Journal of Nonparametric Statistics
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/gnst20

Adaptively weighted kernel regression

Qi Zheng[a], Colin Gallagher[a] & K.B. Kulasekera[a]
[a] Department of Mathematical Sciences, Clemson University, Clemson, SC 29634-0975, USA
Published online: 23 Aug 2013.

To cite this article: Journal of Nonparametric Statistics (2013): Adaptively weighted kernel regression, Journal of Nonparametric Statistics, DOI: 10.1080/10485252.2013.813511

To link to this article: http://dx.doi.org/10.1080/10485252.2013.813511

Journal of Nonparametric Statistics, 2013
http://dx.doi.org/10.1080/10485252.2013.813511

Taylor & Francis
Taylor & Francis Group

Adaptively weighted kernel regression

Qi Zheng*, Colin Gallagher and K.B. Kulasekera

Department of Mathematical Sciences, Clemson University, Clemson, SC 29634-0975, USA

(*Received 14 November 2012; accepted 4 June 2013*)

We develop a new kernel-based local polynomial methodology for nonparametric regression based on optimising a linear combination of several loss functions. Optimal weights for least squares and quantile loss functions can be chosen to provide maximum efficiency and these optimal weights can be estimated from data. The resulting estimators are at least as efficient as those provided by existing procedures, but can be much more efficient for many distributions. The data-based weights adapt to the tails of the error distribution resulting in a procedure which is both robust and resistant. Furthermore, the assumption of homogeneous error variance is not required. To illustrate its practical use, we apply the proposed method to model the motorcycle data.

Keywords: nonparametric regression; composite quantile; efficiency

1. Introduction

Consider a general nonparametric model

$$Y = m(X) + \sigma(X)\epsilon, \tag{1}$$

where Y is the response variable, X is the explanatory variable, $m(\cdot)$ is a smooth nonparametric regression function, $\sigma(X)$ is a smooth function and ϵ is random error with a probability density function symmetric about 0. Without loss of generality, we assume $E[\epsilon_i^2] = 1$, if $E[\epsilon_i^2] < \infty$.

Various methods have been developed to fit this type of model (Watson 1964; Wahba 1990; Fan and Gijbels 1996). It is fairly common to fit the model using weighted least squares (LS) with local polynomial approximation (Fan and Gijbels 1992). However, LS fitting can be very sensitive to heavy-tailed errors and severe outliers. Consequently, LS-based local polynomial regression could fail to produce reliable estimates in some cases. As a result, a lot of literature (Fan, Hu, and Truong 1994; Welsh 1996; Yu and Jones 1998; Jiang and Mack 2001; Chan and Zhang 2004) has been devoted to study robust local polynomial regression. Among those robust regression estimators, quantile regression (Koenker and Bassett 1978) is particularly attractive, since it can provide a more complete model of the relationship between predictors and response variables (Koenker 2005), owns excellent computational properties (Portnoy and Koenker 1997) and has widespread applications (Yu, Liu, and Stander 2003; Chernozhukov 2005). For some error structures, local polynomial quantile regression can be more efficient than the local LS polynomial regression. For example, if the error follows a Laplace distribution, the local median polynomial regression

*Corresponding author. Email: qiz@clemson.edu

has been demonstrated to be the most efficient (Fan et al. 1994; Welsh 1996). In other cases, the local quantile regression could be arbitrarily less efficient than the local LS polynomial regression (e.g. for normal data), resulting from the fact that loss functions of quantile regressions penalise residuals of small magnitude too strongly.

To improve the performance of quantile regression, Koenker and Portnoy (1987) considered L-estimation for linear models. An L-estimator is a weighted average of quantile estimators, which can achieve high efficiency for non-normal data. Bickel (1973) and Koenker (1984) demonstrate that as the number of quantiles used increases, the optimally weighted L-estimator is as efficient as the maximum likelihood estimator. However, it is difficult to find the optimal weights (Portnoy and Koenker 1989), and the computational cost increases dramatically with the number of quantiles. Instead, Zou and Yuan (2008) introduced composite quantile regression (CQR), which equally weights quantile loss functions. Kai, Li, and Zou (2010) adapted composite quantiles to the local polynomial framework. They showed that the local polynomial CQR can significantly improve the estimation efficiency of its local LS counterpart for common non-normal errors. However, the loss in efficiency compared to the LS polynomial regression still exists in many scenarios. In addition to that, it is unclear how many quantiles should be used in the local polynomial CQR. Even for a huge data set, increasing the number of quantiles does not necessarily improve the efficiency of estimates (Kai et al. 2010). Sun, Gai, and Lin (2013) recently proposed the local polynomial-weighted CQR (WCQR), which extends CQR to asymmetric error distributions. Since neither L-estimators nor CQR type estimators incorporate the LS loss, these estimators can require a large number of quantiles to achieve the efficiency especially when the magnitude of errors is small.

In this paper, we attempt to embed the usage of a convex combination of different loss functions into nonparametric kernel regression to obtain a robust estimator with significant improvement in efficiency. This idea is motivated by Bradic, Fan, and Wang (2011), which attempted to produce a robust and efficient estimator for high-dimensional linear regression by minimising composite loss functions simultaneously. However, different from a direct extension of Bradic et al. (2011) to nonparametric regression, we drop the finite second-moment assumption and combine the squared loss with multiple quantile loss functions for symmetric errors and pick weights to optimise the asymptotic efficiency. This results in a method which inherits strengths from both LS and quantile regression methods. We establish the asymptotic properties of the resulting estimator and show that it performs at least as well as the local LS polynomial estimator or the local polynomial CQR type estimators for any error distribution, and can improve the estimation efficiency for many distributions. Furthermore, it achieves the same efficiency as the optimally weighted L-estimator and can achieve higher efficiency than the equally WCQR of Kai et al. (2010). We propose a simple data-driven procedure to select weights for the convex combination and show that the aforementioned asymptotic properties can be achieved by this adaptively weighted local polynomial regression estimator. The adaptively weighted local polynomial estimator is quite robust and works well even if the error distribution does not have a finite variance.

2. The adaptively weighted local polynomial regression

We start by setting up notations. Let $\rho_\tau(t) = \tau 1(t > 0)t - (1-\tau)1(t \le 0)t$ be the check function with quantile index τ. Let $\tau_k = k/(q+1), k = 1, \ldots, q$ be equally spaced quantile indices between 0 and 1. We denote the τ_kth quantile by $q_{\tau_k}, k = 1, \ldots, q$. In particular, let $\tau_0 = 0, q_{\tau_0} = 0$ and $\rho_{\tau_0}(t) = t^2$. We use $F(\cdot)$ and $f(\cdot)$ to denote the cumulative distribution function and probability density function of ϵ_i, respectively. $g_X(\cdot)$ is the marginal density of X. $K(\cdot)$ is a classical kernel function. We also use the following notations $\tau_{k,k'} = \tau_{k\wedge k'} - \tau_k\tau_{k'}$, where $k \wedge k' = \min\{k, k'\}$, and $\tau_{0,k} = E[\epsilon_i 1(\epsilon_i \le q_{\tau_k})]$, for $k = 1, \ldots, q$.

In order to define the adaptively weighted local polynomial regression, let us briefly review the local LS polynomial regression, the local quantile polynomial regression and the local polynomial CQR.

Let (x_i, y_i) be n independently and identically distribution observations. Of interest is to estimate the value of $m(X)$ at x_0. Suppose $m(X)$ is smooth enough to be approximated by a pth-order polynomial in a neighbourhood of x_0, that is, $m(x) \approx \sum_{j=0}^{p} (1/j!) m^{(j)}(x_0)(x - x_0)^j$. The local LS polynomial regression estimator of $(m(x_0), m^{(1)}(x_0), \ldots, m^{(p)}(x_0))$ is defined as the minimiser of the following objective function

$$\min_{a_0, a_1, \ldots, a_p} \sum_{i=1}^{n} \rho_{\tau_0} \left(y_i - \sum_{j=0}^{p} \frac{1}{j!} a_j (x_i - x_0)^j \right) K \left(\frac{x_i - x_0}{h} \right), \tag{2}$$

where h is a smoothing parameter. Fan and Gijbels (1992) demonstrated that the local LS polynomial regression owns several desirable properties: it adapts to a wide variety of design densities, significantly reduces bias at boundary points, and attains high minimax efficiency.

However, the local LS polynomial regression suffers from outliers and heavy-tailed errors. Motivated by its robustness and other good features, several authors (Fan et al. 1994; Welsh 1996; Yu and Jones 1998) advocated the local quantile polynomial regression

$$\min_{a_0, a_1, \ldots, a_p} \sum_{i=1}^{n} \rho_{\tau} \left(y_i - \sum_{j=0}^{p} \frac{1}{j!} a_j (x_i - x_0)^j \right) K \left(\frac{x_i - x_0}{h} \right),$$

for some quantile index τ. Although the local quantile polynomial regression can be applied for more general error structures, it can be arbitrarily inefficient compared to the local LS polynomial regression. To improve the efficiency of the local quantile polynomial regression while maintaining the robustness, Kai et al. (2010) proposed the local polynomial CQR as follows

$$\min_{a_{01}, \ldots, a_{0q}, a_1, \ldots, a_p} \sum_{i=1}^{n} \sum_{k=1}^{q} \rho_{\tau_k} \left(y_i - a_{0k} - \sum_{j=1}^{p} \frac{1}{j!} a_j (x_i - x_0)^j \right) K \left(\frac{x_i - x_0}{h} \right). \tag{3}$$

They showed that the local polynomial CQR can significantly improve the efficiency compared to the local quantile polynomial regression. However, the loss of efficiency of the local polynomial CQR still exists for some commonly seen distributions.

We consider combining the local LS and CQR from Equations (2) and (3) to produce an efficient and robust regression estimator. Let $\theta = (a_{01}, \ldots, a_{0q}, a_0, a_1, \ldots, a_p)$ and denote the solution to the objective function

$$\min_{\theta} \sum_{i=1}^{n} \left[\sum_{k=0}^{q} \beta_k \rho_{\tau_k} \left(y_i - a_{0k} - \sum_{j=0}^{p} \frac{1}{j!} a_j (x_i - x_0)^j \right) \right] K \left(\frac{x_i - x_0}{h} \right), \tag{4}$$

by $\hat{\theta}_\beta = (\hat{a}_{01}, \ldots, \hat{a}_{0q}, \hat{a}_0, \hat{a}_1, \ldots, \hat{a}_p)$. Here, $a_{00} = 0$ and $\beta_0, \ldots, \beta_q$ are well-chosen non-negative weights which adapt to the error structures. The details about how to choose those weights

are presented in Section 3.2. Under the regularity conditions presented in Section 3.1, we can show that

$$\sqrt{nh_n}\left(\frac{\sum_{k=1}^{q}(1/2\sigma)\beta_k f(q_{\tau_k})\hat{a}_{0k}}{\beta_0+\sum_{k=1}^{q}(1/2\sigma)\beta_k f(q_{\tau_k})}+\hat{a}_0-m(x_0)-\frac{1}{2}m^{(2)}(x_0)\mu_2 h_n^2\right)$$
$$\xrightarrow{L} N\left(0,\frac{\nu_0}{4g(x_0)}\frac{V_\beta}{(\beta_0+\sum_{k=1}^{q}(1/2\sigma)\beta_k f(q_{\tau_k}))^2}\right).$$

Therefore, we define the adaptively weighted local polynomial regression estimator as

$$\hat{m}_\beta(x_0)=\frac{(1/2\sigma)\sum_{k=1}^{q}\beta_k f(q_{\tau_k})\hat{a}_{0k}}{\beta_0+(1/2\sigma)\sum_{k=1}^{q}\beta_k f(q_{\tau_k})}+\hat{a}_0,\quad \hat{m}_\beta^{(j)}(x_0)=\hat{a}_j,\quad j=1,\ldots,p. \tag{5}$$

For identification purposes, we set $\sigma\beta_0+\sum_{k=1}^{q}\beta_k f(q_{\tau_k})/2=1$. The regression estimator from Equation (5) becomes

$$\hat{m}_\beta(x_0)=\frac{1}{2}\sum_{k=1}^{q}\beta_k f(q_{\tau_k})\hat{a}_{0k}+\hat{a}_0.$$

This formulation actually provides an advantage. In the following sections, it can be seen that the variances of $\hat{m}_\beta(x_0)$ and $\hat{m}_\beta^{(j)}(x_0)$, $1\le j\le p$ are of similar forms. Consequently, minimising them separately still produces the same optimal weights vector β.

3. Asymptotic properties

In this section, we state primitive regularity conditions and then establish the asymptotic properties of the adaptively weighted local polynomial regression estimator.

3.1. *Regularity Conditions*

To study the asymptotic properties of the adaptively weighted local polynomial regression estimator, the following regularity conditions are assumed throughout the rest of this paper.

(A) $m(\cdot)$ has continuous $(p+2)$th derivative in the neighbourhood of x_0.
(B) f is symmetric about 0 and belongs to the domain of attraction of some stable distribution S; this includes all distributions for which normalised sums converge to a weak limit (Feller 1971).
(C) f is continuous and positive.
(D) $g_X(\cdot)$ is positive and differentiable in the neighbourhood of x_0.
(E) $K(\cdot)$ is a symmetric kernel function with a compact support $[-M,M]$, and satisfies
  (a) $|K(u)|<C_k$,
  (b) $\int_{-M}^{M}K(u)\,\mathrm{d}u=1$,
  (c) $\int_{-M}^{M}u^jK(u)\,\mathrm{d}u=\mu_j$, $\int_{-M}^{M}u^jK^2(u)\,\mathrm{d}u=\nu_j$, $j\ge 0$. In particular, $\mu_j=\nu_j=0$ for odd j.

Regularity conditions A, C, D and E are commonly assumed in the literature (Fan 1992; Yu and Jones 1998; Kai et al. 2010). As is pointed out elsewhere, the assumption that $K(\cdot)$ has a compact support can be relaxed at the cost of more complicated technical proofs. In simulation studies, we exhibit the excellent performance of the proposed estimator with the classical normal kernel. The assumption that f is symmetric about 0 is required in Kai et al. (2010). Although WCQR for asymmetric errors was considered recently in Sun et al. (2013), we still maintain the symmetry

assumption to simplify the complicated proof that the impact of the LS part is negligible when $E[\epsilon_i^2]$ does not exist. However, our estimator can be generalised to asymmetric distributions following Sun et al. (2013).

Up front, under the assumption that $E[\epsilon_i^2] < \infty$, we establish the asymptotic properties of the adaptively weighted local polynomial regression estimator to demonstrate that it is more efficient and hence is favourable to other polynomial regression estimators. Next, we consider $E[\epsilon_i^2] = \infty$ and show that the impact of the LS part in the adaptively weighted local polynomial regression estimator is asymptotically negligible, while the efficiency is preserved under this infinite variance scenario. Therefore, the proposed estimator is a robust and efficient alternative to other polynomial regression estimators.

To avoid complicated statements, we first illustrate our ideas via the i.i.d error models:

$$Y = m(X) + \sigma\epsilon,$$

and then generalise it to heterogeneous error models.

3.2. *Asymptotic properties when $E[\epsilon_i]^2$ exists*

Throughout this subsection, we assume $E[\epsilon_i]^2 < \infty$. To state the asymptotic properties of the adaptively weighted local polynomial regression estimator, we need to introduce the following notations:

Define

$$S(\boldsymbol{\beta}) = \begin{pmatrix} S_{11}(\beta) & S_{12}(\beta) \\ S_{21}(\beta) & S_{22}(\beta) \end{pmatrix},$$

where $S_{11}(\beta)$ is a $q \times q$ diagonal matrix with diagonal elements $\beta_k f(q_{\tau_k})/(2\sigma)$, for $k = 1, \ldots, q$, S_{22} is a $(p+1) \times (p+1)$ matrix with (j, j')-entry $\mu_{(j+j'-2)}$, for $j, j' = 1, \ldots, p+1$, and $S_{12}(\beta) = S_{21}(\beta)^{\mathrm{T}}$ is a $q \times (p+1)$ matrix with (k, j)-entry $\beta_k f(q_{\tau_k})/(2\sigma)\mu_{j-1}$, for $k = 1, \ldots, q$; $j = 1, \ldots, p+1$.

Let

$$V_\beta = 4\beta_0^2\sigma^2 - 4\beta_0 \sum_{k=1}^{q} \beta_k \sigma \tau_{0,k} + \sum_{k,k'=1}^{q} \beta_k \beta_{k'} \tau_{k,k'},$$

and define

$$\Sigma(\boldsymbol{\beta}) = \begin{pmatrix} \Sigma_{11}(\beta) & \Sigma_{12}(\beta) \\ \Sigma_{21}(\beta) & \Sigma_{22}(\beta) \end{pmatrix},$$

where $\Sigma_{11}(\beta)$ is a $q \times q$ matrix with (k, k')-entry $\beta_k \beta_{k'} \nu_0 \tau_{k,k'}$, for $k, k' = 1, \ldots, q$, $\Sigma_{22}(\beta)$ is a $(p+1) \times (p+1)$ matrix with (j, j')th element $V_\beta \nu_{(j+j'-2)}$, for $j, j' = 1, \ldots, p+1$, and $\Sigma_{12}(\beta) = \Sigma_{21}^{\mathrm{T}}(\beta)$ is a $q \times (p+1)$ matrix with (k, j)-entry $(-2\beta_0\beta_k\sigma\tau_{0,k} + \beta_k \sum_{k'=1}^{q} \beta_{k'} \tau_{k,k'})\nu_{(j-1)}$, for $k = 1, \ldots, q; j = 1, \ldots, (p+1)$.

Let $r_{i,p} = m(x_i) - \sum_{j=0}^{p} m^{(j)}(x_0)(x_i - x_0)^j/j!$ be the residual of the Taylor expansion of $m(x_i)$ at x_0, and $\xi_{\beta,i} = -2\beta_0(\sigma\epsilon_i + r_{i,p}) + \sum_{k=1}^{q} \beta_k[1(\epsilon_i \le (\sigma q_{\tau_k} - r_{i,p})/\sigma) - \tau_k]$. We define $W_{\beta,n} = (w_{\beta,01}, \ldots, w_{\beta,0q}, w_{\beta,0}, w_{\beta,1}, \ldots, w_{\beta,p})^{\mathrm{T}}$, where

$$w_{\beta,0k} = \beta_k \frac{1}{\sqrt{nh_n}} \sum_{i=1}^{n} K\left(\frac{x_i - x_0}{h_n}\right)\left[1\left(\epsilon_i \le \frac{\sigma q_{\tau_k} - r_{i,p}}{\sigma}\right) - \tau_k\right], \quad k = 1, \ldots, q;$$

$$w_{\beta,j} = \frac{1}{\sqrt{nh_n}} \sum_{i=1}^{n} K\left(\frac{x_i - x_0}{h_n}\right)\left(\frac{x_i - x_0}{h_n}\right)^j \xi_{\beta,i}, \quad j = 0, \ldots, p.$$

Then, the asymptotic properties of the adaptively weighted local polynomial regression estimator can be established in the following theorem:

THEOREM 3.1 *Suppose assumptions A–E are satisfied. Furthermore, we assume* $E[\epsilon_i^2] < \infty$. *If* $h_n \to 0$ *and* $nh_n \to \infty$, *then for any nonnegative weights vector* $\beta = (\beta_0, \ldots, \beta_q)^{\mathrm{T}}$,

$$\sqrt{nh_n}S(\beta)A_{h_n}(\hat{\theta}_\beta - \theta^*) + \frac{1}{2g(x_0)}E[W_{\beta,n}] \xrightarrow{L} N\left(0, \frac{1}{4g(x_0)}\Sigma(\beta)\right),$$

where $\theta^* = (q_{\tau_1}, \ldots, q_{\tau_q}, m(x_0), m^{(1)}(x_0), \ldots, m^{(p)}(x_0))^{\mathrm{T}}$ *is a vector of true parameters and* A_{h_n} *is a* $(q+1+p) \times (q+1+p)$ *diagonal matrix with diagonal elements* $(1, \ldots, 1, h_n^0/0!, \ldots, h_n^p/p!)$.

As special cases, two corollaries follow immediately.

COROLLARY 3.1 *Under the same assumptions as Theorem* 3.1, *if* $p = 1$, *we have*

$$\sqrt{nh_n}\left[\hat{m}_\beta(x_0) - m(x_0) - \frac{m^{(2)}(x_0)}{2}\mu_2 h_n^2\right] \xrightarrow{L} N\left(0, \frac{\nu_0\sigma^2}{4g(x_0)}V_\beta\right),$$

and the mean squared error (MSE) of $\hat{m}_\beta(x_0)$ *is*

$$\mathrm{MSE}(\hat{m}_\beta(x_0)) = \left(\frac{m^{(2)}(x_0)}{2}\mu_2\right)^2 h_n^4 + \frac{\nu_0\sigma^2}{4g(x_0)}\frac{V_\beta}{nh_n} + o_p\left(h_n^4 + \frac{1}{nh_n}\right). \tag{6}$$

COROLLARY 3.2 *Under the same assumptions as Theorem* 3.1, *if* $p = 1$, *then*

$$\sqrt{nh_n}\left[\hat{m}_\beta^{(1)}(x_0) - m^{(1)}(x_0) - \left(\frac{m^3(x_0)}{6} + \frac{m^{(2)}(x_0)g^{(1)}(x_0)}{2g(x_0)}\right)\frac{\mu_4}{\mu_2}h_n^2\right] \xrightarrow{L} N\left(0, \frac{\nu_2\sigma^2}{4g(x_0)h_n^2\mu_2^2}V_\beta\right),$$

and

$$\mathrm{MSE}(\hat{m}_\beta^{(1)}(x_0)) = \left(\frac{m^{(3)}(x_0)}{6} + \frac{m^{(2)}(x_0)g^{(1)}(x_0)}{2g(x_0)}\right)^2 \frac{\mu_4^2}{\mu_2^2}h_n^4 + \frac{\nu_2\sigma^2}{4g(x_0)\mu_2^2}\frac{V_\beta}{nh_n^3} + o_p\left(h_n^4 + \frac{1}{nh_n^3}\right). \tag{7}$$

If $p = 2$, *then*

$$\sqrt{nh_n}\left[\hat{m}_\beta^{(1)}(x_0) - m^{(1)}(x_0) - \frac{m^{(3)}(x_0)\mu_4}{6\mu_2}h_n^2\right] \xrightarrow{L} N\left(0, \frac{\nu_2\sigma^2}{4g(x_0)h_n^2\mu_2^2}V_\beta\right),$$

and

$$\mathrm{MSE}(\hat{m}_\beta^{(1)}(x_0)) = \left(\frac{m^{(3)}(x_0)}{6}\right)^2 \frac{\mu_4^2}{\mu_2^2}h_n^4 + \frac{\nu_2\sigma^2}{4g(x_0)\mu_2^2}\frac{V_\beta}{nh_n^3} + o_p\left(h_n^4 + \frac{1}{nh_n^3}\right). \tag{8}$$

Equation (6) indicates that the asymptotic MSE depends on β only through V_β. Thus, the optimal weights vector β in the sense of minimising the MSE of $\hat{m}(x_0)$ can be chosen by minimising V_β:

$$\beta_{\mathrm{opt}} = \underset{\beta \geq 0, \boldsymbol{\alpha}^{\mathrm{T}}\beta = 1}{\operatorname{argmin}} V_\beta, \tag{9}$$

where $\alpha = (\sigma, f(q_{\tau_k})/2, \ldots, f(q_{\tau_q})/2)^{\mathrm{T}}$.

Remark 3.1 When $q \to \infty$, if we set $\beta_0 = 0$ and minimise V_β with respect to β, the resulting covariance matrix is the same as that of the nonparametric polynomial L-estimation with optimal weights. Therefore, the proposed method is also as efficient as the maximum likelihood, when $q \to \infty$.

Noting that the MSE of $\hat{m}_\beta^{(1)}(x_0)$ from Equations (7) and (8) only depends on β by V_β as well, then β_{opt} is also optimal for estimating $m^{(1)}(x_0)$. For most practical interests, estimating $m(x_0)$ is the main focus. However, it can be shown that β_{opt} is optimal for estimating all $m^{(j)}(x_0)$ for $j \le p$.

Since Equation (9) is a constrained quadratic minimisation problem, the closed form solution for the optimal weights can be difficult to obtain. However, in some cases, optimal weights can be explicitly found. We provide several examples to show the availability of the optimal weights.

Example 3.1 Let $q = 1, \tau_1 = \frac{1}{2}$ and $p \ge 1$. In other words, we consider the combination of LS and least absolute deviation (LAD), then the optimal weights are

$$\beta_{1,\text{opt}} = \frac{2(1 - \sigma\beta_{0,\text{opt}})}{f(0)}$$

$$\text{and} \quad \beta_{0,\text{opt}} = \begin{cases} 0 & \text{if } \dfrac{1 - 2f(0)E[|\epsilon_i|]}{4f(0)^2 - 4f(0)E[|\epsilon_i|] + 1} < 0, \\ 1 & \text{if } \dfrac{1 - 2f(0)E[|\epsilon_i|]}{4f(0)^2 - 4f(0)E[|\epsilon_i|] + 1} > 1, \\ \dfrac{1}{\sigma} \dfrac{1 - 2f(0)E[|\epsilon_i|]}{4f(0)^2 - 4f(0)E[|\epsilon_i|] + 1} & \text{otherwise.} \end{cases}$$

Example 3.2 If $\epsilon_i \sim N(0, 1)$, then $\beta_{0,\text{opt}} = 1/\sigma$ and $\beta_{k,\text{opt}} = 0$, for all $1 \le k \le q$.

Example 3.3 If $\epsilon_i \sim \text{Laplace}(0, 1/\sqrt{2})$ and q is odd, then $\beta_{l,\text{opt}} = 2f(0)$ for $l = (q+1)/2$, and $\beta_{0,\text{opt}} = \beta_{k,\text{opt}} = 0$, for all $k \ne l$.

Both the local polynomial LS and the local polynomial CQR estimators are special cases of the weighted local polynomial regression. Regardless of the error distribution, the efficiency achieved by choosing the theoretically optimal weights can be no less than that gained by either of those methods. Moreover, the proposed estimator with true optimal weights can be more efficient than the local LS polynomial regression estimator and the local polynomial CQR for some distributions, as Examples 3.1 and 3.2 demonstrate. Theorem 2 in Kai et al. (2010) indicates that as the number of quantiles increases, the asymptotic relative efficiency between CQR and LS converges to 1. For the proposed weighted estimator, increasing the number of quantiles does not impact the efficiency in this way, but can in fact improve the asymptotic efficiency of the estimator.

Although V_β is typically unobservable, we can replace it with a consistent estimator. Let $\tilde{\zeta}_i$ be residuals of a $\sqrt{nh_n}$-consistent preliminary estimation, $i = 1, \ldots, n$. For example, we could use the residuals from the local polynomial median regression, or residuals from the local polynomial CQR, or the residuals from the local LS polynomial regression, if the error terms $\{\epsilon_i\}$ have a finite second moment. We note here that both CQR and WCQR require preliminary estimation as well.

We use the notation $\tilde{T}$ to denote the empirical estimate of T using $\tilde{\zeta}_i$, for some statistic T. Then, $\tilde{V}_\beta = 4\beta_0^2\tilde{\sigma}^2 - 4\beta_0 \sum_{k=1}^q \beta_k \tilde{\sigma}\,\tilde{\tau}_{0,k} + \sum_{k,k'=1}^q \beta_k \beta_{k'} \tau_{k,k'}$, and we can obtain the practically optimal weights vector

$$\hat{\beta} = \underset{\beta \ge 0, \tilde{\alpha}^{\mathrm{T}}\beta = 1}{\text{argmax}}\ \tilde{V}_\beta, \tag{10}$$

where $\tilde{\alpha} = (\tilde{\sigma}, \frac{1}{2}\tilde{f}(\tilde{q}_{\tau_1}), \ldots, \frac{1}{2}\tilde{f}(\tilde{q}_{\tau_q}))^{\mathrm{T}}$. The consistency of $\hat{\beta}$ can easily be verified. We have the following corollary:

COROLLARY 3.3 *Under the same assumptions as Theorem* 3.1,

$$\sqrt{nh_n}S(\beta_{\text{opt}})A_{h_n}(\hat{\theta}_{\hat{\beta}} - \theta^*) + \frac{1}{2g(x_0)}E[W_{\beta_{\text{opt}},n}] \xrightarrow{L} N\left(0, \frac{1}{4g(x_0)}\Sigma(\beta_{\text{opt}})\right),$$

The proposed estimator, using $\hat{\beta}$ obtained from Equation (10) does not suffer from any loss of asymptotic efficiency.

Notice that in $\tilde{V}_\beta$, q_{τ_k}, $\tau_{0,k}$ and $f(q_{\tau_k})$ need to be estimated. Therefore, the number of quantiles should depend on the sample size. When the sample size is small we recommend using only a few quantiles to avoid the impact of introducing too many parameters. On the other hand, if the sample size is large, more quantiles should be adopted. In practice, cross-validation, akaike information criterion (AIC), Bayesian information criterion (BIC) type criteria can be applied to choose the number of quantiles.

3.3. *Asymptotic properties when $E[\epsilon_i^2]$ does not exist*

Since LS may not provide reliable estimates when heavy-tailed errors or outliers appear, in this case, one might use a weight of zero ($\beta_0 = 0$) for the LS part of objective function (4). In practice, we do not know if the variance is finite and we propose picking weights using a numerical solution to the constrained quadratic minimisation problem (10). Therefore, $\hat{\beta}_0$ is not necessarily 0. We would like to find out if the proposed estimator can still be applied. The following theorem answers the aforementioned question.

THEOREM 3.2 *Suppose assumptions A–E are satisfied. Furthermore, we assume $E[\epsilon_i^2]$ does not exist. If $h_n \to 0$ and $nh_n \to \infty$, then*

$$\sqrt{nh_n}S(\beta_{\text{opt}})A_{h_n}(\hat{\theta}_{\hat{\beta}} - \theta^*) + \frac{1}{2g(x_0)}E[W_{\beta_{\text{opt}},n}] \xrightarrow{L} N\left(0, \frac{1}{4g(x_0)}\Sigma(\beta_{\text{opt}})\right).$$

This theorem indicates that $\hat{\beta}_0$ converges to 0 fast enough to make the instability caused by LS negligible. Theorem 3.2 coupled with Theorem 3.1 imply that the adaptively weighted local polynomial regression can be applied universally. In addition, because $\hat{\beta}$ is chosen to adapt to different error distributions, the resulting local polynomial regression estimator is asymptotically more efficient than the local polynomial CQR. Those features make the proposed estimator very appealing in practice.

3.4. *Heterogeneous errors*

In the foregoing sections, we exhibit the desirable theoretical properties of the adaptively weighted local polynomial regression estimator under the homogeneous model. An interesting question naturally arises: 'Can this method be applied to regression problems of which the error sequences are heterogeneous?'

The essential idea of the proposed procedure is to use the residuals from some preliminary method to select approximately optimal weights for the different loss functions. If the error sequences are homogeneous, then all residuals can be employed to establish the error structure. On the other hand, if the errors are heterogeneous, residuals of observations with covariate values closer to x_0, the point of interest, should contribute more to the local error structure estimation. Hence, we can use weighted squares of residuals to estimate $\sigma(\cdot)$ at x_0, where weights are assigned by a kernel function. Take the uniform kernel as an illustration, as $n \to \infty$, the number of observations falling into $[x_0 - h_n, x_0 + h_n]$ is of order nh_n. Therefore, the asymptotic efficiency should

not be impacted by doing the local error structure estimation. In practice, the pilot fit also provides initial bandwidths so that we can manipulate observations falling into the smoothing window to approximate $\sigma(\cdot)$ locally.

THEOREM 3.3 *Under model* (1), *suppose assumptions A–E are satisfied. Furthermore, if* $h_n \to 0$ *and* $nh_n \to \infty$, *then*

$$\sqrt{nh_n}S(\beta_{\text{opt}})A_{h_n}(\hat{\theta}_{\hat{\beta}} - \theta^*) + \frac{1}{2g(x_0)}E[W_{\beta_{\text{opt}},n}] \xrightarrow{L} N\left(0, \frac{1}{4g(x_0)}\Sigma(\beta_{\text{opt}})\right),$$

where $\hat{\beta}$ *is obtained from Equation* (10) *using weighted local residuals.*

In heterogeneous cases, we use the following procedure to get the practical optimal weights.

Step 1 Obtain a $\sqrt{nh_n}$-consistent pilot estimator $\tilde{m}(\cdot)$ for $m(\cdot)$ with bandwidth $h(\cdot)$.
Step 2 Implement the procedure proposed by Ruppert, Wand, Holst, and Hössjer (1997) to construct $\tilde{\sigma}^2(\cdot)$ using weighted squares of residuals from the initial fit.
Step 3 Compute $\tilde{\epsilon}_i = (y_i - \tilde{m}(x_i))/\tilde{\sigma}(x_i)$.
Step 4 Estimate the sample quantiles $\tilde{q}_{\tau_k}, k = 1, \ldots, q$. Then estimate $f(\cdot)$ at quantiles $\tilde{q}_{\tau_k}$, and empirically estimate $\tau_{0,k}$.
Step 5 Construct and minimise $\tilde{V}_\beta$ with respect to β at different x.

A simple fact can facilitate our computation to obtain the optimal weights $\hat{\beta}$ in Step 5. We observe that in heterogeneous cases, V_β and the identification constraint only depend on x through $\sigma(\cdot)\beta_0$. Denote $\sigma(\cdot)\beta_0$ by ω. Then, $V(\beta) = 4\omega^2 - 4\sum_{k=1}^{q}\beta_k\omega\tau_{0,k} + \sum_{k,k'=1}^{q}\beta_k\beta_{k'}\tau_{k,k'}$, and the constraint becomes $\omega + \sum_{k=1}^{q}\beta_k f(q_{\tau_k})/2 = 1$. This indicates $(\beta_{1,\text{opt}}, \ldots, \beta_{q,\text{opt}})$ and ω_{opt} are constant. Therefore, instead of carrying out multiple minimisation problems for different x in Step 5, we can solve a single-minimisation problem to get $(\hat{\beta}_1, \ldots, \hat{\beta}_q)$ and $\hat{\omega}$, and then obtain $\hat{\beta}_0 = \hat{\omega}/\tilde{\sigma}(\cdot)$ at different x. Since $\sigma(\cdot)$ is a smooth function, then the weights are smooth functions of x, and so are the resulting estimators.

According to this procedure, we can also see that compared to CQR, the adaptively weighted kernel regression only needs additional cost for the empirical estimations for $\tau_{0,k}$ and one quadratic minimisation, regardless of the heteroscedasticity of the errors. Therefore, the computation costs of the proposed method and CQR are of the same order.

3.5. *Bandwidth selection*

The performance of local polynomial regression estimators depends crucially on the smoothing parameter h. Obtaining a good bandwidth is very important for the success of the adaptively weighted local polynomial regression estimator. Given a weights vector β, the optimal bandwidth in the sense of minimising $\text{MSE}(\hat{m}_\beta(x_0))$ is

$$h_{\beta,\text{opt}}(x_0) = \left[\frac{1}{(m^{(2)}(x_0))^2}\frac{\nu_0\sigma^2(x_0)}{4g(x_0)\mu_2^2}V_\beta\right]^{1/5} n^{-1/5},$$

and the optimal bandwidth for the local linear regression estimator is

$$h_{\text{LS}}(x_0) = \left[\frac{1}{(m^{(2)}(x_0))^2}\frac{\nu_0\sigma^2(x_0)}{g(x_0)\mu_2^2}\right]^{1/5} n^{-1/5}.$$

It follows that

$$h_{\beta,\mathrm{opt}}(x_0) = \left(\frac{V_\beta}{4}\right)^{1/5} h_{\mathrm{LS}}(x_0) \tag{11}$$

As suggested in Kai et al. (2010), when $E[\epsilon_i^2]$ exists we can exploit this simple relationship to select the optimal bandwidth for the proposed estimator using existing bandwidth selectors for the local linear estimator. When $E[\epsilon_i^2]$ does not exist, we can similarly select the bandwidth via the relationship between the proposed estimator and the local LAD linear estimator. In both cases, we can infer $V_{\tilde{\beta}}$ from preliminary estimates. This procedure works well for both homogeneous and heterogenous errors.

4. Numerical studies and applications

In this section, we conduct a simulation study which evaluates the finite sample performance of the adaptively weighted local polynomial regression estimator. We then apply the proposed estimator to a real data set as a demonstration of its practical use.

4.1. *Simulations*

In our simulation studies, we adopt the settings used in Kai et al. (2010) and Sun et al. (2013). We consider two simulation models.

(1) $Y = \sin(2X) + 2\exp(-16X^2) + 0.5\epsilon$, where $X \sim N(0,1)$,
(2) $Y = X\sin(2\pi X) + (\frac{1}{5} + \cos(2\pi X)/10)\epsilon$, where $X \sim \mathrm{Unif}(0,1)$.

In each model, we consider various distributions for ϵ: $N(0,1)$ and $\mathrm{Unif}(-1,1)$ represent light-tailed errors; Laplace$(0,1)$ represents moderate-tailed errors; a t_3-distribution and a truncated Cauchy$(0,1)$ on $[-10,10]$ represent heavy-tailed errors; a mixture of two normal distributions $0.95N(0,1) + 0.05N(0,\sigma^2)$ with $\sigma = 3, 10$ represent errors with light and severe outliers, respectively. For each combination, we simulated 1000 independent random samples, each consisting of 200 observations.

In our simulations, we assume that the correct choice of homoscedasticity or heteroscedasticity has been made for each of the two models. We use the local LS as the pilot fit, since it can be easily implemented in **R**. We also use the local LAD as the pilot estimation and observe similar simulation results. We present those results in the supplemental material. We compare the proposed method with the classical local linear estimator, the local polynomial CQR and WCQR via evaluating the integrated mean squared errors (IMSE), which is a summation of MSE at L equally spaced grid points over the interval at which the regression function is estimated. For model 1, we estimate $m(x)$ over $[-1.5, 1.5]$ with $L = 200$ and for model 2, we estimate $m(x)$ over [0,1] with $L = 200$.

We consider $q = 5, 9, 19$ for the local polynomial CQR, WCQR and the adaptively weighted local polynomial regression estimator. Additional simulations for $q = 1$ were also conducted and are reported in the supplemental material. We use the normal kernel and select h_{LS} via a plug-in bandwidth selector, dpill, proposed by Ruppert, Sheather, and Wand (1995). For the proposed estimator, we select the bandwidth using Equation (11). The bandwidths for CQR and WCQR are calculated using their relationship to LS. We summarise our simulation results using the ratio of the IMSE (RIMSE) of the local linear estimator over the IMSE of the other estimators. The results are presented in Tables 1 and 2, where CQR_5, CQR_9, CQR_{19}, WCQR_5, WCQR_9, WCQR_{19}, AW_5, AW_9 and AW_{19}, denote the local polynomial CQR, WCQR and the adaptively weighted local polynomial regression estimators with $q = 5, 9, 19$ respectively.

Table 1. The means and standard deviations of RIMSE for model 1.

	Mean	Standard deviation		Mean	Standard deviation		Mean	Standard deviation
$N(0,1)$						LS	–	–
CQR_5	0.944	0.102	$WCQR_5$	0.928	0.128	AW_5	**0.984**	0.064
CQR_9	0.965	0.073	$WCQR_9$	0.943	0.119	AW_9	**0.969**	0.088
CQR_{19}	**0.982**	0.048	$WCQR_{19}$	0.931	0.123	AW_{19}	0.945	0.127
Unif(-1,1)						LS	–	–
CQR_5	0.854	0.058	$WCQR_5$	0.979	0.140	AW_5	**1.019**	0.063
CQR_9	0.924	0.044	$WCQR_9$	**1.174**	0.196	AW_9	1.137	0.165
CQR_{19}	0.971	0.036	$WCQR_{19}$	**1.343**	0.265	AW_{19}	1.216	0.269
Laplace(0,1)						LS	–	–
CQR_5	1.125	0.217	$WCQR_5$	1.193	0.314	AW_5	**1.243**	0.327
CQR_9	1.080	0.143	$WCQR_9$	1.167	0.301	AW_9	**1.243**	0.332
CQR_{19}	1.036	0.086	$WCQR_{19}$	1.095	0.274	AW_{19}	**1.214**	0.321
t_3						LS	–	–
CQR_5	1.383	0.704	$WCQR_5$	1.342	0.619	AW_5	**1.427**	0.744
CQR_9	1.276	0.560	$WCQR_9$	1.284	0.520	AW_9	**1.393**	0.741
CQR_{19}	1.125	0.199	$WCQR_{19}$	1.137	0.434	AW_{19}	**1.329**	0.641
Truncated Cauchy on $[-10, 10]$						LS	–	–
CQR_5	1.319	0.326	$WCQR_5$	1.558	0.529	AW_5	**1.716**	0.612
CQR_9	1.158	0.207	$WCQR_9$	1.510	0.551	AW_9	**1.725**	0.623
CQR_{19}	1.070	0.107	$WCQR_{19}$	1.376	0.495	AW_{19}	**1.696**	0.606
$0.95N(0,1)+0.05N(0,9)$						LS	–	–
CQR_5	**1.113**	0.226	$WCQR_5$	1.065	0.216	AW_5	1.088	0.200
CQR_9	**1.089**	0.168	$WCQR_9$	1.043	0.195	AW_9	1.054	0.205
CQR_{19}	**1.047**	0.095	$WCQR_{19}$	0.945	0.187	AW_{19}	0.982	0.189
$0.95N(0,1)+0.05N(0,100)$						LS	–	–
CQR_5	2.551	1.502	$WCQR_5$	2.232	1.261	AW_5	**2.572**	1.554
CQR_9	2.018	1.016	$WCQR_9$	1.807	1.039	AW_9	**2.153**	1.342
CQR_{19}	1.406	0.459	$WCQR_{19}$	1.486	0.909	AW_{19}	**1.860**	1.276

Bold labels the best estimator among CQR, WCQR and AW with the same number of quantiles.

It appears that the proposed method adapts well to the different error distributions. From Table 1, we can see that although slight loss in efficiency relative to LS exists for the adaptively weighted local polynomial regression estimators, when the error distribution is normal, the proposed estimators outperform LS, CQR and WCQR counterparts for most of the distributions considered. The proposed estimator shows significant improvement over LS, CQR in terms of RIMSE especially when the errors are heavy-tailed or large outliers appear.

Comparing the simulation results for $q = 1$ in the supplemental material, we can see that the proposed estimator with more quantiles has superior performance over those simpler estimators for most error distributions. However, in Table 1, it seems that the estimators with 19 quantiles can be less efficient than the estimators with 5 or 9 quantiles for some distributions, this is due to the fact that introducing too many quantiles into WCQR and the proposed estimators requires more parameter estimation and hence impacts the efficiency for a fixed sample size. However, with the incorporation of LS part, the proposed estimator outperforms the WCQR for most cases. Similar results for model 2 can be observed from Table 2. This indicates that the proposed adaptively weighted estimator still performs well in the presence of heteroscedasticity.

4.2. *A real data analysis*

To illustrate its practical use, we apply the adaptively weighted local polynomial regression to the motorcycle data which also has been studied by Schmidt, Mattern, and Schü (1981), Fan and

Table 2. The means and standard deviations of RIMSE for model 2.

	Mean	Standard deviation		Mean	Standard deviation		Mean	Standard deviation
$N(0,1)$						LS	–	–
CQR_5	0.936	0.171	$WCQR_5$	0.927	0.199	AW_5	**0.990**	0.094
CQR_9	0.962	0.142	$WCQR_9$	0.940	0.191	AW_9	**0.969**	0.143
CQR_{19}	**0.979**	0.099	$WCQR_{19}$	0.915	0.217	AW_{19}	0.949	0.171
Unif(−1,1)						LS	–	–
CQR_5	0.812	0.096	$WCQR_5$	0.998	0.273	AW_5	**1.032**	0.123
CQR_9	0.890	0.072	$WCQR_9$	**1.245**	0.478	AW_9	1.147	0.263
CQR_{19}	0.946	0.064	$WCQR_{19}$	**1.328**	0.692	AW_{19}	1.190	0.483
Laplace(0,1)						LS	–	–
CQR_5	1.214	0.392	$WCQR_5$	1.289	0.530	AW_5	**1.371**	0.578
CQR_9	1.146	0.273	$WCQR_9$	1.246	0.501	AW_9	1.355	0.588
CQR_{19}	1.086	0.178	$WCQR_{19}$	1.148	0.442	AW_{19}	1.246	0.510
t_3						LS	–	–
CQR_5	1.443	0.822	$WCQR_5$	1.456	0.880	AW_5	**1.566**	1.023
CQR_9	1.313	0.741	$WCQR_9$	1.411	0.921	AW_9	**1.537**	0.980
CQR_{19}	1.250	0.499	$WCQR_{19}$	1.265	0.729	AW_{19}	**1.469**	0.911
Truncated Cauchy on [−10, 10]						LS	–	–
CQR_5	1.549	0.647	$WCQR_5$	1.860	1.138	AW_5	**2.118**	1.130
CQR_9	1.306	0.367	$WCQR_9$	1.814	0.992	AW_9	**2.146**	1.274
CQR_{19}	1.179	0.238	$WCQR_{19}$	1.650	0.866	AW_{19}	**1.966**	1.233
$0.95N(0,1)+0.05N(0,9)$						LS	–	–
CQR_5	**1.142**	0.426	$WCQR_5$	1.074	0.381	AW_5	1.102	0.346
CQR_9	**1.101**	0.362	$WCQR_9$	1.043	0.330	AW_9	1.075	0.352
CQR_{19}	**1.083**	0.258	$WCQR_{19}$	0.944	0.344	AW_{19}	0.999	0.338
$0.95N(0,1)+0.05N(0,100)$						LS	–	–
CQR_5	3.480	3.683	$WCQR_5$	2.799	2.909	AW_5	**3.517**	3.317
CQR_9	3.014	2.743	$WCQR_9$	2.355	2.236	AW_9	**3.103**	2.818
CQR_{19}	1.785	0.942	$WCQR_{19}$	1.725	1.648	AW_{19}	**2.444**	2.575

Bold labels the best estimator among CQR, WCQR and AW.

Gijbels (1996) and many others. The covariate X is the recorded time (in milliseconds) after a simulated impact on motorcycles, while the response variable Y stands for the head acceleration (in g) of a test object. We use local the LS, CQR, WCQR and the adaptively weighted kernel regression to fit the regression model. Since the sample size is 133, we chose $q = 5$ to avoid introducing too many parameters. For the local CQR, WCQR and the proposed estimator, we use the local LS as the pilot estimate. Since we do not know if the errors are homogeneous, we apply heteroscedastic version of our estimator and implement the procedure described in Section 3.4 to select the practical weights for the proposed estimator. From Figure 1, we can see that all four procedures provide similar fits, and the local linear fit and the AW_5 fit are almost identical. This coincidence can be explained by the normal plot of the residuals of the pilot fit, which is displayed in Figure 2. We can see that the residuals roughly follow a normal distribution, where LS is more likely to get the best fit. However, the proposed method does as well as the LS. This indicates that the proposed procedure actually adapts to the unknown error structure.

Although there is no outlier in the data set, we artificially create 2 to examine robustness properties, we move the 102th observation from −54.9 to −109.8 and the 112th observation from −21.5 to −86. In Figure 3, we only depict the LS and AW_5, because the CQR and WCQR are demonstrated to be resistant against outliers. We use the local LS fit from the original data set as the baseline to see what impact the introduced outliers bring to those fits. We note that

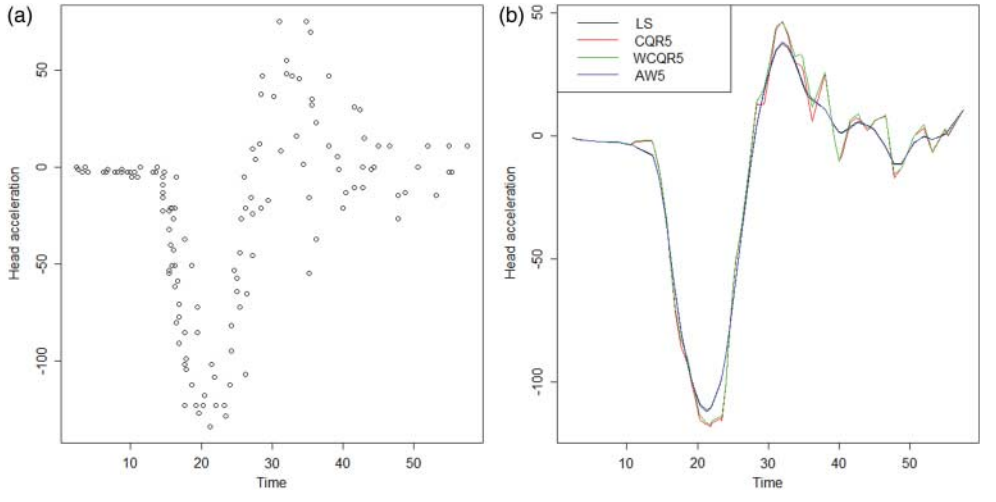

Figure 1. Scatter plot and four fittings.

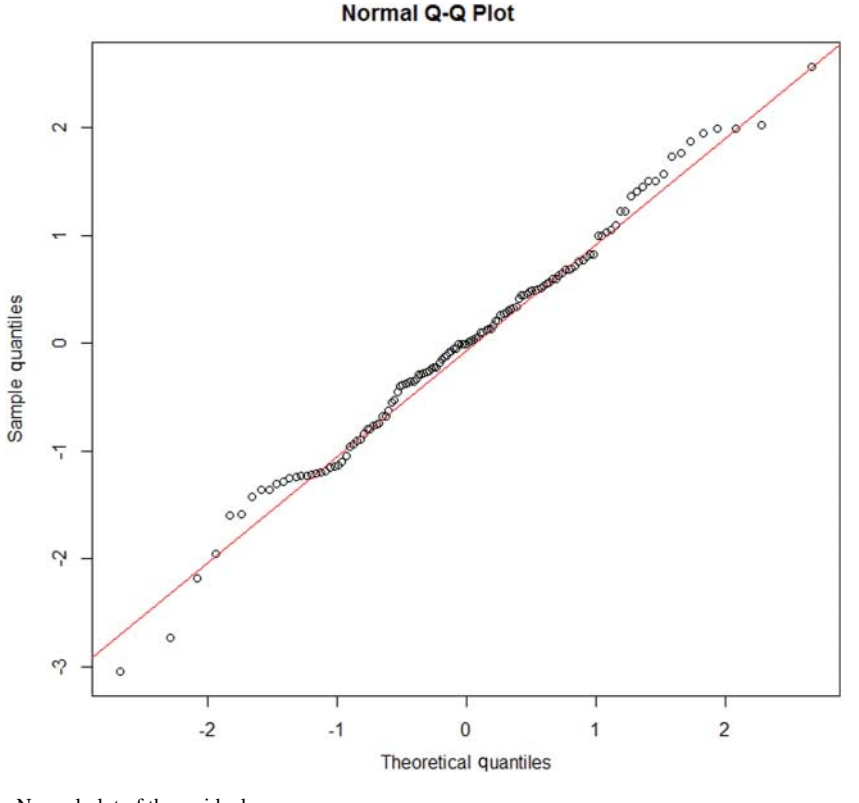

Figure 2. Normal plot of the residuals.

the AW_5 still maintains a similar pattern, while the local linear estimator starts to deviate from the baseline. This demonstrates that the proposed estimator inherits the resistance property from quantile regressions, and enjoys the favoured robustness.

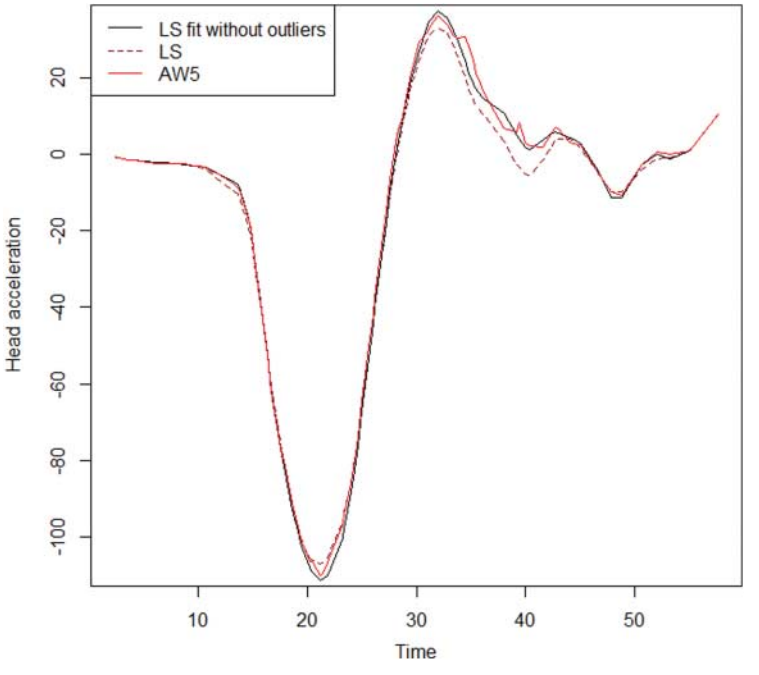

Figure 3. Estimated regression function with introduced outliers.

5. Conclusion

In this paper, we combine the strength of the LS and quantile regression to propose the adaptively weighted local polynomial estimator for nonparametric regression. The novelty of the method is that it adapts to the distribution of the error terms in a regression model. We have explicitly described how data can be used to select weights as well as the bandwidth parameter. It appears that even when the weights are selected from the data, the estimators perform nearly as well as the optimal choice. For example, if the distribution is normal, the method is nearly as efficient as LS, but the method still works well if the errors follow a t-distribution with three degrees of freedom. The estimators compete favourably with equally WCQR. The idea of weighting different objective functions and using the asymptotic efficiency to select the optimal weights can be extended to other situations.

References

Bickel, P.J. (1973), 'On Some Analogues to Linear Combinations of Order Statistics in the Linear Model', *The Annals of Statistics*, 1, 597–616.

Bradic, J., Fan, J., and Wang, W. (2011), 'Penalized Composite Quasi-Likelihood for Ultrahigh-Dimensional Variable Selection', *Journal of the Royal Statistical Society Series B*, 73, Part 2, 325–349.

Chan, S., and Zhang, Z. (2004), 'Robust Local Polynomial Regression Using M-Estimator With Adaptive Bandwidth', in *Proceedings of the International Symposium on Circuits and Systems, 2004*, ISCAS '04, Vancouver, Vol.3.

Chernozhukov, V. (2005), 'Extremal Quantile Regression', *The Annals of Statistics*, 33, 806–839.

Fan, J. (1992), 'Design-Adaptive Nonparametric Regression', *Journal of the American Statistical Association*, 87, 998–1004.

Fan, J., and Gijbels, I. (1992), 'Variable Bandwidth and Local Linear Regression Smoothers', *The Annals of Statistics*, 20, 2008–2036.

Fan, J., and Gijbels, I. (1996), *Local Polynomial Modelling and its Applications*, London: Chapman and Hall.

Fan, J., Hu, T.C., and Truong, Y.K. (1994), 'Robust Non-Parametric Function Estimation', *Scandinavian Journal of Statistics*, 21, 433–446.

Feller, W. (1971), *An Introduction to Probability Theory and its Applications* (Vol. 2), New York: Wiley.
Jiang, J and Mack, Y. P. (2001), 'Robust Local Polynomial Regression for Dependent Data', *Statistica Sinica*, 11, 705–722.
Kai, B., Li, R., and Zou, H. (2010), 'Local Composite Quantile Regression Smoothing: An Efficient and Safe Alternative to Local Polynomial Regression', *Journal of the Royal Statistical Society Series B*, 71, 49–69.
Knight, K. (1998), 'Limiting Distributions for L_1 Regression Estimators Under General Conditions', *The Annals of Statistics*, 26, 755–770.
Koenker, R. (1984), 'A Note on L-estimates for Linear Models', *Statistics and Probability Letters*, 2, 323–325.
Koenker, R. (2005), *Regression Quantiles*, Cambridge: Cambridge University Press.
Koenker, R., and Bassett, G. (1978), 'Regression Quantiles', *Econometrica*, 46, 33–50.
Koenker, R., and Portnoy, S. (1987), 'L-Estimation for Linear Models', *Journal of the American Statistical Association*, 82, 851–857.
Portnoy, S., and Koenker, R. (1989), 'Adaptive L-Estimation for Linear Models', *The Annals of Statistics*, 17, 362–381.
Portnoy, S., and Koenker, R. (1997), 'The Gaussian Hare and The Laplacian Tortoise: Computability of Square-Error Versus Absolute-Error Estimators', *Statistical Science*, 12, 279–300.
Ruppert, D., Sheather, S., and Wand, M.P. (1995), 'An Effective Bandwidth Selector for Local Least Squares Regression', *Journal of the American Statistical Association*, 90, 1257–1270.
Ruppert, D., Wand, M.P., Holst, U., and Hössjer, O. (1997), 'Local Polynomial Variance-Function Estimation', *Technometrics*, 39, 262–273.
Schmidt, G., Mattern, R., and Schü, 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', Final Report Phase III, Project 65, Institue für Rechtsmedizin, Universität Heidelberg, Germany.
Sun, J., Gai, Y., and Lin, L. (2013), 'Weighted Local Linear Composite Quantile Estimation for the Case of General Error Distributions', *Journal of Statistical Planning and Inference*, 143, 1049–1063.
Wahba, G. (1990), *Spline Models for Observational Data*, Philadelphia: SIAM.
Watson, G.S. (1964), 'Smooth Regression Analysis', *Sankhyā: The Indian Journal of Statistics, Series A*, 26, 359–372.
Welsh, A.H. (1996), 'Robust Estimation of Smooth Regression and Spread Functions and their Derivatives', *Statistica Sinica*, 6, 347–366.
Yu, K., and Jones, M.C. (1998), 'Local Linear Quantile Regression', *Journal of the American Statistical Association*, 93, 228–237.
Yu, K., Liu, Z., and Stander, J. (2003), 'Quantile Regression: Applications and Current Research Areas', *Journal of the Royal Statistical Society Series D,* 52, 331–350.
Zou, H., and Yuan, M. (2008), 'Composite Quantile Regression and The Oracle Model Selection Theory', *The Annals of Statistics*, 36, 1108–1126.

Appendix 1

Proof of Theorem 3.1 We sketch the proofs in this section. Detailed proofs are provided in the supplemental material. Let

$$Q_\beta(\theta) = \sum_{i=1}^{n} \left[\sum_{k=0}^{q} \beta_k \rho_{\tau_k} \left(y_i - a_{0k} - \sum_{j=0}^{p} \frac{1}{j!} a_j (x_i - x_0)^j \right) \right] K\left(\frac{x_i - x_0}{h} \right).$$

We can see that minimising $Q_\beta(\theta)$ with respect to θ is equivalent to minimising $Q_\beta(\theta^* + (nh_n)^{-1/2}u) - Q_\beta(\theta^*)$ with respect to u, where $u = (u_{01}, \ldots, u_{0q}, u_0, u_1, \ldots, u_p)^{\mathrm{T}}$ is a $q + 1 + p$ vector.

Let $\Delta_{0,i} = \sum_{j=0}^{p} (1/j!) u_j (x_i - x_0)^j$, and $\Delta_{k,i} = \Delta_{0,i} + u_{0k}$, for $k = 1, \ldots, q$. Applying the identity (Knight 1998),

$$\rho_\tau(x - y) - \rho_\tau(x) = y[1(x \le 0) - \tau] + \int_0^y \{1(x \le z) - 1(x \le 0)\} \, \mathrm{d}z,$$

yields

$$\begin{aligned}
&= \frac{1}{\sqrt{nh_n}} \sum_{i=1}^{n} K\left(\frac{x_i - x_0}{h_n} \right) \left\{ -2\beta_0 \Delta_{0,i} (\sigma \epsilon_i + r_{i,p}) + \sum_{k=1}^{q} \beta_k \left[1\left(\epsilon_i \le \frac{\sigma q_{\tau_k} - r_{i,p}}{\sigma} \right) - \tau_k \right] \Delta_{k,i} \right\} \\
&\quad + \frac{\beta_0}{nh_n} \sum_{i=1}^{n} K\left(\frac{x_i - x_0}{h_n} \right) \Delta_{0,i}^2 \\
&\quad + \sum_{i=1}^{n} K\left(\frac{x_i - x_0}{h_n} \right) \sum_{k=1}^{q} \beta_k \int_0^{\Delta_{k,i}/\sqrt{nh_n}} \left\{ 1\left(\epsilon_i \le \frac{\sigma q_{\tau_k} - r_{i,p} + z}{\sigma} \right) - 1\left(\epsilon_i \le \frac{\sigma q_{\tau_k} - r_{i,p}}{\sigma} \right) \right\} \mathrm{d}z \\
&:= I_1 + I_2 + I_3.
\end{aligned}$$

By some algebra, we can show that $\mathrm{Var}[I_3] \to 0$. Applying Chebyshev's inequality yields $I_3 - E[I_3] \xrightarrow{p} 0$ and

$$E[I_3] = n\sum_{k=1}^{q}\beta_k E\left[Z_{n,k,i}(u)1\left(\left|\frac{\Delta_{k,i}}{\sqrt{nh_n}}\right| \le \eta\right)\right] + n\sum_{k=1}^{q}\beta_k E\left[Z_{n,k,i}(u)1\left(\left|\frac{\Delta_{k,i}}{\sqrt{nh_n}}\right| > \eta\right)\right]. \tag{A1}$$

where $Z_{n,k,i}(u) = K((x_i - x_0)/h_n)\int_0^{\Delta_{k,i}/\sqrt{nh_n}}\{1(\epsilon_i \le (\sigma q_{\tau_k} - r_{i,p} + z)/\sigma) - 1(\epsilon_i \le (\sigma q_{\tau_k} - r_{i,p})/\sigma)\}\,\mathrm{d}z$. Using the same argument as in the proof of $E[I_3^2] \to 0$, we can show that

$$\begin{aligned} &n\sum_{k=1}^{q}\beta_k E\left[Z_{n,k,i}(u)1\left(\left|\frac{\Delta_{k,i}}{\sqrt{nh_n}}\right| \le \eta\right)\right] \\ &\quad = \frac{1}{2}\sum_{k=1}^{q}\beta_k f(q_{\tau_k})g(x_0)\int_{-M}^{M}K(t_i)\left(u_{0k} + \sum_{j=0}^{p}\frac{u_j}{j!}h_n^j t_i^j\right)^2 \mathrm{d}t_i + o(\|A_{h_n}u\|^2), \end{aligned} \tag{A2}$$

where $t_i = (x_i - x_0)/h_n$.

Applying the Cauchy–Schwartz inequality, we have

$$n\sum_{k=1}^{p}\beta_k E\left[Z_{n,k,i}(u)1\left(\left|\frac{\Delta_{k,i}}{\sqrt{nh_n}}\right| > \eta\right)\right] = o(\|A_{h_n}u\|^2) = o(1). \tag{A3}$$

Therefore,

$$I_3 \xrightarrow{p} \frac{1}{2}\sum_{k=1}^{p}\beta_k f(q_{\tau_k})g(x_0)\int_{-M}^{M}K(t_i)\left[u_{0k} + \sum_{j=0}^{p}\frac{u_j}{j!}h_n^j t_i^j\right]^2 \mathrm{d}t_i. \tag{A4}$$

According to the law of large numbers, we know

$$I_2 \xrightarrow{a.s.} \beta_0 g(x_0)\int_{-M}^{M}K(t_i)\left(\sum_{j=0}^{p}\frac{u_j}{j!}h_n^j t_i^j\right)^2 \mathrm{d}t_i. \tag{A5}$$

Since I_1 can be written as $W_{\beta,n}^{\mathrm{T}}A_{h_n}u$, then

$$Q_\beta(\theta^* + (nh_n)^{-1/2}u) - Q_\beta(\theta^*) = g(x_0)u^{\mathrm{T}}A_{h_n}S(\beta)A_{h_n}u + W_{\beta,n}^{\mathrm{T}}A_{h_n}u + o_p(\|A_{h_n}u\|).$$

Let $\hat{u}$ denote the minimiser of $Q(\theta^* + (nh_n)^{-1/2}u) - Q(\theta^*)$, we have

$$S(\beta)A_{h_n}\hat{u} = -\frac{1}{2g(x_0)}W_{\beta,n} + o_p(1).$$

According to the definition of $W_{\beta,n}$, applying Central Limit Theorem (CLT) yields

$$\frac{\alpha^{\mathrm{T}}W_{\beta,n} - E[\alpha^{\mathrm{T}}W_{\beta,n}]}{\sqrt{\mathrm{Var}[\alpha^{\mathrm{T}}W_{\beta,n}]}} \xrightarrow{L} N(0,1),$$

for any nonzero $(1+q+p)\times 1$ vector α. Thus, The Cramer–Wald device provides us

$$[\mathrm{Cov}(W_{\beta,n})]^{-1/2}(W_{\beta,n} - E[W_{\beta,n}]) \xrightarrow{L} N(0, I_{(1+q+p)\times(1+q+p)}),$$

where $\mathrm{Cov}(W_{\beta,n})$ is the covariance matrix of $W_{\beta,n}$. It is easy to check

$$\mathrm{Cov}(W_{\beta,n})] \xrightarrow{p} g(x_0)\Sigma(\beta).$$

Therefore, we have

$$S(\beta)A_{h_n}\hat{u} + \frac{1}{2g(x_0)}E[W_{\beta,n}] \xrightarrow{L} N(0, \frac{1}{4g(x_0)}\Sigma(\beta)). \tag{A6}$$

This completes the proof of Theorem 3.1. ■

Corollaries 3.1 and 3.2 are special cases of Theorem 3.1. We omit the proofs here. Complete proofs can be seen in the supplemental material.

Proof of Corollary 3.3 Consider $Q_{\hat{\beta}}(\theta^* + (nh_n)^{-1/2}u) - Q_{\hat{\beta}}(\theta^*)$. We can write it as

$$Q_\gamma(\theta^* + (nh_n)^{-1/2}u) - Q_\gamma(\theta^*) + Q_{\beta_{\text{opt}}}(\theta^* + (nh_n)^{-1/2}u) - Q_{\beta_{\text{opt}}}(\theta^*),$$

where $\gamma = \hat{\beta} - \beta_{\text{opt}}$. Since β_{opt} is a fixed vector given the error structure, then using the same arguments as in Theorem 3.1, we obtain that

$$Q_{\beta_{\text{opt}}}(\theta^* + (nh_n)^{-1/2}u) - Q_{\beta_{\text{opt}}}(\theta^*) = g(x_0)u^{\mathrm{T}}A_{h_n}S(\beta_{\text{opt}})A_{h_n}u + W^{\mathrm{T}}_{\beta_{\text{opt}},n}A_{h_n}u + o_p(\|A_{h_n}u\|^2).$$

And we have

$$\begin{aligned}
&Q_\gamma(\theta^* + (nh_n)^{-1/2}u) - Q_\gamma(\theta^*) \\
&= \frac{1}{\sqrt{nh_n}}\sum_{i=1}^{n} K\left(\frac{x_i - x_0}{h_n}\right)\left\{-2\gamma_0(\sigma\epsilon_i + r_{i,p})\Delta_{0,i} + \sum_{k=1}^{q}\gamma_k\left[1\left(\epsilon_i \le \frac{\sigma q_{\tau_k} - r_{i,p}}{\sigma}\right) - \tau_k\right]\Delta_{k,i}\right\} \\
&\quad + \frac{\gamma_0}{nh_n}\Delta_{0,i}^2\sum_{i=1}^{n} K\left(\frac{x_i - x_0}{h_n}\right) \\
&\quad + \sum_{i=1}^{n} K\left(\frac{x_i - x_0}{h_n}\right)\sum_{k=1}^{q}\gamma_k\int_0^{\Delta_{k,i}/\sqrt{nh_n}}\left\{1\left(\epsilon_i \le \frac{\sigma q_{\tau_k} - r_{i,p} + z}{\sigma}\right) - 1\left(\epsilon_i \le \frac{\sigma q_{\tau_k} - r_{i,p}}{\sigma}\right)\right\}\mathrm{d}z. \qquad \text{(A7)}
\end{aligned}$$

Since

$$\begin{aligned}
&\sum_{i=1}^{n} K\left(\frac{x_i - x_0}{h_n}\right)\int_0^{\Delta_{k,i}/\sqrt{nh_n}}\left\{1\left(\epsilon_i \le \frac{\sigma q_{\tau_k} - r_{i,p} + z}{\sigma}\right) - 1\left(\epsilon_i \le \frac{\sigma q_{\tau_k} - r_{i,p}}{\sigma}\right)\right\}\mathrm{d}z \\
&\quad \xrightarrow{p} \frac{1}{2\sigma}f(q_{\tau_k})g(x_0)\int_{-M}^{M} K(t_i)\left[u_{0k} + \sum_{j=0}^{p}\frac{u_j}{j!}h_n^j t_i^j\right]^2\mathrm{d}t_i,
\end{aligned}$$

and

$$\frac{1}{nh_n}\sum_{i=1}^{n} K\left(\frac{x_i - x_0}{h_n}\right)\Delta_{0,i}^2 \xrightarrow{a.s.} g(x_0)\int_{-M}^{M} K(t_i)\left[\sum_{j=0}^{p}\frac{u_j}{j!}h_n^j t_i^j\right]^2\mathrm{d}t_i,$$

then the last two terms of Equation (A7) are $o(\|A_{h_n}u\|^2)$. Applying Slutsky's theorem, we can show that the first term of Equation (A7) is $o_p(A_{h_n}u)$. Therefore,

$$Q_{\hat{\beta}}(\theta^* + (nh_n)^{-1/2}u) - Q_{\hat{\beta}}(\theta^*) = g(x_0)u^{\mathrm{T}}A_{h_n}S(\beta_{\text{opt}})A_{h_n}u + W^{\mathrm{T}}_{\beta_{\text{opt}},n}A_{h_n}u + o_p(\|A_{h_n}u\|^2).$$

This completes the proof of Corollary 3.4. ■

In order to prove Theorem 3.2, the following lemmas are needed.

LEMMA A.1 *If assumptions B–E are satisfied, and* $E[\epsilon_i^2]$ *does not exist, then*

(a)

$$\frac{\sum_{i=1}^{n}|\epsilon|}{a_n\sqrt{n}} \to 0,$$

(b)

$$\frac{\sum_{i=1}^{n}\epsilon_i^2}{a_n^2} \to O_p(1),$$

(c)

$$\frac{\sum_{i=1}^{n}(1/\sqrt{h_n})K((X_i - x_0)/h_n)\epsilon_i}{a_n} \le O_p(1),$$

where $\{a_n\}$ *is the norming sequence for* ϵ_i*, such that* $\sum_{i=1}^{n}\epsilon_i/a_n \xrightarrow{L} S$.

LEMMA A.2 *Let* $\hat{\beta}_{0p} = (1/\tilde{\sigma})((1+4\tilde{f}(0)\tilde{\tau}_{0,l})/(4\tilde{f}^2(0)+8\tilde{f}(0)\tilde{\tau}_{0,l}+1)1(((1+4\tilde{f}(0)\tilde{\tau}_{0,l})/(4\tilde{f}^2(0)+8\tilde{f}(0)\tilde{\tau}_{0,l}+1)\geq 0)$, *where* $\tau_l = \frac{1}{2}$. *Then,* $\hat{\beta}_0 \leq \hat{\beta}_{0p}$ *almost surely.*

We omit the proofs here. They can be found in the supplementary material.

Proof of Theorem 3.2 Let $\tilde{\zeta}_i = \sigma\tilde{\epsilon}_i, i = 1,\ldots,n$ denote the residuals of a $\sqrt{nh_n}$-consistent fit. Since σ and ϵ_i are unknown, we use $\tilde{\sigma}^2 = \sum_{i=1}^n \tilde{\zeta}_i^2/n$ to denote the sample variance of $\sigma\epsilon_i$. Then, we have $\tilde{\tau}_{0,l} = -\sum_{i=1}^n |\tilde{\zeta}_i|/(2n\tilde{\sigma})$ and $\tilde{f}(0) = (1/2nb)\sum_{i=1}^n \tilde{\sigma}\,1(|\sigma\tilde{\epsilon}_i| < b)$, for some $b \sim O(n^{-1/5})$.

The key of the proof is to show that the impact from the LS part is negligible. Since $\hat{\beta}_0 \xrightarrow{p} \beta_{0\text{opt}} = 0$, we need to show

$$\tilde{\sigma}\frac{1}{\sqrt{nh_n}}\sum_{i=1}^n K\left(\frac{x_i - x_0}{h_n}\right)\hat{\beta}_0\sigma\epsilon_i = o_p(1), \tag{A8}$$

$$\tilde{\sigma}\frac{1}{\sqrt{nh_n}}\sum_{k=1}^q \hat{\beta}_k \sum_{i=1}^n K\left(\frac{x_i - x_0}{h_n}\right)[1(\epsilon_i < q_{\tau_k}) - \tau_k] = O_p(1). \tag{A9}$$

According to Lemma A.2, if we can show that $\tilde{\sigma}(1/\sqrt{nh_n})\sum_{i=1}^n K((x_i - x_0)/h_n)\hat{\beta}_{0p}\epsilon_i = o_p(1)$, then Equation (A8) can be directly inferred.

$$\begin{aligned}&\tilde{\sigma}\frac{1}{\sqrt{nh_n}}\sum_{i=1}^n K\left(\frac{x_i - x_0}{h_n}\right)\hat{\beta}_{0p}\sigma\epsilon_i \\ &\quad= \frac{1+4\tilde{f}(0)\tilde{\tau}_{0,l}}{4\tilde{f}^2(0)+8\tilde{f}(0)\tilde{\tau}_{0,l}+1}1\left(\frac{1+4\tilde{f}(0)\tilde{\tau}_{0,l}}{4\tilde{f}^2(0)+8\tilde{f}(0)\tilde{\tau}_{0,l}+1}\geq 0\right)\frac{a_n}{\sqrt{n}}\frac{1}{a_n\sqrt{h_n}}\sum_{i=1}^n K\left(\frac{x_i - x_0}{h_n}\right)\sigma\epsilon_i.\end{aligned}$$

Since $E[\epsilon_i^2]$ does not exist, by Lemma A.2, we have

$$\frac{1}{a_n\sqrt{h_n}}\sum_{i=1}^n K\left(\frac{x_i - x_0}{h_n}\right), \quad \sigma\epsilon_i \leq O_p(1)$$

and

$$\begin{aligned}&\frac{1+4\tilde{f}(0)\tilde{\tau}_{0,l}}{4\tilde{f}^2(0)+8\tilde{f}(0)\tilde{\tau}_{0,l}+1}\frac{a_n}{\sqrt{n}} \\ &\quad= \frac{1-(4\sum_{i=1}^n \tilde{\sigma}\,1(|\sigma\tilde{\epsilon}_i| < b)/2nb)(\sum_{i=1}^n |\sigma\tilde{\epsilon}_i|/2n\tilde{\sigma})}{4(\sum_{i=1}^n 1(|\sigma\tilde{\epsilon}_i| < b)/2nb)^2\tilde{\sigma}^2 - (8\sum_{i=1}^n \tilde{\sigma}\,1(|\sigma\tilde{\epsilon}_i| < b)/2nb)(\sum_{i=1}^n |\sigma\tilde{\epsilon}_i|/2n\tilde{\sigma}) + 1}\frac{a_n}{\sqrt{n}} \\ &\quad= \frac{1-(2\sum_{i=1}^n 1(|\sigma\tilde{\epsilon}_i| < b)/2nb)(\sum_{i=1}^n |\sigma\tilde{\epsilon}_i|/a_n\sqrt{n})(a_n/\sqrt{n})}{4(\sum_{i=1}^n 1(|\sigma\tilde{\epsilon}_i| < b)/2nb)^2(\tilde{\epsilon}_i^2/a_n^2)(a_n^2/n) - (4\sum_{i=1}^n 1(|\sigma\tilde{\epsilon}_i| < b)/2nb)(\sum_{i=1}^n |\sigma\tilde{\epsilon}_i|/a_n\sqrt{n})(a_n/\sqrt{n}) + 1}\frac{a_n}{\sqrt{n}} \\ &\quad= \frac{(\sqrt{n}/a_n) - (2\sum_{i=1}^n 1(|\sigma\tilde{\epsilon}_i| < b)/2nb)(\sum_{i=1}^n |\sigma\tilde{\epsilon}_i|/a_n\sqrt{n})}{4(\sum_{i=1}^n 1(|\sigma\tilde{\epsilon}_i| < b)/2nb)^2(\tilde{\epsilon}_i^2/a_n^2) - (4\sum_{i=1}^n 1(|\sigma\tilde{\epsilon}_i| < b)/2nb)(\sum_{i=1}^n |\sigma\tilde{\epsilon}_i|/a_n\sqrt{n})(\sqrt{n}/a_n) + (n/a_n^2)} \\ &\quad\xrightarrow{p} 0.\end{aligned}$$

Consequently,

$$\tilde{\sigma}\frac{1}{\sqrt{nh_n}}\sum_{i=1}^n K\left(\frac{x_i - x_0}{h_n}\right)\hat{\beta}_{0p}\epsilon_i = o_p(1).$$

From the above proof, we can see that as $n \to \infty$,

$$\sum_{k=1}^q \frac{\tilde{f}(\tilde{q}_{\tau_k})}{\tilde{\sigma}}\tilde{\sigma}\hat{\beta}_k = \sum_{k=1}^q \tilde{f}(\tilde{q}_{\tau_k})\hat{\beta}_k = 1 - \tilde{\sigma}\hat{\beta}_0 \to 1.$$

Since $\tilde{f}(\tilde{q}_{\tau_k})/\tilde{\sigma}$ is bounded for $1 \leq k \leq q$, then $\sum_{k=1}^q \tilde{\sigma}\hat{\beta}_k$ is bounded away from 0. There, Equation (A9) can be inferred. This completes the proof of Theorem 3.2. ■

The proof of Theorem 3.3 is essentially the same as for Theorem 3.2. Thus, we omit it here.