

# Local Polynomial Quantile Regression With Parametric Features

Anouar EL GHOUGH and Marc G. GENTON

---

We propose a new approach to conditional quantile function estimation that combines both parametric and nonparametric techniques. At each design point, a global, possibly incorrect, pilot parametric model is locally adjusted through a kernel smoothing fit. The resulting quantile regression estimator behaves like a parametric estimator when the latter is correct and converges to the nonparametric solution as the parametric start deviates from the true underlying model. We give a Bahadur-type representation of the proposed estimator from which consistency and asymptotic normality are derived under an  $\alpha$ -mixing assumption. We also propose a practical bandwidth selector based on the plug-in principle and discuss the numerical implementation of the new estimator. Finally, we investigate the performance of the proposed method via simulations and illustrate the methodology with a data example.

KEY WORDS: Bias reduction; Local polynomial smoothing; Model misspecification; Robustness; Strong mixing sequence.

---

## 1. INTRODUCTION

It is known from the literature that regression function estimators based on least squares are optimal and are equivalent to the maximum likelihood estimators when errors follow a normal distribution. However, in many non-Gaussian (i.e., skewed or heavy-tailed) situations, they are far from optimal and also are very sensitive to modest amounts of outlier contamination. An attractive alternative to the classical regression approach based on the quadratic loss function is the use of the absolute error criterion, which leads to the well-known median regression function or, more generally, the quantile regression method. Since it was introduced by Koenker and Bassett (1978) as a robust (to outliers) and flexible (to error distribution) linear regression method, quantile regression has received considerable interest in both theoretical and applied statistics (see Koenker 2005 and references therein).

### 1.1 Bias Reduction in Kernel Smoothing

Unlike parametric techniques, nonparametric kernel smoothing techniques are well-known flexible methods that can be used without making restrictive assumptions about the form of the unknown target function. In general, their performance depends on the smoothness of the regression function, the sample size  $n$ , the selected kernel (density function), and the bandwidth  $h_n > 0$  that describes the degree of smoothing applied to the data. Many kernel smoothers, including Nadaraya-Watson, local linear (LL), and nearest-neighbor, share the same form for the asymptotic mean squared error (MSE), namely  $(h_n^2 a)^2 + (nh_n)^{-1} b$ , where the first term is the squared asymptotic bias and the second term is the asymptotic variance. The quantities  $a$  and  $b$  depend on the unknown data-generating procedure and on the chosen kernel, but not on  $n$  or  $h_n$ . From this formula, it is clear that for a fixed  $n$ , the bias can be reduced

simply by choosing a small bandwidth ( $h_n \rightarrow 0$ ), although this inevitably will increase the variance of the estimator. At least two approaches have been proposed in the literature to reduce the bias without increasing the instability of the resulting estimator. The first approach aims to improve the bias rate from  $O(h_n^2)$  to  $O(h_n^4)$  by using, for example, higher-order kernels or variable kernel methods. As noted by Jones and Signorini (1997) in the context of density estimation, the merit of such an approach is not clear for finite (small to moderate) sample sizes. The second approach, which we use in the present work, attempts to remove the bias asymptotically by acting only on the leading constant term  $a$  without changing the variance of the estimator. This is particularly interesting, because the decrease in bias allows an increase in the bandwidth and thus the use of more data in the local fit, which also will be beneficial in reducing the variance. One of the most widely used techniques to achieve this goal is to guide the nonparametric regression function by a parametric pilot estimate. To be more precise, denote by  $m(x)$  the unknown objective function and by  $m(x, \theta)$  a given parametric model. Based on the available sample data, start by estimating  $\theta$  by  $\hat{\theta}$ , say, and then plug it into  $m(\cdot, \theta)$  to get  $m(\cdot, \hat{\theta})$ , a parametric global estimator for  $m$ . Even if the parametric model is not adequate throughout the entire range of the data, which is likely the case in practice,  $m(\cdot, \hat{\theta})$  should contain some useful information about  $m$ . Locally and only in regions where  $m(x, \hat{\theta})$  seems to not conform to  $m(x)$ , a kernel smoother, say  $\hat{m}$ , that relies totally on the data, should intervene and adjust the primary approximation, which can be seen as a Bayesian prior. Ideally, the final estimator should take advantage of both parametric and nonparametric methods: it should never do worse than the corresponding purely nonparametric regressor, but it should also adapt automatically to the parametric model if the latter is locally or globally closer to the true underlying curve. An obvious technique for combining parametric and nonparametric fits is to mix them linearly, that is,  $\lambda m(x, \hat{\theta}) + (1 - \lambda)\hat{m}(x)$ . The balance between the two methods is controlled by  $\lambda \in [0, 1]$ , which is a smoothing parameter

---

Anouar El Ghouh is Postdoctoral Fellow, Department of Econometrics, University of Geneva, CH-1211 Geneva 4, Switzerland (E-mail: [Anouar.ElGhouch@metri.unige.ch](mailto:Anouar.ElGhouch@metri.unige.ch)). Marc G. Genton is Professor, Department of Statistics, Texas A&M University, College Station, TX 77843-3143 (E-mail: [genton@stat.tamu.edu](mailto:genton@stat.tamu.edu)). Financial support from the Swiss National Science Foundation (project 116019) is gratefully acknowledged. Genton's research was supported in part by National Science Foundation grants DMS-0504896 and CMG ATM-0620624 and by King Abdullah University of Science and Technology award KUS-C1-016-04. The authors thank the editor, an associate editor, and two anonymous referees for their valuable comments.

that also must be estimated. This method was proposed by Einsporn (1987) and further studied and extended by Fan and Ullah (1999) and Mays, Birch, and Starnes (2001). Recent related work includes that of Glad (1998), Gozalo and Linton (2000), Naito (2004), and Hagmann and Scaillet (2007).

### 1.2 Performance Improvement of Quantile Regression

To the best of our knowledge, the only results in the literature regarding combining different estimators in the context of quantile regression have been published by Su and Ullah (2008) and Shan and Yang (2009). Su and Ullah (2008) proposed a double-smoothing estimator in which a LL primary fit is multiplicatively adjusted by another LL fit. In their report on a quantile model that performs uniformly better in the whole probability interval  $(0, 1)$ , Shan and Yang (2009) proposed pooling together many quantile estimators using a weight function based either on the check loss or on a mixture of check and quadratic loss. Their method includes a tuning parameter,  $\lambda$ , that when chosen correctly leads to optimal performance in terms of an oracle inequality. Our approach and motivations here are completely different. We are particularly interested in the case where a naive parametric estimator (that may be completely misspecified) is available but either fails to adequately fit the observed data or casts doubt on the data's accuracy and efficacy. The parametric estimator is then corrected additively through a  $p$ th-order local polynomial quantile regressor. These considerations yield a consistent and substantially better estimate of the underlying conditional quantile function and its derivatives with a single bandwidth. This bandwidth not only controls the local window size, as is the case for the classical kernel methods, but also adapts the local fit to the global parametric model. Another advantage of the proposed method is that our estimator can be seen as a generalization of the classical local polynomial fit. It shares with the well-known LL smoother ( $p = 1$ ) some good properties, such as small boundary effects, adaptive design, and high minimax efficiency; however, it typically has a smaller MSE and a faster rate of convergence. It also should be noted that, to the best of our knowledge, no such procedure is available in the literature for local polynomial quantile fitting either in the context of strong mixing data, as considered in the present work, or for the special case of LL fit with independent and identically distributed (iid) data samples.

This article is organized as follows. In the next section we describe the estimation methodology and define our guided local polynomial quantile (GLPQ) estimator. In Section 3 we examine some asymptotic results of the proposed approach, including Bahadur representation, consistency, and asymptotic normality. In Section 4 we discuss the problem of choosing the smoothing parameter and suggest a new data-driven procedure based on the plug-in idea. In Section 5 we analyze the finite-sample performance of the proposed estimator via a simulation study using both the asymptotic optimal bandwidth and our data-driven bandwidth. In Section 6 we give an empirical application that illustrates the proposed method. We provide proofs of the asymptotic results in the Appendix.

## 2. THE GUIDED LOCAL POLYNOMIAL QUANTILE ESTIMATOR

The data under consideration here comprises a set of  $n$  replications  $(X_i, Y_i)$  of the bivariate random vectors  $(X, Y)$ , where  $Y$  is the variable of interest and  $X$  is some covariate. The objective function is given by  $Q_\pi(x) = \inf\{t: F_x(t) \geq \pi\}$ , where  $F_x(t)$  is the common conditional distribution function of  $Y|X = x$  and  $\pi \in (0, 1)$ . Equivalently, the quantile function can be written as  $Q_\pi(x) = \arg \min_a \mathbb{E}_x(\varphi_\pi(Y - a))$ , where, from now on,  $\mathbb{E}_x(\cdot)$  represents a shortcut for  $\mathbb{E}(\cdot|X = x)$ ,  $\varphi_\pi(s) = s(\pi - I(s < 0))$  is the check loss, and  $I(\cdot)$  is the usual indicator function. The case where  $\pi = 0.5$  is well known to researchers as the least absolute deviation (LAD) or median regression. The local polynomial (LP) estimator is based on the following Taylor approximation of  $Q_\pi(X_i)$  in the neighborhood of  $x$ :  $Q_\pi(X_i) \approx \sum_{j=0}^p \frac{Q_\pi^{(j)}(x)}{j!} (X_i - x)^j \equiv \tilde{\mathbf{X}}_i^T \boldsymbol{\beta}$ , where  $Q_\pi^{(j)}$  denotes the  $j$ th derivative of  $Q_\pi$ ,  $\tilde{\mathbf{X}}_i = (1, X_i - x, \dots, (X_i - x)^p)^T$ ,  $\boldsymbol{\beta} = (\beta_0, \dots, \beta_p)^T$ , with  $\beta_j = Q_\pi^{(j)}(x)/j!$ , for  $j = 0, \dots, p$ . Note that  $\boldsymbol{\beta}$  depends on  $\pi$  and  $x$ , but we omit this here for notational convenience. The LP estimator  $\hat{\boldsymbol{\beta}} = (\hat{\beta}_0, \dots, \hat{\beta}_p)^T$  is defined by

$$\arg \min_{\mathbf{b} \in \mathbb{R}^{p+1}} \sum_{i=1}^n \varphi_\pi(Y_i - \tilde{\mathbf{X}}_i^T \mathbf{b}) K_{h,i}, \tag{1}$$

where  $K_{h,i} = K((X_i - x)/h_n)$ , with  $K$  a kernel function. Via the weights  $K_{h,i}$ , only the points  $X_i$  near  $x$  contribute effectively to the estimation of  $\boldsymbol{\beta}$ . The LL case ( $p = 1$ ) was studied by Fan, Hu, and Truong (1994) and Yu and Jones (1998).

To motivate our approach, assume that instead of  $Q_\pi(x)$ , we are interested in  $Q_\pi(x) - q(x) \equiv \arg \min_a \mathbb{E}_x(\varphi_\pi(Y - q(X) - a))$ , for some given function  $q$ . One can first estimate  $Q_\pi$  and then subtract  $q$  from it, or directly search the argument that minimizes  $\sum_{i=1}^n \varphi_\pi(Y_i - q(X_i) - \tilde{\mathbf{X}}_i^T \mathbf{b}) K_{h,i}$  with respect to  $\mathbf{b} \in \mathbb{R}^{p+1}$ . An obvious way to get back to  $Q_\pi(x)$  is by minimizing  $\sum_{i=1}^n \varphi_\pi(Y_i - (q(X_i) - q(x)) - \tilde{\mathbf{X}}_i^T \mathbf{b}) K_{h,i}$ . This can be shown to be a valid estimator for  $Q_\pi(x)$  and its derivatives up to the  $p$ th order; however, for reasons that we make clear later, here we suggest replacing  $q(X_i) - q(x)$  with  $r_q(X_i) := q(X_i) - \sum_{j=0}^p \frac{q^{(j)}(x)}{j!} (X_i - x)^j$  in the last equation, provided that  $q^{(p)}(x)$  exists. This leads to a new class of LP estimators given by

$$\arg \min_{\mathbf{b} \in \mathbb{R}^{p+1}} \sum_{i=1}^n \varphi_\pi(Y_i - r_q(X_i) - \tilde{\mathbf{X}}_i^T \mathbf{b}) K_{h,i}, \tag{2}$$

where  $q$  may be any pilot function.

*Remark 1.* If  $q$  is a polynomial function of degree  $d$ , then  $q(X_i) = \sum_{j=0}^d \frac{q^{(j)}(x)}{j!} (X_i - x)^j$ . This implies that for  $d \leq p$ ,  $r_q(X_i) = 0$  and so (1) coincides with (2). This remark has two consequences. First, the classical LP fit is a special case of our estimator. Second, to capture some features of the underlying model through a polynomial start, the latter should be of a high order, i.e.  $d > p$ . In this case,  $r_q(X_i)$  becomes  $\sum_{j=p+1}^d \frac{q^{(j)}(x)}{j!} (X_i - x)^j$  and it will necessarily influence the resulting estimator.

Although any function  $q$  can be used, in practice the user has to make an appropriate choice. This can be done by specifying a parametric model in the following way. For a fixed  $\pi$ , let  $q_\pi(x, \theta) \equiv q_\theta(x)$  be a model for the conditional  $\pi$ -quantile function, where  $\theta$  is a  $d$ -dimensional parameter vector in  $\Theta \subset \mathbb{R}^d$ . There are no special considerations made on the family  $\{q_\pi(x, \theta), \theta \in \Theta\}$ , although it should be constructed by taking into account all the previous knowledge (if any) about the underlying structure. Using the available data and minimizing a certain distance between  $Q_\pi(\cdot)$  and  $q_\pi(\cdot, \theta)$ , an estimator  $\hat{\theta}$  for  $\theta$  is obtained and then plugged into  $q_\pi$ . In our context,  $q_\pi(x, \hat{\theta})$  and its derivatives are considered as a crude approximation for  $Q_\pi^{(j)}(x), j = 0, \dots, p$ . Here as an estimator for  $\theta$  we recommend  $\hat{\theta} = \arg \min_{\theta \in \Theta} n^{-1} \sum_{i=1}^n \varphi_\pi(Y_i - q_\theta(X_i))$ . This approach is largely used in the literature and was originally proposed by Koenker and Bassett (1978). Even if the considered model  $q_\theta$  is incorrectly specified, i.e., there is no  $\theta \in \Theta$  such that  $\mathbb{E}(I(Y - q_\theta(X) \leq 0)) = \pi$ , and under very weak assumptions, see Komunjer (2005) and also Oberhofer and Haupt (2009),  $\hat{\theta}$  converges in probability to  $\theta^* = \arg \min_{\theta \in \Theta} \mathbb{E}(\varphi_\pi(Y - q_\theta(X)))$ . The latter parameter is the best possible value of  $\theta \in \Theta$  with respect to the “distance”  $\varphi_\pi$ .

To conclude, our guided LP quantile (GLPQ) estimator is the minimizer  $\hat{\beta} = (\hat{\beta}_0, \dots, \hat{\beta}_p)^T$ , with respect to  $\mathbf{b} \in \mathbb{R}^{p+1}$ , of

$$\sum_{i=1}^n \varphi_\pi(Y_i - r_i(\hat{\theta}) - \tilde{\mathbf{X}}_i^T \mathbf{b}) K_{h,i}, \tag{3}$$

where  $r_i(\hat{\theta})$  is a shortcut for  $r_{q_{\hat{\theta}}}(X_i)$ . Later we show that using this first estimation step [i.e.,  $r_i(\hat{\theta})$ ] instead of  $r_i(\theta^*)$  has asymptotically no effect on the final fit.

In some situations, many candidate parametric models can be plausible. So a question that merits investigation is how to choose an appropriate model in the first step. As we stated earlier, the model does not need to be the best possible approximation; however, using a completely wrong fit will provide no advantage over the fully nonparametric approach and may even harm the estimation procedure. If one has no prior idea about the structure under investigation, then a preanalysis of the data can help provide some information. A realistic and simple model can be developed using any model selection procedure, including the Akaike information criterion (AIC) (Akaike 1973), the Bayes information criterion (Schwarz 1978), or the deviance information criterion (Spiegelhalter et al. 2002). In Section 5 we show that this approach works very nicely.

### 3. ASYMPTOTIC THEORY

In this section we present some important properties of the proposed estimator, such as consistency and large-sample distribution. As might be expected, the estimator’s performance depends on  $q_\pi(x, \theta^*)$ , the initial “best feasible” parametric model presented in the previous section. Here we introduce some notation, list all of the necessary assumptions, and state and discuss the theoretical results.

The process  $(X_t, Y_t), t = 0, \pm 1, \dots, \pm \infty$ , has the same distribution as  $(X, Y)$  and is stationary  $\alpha$ -mixing. By this, we mean

that if  $\mathcal{F}_I^L (-\infty \leq I, L \leq \infty)$  denotes the  $\sigma$ -field generated by the family  $\{(X_t, Y_t), I \leq t \leq L\}$ , then the mixing coefficient

$$\alpha(t) = \sup_{A \in \mathcal{F}_{-\infty}^0, B \in \mathcal{F}_t^\infty} |P(A \cap B) - P(A)P(B)|$$

converges to 0 as  $t \rightarrow \infty$ . This dependency structure, also known as strong mixing, includes independent and  $m$ -dependent random sequences. Moreover, under some weak conditions, the classical linear and nonlinear ARMA and (G)ARCH time series are strongly mixing (see, e.g., Fan and Yao 2003 and Carrasco and Chen 2002 for further details). Let  $x$  be a fixed point in the interior of the support of  $X$ . Denote by  $f_0(x), f_x(y)$ , and  $f(x, y) = f_0(x)f_x(y)$  the marginal density of  $X$ , the conditional density of  $Y|X = x$  and the joint density of  $(X, Y)$ , respectively, and assume that  $f(x, Q_\pi(x)) > 0$ . Let  $u_j = \int u^j K(u) du, v_j = \int u^j K^2(u) du, \tilde{\mathbf{u}} = (u_{p+1}, \dots, u_{2p+1})^T$ , and

$$\mathbf{\Lambda} = \begin{pmatrix} u_0 & u_1 & \cdots & u_p \\ u_1 & u_2 & \cdots & u_{p+1} \\ \vdots & \vdots & \ddots & \vdots \\ u_p & u_{p+1} & \cdots & u_{2p} \end{pmatrix}$$

and

$$\mathbf{\Omega} = \begin{pmatrix} v_0 & v_1 & \cdots & v_p \\ v_1 & v_2 & \cdots & v_{p+1} \\ \vdots & \vdots & \ddots & \vdots \\ v_p & v_{p+1} & \cdots & v_{2p} \end{pmatrix}.$$

In what follows, the bandwidth  $h_n$  converges to 0 and  $nh_n \rightarrow \infty$ . We require the following assumptions:

*Assumptions (A).*

(A1.a)  $\hat{\theta} - \theta^* = O_p(\delta_n)$ , with  $\delta_n \rightarrow 0$ .

(A1.b)  $u \rightarrow q_\pi(u, \theta)$  have a  $(p + 1)$ th continuous derivative at the point  $u = x$ .

(A1.c) There exists a neighborhood  $J$  of  $x$  such that  $u \rightarrow q_\pi^{(p+1)}(u, \theta)$  is Lipschitz on  $J$ .

(A1.d)  $\theta \rightarrow q_\pi(x, \theta)$  is Lipschitz on  $\Theta$ .

(A2.a)  $\alpha(t) = O(t^\iota)$  for some  $\iota > 2$ .

(A2.b) There exists a neighborhood  $J$  of  $x$  such that  $\sup_{j \geq j_*} \sup_{u, v \in J} f_j(u, v) \leq M_*$ , for some  $j_* \geq 1$  and  $0 < M_* < \infty$ , where  $f_j(u, v), j = 1, 2, \dots$ , denotes the density of  $(X_1, X_{j+1})$ .

(A3.a)  $Q_\pi(u)$  have a  $(p + 1)$ th continuous derivative at the point  $u = x$ .

(A3.b)  $f_0(u)$  and  $f_u(t)$  are continuous at  $x$  and  $(u, t) = (x, Q_\pi(x))$ , respectively.

(A3.c) There exists a neighborhood  $J$  of  $x$  such that  $f'_0$  exists and is Lipschitz on  $J$ .

(A4)  $K$  is a symmetric bounded density that has a bounded support, say  $[-1, 1]$ .

The requirement (A1.a) can be relaxed to the weaker assumption that  $\delta_n = O(1)$ ; however, in this case the bias term resulting from the parametric first step estimator may dominate the global bias term, as can be seen from the formula (4). In the case of the parametric estimation procedure described in the previous section, the conditions under which (A1.a) is fulfilled have been described by Komunjer (2005). All other assumptions are used mainly in the context of dependent nonparametric kernel regression.

The following theorem gives a Bahadur-type representation that facilitates the asymptotic analysis. This result is particularly interesting in our case because, unlike in the mean regression method, there is no explicit mathematical formula for the estimators proposed in the previous section.

*Theorem 1.* Under Assumptions (A), if  $h_n = O(n^{-1/(2p+3)})$ , then

$$\begin{aligned} \mathbf{H}_n(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) - \frac{h_n^{p+1}}{(p+1)!} [Q_\pi^{(p+1)}(x) - q_\pi^{(p+1)}(x, \boldsymbol{\theta}^*)] \boldsymbol{\Lambda}^{-1} \tilde{\mathbf{u}} \\ = \frac{a_n^2}{f(x, Q_\pi(x))} \boldsymbol{\Lambda}^{-1} \sum_{i=1}^n e_i \tilde{\mathbf{X}}_{h,i} K_{h,i} + \mathbf{r}_n, \end{aligned}$$

where  $e_i = \pi - I(Y_i < Q_\pi(X_i))$ ,  $\mathbf{H}_n = \text{diag}(1, h_n, \dots, h_n^p)$ ,  $\tilde{\mathbf{X}}_{h,i} = \mathbf{H}_n^{-1} \tilde{\mathbf{X}}_i$ , and  $\mathbf{r}_n = o_p(a_n) + h_n^{p+1}(O_p(\delta_n) + o_p(1))$ , with  $a_n^{-1} = \sqrt{nh_n}$ .

From this theorem, we now obtain the following results, which state the joint asymptotic normality for the estimators  $\hat{\beta}_j \equiv \hat{Q}_\pi^{(j)}(x)/j!$  of  $Q_\pi^{(j)}(x)/j!$  for  $j = 0, \dots, p$ .

*Theorem 2.* Under the assumptions of Theorem 1, if  $j_* = 1$  [see assumption (A2.b)], then

$$\begin{aligned} \sqrt{nh_n} \left\{ \mathbf{H}_n(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) - \frac{h_n^{p+1}}{(p+1)!} [Q_\pi^{(p+1)}(x) - q_\pi^{(p+1)}(x, \boldsymbol{\theta}^*)] \boldsymbol{\Lambda}^{-1} \tilde{\mathbf{u}} + h_n^{p+1}(O_p(\delta_n) + o_p(1)) \right\} \\ \xrightarrow{\mathcal{L}} \mathcal{N}_{p+1}(\mathbf{0}, \sigma_\pi^2(x) \boldsymbol{\Sigma}), \end{aligned}$$

where  $\sigma_\pi^2(x) = \frac{\pi(1-\pi)}{f_x^2(Q_\pi(x))f_0(x)}$  and  $\boldsymbol{\Sigma} = \boldsymbol{\Lambda}^{-1} \boldsymbol{\Omega} \boldsymbol{\Lambda}^{-1}$ .

Higher-order terms in the asymptotic expansions of bias and variance can be obtained using a Bahadur representation of  $\hat{\boldsymbol{\theta}}$  as given by, for example, He and Shao (1996). Let  $\mu_j$  and  $v_j$  be the  $(j+1)$ th element of the vector  $\boldsymbol{\Lambda}^{-1} \tilde{\mathbf{u}}$  and the  $(j+1)$ th diagonal element of the matrix  $\boldsymbol{\Sigma}$ , for  $j = 0, \dots, p$ . As a corollary we get the individual asymptotic normality for each  $\hat{Q}_\pi^{(j)}(x)$ .

*Corollary 1.* Under the assumptions of Theorem 2,

$$\begin{aligned} \sqrt{nh_n^{2j+1}} \left\{ [\hat{Q}_\pi^{(j)}(x) - Q_\pi^{(j)}(x)] - \frac{h_n^{p+1-j}}{(p+1)!} \mu_j j! [Q_\pi^{(p+1)}(x) - q_\pi^{(p+1)}(x, \boldsymbol{\theta}^*)] + h_n^{p+1-j} O_p(\delta_n) \right\} \\ \xrightarrow{\mathcal{L}} \mathcal{N}(0, \sigma_\pi^2(x) v_j (j!)^2), \end{aligned}$$

for  $j = 0, \dots, p$ .

The asymptotic variance of  $\hat{Q}_\pi(x)$ , given by  $\text{Var}(x) = \frac{v_0}{nh_n} \times \sigma_\pi^2(x)$ , is free from the pilot parametric function and has exactly the same expression as the classical fully nonparametric LP estimator. The essential difference between our approach and the

standard approach appears in the asymptotic bias term, which is given by

$$\begin{aligned} \text{Bias}(x) = \frac{\mu_0}{(p+1)!} [Q_\pi^{(p+1)}(x) - q_\pi^{(p+1)}(x, \boldsymbol{\theta}^*)] h_n^{p+1} + h_n^{p+1} O(\delta_n). \end{aligned} \quad (4)$$

The extra term  $O(\delta_n)$  appearing in the foregoing bias formula is the asymptotic error related to the estimation of  $\boldsymbol{\theta}$ . It becomes negligible whenever  $\hat{\boldsymbol{\theta}}$  converges to the pseudo-true value  $\boldsymbol{\theta}^*$  which is actually our assumption (A1.a). Typically  $\delta_n$  equals  $n^{-1/2}$  under some weak assumptions (see Komunjer 2005). For the fully nonparametric kernel smoother ( $q_\pi \equiv 0, \delta_n \equiv 0$ ), the bias is a constant multiple of  $Q_\pi^{(p+1)}(x)$ , and in our case it is essentially the same constant multiple of  $\Delta(x, \boldsymbol{\theta}^*) = Q_\pi^{(p+1)}(x) - q_\pi^{(p+1)}(x, \boldsymbol{\theta}^*)$ . This clearly indicates how the closeness of  $q_\pi$  to  $Q_\pi$  affects the behavior of the final fit. If  $\Delta(x, \boldsymbol{\theta}) = 0$  [i.e.,  $Q_\pi(x) - q_\pi(x, \boldsymbol{\theta})$  is a polynomial of degree equal to or less than  $p$ ], then the bias vanishes. In other words, our  $p$ th-order LP estimator is unbiased whenever the  $(p+1)$ th derivative of  $Q_\pi(u) - q_\pi(u, \boldsymbol{\theta})$  at  $u = x$  equals 0, whereas the standard  $p$ th-order LP smoother is unbiased if and only if  $Q_\pi^{(p+1)}(x) = 0$ . In the unbiased case, an arbitrarily large bandwidth can be used to obtain an estimator with the minimum possible variance. More generally, in the regions of the data where the parametric start is close to  $Q_\pi$  in the sense that  $|\Delta(x, \hat{\boldsymbol{\theta}})| < |Q_\pi^{(p+1)}(x)|$ , our estimator is better than the standard estimator.

For a given  $x$ , assume that  $|\Delta(x, \hat{\boldsymbol{\theta}})| \approx Cn^{-\epsilon}$  for some  $C$  and  $\epsilon > 0$ . Ignoring constants and higher-order terms, the MSE takes the form  $h_n^{2(p+1)} n^{-2\epsilon} + (nh_n)^{-1}$ , which attains its minimum value at  $h_n \sim n^{-(1-2\epsilon)/(2(p+1)+1)}$ . This implies that for the proposed estimator, the optimum value of the MSE is proportional to  $n^{-(2\epsilon+2(p+1))/(2(p+1)+1)}$  which, as  $\epsilon \rightarrow 1/2$ , converges to  $n^{-1}$ , the well-known convergence rate for a correctly specified parametric estimator. The classical LP smoother achieves this optimal rate of convergence if and only if  $Q_\pi^{(p+1)}(x) = 0$ .

Regarding the order of the Taylor expansion used in the estimation procedure, first note that the method allows us to get a consistent estimator for the derivatives of the quantile function, which may be used to, for example, estimate the bias or construct a confidence interval for  $Q_\pi(x)$ . Second, it is clear from (4) that increasing the value of  $p$  helps reduce the bias rate in a similar manner as with higher-order kernels. In practice, however, increasing the value of  $p$  comes with the numerical difficulties and poor accuracy related to high-order derivative estimation. The case where  $p = 1$  corresponds to the guided LL quantile (GLLQ) estimator, whose asymptotic MSE is given by

$$\text{MSE}(x) = u_2^2 h_n^4 [Q_\pi^{(2)}(x) - q_\pi^{(2)}(x, \boldsymbol{\theta}^*)]^2 / 4 + v_0 \sigma_\pi^2(x) / (nh_n). \quad (5)$$

The smoothing parameter that minimizes this expression is

$$h_n^5(x) = \frac{v_0}{u_2^2} \frac{\sigma_\pi^2(x)}{[Q_\pi^{(2)}(x) - q_\pi^{(2)}(x, \boldsymbol{\theta}^*)]^2} n^{-1}. \quad (6)$$

From this formula, it can be seen that if the parametric estimate is actually a poor approximation of the quantile function, then  $h_n$  will be very small. In this case, because  $|X_i - x| \leq h_n$ ,

$r_i(\hat{\theta}) = q_\pi(X_i, \hat{\theta}) - q_\pi(x, \hat{\theta}) - (X_i - x)q_\pi^{(1)}(x, \hat{\theta}) \approx 0$ , and the final estimate in (3) should almost coincide with the standard LL estimator. On the other hand, as  $q_\pi(x, \hat{\theta}) \rightarrow Q_\pi(x)$ ,  $h_n$  becomes larger, and the local correction induced by the kernel smoother becomes increasingly negligible.

*Remark 2.* For simplicity, we have considered only the univariate case; however, the extension of the new method to the multivariate case is particularly interesting. Assume that  $\mathbf{X} \in \mathbb{R}^s$  and let  $Q_\pi(\mathbf{x})$  be the multivariate conditional quantile function of  $Y$  given that  $\mathbf{X} = \mathbf{x}$ . Denote by  $\mathbf{Q}'_\pi(\mathbf{x})$  and  $\mathbf{Q}''_\pi(\mathbf{x})$  the gradient vector and the Hessian matrix of  $Q_\pi(\mathbf{x})$ , respectively. Let  $q_\pi(\mathbf{x}, \hat{\theta})$  be a (multivariate) parametric approximation of  $Q_\pi(\mathbf{x})$  and denote by  $\mathbf{q}'_\pi(\mathbf{x})$  and  $\mathbf{q}''_\pi(\mathbf{x})$  its gradient vector and Hessian matrix, respectively. Put  $r_i(\hat{\theta}) = q_\pi(\mathbf{X}_i, \hat{\theta}) - q_\pi(\mathbf{x}, \hat{\theta}) - (\mathbf{q}'_\pi(\mathbf{x}, \hat{\theta}))^T(\mathbf{X}_i - \mathbf{x})$ . A multivariate GLLQ estimator of  $(Q_\pi(\mathbf{x}), \mathbf{Q}'_\pi(\mathbf{x})) \in \mathbb{R} \times \mathbb{R}^s$  is  $(\hat{Q}_\pi(\mathbf{x}), \hat{\mathbf{Q}}'_\pi(\mathbf{x})) = \arg \min_{(b_0, \mathbf{b}_1) \in \mathbb{R} \times \mathbb{R}^s} \sum_{i=1}^n \varphi_\pi(Y_i - r_i(\hat{\theta}) - b_0 - \mathbf{b}_1^T(\mathbf{X}_i - \mathbf{x}))\mathcal{K}_{h,i}$ , with  $\mathcal{K}(\mathbf{x}) = \prod_{j=1}^s K(x_j)$ . From a technical standpoint, the asymptotic theory in this case closely parallels the theory for the univariate case—namely, under appropriate smoothness conditions, if the bandwidth  $h_n$  satisfies  $h_n \rightarrow 0$ ,  $nh_n^s \rightarrow \infty$  and  $nh_n^{s+4} = O(1)$ , then  $(\hat{Q}_\pi(\mathbf{x}) - Q_\pi(\mathbf{x}))$  is asymptotically Gaussian with asymptotic mean and variance given by  $\frac{h_n^2}{2} \text{tr}(\mathbf{Q}''_\pi(\mathbf{x}) - \mathbf{q}''_\pi(\mathbf{x}, \theta^*))u_2 + h_n^2 O(\delta_n)$  and  $\frac{v_0}{nh_n^s} \sigma_\pi^2(\mathbf{x})$ . But from a practical standpoint, the MSE increases rapidly with the dimension  $s$ , leading to the so-called “curse of dimensionality.” As for the univariate case, a bias reduction occurs whenever  $|\text{tr}(\mathbf{Q}''_\pi(\mathbf{x}) - \mathbf{q}''_\pi(\mathbf{x}, \hat{\theta}))| \leq |\text{tr}(\mathbf{Q}''_\pi(\mathbf{x}))|$ , which suggests that the guided smoother may be of greater benefit for high-dimensional problems.

#### 4. BANDWIDTH SELECTION

The theoretical results given in the previous section are very encouraging; however, in practice, applying the proposed method requires choosing the bandwidth parameter. As for other smoothing techniques, the performance of our guided smoother may be destroyed by a really bad choice of bandwidth. An extensive literature addresses this problematic subject, especially in the context of nonparametric mean regression. The classical techniques used for mean kernel smoothing, such as cross-validation, plug-in, rule-of-thumbs, and bootstrap, also can be used (after adaptation) to select the bandwidth for quantile regression. (For more details, see Yu and Jones 1998; Zheng and Yang 1998; Leung 2005, and references therein.) Here we adopt the plug-in principle and thus use the simple expression of the optimal asymptotic bandwidth given by (6). Nevertheless, this theoretical formula involves the following unknown quantities: the quantile function,  $Q_\pi(x)$ ; the design density,  $f_0(x)$ ; the conditional density,  $f_x(y)$ ; the second derivative of the regression function,  $Q_\pi^{(2)}(x)$ ; and, for the guided LL smoother, the second derivative of the pseudo-true parametric quantile function. The latter can be easily estimated by  $q_\pi^{(2)}(x, \hat{\theta})$ , the second derivative of  $q_\pi(x, \hat{\theta})$  with respect to  $x$ , where  $\hat{\theta} = \arg \min_\theta \sum_{i=1}^n \varphi_\pi(Y_i - q_\pi(X_i, \theta))$  and  $q_\pi(x, \theta)$  is the parametric model start. For the other unknown quantities, two approaches are mainly used in the literature: make reference to a normal case or use some nonparametric pilot estimate.

The first of these approaches is very simple but may lead to a bad selection of the bandwidth, and thus to a poor estimator, when the data do not match the normal assumption. The second approach is more elaborate but entails the difficulty that each pilot estimate has at least one smoothing parameter that must be selected. Cross-validation or the bootstrap can be used to avoid the drawbacks of the plug-in method, but this will substantially increase the computation time. For these reasons, we propose a new approach that combines the parametric and the nonparametric techniques. The idea is very simple: Instead of using a nonparametric pilot estimator, we use a fully parametric approach to approximate all of the unknown quantities except  $Q_\pi^{(2)}(x)$ . For the latter, a local cubic smoother,  $\hat{Q}_\pi^{(2)}(x)$ , as given by eq. (1), with  $p = 3$ , is used at the pilot stage. To calculate the pilot bandwidth, we simply use the routine *gkerns* from the R package *lokern* (Herrmann and Mächler 2003). An appropriate estimate for the distribution functions  $f_0$  and  $f_x$  can be chosen using the AIC and the R package *gamlss* (Stasinopoulos, Rigby, and Akantziliotou 2008). The latter contains a large collection of distribution families that might be fitted to the data. Instead of a pilot bandwidth, the analyst need only choose a set of candidate (likely incorrect) models to use. In the sequel (see the simulation study in Sec. 5), we considered the normal distribution, the power exponential distribution, and the skew- $t$  distribution (see, e.g., Azzalini and Genton 2008). The resulting estimators  $\hat{f}_0(x)$  and  $\hat{f}_x(\tilde{q}_\pi(x, \hat{\theta}))$ , with  $\tilde{q}_\pi(x, \theta)$  either the parametric start or any other data-driven parametric model (elected via, e.g., the AIC), together with  $q_\pi^{(2)}(x, \hat{\theta})$  and  $\hat{Q}_\pi^{(2)}(x)$ , are then plugged into (6) to get  $\hat{h}_n$ , our data-driven bandwidth. Next we demonstrate the efficiency of this automatic procedure.

*Remark 3.* Corollary 1 may be used to build pointwise confidence intervals for the quantile function. To avoid an explicit correction of the bias associated with  $\hat{Q}_\pi(x)$ , the optimal bandwidth given by (6) should not be used. Instead, it is better to undersmooth the estimator (i.e., choose a smaller bandwidth  $h_n$  that satisfies  $nh_n^5 \rightarrow 0$ ), so that the bias asymptotically vanishes. But decreasing the bandwidth will inevitably increase the length of the resulting confidence intervals, especially those based on the classical LL estimator. For our guided smoother, the decrease in the bias due to the parametric start allows the use of a larger bandwidth, which leads to narrower confidence intervals. The magnitude of this reduction in length depends on how well the parametric start approximates locally the regression curve. For a given  $\alpha \in (0, 1)$ , let  $z_\alpha$  be the upper  $\alpha/2$ -quantile of the standard normal distribution. As a Wald-type confidence interval for  $Q_\pi(x)$  based on the GLLQ estimator, we suggest  $\hat{Q}_\pi(x) \pm z_\alpha \sqrt{\frac{v_0}{nh_n} \hat{\sigma}_\pi^2(x)}$ , where  $\hat{\sigma}_\pi^2(x) = \frac{\pi(1-\pi)}{f_x^2(\tilde{q}_\pi(x, \hat{\theta}))f_0(x)}$ .

#### 5. SIMULATION STUDY

In this section we investigate the finite-sample performance of the proposed method using a Monte Carlo simulation study. Because there is no explicit formula for the estimator, we first need an efficient optimization routine to solve the mathematical minimization problem imposed by the definition of the quantile estimator. Using the fact that the kernel function  $K$  is non-negative, we write (3) as  $\sum_{i=1}^n \varphi_\pi(\tilde{Y}_{K,i} - \tilde{\mathbf{X}}_{K,i}^T \mathbf{b})$ , with  $\tilde{Y}_{K,i} = (Y_i - r_i(\hat{\theta}))\mathcal{K}_{h,i}$  and  $\tilde{\mathbf{X}}_{K,i} = \tilde{\mathbf{X}}_i \mathcal{K}_{h,i}$ . This linear parameterization

allows the use of any minimization algorithm available in the literature for parametric quantile regression, including the interior point algorithm of Koenker and Park (1996), the smoothing algorithm of Chen (2007), and the majorize–minimize algorithm of Hunter and Lange (2000). Here we choose the latter for its simplicity and numerical stability.

The data are generated according to the equation  $Y = m_\lambda(X) + \epsilon$ , where  $X$  is a uniform variable on  $[-1.1, 2.1]$ ,  $\epsilon$  is  $\mathcal{N}(0, 1)$ , and

$$m_\lambda(x) = 10 - 6x^2 + 2.8x^3 + \lambda r(x), \tag{7}$$

with  $r(x)$  either  $\exp(-4(x-1)^2)$  (model M1) or  $\sin((\pi/2.25) \times (x-1)^2)$  (model M2).

Our objective is to compare the GLLQ estimator with both the fully parametric and the fully nonparametric competitors. Toward this end, we consider five conditional quantile estimating methods:

- LLQ: The standard LLQ estimator
- PQ<sub>1</sub>: A parametric estimator  $q(X, \hat{\theta})$ , where

$$\hat{\theta} = \arg \min_{\theta} \varphi_{\pi}(Y_i - q(X_i, \theta)),$$

with  $q(X, \theta)$  an (unknown) third-order polynomial

- GLLQ<sub>1</sub>: The LLQ estimator guided by  $q(X, \hat{\theta})$
- PQ<sub>2</sub>: Similar to PQ<sub>1</sub>, but with the parametric model a polynomial with order  $p \in [1, 20]$  selected by the data using the AIC
- GLLQ<sub>2</sub>: The LLQ estimator guided by the data-driven polynomial model used in PQ<sub>2</sub>.

The parameter  $\lambda$  in (7) can be seen as a misspecification parameter that controls the deviation of the parametric guide from the true data-generating equations. The case where  $\lambda = 0$  [i.e.,  $m_\lambda(x) = 10 - 6x^2 + 2.8x^3$ ], is the ideal situation not only for the fully parametric method PQ<sub>1</sub>, but also for the guided smoothers GLLQ<sub>1</sub> and GLLQ<sub>2</sub>, both of which are based on a polynomial start. Whenever  $\lambda \neq 0$ , the polynomial guide used in GLLQ<sub>1</sub> and GLLQ<sub>2</sub> is incorrect. In addition, as  $\lambda$  increases, the data structure becomes more complicated, and approximating the true curve by a polynomial becomes increasingly difficult. In such a case, the fully nonparametric kernel smoother LLQ should perform better. To allow comparisons of situations of correct parametric specification, approximately correct parametric specification and a wrong parametric model,  $\lambda$  is varied as 0, 2, 6, 10, 20. For each scenario,  $N = 1000$  samples of size  $n = 100$  are generated. A bandwidth parameter,  $h_n$ , and a kernel function,  $K(\cdot)$ , are needed for the three smoothing methods, LLQ, GLLQ<sub>1</sub>, and GLLQ<sub>2</sub>. As a bandwidth, we use our data-driven smoothing parameter  $\hat{h}_n$  as described in the previous section. We also set  $K(\cdot)$  as the Epanechnikov kernel function. For each set of data, we evaluate the different estimators at 61 equally spaced locations,  $x_i$ , taken from  $-1$  to  $2$ . At every data point  $x_i$ ,  $i = 1, \dots, 61$ , we approximate the bias by  $B_i = N^{-1} \sum_{k=1}^N (\hat{Q}_{\pi,k}(x_i) - Q_{\pi}(x_i))$  and the variance by  $V_i = N^{-1} \sum_{k=1}^N (\hat{Q}_{\pi,k}(x_i) - N^{-1} \sum_{k=1}^N \hat{Q}_{\pi,k}(x_i))^2$ , where  $\hat{Q}_{\pi,k}$  is the estimated conditional quantile for the  $k$ th replication and  $Q_{\pi}$  is the true quantile function.

### 5.1 Median Function With iid Data

Table 1 reports the averaged squared bias,  $\mathbb{B}ias^2 = 61^{-1} \times \sum_{i=1}^{61} B_i^2$ , the averaged variance,  $\mathbb{V}ar = 61^{-1} \sum_{i=1}^{61} V_i$ , and the average MSE,  $AMSE = \mathbb{B}ias^2 + \mathbb{V}ar$ , for the median function with data generated according to an iid process. Note that  $AMSE = N^{-1} \sum_{k=1}^N ASE_k$ , with  $ASE_k = 61^{-1} \sum_{i=1}^{61} (\hat{Q}_{\pi,k}(x_i) - Q_{\pi}(x_i))^2$ . As an assessment of uncertainty about the  $AMSE$ , Table 1 reports the Monte Carlo standard error of  $AMSE$ . The table also reports the values of  $AMSE^*$ , the  $AMSE$  obtained using the theoretical asymptotic formula of the bandwidth given in (6). We start by analyzing the case of a correctly specified parametric model ( $\lambda = 0$ ). As expected, this is the only example for which we get the best results by using the fully parametric method, PQ<sub>1</sub>. Interestingly, the  $AMSE$  performance of the two guided LL estimators (GLLQ<sub>1</sub> and GLLQ<sub>2</sub>) is quite similar to that of the parametric estimator  $q(x, \hat{\theta})$ , with the GLLQ<sub>1</sub> method having a slight advantage. With a loss of efficiency  $> 60\%$ , the purely nonparametric estimator LLQ is significantly worse than all of the other methods. The observed bias for the LLQ estimator is around 11 times that for GLLQ<sub>1</sub> and GLLQ<sub>2</sub>. Regarding the variance, the LLQ also behaves clearly worse than GLLQ<sub>1</sub> and GLLQ<sub>2</sub>. The superiority of the proposed method is illustrated in plots (a1) and (a2) in Figure 1, which displays boxplots of the estimated values using the LLQ and GLLQ<sub>2</sub> methods at different data points together with the true curve.

For a small to moderate misspecified parametric start, we can see that the GLLQ<sub>1</sub> and the GLLQ<sub>2</sub> methods are significantly superior to both the purely parametric and the purely nonparametric estimators. Because the GLLQ<sub>1</sub> estimator relies on  $q(x, \theta)$ , the given (fixed) parametric model, its  $AMSE$  performance is slightly better than the GLLQ<sub>2</sub> estimator when  $q(x, \theta)$  is “close” to the true underlying structure. This is due mainly to a relatively larger variance term for GLLQ<sub>2</sub>, which might be explained by the variations related to the data-driven polynomial guide. As the misspecification becomes stronger, the GLLQ<sub>2</sub> method shows considerable improvement and becomes the best approach. In addition, it reduces the bias considerably and performs uniformly better than LLQ (the LL smoother). Compared with PQ<sub>2</sub> (the data-driven parametric model), GLLQ<sub>2</sub> significantly reduces the  $AMSE$  (10% to 45% in general), especially when the data structure becomes complicated; see, for example, the model M2 with  $\lambda = 10$ . At any given value  $x$  of  $X$ , the fraction of data falling on or below  $\hat{Q}_{\pi}(x)$  ideally should be close to  $\pi$ . Thus another way to quantify the quality of the regression quantile estimator  $\hat{Q}_{\pi}(x)$  is by calculating the “coverage probability”  $P(Y \leq \hat{Q}_{\pi}(x) | X = x)$ , which in our case is given by  $\Phi(\hat{Q}_{\pi}(x) - m_\lambda(x))$ , where  $\Phi$  is the standard normal distribution function. Figure 2 shows the boxplots of  $\Phi(\hat{Q}_{1/2,k}(x) - m_\lambda(x))$ ,  $k = 1, \dots, N$ , for different values of  $x$  and  $\lambda$ . Again, it can be seen that GLLQ<sub>2</sub> is the best estimator under this performance criterion. The coverage probability and MSE behavior [see Figures 1(b) and 2] demonstrate that our method performs better at the middle of the  $X$ -domain than at the boundary region. This property, common to all kernel smoothing techniques, is due to the increased stochastic variability in the tails of the data. At the boundary region, the classical LLQ estimator may behave better than the guided one. This is typically the case when very

Table 1. Averaged squared bias  $\mathbb{B}ias^2$ , averaged variance  $\mathbb{V}ar$ ,  $AMSE$  (and its standard error) and  $AMSE^*$  for  $\hat{Q}_{0.5}$  with iid data, sample size  $n = 100$ , and  $N = 1000$  replicates

$\lambda$	Method	Model M1				Model M2			
		$10^2 \times$				$10^2 \times$			
		$\mathbb{B}ias^2$	$\mathbb{V}ar$	$AMSE (SE)$	$AMSE^*$	$\mathbb{B}ias^2$	$\mathbb{V}ar$	$AMSE (SE)$	$AMSE^*$
0	LLQ	1.786	14.16	15.94 (0.22)	11.91	1.786	14.16	15.94 (0.22)	11.91
	PQ <sub>1</sub>	0.003	5.62	<b>5.63</b> (0.10)	–	0.003	5.62	<b>5.63</b> (0.10)	–
	GLLQ <sub>1</sub>	0.015	9.26	9.27 (0.15)	5.82	0.015	9.26	9.27 (0.15)	5.82
	PQ <sub>2</sub>	0.003	5.67	5.68 (0.11)	–	0.003	5.67	5.68 (0.11)	–
	GLLQ <sub>2</sub>	0.013	9.36	9.38 (0.15)	6.06	0.013	9.36	9.38 (0.15)	6.06
2	LLQ	1.846	14.60	16.45 (0.21)	13.22	2.220	16.81	19.03 (0.24)	13.82
	PQ <sub>1</sub>	9.049	6.24	15.29 (0.12)	–	25.768	9.42	35.18 (0.24)	–
	GLLQ <sub>1</sub>	1.016	10.79	<b>11.81</b> (0.16)	9.29	1.296	12.27	13.56 (0.17)	<i>11.21</i>
	PQ <sub>2</sub>	2.444	11.56	14.00 (0.16)	–	4.461	15.85	20.31 (0.44)	–
	GLLQ <sub>2</sub>	0.510	12.07	12.58 (0.17)	10.73	0.435	12.72	<b>13.15</b> (0.18)	11.40
6	LLQ	2.568	16.19	18.76 (0.22)	16.91	2.859	22.08	24.94 (0.28)	19.83
	PQ <sub>1</sub>	80.994	12.96	93.95 (0.34)	–	215.581	48.20	263.78 (1.12)	–
	GLLQ <sub>1</sub>	2.502	14.71	17.21 (0.21)	<i>14.15</i>	2.371	18.99	21.36 (0.24)	18.71
	PQ <sub>2</sub>	6.299	24.15	30.45 (0.53)	–	2.899	21.20	24.10 (0.22)	–
	GLLQ <sub>2</sub>	0.826	15.71	<b>16.53</b> (0.22)	14.64	0.719	15.50	<b>16.22</b> (0.20)	<i>14.73</i>
10	LLQ	2.711	18.97	21.68 (0.24)	17.76	3.281	28.53	31.81 (0.33)	25.31
	PQ <sub>1</sub>	227.978	31.80	259.78 (0.93)	–	597.258	139.03	736.29 (3.46)	–
	GLLQ <sub>1</sub>	2.729	18.19	20.91 (0.22)	18.07	2.973	25.26	28.24 (0.31)	24.90
	PQ <sub>2</sub>	1.381	16.98	18.36 (0.25)	–	2.962	24.32	27.28 (0.31)	–
	GLLQ <sub>2</sub>	0.545	16.59	<b>17.13</b> (0.22)	<i>15.54</i>	0.713	17.68	<b>18.39</b> (0.22)	<i>16.53</i>
20	LLQ	3.119	26.26	29.38 (0.31)	23.89	4.254	40.64	44.90 (0.44)	40.78
	PQ <sub>1</sub>	951.273	138.10	1089.37 (4.69)	–	2410.145	568.69	2978.83 (13.07)	–
	GLLQ <sub>1</sub>	3.390	25.66	29.05 (0.31)	27.17	4.271	39.54	43.81 (0.45)	40.57
	PQ <sub>2</sub>	3.079	18.55	21.63 (0.25)	–	0.874	28.15	29.02 (0.31)	–
	GLLQ <sub>2</sub>	0.912	19.20	<b>20.12</b> (0.23)	<i>18.10</i>	0.278	21.55	<b>21.83</b> (0.26)	<i>18.47</i>
		$Y = 20 \exp(-4(X - 1)^2) + \epsilon$				$Y = 20 \sin((\pi/2.25)(X - 1)^2) + \epsilon$			
		$10^2 \times$				$10^2 \times$			
Method		$\mathbb{B}ias^2$	$\mathbb{V}ar$	$AMSE (SE)$	$AMSE^*$	$\mathbb{B}ias^2$	$\mathbb{V}ar$	$AMSE (SE)$	$AMSE^*$
LLQ		2.504	25.94	28.45 (0.30)	21.07	4.591	43.55	48.14 (0.51)	40.29
PQ <sub>1</sub>		944.280	132.69	1076.97 (4.23)	–	2374.550	563.56	2938.11 (13.04)	–
GLLQ <sub>1</sub>		3.169	25.11	28.28 (0.31)	25.41	4.610	41.88	46.49 (0.51)	42.00
PQ <sub>2</sub>		2.933	18.26	21.19 (0.24)	–	0.817	36.87	37.68 (0.30)	–
GLLQ <sub>2</sub>		0.952	19.10	<b>20.05</b> (0.24)	<i>18.56</i>	0.206	21.87	<b>22.08</b> (0.26)	<i>19.12</i>

NOTE: Values in bold (italics) are the minimum observed values of the  $AMSE$  ( $AMSE^*$ ).

few observations are available at the end of the sampled data. Better results can be obtained if for each individual datum, the candidate model can be visualized and, if necessary, adjusted globally or locally before being used as a pilot. Figure 1(c) shows how the optimal smoothing parameter changes with  $x$  for LLQ, GLLQ<sub>1</sub>, and GLLQ<sub>2</sub>. From this plot, it can be seen that both GLLQ<sub>1</sub> and GLLQ<sub>2</sub> attain their minimum MSE at a much larger value than the LLQ estimator. That GLLQ<sub>1</sub> and GLLQ<sub>2</sub> have a similar optimal bandwidth is not the rule, but rather an exception; in fact, as  $\lambda$  increases, the optimal bandwidth for GLLQ<sub>2</sub> decreases and approaches the optimal bandwidth for LLQ.

In terms of the usefulness of our automatic bandwidth selection procedure, we found that in general, the resulting values for the three smoothers (LLQ, GLLQ<sub>1</sub>, and GLLQ<sub>2</sub>) obtained using  $\hat{h}_n$  were quite close to the optimal ones obtained

using  $h_n$ , the asymptotic optimal theoretical bandwidth. That is why the difference in the AMSEs  $|AMSE - AMSE^*|$  remained very small. In terms of the bias, no significant efficiency loss resulted by using  $\hat{h}_n$  instead of  $h_n$ . In other words, our data-driven bandwidth affects (increases) mainly the variance, not the bias. The loss of efficiency from estimating the bandwidth is larger for the classical LL estimator. This indicates a greater robustness of GLLQ<sub>1</sub> and GLLQ<sub>2</sub> to bandwidth misspecification. This also explains the fact that when the parametric guide is actually a poor approximation of the true underlying model (see the cases  $\lambda = 10$  and  $\lambda = 20$  with model M1), the classical LLQ smoother is superior to GLLQ<sub>1</sub> in terms of  $AMSE^*$  but not in terms of  $AMSE$ . Comparing  $AMSE$  and  $AMSE^*$  also shows that the  $AMSE$  performance of GLLQ<sub>2</sub> using  $\hat{h}_n$  is uniformly better than that of LLQ using the optimal theoretical bandwidth. This

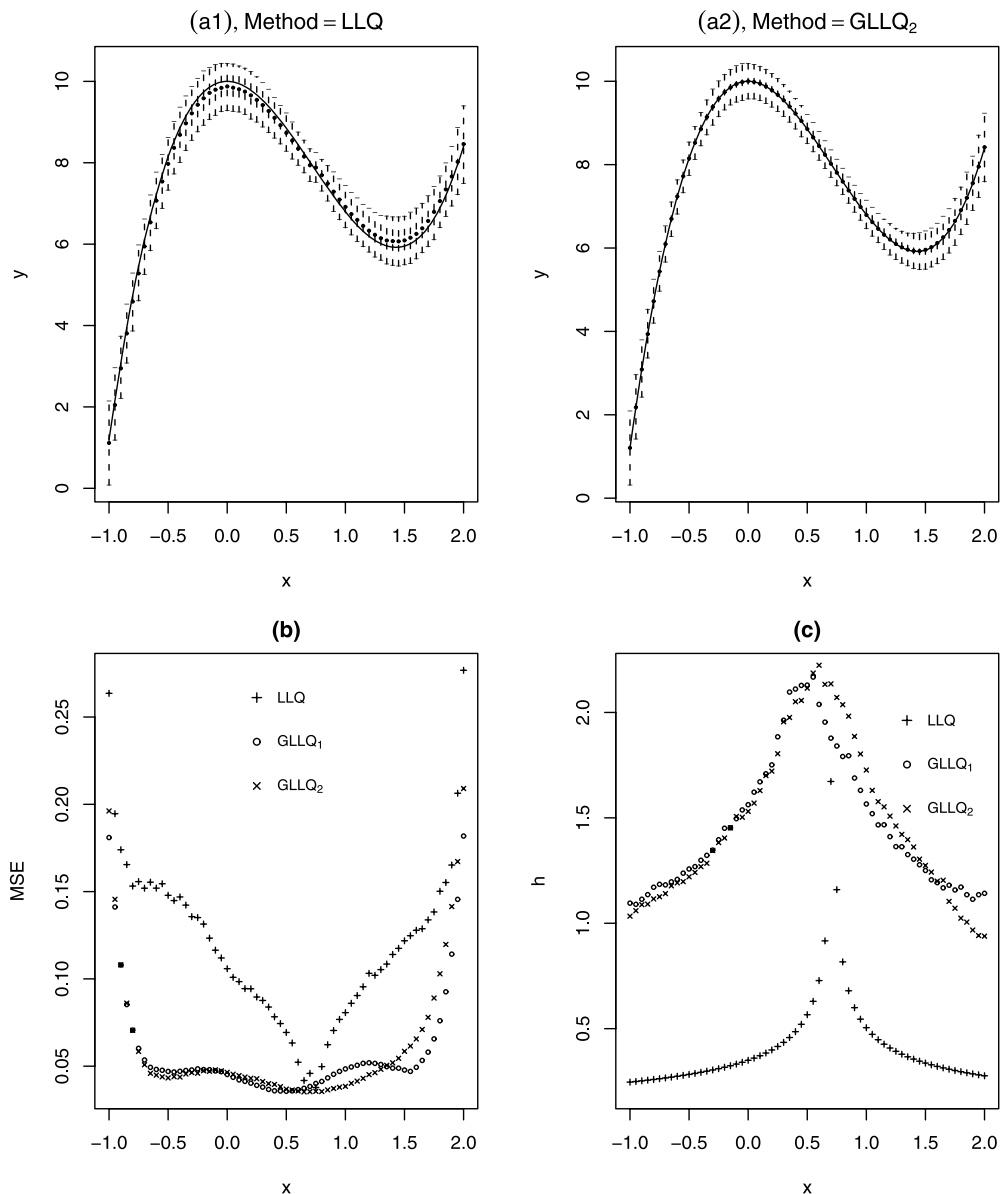


Figure 1. Boxplots of the estimated values using the asymptotically optimal bandwidth [(a1) and (a2)], with the true regression curve given by the solid curve; the *MSE* as a function of  $x$  (b); and the mean optimal values of  $h_n$  at different values of  $x$  (c).  $\lambda = 0, n = 100$ .

definitely demonstrates the advantages of including information from a parametric model into nonparametric estimates.

Finally, to check the performance of the proposed method under a completely wrong parametric start, we ran the same study using as a data-generating equation the model  $Y = 20r(X) + \epsilon$ , with  $r(x)$  either  $\exp(-4(x - 1)^2)$  or  $\sin((\pi/2.25)(x - 1)^2)$ . The factor 20 in this equation guarantees the same scale as in (7) and thus facilitates a comparison with the other experiments. As can be seen at the end of Table 1, the results are very close to those obtained using eq. (7) with  $\lambda = 20$ . This clearly demonstrates our proposed method’s superiority and its very high robustness against a misspecified parametric guide.

### 5.2 Other Scenarios

In this section we briefly discuss the performance of our method under situations not considered earlier. For the sake of brevity, we report only the results for model M2 with some se-

lected (representative) values of  $\lambda$ . All of the results are summarized in Table 2.

We ran the entire simulation study as described in the previous section with  $n = 50$  and  $n = 200$  and compared the results with the corresponding findings listed in Table 1. In general, the findings for the reference case ( $n = 100$ ) remained the same for both small ( $n = 50$ ) and large ( $n = 200$ ) sample sizes.

In addition to the median case, we also investigated two other quantiles:  $\pi = 0.05$  and  $\pi = 0.95$ . Table 2 shows that GLLQ<sub>1</sub> and GLLQ<sub>2</sub> are superior to the classical LLQ smoother, with a clear advantage to the GLLQ<sub>2</sub> estimator. Globally, these new results match those obtained for the median functions, but with two main differences. First,  $AMSE^*$  often exceeds  $AMSE$ , indicating that using the data-driven bandwidth is sometimes more beneficial than using the asymptotic theoretical optimal bandwidth. Second, the practical performance of all tested estima-

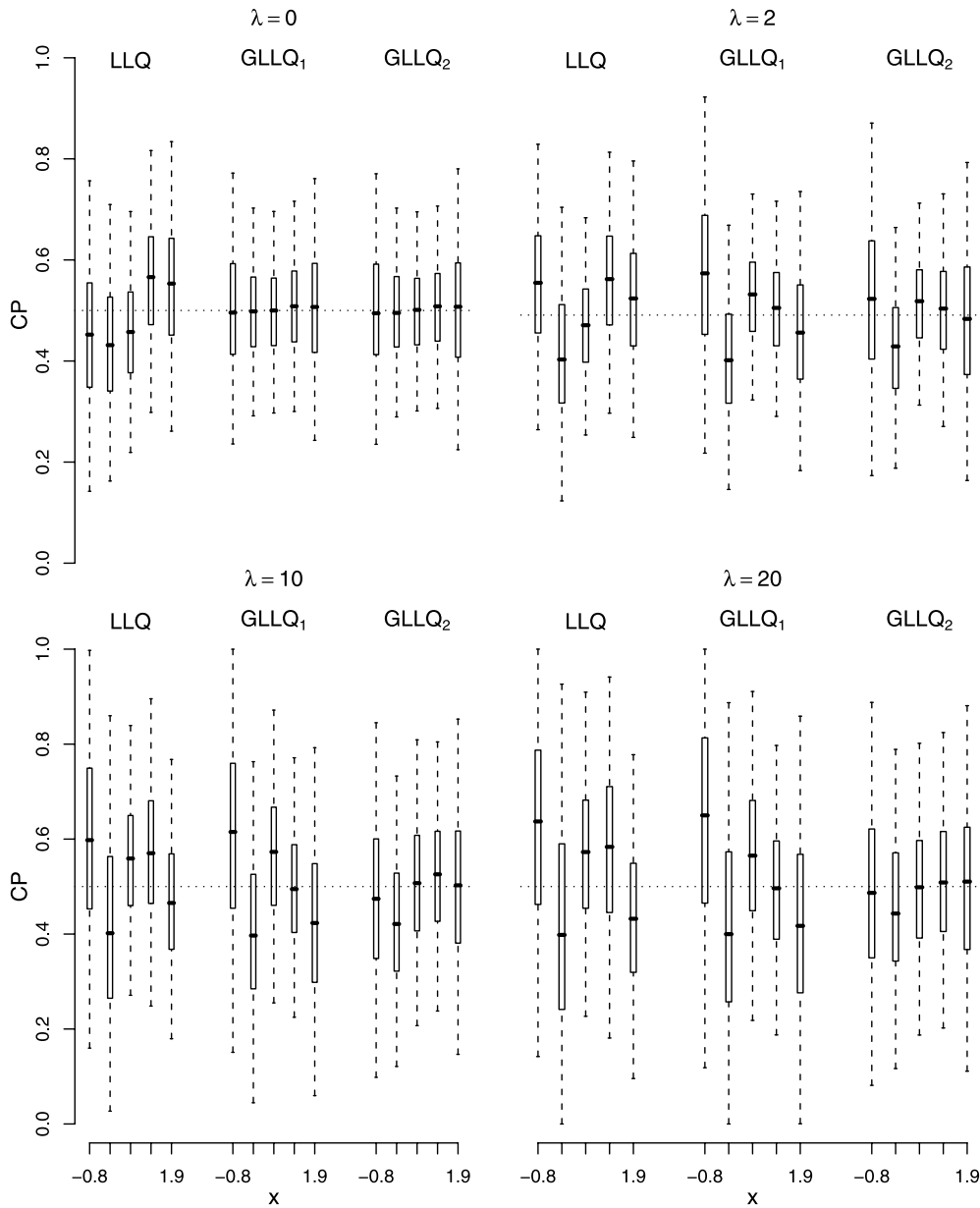


Figure 2. Coverage probability for the conditional median obtained under model M2 for different values of  $x$  and  $\lambda$ .

tors becomes worse as  $\pi$  approaches 0 or 1. This is a known feature in quantile regression.

To check the behavior of the proposed estimator with dependent data, we tested the model M2 under the following scenarios involving autoregressive  $[AR(\varrho)]$  processes of order 1 and autocorrelation parameter  $\varrho$ : (a)  $X$  is drawn from  $AR(0.5)$  and  $\epsilon$  is iid; (b)  $X$  is drawn from  $AR(0.8)$  and  $\epsilon$  is iid; and (c)  $X$  is drawn from  $AR(0.8)$  and  $\epsilon$  drawn from  $AR(0.5)$ . The results for the first case were similar to those obtained with iid data (see Table 1), and so we do not give them here. For the second and third cases, some results are given in Table 2. Again, we obtained somewhat better results with our procedure. As expected, the dependence deteriorated the practical performance of all the estimators. A larger sample size is needed to achieve better performance. Finally, we also noticed that in general, the loss of efficiency due to the bandwidth selection increased as the dependency in the data increased. A sophis-

ticated bandwidth selection procedure that takes into account the dependence structure of the data, such as that proposed by Francisco-Fernandez, Opsomer, and Vilar-Fernandez (2004), may lead to better results. (See also Francisco-Fernandez and Vilar-Fernandez 2005 for more about bandwidth selection under correlated data.)

### 6. DATA ANALYSIS

To illustrate the method on a real example, we now analyze the so-called motorcycle data set given by Härdle (1990). The covariate  $X$  is the time (in milliseconds) after a simulated impact with a motorcycle, and the response variable  $Y$  is the head acceleration due to gravity (in g) of a postmortem human test object. A more detailed description of this data set was provided by Schmidt, Mattern, and Schüler (1981). The sample size is  $n = 133$ . The observations are correlated, and they are all sub-

Table 2. *AMSE*, averaged variance  $\text{Var}$ , and *AMSE\** for  $\hat{Q}_\pi$  under different scenarios with  $N = 1000$  replicates

$\lambda$	$10^2 \times$	LLQ	GLLQ <sub>1</sub>	GLLQ <sub>2</sub>	LLQ	GLLQ <sub>1</sub>	GLLQ <sub>2</sub>
iid case with $\pi = 0.5$							
		$n = 50$			$n = 200$		
0	<i>AMSE</i>	31.55	<b>21.26</b>	21.36	8.09	4.20	<b>4.19</b>
	$\text{Var}$	29.61	21.23	21.33	7.05	4.20	4.19
	<i>AMSE*</i>	22.24	<i>11.80</i>	12.53	6.39	2.86	2.91
2	<i>AMSE</i>	38.45	28.17	<b>27.56</b>	8.95	6.85	<b>6.16</b>
	$\text{Var}$	35.42	26.17	26.50	7.76	6.05	5.94
	<i>AMSE*</i>	27.23	<i>21.40</i>	22.54	7.27	6.05	5.79
10	<i>AMSE</i>	77.97	73.64	<b>48.31</b>	13.73	13.48	<b>8.95</b>
	$\text{Var}$	69.94	64.15	47.20	12.20	11.95	8.49
	<i>AMSE*</i>	62.11	66.31	<i>41.86</i>	12.20	12.22	8.23
20	<i>AMSE</i>	124.15	134.34	<b>63.93</b>	18.83	18.57	<b>9.93</b>
	$\text{Var}$	104.67	109.71	61.51	16.92	16.63	9.67
	<i>AMSE*</i>	124.56	144.09	<i>58.15</i>	17.22	17.32	8.51
iid case with $n = 100$							
		$\pi = 0.05$			$\pi = 0.95$		
0	<i>AMSE</i>	50.40	19.70	<b>19.25</b>	35.57	19.68	<b>19.17</b>
	$\text{Var}$	42.04	18.78	18.33	28.45	18.36	17.85
	<i>AMSE*</i>	34.17	18.18	<i>17.92</i>	37.28	18.22	<i>17.71</i>
10	<i>AMSE</i>	136.4	74.6	<b>39.16</b>	89.75	72.11	<b>40.21</b>
	$\text{Var}$	103.91	48.79	36.02	67.68	51.33	35.04
	<i>AMSE*</i>	84.72	86.43	<i>49.05</i>	76.57	78.57	<i>50.76</i>
Dependent case with $n = 100, \pi = 0.5$							
		$X \sim AR(0.8); \epsilon \sim \text{iid}$			$X \sim AR(0.8); \epsilon \sim AR(0.5)$		
0	<i>AMSE</i>	17.67	<b>10.65</b>	10.66	21.29	<b>14.51</b>	14.54
	$\text{Var}$	16.20	10.64	10.65	19.89	14.49	14.52
	<i>AMSE*</i>	13.03	<i>6.71</i>	7.02	16.89	<i>10.79</i>	11.29
10	<i>AMSE</i>	35.63	33.07	<b>23.10</b>	41.14	37.12	<b>25.77</b>
	$\text{Var}$	32.48	29.64	22.66	37.93	33.50	25.36
	<i>AMSE*</i>	27.81	30.26	<i>19.44</i>	31.35	33.85	<i>23.19</i>

NOTE: Values in bold (italics) are the minimum observed values of the *AMSE* (*AMSE\**).

ject to error. In addition, the variance of the data is not constant (see Figure 3). Although the homoscedasticity assumption is not needed, we prefer to stabilize the variance function. To do so, we first added 200 to all  $Y_i$ 's to get positive values, and then applied the log transformation. Thus in our numerical analysis, we used  $V = \log(Y + 200)$  instead of  $Y$  as the response variable. After estimating  $\hat{Q}_\pi^V(x)$ , the conditional quantile of  $V|X = x$ , we used the equivariant properties of regression quantiles to return to the variable of interest,  $Y$ ; that is, our final estimator of  $Q_\pi^Y(x)$ , the quantile regression function of  $Y$ , is given by  $\exp(\hat{Q}_\pi^V(x)) - 200$ .

Figure 3(a) displays the guided LL estimator of the median curve using a piecewise polynomial function with break points  $x_1 = 15.4$  and  $x_2 = 42.8$  as a parametric start. The order of each polynomial (a total of three) was selected via the AIC. The figure also shows the classical LL estimator of the median function. Although the two methods capture the functional dependency between  $X$  and  $Y$  very well, obviously GLLQ considerably reduces the variation in the estimate, especially at higher time points, where the data become more sparse. Fig-

ure 3(b) uses the same guide as in Figure 3(a) but also provides a 95% confidence interval for the true regression curve calculated as described in Remark 3. The main appealing aspect here is that the confidence interval becomes larger and larger as we approach the boundary region, because of the inflation in the variance estimate. Finally, Figure 3(c) plots the estimator of the median curve together with two other quantiles,  $\pi = 0.25$  and  $\pi = 0.75$ .

To compare LLQ and GLLQ in terms of ‘‘coverage probability,’’ because the true conditional function is unknown, we began by randomly splitting the data into two subsets: 80% of observations into training data (to estimate the conditional median curve) and 20% of observations into evaluation data [to empirically evaluate  $P(Y \leq \hat{Q}_{0.5}(X))$ ]. We then repeated this procedure 1000 times and computed the average performance (and its standard error). For the classical LL smoother, we obtained 48.1% (0.34%), and using our new method, we obtained 49.9% (0.36%). Although these values are close to each other, they still indicate the better performance of the GLLQ.

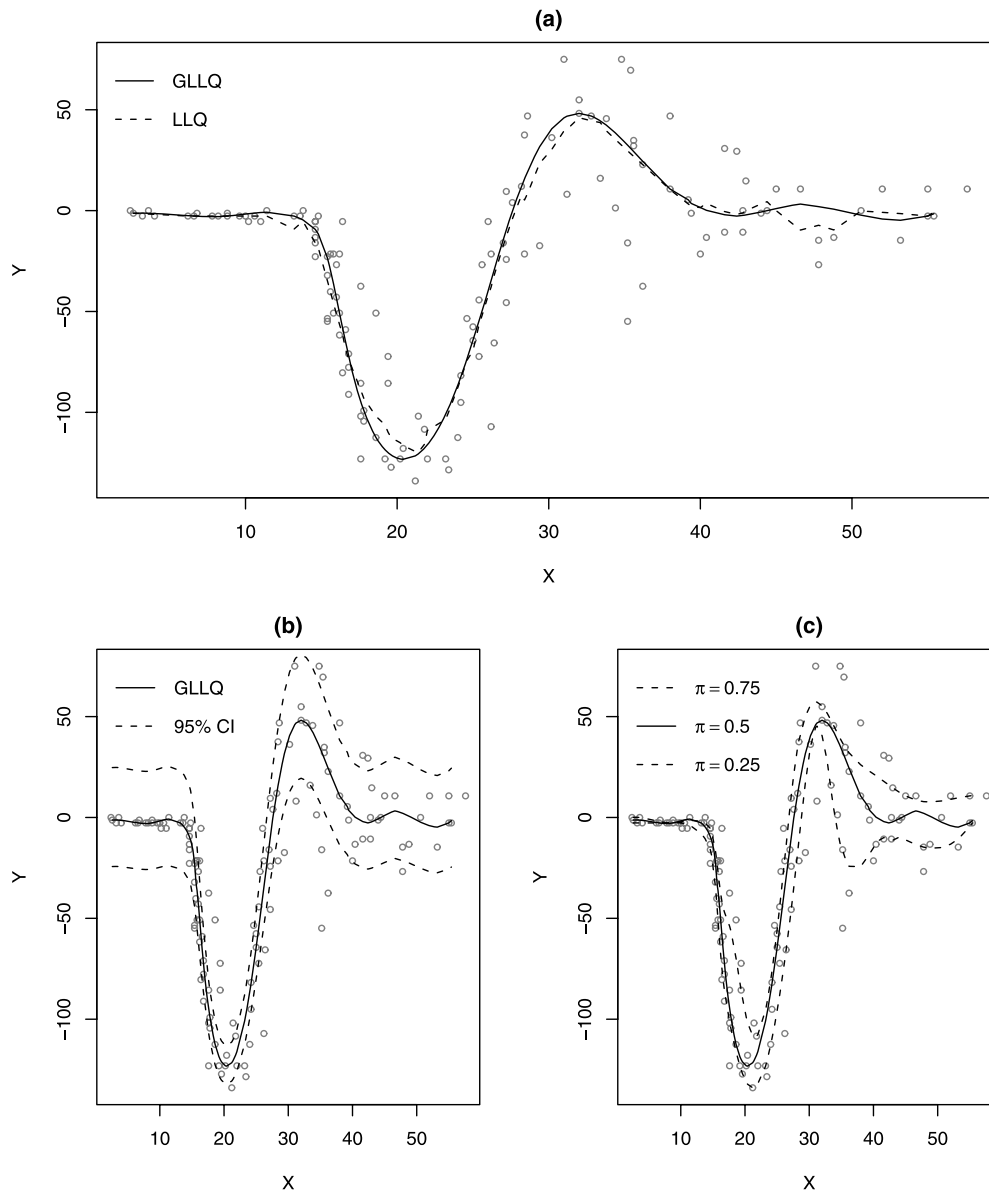


Figure 3. The LL median estimator and the guided LL median fit (a), the 95% confidence interval for the median function (b), and the guided LLQ estimators for  $\pi = 0.5$ ,  $\pi = 0.25$ , and  $\pi = 0.75$  (c).

APPENDIX

A.1 Proof of Theorem 1

We start by introducing some notation. Define  $a_u^1 = \sum_{j=0}^p \beta_j (u - x)^j$ . For a given  $\vartheta \in \mathbb{R}^{p+1}$ , let  $a_u^2(\vartheta) = a_n \vartheta^T (1, (u - x)/h_n, \dots, ((u - x)/h_n)^p)^T$ . We will use  $a_i^1 \equiv \tilde{\mathbf{X}}_{h,i}^T \boldsymbol{\beta}$  and  $a_i^2(\vartheta) \equiv a_n \vartheta^T \tilde{\mathbf{X}}_{h,i}$  as a shortcut for  $a_{X_i}^1$  and  $a_{X_i}^2(\vartheta)$ , respectively. Let  $\hat{\vartheta} = a_n^{-1} \mathbf{H}_n (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$  or, equivalently [see (3)],  $\hat{\vartheta} = \arg \min_{\vartheta} \sum_{i=1}^n \varphi_{\pi}(\tilde{Y}_i - a_i^2(\vartheta) - r_i(\hat{\boldsymbol{\theta}})) K_{h,i}$ , with  $\tilde{Y}_i = Y_i - a_i^1$ . Let  $\mathbf{V}_n(\vartheta, \boldsymbol{\theta}) = a_n \sum_{i=1}^n [\pi - I(\tilde{Y}_i - a_i^2(\vartheta) < r_i(\boldsymbol{\theta}))] \tilde{\mathbf{X}}_{h,i} K_{h,i}$ . This is the partial “derivative” of  $-\sum_{i=1}^n \varphi_{\pi}(\tilde{Y}_i - a_i^2(\vartheta) - r_i(\hat{\boldsymbol{\theta}})) K_{h,i}$  with respect to  $\vartheta$ . Hereafter,  $C$  designates a generic constant that may change from line to line. First, we show the following:

*Lemma 1.* For any  $0 < M < \infty$ , under the assumptions of Theorem 1,

$$\sup_{\|\boldsymbol{\theta}\| \leq M} \|\mathbf{V}_n(\vartheta, \hat{\boldsymbol{\theta}}) - \mathbf{V}_n(\vartheta, \boldsymbol{\theta}^*)\| = o_p(1) + O_p(a_n^{-1} \tilde{\delta}_n),$$

with  $\tilde{\delta}_n = h_n^{p+1} (h_n + \delta_n)$ .

*Proof.* Using (A4) and the fact that  $|I(y < b) - I(y < a)| \leq I(|y - a| \leq |b - a|)$ , we have  $\|\mathbf{V}_n(\vartheta, \hat{\boldsymbol{\theta}}) - \mathbf{V}_n(\vartheta, \boldsymbol{\theta}^*)\| \leq a_n \sum_{i=1}^n I(|\tilde{Y}_i - a_i^2(\vartheta) - r_i(\boldsymbol{\theta}^*)| \leq |r_i(\hat{\boldsymbol{\theta}}) - r_i(\boldsymbol{\theta}^*)|) K_{h,i}$ . By (A1.b) and Taylor’s expansion, there exist  $0 < \eta_1, \eta_2 < 1$  such that, as  $n \rightarrow \infty$ ,

$$\begin{aligned} |r_i(\hat{\boldsymbol{\theta}}) - r_i(\boldsymbol{\theta}^*)| &= \left| \frac{(X_i - x)^{p+1}}{(p+1)!} [q_{\pi}^{(p+1)}(x + \eta_1(X_i - x), \hat{\boldsymbol{\theta}}) \right. \\ &\quad \left. - q_{\pi}^{(p+1)}(x + \eta_2(X_i - x), \boldsymbol{\theta}^*)] \right| \\ &\leq C h_n^{p+1} (h_n + \delta_n) = C \tilde{\delta}_n, \end{aligned}$$

where in the last inequality we use the decomposition  $q_{\pi}^{(p+1)}(u_1, \hat{\boldsymbol{\theta}}) - q_{\pi}^{(p+1)}(u_2, \boldsymbol{\theta}^*) = [q_{\pi}^{(p+1)}(u_1, \hat{\boldsymbol{\theta}}) - q_{\pi}^{(p+1)}(x, \hat{\boldsymbol{\theta}})] + [q_{\pi}^{(p+1)}(x, \hat{\boldsymbol{\theta}}) - q_{\pi}^{(p+1)}(x, \boldsymbol{\theta}^*)] + [q_{\pi}^{(p+1)}(x, \boldsymbol{\theta}^*) - q_{\pi}^{(p+1)}(u_2, \boldsymbol{\theta}^*)]$ , together with assumptions (A1.a), (A1.c), (A1.d), and (A4). It follows that  $\|\mathbf{V}_n(\vartheta, \hat{\boldsymbol{\theta}}) - \mathbf{V}_n(\vartheta, \boldsymbol{\theta}^*)\| \leq a_n \sum_{i=1}^n I(|\tilde{Y}_i - a_i^2(\vartheta) - r_i(\boldsymbol{\theta}^*)| \leq C \tilde{\delta}_n) K_{h,i} := S_n(\vartheta)$ .

Now to prove Lemma 1, it is sufficient to show that

$$\sup_{\|\boldsymbol{\theta}\| \leq M} S_n(\boldsymbol{\theta}) = o_p(1) + O_p(a_n^{-1} \tilde{\delta}_n).$$

To do this, we use a chaining argument; thus we need to demonstrate the following:

(S1) For any  $\boldsymbol{\theta}$  such that  $\|\boldsymbol{\theta}\| \leq M$ ,  $S_n(\boldsymbol{\theta}) = O_p(a_n^{-1} \tilde{\delta}_n)$ .

(S2) For any  $\boldsymbol{\theta}$  and  $\tilde{\boldsymbol{\theta}}$  such that  $\|\boldsymbol{\theta}\| \leq M$  and  $\|\boldsymbol{\theta} - \tilde{\boldsymbol{\theta}}\| \leq C\tau$ ,  $|S_n(\boldsymbol{\theta}) - S_n(\tilde{\boldsymbol{\theta}})| \leq S_{n,\tau}^{\pm}(\tilde{\boldsymbol{\theta}})$ , where  $S_{n,\tau}^{\pm}(\tilde{\boldsymbol{\theta}})$  is a quantity that is free from the parameter  $\boldsymbol{\theta}$  and satisfying  $S_{n,\tau}^{\pm}(\tilde{\boldsymbol{\theta}}) = O_p(\tau)$ .

*Proof of (S1).* By Taylor's expansion, there exists an  $0 < \eta < 1$  such that

$$\begin{aligned} & \mathbb{E}[a_n S_n(\boldsymbol{\theta})] \\ &= na_n^2 \int [F_u(a_u^1 + r_u(\boldsymbol{\theta}^*) + a_u^2(\boldsymbol{\theta}) + C\tilde{\delta}_n) \\ & \quad - F_u(a_u^1 + r_u(\boldsymbol{\theta}^*) + a_u^2(\boldsymbol{\theta}) - C\tilde{\delta}_n)] K\left(\frac{u-x}{h}\right) f_0(u) du \\ &= 2C\tilde{\delta}_n \frac{1}{h_n} \int f_u(a_u^1 + r_u(\boldsymbol{\theta}^*) + a_u^2(\boldsymbol{\theta}) - C\tilde{\delta}_n + 2C\eta\tilde{\delta}_n) \\ & \quad \times K\left(\frac{u-x}{h}\right) f_0(u) du. \end{aligned}$$

Note that  $\tilde{\delta}_n \xrightarrow{n \rightarrow \infty} 0$ ,  $|a_u^2(\boldsymbol{\theta})| \leq Ca_n \xrightarrow{n \rightarrow \infty} 0$ ,  $r_u(\boldsymbol{\theta}^*) \xrightarrow{u \rightarrow x} 0$ , and  $a_u^1 \xrightarrow{u \rightarrow x} \beta_0$ . Using the fact that for any function  $g$  continuous at  $x$ ,  $h_n^{-1} \int g(u) \left(\frac{u-x}{h_n}\right)^i K^j\left(\frac{u-x}{h_n}\right) du \rightarrow g(x) \int u^i K^j(u) du$ ,  $i = 0, 1, \dots$ ,  $j = 1, 2, \dots$ , we conclude, by assumption (A3.b), that  $\mathbb{E}[S_n(\boldsymbol{\theta})] = O(a_n^{-1} \tilde{\delta}_n)$ . It remains to show that  $\text{Var}[S_n(\boldsymbol{\theta})] = o(1)$ . Let  $C_j(\boldsymbol{\theta}) = \text{Cov}(I(|\tilde{Y}_i - a_i^2(\boldsymbol{\theta}) - r_i(\boldsymbol{\theta}^*)| \leq C\tilde{\delta}_n) K_{h,1}, I(|\tilde{Y}_{j+1} - a_{j+1}^2(\boldsymbol{\theta}) - r_{j+1}(\boldsymbol{\theta}^*)| \leq C\tilde{\delta}_n) K_{h,j+1})$ ,  $j = 1, 2, \dots$ . By stationarity,  $\text{Var}[S_n(\boldsymbol{\theta})] = na_n^2 \{\text{Var}[I(|\tilde{Y}_i - a_i^2(\boldsymbol{\theta}) - r_i(\boldsymbol{\theta}^*)| \leq C\tilde{\delta}_n) K_{h,i}] + 2 \sum_{j=1}^n (1-j/n) C_j(\boldsymbol{\theta})\}$ . Observe that

$$\begin{aligned} & \text{Var}[I(|\tilde{Y}_i - a_i^2(\boldsymbol{\theta}) - r_i(\boldsymbol{\theta}^*)| \leq C\tilde{\delta}_n) K_{h,i}] \\ & \leq \mathbb{E}[I(|\tilde{Y}_i - a_i^2(\boldsymbol{\theta}) - r_i(\boldsymbol{\theta}^*)| \leq C\tilde{\delta}_n) K_{h,i}^2] \\ &= 2C\tilde{\delta}_n \int f_u(a_u^1 + r_u(\boldsymbol{\theta}^*) + a_u^2(\boldsymbol{\theta}) - C\tilde{\delta}_n + 2C\eta\tilde{\delta}_n) \\ & \quad \times K^2\left(\frac{u-x}{h}\right) f_0(u) du \quad \text{for some } 0 < \eta < 1 \\ &= O(\tilde{\delta}_n h_n) = o(h_n). \end{aligned}$$

Now, by the Cauchy-Schwartz inequality,

$$\begin{aligned} |C_j(\boldsymbol{\theta})| & \leq \text{Var}[I(|\tilde{Y}_1 - a_1^2(\boldsymbol{\theta}) - r_1(\boldsymbol{\theta}^*)| \leq C\tilde{\delta}_n) K_{h,1}] \\ &= o(h_n) \quad \text{for } j = 1, 2, \dots \end{aligned}$$

Also,

$$\begin{aligned} |C_j(\boldsymbol{\theta})| & \leq \mathbb{E}[I(|\tilde{Y}_1 - a_1^2(\boldsymbol{\theta}) - r_1(\boldsymbol{\theta}^*)| \leq C\tilde{\delta}_n) \\ & \quad \times I(|\tilde{Y}_{j+1} - a_{j+1}^2(\boldsymbol{\theta}) - r_{j+1}(\boldsymbol{\theta}^*)| \leq C\tilde{\delta}_n) K_{h,j+1} K_{h,1}] \\ & \quad + \{\mathbb{E}[I(|\tilde{Y}_1 - a_1^2(\boldsymbol{\theta}) - r_1(\boldsymbol{\theta}^*)| \leq C\tilde{\delta}_n) K_{h,1}]\}^2 \\ & \leq C\{\mathbb{E}[K_{h,j+1} K_{h,1}] + (\mathbb{E}[K_{h,1}])^2\}. \end{aligned}$$

By assumption (A2.b), for any  $j \geq j_*$ ,  $\mathbb{E}[K_{h,j+1} K_{h,1}] = \iint K\left(\frac{u_1-x}{h_n}\right) \times K\left(\frac{u_2-x}{h_n}\right) f_j(u_1, u_2) du_1 du_2 \leq M_* [\int K\left(\frac{u_1-x}{h_n}\right)]^2 = O(h_n^2)$ . By assumption (A3.b),  $\mathbb{E}[K_{h,1}] = O(h_n)$ . Thus  $|C_j(\boldsymbol{\theta})| = O(h_n^2)$ , for any  $j \geq j_*$ . It follows that for some  $0 < k_n \rightarrow \infty$ ,  $|\sum_{j=1}^n (1-j/n) C_j(\boldsymbol{\theta})| \leq$

$\sum_{j=1}^{j_*} |C_j(\boldsymbol{\theta})| + \sum_{j=j_*+1}^{k_n} |C_j(\boldsymbol{\theta})| + \sum_{j \geq k_n+1} |C_j(\boldsymbol{\theta})| = o(h_n) + O(k_n h_n^2) + O(k_n^{1-\nu})$ , where in the last equality we use our assumption (A2.a) and Billingsley's inequality (see, e.g., corollary 1.1 in Bosq 1998). We conclude that  $\text{Var}[S_n(\boldsymbol{\theta})] = o(1) + O(k_n h_n) + O(k_n^{1-\nu} h_n^{-1})$ , which converges to 0 by choosing an appropriate  $k_n$ .

*Proof of (S2).* It is easy to check that  $|S_n(\boldsymbol{\theta}) - S_n(\tilde{\boldsymbol{\theta}})| \leq S_{n,\tau}^-(\tilde{\boldsymbol{\theta}}) + S_{n,\tau}^+(\tilde{\boldsymbol{\theta}}) := S_{n,\tau}^{\pm}(\tilde{\boldsymbol{\theta}})$ , with  $S_{n,\tau}^-(\tilde{\boldsymbol{\theta}}) = a_n \sum_{i=1}^n I(|\tilde{Y}_i - r_i(\boldsymbol{\theta}^*) - a_i^2(\tilde{\boldsymbol{\theta}}) - C\tilde{\delta}_n| \leq \tau a_n) K_{h,i}$  and  $S_{n,\tau}^+(\tilde{\boldsymbol{\theta}}) = a_n \sum_{i=1}^n I(|\tilde{Y}_i - r_i(\boldsymbol{\theta}^*) - a_i^2(\tilde{\boldsymbol{\theta}}) + C\tilde{\delta}_n| \leq \tau a_n) K_{h,i}$ . Following the same procedure as before for  $S_n(\boldsymbol{\theta})$ , we can show that  $\mathbb{E}[S_{n,\tau}^{\pm}(\tilde{\boldsymbol{\theta}})] = O(\tau)$ ,  $\text{Var}[S_{n,\tau}^{\pm}(\tilde{\boldsymbol{\theta}})] = o(1)$ ,  $\mathbb{E}[S_{n,\tau}^+(\tilde{\boldsymbol{\theta}})] = O(\tau)$  and  $\text{Var}[S_{n,\tau}^+(\tilde{\boldsymbol{\theta}})] = o(1)$ . This concludes the proof of (S2).

Now, to complete the proof of Lemma 1, we use Bickel's chaining approach (Bickel 1975). We decompose the cube  $\mathcal{K} = \{\mathbf{v} \in \mathbb{R}^{p+1} : \|\mathbf{v}\| \leq M\}$  into cubes with vertices on the grid of points  $\{j_0 \tau M, \dots, j_p \tau M\}$ , with  $j_i = 0, \pm 1, \dots, \pm \lfloor 1/\tau \rfloor + 1$ . Let  $\mathbf{v}_{\boldsymbol{\theta}}$  be the lowest vertex of the cube containing  $\boldsymbol{\theta} \in \mathcal{K}$ . Note that  $|S_n(\boldsymbol{\theta}) - S_n(\mathbf{v}_{\boldsymbol{\theta}})| \leq C\tau$  and that  $\{\mathbf{v}_{\boldsymbol{\theta}} : \|\boldsymbol{\theta}\| \leq M\}$  is finite. Thus  $\sup |S_n(\boldsymbol{\theta})| \leq \sup |S_n(\mathbf{v}_{\boldsymbol{\theta}})| + \sup |S_n(\mathbf{v}_{\boldsymbol{\theta}})| \leq \max |S_{n,\tau}^{\pm}(\mathbf{v}_{\boldsymbol{\theta}})| + \max |S_n(\mathbf{v}_{\boldsymbol{\theta}})| = O_p(\tau) + O_p(a_n^{-1} \tilde{\delta}_n)$ . The result of Lemma 1 follows by letting  $\tau$  go to 0.

We now continue our demonstration of Theorem 1.

*Lemma 2.* Under the assumptions of Theorem 1, we have the following:

- (a)  $\|\mathbf{V}_n(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\theta}})\| = O_p(a_n)$ .
- (b)  $-\boldsymbol{\theta}^T \mathbf{V}_n(\lambda \boldsymbol{\theta}, \hat{\boldsymbol{\theta}}) \geq -\boldsymbol{\theta}^T \mathbf{V}_n(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}})$ , for any  $\lambda \geq 1$  and  $\boldsymbol{\theta} \in \mathbb{R}^{p+1}$ .
- (c)  $\|\mathbf{V}_n(\mathbf{0}, \boldsymbol{\theta}^*)\| = O_p(1)$ .
- (d)  $\sup_{\|\boldsymbol{\theta}\| \leq M} \|\mathbb{E}[\mathbf{V}_n(\boldsymbol{\theta}, \boldsymbol{\theta}^*) - \mathbf{V}_n(\mathbf{0}, \boldsymbol{\theta}^*)] + \mathbf{D}_x \boldsymbol{\theta}\| = o(1)$ , with  $\mathbf{D}_x = f(x, \beta_0) \boldsymbol{\Lambda}$ .
- (e)  $\sup_{\|\boldsymbol{\theta}\| \leq M} \|\mathbb{E}[\mathbf{V}_n(\boldsymbol{\theta}, \boldsymbol{\theta}^*) - \mathbf{V}_n(\mathbf{0}, \boldsymbol{\theta}^*)] - \mathbb{E}[\mathbf{V}_n(\boldsymbol{\theta}, \boldsymbol{\theta}^*) - \mathbf{V}_n(\mathbf{0}, \boldsymbol{\theta}^*)]\| = o_p(1)$ .

*Proof.* Part (a) is a direct application of the following lemma, the proof of which was given by Ruppert and Carroll (1980).

*Lemma 3.* For any random vectors  $\mathbf{X}_t \in \mathbb{R}^d$  and  $(A_t, B_t)^T \in \mathbb{R}^2$ ,  $t = 1, \dots, n$ , let  $\boldsymbol{\theta}_n = \arg \min_{\boldsymbol{\theta} \in \mathbb{R}^d} \sum_t \varphi_{\pi}(A_t - \boldsymbol{\theta}^T \mathbf{X}_t) B_t$ . If  $B_t \geq 0$ ,  $\mathbf{X}_t$  is continuous and  $\|\boldsymbol{\theta}_n\| < \infty$  then, with probability 1,  $\|\sum_t \mathbf{X}_t [\pi - I(A_t < \boldsymbol{\theta}_n^T \mathbf{X}_t)] B_t\| \leq d \max_t \|B_t \mathbf{X}_t\|$ .

Part (b) follows from the fact that  $\lambda \rightarrow -\boldsymbol{\theta}^T \mathbf{V}_n(\lambda \boldsymbol{\theta}, \hat{\boldsymbol{\theta}})$  is a nondecreasing function.

To prove part (c), we first note that  $\mathbf{V}_n(\mathbf{0}, \boldsymbol{\theta}^*) = (\mathbf{V}_n^{(0)}, \dots, \mathbf{V}_n^{(p)})$ , with  $\mathbf{V}_n^{(j)} = a_n \sum_{i=1}^n [\pi - I(\tilde{Y}_i < r_i(\boldsymbol{\theta}^*))] \left(\frac{X_{i,j}-x}{h}\right)^j K_{h,i}$ ,  $j = 0, \dots, p$ . We observe that

$$\begin{aligned} \mathbb{E}[\mathbf{V}_n^{(j)}] &= na_n \int [F_u(Q_{\pi}(u)) - F_u(a_u^1 + r_u(\boldsymbol{\theta}^*))] \\ & \quad \times f_0(u) \left(\frac{u-x}{h}\right)^j K\left(\frac{u-x}{h}\right) du. \end{aligned}$$

By Taylor's expansion, there exist  $0 < \eta_1, \eta_2, \eta_3 < 1$  such that:  $F_u(Q_{\pi}(u)) - F_u(a_u^1 + r_u(\boldsymbol{\theta}^*)) = (Q_{\pi}(u) - a_u^1 - r_u(\boldsymbol{\theta}^*)) f_u(a_u^1 + r_u(\boldsymbol{\theta}^*) + \eta_1(Q_{\pi}(u) - a_u^1 - r_u(\boldsymbol{\theta}^*)))$ ,  $Q_{\pi}(u) - a_u^1 = \frac{(u-x)^{p+1}}{(p+1)!} \times Q_{\pi}^{(p+1)}(x + \eta_2(u-x))$ , and  $r_u(\boldsymbol{\theta}^*) = \frac{(u-x)^{p+1}}{(p+1)!} q_{\pi}^{(p+1)}(x + \eta_3(u-x), \boldsymbol{\theta}^*)$ . Thus

$$\begin{aligned} \mathbb{E}[\mathbf{V}_n^{(j)}] &= a_n^{-1} \frac{h^{p+1}}{(p+1)!} \frac{1}{h_n} \int [Q_{\pi}^{(p+1)}(x + \eta_2(u-x)) \\ & \quad - q_{\pi}^{(p+1)}(x + \eta_3(u-x), \boldsymbol{\theta}^*)] \\ & \quad \times f_u(a_u^1 + r_u(\boldsymbol{\theta}^*) + \eta_1(Q_{\pi}(u) - a_u^1 - r_u(\boldsymbol{\theta}^*))) f_0(u) \end{aligned}$$

$$\begin{aligned} & \times \left(\frac{u-x}{h}\right)^j K\left(\frac{u-x}{h}\right) du \\ & = a_n^{-1} \frac{h^{p+1}}{(p+1)!} \left\{ [Q_\pi^{(p+1)}(x) - q_\pi^{(p+1)}(x, \theta^*)] \right. \\ & \quad \left. \times f(x, \beta_0) u_{j+p+1} + o(1) \right\}. \end{aligned}$$

In contrast, following a similar approach as before for  $S_n(\hat{\theta})$ , it is easy to check that  $\text{Var}[\mathbf{V}_n^{(j)}] = o(1)$  for  $j = 0, \dots, p$ . Thus we have shown that

$$\begin{aligned} \mathbb{E}[\mathbf{V}_n(\mathbf{0}, \theta^*)] & = a_n^{-1} \frac{h^{p+1}}{(p+1)!} \\ & \quad \times [Q_\pi^{(p+1)}(x) - q_\pi^{(p+1)}(x, \theta^*)] f(x, \beta_0) \tilde{\mathbf{u}} \\ & \quad + o(a_n^{-1} h^{p+1}) \end{aligned} \tag{A.1}$$

and  $\text{Var}[\mathbf{V}_n(\mathbf{0}, \theta^*)] = o(1)$ , which concludes the proof of part (c).

To prove part (d), by Taylor’s expansion, we can easily check that there exists  $0 < \eta < 1$  such that

$$\begin{aligned} \mathbb{E}[\mathbf{V}_n(\hat{\theta}, \theta^*) - \mathbf{V}_n(\mathbf{0}, \theta^*)] & = -h_n^{-1} \mathbb{E}[\tilde{\mathbf{X}}_{h,i} \tilde{\mathbf{X}}_{h,i}^T f_{X_i}(a_i^1 + r_i(\theta^*) + \eta a_i^2(\hat{\theta})) K_{h,i}] \hat{\theta}. \end{aligned}$$

Using the fact that  $\sup_{|u-x| \leq h_n, \|\hat{\theta}\| \leq M} |f_u(a_u^1 + r_u(\theta^*) + \eta a_u^2(\hat{\theta})) - f_x(\beta_0)| \rightarrow 0$ , it follows that, uniformly in  $\{\hat{\theta} : \|\hat{\theta}\| \leq M\}$ ,  $\mathbb{E}[\mathbf{V}_n(\hat{\theta}, \theta^*) - \mathbf{V}_n(\mathbf{0}, \theta^*)] = -h_n^{-1} \mathbb{E}[\tilde{\mathbf{X}}_{h,i} \tilde{\mathbf{X}}_{h,i}^T f_x(\beta_0) K_{h,i}] \hat{\theta} + o(1)$ , which leads to the desired result by observing that  $h_n^{-1} \mathbb{E}[\tilde{\mathbf{X}}_{h,i} \tilde{\mathbf{X}}_{h,i}^T f_x(\beta_0) K_{h,i}] \rightarrow \mathbf{D}_x$ .

We omit the proof of part (e), because it follows nearly the same lines as in the classical case of the fully nonparametric LLQ estimator by using a chaining argument, as was done in the proof of Lemma 1.

Now we complete the proof of Theorem 1. By Lemma 1 and parts (d) and (e) of Lemma 2, we get

$$\begin{aligned} & \sup_{\|\hat{\theta}\| \leq M} \|\mathbf{V}_n(\hat{\theta}, \hat{\theta}) + \mathbf{D}_x \hat{\theta} - \mathbf{V}_n(\mathbf{0}, \theta^*)\| \\ & \leq \sup_{\|\hat{\theta}\| \leq M} \|\mathbf{V}_n(\hat{\theta}, \hat{\theta}) - \mathbf{V}_n(\hat{\theta}, \theta^*)\| \\ & \quad + \sup_{\|\hat{\theta}\| \leq M} \|\mathbb{E}[\mathbf{V}_n(\hat{\theta}, \theta^*) - \mathbf{V}_n(\mathbf{0}, \theta^*)] + \mathbf{D}_x \hat{\theta}\| \\ & \quad + \sup_{\|\hat{\theta}\| \leq M} \|\mathbf{V}_n(\hat{\theta}, \theta^*) - \mathbf{V}_n(\mathbf{0}, \theta^*)\| \\ & \quad - \mathbb{E}[\mathbf{V}_n(\hat{\theta}, \theta^*) - \mathbf{V}_n(\mathbf{0}, \theta^*)]\| \\ & = o_p(1) + O_p(a_n^{-1} \tilde{\delta}_n) = O_p(a_n^{-1} h_n^{p+1} \delta_n) + o_p(1). \end{aligned}$$

This, together with parts (a), (b), and (c) of Lemma 2 and lemma A.4 in Koenker and Zhao (1996), leads to  $\hat{\beta} = \mathbf{D}_x^{-1} \mathbf{V}_n(\mathbf{0}, \theta^*) + O_p(a_n^{-1} \times h_n^{p+1} \delta_n) + o_p(1)$  or, equivalently,  $\mathbf{H}_n(\hat{\beta} - \beta) = a_n \mathbf{D}_x^{-1} \mathbf{V}_n(\mathbf{0}, \theta^*) + O_p(h_n^{p+1} \delta_n) + o_p(a_n)$ . Observe that  $\mathbf{V}_n(\mathbf{0}, \theta^*) = a_n \sum_{i=1}^n e_i \times \tilde{\mathbf{X}}_{h,i} K_{h,i} + \mathbf{B}_n(\theta^*)$ , with  $\mathbf{B}_n(\theta^*) = a_n \sum_{i=1}^n [I(Y_i < Q_\pi(X_i)) - I(\tilde{Y}_i < r_i(\theta^*))] \tilde{\mathbf{X}}_{h,i} K_{h,i}$ . It is easy to check that  $\mathbb{E}[\mathbf{B}_n(\theta^*)] = \mathbb{E}[\mathbf{V}_n(\mathbf{0}, \theta^*)]$  and  $\text{Var}[\mathbf{B}_n(\theta^*)] = o(1)$ . Using (A.1), we conclude that

$$\begin{aligned} \mathbf{H}_n(\hat{\beta} - \beta) & = \frac{h^{p+1}}{(p+1)!} [Q_\pi(x) - q_\pi(x, \theta^*)] \Lambda^{-1} \tilde{\mathbf{u}} \\ & \quad + \frac{a_n^2}{f(x, \beta_0)} \Lambda^{-1} \sum_{i=1}^n e_i \tilde{\mathbf{X}}_{h,i} K_{h,i} + \mathbf{r}_n, \end{aligned}$$

which achieves the proof of Theorem 1.

## A.2 Proof of Theorem 2

To prove Theorem 2, we need only show that  $a_n \Lambda^{-1} \sum_{i=1}^n e_i \tilde{\mathbf{X}}_{h,i} \times K_{h,i} \rightarrow \mathcal{N}_{p+1}(\mathbf{0}, \pi(1 - \pi) f_0(x) \Sigma)$ . This can be done using the Cramér–Wold device and the well-known small-blocks and large-blocks techniques in a very similar way as for the classical fully nonparametric LP mean regression (see Masry and Fan 1997). We omit the details here.

[Received July 2008. Revised March 2009.]

## REFERENCES

Akaike, H. (1973), “Information Theory and an Extension of the Maximum Likelihood Principle,” in *Second International Symposium on Information Theory (Tsahkadsor, 1971)*, Budapest: Akadémiai Kiadó, pp. 267–281.

Azzalini, A., and Genton, M. G. (2008), “Robust Likelihood Methods Based on the Skew- $t$  and Related Distributions,” *International Statistical Review*, 76, 106–129.

Bickel, P. J. (1975), “One-Step Huber Estimates in the Linear Model,” *Journal of the American Statistical Association*, 70, 428–434.

Bosq, D. (1998), *Nonparametric Statistics for Stochastic Processes*, New York: Springer.

Carrasco, M., and Chen, X. (2002), “Mixing and Moment Properties of Various GARCH and Stochastic Volatility Models,” *Econometric Theory*, 18, 17–39.

Chen, C. (2007), “A Finite Smoothing Algorithm for Quantile Regression,” *Journal of Computational and Graphical Statistics*, 16, 136–164.

Einsporn, R. (1987), “A Link Between Least Squares Regression and Nonparametric Curve Estimation,” Ph.D. dissertation, Virginia Tech, Blacksburg, VA.

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

Fan, J., Hu, T. C., and Truong, Y. K. (1994), “Robust Nonparametric Function Estimation,” *The Scandinavian Journal of Statistics*, 21, 433–446.

Fan, Y., and Ullah, A. (1999), “Asymptotic Normality of a Combined Regression Estimator,” *Journal of Multivariate Analysis*, 71, 191–240.

Francisco-Fernandez, M., and Vilar-Fernandez, J. (2005), “Bandwidth Selection for the Local Polynomial Estimator Under Dependence: A Simulation Study,” *Computational Statistics*, 20, 539–558.

Francisco-Fernandez, M., Opsomer, J., and Vilar-Fernandez, J. (2004), “Plug-in Bandwidth Selector for Local Polynomial Regression Estimator With Correlated Errors,” *Journal of Nonparametric Statistics*, 16, 127–151.

Glad, I. K. (1998), “Parametrically Guided Nonparametric Regression,” *The Scandinavian Journal of Statistics. Theory and Applications*, 25, 649–668.

Gozalo, P., and Linton, O. (2000), “Local Nonlinear Least Squares: Using Parametric Information in Nonparametric Regression,” *Journal of Econometrics*, 99, 63–106.

Hagmann, M., and Scaillet, O. (2007), “Local Multiplicative Bias Correction for Asymmetric Kernel Density Estimators,” *Journal of Econometrics*, 141, 213–249.

Härdle, W. (1990), *Applied Nonparametric Regression*. Econometric Society Monographs, Vol. 19, Cambridge: Cambridge University Press.

He, X., and Shao, Q.-M. (1996), “A General Bahadur Representation of  $M$ -Estimators and Its Application to Linear Regression With Nonstochastic Designs,” *The Annals of Statistics*, 24, 2608–2630.

Herrmann, E., and Mächler, M. (2003), “lोकern: Kernel Regression Smoothing With Local or Global Plug-in Bandwidth,” R package version 1.0-4.

Hunter, D. R., and Lange, K. (2000), “Quantile Regression via an MM Algorithm,” *Journal of Computational and Graphical Statistics*, 9, 60–77.

Jones, M. C., and Signorini, D. F. (1997), “A Comparison of Higher-Order Bias Kernel Density Estimators,” *Journal of the American Statistical Association*, 92, 1063–1073.

Koenker, R. (2005), *Quantile Regression*. Econometric Society Monographs, Vol. 38, Cambridge: Cambridge University Press.

Koenker, R., and Bassett, G., Jr. (1978), “Regression Quantiles,” *Econometrica*, 46, 33–50.

Koenker, R., and Park, B. J. (1996), “An Interior Point Algorithm for Nonlinear Quantile Regression,” *Journal of Econometrics*, 71, 265–283.

Koenker, R., and Zhao, Q. (1996), “Conditional Quantile Estimation and Inference for ARCH Models,” *Econometric Theory*, 12, 793–813.

Komunjer, I. (2005), “Quasi-Maximum Likelihood Estimation for Conditional Quantiles,” *Journal of Econometrics*, 128, 137–164.

Leung, D. H.-Y. (2005), “Cross-Validation in Nonparametric Regression With Outliers,” *The Annals of Statistics*, 33, 2291–2310.

Masry, E., and Fan, J. (1997), “Local Polynomial Estimation of Regression Functions for Mixing Processes,” *The Scandinavian Journal of Statistics*, 24, 165–179.

- Mays, J. E., Birch, J. B., and Starnes, B. A. (2001), "Model Robust Regression: Combining Parametric, Nonparametric, and Semiparametric Methods," *Journal of Nonparametric Statistics*, 13, 245–277.
- Naito, K. (2004), "Semiparametric Density Estimation by Local  $L_2$ -Fitting," *The Annals of Statistics*, 32, 1162–1191.
- Oberhofer, W., and Haupt, H. (2009), "Asymptotic Theory for Nonlinear Quantile Regression Under Weak Dependence," *Econometric Theory*, to appear.
- Ruppert, D., and Carroll, R. J. (1980), "Trimmed Least Squares Estimation in the Linear Model," *Journal of the American Statistical Association*, 75, 828–838.
- Schmidt, G., Mattern, R., and Schüler, F. (1981), "Biomechanical Investigation to Determine Physical and Traumatological Differentiation Criteria for the Maximum Load Capacity of Head and Vertebral Column With and Without Protective Helmet Under Effects of Impact," technical report, Final Report Phase III, Project 65, EEC Research Program on Biomechanics of Impacts, Institut für Rechtsmedizin, Universität Heidelberg, Heidelberg, Germany.
- Schwarz, G. (1978), "Estimating the Dimension of a Model," *The Annals of Statistics*, 6, 461–464.
- Shan, K., and Yang, Y. (2009), "Combining Regression Quantile Estimators," *Statistica Sinica*, 19, 1171–1191.
- Spiegelhalter, D. J., Best, N. G., Carlin, B. P., and van der Linde, A. (2002), "Bayesian Measures of Model Complexity and Fit," *Journal of the Royal Statistical Society, Ser. B*, 64, 583–639.
- Stasinopoulos, M., Rigby, B., and Akantziliotou, C. (2008), "gamlss: Generalized Additive Models for Location Scale and Shape," R package version 1.9-2.
- Su, L., and Ullah, A. (2008), "Nonparametric Prewhitening Estimators for Conditional Quantiles," *Statistica Sinica*, 18, 1131–1152.
- Yu, K., and Jones, M. C. (1998), "Local Linear Quantile Regression," *Journal of the American Statistical Association*, 93, 228–237.
- Zheng, Z. G., and Yang, Y. (1998), "Cross-Validation and Median Criterion," *Statistica Sinica*, 8, 907–921.