{\beta})\mathbf{Z}|\mathbf{X})\mathbf{Z}^{\mathrm{T}}|\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}\right] \times \left\{E[\mathbf{Z}f(\mathbf{g}^{\mathrm{T}}(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta};\boldsymbol{\beta})\mathbf{Z}|\mathbf{X})\mathbf{Z}^{\mathrm{T}}|\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}]\right\}^{-1}.$$

It is also easy to see that the right-hand size above is the minimizer of

$$\min_{\mathbf{H}\in R^{p\times(p+1)}} E\left[f(\mathbf{g}^{\mathrm{T}}(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta};\boldsymbol{\beta})\mathbf{Z}|\mathbf{X})\|\mathbf{X}\mathbf{Z}^{\mathrm{T}}\mathbf{g}'(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta};\boldsymbol{\beta}) - \mathbf{H}\mathbf{Z}\|^2|\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}\right],$$

which can be written as $\mathbf{H}(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta};\boldsymbol{\beta})$ to make the dependence explicit. This can be interpreted as a kind of projection in some space of functions as we elucidate now. Such projection is often used in semiparametric models to "orthogonalize" the parametric part and the nonparametric part. Let $\mathcal{M}_{\boldsymbol{\beta}} = \{m : m(\mathbf{x}) = \mathbf{h}^{\mathrm{T}}(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta})\mathbf{z}, Em^2(\mathbf{X}) < \infty\}$ be the space of functions taking the adaptive varying-coefficient form with fixed $\boldsymbol{\beta}$. For any random variable W, $E[W|\mathcal{M}_{\boldsymbol{\beta}}]$ denotes the projection of W on $\mathcal{M}_{\boldsymbol{\beta}}$ in the sense that $E[W|\mathcal{M}_{\boldsymbol{\beta}}]$ is the minimizer of $E\left[f(\mathbf{g}^{\mathrm{T}}(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta};\boldsymbol{\beta})\mathbf{Z}|\mathbf{X})(W - m(\mathbf{X}))^2\right]$ over $m \in \mathcal{M}_{\boldsymbol{\beta}}$. This definition can be extended to the case $\mathbf{W} = (W_1, \ldots, W_p)^{\mathrm{T}}$ is a random vector, by $E[\mathbf{W}|\mathcal{M}_{\boldsymbol{\beta}}] = \left(E[W_1|\mathcal{M}_{\boldsymbol{\beta}}], \ldots, E[W_p|\mathcal{M}_{\boldsymbol{\beta}}]\right)^{\mathrm{T}}$.

By our previous definition, $-\frac{\partial \mathbf{g}^{\mathrm{T}}(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta};\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}\mathbf{Z}$ can also be written as $E\left[\mathbf{X}\mathbf{Z}^{\mathrm{T}}\mathbf{g}'(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta};\boldsymbol{\beta})|\mathcal{M}_{\boldsymbol{\beta}}\right]$. Plugging into (2), we get the estimating equations (EE)

$$\left(\mathbf{X}\mathbf{Z}^{\mathrm{T}}\mathbf{g}'(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta};\boldsymbol{\beta}) - E\left[\mathbf{X}\mathbf{Z}^{\mathrm{T}}\mathbf{g}'(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta};\boldsymbol{\beta})|\mathcal{M}_{\boldsymbol{\beta}}\right]\right)\psi_\tau\left(Y - \mathbf{g}^{\mathrm{T}}(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta};\boldsymbol{\beta})\mathbf{Z}\right) = \mathbf{0}.$$

We note in passing that if $\mathbf{g}$ were known, the estimating equations would be naturally defined as

$$\mathbf{X}\mathbf{Z}^{\mathrm{T}}\mathbf{g}'(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta};\boldsymbol{\beta})\psi_\tau(Y - \mathbf{g}^{\mathrm{T}}(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta};\boldsymbol{\beta})) = \mathbf{0},$$

and thus our EE can be regarded as a bias-corrected EE taking into account that $\mathbf{g}$ is unknown.

The parameter space $\mathcal{B} = \{\boldsymbol{\beta} : \|\boldsymbol{\beta}\| = 1, \beta_1 > 0\}$ is a compact set. So far in the derivation we have ignored the constraint $\|\boldsymbol{\beta}\| = 1$. To take into account this, we use the delete-one-component method of Cui et al. (2011); Yu and Ruppert (2002). Assuming $\beta_1 > 0$, we can write $\boldsymbol{\beta} = ((1-\|\boldsymbol{\beta}^{(-1)}\|^2)^{1/2}, \beta_2, \ldots, \beta_p)^{\mathrm{T}}$ where $\boldsymbol{\beta}^{(-1)} = (\beta_2, \ldots, \beta_p)^{\mathrm{T}}$ is $\boldsymbol{\beta}$ without the first component. Thus $\boldsymbol{\beta}$ is a function of $\boldsymbol{\beta}^{(-1)}$. The $p \times (p-1)$ Jacobian matrix is

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

where $\mathbf{I}_{(p-1)\times(p-1)}$ is the $(p-1)\times(p-1)$ identity matrix. Note $\mathbf{J}$ is actually a function of $\boldsymbol{\beta}$. Thus finally, our estimating equations are given by

$$\mathbf{J}^{\mathrm{T}}\left(\mathbf{X}\mathbf{Z}^{\mathrm{T}}\mathbf{g}'(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta};\boldsymbol{\beta}) - E[\mathbf{X}\mathbf{Z}^{\mathrm{T}}\mathbf{g}'(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta};\boldsymbol{\beta})|\mathcal{M}_{\boldsymbol{\beta}}]\right)\psi_\tau\left(Y - \mathbf{g}^{\mathrm{T}}(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta};\boldsymbol{\beta})\right) = \mathbf{0}.$$

The delete-one-component method is only used in the derivation of the estimating equations and theoretical proofs, while for computation we use the fixed-point algorithm as explained later.

2.2 Estimation and asymptotics

Quantile regression is typically performed by minimizing

$$\sum_i \rho_\tau\left(Y_i - \mathbf{g}^{\mathrm{T}}(\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta})\mathbf{Z}_i\right).$$

Since $\mathbf{g}$ is unknown, we approximate the $p+1$ functions using B-splines. We assume that $\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}$ are contained in some interval $[a, b]$ (in practice the interval boundary can be set to be the minimum and maximum value of $\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}$ based on the current estimates). Let $t_0 = a < t_1 < \cdots < t_{K'} < b = t_{K'+1}$ be a partition of $[a, b]$ into subintervals $[t_k, t_{k+1}), k = 0, \ldots, K'$ with K' internal knots. We only restrict our attention to equally spaced knots although data-driven choice can be considered such as putting knots at certain sample quantiles of the observed values. 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 $\{B_1(x), \ldots, B_K(x)\}$ with $K = K' + s$. We assume the basis functions are normalized to have $\sum_k B_k(x) \equiv \sqrt{K}$. This is not essential but simplifies some of the expressions in the proof. With the spline basis defined, we can approximate $\mathbf{g}(x) \approx \boldsymbol{\Theta}^{\mathrm{T}}\mathbf{B}(x)$ where $\boldsymbol{\Theta} = (\boldsymbol{\theta}_0, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_p)$ is a $K \times (p+1)$ matrix of spline coefficients and we denote $\mathbf{B}(x) = (B_1(x), \ldots, B_K(x))^{\mathrm{T}}$.

Given i.i.d. observations, the functions $\mathbf{g}(u;\boldsymbol{\beta})$ can be estimated by $\widehat{\mathbf{g}}(u;\boldsymbol{\beta}) = \mathbf{B}^{\mathrm{T}}(u)\widehat{\boldsymbol{\Theta}}(\boldsymbol{\beta})$ where $\widehat{\boldsymbol{\Theta}}(\boldsymbol{\beta})$ is the minimizer of

$$\min_{\boldsymbol{\Theta}} \sum_i \rho_\tau\left(Y_i - \mathbf{B}^{\mathrm{T}}(\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta})\boldsymbol{\Theta}\mathbf{Z}_i\right),$$

for a given $\boldsymbol{\beta}$, and $\mathbf{g}'(u;\boldsymbol{\beta})$ is estimated by $\widehat{\mathbf{g}}'(u;\boldsymbol{\beta})$, the derivative of $\widehat{\mathbf{g}}(u;\boldsymbol{\beta})$ with respect to u. We similarly estimate $\mathbf{H}(u,\boldsymbol{\beta})$ by

$$\widehat{\mathbf{H}}(u;\boldsymbol{\beta}) = \widehat{\boldsymbol{\Gamma}}\left(\mathbf{I}\otimes\mathbf{B}(u)\right), \tag{6}$$

where $\otimes$ denotes the Kronecker product and $\widehat{\boldsymbol{\Gamma}}$ is the minimizer of

$$\sum_i f\left(\mathbf{g}^{\mathrm{T}}(\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta};\boldsymbol{\beta})\mathbf{Z}_i|\mathbf{X}_i\right)\|\mathbf{X}_i\mathbf{Z}_i^{\mathrm{T}}\widehat{\mathbf{g}}'(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta};\boldsymbol{\beta}) - \boldsymbol{\Gamma}\left(\mathbf{Z}_i\otimes\mathbf{B}(\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta})\right)\|^2,$$

where $\boldsymbol{\Gamma} \in R^{p\times(p+1)K}$ (this is same as minimizing over $\boldsymbol{\Gamma}$ row-by-row). We assume that the density $f(\mathbf{g}^{\mathrm{T}}(\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta};\boldsymbol{\beta})\mathbf{Z}_i|\mathbf{X}_i)$ is known in our theoretical studies. The theory still holds if $f(\mathbf{g}^{\mathrm{T}}(\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta};\boldsymbol{\beta})\mathbf{Z}_i|\mathbf{X}_i)$ can be consistently estimated (Wang et al. 2009).

Finally, we obtain the following sample-based estimating equations for $\boldsymbol{\beta}$:

$$\Phi_n(\boldsymbol{\beta};\widehat{\mathbf{m}}) = \frac{1}{n}\sum_i \mathbf{J}^{\mathrm{T}}\left(\mathbf{X}_i\mathbf{Z}_i^{\mathrm{T}}\widehat{\mathbf{g}}'(\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta};\boldsymbol{\beta}) - \widehat{\mathbf{H}}(\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta};\boldsymbol{\beta})\mathbf{Z}_i\right) \times\psi_\tau\left(Y_i - \widehat{\mathbf{g}}^{\mathrm{T}}(\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta};\boldsymbol{\beta})\mathbf{Z}_i\right), \tag{7}$$

where $\widehat{\mathbf{m}} = (\widehat{\mathbf{g}}, \widehat{\mathbf{H}})$ denotes all nonparametric functions that are estimated. $\widehat{\mathbf{m}}$ implicitly depends on $\boldsymbol{\beta}$, and thus, sometimes we will write $\widehat{\mathbf{m}}(.;\widehat{\boldsymbol{\beta}})$ to make the dependence explicit. Similarly we can write $\mathbf{m} = \mathbf{m}(.,\boldsymbol{\beta}) = (\mathbf{g}(.;\boldsymbol{\beta}), \mathbf{H}(.;\boldsymbol{\beta}))$.

Since $\Phi_n(\boldsymbol{\beta};\widehat{\mathbf{m}})$ is a discontinuous function of $\boldsymbol{\beta}$ (inherited from the discontinuity of $\psi_\tau(.)$), we formally define the estimator as any $\widehat{\boldsymbol{\beta}}$ that satisfies (the infimum may not be achieved by any $\boldsymbol{\beta}$ due to discontinuity)

$$\|\Phi_n(\widehat{\boldsymbol{\beta}};\widehat{\mathbf{m}})\| \le \inf_{\boldsymbol{\beta}} \|\Phi_n(\boldsymbol{\beta};\widehat{\mathbf{m}}(.;\boldsymbol{\beta}))\| + o_p\left(n^{-1/2}\right). \tag{8}$$

We note here that we always implicitly think of $\boldsymbol{\beta}$ as a function of $\boldsymbol{\beta}^{(-1)}$. For example, (8) is regarded as minimization over $\boldsymbol{\beta}^{(-1)}$. Mathematically this is the same as constrained minimization over $\mathcal{B}$. Let

$$\phi_{\boldsymbol{\beta},\mathbf{m}} = \mathbf{J}^{\mathrm{T}}\left(\mathbf{X}\mathbf{Z}^{\mathrm{T}}\mathbf{g}'(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta};\boldsymbol{\beta}) - \mathbf{H}(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta};\boldsymbol{\beta})\mathbf{Z}\right)\psi_\tau\left(Y - \mathbf{g}^{\mathrm{T}}(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta};\boldsymbol{\beta})\mathbf{Z}\right),$$

and $\Phi(\boldsymbol{\beta};\mathbf{m}) = E\phi_{\boldsymbol{\beta},\mathbf{m}}$.

Now we define appropriate classes of smooth functions. For $\alpha > 0$, we consider $\mathcal{C}^\alpha(M)$, the class of univariate functions on a bounded set that possess uniformly bounded derivatives up to order $\underline{\alpha}$ (the greatest integer strictly smaller than α) and whose $\underline{\alpha}$th derivative is Lipschitz of order $\alpha - \underline{\alpha}$. More specifically, let $f^{(k)}$ be the kth derivative of f and $\|f\|_\alpha = \max_{k\le\underline{\alpha}}\sup_x|f^{(k)}(x)| + \sup_{x,y}\frac{|f^{(\underline{\alpha})}(x)-f^{(\underline{\alpha})}(y)|}{|x-y|^{\alpha-\underline{\alpha}}}$; then, $\mathcal{C}^\alpha(M) = \{f : \|f\|_\alpha \le M\}$.

We impose the following assumptions.

(A1) $\mathbf{X} \in R^p$ has a bounded joint density supported on a bounded convex set.

(A2) The space of parameters is $\mathcal{B} = \{\boldsymbol{\beta} : \|\boldsymbol{\beta}\| = 1, \beta_1 > 0\}$. Uniformly in $\mathbf{X}$ and $\boldsymbol{\beta}$, the conditional density of Y at $\mathbf{g}^{\mathrm{T}}(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta};\boldsymbol{\beta})\mathbf{Z}$ is bounded and bounded away from zero, and the derivative of the conditional density at $\mathbf{g}^{\mathrm{T}}(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta};\boldsymbol{\beta})\mathbf{Z}$ is bounded.

(A3) As functions of u, entries of $\mathbf{g}(u;\boldsymbol{\beta})$ are in $\mathcal{C}^\alpha(M)$ for some $\alpha > 5/2$ and $M > 0$, and entries of $\mathbf{H}(u;\boldsymbol{\beta})$ are in $\mathcal{C}^{\alpha'}(M)$ for some $\alpha' > 1/2$. Entries of $\mathbf{g}(u;\boldsymbol{\beta})$, $\mathbf{g}'(u;\boldsymbol{\beta})$, $\mathbf{H}(u;\boldsymbol{\beta})$ are continuous functions of $\boldsymbol{\beta}$.

(A4) The true parameter $\boldsymbol{\beta}_0$ is the unique minimizer of $\|\Phi(\boldsymbol{\beta},\mathbf{m}(.;\boldsymbol{\beta}))\|$. For any $\epsilon > 0$, we have $\inf_{\|\boldsymbol{\beta}-\boldsymbol{\beta}_0\|\ge\epsilon}\|\Phi(\boldsymbol{\beta},\mathbf{m}(.;\boldsymbol{\beta}))\| - \|\Phi(\boldsymbol{\beta}_0,\mathbf{m}(.;\boldsymbol{\beta}_0))\| > 0$.

(A5) The matrix $E\left[\left(\mathbf{J}^{\mathrm{T}}\mathbf{X}\mathbf{Z}^{\mathrm{T}}\mathbf{g}'(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta};\boldsymbol{\beta}) - \mathbf{J}^{\mathrm{T}}E[\mathbf{X}\mathbf{Z}^{\mathrm{T}}\mathbf{g}'(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta};\boldsymbol{\beta})|\mathcal{M}_{\boldsymbol{\beta}}]\right)^{\otimes 2}\right]$ has eigenvalues bounded and bounded away from zero, uniformly over $\boldsymbol{\beta} \in \mathcal{B}$, where $\mathbf{A}^{\otimes 2} = \mathbf{A}\mathbf{A}^{\mathrm{T}}$ for any matrix $\mathbf{A}$.

(A6) Suppose the number of basis functions is set to be $K \sim n^{1/(2\alpha+1)}$ for estimation of $\mathbf{g}$, while $K \sim n^{1/(2\alpha'+1)}$ for the estimation of $\mathbf{H}$.

Assumption (A1) guarantees that $\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}$ is supported on a bounded interval, which makes it suitable to use the polynomial splines for estimation. Assumption (A2) on conditional density is commonly used in quantile regression (He and Shi 1994; Wang et al. 2009). Smoothness condition (A3) is used in Proposition 1 (in Appendix) which is in turn used in our proof of consistency and asymptotic normality. Both (A4) and (A5) are identifiability assumptions. The choice of the number of basis functions is the theoretically optimal one according to the different smoothness assumptions for $\mathbf{g}$ and $\mathbf{H}$, respectively. However, to simplify notation, we denote both basis functions by $\mathbf{B} = (B_1, \ldots, B_K)^{\mathrm{T}}$ for simplicity of notation.

The following is the main result that establishes the asymptotic normality of $\widehat{\boldsymbol{\beta}}$.

Theorem 1 *Under assumptions* (A1)–(A6) *we have*

$$\sqrt{n}(\widehat{\boldsymbol{\beta}}^{(-1)} - \boldsymbol{\beta}_0^{(-1)}) \xrightarrow{d} N(\mathbf{0}, \boldsymbol{\Psi}_1^{-1}\boldsymbol{\Psi}_2\boldsymbol{\Psi}_1^{-1}),$$

where $\boldsymbol{\Psi}_1 = E\left[f(\mathbf{g}^{\mathrm{T}}(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}_0)\mathbf{Z}|\mathbf{X})\left(\mathbf{J}^{\mathrm{T}}\mathbf{X}\mathbf{Z}^{\mathrm{T}}\mathbf{g}'(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}_0) - E\left[\mathbf{J}^{\mathrm{T}}\mathbf{X}\mathbf{Z}^{\mathrm{T}}\mathbf{g}'(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}_0)\mathbf{Z}|\mathcal{M}_{\boldsymbol{\beta}_0}\right]\right)^{\otimes 2}\right]$ *and* $\boldsymbol{\Psi}_2 = \tau(1-\tau)E\left[\left(\mathbf{J}^{\mathrm{T}}\mathbf{X}\mathbf{Z}^{\mathrm{T}}\mathbf{g}'(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}_0) - E\left[\mathbf{J}^{\mathrm{T}}\mathbf{X}\mathbf{Z}^{\mathrm{T}}\mathbf{g}'(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}_0)\mathbf{Z}|\mathcal{M}_{\boldsymbol{\beta}_0}\right]\right)^{\otimes 2}\right]$, *in which* $\mathbf{J}$ *is evaluated at* $\boldsymbol{\beta}_0$. *By the Delta method, as an immediate corollary,*

$$\sqrt{n}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \xrightarrow{d} N(\mathbf{0}, \mathbf{J}\boldsymbol{\Psi}_1^{-1}\boldsymbol{\Psi}_2\boldsymbol{\Psi}_1^{-1}\mathbf{J}^{\mathrm{T}}),$$

In the proof of the theorem, we make use of Proposition 1, which also directly gives as a corollary the rate of convergence of the unknown coefficient functions $\mathbf{g}$.

Corollary 1 *Under assumptions* (A1)–(A6), *we have*

$$\left\|\widehat{\mathbf{g}}(.;\widehat{\boldsymbol{\beta}}) - \mathbf{g}(.)\right\|^2 := \sum_{j=0}^{p}\|g_j(.;\widehat{\boldsymbol{\beta}}) - g_j(.)\|^2 = O_p\left(n^{-2\alpha/(2\alpha+1)}(rm\log n)\right),$$

where $\|.\|$ *denotes the* L_2 *norm of functions, as well as* l_2 *norm for vectors.*

3 Numerical Experiments

3.1 Estimation algorithm

In this subsection, we present the fixed-point algorithm first proposed by Cui et al. (2011) and adapt it to our adaptive varying coefficient linear models.

Specifically, for the estimating Eq. (7), we can rewrite them as

$$\Phi_n(\boldsymbol{\beta}; \widehat{\mathbf{m}}) = \mathbf{J}^{\mathrm{T}} \mathbf{G}(\boldsymbol{\beta}; \widehat{\mathbf{m}}) \text{ with } \mathbf{G}(\boldsymbol{\beta}; \widehat{\mathbf{m}}) = (G_1(\boldsymbol{\beta}; \widehat{\mathbf{m}}), \ldots, G_p(\boldsymbol{\beta}; \widehat{\mathbf{m}}))^{\mathrm{T}},$$

where

$$G_j(\boldsymbol{\beta}; \widehat{\mathbf{m}}) = \frac{1}{n}\sum_{i=1}^{n}(X_{ij}\mathbf{Z}_i^{\mathrm{T}}\widehat{\mathbf{g}}'(\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}; \boldsymbol{\beta}) - \widehat{\mathbf{H}}_j(\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}; \boldsymbol{\beta})\mathbf{Z}_i)\psi_\tau(Y_i - \widehat{\mathbf{g}}^{\mathrm{T}}(\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}; \boldsymbol{\beta})\mathbf{Z}_i),$$

and $\widehat{\mathbf{H}}_j(\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}; \boldsymbol{\beta})$ is the jth row of $\widehat{\mathbf{H}}(\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}; \boldsymbol{\beta})$, $j = 1, \ldots, p$. The estimating equations $\Phi_n(\boldsymbol{\beta}; \widehat{\mathbf{m}}) = \mathbf{0}$ can be written as

$$\begin{cases} -\beta_2 G_1(\boldsymbol{\beta}; \widehat{\mathbf{m}})/\sqrt{1 - \|\boldsymbol{\beta}^{(-1)}\|^2} + G_2(\boldsymbol{\beta}; \widehat{\mathbf{m}}) = 0, \\ -\beta_3 G_1(\boldsymbol{\beta}; \widehat{\mathbf{m}})/\sqrt{1 - \|\boldsymbol{\beta}^{(-1)}\|^2} + G_3(\boldsymbol{\beta}; \widehat{\mathbf{m}}) = 0, \\ \vdots \\ -\beta_p G_1(\boldsymbol{\beta}; \widehat{\mathbf{m}})/\sqrt{1 - \|\boldsymbol{\beta}^{(-1)}\|^2} + G_p(\boldsymbol{\beta}; \widehat{\mathbf{m}}) = 0, \end{cases}$$

which leads to

$$\begin{cases} \beta_1 = |G_1(\boldsymbol{\beta}; \widehat{\mathbf{m}})|/\|\mathbf{G}(\boldsymbol{\beta}; \widehat{\mathbf{m}})\|^2, & \\ \beta_j^2 = G_j^2(\boldsymbol{\beta}; \widehat{\mathbf{m}})/\|\mathbf{G}(\boldsymbol{\beta}; \widehat{\mathbf{m}})\|^2, & 2 \le j \le p, \\ \mathrm{sign}\{\beta_j G_1(\boldsymbol{\beta}; \widehat{\mathbf{m}})\} = \mathrm{sign}\{G_j(\boldsymbol{\beta}; \widehat{\mathbf{m}})\}, & 2 \le j \le p. \end{cases}$$

The above system can be also written as

$$\boldsymbol{\beta}\frac{G_1(\boldsymbol{\beta}; \widehat{\mathbf{m}})}{\|\mathbf{G}(\boldsymbol{\beta}; \widehat{\mathbf{m}})\|} = \frac{|G_1(\boldsymbol{\beta}; \widehat{\mathbf{m}})|}{\|\mathbf{G}(\boldsymbol{\beta}; \widehat{\mathbf{m}})\|} \times \frac{\mathbf{G}(\boldsymbol{\beta}; \widehat{\mathbf{m}})}{\|\mathbf{G}(\boldsymbol{\beta}; \widehat{\mathbf{m}})\|}. \tag{9}$$

Since $\|\mathbf{G}(\boldsymbol{\beta}; \widehat{\mathbf{m}})\|$ may sometimes be a small value, which leads to the algorithm unstable. For fast convergence and robustness of the fixed-point algorithm, add $C\boldsymbol{\beta}$ to both sides of (9) and after transformation, we obtain

$$\boldsymbol{\beta} = \frac{C}{G_1(\boldsymbol{\beta}; \widehat{\mathbf{m}})/\|\mathbf{G}(\boldsymbol{\beta}; \widehat{\mathbf{m}})\| + C}\boldsymbol{\beta} + \frac{|G_1(\boldsymbol{\beta}; \widehat{\mathbf{m}})|/\|\mathbf{G}(\boldsymbol{\beta}; \widehat{\mathbf{m}})\|^2}{G_1(\boldsymbol{\beta}; \widehat{\mathbf{m}})/\|\mathbf{G}(\boldsymbol{\beta}; \widehat{\mathbf{m}})\| + C}\mathbf{G}(\boldsymbol{\beta}; \widehat{\mathbf{m}}), \tag{10}$$

where C is a constant which can be properly chosen such that $G_1(\boldsymbol{\beta}; \widehat{\mathbf{m}})/\|\mathbf{G}(\boldsymbol{\beta}; \widehat{\mathbf{m}})\| + C \neq 0$. Further discussions on choosing the constant C are referred to Cui et al. (2011).

Using (10), we can iteratively perform

$$\boldsymbol{\beta} \leftarrow \frac{C}{G_1(\boldsymbol{\beta}; \widehat{\mathbf{m}})/\|\mathbf{G}(\boldsymbol{\beta}; \widehat{\mathbf{m}})\| + C}\boldsymbol{\beta} + \frac{|G_1(\boldsymbol{\beta}; \widehat{\mathbf{m}})|/\|\mathbf{G}(\boldsymbol{\beta}; \widehat{\mathbf{m}})\|^2}{G_1(\boldsymbol{\beta}; \widehat{\mathbf{m}})/\|\mathbf{G}(\boldsymbol{\beta}; \widehat{\mathbf{m}})\| + C}\mathbf{G}(\boldsymbol{\beta}; \widehat{\mathbf{m}}),$$

until convergence. All numerical results below are produced by using the linear regression estimator (normalized to have unit norm) as the initial estimator. Our experience is that that the algorithm seems insensitive to the choice of initial estimators and the algorithm always converges to a small neighborhood of the true parameter in the simulations.

Within each iteration, for given $\boldsymbol{\beta}$, spline coefficients of $g_j(\cdot)$, $j = 0, \ldots, p$ can be obtained using R package *quantreg*. The projection estimate $\widehat{\mathbf{H}}(u; \boldsymbol{\beta})$ in (6) is obtained by the least squares estimate with weights $f(\mathbf{g}^{\mathrm{T}}(\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}; \boldsymbol{\beta})\mathbf{Z}_i|\mathbf{X}_i)$, $i = 1, \ldots, n$. Here, we adopt the difference quotient method of Hendricks and Koenker (1992),

$$\hat{f}(\mathbf{g}^{\mathrm{T}}(\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}; \boldsymbol{\beta})\mathbf{Z}_i|\mathbf{X}_i) = 2h_n \left\{\hat{\mathbf{g}}_{\tau+h_n}^{\mathrm{T}}(\mathbf{X}_i^{\mathrm{T}}\hat{\boldsymbol{\beta}}_{\tau+h_n})\mathbf{Z}_i - \hat{\mathbf{g}}_{\tau-h_n}^{\mathrm{T}}(\mathbf{X}_i^{\mathrm{T}}\hat{\boldsymbol{\beta}}_{\tau-h_n})\mathbf{Z}_i\right\}^{-1}, \tag{11}$$

where $\hat{\boldsymbol{\beta}}_\tau$ and $\hat{\mathbf{g}}_\tau(\cdot)$ denote the estimators at quantile level τ (obtained when values of ones are used in place of the conditional density for simplicity) and h_n is a bandwidth parameter tending to zero as $n \to \infty$. In our numerical studies, we choose

$$\mathrm{h}_n = 1.57n^{-1/3}\left(1.5\phi^2\{\Phi^{-1}(\tau)\}/(2\{\Phi^{-1}(\tau)\}^2+1)\right)^{2/3}$$

following Hall and Sheather (1988), where $\phi(\cdot)$ and $\Phi(\cdot)$ are the pdf and cdf of the standard normal distribution.

We also need to select the number of knots and the order of splines to approximate the nonparametric functions. In practice, the cubic spline ($s = 4$) is used and the number of knots can be chosen by minimizing the following Schwarz Information Criterion (SIC, Horowitz and Lee 2005)

$$\hat{K} = \operatorname{argmin}_K \mathrm{SIC}(K),$$

where

$$\mathrm{SIC}(K) = \log\left(\sum_{i=1}^n \rho_\tau\left(Y_i - \sum_{j=0}^p Z_{ij}\mathbf{B}^{\mathrm{T}}(\mathbf{X}_i^{\mathrm{T}}\hat{\boldsymbol{\beta}})\hat{\boldsymbol{\theta}}_j\right)\right) + \log(n)\times((p+1)K)/(2n),$$

where $\hat{\boldsymbol{\theta}}_j$, $j = 0, 1, \ldots, p$ denote the estimated spline coefficients for a given K.

Finally, based on the density estimate given in (11) and Theorem 1, we can obtain the estimate of variance–covariance matrix for index parameter $\boldsymbol{\beta}$ by the following sandwich formula

$$\hat{\boldsymbol{\Sigma}}_\beta = \frac{1}{n}\mathbf{J}\hat{\boldsymbol{\Psi}}_1^{-1}\hat{\boldsymbol{\Psi}}_2\hat{\boldsymbol{\Psi}}_1^{-1}\mathbf{J}^{\mathrm{T}}, \tag{12}$$

where

$$\boldsymbol{\Psi}_1 = \frac{1}{n}\sum_{i=1}^n \hat{f}(\hat{\mathbf{g}}^{\mathrm{T}}(\mathbf{X}_i^{\mathrm{T}}\hat{\boldsymbol{\beta}})\mathbf{Z}_i|\mathbf{X}_i)(\mathbf{J}^{\mathrm{T}}\mathbf{X}_i\mathbf{Z}_i^{\mathrm{T}}\hat{\mathbf{g}}'(\mathbf{X}_i^{\mathrm{T}}\hat{\boldsymbol{\beta}}) - \mathbf{J}^{\mathrm{T}}\widehat{\mathbf{H}}(\mathbf{X}_i^{\mathrm{T}}\widehat{\boldsymbol{\beta}}; \widehat{\boldsymbol{\beta}})\mathbf{Z}_i)^{\otimes 2}$$

and

$$\boldsymbol{\Psi}_2 = \tau(1-\tau)\frac{1}{n}\sum_{i=1}^{n}\left(\mathbf{J}^{\mathrm{T}}\mathbf{X}_i\mathbf{Z}_i^{\mathrm{T}}\hat{\mathbf{g}}'(\mathbf{X}_i^{\mathrm{T}}\hat{\boldsymbol{\beta}}) - \mathbf{J}^{\mathrm{T}}\widehat{\mathbf{H}}(\mathbf{X}_i^{\mathrm{T}}\widehat{\boldsymbol{\beta}};\widehat{\boldsymbol{\beta}})\mathbf{Z}_i\right)^{\otimes 2},$$

in which $\mathbf{J}$ is evaluated at $\hat{\boldsymbol{\beta}}$.

3.2 Simulation studies

Consider the following model

$$Y_i = g_0(\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}) + \sum_{j=1}^{p} g_j(\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta})X_{ij} + 0.1\exp(\kappa\cdot\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta})\times e_i,\ i = 1,\ldots,n,$$

where the true parameter $\boldsymbol{\beta} = (1, 2, 0, 2, \underbrace{0,\ldots,0}_{p-4})^{\mathrm{T}}$, and the nonparametric functions

$$g_0(u) = 3\exp(-u^2),\ g_1(u) = 1.8u^2,\ g_3(u) = 2\sin(\pi u),$$

and

$$g_2(u) \equiv 0,\quad g_j(u) \equiv 0,\quad j = 4,\ldots,p.$$

The covariates $\mathbf{X}_i = (X_{i1},\ldots,X_{ip})^T$ are independent random vectors uniformly distributed on $[-1, 1]^p$, and the error e_i are independently sampled from standard normal distribution $N(0, 1)$ or t(3), the Student's t distribution with degrees of freedom 3. In addition, the quantity κ equals 0 or 1 corresponding to homoscedastic model (HM) and heteroscedastic model (HT), respectively. Here, we focus on the quantile levels at $\tau = 0.1, 0.25$ and 0.5, and conduct the simulations with sample size $n = 400$, each with 500 replications.

For the dimension of covariates $\mathbf{X}$, we consider three cases with $p = 4, 8, 12$. To assess the finite sample performance of the estimated index parameter, we report the root mean squared errors denoted as $\mathrm{RMSE}_{\boldsymbol{\beta}} = \sqrt{\frac{1}{p}\sum_{j=1}^{p}(\hat{\beta}_j - \beta_j)^2}$. For the nonparametric functions, we also use the root mean squared errors defined as

$$\mathrm{RMSE}_g = \frac{1}{p+1}\sum_{j=0}^{p}\sqrt{\frac{1}{n_{\mathrm{grid}}}\sum_{k=1}^{n_{\mathrm{grid}}}(\hat{g}_j(u_k) - g_j(u_k))^2},$$

where $\{u_k : k = 1,\ldots,n_{\mathrm{grid}}\}$ are regular grid points on $[\min_i \mathbf{X}_i^{\mathrm{T}}\hat{\boldsymbol{\beta}}, \max_i \mathbf{X}_i^{\mathrm{T}}\hat{\boldsymbol{\beta}}]$ with $n_{\mathrm{grid}} = 100$. To illustrate the robustness of the median regression (quantile regression when $\tau = 0.5$, we also report the results of adaptive varying-coefficient linear model with quadratic loss (mean regression), for which we again used polynomial splines.

Table 1 Simulation results at three quantile levels with $\boldsymbol{\beta} = (1, 2, 0, 2, 0, \ldots, 0)^{\mathrm{T}}$

κ	p	τ	$N(0, 1)$		$t(3)$	
			$\text{RMSE}_{\boldsymbol{\beta}}$	RMSE_g	$\text{RMSE}_{\boldsymbol{\beta}}$	RMSE_g
$\kappa = 0$	4	0.1	0.0051 (0.0023)	0.3556 (0.1850)	0.0083 (0.0035)	0.3827 (0.3340)
		0.25	0.0039 (0.0016)	0.3339 (0.1619)	0.0050 (0.0023)	0.3820 (0.3252)
		0.5	0.0037 (0.0016)	0.2964 (0.1241)	0.0042 (0.0019)	0.3112 (0.1261)
	8	0.1	0.0059 (0.0036)	0.3290 (0.1699)	0.0086 (0.0105)	0.3844 (0.3186)
		0.25	0.0048 (0.0014)	0.3120 (0.1591)	0.0059 (0.0018)	0.3801 (0.3113)
		0.5	0.0043 (0.0013)	0.2541 (0.1325)	0.0049 (0.0014)	0.2840 (0.1402)
	12	0.1	0.0138 (0.0266)	0.3430 (0.3180)	0.0199 (0.0330)	0.3967 (0.3824)
		0.25	0.0057 (0.0097)	0.3279 (0.2963)	0.0102 (0.0145)	0.3937 (0.3744)
		0.5	0.0048 (0.0011)	0.2479 (0.1090)	0.0061 (0.0063)	0.2600 (0.1123)
$\kappa = 1$	4	0.1	0.0048 (0.0023)	0.3622 (0.2871)	0.0077 (0.0034)	0.4089 (0.3821)
		0.25	0.0037 (0.0016)	0.3602 (0.2762)	0.0047 (0.0023)	0.4008 (0.3765)
		0.5	0.0035 (0.0016)	0.3552 (0.1697)	0.0040 (0.0019)	0.3937 (0.1954)
	8	0.1	0.0113 (0.0564)	0.3513 (0.2699)	0.0091 (0.0033)	0.4205 (0.3809)
		0.25	0.0045 (0.0015)	0.3495 (0.2614)	0.0057 (0.0018)	0.4200 (0.3764)
		0.5	0.0041 (0.0013)	0.2773 (0.1475)	0.0048 (0.0014)	0.3449 (0.1791)
	12	0.1	0.0156 (0.0261)	0.3751 (0.3412)	0.0243 (0.0367)	0.4380 (0.4186)
		0.25	0.0070 (0.0187)	0.3720 (0.3352)	0.0082 (0.0152)	0.4373 (0.4147)
		0.5	0.0046 (0.0012)	0.2805 (0.1551)	0.0054 (0.0023)	0.3635 (0.2072)

The values in the parentheses are the sample standard errors computed based on the 500 replications

The simulation results for the quantile model are shown in Table 1, while those for the mean regression model are shown in Table 2. From Table 1, we see that the performance of our proposed estimation procedure is reasonable for all cases, for both homoscedastic and heteroscedastic models at three different quantile levels, even when the dimension p is not too small. Empirically, the algorithm converges well in all cases. We also note that among the three quantile levels, the errors for $\tau = 0.1$ are the largest and the errors for $\tau = 0.5$ are the smallest, suggesting tail of the response distribution is more difficult to estimate.

To better illustrate the difference between quantile method ($\tau = 0.5$) and mean regression for each setting, the boxplots of both $\text{RMSE}_{\boldsymbol{\beta}}$ and RMSE_g are shown in Fig. 1–2. We see that comparing mean regression with median regression, the performance of our proposed procedure is significantly better than mean regression for t errors, which is as expected due to robustness of median regression.

Finally, according to a referee's suggestion, we report the results of an additional simulation study in Table 3 where the index parameter $\boldsymbol{\beta} = (1, 2, 0, 2, \underbrace{0.25, \ldots, 0.25}_{p-4})^{\mathrm{T}}$ with $p = 8$ and 12, and the other settings are the same as above, to illustrate the case when the number of nonzero parameters increases. As we can see, our proposed estimation procedure still works well, with larger dimensionality associated with larger errors.

Table 2 Simulation results for mean regression of adaptive varying-coefficient linear model with $\boldsymbol{\beta} = (1, 2, 0, 2, 0, \ldots, 0)^{\mathrm{T}}$

κ	p	$N(0, 1)$		$t(3)$	
		RMSE$_{\boldsymbol{\beta}}$	RMSE$_g$	RMSE$_{\boldsymbol{\beta}}$	RMSE$_g$
$\kappa = 0$	4	0.0029 (0.0012)	0.2814 (0.1174)	0.0050 (0.0020)	0.3396 (0.1676)
	8	0.0035 (0.0010)	0.2296 (0.0992)	0.0056 (0.0024)	0.3062 (0.1919)
	12	0.0037 (0.0009)	0.2156 (0.0861)	0.0074 (0.0034)	0.2884 (0.1488)
$\kappa = 1$	4	0.0036 (0.0015)	0.3598 (0.1594)	0.0068 (0.0041)	0.4799 (0.2637)
	8	0.0045 (0.0014)	0.2784 (0.1342)	0.0075 (0.0049)	0.4245 (0.3637)
	12	0.0047 (0.0011)	0.2811 (0.1440)	0.0088 (0.0064)	0.4568 (0.3300)

The values in the parentheses are the sample standard errors computed based on the 500 replications

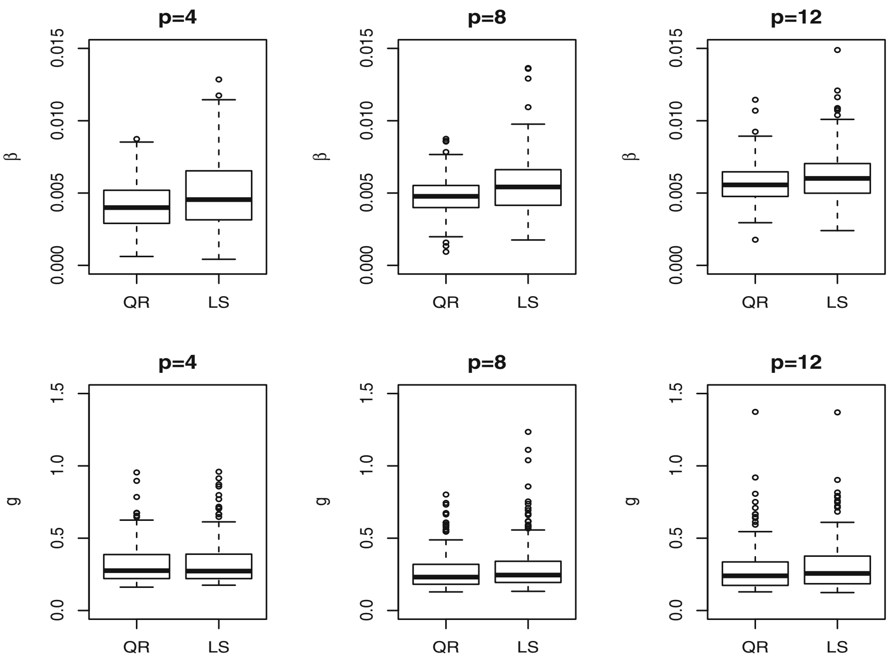

Fig. 1 The boxplots of RMSE for quantile regression method (QR, with $\tau = 0.5$) and least squares method (LS) with homoscedastic t errors, where the two rows corresponds to RMSE$_{\boldsymbol{\beta}}$ and RMSE$_g$, respectively

3.3 Data application

Forest fire has been a major environmental issue, which leads to economic and ecologic damage and endangers human lives. Fast detection is a key element for controlling such phenomenon (Cortez and Morais 2007). In this subsection, we will apply our proposed quantile adaptive varying-coefficient model to the forest fires data.

The data were collected from the Montesinho natural park, which lies in Trás-os-Montes northeast region of Portugal, from January 2000 to December 2003 and

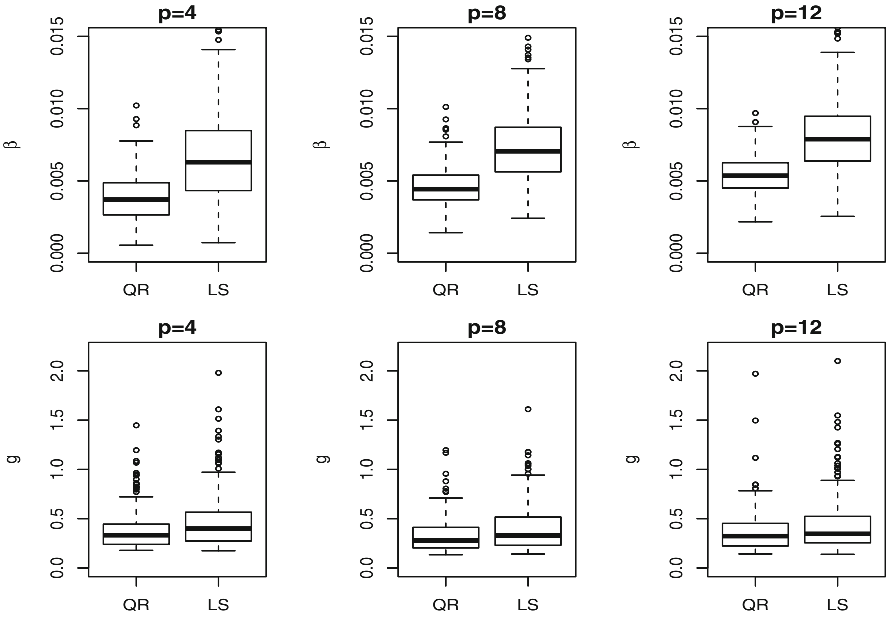

Fig. 2 The boxplots of RMSE for quantile regression method (QR, with $\tau = 0.5$) and least squares method (LS) with heterogeneous t errors, where the two rows corresponds to $\text{RMSE}_{\boldsymbol{\beta}}$ and RMSE_g, respectively

Table 3 Simulation results at three quantile levels with $\boldsymbol{\beta} = (1, 2, 0, 2, 0.25, \ldots, 0.25)^{\mathrm{T}}$

κ	p	τ	$N(0, 1)$		$t(3)$	
			$\text{RMSE}_{\boldsymbol{\beta}}$	RMSE_g	$\text{RMSE}_{\boldsymbol{\beta}}$	RMSE_g
$\kappa = 0$	8	0.1	0.0059 (0.0017)	0.3231 (0.1980)	0.0095 (0.0033)	0.4892 (0.3498)
		0.25	0.0047 (0.0013)	0.2855 (0.1670)	0.0059 (0.0017)	0.3083 (0.1544)
		0.5	0.0044 (0.0012)	0.2683 (0.1323)	0.0050 (0.0015)	0.2983 (0.1473)
	12	0.1	0.0198 (0.0274)	0.5899 (0.4875)	0.0333 (0.0568)	0.8141 (0.8212)
		0.25	0.0072 (0.0143)	0.3163 (0.2000)	0.0083 (0.0154)	0.3863 (0.2742)
		0.5	0.0064 (0.0181)	0.3090 (0.1879)	0.0056 (0.0014)	0.3262 (0.1917)
$\kappa = 1$	8	0.1	0.0058 (0.0019)	0.4052 (0.2556)	0.0093 (0.0036)	0.6092 (0.4260)
		0.25	0.0045 (0.0014)	0.3426 (0.1867)	0.0055 (0.0018)	0.4292 (0.2299)
		0.5	0.0043 (0.0012)	0.3239 (0.1699)	0.0049 (0.0016)	0.3762 (0.2025)
	12	0.1	0.0239 (0.0321)	0.6679 (0.6139)	0.0303 (0.0418)	1.0062 (0.9705)
		0.25	0.0081 (0.0226)	0.3965 (0.2777)	0.0084 (0.0168)	0.4950 (0.3373)
		0.5	0.0052 (0.0095)	0.3759 (0.2304)	0.0056 (0.0015)	0.4394 (0.2842)

The values in the parentheses are the sample standard errors computed based on the 500 replications

available at http://www.dsi.uminho.pt/~pcortez/forestfires/. The response variable of interest is the burned area (in hectares); we will use the adaptive varying-coefficient model to explore the relationship between burned area and some predictors at different

Table 4 The estimate of index parameter under two different quantile levels

		β_1	β_2	β_3	β_4	β_5	β_6	β_7
$\tau = 0.5$	Estimate	0.6954	0.1081	0.5248	0.2777	0.0136	0.0227	−0.3892
	Std	0.0172	0.0102	0.0124	0.0141	0.0555	0.0137	0.0484
$\tau = 0.95$	Estimate	0.1289	−0.6642	0.3712	−0.4168	−0.0751	0.1245	0.4578
	Std	0.0005	0.0006	0.0003	0.0005	0.0011	0.0002	0.0015

quantile levels. These include seven forest fire weather index: fine fuel moisture code (FFMC), duff moisture code (DMC), drought code (DC), initial spread index (ISI), outside temperature (TEMP, in ^oC), outside relative humidity (RH, in %) and outside wind speed (WIND in km/h). There are a total of 517 observations.

The distribution of burned area is severely right-skewed, and thus, we use $\log(\text{burned area} + 1)$ as the response. The estimated index parameters using the proposed adaptive varying-coefficient linear model are shown in Table 4 for $\tau = 0.5$ and $\tau = 0.95$. The standard errors are obtained based on asymptotic normality. All the seven variables are significant for $\tau = 0.95$, and two variables, i.e., TEMP and RH, are insignificant when $\tau = 0.5$. Obviously the estimates at the tail are very different from the estimates at the center of the response distribution and such information cannot be gained by mean regression. Figures 3 and 4 show the estimated curves of eight nonparametric functions for $\tau = 0.5$ and $\tau = 0.95$, and the 95% confidence band for each function is also shown, which is obtained by bootstrap method. Visually the bands for $\tau = 0.95$ are generally wider. If we think of $\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}$ as an overall weather index, we can interpret the curves based on this constructed index. For example, we see that at $\tau = 0.5$, g_1 is flat for most of the index values while it increases suddenly when the index is very large. This means when the overall weather index is extreme (large), the effect of FFMC (the variable X_1) becomes more obvious.

4 Discussion

In this paper, we proposed a profiled estimating equations approach for estimation in adaptive varying-coefficient quantile models. The quantile model allows us to obtain a more complete picture of the conditional distribution of a response variable given covariates. We use polynomial splines to estimate the unknown nonparametric coefficient functions and derive the bias-corrected estimating equations that do not require undersmoothing of the nonparametric functions. A fixed-point algorithm is used in our implementation to obtain the estimator for the index vector.

In our experience, the fixed-point algorithm seems to be quite insensitive to the choice of initial estimators. It is known that the fixed-point algorithm converges if the iteration is contractive. However, it seems this is generally not true. Thus whether there is theoretical guarantee for the fixed-point algorithm, we use here is an open question.

A problem of interest is penalized variable selection when the covariate vector is high dimensional. For linear quantile regression, this has been studied in Belloni and

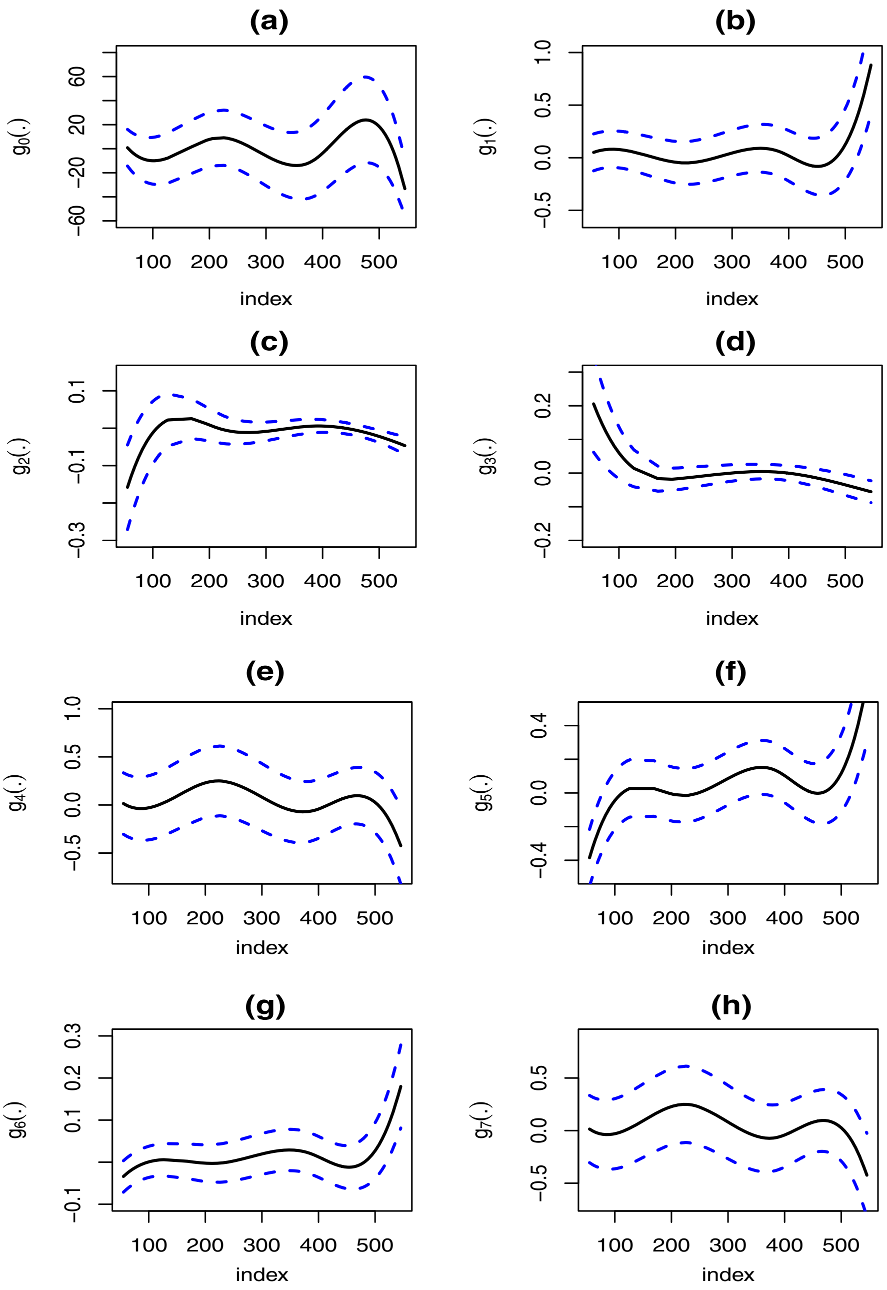

Fig. 3 The estimated curves for nonparametric functions when $\tau = 0.5$

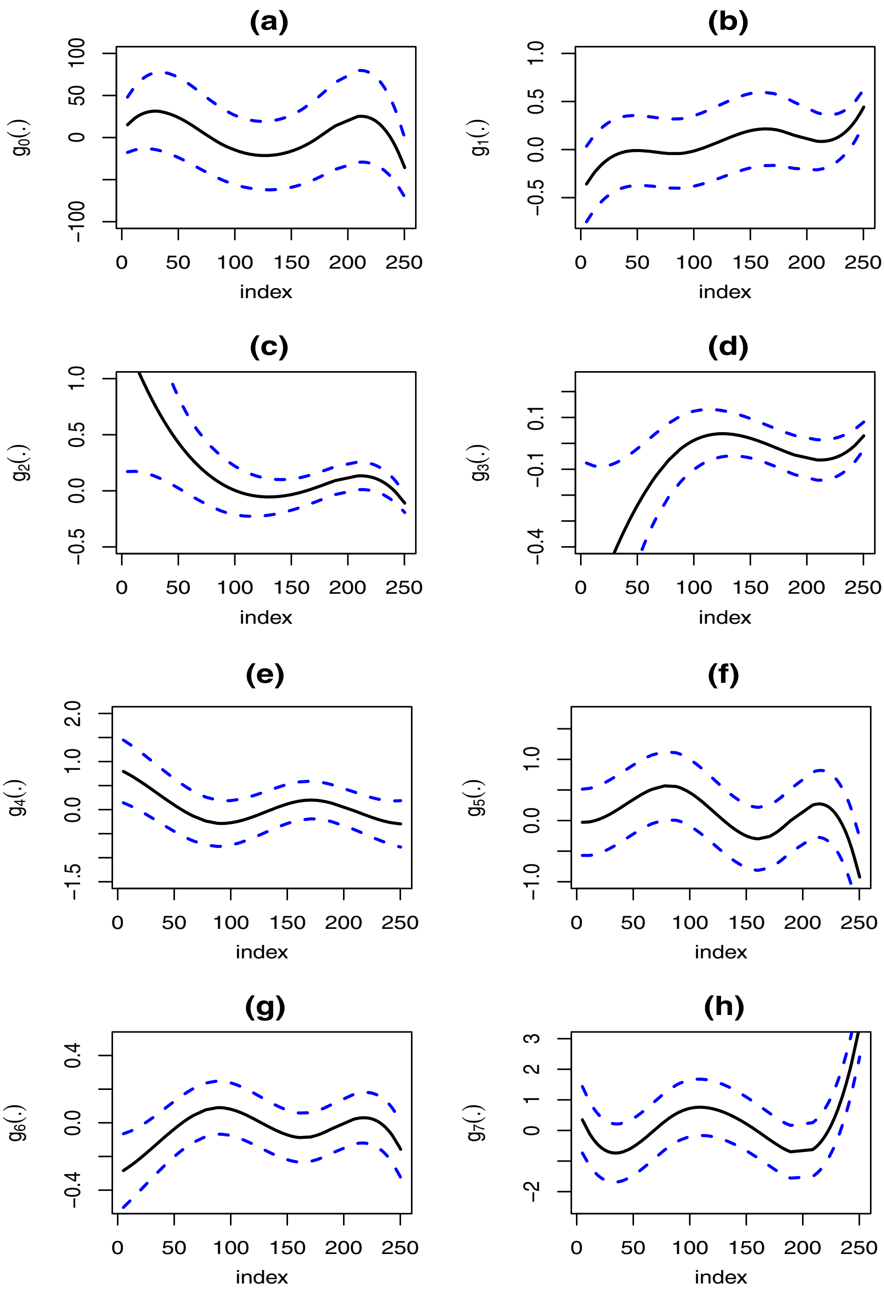

Fig. 4 The estimated curves for nonparametric functions when $\tau = 0.95$

Chernozhukov (2011); Fan et al. (2014a); Wang et al. (2012), while Sherwood and Wang (2016); Tang et al. (2013) considered high-dimensional partially linear quantile models.

Another related problem is whether one can estimate the multiple quantiles simultaneous while borrowing information cross-quantile levels for estimation and variable selection. Some efforts along this direction for different setups have appeared in Jiang et al. (2013); Yang and He (2012) and more recently in Sherwood and Wang (2016).

Acknowledgements We sincerely thank the AE and two anonymous reviewers for their insightful comments which have led to significant improvement of the paper. The research of Zhao was supported in part by National Social Science Fund of China (15BTJ027), and the research of Lian was supported by a start up Grant (No. 7200521/MA) from the City University of Hong Kong.

Appendix: Technical proofs

In our proofs, C denotes a generic positive constant which can take different values even on the same line.

Proof of Theorem 1 Consider the class of functions

$$\begin{aligned}\mathcal{F} = \Big\{ &\left(\mathbf{x}\mathbf{z}^{\mathrm{T}}\mathbf{g}'(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}) - \mathbf{H}(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta})\mathbf{z}\right)\left(\tau - I\left\{y - \mathbf{g}^{\mathrm{T}}(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta})\mathbf{z} \le 0\right\}\right), \boldsymbol{\beta} \in \mathcal{B},\\ &\text{and for some } \alpha > 3/2, \alpha' > 1/2, \text{ and } M > 0,\\ &\text{entries of } \mathbf{g} \text{ are in } \mathcal{C}^{\alpha}(M), \text{ and entries of } \mathbf{H} \text{ are in } \mathcal{C}^{\alpha'}(M)\Big\}.\end{aligned}$$

For $\alpha > 1/2$, define

$$\mathcal{F}_1 = \left\{h(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}) : \boldsymbol{\beta} \in \mathcal{B}, h \in \mathcal{C}^{\alpha}(M)\right\}.$$

For any fixed $\boldsymbol{\beta}$, the class $\mathcal{F}_1(\boldsymbol{\beta}) := \{h(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}) : h \in \mathcal{C}^{\alpha}(R)\}$ has entropy $\log N(\delta, \mathcal{C}^{\alpha}(M), \|.\|_{\infty}) \le C\delta^{-1/\alpha}$ by Theorem 2.7.1 of van der Vaart and Wellner (1996). Since $\|h(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}) - h(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}')\|_{\infty} \le C\|\boldsymbol{\beta} - \boldsymbol{\beta}\|^{s}$ for some $s > 1/2$, it is easy to see that the δ-entropy of $\mathcal{F}_1$ in L_{∞} norm is bounded by the sum of $C\delta^{-1/\alpha}$ and the $\delta^{1/s}$-entropy of $\mathcal{B}$ in Euclidean norm, with the latter being $C\log(1/\delta)$. Thus by Theorem 19.14 of van der Vaart (1998) and that $\mathbf{x}$ lies in a compact set, $\mathcal{F}_1$ is a Donsker class.

Furthermore, consider

$$\mathcal{F}_2 = \left\{I\{y - \mathbf{g}^{\mathrm{T}}(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta})\mathbf{z} \le 0\} : \boldsymbol{\beta} \in \mathcal{B}, \mathbf{g} \in \mathcal{C}^{\alpha}(M)\right\}.$$

We have

$$
\begin{aligned}
&E\left(I\left\{y-\mathbf{g}_1^{\mathrm{T}}(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_1)\mathbf{z}\le 0\right\}-I\left\{y-\mathbf{g}_2^{\mathrm{T}}(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_2)\mathbf{z}\le 0\right\}\right)^2\\
&\quad\le CE|\mathbf{g}_1^{\mathrm{T}}(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_1)\mathbf{z}-\mathbf{g}_2^{\mathrm{T}}(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_2)\mathbf{z}|\\
&\quad\le C\left(E\left[\mathbf{g}_1(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_1)-\mathbf{g}_2(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_2)\right]^2\right)^{1/2}.
\end{aligned}
$$

Thus $\log N(\delta,\mathcal{F}_2,L_2)\le C\log N(C\delta^2,\mathcal{F}_1,L_2)\le C\delta^{-2/\alpha}$ and $\mathcal{F}_2$ is Donsker if $\alpha>1$. Combining that $\mathcal{F}_1$ and $\mathcal{F}_2$ are Donsker classes, it is easy to see that $\mathcal{F}$ is also a Donsker class.

First we prove consistency. Let $F(.|\mathbf{X})$ be the conditional cumulative distribution function of Y. Uniformly for all $\boldsymbol{\beta}\in\mathcal{B}$, we have

$$
\begin{aligned}
&\Phi(\boldsymbol{\beta};\widehat{\mathbf{m}}(;\boldsymbol{\beta}))-\Phi(\boldsymbol{\beta};\mathbf{m}(;\boldsymbol{\beta}))\\
&\quad=\mathbf{J}^{\mathrm{T}}E\left(\mathbf{X}\mathbf{Z}^{\mathrm{T}}\widehat{\mathbf{g}}'(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta};\boldsymbol{\beta})-\widehat{\mathbf{H}}(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta};\boldsymbol{\beta})\mathbf{Z}\right)(\tau-F\left(\widehat{\mathbf{g}}^{\mathrm{T}}(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta};\boldsymbol{\beta})\mathbf{Z}|\mathbf{X}\right)\\
&\qquad-\mathbf{J}^{\mathrm{T}}E\left(\mathbf{X}\mathbf{Z}^{\mathrm{T}}\mathbf{g}'(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta};\boldsymbol{\beta})-\mathbf{H}(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta};\boldsymbol{\beta})\mathbf{Z}\right)\left(\tau-F\left(\mathbf{g}^{\mathrm{T}}(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta};\boldsymbol{\beta})\mathbf{Z}|\mathbf{X}\right)\right)\\
&\quad=\mathbf{J}^{\mathrm{T}}E\left(\mathbf{X}\mathbf{Z}^{\mathrm{T}}\left(\widehat{\mathbf{g}}'(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta};\boldsymbol{\beta})-\mathbf{g}'(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta};\boldsymbol{\beta})\right)-\left(\widehat{\mathbf{H}}(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta})\right.\right.\\
&\qquad\left.\left.-\mathbf{H}(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta})\right)\mathbf{Z}\right)(\tau-F\left(\widehat{\mathbf{g}}^{\mathrm{T}}(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta};\boldsymbol{\beta})\mathbf{Z}|\mathbf{X}\right)\\
&\qquad-\mathbf{J}^{\mathrm{T}}E\left(\mathbf{X}\mathbf{Z}^{\mathrm{T}}\mathbf{g}'(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta};\boldsymbol{\beta})-\mathbf{H}(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta})\mathbf{Z}\right)\left(F\left(\widehat{\mathbf{g}}^{\mathrm{T}}(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta};\boldsymbol{\beta})\mathbf{Z}|\mathbf{X}\right)\right.\\
&\qquad\left.-F\left(\mathbf{g}^{\mathrm{T}}(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta};\boldsymbol{\beta})\mathbf{Z}|\mathbf{X}\right)\right)\\
&\quad=o_p(1),
\end{aligned}
\tag{13}
$$

using Proposition 1. Furthermore, by the Glivenko–Cantelli Theorem ($\mathcal{F}$ is Donsker implies it is Glivenko–Cantelli), $\sup_{f\in\mathcal{F}}(P_n-P)f=o_p(1)$. Thus uniformly for $\boldsymbol{\beta}\in\mathcal{B}$,

$$
\|\Phi_n(\boldsymbol{\beta};\widehat{\mathbf{m}}(.;\boldsymbol{\beta}))-\Phi(\boldsymbol{\beta};\widehat{\mathbf{m}}(.;\boldsymbol{\beta})\|=o_p(1).
\tag{14}
$$

Thus

$$
\begin{aligned}
&\|\Phi(\widehat{\boldsymbol{\beta}};\mathbf{m}(.;\widehat{\boldsymbol{\beta}})\|\\
&\quad=\|\Phi(\widehat{\boldsymbol{\beta}};\widehat{\mathbf{m}}(.;\widehat{\boldsymbol{\beta}})\|+o_p(1)\\
&\quad=\|\Phi_n(\widehat{\boldsymbol{\beta}};\widehat{\mathbf{m}}(.;\widehat{\boldsymbol{\beta}})\|+o_p(1)\\
&\quad\le\|\Phi_n(\boldsymbol{\beta}_0;\widehat{\mathbf{m}}(.;\boldsymbol{\beta}_0)\|+o_p(1)\\
&\quad=\|\Phi(\boldsymbol{\beta}_0;\widehat{\mathbf{m}}(.;\boldsymbol{\beta}_0)\|+o_p(1)\\
&\quad=\|\Phi(\boldsymbol{\beta}_0;\mathbf{m}(.;\boldsymbol{\beta}_0)\|+o_p(1),
\end{aligned}
$$

where the first and the last equality used (13), the second and third equality used (14) and the inequality follows from the definition of $\widehat{\boldsymbol{\beta}}$ as an approximate minimizer of

$\|\Phi(\boldsymbol{\beta}, \widehat{\mathbf{m}}(.; \boldsymbol{\beta}))\|$. Thus by assumption (A4), $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\| < \epsilon$ with probability approaching one, for any $\epsilon > 0$. This shows $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\| = o_p(1)$.

Now we consider asymptotic normality. For readability, we split the proof into several steps.

Step 1 By consistency, Lemma 19.24 in van der Vaart (1998) then implies that

$$\mathbb{G}_n(\phi_{\widehat{\boldsymbol{\beta}},\widehat{\mathbf{m}}} - \phi_{\boldsymbol{\beta}_0,\mathbf{m}}) = o_p(1), \tag{15}$$

where $\mathbb{G}_n = \sqrt{n}(P_n - P)$ is the empirical process.

Step 2 We show

$$\sqrt{n}P(\phi_{\widehat{\boldsymbol{\beta}},\widehat{\mathbf{m}}} - \phi_{\boldsymbol{\beta}_0,\mathbf{m}}) = \boldsymbol{\Psi}_1\sqrt{n}(\widehat{\boldsymbol{\beta}}^{(-1)} - \boldsymbol{\beta}_0^{(-1)}) + o_p(\sqrt{n}(\widehat{\boldsymbol{\beta}}^{(-1)} - \boldsymbol{\beta}_0^{(-1)})) + o_p(1). \tag{16}$$

In the proof of (16), we write $\widehat{\boldsymbol{\beta}}$ as $\boldsymbol{\beta}$ for simplicity of notation. Writing $P(\phi_{\boldsymbol{\beta},\widehat{\mathbf{m}}} - \phi_{\boldsymbol{\beta}_0,\mathbf{m}}) = P(\phi_{\boldsymbol{\beta},\widehat{\mathbf{m}}} - \phi_{\boldsymbol{\beta},\mathbf{m}}) + P(\phi_{\boldsymbol{\beta},\mathbf{m}} - \phi_{\boldsymbol{\beta}_0,\mathbf{m}})$, we first compute $P(\phi_{\boldsymbol{\beta},\widehat{\mathbf{m}}} - \phi_{\boldsymbol{\beta},\mathbf{m}})$ as follows.

$$\begin{aligned}
&P(\phi_{\boldsymbol{\beta},\widehat{\mathbf{m}}} - \phi_{\boldsymbol{\beta},\mathbf{m}}) \\
&\quad = \mathbf{J}^{\mathrm{T}}E\left(\mathbf{X}\mathbf{Z}^{\mathrm{T}}\widehat{\mathbf{g}}'(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta};\boldsymbol{\beta}) - \widehat{\mathbf{H}}(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta};\boldsymbol{\beta})\mathbf{Z}\right)\left(\tau - F\left(\widehat{\mathbf{g}}^{\mathrm{T}}(\mathbf{X}^{\mathrm{T}}\widehat{\boldsymbol{\beta}};\widehat{\boldsymbol{\beta}})\mathbf{Z}|\mathbf{X}\right)\right) \\
&\qquad -\mathbf{J}^{\mathrm{T}}E\left(\mathbf{X}\mathbf{Z}^{\mathrm{T}}\mathbf{g}'(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta};\boldsymbol{\beta}) - \mathbf{H}(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta};\boldsymbol{\beta})\mathbf{Z}\right)\left(\tau - F\left(\mathbf{g}^{\mathrm{T}}(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta};\boldsymbol{\beta})\mathbf{Z}|\mathbf{X}\right)\right) \\
&\quad = \mathbf{J}^{\mathrm{T}}E\left(\mathbf{X}\mathbf{Z}^{\mathrm{T}}\left(\widehat{\mathbf{g}}'(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta};\boldsymbol{\beta}) - \mathbf{g}'(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta};\boldsymbol{\beta})\right) - \left(\widehat{\mathbf{H}}(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta};\boldsymbol{\beta}) - \mathbf{H}(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta};\boldsymbol{\beta})\right)\mathbf{Z}\right) \\
&\qquad \cdot\left(\tau - F\left(\widehat{\mathbf{g}}^{\mathrm{T}}(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta};\boldsymbol{\beta})\mathbf{Z}|\mathbf{X}\right)\right) \\
&\qquad -\mathbf{J}^{\mathrm{T}}E\left(\mathbf{X}\mathbf{Z}^{\mathrm{T}}\mathbf{g}'(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta};\boldsymbol{\beta}) - \mathbf{H}(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta};\boldsymbol{\beta})\mathbf{Z}\right)\left(F\left(\widehat{\mathbf{g}}^{\mathrm{T}}(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta};\boldsymbol{\beta})\mathbf{Z}|\mathbf{X}\right)\right. \\
&\qquad \left. -F\left(\mathbf{g}^{\mathrm{T}}(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta};\boldsymbol{\beta})\mathbf{Z}|\mathbf{X}\right)\right) \\
&\quad = \mathbf{J}^{\mathrm{T}}E\left(\mathbf{X}\mathbf{Z}^{\mathrm{T}}\left(\widehat{\mathbf{g}}'(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta};\boldsymbol{\beta}) - \mathbf{g}'(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta};\boldsymbol{\beta})\right) - \left(\widehat{\mathbf{H}}(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta};\boldsymbol{\beta}) - \mathbf{H}(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta};\boldsymbol{\beta})\right)\mathbf{Z}\right) \\
&\qquad \cdot\left(F\left(\mathbf{g}^{\mathrm{T}}(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta};\boldsymbol{\beta})\mathbf{Z}|\mathbf{X}\right) - F\left(\widehat{\mathbf{g}}^{\mathrm{T}}(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta};\boldsymbol{\beta})\mathbf{Z}|\mathbf{X}\right)\right) \\
&\qquad +\mathbf{J}^{\mathrm{T}}E\left(\mathbf{X}\mathbf{Z}^{\mathrm{T}}\left(\widehat{\mathbf{g}}'(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta};\boldsymbol{\beta}) - \mathbf{g}'(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta};\boldsymbol{\beta})\right) - \left(\widehat{\mathbf{H}}(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta};\boldsymbol{\beta}) - \mathbf{H}(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta};\boldsymbol{\beta})\right)\mathbf{Z}\right) \\
&\qquad \cdot\left(F\left(\mathbf{g}^{\mathrm{T}}(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}_0;\boldsymbol{\beta}_0)\mathbf{Z}|\mathbf{X}\right) - F\left(\mathbf{g}^{\mathrm{T}}(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta};\boldsymbol{\beta})\mathbf{Z}|\mathbf{X}\right)\right) \\
&\qquad +\mathbf{J}^{\mathrm{T}}E\left(\mathbf{X}\mathbf{Z}^{\mathrm{T}}\left(\widehat{\mathbf{g}}'(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta};\boldsymbol{\beta}) - \mathbf{g}'(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta};\boldsymbol{\beta})\right) - \left(\widehat{\mathbf{H}}(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta};\boldsymbol{\beta}) - \mathbf{H}(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta};\boldsymbol{\beta})\right)\mathbf{Z}\right) \\
&\qquad \cdot\left(\tau - F\left(\mathbf{g}^{\mathrm{T}}(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}_0;\boldsymbol{\beta}_0)\mathbf{Z}|\mathbf{X}\right)\right) \\
&\qquad -\mathbf{J}^{\mathrm{T}}E\left(\mathbf{X}\mathbf{Z}^{\mathrm{T}}\mathbf{g}'(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta};\boldsymbol{\beta}) - \mathbf{H}(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta};\boldsymbol{\beta})\mathbf{Z}\right)\left(F\left(\widehat{\mathbf{g}}^{\mathrm{T}}(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta};\boldsymbol{\beta})\mathbf{Z}|\mathbf{X}\right)\right. \\
&\qquad \left. -F\left(\mathbf{g}^{\mathrm{T}}(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta};\boldsymbol{\beta})\mathbf{Z}|\mathbf{X}\right)\right).
\end{aligned}$$

By Proposition 1, the first term above is $o(n^{-1/2})$ and the second term is $o_p(\|\boldsymbol{\beta}-\boldsymbol{\beta}_0\|)$. The third term is actually zero since $F\left(\mathbf{g}^{\mathrm{T}}(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}_0;\boldsymbol{\beta}_0)\mathbf{Z}|\mathbf{X}\right)=\tau$. Finally, the last term is, by Taylor's expansion

$$
\begin{aligned}
&\mathbf{J}^{\mathrm{T}}E\left(\mathbf{X}\mathbf{Z}^{\mathrm{T}}\mathbf{g}'(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta};\boldsymbol{\beta})-\mathbf{H}(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta};\boldsymbol{\beta})\mathbf{Z}\right)f\left(\mathbf{g}^{\mathrm{T}}(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta};\boldsymbol{\beta})\mathbf{Z}|\mathbf{X}\right)\\
&\quad\times\left(\widehat{\mathbf{g}}^{\mathrm{T}}(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta};\boldsymbol{\beta})\mathbf{Z}-\mathbf{g}^{\mathrm{T}}(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta};\boldsymbol{\beta})\mathbf{Z}\right)\\
&\quad+O_p\left(\left(\widehat{\mathbf{g}}^{\mathrm{T}}(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta};\boldsymbol{\beta})\mathbf{Z}-\mathbf{g}^{\mathrm{T}}(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta};\boldsymbol{\beta})\mathbf{Z}\right)^2\right).
\end{aligned}
$$

The first term above is zero since $\mathbf{H}(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta};\boldsymbol{\beta})\mathbf{Z}$ is the projection of $\mathbf{X}\mathbf{Z}^{\mathrm{T}}\mathbf{g}'(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta};\boldsymbol{\beta})$ onto $\mathcal{M}_{\boldsymbol{\beta}}$ while $\widehat{\mathbf{g}}^{\mathrm{T}}(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta};\boldsymbol{\beta})\mathbf{Z}-\mathbf{g}^{\mathrm{T}}(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta};\boldsymbol{\beta})\mathbf{Z}\in\mathcal{M}_{\boldsymbol{\beta}}$. The second term above is $o_p(n^{-1/2})$ by Proposition 1.

Now we compute $P(\phi_{\boldsymbol{\beta},\mathbf{m}}-\phi_{\boldsymbol{\beta}_0,\mathbf{m}})$. We have

$$
\begin{aligned}
&E[\phi_{\boldsymbol{\beta},\mathbf{m}}-\phi_{\boldsymbol{\beta}_0,\mathbf{m}}|\mathbf{X}]\\
&=\mathbf{J}^{\mathrm{T}}\left(\mathbf{X}\mathbf{Z}^{\mathrm{T}}\mathbf{g}'(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta};\boldsymbol{\beta})-\mathbf{H}(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta};\boldsymbol{\beta})\mathbf{Z}\right)\left(\tau-F\left(\mathbf{g}^{\mathrm{T}}(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta};\boldsymbol{\beta})\mathbf{Z}\right)\right)\\
&=\mathbf{J}^{\mathrm{T}}\left(\mathbf{X}\mathbf{Z}^{\mathrm{T}}\mathbf{g}'(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta};\boldsymbol{\beta})-\mathbf{H}(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta};\boldsymbol{\beta})\mathbf{Z}\right)\left(F\left(\mathbf{g}^{\mathrm{T}}(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}_0;\boldsymbol{\beta}_0)\mathbf{Z}\right)-F\left(\mathbf{g}^{\mathrm{T}}(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta};\boldsymbol{\beta})\mathbf{Z}\right)\right)\\
&=\mathbf{J}^{\mathrm{T}}\left(\mathbf{X}\mathbf{Z}^{\mathrm{T}}\mathbf{g}'(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta};\boldsymbol{\beta})-\mathbf{H}(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta};\boldsymbol{\beta})\mathbf{Z}\right)\\
&\quad f\left(\mathbf{g}^{\mathrm{T}}(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta};\boldsymbol{\beta})\mathbf{Z}|\mathbf{X}\right)\left(\mathbf{g}^{\mathrm{T}}(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}_0;\boldsymbol{\beta}_0)-\mathbf{g}^{\mathrm{T}}(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta};\boldsymbol{\beta})\right)\mathbf{Z}+o_p\left(n^{-1/2}\right)\\
&=-\mathbf{J}^{\mathrm{T}}\left(\mathbf{X}\mathbf{Z}^{\mathrm{T}}\mathbf{g}'(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta};\boldsymbol{\beta})-\mathbf{H}(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta};\boldsymbol{\beta})\mathbf{Z}\right)f\left(\mathbf{g}^{\mathrm{T}}(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta};\boldsymbol{\beta})\mathbf{Z}|\mathbf{X}\right)\\
&\quad\frac{d\mathbf{g}^{\mathrm{T}}}{d\boldsymbol{\beta}}\mathbf{Z}(\boldsymbol{\beta}-\boldsymbol{\beta}_0)+o_p\left(n^{-1/2}\right)\\
&=-\left(\mathbf{J}^{\mathrm{T}}\left(\mathbf{X}\mathbf{Z}^{\mathrm{T}}\mathbf{g}'(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta};\boldsymbol{\beta})-\mathbf{H}(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta};\boldsymbol{\beta})\mathbf{Z}\right)\right)^{\otimes 2}f\left(\mathbf{g}^{\mathrm{T}}(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta};\boldsymbol{\beta})\mathbf{Z}|\mathbf{X}\right)\left(\boldsymbol{\beta}^{(-1)}-\boldsymbol{\beta}_0^{(-1)}\right)\\
&\quad+o_p\left(\|\boldsymbol{\beta}-\boldsymbol{\beta}_0\|\right)+o_p\left(n^{-1/2}\right)\\
&=-\left(\mathbf{J}^{\mathrm{T}}\left(\mathbf{X}\mathbf{Z}^{\mathrm{T}}\mathbf{g}'(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}_0;\boldsymbol{\beta}_0)-\mathbf{H}(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}_0;\boldsymbol{\beta}_0)\mathbf{Z}\right)\right)^{\otimes 2}f\left(\mathbf{g}^{\mathrm{T}}(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}_0)\mathbf{Z}|\mathbf{X}\right)\\
&\quad\times\left(\boldsymbol{\beta}^{(-1)}-\boldsymbol{\beta}_0^{(-1)}\right)+o_p\left(\|\boldsymbol{\beta}-\boldsymbol{\beta}_0\|\right)+o_p\left(n^{-1/2}\right),
\end{aligned}
$$

where in the second to last line, we used the identity

$$
\frac{d\mathbf{g}^{\mathrm{T}}(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta};\boldsymbol{\beta})}{d\boldsymbol{\beta}}\mathbf{Z}=\mathbf{X}\mathbf{Z}^{\mathrm{T}}\mathbf{g}'(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta};\boldsymbol{\beta})-E\left[\mathbf{X}\mathbf{Z}^{\mathrm{T}}\mathbf{g}'(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta};\boldsymbol{\beta})|\mathcal{M}_{\boldsymbol{\beta}}\right],
$$

as derived in the previous subsection.

Step 3 Let $\widetilde{\boldsymbol{\beta}}^{(-1)}=\boldsymbol{\beta}_0^{(-1)}-\boldsymbol{\Psi}_1^{-1}P_n\phi_{\boldsymbol{\beta}_0,\mathbf{m}}$, we have

$$
\mathbb{G}_n\left(\phi_{\widetilde{\boldsymbol{\beta}},\widehat{\mathbf{m}}}-\phi_{\boldsymbol{\beta}_0,\mathbf{m}}\right)=o_p(1).
$$

In fact, it is easy to see by central limit theorem that $\|\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\| = O_p\left(n^{-1/2}\right)$ and the proof is similar to Step 1 (actually only $\|\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\| = o_p(1)$ is needed here).

Step 4 $\sqrt{n} P_n \phi_{\widetilde{\boldsymbol{\beta}},\widehat{\mathbf{m}}} = o_p(1)$.

Rewriting the result in Step 3 as

$$\sqrt{n} P_n \phi_{\widetilde{\boldsymbol{\beta}},\widehat{\mathbf{m}}} = \sqrt{n} P\left(\phi_{\widetilde{\boldsymbol{\beta}},\widehat{\mathbf{m}}} - \phi_{\boldsymbol{\beta}_0,\mathbf{m}}\right) + \sqrt{n} P_n \phi_{\boldsymbol{\beta}_0,\mathbf{m}} + o_p(1),$$

using the same arguments as in Step 2, we have

$$\sqrt{n} P\left(\phi_{\widetilde{\boldsymbol{\beta}},\widehat{\mathbf{m}}} - \phi_{\boldsymbol{\beta}_0,\mathbf{m}}\right) = \boldsymbol{\Psi}_1 \sqrt{n}(\widetilde{\boldsymbol{\beta}}^{(-1)} - \boldsymbol{\beta}_0^{(-1)}) + o_p(1),$$

and thus

$$\sqrt{n} P_n \phi_{\widetilde{\boldsymbol{\beta}},\widehat{\mathbf{m}}} = \boldsymbol{\Psi}_1 \sqrt{n}\left(\widetilde{\boldsymbol{\beta}}^{(-1)} - \boldsymbol{\beta}_0^{(-1)}\right) + \sqrt{n} P_n \phi_{\boldsymbol{\beta}_0,\mathbf{m}} + o_p(1) = o_p(1),$$

by the definition of $\widetilde{\boldsymbol{\beta}}$.

Step 5 Finish the proof. Since $\widehat{\boldsymbol{\beta}}$ minimizes $\|\sqrt{n} P_n \phi_{\boldsymbol{\beta},\widehat{\mathbf{m}}}\|$ (up to an $o_p(1)$ term), we have $|\sqrt{n} P_n \phi_{\widehat{\boldsymbol{\beta}},\widehat{\mathbf{m}}}| \le |\sqrt{n} P_n \phi_{\widetilde{\boldsymbol{\beta}},\widehat{\mathbf{m}}}| + o_p(1) = o_p(1)$.

Thus (15) can be rewritten as

$$\sqrt{n} P\left(\phi_{\widehat{\boldsymbol{\beta}},\widehat{\mathbf{m}}} - \phi_{\boldsymbol{\beta}_0,\mathbf{m}}\right) = \sqrt{n} P_n \phi_{\boldsymbol{\beta}_0,\mathbf{m}} + o_p(1). \tag{17}$$

Using the result in Step 2 on the left-hand side of (17), we have

$$\sqrt{n}\left(\widehat{\boldsymbol{\beta}}^{(-1)} - \boldsymbol{\beta}_0^{(-1)}\right) = \boldsymbol{\Psi}_1^{-1} \sqrt{n} P_n \phi_{\boldsymbol{\beta}_0,\mathbf{m}} + o_p\left(\sqrt{n}\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\right)\right) + o_p(1).$$

This implies root-n consistency of $\widehat{\boldsymbol{\beta}}^{(-1)}$ as well as the asymptotic normality. □

Proposition 1 *For $\alpha > 5/2$, $\alpha' > 1/2$,*

$$\sup_{\boldsymbol{\beta}\in\mathcal{B}} \|\widehat{\mathbf{g}}(.;\boldsymbol{\beta}) - \mathbf{g}(.;\boldsymbol{\beta})\| = O_p\left(n^{-\alpha/(2\alpha+1)}(\log n)^{1/2}\right),$$
$$\sup_{\boldsymbol{\beta}\in\mathcal{B}} \left\|\widehat{\mathbf{g}}'(.;\boldsymbol{\beta}) - \mathbf{g}'(.;\boldsymbol{\beta})\right\| = O_p\left(n^{-(\alpha-1)/(2\alpha+1)}(\log n)^{1/2}\right),$$

and

$$\sup_{\boldsymbol{\beta}\in\mathcal{B}} \left\|\widehat{\mathbf{H}}(.;\boldsymbol{\beta}) - \mathbf{H}(.;\boldsymbol{\beta})\right\| = O_p\left(\max\left\{n^{-\alpha'/(2\alpha'+1)}, n^{-(\alpha-1)/(2\alpha+1)}\right\}(\log n)^{1/2}\right).$$

In particular, all rates above are of order $o_p(n^{-1/4})$. The term $(\log n)^{1/2}$ can be roughly regarded as the cost of taking supremum over $\boldsymbol{\beta} \in \mathcal{B}$.

Proof For illustration, we will only prove the first rate and the second is easily derived by the first. The third rate is also easier since $\widehat{\mathbf{H}}$ is obtained from minimizing a smooth weighted least square (unlike quantile regression). If $\mathbf{g}'$ were known, we would have the standard nonparametric rate $n^{-\alpha'/(2\alpha'+1)}$ for $\widehat{\mathbf{H}}$ (up to a logarithmic term). The other term is due to that $\mathbf{g}'$ is estimated by $\widehat{\mathbf{g}}'$, contributing a term of $O_p(n^{-(\alpha-1)/(2\alpha+1)}\}(\log n)^{1/2})$. Since the arguments are standard, the details for the rates of $\widehat{\mathbf{H}}$ are thus omitted.

Now we set out to show the uniform convergence rate of $\widehat{\mathbf{g}}$. Note that $\widehat{\mathbf{g}}(u, \boldsymbol{\beta}) = \widehat{\boldsymbol{\Theta}}^{\mathrm{T}}\mathbf{B}(u)$ where $\widehat{\boldsymbol{\Theta}} = \widehat{\boldsymbol{\Theta}}(\boldsymbol{\beta})$ is the minimizer of

$$\min_{\boldsymbol{\Theta}} \sum_i \rho_\tau \left(Y_i - \mathbf{B}^{\mathrm{T}}(\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta})\boldsymbol{\Theta}\mathbf{Z}_i \right).$$

Let $\boldsymbol{\Theta}_0 = \boldsymbol{\Theta}_0(\boldsymbol{\beta})$ be such that $\|\boldsymbol{\Theta}_0^{\mathrm{T}}\mathbf{B}(.) - \mathbf{g}(.; \boldsymbol{\beta})\|_\infty \le CK^{-d}$. Define $\boldsymbol{\theta} = \mathrm{vec}(\boldsymbol{\Theta})$, $\widehat{\boldsymbol{\theta}} = \mathrm{vec}(\widehat{\boldsymbol{\Theta}})$, $\boldsymbol{\theta}_0(\boldsymbol{\beta}) = \mathrm{vec}(\boldsymbol{\Theta}_0(\boldsymbol{\beta}))$. Note that $\mathbf{B}^{\mathrm{T}}(\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta})\boldsymbol{\Theta}\mathbf{Z}_i$ can also be written as $(\mathbf{Z}_i \otimes \mathbf{B}^{\mathrm{T}}(\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}))^{\mathrm{T}}\boldsymbol{\theta}$.

The general strategy of proof is similar to that in He and Shi (1994). However, besides that we have a single-index model instead of a simple univariate nonparametric regression, it turns out to be nontrivial to deal with supremum over $\boldsymbol{\beta}$ and this involves an important modification of the arguments used in He and Shi (1994). The proof of Proposition 1 is complete by combining Lemmas 1, 3 and 4 below. □

Define $m_i(\boldsymbol{\beta}) = \mathbf{g}^{\mathrm{T}}(\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}; \boldsymbol{\beta})\mathbf{Z}$ and $e_i(\boldsymbol{\beta}) = Y_i - m_i(\boldsymbol{\beta})$. Note that $\tau - I\{e_i(\boldsymbol{\beta}) \le 0\}$ does not have mean zero in general (unless $\boldsymbol{\beta} = \boldsymbol{\beta}_0$) but $\mathbf{Z}(\tau - I\{e_i(\boldsymbol{\beta}) \le 0\})$ still has mean zero, as in (4), which is sufficient for our purpose.

Lemma 1 *Let* $r_n = \left(\sqrt{K/n} + K^{-\alpha}\right)(\log n)^{1/2}$.

$$\begin{aligned}
\sup_{\boldsymbol{\beta}\in\mathcal{B}, \|\boldsymbol{\theta}-\boldsymbol{\theta}_0(\boldsymbol{\beta})\|\le Cr_n} & \sum_{i=1}^n \rho_\tau \left(Y_i - \left(\mathbf{Z}_i \otimes \mathbf{B}(\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta})\right)^{\mathrm{T}} \boldsymbol{\theta} \right) \\
& - \sum_{i=1}^n \rho_\tau \left(Y_i - \left(\mathbf{Z}_i \otimes \mathbf{B}(\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta})\right)^{\mathrm{T}} \boldsymbol{\theta}_0(\boldsymbol{\beta}) \right) \\
& + \sum_{i=1}^n \left(\mathbf{Z}_i \otimes \mathbf{B}(\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta})\right)^{\mathrm{T}} (\boldsymbol{\theta} - \boldsymbol{\theta}_0(\boldsymbol{\beta})) (\tau - I\{e_i(\boldsymbol{\beta}) \le 0\}) \\
& - E\sum_{i=1}^n \rho_\tau \left(Y_i - \left(\mathbf{Z}_i \otimes \mathbf{B}(\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta})\right)^{\mathrm{T}} \boldsymbol{\theta} \right) \\
& + E\sum_{i=1}^n \rho_\tau \left(Y_i - \left(\mathbf{Z}_i \otimes \mathbf{B}(\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta})\right)^{\mathrm{T}} \boldsymbol{\theta}_0(\boldsymbol{\beta}) \right) = o_p\left(nr_n^2\right),
\end{aligned}$$

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

Proof As in He and Shi (1994), in the proof we consider median regression with $\tau = 1/2$, $\rho_\tau(u) = |u|/2$ and the general case can be shown in the same way. For any $\boldsymbol{\beta} \in \mathcal{B}$, let $\mathcal{N}_{\boldsymbol{\beta}} = \left\{\boldsymbol{\theta}^{(1)}(\boldsymbol{\beta}), \ldots, \boldsymbol{\theta}^{(N)}(\boldsymbol{\beta})\right\}$ be a δ_n covering of $\{\boldsymbol{\theta} : \|\boldsymbol{\theta} - \boldsymbol{\theta}_0(\boldsymbol{\beta})\| \le Cr_n\}$, with size bounded by $N \le (C/\delta_n)^{CK}$ (see, for example, Lemma 2.5 of van der Geer (2000) for the bound) and thus $\log N \le CK \log n$ if we choose $\delta_n \sim n^{-a}$ for some $a > 0$ (we will choose a to be large enough). Let $(\boldsymbol{\beta}^{(1)}, \ldots, \boldsymbol{\beta}^{(N')})$ be a δ_n covering of $\mathcal{B}$ (it is well known that $\log N' \le C \log n$) and $\mathcal{N} = \cup_{1 \le j \le N'} \{\boldsymbol{\beta}^{(j)}\} \times \mathcal{N}_{\boldsymbol{\beta}^{(j)}}$. We denote all elements of $\mathcal{N}$ by $(\boldsymbol{\beta}_s, \boldsymbol{\theta}_s)$, $1 \le s \le S$ with $S \le CK \log n$.

Define $M_{ni}(\boldsymbol{\beta}, \boldsymbol{\theta}) = \frac{1}{2}|Y_i - (\mathbf{Z}_i \otimes \mathbf{B}(\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}))^{\mathrm{T}}\boldsymbol{\theta}| - \frac{1}{2}|Y_i - \left(\mathbf{Z}_i \otimes \mathbf{B}(\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta})\right)^{\mathrm{T}} \boldsymbol{\theta}_0(\boldsymbol{\beta})| + \left(\mathbf{Z}_i \otimes \mathbf{B}(\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta})\right)^{\mathrm{T}} (\boldsymbol{\theta} - \boldsymbol{\theta}_0(\boldsymbol{\beta}))\,(1/2 - I\{e_i(\boldsymbol{\beta}) \le 0\})$, and $M_n(\boldsymbol{\beta}, \boldsymbol{\theta}) = \sum_{i=1}^n M_{ni}(\boldsymbol{\beta}, \boldsymbol{\theta})$. Next we claim that for any $\boldsymbol{\beta}$ and any $\boldsymbol{\theta}$ with $\|\boldsymbol{\theta} - \boldsymbol{\theta}_0(\boldsymbol{\beta})\| \le Cr_n$, there exists some $\boldsymbol{\beta}_s, \boldsymbol{\theta}_s \in \mathcal{N}$ such that

$$|M_n(\boldsymbol{\beta}, \boldsymbol{\theta}) - EM_n(\boldsymbol{\beta}, \boldsymbol{\theta}) - M_n(\boldsymbol{\beta}_s, \boldsymbol{\theta}_s) + EM_n(\boldsymbol{\beta}_s, \boldsymbol{\theta}_s)| = o_p(nr_n^2). \tag{18}$$

By the construction of $\mathcal{N}$, it is obvious that we can find $(\boldsymbol{\beta}_s, \boldsymbol{\theta}_s) \in \mathcal{N}$ such that $\|\boldsymbol{\beta} - \boldsymbol{\beta}_s\| + \|\boldsymbol{\theta} - \boldsymbol{\theta}_s\| \le C\delta_n$. In He and Shi (1994), the $\boldsymbol{\beta}$ in the indicator function $I\{e_i(\boldsymbol{\beta}) \le 0\}$ in the definition of $M_n(\boldsymbol{\beta}, \boldsymbol{\beta})$ is actually $\boldsymbol{\beta}_0$, and $M_n(\boldsymbol{\beta}, \boldsymbol{\theta})$ is Lipschitz in $(\boldsymbol{\beta}, \boldsymbol{\theta})$ and thus (18) is trivially satisfied when $\delta_n \sim n^{-a}$ with a sufficiently large. Here proving (18) however is nontrivial since $I\{e_i(\boldsymbol{\beta}) \le 0\}$ is not continuous in $\boldsymbol{\beta}$. We deal with this term in Lemma 2 below. With the help of Lemma 2 dealing with the indicator function, and that all other terms in the definition of $M_n(\boldsymbol{\beta}, \boldsymbol{\theta})$ are Lipschitz continuous, we have that (18) holds.

By (18), we only need to show that

$$\sup_{(\boldsymbol{\beta}_s, \boldsymbol{\theta}_s) \in \mathcal{N}} |M_n(\boldsymbol{\beta}_s, \boldsymbol{\theta}_s) - EM_n(\boldsymbol{\beta}_s, \boldsymbol{\theta}_s)| = o_p(nr_n^2).$$

By simple algebra

$$\begin{aligned}
|M_{ni}(\boldsymbol{\beta}, \boldsymbol{\theta})| &= \Bigg|\frac{1}{2}|Y_i - \left(\mathbf{Z}_i \otimes \mathbf{B}(\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta})\right)^{\mathrm{T}} \boldsymbol{\theta}| - \frac{1}{2}|Y_i - \left(\mathbf{Z}_i \otimes \mathbf{B}(\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta})\right)^{\mathrm{T}} \boldsymbol{\theta}_0(\boldsymbol{\beta})| \\
&\quad + \left(\mathbf{Z}_i \otimes \mathbf{B}(\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta})\right)^{\mathrm{T}} (\boldsymbol{\theta} - \boldsymbol{\theta}_0(\boldsymbol{\beta}))\,(1/2 - I\{e_i \le 0\})\Bigg| \\
&= \Bigg|\frac{1}{2}|e_i(\boldsymbol{\beta}) + m_i(\boldsymbol{\beta}) - \left(\mathbf{Z}_i \otimes \mathbf{B}(\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta})\right)^{\mathrm{T}} \boldsymbol{\theta}| \\
&\quad - \frac{1}{2}|e_i(\boldsymbol{\beta}) + m_i(\boldsymbol{\beta}) - \left(\mathbf{Z}_i \otimes \mathbf{B}(\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta})\right)^{\mathrm{T}} \boldsymbol{\theta}_0(\boldsymbol{\beta})| \\
&\quad + \left(\mathbf{Z}_i \otimes \mathbf{B}(\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta})\right)^{\mathrm{T}} (\boldsymbol{\theta} - \boldsymbol{\theta}_0(\boldsymbol{\beta}))\,(1/2 - I\{e_i \le 0\})\Bigg| \\
&\le |(\mathbf{Z}_i \otimes \mathbf{B}(\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}))^{\mathrm{T}}(\boldsymbol{\theta} - \boldsymbol{\theta}_0(\boldsymbol{\beta}))| \cdot
\end{aligned}$$

$$\times I\{|e_i| \le |\left(\mathbf{Z}_i \otimes \mathbf{B}(\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta})\right)^{\mathrm{T}} (\boldsymbol{\theta} - \boldsymbol{\theta}_0(\boldsymbol{\beta}))| + |m_i(\boldsymbol{\beta}) - \left(\mathbf{Z}_i \otimes \mathbf{B}(\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta})\right)^{\mathrm{T}} \boldsymbol{\theta}_0(\boldsymbol{\beta})|\}.$$

Thus

$$|M_{ni}(\boldsymbol{\beta}, \boldsymbol{\theta})| \le C\sqrt{K} r_n =: A,$$

where we used that $\|\mathbf{B}(x)\| \le C\sqrt{K}$ at any fixed point $x \in [a, b]$.

Furthermore, we have

$$\begin{aligned} E|M_{ni}(\boldsymbol{\beta}, \boldsymbol{\theta})|^2 &\le C(\sqrt{K} r_n) E\left|\left(\mathbf{Z}_i \otimes \mathbf{B}(\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta})\right)^{\mathrm{T}} (\boldsymbol{\theta} - \boldsymbol{\theta}_0(\boldsymbol{\beta}))\right|^2 \\ &\le C(\sqrt{K} r_n)(r_n^2) =: D^2. \end{aligned} \tag{19}$$

Using Bernstein's inequality, together with union bound, we have

$$P\left(\sup_{(\boldsymbol{\beta},\boldsymbol{\theta})\in\mathcal{N}} |M_n(\boldsymbol{\beta}, \boldsymbol{\theta}) - EM_n(\boldsymbol{\beta}, \boldsymbol{\theta})| > a\right) \le C \exp\left\{-\frac{a^2}{aA + nD^2} - CK \log n\right\}.$$

The right-hand side converges to zero with $a = O\left(\max\left\{K^{3/2} r_n \log n, \sqrt{nK^{3/2} r_n^3 \log n}\right\}\right) = o\left(nr_n^2\right)$. □

Lemma 2

$$\begin{aligned} \sup_{1\le s\le S, \|\boldsymbol{\beta}-\boldsymbol{\beta}_s\|+\|\boldsymbol{\theta}-\boldsymbol{\theta}_s\|\le\delta_n} &\left|\left(\mathbf{Z}_i \otimes \mathbf{B}(\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta})\right)^{\mathrm{T}} (\boldsymbol{\theta} - \boldsymbol{\theta}_0(\boldsymbol{\beta}))\, I\left\{e_i(\boldsymbol{\beta}) \le 0\right\}\right. \\ &\left. - \left(\mathbf{Z}_i \otimes \mathbf{B}(\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}_s)\right)^{\mathrm{T}} \left(\boldsymbol{\theta}_s - \boldsymbol{\theta}_0(\boldsymbol{\beta}_s)\right) I\left\{e_i(\boldsymbol{\beta}_s) \le 0\right\}\right| \\ &= o_p\left(nr_n^2\right). \end{aligned}$$

Proof Writing $W_i(\boldsymbol{\beta}, \boldsymbol{\theta}) = \left(\mathbf{Z}_i \otimes \mathbf{B}(\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta})\right)^{\mathrm{T}} (\boldsymbol{\theta} - \boldsymbol{\theta}_0(\boldsymbol{\beta}))$, we only need to show that

$$\begin{aligned} \sup_{1\le s\le S, \|\boldsymbol{\beta}-\boldsymbol{\beta}_s\|+\|\boldsymbol{\theta}-\boldsymbol{\theta}_s\|\le\delta_n} &\left|\sum_i W_i(\boldsymbol{\beta}, \boldsymbol{\theta})(I\left\{e_i(\boldsymbol{\beta}) \le 0\right\} - F\left(m_i(\boldsymbol{\beta})|\mathbf{X}_i\right)\right. \\ &\left. - \sum_i W_i(\boldsymbol{\beta}_s, \boldsymbol{\theta}_s)\left(I\left\{e_i(\boldsymbol{\beta}_s) \le 0\right\} - F\left(m_i(\boldsymbol{\beta}_s)|\mathbf{X}_i\right)\right)\right| \\ &= o_p\left(nr_n^2\right). \end{aligned} \tag{20}$$

Obviously (20) is implied by

$$\sup_{1\le s\le S,\|\boldsymbol{\beta}-\boldsymbol{\beta}_s\|+\|\boldsymbol{\theta}-\boldsymbol{\theta}_s\|\le\delta_n}\Big|\sum_i W_i(\boldsymbol{\beta}_s,\boldsymbol{\theta}_s)\left(I\left\{e_i(\boldsymbol{\beta})\le 0\right\}\right.$$
$$\left.-I\left\{e_i(\boldsymbol{\beta}_s)\le 0\right\}-F\left(m_i(\boldsymbol{\beta})|\mathbf{X}_i\right)+F\left(m_i(\boldsymbol{\beta}_s)|\mathbf{X}_i\right)\right)\Big|$$
$$=o_p\left(nr_n^2\right). \tag{21}$$

Assume $W_i(\boldsymbol{\beta}_s,\boldsymbol{\theta}_s)>0$ for now and we first show (21) without the absolute value on the left-hand side. Since $\|\boldsymbol{\beta}-\boldsymbol{\beta}_s\|\le\delta_n$, by our assumption we have $|e_i(\boldsymbol{\beta})-e_i(\boldsymbol{\beta}_s)|\le C\delta_n$. By the monotonicity of the function $t\to I\left\{e_i(\boldsymbol{\beta})\le t\right\}$, the left-hand side of (21) is bounded by

$$\sum_i W_i(\boldsymbol{\beta}_s,\boldsymbol{\theta}_s)\left(I\left\{e_i(\boldsymbol{\beta}_s)\le C\delta_n\right\}-I\left\{e_i(\boldsymbol{\beta}_s)\le 0\right\}\right.$$
$$\left.-F\left(m_i(\boldsymbol{\beta})|\mathbf{X}_i\right)+F\left(m_i(\boldsymbol{\beta}_s)|\mathbf{X}_i\right)\right)$$
$$=\sum_i W_i\left(\boldsymbol{\beta}_s,\boldsymbol{\theta}_s\right)(I\left\{e_i(\boldsymbol{\beta}_s)\le C\delta_n\right\}-I\left\{e_i(\boldsymbol{\beta}_s)\le 0\right\}$$
$$-F\left(m_i(\boldsymbol{\beta}_s)+C\delta_n|\mathbf{X}_i\right)+F\left(m_i(\boldsymbol{\beta}_s)|\mathbf{X}_i\right))$$
$$+\sum_i W_i(\boldsymbol{\beta}_s,\boldsymbol{\theta}_s)\left(F\left(m_i(\boldsymbol{\beta}_s)+C\delta_n|\mathbf{X}_i\right)-F\left(m_i(\boldsymbol{\beta})|\mathbf{X}_i\right)\right). \tag{22}$$

The first term of (22) is $o_p\left(nr_n^2\right)$ which follows easily from Bernstein's inequality, the union bound, and that $\delta_n\sim n^{-a}$ for a sufficiently large. The second term of (22) is also $o_p\left(nr_n^2\right)$ since $\|\boldsymbol{\beta}-\boldsymbol{\beta}_s\|\le\delta_n$. Obviously, using $I\left\{e_i(\boldsymbol{\beta})\le 0\right\}\ge I\left\{e_i(\boldsymbol{\beta}_s)\le -C\delta_n\right\}$, we can also show that (22) is $o_p\left(nr_n^2\right)$ if we change the sign.

So far we have assumed that $W_i(\boldsymbol{\beta}_s,\boldsymbol{\theta}_s)>0$. In general, we can consider the positive part and the negative part of $W_i(\boldsymbol{\beta}_s,\boldsymbol{\theta}_s)$ separately and the proof is complete. □

Lemma 3 *For $L>0$ large enough*

$$\inf_{\boldsymbol{\beta}\in\mathcal{B},\|\boldsymbol{\theta}-\boldsymbol{\theta}_0(\boldsymbol{\beta})\|=Lr_n}\sum_i E\rho_\tau\left(e_i(\boldsymbol{\beta})+m_i(\boldsymbol{\beta})-\left(\mathbf{Z}_i\otimes\mathbf{B}(\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta})\right)^{\mathrm{T}}\boldsymbol{\theta}\right)$$
$$-\sum_i E\rho_\tau\left(e_i(\boldsymbol{\beta})+m_i(\boldsymbol{\beta})-\left(\mathbf{Z}_i\otimes\mathbf{B}(\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta})\right)^{\mathrm{T}}\boldsymbol{\theta}_0(\boldsymbol{\beta})\right)$$
$$\ge L^2Cnr_n^2.$$

Proof Applying the Knight's identity $\rho_\tau(x-y)-\rho_\tau(x)=-y\left(\tau-I\{x\le 0\}\right)+\int_0^y\left(I\{x\le t\}-I\{x\le 0\}\right)dt$ twice on the two terms, we have that

$$E\sum_{i=1}^{n}\rho_\tau\left(e_i(\boldsymbol{\beta})+m_i(\boldsymbol{\beta})-\left(\mathbf{Z}_i\otimes\mathbf{B}(\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta})\right)^{\mathrm{T}}\boldsymbol{\theta}\right)$$
$$-E\sum_{i=1}^{n}\rho_\tau\left(e_i(\boldsymbol{\beta})+m_i(\boldsymbol{\beta})-\left(\mathbf{Z}_i\otimes\mathbf{B}(\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta})\right)^{\mathrm{T}}\boldsymbol{\theta}_0(\boldsymbol{\beta})\right)$$
$$=\sum_i\int_{(\mathbf{Z}_i\otimes\mathbf{B}(\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}))^{\mathrm{T}}\boldsymbol{\theta}_0(\boldsymbol{\beta})-m_i(\boldsymbol{\beta})}^{(\mathbf{Z}_i\otimes\mathbf{B}(\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}))^{\mathrm{T}}\boldsymbol{\theta}-m_i(\boldsymbol{\beta})}F\left(m_i(\boldsymbol{\beta})+t|\mathbf{X}_i\right)-F\left(m_i(\boldsymbol{\beta})|\mathbf{X}_i\right)dt$$
$$\geq C\sum_i f\left(m_i(\boldsymbol{\beta})|\mathbf{X}_i\right)\left[\left(\left(\mathbf{Z}_i\otimes\mathbf{B}(\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta})\right)^{\mathrm{T}}(\boldsymbol{\theta}-\boldsymbol{\theta}_0(\boldsymbol{\beta}))\right)^2\right.$$
$$\left.+2\left(\left(\mathbf{Z}_i\otimes\mathbf{B}(\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta})\right)^{\mathrm{T}}(\boldsymbol{\theta}-\boldsymbol{\theta}_0(\boldsymbol{\beta}))\right)\left(\left(\mathbf{Z}_i\otimes\mathbf{B}(\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta})\right)^{\mathrm{T}}\boldsymbol{\theta}_0(\boldsymbol{\beta})-m_i(\boldsymbol{\beta})\right)\right].$$

Combining

$$\sum_i\left(\left(\mathbf{Z}_i\otimes\mathbf{B}(\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta})\right)^{\mathrm{T}}(\boldsymbol{\theta}-\boldsymbol{\theta}_0(\boldsymbol{\beta}))\right)^2\geq CL^2nr_n^2,$$

and

$$\sum_i\left(\left(\mathbf{Z}_i\otimes\mathbf{B}(\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta})\right)^{\mathrm{T}}(\boldsymbol{\theta}-\boldsymbol{\theta}_0(\boldsymbol{\beta}))\right)\left(\left(\mathbf{Z}_i\otimes\mathbf{B}(\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta})\right)^{\mathrm{T}}\boldsymbol{\theta}_0(\boldsymbol{\beta})-m_i(\boldsymbol{\beta})\right)$$
$$\leq CLnr_nK^{-d},$$

we get the statement of the lemma if L is large enough. □

Lemma 4

$$\sup_{\boldsymbol{\beta}\in\mathcal{B},\|\boldsymbol{\theta}-\boldsymbol{\theta}_0(\boldsymbol{\beta})\|=Lr_n}\sum_i\left(\mathbf{Z}_i\otimes\mathbf{B}(\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta})\right)^{\mathrm{T}}(\boldsymbol{\theta}-\boldsymbol{\theta}_0)\left(\tau-I\left\{e_i(\boldsymbol{\beta})\leq 0\right\}\right)$$
$$=L\cdot O_p\left(nr_n^2\right).$$

Proof By Lemma 2, we only need to consider supremum over $(\boldsymbol{\beta}_s,\boldsymbol{\theta}_s)\in\mathcal{N}$. For fixed $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$, using $\sum_i\left(\left(\mathbf{Z}_i\otimes\mathbf{B}(\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta})\right)^{\mathrm{T}}(\boldsymbol{\theta}-\boldsymbol{\theta}_0(\boldsymbol{\beta}))\right)^2=O_p\left(L^2nr_n^2\right)$, and $|\left(\mathbf{Z}_i\otimes\mathbf{B}(\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta})\right)^{\mathrm{T}}(\boldsymbol{\theta}-\boldsymbol{\theta}_0(\boldsymbol{\beta}))|\leq CLr_n\sqrt{K}$, we have, by Bernstein's inequality,

$$P\left(\sum_i\left(\mathbf{Z}_i\otimes\mathbf{B}(\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta})\right)^{\mathrm{T}}(\boldsymbol{\theta}-\boldsymbol{\theta}_0)\left(\tau-I\left\{e_i(\boldsymbol{\beta})\leq 0\right\}\right)>a\right\}$$
$$\leq C\exp\left\{-\frac{a^2}{aL\sqrt{K}r_n+nL^2r_n^2}\right\}.$$

Thus

$$P\left(\sup_{(\boldsymbol{\beta},\boldsymbol{\theta})\in\mathcal{N}}\sum_i\left(\mathbf{Z}_i\otimes\mathbf{B}(\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta})\right)^{\mathrm{T}}(\boldsymbol{\theta}-\boldsymbol{\theta}_0)\left(\tau-I\left\{e_i(\boldsymbol{\beta})\le 0\right\}\right)>a\right\}$$
$$\le C\exp\left\{-\frac{a^2}{aL\sqrt{K}r_n+nL^2r_n^2}-CK\log n\right\},$$

which implies $\sup_{(\boldsymbol{\beta},\boldsymbol{\theta})\in\mathcal{N}}\sum_i\left(\mathbf{Z}_i\otimes\mathbf{B}(\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta})\right)^{\mathrm{T}}(\boldsymbol{\theta}-\boldsymbol{\theta}_0)(\tau-I\left\{e_i(\boldsymbol{\beta})\le 0\right\}) = LO_p(\sqrt{n}r_n\sqrt{K\log n}) = LO_p\left(nr_n^2\right)$. □

References

Belloni, A., Chernozhukov, V. (2011). l1-penalized quantile regression in high-dimensional sparse models. *The Annals of Statistics*, *39*, 82–130.

Bondell, H. D., Reich, B. J., Wang, H. (2010). Noncrossing quantile regression curve estimation. *Biometrika*, *97*, 825–838.

Cai, Z., Xiao, Z. (2012). Semiparametric quantile regression estimation in dynamic models with partially varying coefficients. *Journal of Econometrics*, *167*, 413–425.

Carroll, R. J., Fan, J., Gijbels, I., Wand, M. P. (1997). Generalized partially linear single-index models. *Journal of the American Statistical Association*, *92*, 477–489.

Chen, R., Tsay, R. S. (1993). Functional-coefficient autoregressive models. *Journal of the American Statistical Association*, *88*, 298–308.

Cortez, P., Morais, A. (2007). A data mining approach to predict forest fires using meteorological data. In J. Neves, M. F. Santos and J. Machado (Eds.), *New trends in artificial intelligence, Proceedings of the 13th EPIA 2007 - Portuguese Conference on Artificial Intelligence*, 512–523.

Cui, X., Haerdle, W. K., Zhu, L. (2011). The efm approach for single-index models. *The Annals of Statistics*, *39*, 1658–1688.

Fan, J., Fan, Y., Barut, E. (2014a). Adaptive robust variable selection. *Annals of Statistics*, *42*, 324–351.

Fan, J., Ma, Y., Dai, W. (2014b). Nonparametric independence screening in sparse ultra-high dimensional varying coefficient models. *Journal of the American Statistical Association*, *109*, 1270–1284.

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.

Fan, J. Q., Zhang, J. T. (2000). Two-step estimation of functional linear models with applications to longitudinal data. *Journal of the Royal Statistical Society Series B-Statistical Methodology*, *62*, 303–322.

Fan, J. Q., Zhang, W. Y. (1999). Statistical estimation in varying coefficient models. *Annals of Statistics*, *27*, 1491–1518.

Hall, P., Sheather, S. J. (1988). On the distribution of a studentized quantile. *Journal of the Royal Statistical Society*: *Series B (Methodological)*, *50*, 381–391.

Hastie, T., Tibshirani, R. (1993). Varying-coefficient models. *Journal of the Royal Statistical Society Series B-Methodological*, *55*, 757–796.

He, X., Shi, P. (1994). Convergence rate of b-spline estimators of nonparametric conditional quantile functions. *Journal of Nonparametric Statistics*, *3*, 299–308.

Hendricks, W., Koenker, R. (1992). Hierarchical spline models for conditional quantiles and the demand for electricity. *Journal of the American Statistical Association*, *87*, 58–68.

Hoover, D. R., Rice, J. A., Wu, C. O., Yang, L. P. (1998). Nonparametric smoothing estimates of time-varying coefficient models with longitudinal data. *Biometrika*, *85*, 809–822.

Horowitz, J. L., Lee, S. (2005). Nonparametric estimation of an additive quantile regression model. *Journal of the American Statistical Association*, *100*, 1238–1249.

Hu, Y., Gramacy, R., Lian, H. (2013). Bayesian quantile regression for single-index models. *Statistics and Computing*, *23*, 437–454.

Huang, J. H. Z., Wu, C. O., Zhou, L. (2002). Varying-coefficient models and basis function approximations for the analysis of repeated measurements. *Biometrika*, *89*, 111–128.

Jiang, L., Wang, H. J., Bondell, H. D. (2013). Interquantile shrinkage in regression models. *Journal of Computational and Graphical Statistics*, *22*, 970–986.
Kim, M. (2007). Quantile regression with varying coefficients. *Annals of Statistics*, *35*, 92–108.
Koenker, R., Bassett, G, Jr. (1978). Regression quantiles. *Econometrica*: *Journal of the Econometric Society*, *1*, 33–50.
Koenker, R., Ng, P., Portnoy, S. (1994). Quantile smoothing splines. *Biometrika*, *81*, 673–680.
Kong, E., Xia, Y. (2012). A single-index quantile regression model and its estimation. *Econometric Theory*, *28*, 730–768.
Kottas, A., Krnjajic, M. (2009). Bayesian semiparametric modelling in quantile regression. *Scandinavian Journal of Statistics*, *36*, 297–319.
Lai, P., Wang, Q., Lian, H. (2012). Bias-corrected gee estimation and smooth-threshold gee variable selection for single-index models with clustered data. *Journal of Multivariate Analysis*, *105*, 422–432.
Lee, S. (2003). Efficient semiparametric estimation of a partially linear quantile regression model. *Econometric Theory*, *19*, 1–31.
Lian, H. (2012a). Semiparametric estimation of additive quantile regression models by two-fold penalty. *Journal of Business & Economic Statistics*, *30*, 337–350.
Lian, H. (2012b). Variable selection for high-dimensional generalized varying-coefficient models. *Statistica Sinica*, *22*, 1563–1588.
Liu, J., Li, R., Wu, R. (2014). Feature selection for varying coefficient models with ultrahigh-dimensional covariates. *Journal of the American Statistical Association*, *109*, 266–274.
Lu, Z., Tjstheim, D., Yao, Q. (2007). Adaptive varying-coefficient linear models for stochastic processes: asymptotic theory. *Statistica Sinica*, *17*, 177–198.
Reich, B., Bondell, H., Wang, H. (2010). Flexible bayesian quantile regression for independent and clustered data. *Biostatistics*, *11*, 337–352.
Sherwood, B., Wang, L. (2016). Partially linear additive quantile regression in ultra-high dimension. *Annals of Statistics*, *44*, 288–317.
Tang, Y., Wang, H. J., Zhu, Z. (2013). Variable selection in quantile varying coefficient models with longitudinal data. *Computational Statistics & Data Analysis*, *57*, 435–449.
Tokdar, S., Kadane, J. B. (2011). Simultaneous linear quantile regression: A semiparametric bayesian approach. *Bayesian Analysis*, *6*, 1–22.
van der Geer, S. A. (2000). *Empirical processes in M-estimation*. Cambridge: Cambridge University Press.
van der Vaart, A. W. (1998). *Asymptotic statistics*. Cambridge: Cambridge University Press.
van der Vaart, A. W., Wellner, J. A. (1996). *Weak convergence and empirical processes*. New York: Springer.
Wang, H. J., Zhu, Z., Zhou, J. (2009). Quantile regression in partially linear varying coefficient models. *The Annals of Statistics*, *37*, 3841–3866.
Wang, J.-L., Xue, L., Zhu, L., Chong, Y. S. (2010). Estimation for a partial-linear single-index model. *Annals of Statistics*, *38*, 246–274.
Wang, L., Wu, Y., Li, R. (2012). Quantile regression for analyzing heterogeneity in ultra-high dimension. *Journal of the American Statistical Association*, *107*, 214–222.
Wei, F., Huang, J., Li, H. Z. (2011). Variable selection and estimation in high-dimensional varying-coefficient models. *Statistica Sinica*, *21*, 1515–1540.
Wu, T. Z., Yu, K., Yu, Y. (2010). Single-index quantile regression. *Journal of Multivariate Analysis*, *101*, 1607–1621.
Wu, Y., Liu, Y. (2009). Variable selection in quantile regression. *Statistica Sinica*, *19*, 801–817.
Xia, Y., Li, W. (1999). On single-index coefficient regression models. *Journal of the American Statistical Association*, *94*, 1275–1285.
Xue, L., Qu, A. (2012). Variable selection in high-dimensional varying-coefficient models with global optimality. *The Journal of Machine Learning Research*, *13*, 1973–1998.
Yang, Y., He, X. (2012). Bayesian empirical likelihood for quantile regression. *The Annals of Statistics*, *40*, 1102–1131.
Yu, K., Jones, M. (1998). Local linear quantile regression. *Journal of the American Statistical Association*, *93*, 228–237.
Yu, K., Moyeed, R. (2001). Bayesian quantile regression. *Statistics & Probability Letters*, *54*, 437–447.
Yu, Y., Ruppert, D. (2002). Penalized spline estimation for partially linear single-index models. *Journal of the American Statistical Association*, *97*, 1042–1054.
Zhu, L., Huang, M., Li, R. (2012). Semiparametric quantile regression with high-dimensional covariates. *Statistica Sinica*, *22*, 1379–1401.

Zhu, L., Lin, L., Cui, X., Li, G. (2010). Bias-corrected empirical likelihood in a multi-link semiparametric model. *Journal of Multivariate Analysis*, *101*, 850–868.