Journal of Machine Learning Research 20 (2019) 1-49 Submitted 6/17; Revised 2/19; Published 5/19

Robust Estimation of Derivatives Using Locally Weighted Least Absolute Deviation Regression

WenWu Wang WENGEWSH@SINA.COM
School of Statistics
Qufu Normal University
Jingxuan West Road, Qufu, Shandong, China
Ping Yu PINGYU@HKU.HK
Faculty of Business and Economics
University of Hong Kong
Pokfulam Road, Hong Kong
Lu Lin LINLU@SDU.EDU.CN
Zhongtai Securities Institute for Financial Studies
Shandong University
Jinan, Shandong, China
Tiejun Tong TONGT@HKBU.EDU.HK
Department of Mathematics
Hong Kong Baptist University
Kowloon Tong, Hong Kong

Editor: Zhihua Zhang

Abstract

In nonparametric regression, the derivative estimation has attracted much attention in recent years due to its wide applications. In this paper, we propose a new method for the derivative estimation using the locally weighted least absolute deviation regression. Different from the local polynomial regression, the proposed method does not require a finite variance for the error term and so is robust to the presence of heavy-tailed errors. Meanwhile, it does not require a zero median or a positive density at zero for the error term in comparison with the local median regression. We further show that the proposed estimator with random difference is asymptotically equivalent to the (infinitely) composite quantile regression estimator. In other words, running one regression is equivalent to combining infinitely many quantile regressions. In addition, the proposed method is also extended to estimate the derivatives at the boundaries and to estimate higher-order derivatives. For the equidistant design, we derive theoretical results for the proposed estimators, including the asymptotic bias and variance, consistency, and asymptotic normality. Finally, we conduct simulation studies to demonstrate that the proposed method has better performance than the existing methods in the presence of outliers and heavy-tailed errors, and analyze the Chinese house price data for the past ten years to illustrate the usefulness of the proposed method.

Keywords: composite quantile regression, differenced method, LowLAD, LowLSR, outlier and heavy-tailed error, robust nonparametric derivative estimation

1. Introduction

The derivative estimation is an important problem in nonparametric regression and it has applications in a wide range of fields. For instance, when analyzing human growth data (Müller, 1988; Ramsay and Silverman, 2002) or maneuvering target tracking data (Li and Jilkov, 2003, 2010), the first- and second-order derivatives of the height as a function of time are two important parameters, with the first-order derivative representing the speed and the second-order derivative representing the acceleration. The derivative estimates are also needed in change-point problems, e.g., for exploring the structures of curves (Chaudhuri and Marron, 1999; Gijbels and Goderniaux, 2005), for detecting the extremum of derivatives (Newell et al., 2005), for characterizing submicroscopic nanoparticle from scattering data (Charnigo et al., 2007, 2011a), for comparing regression curves (Park and Kang, 2008), for detecting abrupt climate changes (Matyasovszky, 2011), and for inferring the cell growth rates (Swain et al., 2016).

In the existing literature, one usually obtains the derivative estimates as a by-product by taking the derivative of a nonparametric fit of the regression function. There are three main approaches for the derivative estimation: smoothing spline, local polynomial regression, and differenced estimation. For smoothing spline, the derivatives are estimated by taking derivatives of the spline estimation of the regression function (Stone, 1985; Zhou and Wolfe, 2000). For local polynomial regression, a polynomial using the Taylor expansion is fitted locally by the kernel method (Ruppert and Wand, 1994; Fan and Gijbels, 1996; Delecroix and Rosa, 1996). These two methods both require an estimate of the regression function. As pointed out in Wang and Lin (2015), when the regression function estimator achieves the optimal rate of convergence, the corresponding derivative estimators may fail to achieve the rate. In other words, minimizing the mean square error of the regression function estimator does not necessarily guarantee the derivatives be optimally estimated (Wahba and Wang, 1990; Charnigo et al., 2011b).

For the differenced estimation, Müller et al. (1987) and Härdle (1990) proposed a cross-validation method to estimate the first-order derivative without estimating the regression function. Unfortunately, their method may not perform well in practice as the variance of their estimator is proportional to n^2 when the design points are equally spaced. Observing this shortcoming, Charnigo et al. (2011b) and De Brabanter et al. (2013) proposed a variance-reducing estimator for the derivative function called the empirical derivative that is essentially a linear combination of the symmetric difference quotients. They further derived the order of the asymptotic bias and variance, and established the consistency of the empirical derivative. Wang and Lin (2015) represented the empirical derivative as a local constant estimator in locally weighted least squares regression (LowLSR), and proposed a new estimator for the derivative function to reduce the estimation bias in both valleys and peaks of the true derivative function. More recently, Dai et al. (2016) generalized equidistant design to non-equidistant design, and Liu and De Brabanter (2018) further generalized the existing work to random design.

The aforementioned differenced derivative estimators are all based on the least squares (LS) method. Although elegant, the least squares method is not robust to outliers (Huber and Ronchetti, 2009). To overcome this problem, various robust methods have been proposed in the literature to improve the estimation of the regression function, see, for ex-

ample, kernel M-smoother (Härdle and Gasser, 1984), local least absolute deviation (LAD) (Fan and Hall, 1994; Wang and Scott, 1994), and locally weighted least squares (Cleveland, 1979; Ruppert and Wand, 1994) among others. In contrast, little attention has been paid to improving the derivative estimation except for the parallel developments of the above remedies (Härdle and Gasser, 1985; Welsh, 1996; Boente and Rodriguez, 2006), so call for a better solution.

In this paper, we propose a locally weighted least absolute deviation (LowLAD) method by combining the differenced method and the L_1 regression systematically. Over a neighborhood centered at a fixed point, we first obtain a sequence of linear regression representation in which the derivative is the intercept term. We then estimate the derivative by minimizing the sum of weighted absolute errors. By repeating this local fitting over a grid of points, we can obtain the derivative estimates on a discrete set of points. Finally, the entire derivative function is obtained by applying the local polynomial regression or the cubic spline interpolation.

The rest of the paper is organized as follows. Section 2 presents the motivation, the first-order derivative estimator and its theoretical properties, including the asymptotic bias and variance, consistency, and asymptotic normality. Section 3 studies the relation between the LowLAD estimator and the existing estimators. In particular, we show that the LowLAD estimator with random difference is asymptotically equivalent to the (infinitely) composite quantile regression estimator. Section 4 derives the first-order derivative estimation at the boundaries of the domain, and Section 5 generalizes the proposed method to estimate the higher-order derivatives. In Section 6 we conduct extensive simulation studies to assess the finite-sample performance of the proposed estimators and compare them with the existing competitors; we also apply our method to a real data set to illustrate its usefulness in practice. Finally, we conclude the paper with some discussions in Section 7, and provide the proofs of the theoretical results in six Appendices.

A word on notation: $\doteq$ means that the higher-order terms are omitted, and $\approx$ means an approximate result with up to two decimal digits.

2. First-Order Derivative Estimation

Combining the differenced method and the L_1 regression, we propose the LowLAD regression to estimate the first-order derivative. The new method inherits the advantage of the differenced method and also the robustness of the L_1 method.

2.1. Motivation

Consider the nonparametric regression model

$$Y_i = m(x_i) + \epsilon_i, \quad 1 \le i \le n, \tag{1}$$

where $x_i = i/n$ is the design point, Y_i is the response variable, $m(\cdot)$ is an unknown regression function, and ϵ_i are independent and identically distributed (iid) random errors with a continuous density $f(\cdot)$.

We first define first-order symmetric (about i) difference quotient (Charnigo et al., 2011b; De Brabanter et al., 2013) as

$$Y_{ij}^{(1)} = \frac{Y_{i+j} - Y_{i-j}}{x_{i+j} - x_{i-j}}, \quad 1 \le j \le k, \tag{2}$$

where k is a positive integer, and then decompose $Y_{ij}^{(1)}$ into two parts as

$$Y_{ij}^{(1)} = \frac{m(x_{i+j}) - m(x_{i-j})}{2j/n} + \frac{\epsilon_{i+j} - \epsilon_{i-j}}{2j/n}, \quad 1 \le j \le k. \tag{3}$$

On the right hand side of (3), the first term contains the bias information of the true derivative, and the second term contains the variance information. By Wang and Lin (2015), the first-order derivative estimation based on the third-order Taylor expansion usually outperforms the estimation based on the first-order Taylor expansion due to bias correction. For the same reason, we assume that $m(\cdot)$ is three times continuously differentiable on $[0, 1]$. By the Taylor expansion, we obtain

$$\frac{m(x_{i+j}) - m(x_{i-j})}{2j/n} = m^{(1)}(x_i) + \frac{m^{(3)}(x_i)}{6}\frac{j^2}{n^2} + o\left(\frac{j^2}{n^2}\right), \tag{4}$$

where the estimation bias is contained in the remainder term of the Taylor expansion.

By (3) and (4), we have

$$Y_{ij}^{(1)} = m^{(1)}(x_i) + \frac{m^{(3)}(x_i)}{6}\frac{j^2}{n^2} + \frac{\epsilon_{i+j} - \epsilon_{i-j}}{2j/n} + o\left(\frac{j^2}{n^2}\right). \tag{5}$$

In Proposition 11 (see Appendix A), we show that the median of $\epsilon_{i+j} - \epsilon_{i-j}$ is always zero, no matter whether the median of ϵ_i is zero or not. As a result, for any fixed $k = o(n)$, we have

$$\text{Median}[Y_{ij}^{(1)}] = m^{(1)}(x_i) + \frac{m^{(3)}(x_i)}{6} d_j^2 + o\left(d_j^2\right), \quad 1 \le j \le k, \tag{6}$$

where $d_j = j/n$. We treat (6) as a linear regression with d_j^2 and $Y_{ij}^{(1)}$ as the independent and dependent variables, respectively. In the presence of heavy-tailed errors, we propose to estimate $m^{(1)}(x_i)$ as the intercept of the linear regression using the LowLAD method.

2.2. Estimation Methodology

In order to derive the estimation bias, we further assume that $m(\cdot)$ is five times continuously differentiable, that is, the regression function is two degrees smoother than our postulated model due to the equidistant design. Following the paradigm of Draper and Smith (1981) and Wang and Scott (1994), we discard the higher-order terms of $m(\cdot)$ and assume locally that the approximate model is

$$Y_{ij}^{(1)} = \beta_{i1} + \beta_{i3} d_j^2 + \beta_{i5} d_j^4 + \zeta_{ij},$$

where $\beta_i = (\beta_{i1}, \beta_{i3}, \beta_{i5})^T = (m^{(1)}(x_i), m^{(3)}(x_i)/6, m^{(5)}(x_i)/120)^T$ are the unknown coefficients of the true underlying quintic function, and $\zeta_{ij} = (\epsilon_{i+j} - \epsilon_{i-j})/(2j/n)$ with

Median$[\zeta_{ij}] = 0$. Under the assumption of the approximate model, β_{i1} can be estimated as

$$\min_b \sum_{j=1}^{k} w_j \big|Y_{ij}^{(1)} - (b_{i1} + b_{i3}d_j^2 + b_{i5}d_j^4)\big| = \min_b \sum_{j=1}^{k} \big|\tilde{Y}_{ij}^{(1)} - (b_{i1}d_j + b_{i3}d_j^3 + b_{i5}d_j^5)\big|,$$

where $w_j = d_j$ are the weights, $b = (b_{i1}, b_{i3}, b_{i5})^T$, and $\tilde{Y}_{ij}^{(1)} = (Y_{i+j} - Y_{i-j})/2$. Accordingly, the approximate model can be rewritten as

$$\tilde{Y}_{ij}^{(1)} = \beta_{i1}d_j + \beta_{i3}d_j^3 + \beta_{i5}d_j^5 + \tilde{\zeta}_{ij},$$

where $\tilde{\zeta}_{ij} = (\epsilon_{i+j} - \epsilon_{i-j})/2$ are iid random errors with Median$[\tilde{\zeta}_{ij}] = 0$ and a continuous, symmetric density $g(\cdot)$ which is positive in a neighborhood of zero (see Appendix A).

Rather than the best L_1 quintic fitting, we search for the best L_1 cubic fitting to $\tilde{Y}_{ij}^{(1)}$. Specifically, we estimate the model by LowLAD:

$$(\hat{\beta}_{i1}, \hat{\beta}_{i3}) = \arg\min_b \sum_{j=1}^{k} \big|\tilde{Y}_{ij}^{(1)} - b_{i1}d_j - b_{i3}d_j^3\big|$$

with $b = (b_{i1}, b_{i3})^T$, and define the LowLAD estimator of $m^{(1)}(x_i)$ as

$$\hat{m}_{\text{LowLAD}}^{(1)}(x_i) = \hat{\beta}_{i1}. \tag{7}$$

The following theorem states the asymptotic behavior of $\hat{\beta}_{i1}$.

Theorem 1 *Assume that ϵ_i are iid random errors with a continuous bounded density $f(\cdot)$. Then as $k \to \infty$ and $k/n \to 0$, $\hat{\beta}_{i1}$ in (7) is asymptotically normally distributed with*

$$\text{Bias}[\hat{\beta}_{i1}] = -\frac{m^{(5)}(x_i)}{504}\frac{k^4}{n^4} + o\left(\frac{k^4}{n^4}\right), \quad \text{Var}[\hat{\beta}_{i1}] = \frac{75}{16g(0)^2}\frac{n^2}{k^3} + o\left(\frac{n^2}{k^3}\right),$$

where $g(0) = 2\int_{-\infty}^{\infty} f^2(x)\mathrm{d}x$. The optimal k that minimizes the asymptotic mean square error (AMSE) is

$$k_{opt} \approx 3.26\left(\frac{1}{g(0)^2 m^{(5)}(x_i)^2}\right)^{1/11} n^{10/11},$$

and, consequently, the minimum AMSE is

$$\text{AMSE}[\hat{\beta}_{i1}] \approx 0.19\left(\frac{m^{(5)}(x_i)^6}{g(0)^{16}}\right)^{1/11} n^{-8/11}.$$

In the local median regression (see Section 3.1 below for its definition), the density $f(\cdot)$ is usually assumed to have a zero median and a positive $f(0)$ value. While in Theorem 1, we only require a continuity condition on the density $f(\cdot)$. In addition, the variance of the LowLAD estimator depends on $g(0) = 2\int_{-\infty}^{\infty} f^2(x)\mathrm{d}x$ which is always positive, while the variance of the local median estimator relies on a single value $f(0)$ only. In this sense, the LowLAD estimator is more robust than the local median estimator.

2.3. LowLAD with Random Difference

To further improve the estimation efficiency, we propose the LowLAD with random difference referred to as RLowLAD. First, define the first-order random difference sequence as

$$Y_{ijl} = Y_{i+j} - Y_{i+l}, \quad -k \le j, l \le k,$$

where we implicitly exclude j and l being 0 to be comparable to the LowLAD estimator. Second, define the RLowLAD estimator as

$$\hat{m}^{(1)}_{\text{RLowLAD}}(x_i) = \hat{\beta}^{\text{RLowLAD}}_{i1}, \tag{8}$$

where

$$\begin{aligned}&(\hat{\beta}^{\text{RLowLAD}}_{i1}, \hat{\beta}^{\text{RLowLAD}}_{i2}, \hat{\beta}^{\text{RLowLAD}}_{i3}, \hat{\beta}^{\text{RLowLAD}}_{i4})\\ &= \arg\min_{b} \sum_{j=-k}^{k} \sum_{l=-k, l<j}^{k} \left| Y_{ijl} - b_{i1}(d_j - d_l) - b_{i2}(d_j^2 - d_l^2) - b_{i3}(d_j^3 - d_l^3) - b_{i4}(d_j^4 - d_l^4) \right|\end{aligned}$$

with the true value $\beta_i = (\beta_{i1}, \beta_{i2}, \beta_{i3}, \beta_{i4})^T = (m^{(1)}(x_i), m^{(2)}(x_i)/2, m^{(3)}(x_i)/6, m^{(4)}(x_i)/24)^T$ and $b = (b_{i1}, b_{i2}, b_{i3}, b_{i4})^T$.

Theorem 2 *Under the assumptions of Theorem 1, the bias and variance of the RLowLAD estimator in* (8) *are, respectively,*

$$\text{Bias}[\hat{\beta}^{\text{RLowLAD}}_{i1}] = -\frac{m^{(5)}(x_i)}{504}\frac{k^4}{n^4} + o\left(\frac{k^4}{n^4}\right), \quad \text{Var}[\hat{\beta}^{\text{RLowLAD}}_{i1}] = \frac{75}{24g(0)^2}\frac{n^2}{k^3} + o\left(\frac{n^2}{k^3}\right).$$

By replacing the symmetric difference with the random difference, we improve the estimation accuracy with the variance from $\frac{75}{16g(0)^2}\frac{n^2}{k^3}$ to $\frac{75}{24g(0)^2}\frac{n^2}{k^3}$. While the cost is to increase the computation complexity from $O(n)$ to $O(n^2)$.

3. Comparison with Existing First-Order Derivative Estimators

We first review some existing methods for the first-order derivative estimation which are based on either the least squares regression or the quantile regression. Both methods can be used to estimate the first-order derivative of $m(\cdot)$ due to the special structure of model (1). Note that $\mathrm{E}[Y_i|x_i] = m(x_i) + \mathrm{E}[\epsilon_i]$, and $\mathrm{Q}_\tau(Y_i|x_i) = m(x_i) + \mathrm{Q}_\tau(\epsilon_i)$ because ϵ_i is independent of x_i, where $\mathrm{Q}_\tau(\epsilon_i)$ is the τth unconditional quantile of ϵ_i. Although $\mathrm{E}[Y_i|x_i]$ may not equal to $\mathrm{Q}_\tau(Y_i|x_i)$, their derivatives must be the same as the corresponding derivatives of $m(x_i)$. To make the existing estimators comparable to our LowLAD and RLowLAD estimators, we use the uniform kernel and the same order Taylor expansion throughout this section.

3.1. LS, LowLSR and LAD Estimators

In Fan and Gijbels (1996), the first-order derivative is estimated by the least squares method:

$$(\hat{\alpha}^{\text{LS}}_{i0}, \hat{\alpha}^{\text{LS}}_{i1}, \hat{\alpha}^{\text{LS}}_{i2}, \hat{\alpha}^{\text{LS}}_{i3}, \hat{\alpha}^{\text{LS}}_{i4}) = \arg\min_{\alpha} \sum_{j=-k}^{k} \left(Y_{i+j} - \alpha_{i0} - \alpha_{i1}d_j - \alpha_{i2}d_j^2 - \alpha_{i3}d_j^3 - \alpha_{i4}d_j^4\right)^2,$$

where $\alpha = (\alpha_{i0}, \alpha_{i1}, \alpha_{i2}, \alpha_{i3}, \alpha_{i4})^T$. For ease of comparison, we exclude the point $j = 0$ so that the same Y_i's are used as in the LowLAD estimation. Define the least squares (LS) estimator as

$$\hat{m}^{(1)}_{\text{LS}}(x_i) = \hat{\alpha}^{\text{LS}}_{i1}. \tag{9}$$

Corollary 3 *Under the assumptions of Theorem 1, the bias and variance of the LS estimator in* (9) *are, respectively,*

$$\text{Bias}[\hat{\alpha}^{\text{LS}}_{i1}] = -\frac{m^{(5)}(x_i)}{504}\frac{k^4}{n^4} + o\left(\frac{k^4}{n^4}\right), \quad \text{Var}[\hat{\alpha}^{\text{LS}}_{i1}] = \frac{75\sigma^2}{8}\frac{n^2}{k^3} + o\left(\frac{n^2}{k^3}\right).$$

To obtain the optimal convergence rate of the derivative estimation, Wang and Lin (2015) proposed the LowLSR estimator:

$$\hat{m}^{(1)}_{\text{LowLSR}}(x_i) = \hat{\alpha}^{\text{LowLSR}}_{i1}, \tag{10}$$

where

$$(\hat{\alpha}^{\text{LowLSR}}_{i1}, \hat{\alpha}^{\text{LowLSR}}_{i3}) = \arg\min_{\alpha_{i1}, \alpha_{i3}} \sum_{j=1}^{k} \left(\tilde{Y}^{(1)}_{ij} - \alpha_{i1} d_j - \alpha_{i3} d_j^3\right)^2.$$

Corollary 4 *Under the assumptions of Theorem 1, the bias and variance of the LowLSR estimator in* (10) *are, respectively,*

$$\text{Bias}[\hat{\alpha}^{\text{LowLSR}}_{i1}] = -\frac{m^{(5)}(x_i)}{504}\frac{k^4}{n^4} + o\left(\frac{k^4}{n^4}\right), \quad \text{Var}[\hat{\alpha}^{\text{LowLSR}}_{i1}] = \frac{75\sigma^2}{8}\frac{n^2}{k^3} + o\left(\frac{n^2}{k^3}\right).$$

Wang and Scott (1994) proposed the local polynomial least absolute deviation (LAD) estimator:

$$\hat{m}^{(1)}_{\text{LAD}}(x_i) = \hat{\beta}^{\text{LAD}}_{i1}, \tag{11}$$

where

$$(\hat{\beta}^{\text{LAD}}_{i0}, \hat{\beta}^{\text{LAD}}_{i1}, \hat{\beta}^{\text{LAD}}_{i2}, \hat{\beta}^{\text{LAD}}_{i3}, \hat{\beta}^{\text{LAD}}_{i4}) = \arg\min_{b} \sum_{j=-k}^{k} \left|Y_{i+j} - b_{i0} - b_{i1} d_j^1 - b_{i2} d_j^2 - b_{i3} d_j^3 - b_{i4} d_j^4\right|$$

with $b = (b_{i0}, b_{i1}, b_{i2}, b_{i3}, b_{i4})^T$.

Corollary 5 *Under the assumptions of Theorem 1, the bias and variance of the LAD estimator in* (11) *are, respectively,*

$$\text{Bias}[\hat{\beta}^{\text{LAD}}_{i1}] = -\frac{m^{(5)}(x_i)}{504}\frac{k^4}{n^4} + o\left(\frac{k^4}{n^4}\right), \quad \text{Var}[\hat{\beta}^{\text{LAD}}_{i1}] = \frac{75}{32 f(0)^2}\frac{n^2}{k^3} + o\left(\frac{n^2}{k^3}\right).$$

There is one key difference between the LS method and the LAD method. For the LS method, the LS estimator and the LowLSR estimator are asymptotically equivalent; while for the LAD method, the asymptotic variances of the LAD estimator and the LowLAD estimator are very different, although their asymptotic biases are the same. We provide the reasons for this key difference in Appendix F.

3.2. Quantile Regression Estimators

Quantile regression (Koenker and Bassett, 1978; Koenker, 2005) exploits the distribution information of the error term to improve the estimation efficiency. Following the composite quantile regression (CQR) in Zou and Yuan (2008), Kai et al. (2010) proposed the local polynomial CQR estimator. In general, the local polynomial CQR estimator of $m^{(1)}(x_i)$ is defined as

$$\hat{m}^{(1)}_{\mathrm{CQR}}(x_i) = \hat{\gamma}^{\mathrm{CQR}}_{i1}, \tag{12}$$

where

$$\begin{aligned}&(\left\{\hat{\gamma}^{\mathrm{CQR}}_{i0h}\right\}_{h=1}^{q}, \hat{\gamma}^{\mathrm{CQR}}_{i1}, \hat{\gamma}^{\mathrm{CQR}}_{i2}, \hat{\gamma}^{\mathrm{CQR}}_{i3}, \hat{\gamma}^{\mathrm{CQR}}_{i4}) \\ &= \arg\min_{\gamma} \sum_{h=1}^{q} \left(\sum_{j=-k}^{k} \rho_{\tau_h}(Y_{i+j} - \gamma_{i0h} - \gamma_{i1} d_j^1 - \gamma_{i2} d_j^2 - \gamma_{i3} d_j^3 - \gamma_{i4} d_j^4) \right),\end{aligned}$$

with $\gamma = \left(\{\gamma_{i0h}\}_{h=1}^{q}, \gamma_{i1}, \gamma_{i2}, \gamma_{i3}, \gamma_{i4}\right)^T$, $\rho_\tau(x) = \tau x - xI(x < 0)$ is the check function, and $\tau_h = h/(q+1)$.

Corollary 6 *Under the assumptions of Theorem 1, the bias and variance of the CQR estimator in* (12) *are, respectively,*

$$\mathrm{Bias}[\hat{\gamma}^{\mathrm{CQR}}_{i1}] = -\frac{m^{(5)}(x_i)}{504}\frac{k^4}{n^4} + o\left(\frac{k^4}{n^4}\right), \quad \mathrm{Var}[\hat{\gamma}^{\mathrm{CQR}}_{i1}] = \frac{75R_1(q)}{8}\frac{n^2}{k^3} + o\left(\frac{n^2}{k^3}\right),$$

where $R_1(q) = \sum_{l=1}^{q}\sum_{l'=1}^{q} \tau_{ll'} / \{\sum_{l=1}^{q} f(c_l)\}^2$, $c_l = F^{-1}(\tau_l)$, *and* $\tau_{ll'} = \min\{\tau_l, \tau_{l'}\} - \tau_l\tau_{l'}$. *As* $q \to \infty$,

$$R_1(q) \to \frac{1}{12(E[f(\epsilon)])^2} = \frac{1}{3g(0)^2}, \quad \mathrm{Var}[\hat{\gamma}^{\mathrm{CQR}}_{i1}] = \frac{75}{24g(0)^2}\frac{n^2}{k^3} + o\left(\frac{n^2}{k^3}\right).$$

Zhao and Xiao (2014) proposed the weighted quantile average (WQA) estimator for the regression function in nonparametric regression, an idea originated from Koenker (1984). We now extend the WQA method to estimate $m^{(1)}(x_i)$ using the local polynomial quantile regression. Specifically, we define

$$\hat{m}^{(1)}_{\mathrm{WQA}}(x_i) = \sum_{h=1}^{q} w_h \hat{\gamma}^{\mathrm{WQA}}_{i1h}, \tag{13}$$

where $\sum_{h=1}^{q} w_h = 1$, and

$$\begin{aligned}&(\hat{\gamma}^{\mathrm{WQA}}_{i0h}, \hat{\gamma}^{\mathrm{WQA}}_{i1h}, \hat{\gamma}^{\mathrm{WQA}}_{i2h}, \hat{\gamma}^{\mathrm{WQA}}_{i3h}, \hat{\gamma}^{\mathrm{WQA}}_{i4h}) \\ &= \arg\min_{\gamma_h} \left(\sum_{j=-k}^{k} \rho_{\tau_h}(Y_{i+j} - \gamma_{i0h} - \gamma_{i1h} d_j^1 - \gamma_{i2h} d_j^2 - \gamma_{i3h} d_j^3 - \gamma_{i4h} d_j^4) \right)\end{aligned}$$

with $\gamma_h = (\gamma_{i0h}, \gamma_{i1h}, \gamma_{i2h}, \gamma_{i3h}, \gamma_{i4h})^T$.

Corollary 7 *Under the assumptions of Theorem 1, the bias and variance of the WQA estimator in* (13) *are, respectively,*

$$\text{Bias}[\hat{m}^{(1)}_{\text{WQA}}(x_i)] = -\frac{m^{(5)}(x_i)}{504}\frac{k^4}{n^4} + +o\left(\frac{k^4}{n^4}\right), \quad \text{Var}[\hat{m}^{(1)}_{\text{WQA}}(x_i)] = \frac{75R_2(q|\mathbf{w})}{8}\frac{n^2}{k^3} + o\left(\frac{n^2}{k^3}\right),$$

where $R_2(q|\mathbf{w}) = \mathbf{w}^T\mathbf{H}\mathbf{w}$ *with* $\mathbf{w} = (w_1, \cdots, w_q)^T$, *and* $\mathbf{H} = \left\{\frac{\tau_{ll'}}{f(c_l)f(c_{l'})}\right\}_{1\le l,l'\le q}$. *The optimal weights are given by* $\mathbf{w}^* = \frac{\mathbf{H}^{-1}e_q}{e_q^T\mathbf{H}^{-1}e_q}$, *where* $e_q = (1, \ldots, 1)^T_{q\times 1}$, *and* $R_2(q|\mathbf{w}^*) = (e_q^T\mathbf{H}^{-1}e_q)^{-1}$. *As* $q \to \infty$, *under the regularity assumptions in Theorem 6.2 of Zhao and Xiao (2014), the variance of the optimal CQR is*

$$\text{Var}[\hat{m}^{(1)}_{\text{WQA}}(x_i)] = \frac{75I(f)^{-1}}{8}\frac{n^2}{k^3} + o\left(\frac{n^2}{k^3}\right),$$

where $I(f)$ *is the Fisher information of* f.

	LS	LowLSR	LAD	CQR	WQA	LowLAD	RLowLAD
s^2	σ^2	σ^2	$\frac{1}{4f(0)^2}$	$R_1(q)$	$R_2(q)$	$\frac{1}{2g(0)^2}$	$\frac{1}{3g(0)^2}$
Asymptotic s^2	σ^2	σ^2	$\frac{1}{4f(0)^2}$	$\frac{1}{3g(0)^2}$	$I(f)^{-1}$	$\frac{1}{2g(0)^2}$	$\frac{1}{3g(0)^2}$

Table 1: s^2 in the variances $\frac{75n^2}{8k^3}s^2$ of the existing first-order derivative estimators.

For the LS, LowLSR, LAD, CQR, WQA, LowLAD and RLowLAD estimators with the same k, their asymptotic biases are all the same. In contrast, their asymptotic variances are $\frac{75n^2}{8k^3}s^2$ with s^2 being $\sigma^2, \sigma^2, \frac{1}{4f(0)^2}, R_1(q), R_2(q)$, $\frac{1}{2g(0)^2}$ and $\frac{1}{3g(0)^2}$, as listed in Table 1. From the kernel interpretation of the differenced estimator by Wang and Yu (2017), we expect an equivalence between the LS and LowLSR estimators. As $q \to \infty$, we have $R_1(q) \to 1/\{3g(0)^2\}$ and $R_2(q) \to I(f)^{-1}$, and hence the WQA estimator is the most efficient estimator as q becomes large. Nevertheless, it requires the error density function to be known in advance to carry out the most efficient WQA estimator. Otherwise, a two-step procedure is needed, where the first step estimates the error density. For a fixed q, the three quantile-based estimators (i.e., LAD, CQR, WQA) depend only on the density values of $f(\cdot)$ at finite quantile points, whose behaviors are uncertain and may not be reliable. In contrast, our new estimators rely on $g(0) = 2E[f(x)]$, which includes all information on the density $f(\cdot)$ and hence is more robust.

3.3. Relationship Among the CQR, WQA and RLowLAD Estimators

From the asymptotic variances, we can see that the RLowLAD estimator is asymptotically equivalent to the infinitely CQR estimator. Why can this happen? Intuitively, they use the same information in different ways. First, in infinitely CQR, all local data (i.e., data at all quantiles) are employed to estimate the same parameter $m^{(1)}(x_i)$, which is the same as in RLowLAD. Second, in infinitely CQR, we first use data horizontally (at fixed τ) and then combine data vertically (across τ), while in RLowLAD, we first combine data vertically since

the distribution of the error term of Y_{ijl} is $\int f\left(F^{-1}(\tau)\right)d\tau$, and then run a single regression horizontally (at $\tau = 0.5$). It is interesting (and may be also surprising) to see that these two different ways of using information have the same efficiency. The RLowLAD estimation is more powerful in practice by noticing that a single differencing in Y_{ijl} is equivalent to combining all (infinitely many) quantiles.

It should be emphasized that the ways of combining the infinitely many quantiles in CQR and WQA are different. Suppose we use the equal weights $\mathbf{w}_E = \left(\frac{1}{q}, \cdots, \frac{1}{q}\right)^T$ in WQA. Such a weighting scheme is parallel to CQR where an equal weight is imposed on each check function. Then from Theorem 2 of Kai et al. (2010), $R_2(q|\mathbf{w}_E) \to 1$ as $q \to \infty$; while $R_1(q) \to \frac{1}{3g(0)^2}$. From Table 2 of Kai et al. (2010), $\frac{1}{3g(0)^2} < 1$ for most distributions (except $N(0,1)$). Why can this difference happen? Note that

$$\left\{\hat{\gamma}_{ih}^{\text{WQA}}\right\}_{h=1}^{q} = \arg\min_{\gamma_1,\cdots,\gamma_q}\left(\sum_{h=1}^{q}\sum_{j=-k}^{k}\rho_{\tau_h}(Y_{i+j} - \gamma_{i0h} - \gamma_{i1h}d_j^1 - \gamma_{i2h}d_j^2 - \gamma_{i3h}d_j^3 - \gamma_{i4h}d_j^4)\right)$$

where $\hat{\gamma}_{ih}^{\text{WQA}} = (\hat{\gamma}_{i0h}^{\text{WQA}}, \hat{\gamma}_{i1h}^{\text{WQA}}, \hat{\gamma}_{i2h}^{\text{WQA}}, \hat{\gamma}_{i3h}^{\text{WQA}}, \hat{\gamma}_{i4h}^{\text{WQA}})^T$, so the CQR estimator is a constrained WQA estimator with the constraints being that the slopes at different quantiles must be the same. On the other hand, the WQA estimator can be interpreted as a minimum distance estimator,

$$\arg\min_{\gamma}\left(\hat{\gamma}_{i1}^{\text{WQA}} - \gamma e_q\right)^T \mathbf{W}\left(\hat{\gamma}_{i1}^{\text{WQA}} - \gamma e_q\right) = \frac{e_q^T\mathbf{W}\hat{\gamma}_{i1}^{\text{WQA}}}{e_q^T\mathbf{W}e_q} = \mathbf{w}^T\hat{\gamma}_{i1}^{\text{WQA}},$$

where $\hat{\gamma}_{i1}^{\text{WQA}} = \left(\hat{\gamma}_{i11}^{\text{WQA}}, \cdots, \hat{\gamma}_{i1q}^{\text{WQA}}\right)^T$, $\mathbf{W}$ is a symmetric weight matrix, and $\mathbf{w} = \frac{\mathbf{W}e_q}{e_q^T\mathbf{W}e_q}$. When $\mathbf{W} = \mathbf{I}_q$, the $q \times q$ identity matrix, we get the equally-weighted WQA estimator; when $\mathbf{W} = \mathbf{H}^{-1}$, we get the optimally weighted WQA estimator; and when

$$\mathbf{W} = a\mathbf{I}_q + (1-a)\mathbf{H}^{-1} \neq \mathbf{I}_q,$$

we get an estimator that is asymptotically equivalent to the CQR estimator, where

$$a = \frac{-(q-B)(1-BC) + \sqrt{(q^2 - AB)(1-BC)}}{A - (2q-B)(1-BC) - q^2C} \neq 1$$

with $A = e_q^T\mathbf{H}e_q = q^2R_2(q|\mathbf{w}_E)$, $B = R_2(q|\mathbf{w}^*)^{-1}$ and $C = R_1(q)$. For example, if $\varepsilon \sim N(0,1)$, then when $q = 5$, $a = -0.367$; when $q = 9$, $a = -0.165$; when $q = 19$, $a = -0.067$; and when $q = 99$, $a = -0.011$. This is why imposing constraints directly on the objective function (i.e., the CQR estimator) or on the resulting estimators (i.e., the WQA estimator) would generate different estimators. The RLowLAD estimator and the CQR estimator have the same asymptotic variance because both of them impose constraints directly on the objective function.

3.4. Asymptotic Relative Efficiency

In this subsection, we study the Asymptotic Relative Efficiency (ARE) of the RLowLAD estimator with respect to the LowLSR and LAD estimators by comparing their asymptotic variances and AMSEs.

Since all estimators have the same bias, the comparison of their variances becomes important. We define the variance ratios of the LowLSR and LAD estimators relative to the RLowLAD estimator as

$$\mathrm{R}_{\mathrm{LowLSR}} = \frac{\mathrm{Var}(\hat{m}_{\mathrm{LowLSR}^{(1)}})}{\mathrm{Var}(\hat{m}^{(1)}_{\mathrm{RLowLAD}})} = 3\sigma^2 g(0)^2,$$
$$\mathrm{R}_{\mathrm{LAD}} = \frac{\mathrm{Var}(\hat{m}^{(1)}_{\mathrm{LAD}})}{\mathrm{Var}(\hat{m}^{(1)}_{\mathrm{RLowLAD}})} = \frac{3g(0)^2}{4f(0)^2}.$$

In addition, the overall performance of an estimator is usually measured by its AMSE, so we define the AREs based on AMSE as

$$\mathrm{ARE}_{\mathrm{LowLSR}} = \frac{\mathrm{AMSE}(\hat{m}^{(1)}_{\mathrm{LowLSR}})}{\mathrm{AMSE}(\hat{m}^{(1)}_{\mathrm{RLowLAD}})},$$
$$\mathrm{ARE}_{\mathrm{LAD}} = \frac{\mathrm{AMSE}(\hat{m}^{(1)}_{\mathrm{LAD}})}{\mathrm{AMSE}(\hat{m}^{(1)}_{\mathrm{RLowLAD}})}.$$

The LowLSR estimator has the AMSE

$$\mathrm{AMSE}(\hat{m}^{(1)}_{\mathrm{LowLSR}}) = \left\{ \frac{m^{(5)}(x_i)}{504} \frac{k^4}{n^4} \right\}^2 + \frac{75\sigma^2}{8} \frac{n^2}{k^3},$$

and the optimal k minimizing the AMSE is

$$k^{opt}_{\mathrm{LowLSR}} = \left\{ \frac{893025\sigma^2}{(m^{(5)}(x_i))^2} \right\}^{1/11} n^{10/11}$$

Similarly, we have

$$k^{opt}_{\mathrm{RLowLAD}} = \left\{ \frac{893025}{3g(0)^2(m^{(5)}(x_i))^2} \right\}^{1/11} n^{10/11} = R^{-1/11}_{\mathrm{LowLSR}} k^{opt}_{\mathrm{LowLSR}}.$$

As $n \to \infty$, we can show

$$\mathrm{ARE}_{\mathrm{LowLSR}} = {\mathrm{R}_{\mathrm{LowLSR}}}^{8/11},$$
$$\mathrm{ARE}_{\mathrm{LAD}} = {\mathrm{R}_{\mathrm{LAD}}}^{8/11}.$$

Since the variance ratio and ARE have a close relationship, we only report the variance ratios. We consider eight distributions for the random errors: the normal distribution $N(0, 1^2)$, the Laplace (double exponential) distribution, the Logistic distribution, the t distribution with 3 degrees of freedom, the mixed normal distribution $0.9N(0, 1^2)+0.1N(0, 3^2)$ and $0.9N(0, 1^2)+0.1N(0, 10^2)$, the Cauchy distribution, the mixed double gamma distribution $0.9Gamma(0, 1) + 0.1Gamma(1, 1)$ and $0.9Gamma(0, 1) + 0.1Gamma(3, 1)$, and the bimodal distribution $0.5N(-1, 1) + 0.5N(1, 1)$ and $0.5N(-3, 1) + 0.5N(3, 1)$, which were adopted in the robust location estimation by Koenker and Bassett (1978) and the variable selection by Zou and Yuan (2008).

	$f(0)$	$g(0)$	σ^2	$\mathrm{R}_{\mathrm{LowLSR}}$	$\mathrm{R}_{\mathrm{LAD}}$
$N(0,1)$	0.40	0.56	1	0.95	1.5
$0.9N(0,1)+0.1N(0,3^2)$	0.37	0.50	1.8	1.37	1.38
$0.9N(0,1)+0.1N(0,10^2)$	0.36	0.47	10.9	7.28	1.27
$t(3)$	0.37	0.46	3	1.90	1.17
Laplace	0.50	0.50	2	1.50	0.75
Logistic	0.25	0.33	3.29	1.10	1.33
Cauchy	0.32	0.32	∞	∞	0.75
$0.9Gamma(0,1)+0.1Gamma(1,1)$	0.50	0.45	1.1	1.63	0.68
$0.9Gamma(0,1)+0.1Gamma(3,1)$	0.46	0.42	1.3	2.44	0.68
$0.5N(-1,1)+0.5N(1,1)$	0.15	0.39	2	0.89	5.18
$0.5N(-3,1)+0.5N(3,1)$	4.92×10^{-5}	0.28	10	2.39	2.46×10^{7}

Table 2: Variance ratios for different error distributions.

Table 2 lists the variance ratios for these eight error distributions which are derived in Appendix E. From Table 2, a few interesting results can be drawn. First of all, the variance of the RLowLAD estimator is usually smaller and more robust than that of the LowLSR and LAD estimators in most cases. Secondly, the LAD estimator is not robust since it relies on the density value at one point 0, so when facing the bimodal errors the variance is huge. Thirdly, the minimum value of $\mathrm{R}_{\mathrm{LowLSR}}$ in Table 2 is 0.89. Actually, there is an exact lower bound for $\mathrm{R}_{\mathrm{LowLSR}}$ as stated in Theorem 4 of Kai et al. (2010). For completeness, we repeat their Theorem 4 in the following Lemma 8.

Lemma 8 *Let $\mathcal{F}$ denote the class of error distributions with mean 0 and variance 1. Then*

$$\inf_{f\in\mathcal{F}} \mathrm{R}_{\mathrm{LowLSR}}(f) \approx 0.86.$$

The lower bound is reached if and only if the error follows the rescaled beta$(2,2)$ distribution. Thus,

$$\inf_{f\in\mathcal{F}} \mathrm{ARE}_{\mathrm{LowLSR}}(f) = \mathrm{R}_{\mathrm{LowLSR}}^{8/11} \approx 0.90.$$

In other words, the potential efficiency loss of the RLowLAD estimator relative to the LowLSR estimator is at most 10%.

In Appendix E, we further illustrate the trade-off between the sharp-peak and heavy-tailed errors using three error distributions.

4. Derivative Estimation at Boundaries

At the left boundary with $2 \le i \le k$, the bias and variance of the LowLAD estimator are $-m^{(5)}(x_i)(i-1)^4/(504n^4)$ and $75n^2/(16g(0)^2(i-1)^3)$, respectively. At the endpoint $i=1$, the LowLAD estimator is not well defined. Similar results hold for the estimation at the right boundary with $n-k+1 \le i \le n-1$. In this section, we propose an asymmetrical LowLAD method (As-LowLAD) to reduce the estimation variance as well as to improve the finite-sample performance at the boundaries.

Assume that $m(\cdot)$ is twice continuously differentiable on $[0,1]$ and Median$[\epsilon_i]=0$. For $1\le i\le k$, we define the asymmetric lag-j first-order difference quotients as

$$Y_{ij}^{<1>}=\frac{Y_{i+j}-Y_i}{x_{i+j}-x_i},\quad -(i-1)\le j\le k,\quad j\neq 0.$$

By decomposing $Y_{ij}^{<1>}$ into two parts, we have

$$\begin{aligned}Y_{ij}^{<1>}&=\frac{m(x_{i+j})-m(x_i)}{j/n}+\frac{\epsilon_{i+j}-\epsilon_i}{j/n}\\&=m^{(1)}(x_i)+(-\epsilon_i)d_j^{-1}+\frac{\epsilon_{i+j}}{j/n}+\frac{m^{(2)}(x_i)}{2!}\frac{j}{n}+o\left(\frac{j}{n}\right),\end{aligned}$$

where ϵ_i is fixed as j changes. Thus, Median$(Y_{ij}^{<1>}\big|\epsilon_i)=m^{(1)}(x_i)+(-\epsilon_i)d_j^{-1}+\frac{m^{(2)}(x_i)}{2!}\frac{j}{n}$. Ignoring the last term, we can rewrite the above model as

$$Y_{ij}^{<1>}=\beta_{i1}+\beta_{i0}d_j^{-1}+\delta_{ij}+o(1),\quad -(i-1)\le j\le k,\quad j\neq 0,$$

where $(\beta_{i0},\beta_{i1})^T=(-\epsilon_i,m^{(1)}(x_i))^T$, $\delta_{ij}=\frac{\epsilon_{i+j}}{j/n}$. By the LowLAD method, the regression coefficients can be estimated as

$$\begin{aligned}\left(\hat\beta_{i0},\hat\beta_{i1}\right)&=\arg\min_b\sum_{j=-(i-1)}^{k}\big|Y_{ij}^{<1>}-b_{i1}-b_{i0}d_j^{-1}\big|w_j\\&=\arg\min_b\sum_{j=-(i-1)}^{k}\big|\tilde Y_{ij}^{<1>}-b_{i0}-b_{i1}d_j\big|,\end{aligned}\tag{14}$$

where $w_j=d_j$, $b=(b_{i0},b_{i1})^T$, and $\tilde Y_{ij}^{<1>}=Y_{i+j}-Y_i$. The As-LowLAD estimator of $m^{(1)}(x_i)$ is $\hat\beta_{i1}$.

Similarly to Theorem 1, we can prove the asymptotic normality for $\hat\beta_{i1}$. The following theorem states its asymptotic bias and variance.

Theorem 9 *Assume that ϵ_i are iid random errors with median 0 and a continuous, positive density $f(\cdot)$ in a neighborhood of zero. Furthermore, assume that $m(\cdot)$ is twice continuously differentiable on $[0,1]$. Then for each $1\le i\le k$, the leading terms of the bias and variance of $\hat\beta_{i1}$ in (14) are, respectively,*

$$\begin{aligned}\mathrm{Bias}[\hat\beta_{i1}]&=\frac{m^{(2)}(x_i)}{2}\frac{k^4+2k^3i-2ki^3-i^4}{n(k^3+3k^2i+3ki^2+i^3)},\\\mathrm{Var}[\hat\beta_{i1}]&=\frac{3}{f(0)^2}\frac{n^2}{k^3+3k^2i+3ki^2+i^3}.\end{aligned}$$

For the estimation at the boundaries, Wang and Lin (2015) proposed a one-side LowLSR (OS-LowLSR) estimator of the first-order derivative, with the bias and variance being $m^{(2)}(x_i)k/(2n)$ and $12\sigma^2n^2/k^3$, respectively. In contrast, if we consider the one-side LowLAD (OS-LowLAD) estimator of the first-order derivative, its bias and variance are $m^{(2)}(x_i)k/(2n)$

and $3n^2/(f(0)^2k^3)$, respectively. Note that the two biases are the same, while the variances are different with the former related to σ^2 and the latter related to $f(0)$. Note also that different from the LowLAD estimator at the interior point, the variance of the OS-LowLAD estimator involves $f(0)$ rather than $g(0)$.

Theorem 9 shows that our estimator has a smaller bias than the OS-LowLSR and OS-LowLAD estimators. In the special case $i = k$, the bias disappears (in fact it reduces to the higher-order term $O(k^2/n^2)$). With normal errors, when $1 < i < \lfloor 0.163k \rfloor$, the order of variances of the three estimators is $\mathrm{Var}(\hat{m}^{(1)}_{\mathrm{OS-LAD}}(x_i)) > \mathrm{Var}(\hat{m}^{(1)}_{\mathrm{AS-LAD}}(x_i)) > \mathrm{Var}(\hat{m}^{(1)}_{\mathrm{OS-LSR}}(x_i))$, where for a real number x, $\lfloor x \rfloor$ means the largest integer less than x; when $\lfloor 0.163k \rfloor < i < k$, the order of variances of the three estimators is $\mathrm{Var}(\hat{m}^{(1)}_{\mathrm{OS-LSR}}(x_i)) > \mathrm{Var}(\hat{m}^{(1)}_{\mathrm{OS-LAD}}(x_i)) > \mathrm{Var}(\hat{m}^{(1)}_{\mathrm{AS-LAD}}(x_i))$. As i approaches k, the variance of the As-LowLAD estimator is reduced to one-eighth of the variance of the OS-LowLAD estimator, and is much smaller than the variance of the OS-LowLSR estimator.

Up to now, we have the first-order derivative estimators $\{\hat{m}^{(1)}(x_i)\}_{i=1}^n$ on the discrete points $\{x_i\}_{i=1}^n$. To estimate the first-order derivative function, we suggest two strategies for different noise levels of the derivative data: the cubic spline interpolation in Knott (2000) for 'good' derivative estimators, and the local polynomial regression in Brown and Levine (2007) and De Brabanter et al. (2013) for 'bad' derivative estimators. Here, the terms 'good' and 'bad' indicate small and large estimation variances of the derivative estimators, respectively.

5. Second- and Higher-Order Derivative Estimation

In this section, we generalize our robust method for the first-order derivative estimation to the second- and higher-order derivatives estimation.

5.1. Second-Order Derivative Estimation

Define the second-order difference quotients as

$$Y_{ij}^{(2)} = \frac{Y_{i-j} - 2Y_i + Y_{i+j}}{j^2/n^2}, \quad 1 \le j \le k, \tag{15}$$

and assume that $m(\cdot)$ is six times continuously differentiable. Then we can decompose (15) into two parts and simplify it by the Taylor expansion as

$$\begin{aligned} Y_{ij}^{(2)} &= \frac{m(x_{i-j}) - 2m(x_i) + m(x_{i+j})}{j^2/n^2} + \frac{\epsilon_{i-j} - 2\epsilon_i + \epsilon_{i+j}}{j^2/n^2} \\ &= m^{(2)}(x_i) + \frac{m^{(4)}(x_i)}{12}\frac{j^2}{n^2} + \frac{m^{(6)}(x_i)}{360}\frac{j^4}{n^4} + o\left(\frac{j^4}{n^4}\right) + \frac{\epsilon_{i-j} - 2\epsilon_i + \epsilon_{i+j}}{j^2/n^2}. \end{aligned}$$

Since i is fixed as j varies, the conditional expectation of $Y_{ij}^{(2)}$ given ϵ_i is

$$\mathrm{E}[Y_{ij}^{(2)} \big| \epsilon_i] = m^{(2)}(x_i) + \frac{m^{(4)}(x_i)}{12}\frac{j^2}{n^2} + \frac{m^{(6)}(x_i)}{360}\frac{j^4}{n^4} + o\left(\frac{j^4}{n^4}\right) + (-2\epsilon_i)\frac{n^2}{j^2}.$$

This results in the true regression model as

$$Y_{ij}^{(2)} = \alpha_{i2} + \alpha_{i4} d_j^2 + \alpha_{i6} d_j^4 + o\left(d_j^4\right) + \alpha_{i0} d_j^{-2} + \delta_{ij}, \quad 1 \le j \le k,$$

where $\alpha_i = (\alpha_{i0}, \alpha_{i2}, \alpha_{i4}, \alpha_{i6})^T = (-2\epsilon_i, m^{(2)}(x_i), m^{(4)}(x_i)/12, m^{(6)}(x_i)/360)^T$, and the errors $\delta_{ij} = \frac{\epsilon_{i+j}+\epsilon_{i-j}}{j^2/n^2}$. If ϵ_i has a symmetric density about zero, then $\tilde{\delta}_{ij} = \frac{j^2}{n^2}\delta_{ij} = \epsilon_{i+j} + \epsilon_{i-j}$ has median 0 (see Appendix D). Following the similar procedure as in the first-order derivative estimation, our LowLAD estimator of $m^{(2)}(x_i)$ is defined as

$$\hat{m}^{(2)}_{\text{LowLAD}}(x_i) = \hat{\alpha}_{i2}, \tag{16}$$

where

$$\begin{aligned}(\hat{\alpha}_{i0}, \hat{\alpha}_{i2}, \hat{\alpha}_{i4})^T &= \arg\min_a \sum_{j=1}^{k} \left|Y_{ij}^{(2)} - (a_{i2} + a_{i4} d_j^2 + a_{i0} d_j^{-2})\right| W_j \\ &= \arg\min_a \sum_{j=1}^{k} \left|\tilde{Y}_{ij}^{(2)} - (a_{i0} + a_{i2} d_j^2 + a_{i4} d_j^4)\right|\end{aligned}$$

with $a = (a_{i0}, a_{i2}, a_{i4})^T$, $W_j = d_j^2$, and $\tilde{Y}_{ij}^{(2)} = Y_{i-j} - 2Y_i + Y_{i+j}$. The following theorem shows that $\hat{m}^{(2)}_{\text{LowLAD}}(x_i)$ behaves similarly as $\hat{m}^{(1)}_{\text{LowLAD}}(x_i)$.

Theorem 10 *Assume that ϵ_i are iid random errors whose density $f(\cdot)$ is continuous and symmetric about zero. Then as $k \to \infty$ and $k/n \to 0$, $\hat{\alpha}_{i2}$ in (16) is asymptotically normally distributed with*

$$\text{Bias}[\hat{\alpha}_{i2}] = -\frac{m^{(6)}(x_i)}{792}\frac{k^4}{n^4} + o\left(\frac{k^4}{n^4}\right), \quad \text{Var}[\hat{\alpha}_{i2}] = \frac{2205}{16h(0)^2}\frac{n^4}{k^5} + o\left(\frac{n^4}{k^5}\right),$$

where $h(0) = \int_{-\infty}^{\infty} f^2(x)\mathrm{d}x$. The optimal k that minimizes the AMSE is

$$k_{opt} \approx 3.93 \left(\frac{1}{h(0)^2 m^{(6)}(x_i)^2}\right)^{1/13} n^{12/13},$$

and, consequently, the minimum AMSE is

$$\text{AMSE}[\hat{\alpha}_{i2}] \approx 0.24 \left(\frac{m^{(6)}(x_i)^{10}}{h(0)^{16}}\right)^{1/13} n^{-8/13}.$$

5.2. Higher-Order Derivative Estimation

We now propose a robust method for estimating the higher-order derivatives $m^{(l)}(x_i)$ with $l > 2$ via a two-step procedure. In the first step, we construct a sequence of symmetric difference quotients in which the higher-order derivative is the intercept of the linear regression derived by the Taylor expansion, and in the second step, we estimate the higher-order derivative using the LowLAD method.

When l is odd, let $d = (l+1)/2$. We linearly combine $m(x_{i\pm j})$ subject to

$$\sum_{h=1}^{d}[a_{jd+h}m(x_{i+jd+h}) + a_{-(jd+h)}m(x_{i-(jd+h)})] = m^{(l)}(x_i) + O\left(\frac{j}{n}\right), \quad 0 \le j \le k,$$

where k is a positive integer. We can derive a total of $2d$ equations through the Taylor expansion to solve out the $2d$ unknown parameters. Define

$$Y_{ij}^{(l)} = \sum_{h=1}^{d}[a_{jd+h}Y_{i+jd+h} + a_{-(jd+h)}Y_{i-(jd+h)}],$$

and consider the linear regression

$$Y_{ij}^{(l)} = m^{(l)}(x_i) + \delta_{ij}, \quad 0 \le j \le k,$$

where $\delta_{ij} = \sum_{h=1}^{d}[a_{i,jd+h}\epsilon_{i+jd+h} + a_{i,-(jd+h)}\epsilon_{i-(jd+h)}] + O(j/n)$.

When l is even, let $d = l/2$. We linearly combine $m(x_{i\pm j})$ subject to

$$b_j m(x_i) + \sum_{h=1}^{d}[a_{jd+h}m(x_{i+jd+h}) + a_{-(jd+h)}m(x_{i-(jd+h)})] = m^{(l)}(x_i) + O\left(\frac{j}{n}\right), \quad 0 \le j \le k,$$

where k is a positive integer. We can derive a total of $2d+1$ equations through the Taylor expansion to solve out the $2d+1$ unknown parameters. Define

$$Y_{ij}^{(l)} = b_j Y_i + \sum_{h=1}^{d}[a_{jd+h}Y_{i+jd+h} + a_{-(jd+h)}Y_{i-(jd+h)}],$$

and consider the linear regression

$$Y_{ij}^{(l)} = m^{(l)}(x_i) + b_j\epsilon_i + \delta_{ij}, \quad 0 \le j \le k,$$

where $\delta_{ij} = \sum_{h=1}^{d}[a_{i,jd+h}\epsilon_{i+jd+h} + a_{i,-(jd+h)}\epsilon_{i-(jd+h)}] + O(j/n)$.

When k is large, we suggest to keep the j^2/n^2 term as in (6) to reduce the estimation bias. If $\sum_{h=1}^{d}[a_{i,jd+h}\epsilon_{i+jd+h} + a_{i,-(jd+h)}\epsilon_{i-(jd+h)}]$ has median zero, then we can obtain the higher-order derivative estimators by the LowLAD method and deduce their asymptotic properties by similar arguments as in the previous sections. To save space, we omit the technical details in this paper.

6. Simulation Studies and Empirical Application

In this section, we conduct simulations to evaluate the finite-sample performance of our first- and second-order derivative estimators and compare them with some existing estimators. We also apply our method to a real data set to illustrate its usefulness in practice.

6.1. First-Order Derivative Estimators

We first consider the following three regression functions:

$$m_0(x) = (x) = \sqrt{x(1-x)}\sin((2.1\pi)/(x+0.05)), \quad x \in [0.25, 1],$$
$$m_1(x) = \sin(2\pi x) + \cos(2\pi x) + \log(4/3 + x), \quad x \in [-1, 1],$$
$$m_2(x) = 32e^{-8(1-2x)^2}(1-2x), \quad x \in [0, 1].$$

These three functions were also considered in Hall (2010) and De Brabanter et al. (2013).

For normal errors, we consider the function $m_0(x)$ and compare the LowLAD, RLowLAD and LowLSR estimators. The data set of size 300 is generated from model (1) with errors $\epsilon_i \overset{iid}{\sim} N(0, 0.1^2)$ and is plotted in Figure 1(a). Figure 1 (2) displays the LowLAD (RLowLAD) estimator (use the R package 'L1pack' in Osorio (2015)) and the LowLSR estimator with $k \in \{6, 12, 25, 30, 50\}$. When k is small (see Figure 1(b) and 1(c)), both estimators are noise-corrupted versions of the true first-order derivatives; as k becomes larger (see Figure 1(d)-(f)), our estimator provides a similar performance as the LowLSR estimator. Furthermore, by combining the left part of Figure 1(d), the middle part of 1(e) and the right part of 1(f), more accurate derivative estimators can be obtained for practical use.

In addition, note that the three estimators have the same variation trend, whereas the LowLAD estimator has a slightly large oscillation and the RLowLAD estimator has a similar performance compared to the LowLSR estimator. These simulation results coincide with the theoretical results: the three estimators have the same bias, which explains the same variation trend; the variance ratios are $\mathrm{Var}(\hat{m}^{(1)}_{\mathrm{LowLSR}})/\mathrm{Var}(\hat{m}^{(1)}_{\mathrm{LowLAD}}) \approx 0.64$ and $\mathrm{Var}(\hat{m}^{(1)}_{\mathrm{LowLSR}})/\mathrm{Var}(\hat{m}^{(1)}_{\mathrm{RLowLAD}}) \approx 0.95$, which explains the oscillation performance.

Next, we consider the non-normal errors: 90% of the errors come from $\epsilon \sim N(0, \sigma^2)$ with $\sigma = 0.1$, and the remaining 10% come from $\epsilon \sim N(0, \sigma_0^2)$ with $\sigma_0 = 1$ or 10 corresponding to the low or high contamination level. Figures 3 and 4 present the finite-sample performance of the first-order derivative estimators for the regression functions m_1 and m_2, respectively. They show that the estimated curves of the first-order derivative based on LowLAD fit the true curves more accurately than LowLSR in the presence of heavy-tailed errors. The heavier the tail, the more significant the improvement.

We also compute the mean absolute errors to further assess the performance of the four methods, i.e., LowLAD, RLowLAD, LowLSR and LAD. Since the oscillation of a periodic function depends on its frequency and amplitude, we consider the sine function in the following form as the regression function,

$$m_3(x) = A\sin(2\pi f x), \quad x \in [0, 1].$$

The errors are generated in the above contaminated way. We consider two sample sizes: $n = 100$ and 500, two standard deviations: $\sigma = 0.1$ and 0.5, two contaminated standard deviations: $\sigma_0 = 1$ and 10, two frequencies: $f = 0.5$ and 1, and two amplitudes: $A = 1$ and 10.

We use the following adjusted mean absolute error (AMAE) as the criterion of performance evalution:

$$\mathrm{AMAE}(k) = \frac{1}{n-2k}\sum_{i=k+1}^{n-k}\left|\hat{m}^{(1)}(x_i) - m^{(1)}(x_i)\right|.$$

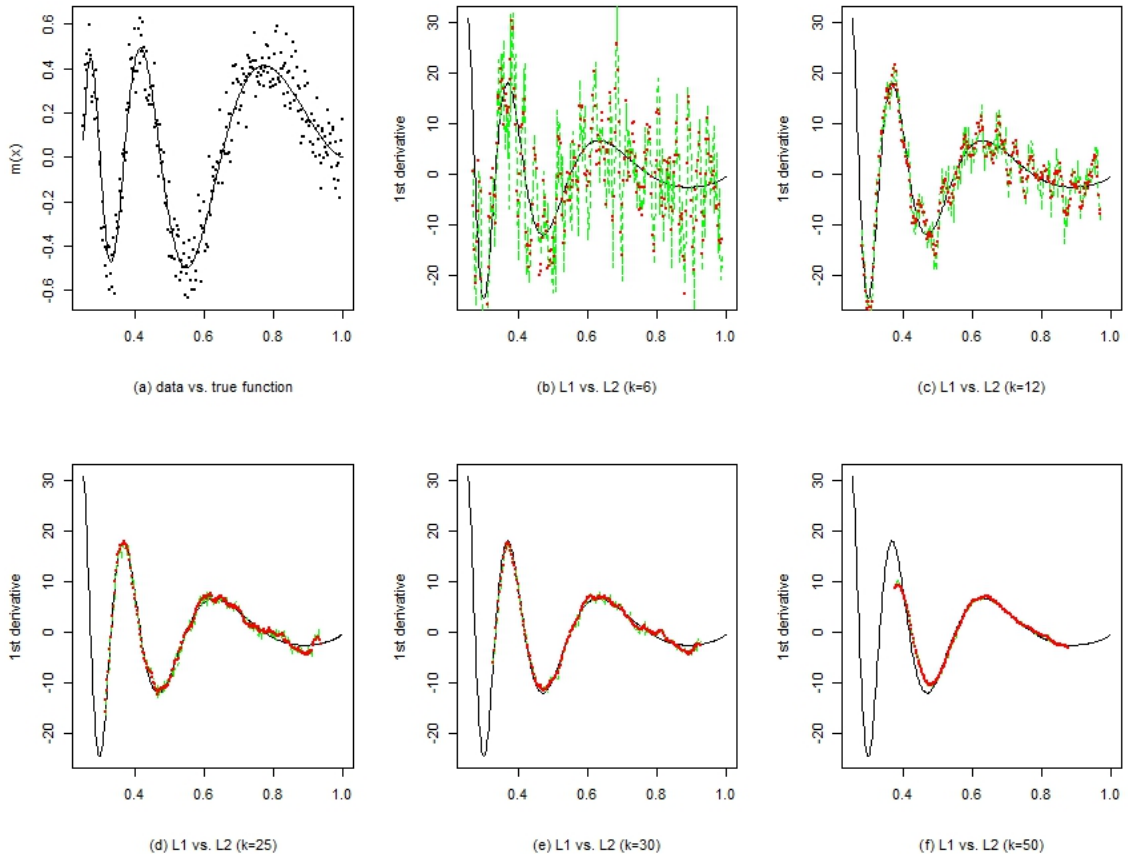

Figure 1: The comparison between the LowLAD and LowLSR estimators. (a) Simulated data set of size 300 from model (1) with equidistant $x_i \in [0.25, 1]$, $\epsilon_i \overset{iid}{\sim} N(0, 0.1^2)$, and the true regression function $m_0(x)$ (bold line). (b)-(f) The first-order LowLAD derivative estimators (green points) and the first-order LowLSR derivative estimators (red dashed line) for $k \in \{6, 9, 12, 25, 30, 50\}$. As a reference, the true first-order derivative fucntion is also plotted (bold line).

Due to the heavy computation (for example, it needs more than 48 hours for the case $n = 500$ and $k = n/5 = 100$ based on 1000 repetitions on our personal computer), we choose $k = n/5$ uniformly.

Table 3 reports the simulation results based on 1000 repetitions. The numbers outside and inside the parentheses represent the mean and standard deviation of the AMAE, respectively. It is evident that RLowLAD performs uniformly better than LowLAD and performs the best for most of cases. In particular for the cases with $\sigma_0 = 2\sigma$, the contamination is very light and thus LowLSR is better than LowLAD; while for the cases with $\sigma_0 = 10\sigma$, the contamination is very heavy and thus LowLSR is worse than LowLAD. These simulation results coincide with the theoretical results.

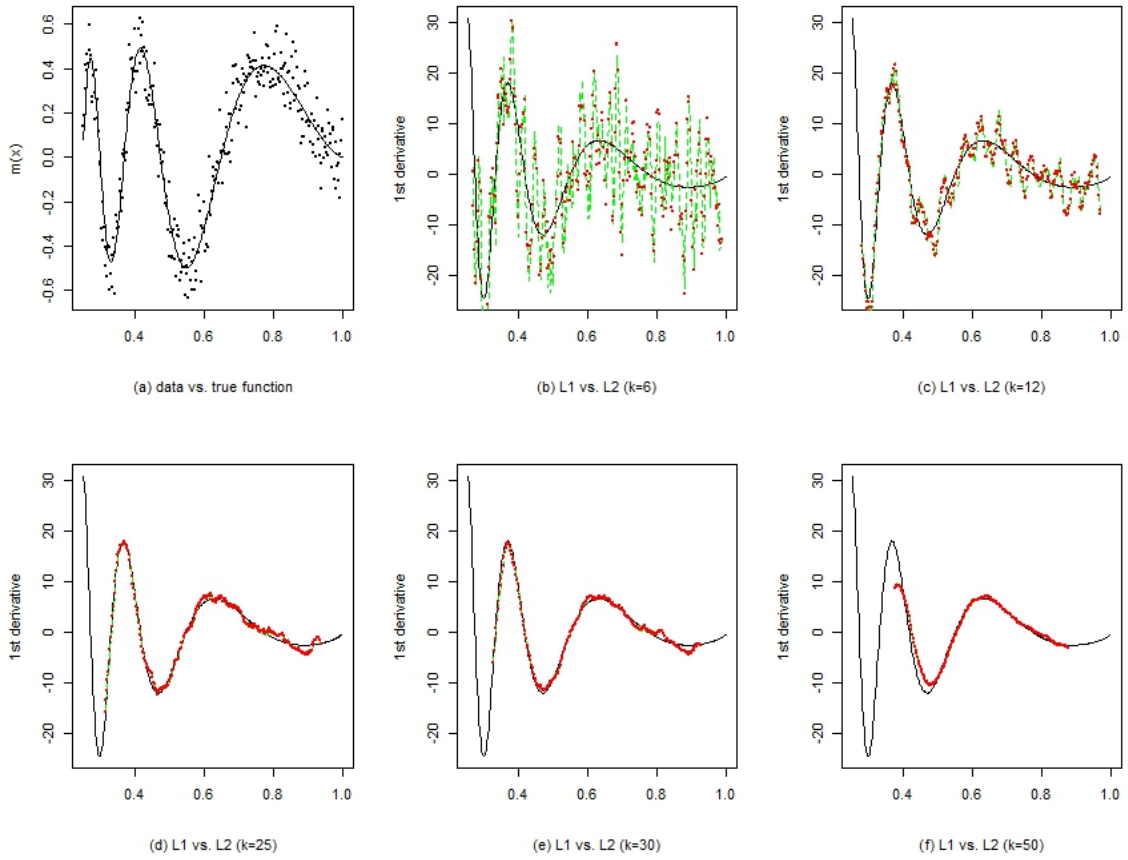

Figure 2: The comparison between the RLowLAD and LowLSR estimators for the same data set as in Figure 1

6.2. Second-Order Derivative Estimators

To assess the finite-sample performance of the second-order derivative estimators, we consider the same regression functions as in Section 6.1. Figures 5 and 6 present the estimated curves of the second-order derivatives of m_1 and m_2, respectively. It shows that our LowLAD estimator fits the true curves more accurately than the LowLSR estimator in all settings.

We further compare LowLAD with two other well-known methods: the local polynomial regression with $p = 5$ (use R package 'locpol' in Cabrera (2012)) and the penalized smoothing splines with $norder = 6$ and $method = 4$ (use R package 'pspline' in Ramsay and Ripley (2013)). For simplicity, we consider the simple version of m_3 with $A = 5$ and $f = 1$:

$$m_4(x) = 5\sin(2\pi x), \quad x \in [0, 1].$$

We let $n = 500$, and generate the errors in the same way as in Section 6.1. With 1000 repetitions, the simulation results are reported in Figures 7 and 8 which indicate that our

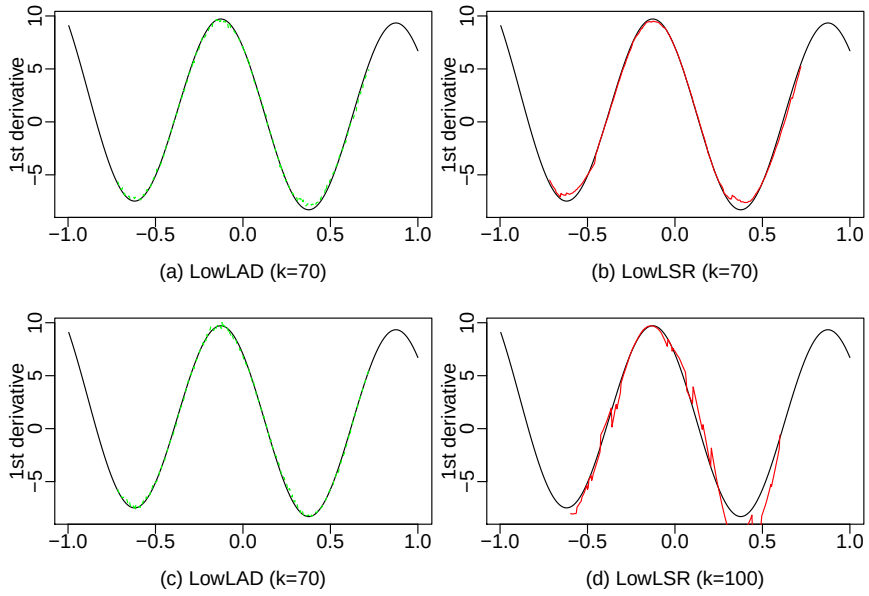

Figure 3: (a-b) The true first-order derivative function (bold line), LowLAD (green line) and LowLSR estimators (red line). Model (1) with equidistant $x_i \in [-1, 1]$, regression function m_1, and $\epsilon \sim 90\%N(0, 0.1^2) + 10\%N(0, 1^2)$. (c-d) The same designs as in (a-b) except $\epsilon \sim 90\%N(0, 0.1^2) + 10\%N(0, 10^2)$.

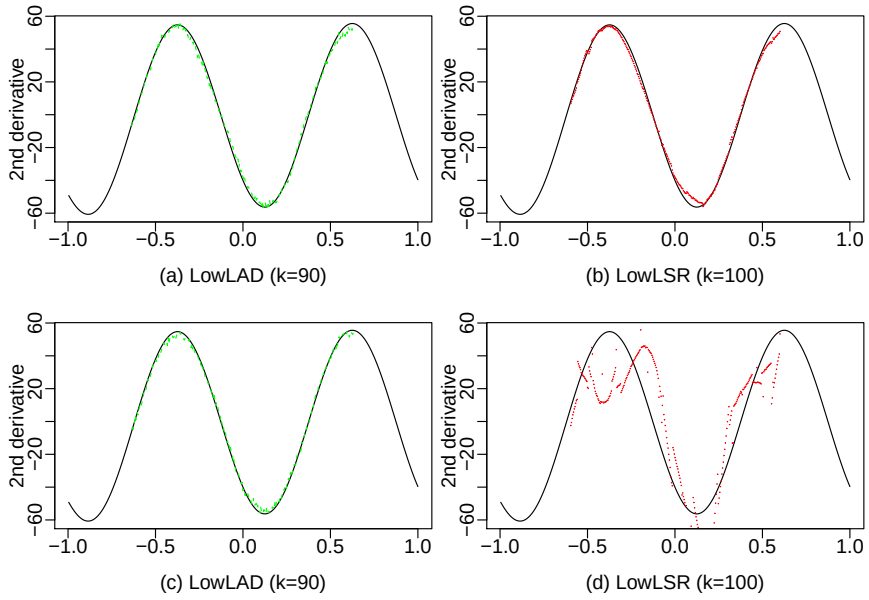

Figure 4: (a-d) The true first-order derivative function (bold line), LowLAD (green line) and LowLSR estimators (red line). The same designs as in Figure 3 except the regression function being m_2.

robust estimator is superior to the existing methods in the presence of sharp-peak and heavy-tailed errors.

6.3. House Price of China in Latest Ten Years

In reality, there are many data sets recorded by year, month, week, day, hour, minute, etc. For example, human growth is usually recorded by year, and temperature is recorded by hour, day or month. In this section, we apply RLowLAD to the data set of house price in two cities of China, i.e., Beijing and Jinan. We collect these monthly data from the web: http://www.creprice.cn/ (see Figure 9), which last from January 2008 to July 2018 and have size 127. We analyze this date set in two steps. Firstly, we apply our method to estimate the first-order derivative with $k = 8$ for RLowLAD and $k = 6$ for lower-order RLowLAD, where the lower-order means that we conduct the Taylor expansion to order 2 instead of order 4. Secondly, we define the relative growth rate as the ratio between the RLowLAD estimator and the house price at the same month, and then plot the relative growth rates in Figures 10 and 11. In the last ten years, the house price goes through tricycle fast increasing, and the monthly growth rate is larger than 0 most of the time with the maximum value at about 0.05.

7. Conclusion and Extensions

In this paper, we propose a robust differenced method for estimating the first- and higher-order derivatives of the regression function in nonparametric models. The new method consists of two main steps: first construct a sequence of symmetric difference quotients, and second estimate the derivatives using the LowLAD regression. The main contributions are as follows:

(1) Unlike LAD, our proposed LowLAD has the unique property of *double robustness* (or *robustness*^{2}. Specifically, it is robust not only to heavy-tailed error distributions (like LAD), but also to low density of the error term at a specific quantile (LAD needs a high value of the error density at median; otherwise, the relative efficiency of LAD can be arbitrarily small compared with LowLAD). Following Theorem 1, the asymptotic variance of the LowLAD estimator includes the term $g(0) = 2\int_{-\infty}^{\infty} f^2(x)dx = 2\int_{-\infty}^{\infty} f(F^{-1}(\tau))d\tau$, which implies that we are able to utilize the information of the whole error density. While for the LAD estimator, its variance depends on a single value $f(0)$ only. In this sense, the LowLAD estimator is more robust than the LAD estimator.

(2) Our proposed LowLAD does not require the error distribution to have a zero median, and so is more flexible than LAD. To be more specific, our symmetric differenced errors are guaranteed to have a zero median and a positive symmetric density in a neighborhood of zero, regardless of whether or not the distribution of the original error is symmetric. While for LAD, we must require the error distribution to have a zero median, and consequently, the practical usefulness of LAD will be rather limited.

(3) More surprisingly, as an extension of LowLAD, our proposed RLowLAD based on random difference is asymptotically equivalent to the infinitely composite quantile regres-

sion (CQR) estimator. *In other words, running one RLowLAD regression is equivalent to combining infinitely many quantile regressions.*

(4) Lastly, it is also worthwhile to mention that the differences between LowLAD and LAD are strikingly distinct from the differences between LowLSR and LS. For the same data and the same tuning parameter k, we have LS = LowLSR, whereas LAD $\neq$ LowLAD. What is more, RLowLAD is able to further improve the estimation efficiency compared with LowLAD, while RLowLSR, the LS counterpart of RLowLAD, is not able to improve efficiency relative to LowLSR.

LowLAD is a new idea to explore the information of density function by combining first-order difference and LAD. We can adopt the third-order symmetric difference $\{(Y_{i+j} - Y_{i-j}) + (Y_{i+l} - Y_{i-l})\}$ or the third-order random difference $\{(Y_{i+j} + Y_{i+l}) - (Y_{i+u} + Y_{i+v})\}$, even higher-order difference, to explore the information of density function. Whether and how to achieve the Cramer-Rao Lower bound deserves further study. These questions would be investigated in a separate paper.

In this paper, we focus on the derivative estimation with fixed designs and iid errors. With minor technical extensions, the proposed method can be extended to random designs with heteroskedastic errors. Further extensions to linear model, high-dimensional model for variable selection, semiparametric model, and change-point detection are also possible.

Acknowledgements

We would like to thank two anonymous reviewers and action editor for their constructive comments on improving the quality of the paper. Wang's work was supported by Qufu Normal University, University of Hong Kong, Hong Kong Baptist University. Lin's work was supported by NNSF projects of China.

Appendix A. Proof of Theorem 1

Proposition 11 *If ϵ_i are iid with a continuous, positive density $f(\cdot)$ in a neighborhood of the median, then $\tilde{\zeta}_{ij} = (\epsilon_{i+j} - \epsilon_{i-j})/2$ (j=1,..., k) are iid with median 0 and a continuous, positive, symmetric density $g(\cdot)$, where*

$$g(x) = 2\int_{-\infty}^{\infty} f(2x+\epsilon)f(\epsilon)\mathrm{d}\epsilon.$$

Proof of Proposition 11 The distribution of $\tilde{\zeta}_{ij} = (\epsilon_{i+j} - \epsilon_{i-j})/2$ is

$$\begin{aligned} F_{\tilde{\zeta}_{ij}}(x) &= P((\epsilon_{i+j} - \epsilon_{i-j})/2 \le x) \\ &= \iint_{\epsilon_{i+j} \le 2x+\epsilon_{i-j}} f(\epsilon_{i+j})f(\epsilon_{i-j})\mathrm{d}\epsilon_{i+j}\mathrm{d}\epsilon_{i-j} \\ &= \int_{-\infty}^{\infty}\{\int_{-\infty}^{2x+\epsilon_{i-j}} f(\epsilon_{i+j})\mathrm{d}\epsilon_{i+j}\}f(\epsilon_{i-j})\mathrm{d}\epsilon_{i-j} \\ &= \int_{-\infty}^{\infty} F(2x+\epsilon_{i-j})f(\epsilon_{i-j})\mathrm{d}\epsilon_{i-j}. \end{aligned}$$

Then the density of $\tilde{\zeta}_{ij}$ is

$$g(x) \triangleq \frac{\mathrm{d}F_{\tilde{\zeta}_{ij}}(x)}{\mathrm{d}x} = 2\int_{-\infty}^{\infty} f(2x+\epsilon_{i-j})f(\epsilon_{i-j})\mathrm{d}\epsilon_{i-j}.$$

The density $g(\cdot)$ is symmetric due to

$$\begin{aligned} g(-x) &= 2\int_{-\infty}^{\infty} f(-2x+\epsilon_{i-j})f(\epsilon_{i-j})\mathrm{d}\epsilon_{i-j} \\ &= 2\int_{-\infty}^{\infty} f(\epsilon_{i-j})f(\epsilon_{i-j}+2x)\mathrm{d}\epsilon_{i-j} \\ &= g(x). \end{aligned}$$

Therefore, we have

$$\begin{aligned} F_{\tilde{\zeta}_{ij}}(0) &= \int_{-\infty}^{\infty} F(\epsilon_{i-j})f(\epsilon_{i-j})\mathrm{d}\epsilon_{i-j} = \frac{1}{2}F^2(\epsilon_{i-j})\mid_{-\infty}^{\infty} = \frac{1}{2}, \\ g(0) &= 2\int_{-\infty}^{\infty} f^2(\epsilon_{i-j})\mathrm{d}\epsilon_{i-j}. \end{aligned}$$

■

Proof of Theorem 1 Rewrite the objective function as

$$S_n(b) = \frac{1}{n}\sum_j f_n\left(\tilde{Y}_{ij}|b\right),$$

where

$$f_n\left(\tilde{Y}_{ij}|b\right) = \left|\tilde{Y}_{ij}^{(1)} - b_{i1}d_j - b_{i3}d_j^3\right| \frac{1}{h} 1(0 < d_j \leq h)$$

with $b = (b_{i1}, b_{i3})^T$ and $h = k/n$. Define $X_j = \left(d_j, d_j^3\right)^T$ and $H =$diag$\{h, h^3\}$. Note that $\arg\min_b S_n(b) = \arg\min_b [S_n(b) - S_n(\beta)]$, where $\beta = \left(m^{(1)}(x_i), m^{(3)}(x_i)/6\right)^T$.

We first show that $H\left(\widehat{\beta} - \beta\right) = o_p(1)$, where $\widehat{\beta} = (\hat{\beta}_{i1}, \hat{\beta}_{i3})^T$. We use Lemma 4 of Porter and Yu (2015) to show the consistency. Essentially, we need to show that

(i) $\sup_{b\in\mathcal{B}} |S_n(b) - S_n(\beta) - \mathrm{E}[S_n(b) - S_n(\beta)]| \xrightarrow{p} 0$,

(ii) $\inf_{\|H(b-\beta)\|>\delta} \mathrm{E}[S_n(b) - S_n(\beta)] > \varepsilon$ for n large enough, where $\mathcal{B}$ is a compact parameter space for β, and δ and ε are fixed positive small numbers.

We use Lemma 2.8 of Pakes and Pollard (1989) to show (i), where

$$\mathcal{F}_n = \left\{f_n\left(\tilde{Y}|b\right) - f_n\left(\tilde{Y}|\beta\right) : b \in \mathcal{B}\right\}.$$

Note that $\mathcal{F}_n$ is Euclidean (see, e.g., Definition 2.7 of Pakes and Pollard (1989) for the definiton of an Euclidean-class of functions) by applying Lemma 2.13 of Pakes and Pollard (1989), where $\alpha = 1$, $f(\cdot, t_0) = 0$, $\phi(\cdot) = \|X_j\| \frac{1}{h} 1(0 < d_j \leq h)$ and the envelope function is $F_n(\cdot) = M\phi(\cdot)$ for some finite constant M. Since $\mathrm{E}[F_n] = \mathrm{E}\left[\|X_j\| \frac{1}{h} 1(0 < d_j \leq h)\right] = O(h) < \infty$, Lemma 2.8 of Pakes and Pollard (1989) implies

$$\sup_{b\in\mathcal{B}} |S_n(b) - S_n(\beta) - E[S_n(b) - S_n(\beta)]| \xrightarrow{p} 0.$$

As to $\inf_{\|H(b-\beta)\|>\delta} E[S_n(b) - S_n(\beta)]$, by Proposition 1 of Wang and Scott (1994),

$$\begin{aligned}
&\mathrm{E}[S_n(b) - S_n(\beta)] \\
\doteq &\frac{1}{n}\sum_j g(0)\left[X_j^T H^{-1} H(b-\beta)\right]^2 \frac{1}{h} 1(0 < d_j \leq h) \\
&-\frac{1}{n}\sum_j 2g(0)\left[\frac{m(d_{i+j}) - m(d_{i-j})}{2} - X_j^T\beta\right]\left[X_j^T H^{-1} H(b-\beta)\right]\frac{1}{h} 1(0 < d_j \leq h) \\
\gtrsim &\delta^2 - h^5\delta,
\end{aligned}$$

where $\doteq$ means that the higher-order terms are omitted, and $\gtrsim$ means the left side is bounded blow by a constant times the right side.

We then derive the asymptotic distribution of $\sqrt{nh}H\left(\widehat{\beta} - \beta\right)$ by applying the empirical process technique. First, the first order conditions can be written as

$$\frac{1}{n}\sum_j \text{sign}\left(\tilde{Y}_{ij}^{(1)} - Z_j^T H\widehat{\beta}\right) Z_j \frac{\sqrt{h}}{h} 1(0 < d_j \leq h) = o_p(1),$$

which is denoted as

$$\frac{1}{n}\sum_j f_n'(\tilde{Y}_{ij}^{(1)}|\widehat{\beta}) \triangleq S_n'\left(\widehat{\beta}\right) = o_p(1),$$

where $Z_j = H^{-1}X_j$. By Example 2.9 of Pakes and Pollard (1989), $\mathcal{F}_n'$ forms an Euclidean-class of functions with envelope $F_n' = \|Z_j\|\frac{\sqrt{h}}{h}1(0 < d_j \le h)$, where $\mathcal{F}_n' = \left\{f_n'(\tilde{Y}_{ij}^{(1)}|b) : b \in \mathcal{B}\right\}$, and $E\left[F_n'^2\right] < \infty$. So by Lemma 2.17 of Pakes and Pollard (1989) and $H\left(\widehat{\beta}-\beta\right) = o_p(1)$,

$$\mathbb{G}_n\left(f_n'(\tilde{Y}_{ij}^{(1)}|\widehat{\beta})\right) = \mathbb{G}_n\left(f_n'(\tilde{Y}_{ij}^{(1)}|\beta)\right) + o_p(1),$$

where $\mathbb{G}_n(f) = \sqrt{n}(P_n - P)f$ is the standardized empirical process, and P_n is the empirical measure of the original data. Since

$$\sqrt{n}\sum_j\left(\mathrm{E}\left[f_n'(\tilde{Y}_{ij}^{(1)}|\widehat{\beta})\right] - \mathrm{E}\left[f_n'(\tilde{Y}_{ij}^{(1)}|\beta)\right]\right) \doteq -\sqrt{nh}\frac{2g(0)}{nh}\sum_j Z_j Z_j^T H\left(\widehat{\beta}-\beta\right),$$

and

$$\begin{aligned}&\frac{1}{nh}\sum_j \mathrm{sign}\left(\tilde{Y}_{ij}^{(1)} - Z_j^T H\beta\right)Z_j 1(0 < d_j \le h)\\ -&\frac{1}{nh}\sum_j \mathrm{sign}\left(\tilde{Y}_{ij}^{(1)} - \frac{m(d_{i+j}) - m(d_{i-j})}{2}\right)Z_j 1(0 < d_j \le h) \doteq \frac{2g(0)}{nh}\sum_j Z_j d_j^5 \frac{m^{(5)}(x_i)}{5!},\end{aligned}$$

we have

$$\begin{aligned}&\sqrt{nh}\left(H\left(\widehat{\beta}-\beta\right) - \left[\frac{1}{nh}\sum_j Z_j Z_j^T\right]^{-1}\frac{1}{nh}\sum_j Z_j d_j^5\frac{m^{(5)}(x_i)}{5!}\right)\\ \doteq&\frac{1}{2g(0)}\left[\frac{1}{nh}\sum_j Z_j Z_j^T\right]^{-1}\frac{1}{\sqrt{nh}}\sum_j \mathrm{sign}\left(\tilde{\zeta}_{ij}^{(1)}\right)Z_j 1(0 < d_j \le h).\end{aligned}$$

In other words,

$$2g(0)V_k\left(\widehat{\beta}-\beta - V_k^{-2}\sum_{j=1}^k X_j d_j^5\frac{m^{(5)}(x_i)}{5!}\right) \doteq V_k^{-1}\sum_{j=1}^k \mathrm{sign}\left(\tilde{\zeta}_{ij}^{(1)}\right)X_j,$$

where $V_k = (\sum_{j=1}^k X_j X_j^T)^{1/2}$ is a symmetric positive definite matrix. By Cramér-Wold device and Lyapunov CLT, we complete the proof of asymptotic normality.

The bias of $\hat{\beta}_{i1}$ is

$$\mathrm{Bias}[\hat{\beta}_{i1}] \doteq [1,0]\left(\sum_j X_j X_j^T\right)^{-1}\sum_{j=1}^k X_j d_j^5\frac{m^{(5)}(x_i)}{5!} = -\frac{m^{(5)}(x_i)}{504}\frac{k^4}{n^4}.$$

and the variance is

$$\mathrm{Var}[\hat{\beta}_{i1}] \doteq \frac{1}{4g(0)^2}[1,0]\left(\sum_j X_j X_j^T\right)^{-1}\begin{bmatrix}1\\0\end{bmatrix} \approx \frac{75}{16g(0)^2}\frac{n^2}{k^3}.$$

Combining the squared bias and the variance, we obtain the AMSE

$$\mathrm{AMSE}[\hat{\beta}_{i1}] = \frac{m^{(5)}(x_i)^2}{504^2}\frac{k^8}{n^8} + \frac{75}{16g(0)^2}\frac{n^2}{k^3}. \tag{17}$$

To minimize (17) with respect to k, we take the first-order derivative of (17) and yield the gradient as

$$\frac{\mathrm{dAMSE}[\hat{\beta}_{i1}]}{\mathrm{d}k} = \frac{m^{(5)}(x_i)^2}{31752}\frac{k^7}{n^8} - \frac{225}{16g(0)^2}\frac{n^2}{k^4}.$$

Our optimization problem is to solve $\frac{\mathrm{dAMSE}[\hat{\beta}_{i1}]}{\mathrm{d}k} = 0$. So we obtain

$$k_{opt} = \left(\frac{893025}{2g(0)^2 m^{(5)}(x_i)^2}\right)^{1/11} n^{10/11} \approx 3.26\left(\frac{1}{g(0)^2 m^{(5)}(x_i)^2}\right)^{1/11} n^{10/11},$$

and

$$\mathrm{AMSE}[\hat{\beta}_{i1}] \approx 0.19(m^{(5)}(x_i)^6/g(0)^{16})^{1/11} n^{-8/11}.$$

Appendix B. Proof of Theorem 2

Rewrite the objective function as a U-process,

$$S_n(b) = \sum_{l<j} f_n\left(Y_{i+j}, Y_{i+l}|b\right),$$

where

$$\begin{aligned} f_n\left(Y_{i+j}, Y_{i+l}|b\right) = &\left|Y_{i+j} - Y_{i+l} - b_1(d_j - d_l) - b_{i2}(d_j^2 - d_l^2) - b_3(d_j^3 - d_l^3) - b_4(d_j^4 - d_l^4)\right| \\ &\cdot \frac{1}{h^2}1(0 < |d_j| \le h)1(0 < |d_l| \le h) \end{aligned}$$

with $b = (b_{i1}, b_{i2}, b_{i3}, b_{i4})^T$ and $h = k/n$. Define $U_n = \frac{2}{n(n-1)}S_n(b)$, $H = \mathrm{diag}\{h, h^2, h^3, h^4\}$ and $X_{jl} = \left(d_j - d_l, d_j^2 - d_l^2, d_j^3 - d_l^3, d_j^4 - d_l^4\right)^T$. Note that $\arg\min_b S_n(b) = \arg\min_b U_n(b) = \arg\min_b [U_n(b) - U_n(\beta)]$, where $\beta = \left(m^{(1)}(x_i), m^{(2)}(x_i)/2!, m^{(3)}(x_i)/3!, m^{(4)}(x_i)/4!\right)^T$.

We first show that $H\left(\widehat{\beta} - \beta\right) = o_p(1)$, where $\widehat{\beta} = (\hat{\beta}_{i1}^{\mathrm{RLowLAD}}, \hat{\beta}_{i2}^{\mathrm{RLowLAD}}, \hat{\beta}_{i3}^{\mathrm{RLowLAD}}, \hat{\beta}_{i4}^{\mathrm{RLowLAD}})^T$. We use Lemma 4 of Porter and Yu (2015) to show the consistency. Essentially, we need to show that

(i) $\sup_{b\in\mathcal{B}} |U_n(b) - U_n(\beta) - \mathrm{E}[U_n(b) - U_n(\beta)]| \xrightarrow{p} 0$,

(ii) $\inf_{\|H(b-\beta)\|>\delta} \mathrm{E}\left[U_n(b) - U_n(\beta)\right] > \varepsilon$ for n large enough, where $\mathcal{B}$ is a compact parameter space for β, and δ and ε are fixed positive small numbers.

We use Theorem A.2 of Ghosal et al. (2000) to show (i), where

$$\mathcal{F}_n = \left\{ f_n\left(Y_{i+j}, Y_{i+l}|b\right) - f_n\left(Y_{i+j}, Y_{i+l}|\beta\right) : b \in \mathcal{B} \right\}.$$

Note that $\mathcal{F}_n$ forms an Euclidean-class of functions by applying Lemma 2.13 of Pakes and Pollard (1989), where $\alpha = 1$, $f(\cdot, t_0) = 0$, $\phi(\cdot) = \|X_{jl}\| \frac{1}{h^2} 1(|d_j| \leq h) 1(|d_l| \leq h)$ and the envelope function is $F_n(\cdot) = M\phi(\cdot)$ for some finite constant M. It follows that

$$N\left(\varepsilon \left\|F_n\right\|_{Q,2}, \mathcal{F}_n, L_2(Q)\right) \lesssim \varepsilon^{-C}$$

for any probability measure Q and some positive constant C, where $\lesssim$ means the left side is bounded by a constant times the right side. Hence,

$$\frac{1}{n}\mathrm{E}\left[\int_0^\infty \log N\left(\varepsilon, \mathcal{F}_n, L_2\left(U_2^n\right)\right) d\varepsilon\right] \lesssim \frac{1}{n}\sqrt{\mathrm{E}\left[F_n^2\right]} \int_0^\infty \log \frac{1}{\varepsilon} d\varepsilon = O\left(\frac{1}{n}\right),$$

where U_2^n is the random discrete measure putting mass $\frac{1}{n(n-1)}$ on each of the points (Y_{i+j}, Y_{i+l}). Next, by Lemma A.2 of Ghosal et al. (2000), the projections

$$\overline{f}_n\left(Y_{i+j}|b\right) = \int f_n\left(Y_{i+j}, Y_{i+l}|b\right) dF_{Y_{i+l}}(Y_{i+l})$$

satisfy

$$\sup_Q N\left(\varepsilon \left\|\overline{F}_n\right\|_{Q,2}, \overline{\mathcal{F}}_n, L_2(Q)\right) \lesssim \varepsilon^{-2C},$$

where $\overline{\mathcal{F}}_n = \left\{\overline{f}_n\left(Y_{i+j}|b\right) - \overline{f}_n\left(Y_{i+j}|\beta\right) : b \in \mathcal{B}\right\}$, and $\overline{F}_n$ is an envelope of $\overline{\mathcal{F}}_n$. Thus

$$\frac{1}{\sqrt{n}}\mathrm{E}\left[\int_0^\infty \log N\left(\varepsilon, \overline{\mathcal{F}}_n, L_2\left(P_n\right)\right) d\varepsilon\right] \lesssim \frac{1}{\sqrt{n}}\sqrt{E\left[\overline{F}_n^2\right]} \int_0^\infty \log \frac{1}{\varepsilon} d\varepsilon = O\left(\frac{1}{\sqrt{n}}\right).$$

By Theorem A.2 and Markov's inequality, $\sup_{b\in\mathcal{B}} \left|U_n(b) - U_n(\beta) - \mathrm{E}\left[U_n(b) - U_n(\beta)\right]\right| \xrightarrow{p} 0$.

As to $\inf_{\|H(b-\beta)\|>\delta} E\left[U_n(b) - U_n(\beta)\right]$, by Proposition 1 of Wang and Scott (1994),

$$\begin{aligned}
&\mathrm{E}\left[U_n(b) - U_n(\beta)\right] \\
\doteq &\frac{2}{n(n-1)} \sum_{l<j} \frac{g(0)}{2} \left[X_{jl}^T H^{-1} H\left(b-\beta\right)\right]^2 \frac{1}{h^2} 1(0 < |d_j| \leq h) 1(0 < |d_l| \leq h) \\
&- \frac{2}{n(n-1)} \sum_{l<j} g(0) \left[m\left(d_{i+j}\right) - m(d_{i+l}) - X_{jl}^T \beta\right] \left[X_{jl}^T H^{-1} H\left(b-\beta\right)\right] \\
&\frac{1}{h^2} 1(0 < |d_j| \leq h) 1(0 < |d_l| \leq h) \\
\gtrsim &\delta^2 - h^5 \delta.
\end{aligned}$$

We then derive the asymptotic distribution of $\sqrt{nh}H\left(\widehat{\beta}-\beta\right)$. First, by Theorem A.1 of Ghosal et al. (2000), we approximate the first order conditions by an empirical process. Second, we derive the asymptotic distribution of $\sqrt{nh}H\left(\widehat{\beta}-\beta\right)$ by applying the empirical process technique.

First, the first order conditions can be written as

$$\frac{2}{n(n-1)}\sum_{l<j}\text{sign}\left(Y_{i+j}-Y_{i+l}-Z_{jl}^{T}H\widehat{\beta}\right)Z_{jl}\frac{\sqrt{h}}{h^{2}}1(0<|d_{j}|\leq h)1(0<|d_{l}|\leq h)=o_{p}\left(1\right),$$

which is denoted as

$$\frac{2}{n(n-1)}\sum_{l<j}f_{n}^{\prime}(Y_{i+j},Y_{i+l}|\widehat{\beta})\triangleq\frac{2}{n(n-1)}S_{n}^{\prime}\left(\widehat{\beta}\right)=o_{p}\left(1\right),$$

where $Z_{jl}=H^{-1}X_{jl}$. By Example 2.9 of Pakes and Pollard (1989), $\mathcal{F}_{n}^{\prime}$ forms an Euclidean-class of functions with envelope $F_{n}^{\prime}=\|Z_{jl}\|\frac{\sqrt{h}}{h^{2}}1(|d_{j}|\leq h)1(|d_{l}|\leq h)$, where

$$\mathcal{F}_{n}^{\prime}=\left\{f_{n}^{\prime}(Y_{i+j},Y_{i+l}|b):b\in\mathcal{B}\right\},$$

so

$$N\left(\varepsilon\left\|F_{n}^{\prime}\right\|_{Q,2},\mathcal{F}_{n}^{\prime},L_{2}\left(Q\right)\right)\lesssim\varepsilon^{-V}$$

for any probability measure Q and some positive constant V. By Theorem A.1 and the discussion following Theorem A.1 and A.2 in Ghosal et al. (2000), it follows that

$$\begin{aligned}&n\text{E}\left[\sup_{f_{n}^{\prime}\in\mathcal{F}_{n}^{\prime}}\left|U_{2}^{n}f_{n}^{\prime}-2P_{n}\left[\overline{E}_{2}\left[f_{n}^{\prime}(Y_{i+j},Y_{i+l}|b)\right]\right]-\overline{E}\left[f_{n}^{\prime}(Y_{i+j},Y_{i+l}|b)\right]\right|\right]\\ \lesssim&E\left[\int_{0}^{\infty}\log N\left(\varepsilon,\mathcal{F}_{n}^{\prime},L_{2}\left(U_{2}^{n}\right)\right)d\varepsilon\right]\lesssim\int_{0}^{1}\log\left(\varepsilon^{-V}\right)d\varepsilon\sqrt{\text{E}\left[\left(F_{n}^{\prime}\right)^{2}\right]}\lesssim h^{-1/2},\end{aligned}$$

where $\overline{E}_{2}\left[\cdot\right]$ takes expectation on Y_{i+l} and also averages over d_{l}, and $\overline{E}\left[\cdot\right]$ takes expectation on (Y_{i+j},Y_{i+l}) and also averages over (d_{j},d_{l}). As a result,

$$\sqrt{n}\sup_{f_{n}^{\prime}\in\mathcal{F}_{n}^{\prime}}\left|U_{2}^{n}f_{n}^{\prime}-2P_{n}\left[\overline{E}_{2}\left[\cdot\right]\left[f_{n}^{\prime}(Y_{i+j},Y_{i+l}|b)\right]\right]+\overline{E}\left[f_{n}^{\prime}(Y_{i+j},Y_{i+l}|b)\right]\right|=o_{p}(1),$$

which implies

$$\sqrt{n}\left(2P_{n}\left[\overline{E}_{2}\left[f_{n}^{\prime}(Y_{i+j},Y_{i+l}|\widehat{\beta})\right]\right]-\overline{E}\left[f_{n}^{\prime}(Y_{i+j},Y_{i+l}|\widehat{\beta})\right]\right)=o_{p}(1),$$

where

$$\begin{aligned}&\overline{E}_{2}\left[f_{n}^{\prime}(Y_{i+j},Y_{i+l}|b)\right]\\ =&\frac{\sqrt{h}}{nh^{2}}\sum_{l}\left[2F_{\epsilon}\left(Y_{i+j}-m(d_{i+l})-Z_{jl}^{T}Hb\right)-1\right]Z_{jl}1(0<|d_{j}|\leq h)1(0<|d_{l}|\leq h)\end{aligned}$$

with $F_\epsilon(\cdot)$ being the cumulative distribution function of ϵ. In other words,

$$2\mathbb{G}_n\left(\overline{E}_2\left[f_n'(Y_{i+j},Y_{i+l}|\widehat{\beta})\right]\right)+\sqrt{nE}\left[f_n'(Y_{i+j},Y_{i+l}|\widehat{\beta})\right]=o_p(1).$$

By Lemma 2.13 of Pakes and Pollard (1989), $\mathcal{F}'_{1n}=\left\{\overline{E}_2\left[f_n'(Y_{i+j},Y_{i+l}|b)\right]:b\in\mathcal{B}\right\}$ is Euclidean with envelope $F_{1n}=\frac{\sqrt{h}}{nh^2}\sum_l\|Z_{jl}\|\,1(0<|d_j|\le h)1(0<|d_l|\le h)$, so by Lemma 2.17 of Pakes and Pollard (1989) and $H\left(\widehat{\beta}-\beta\right)=o_p(1)$,

$$\mathbb{G}_n\left(\overline{E}_2\left[f_n'(Y_{i+j},Y_{i+l}|\widehat{\beta})\right]\right)=\mathbb{G}_n\left(\overline{E}_2\left[f_n'(Y_{i+j},Y_{i+l}|\beta)\right]\right)+o_p(1).$$

As a result,

$$\begin{aligned}
&2\mathbb{G}_n\left(\overline{E}_2\left[f_n'(Y_{i+j},Y_{i+l}|\beta)\right]\right)+\sqrt{nE}\left[f_n'(Y_{i+j},Y_{i+l}|\widehat{\beta})\right]\\
=&2\sqrt{n}P_n\left(\overline{E}_2\left[f_n'(Y_{i+j},Y_{i+l}|\beta)\right]\right)-2\sqrt{nE}\left[f_n'(Y_{i+j},Y_{i+l}|\beta)\right]\\
+&\sqrt{n}\left(\overline{E}\left[f_n'(Y_{i+j},Y_{i+l}|\widehat{\beta})\right]-\overline{E}\left[f_n'(Y_{i+j},Y_{i+l}|\beta)\right]\right)+\sqrt{nE}\left[f_n'(Y_{i+j},Y_{i+l}|\beta)\right]\\
=&2\sqrt{n}P_n\overline{E}_2\left[f_n'(Y_{i+j},Y_{i+l})\right]+2\sqrt{n}P_n\left(\overline{E}_2\left[f_n'(Y_{i+j},Y_{i+l}|\beta)\right]-\overline{E}_2\left[f_n'(Y_{i+j},Y_{i+l})\right]\right)\\
+&\sqrt{n}\left(\overline{E}\left[f_n'(Y_{i+j},Y_{i+l}|\widehat{\beta})\right]-\overline{E}\left[f_n'(Y_{i+j},Y_{i+l}|\beta)\right]\right)-\sqrt{nE}\left[f_n'(Y_{i+j},Y_{i+l}|\beta)\right]\\
=&2\sqrt{n}P_n\overline{E}_2\left[f_n'(Y_{i+j},Y_{i+l})\right]+\sqrt{n}\left(\overline{E}\left[f_n'(Y_{i+j},Y_{i+l}|\widehat{\beta})\right]-\overline{E}\left[f_n'(Y_{i+j},Y_{i+l}|\beta)\right]\right)\\
+&\sqrt{nE}\left[f_n'(Y_{i+j},Y_{i+l}|\beta)\right]\\
=&o_p(1),
\end{aligned}$$

where

$$\overline{E}_2\left[f_n'(Y_{i+j},Y_{i+l})\right]=\frac{\sqrt{h}}{nh^2}\sum_l\left[2F_\epsilon\left(\epsilon_{i+j}\right)-1\right]Z_{jl}1(0<|d_j|\le h)1(0<|d_l|\le h)$$

satisfies $E\left[\overline{E}_2\left[f_n'(Y_{i+j},Y_{i+l})\right]\right]=0$, and the second to last equality is from

$$\sqrt{n}P_n\left(\overline{E}_2\left[f_n'(Y_{i+j},Y_{i+l}|\beta)\right]-\overline{E}_2\left[f_n'(Y_{i+j},Y_{i+l})\right]\right)\doteq\sqrt{nE}\left[f_n'(Y_{i+j},Y_{i+l}|\beta)\right].$$

Since

$$\begin{aligned}
\overline{E}\left[f_n'(Y_{i+j},Y_{i+l}|b)\right]=&\frac{\sqrt{h}}{n^2h^2}\sum_{l,j}Z_{jl}1(0<|d_j|\le h)1(0<|d_l|\le h)\\
&\cdot\left[2\int F_\epsilon\left(\epsilon+m(d_{i+j})-m(d_{i+l})-Z_{jl}^THb\right)-1\right]f\left(\epsilon\right)d\epsilon,
\end{aligned}$$

we have

$$\sqrt{n}\left(\overline{E}\left[f_n'(Y_{i+j},Y_{i+l}|\widehat{\beta})\right]-\overline{E}\left[f_n'(Y_{i+j},Y_{i+l}|\beta)\right]\right)\doteq-\sqrt{nh}g(0)\left(\frac{1}{n^2h^2}\sum_{l,j}Z_{jl}Z_{jl}^T\right)H\left(\widehat{\beta}-\beta\right),$$

$$\sqrt{nE}\left[f_n'(Y_{i+j},Y_{i+l}|\beta)\right]\doteq\sqrt{nh}g(0)\frac{1}{n^2h^2}\sum_{l,j}Z_{jl}\left(d_j^5-d_l^5\right)\frac{m^{(5)}(x_i)}{5!}.$$

In summary,

$$\sqrt{nh}\left(H\left(\widehat{\beta}-\beta\right)-\left(\frac{1}{n^2h^2}\sum_{l,j}Z_{jl}Z_{jl}^T\right)^{-1}\left(\frac{1}{n^2h^2}\sum_{l,j}Z_{jl}\left(d_j^5-d_l^5\right)\right)\frac{m^{(5)}(x_i)}{5!}\right)$$
$$\doteq 2g(0)^{-1}\left(\frac{1}{n^2h^2}\sum_{l,j}Z_{jl}Z_{jl}^T\right)^{-1}\sqrt{n}P_n\overline{E}_2\left[f_n'(Y_{i+j},Y_{i+l})\right],$$

Thus, the asymptotic bias is

$$e^TH^{-1}\left(\frac{1}{n^2h^2}\sum_{l,j}Z_{jl}Z_{jl}^T\right)^{-1}\left(\frac{1}{n^2h^2}\sum_{l,j}Z_{jl}\left(d_j^5-d_l^5\right)\right)\frac{m^{(5)}(x_i)}{5!}=-\frac{m^{(5)}(x_i)}{504}\frac{k^4}{n^4},$$

and the asymptotic variance is

$$\frac{4}{kg(0)^2}e^TH^{-1}G^{-1}VG^{-1}H^{-1}e=\frac{75}{24g(0)^2}\frac{n^2}{k^3},$$

where $e=(1,0,0,0)^T$, $G=\frac{1}{k^2}\sum_{l,j}Z_{jl}Z_{jl}^T$, and $V=\frac{1}{3k}\sum_{j=-k}^{k}(\frac{1}{k}\sum_{l=-k}^{k}Z_{jl})(\frac{1}{k}\sum_{l=-k}^{k}Z_{jl})^T$ with $\mathrm{Var}\left(2F_\epsilon\left(\epsilon_{i+j}\right)-1\right)=1/3$.

Appendix C. Proof of Theorem 9

Proof of Theorem 9 Following the proof of Theorem 1, the leading term of the bias is

$$\mathrm{Bias}[\hat{\beta}_{i11}]=\frac{m^{(2)}(x_i)}{2}\frac{k^4+2k^3i-2ki^3-i^3}{n(k^3+3k^2i+3ki^2+i^3)},$$

and the leading term of the variance is

$$\mathrm{Var}[\hat{\beta}_{i11}]=\frac{3}{f(0)^2}\frac{n^2}{k^3+3k^2i+3ki^2+i^3}.$$

■

Appendix D. Proof of Theorem 10

Proposition 12 *If the errors ϵ_i are iid with a symmetric (about 0), continuous, positive density function $f(\cdot)$, then $\tilde{\delta}_{ij}=\epsilon_{i+j}+\epsilon_{i-j}$ (j=1, ..., k) are iid with $Median[\tilde{\delta}_{ij}]=0$ and a continuous, positive density $h(\cdot)$ in a neighborhood of 0, where*

$$h(x)=\int_{-\infty}^{\infty}f(x-\epsilon)f(\epsilon)\mathrm{d}\epsilon.$$

Proof of Proposition 12 The distribution of $\tilde{\tilde{\delta}}_{ij} = \epsilon_{i+j} + \epsilon_{i-j}$ is

$$\begin{aligned} F_{\tilde{\tilde{\delta}}_{ij}}(x) =& P(\epsilon_{i+j} + \epsilon_{i-j} \leq x) \\ =& \iint_{\epsilon_{i+j} \leq x - \epsilon_{i-j}} f(\epsilon_{i+j}) f(\epsilon_{i-j}) \mathrm{d}\epsilon_{i+j} \mathrm{d}\epsilon_{i-j} \\ =& \int_{-\infty}^{\infty} \{ \int_{-\infty}^{x-\epsilon_{i-j}} f(\epsilon_{i+j}) \mathrm{d}\epsilon_{i+j} \} f(\epsilon_{i-j}) \mathrm{d}\epsilon_{i-j} \\ =& \int_{-\infty}^{\infty} F(x - \epsilon_{i-j}) f(\epsilon_{i-j}) \mathrm{d}\epsilon_{i-j}. \end{aligned}$$

Then the density of $\tilde{\tilde{\delta}}_{ij}$ is

$$h(x) \triangleq \frac{\mathrm{d}F_{\tilde{\tilde{\delta}}_{ij}}(x)}{\mathrm{d}x} = \int_{-\infty}^{\infty} f(x - \epsilon_{i-j}) f(\epsilon_{i-j}) \mathrm{d}\epsilon_{i-j}.$$

By the symmetry of the density function, we have

$$\begin{aligned} F_{\tilde{\tilde{\delta}}_{ij}}(0) =& \int_{-\infty}^{\infty} F(-\epsilon_{i-j}) f(\epsilon_{i-j}) \mathrm{d}\epsilon_{i-j} \\ =& \int_{-\infty}^{\infty} (1 - F(\epsilon_{i-j})) f(\epsilon_{i-j}) \mathrm{d}\epsilon_{i-j} \\ =& (F - \frac{1}{2} F^2(\epsilon_{i-j})) \mid_{-\infty}^{\infty} \\ =& \frac{1}{2}, \\ h(0) =& \int_{-\infty}^{\infty} f^2(\epsilon_{i-j}) \mathrm{d}\epsilon_{i-j}. \end{aligned}$$

■

Proof of Theorem 10 Following the proof of Theorem 1, the asymptotic bias is

$$\mathrm{Bias}[\hat{\alpha}_{i2}] = -\frac{m^{(6)}(x_i)}{792} \frac{k^4}{n^4} + o\left(\frac{k^4}{n^4}\right),$$

and the asymptotic variance of $\hat{\alpha}_{i1}$ is

$$\mathrm{Var}[\hat{\alpha}_{i2}] = \frac{2205}{64 h(0)^2} \frac{n^4}{k^5} + o\left(\frac{n^4}{k^5}\right).$$

Combining the squared bias and the variance, we obtain the AMSE as

$$\mathrm{AMSE}[\hat{\alpha}_{i2}] = \frac{m^{(6)}(x_i)^2}{792^2} \frac{k^8}{n^8} + \frac{2205}{64 h(0)^2} \frac{n^4}{k^5}. \tag{18}$$

To minimize (18) with respect to k, we take the first-order derivative of (18) and yield the gradient as

$$\frac{\mathrm{dAMSE}[\hat{\alpha}_{i2}]}{\mathrm{d}k} = \frac{m^{(6)}(x_i)^2}{78408}\frac{k^7}{n^8} - \frac{11025}{16h(0)^2}\frac{n^4}{k^6}.$$

Now the optimization problem is to solve $\frac{\mathrm{dAMSE}[\hat{\alpha}_{i2}]}{\mathrm{d}k} = 0$. So we obtain

$$k_{opt} = \left(\frac{108056025}{8h(0)^2 m^{(6)}(x_i)^2}\right)^{1/13} n^{12/13} \approx 3.54\left(\frac{1}{h(0)^2 m^{(6)}(x_i)^2}\right)^{1/13} n^{12/13},$$

and

$$\mathrm{AMSE}[\hat{\alpha}_{i2}] \approx 0.29(m^{(6)}(x_i)^{10}/h(0)^{16})^{1/13} n^{-8/13}.$$

■

Appendix E. Variance Ratios for Popular Distributions

Variance Ratios for Eight Error Distributions

In this subsection, we investigate the variance ratio of the RLowLAD estimator with respect to the LowLSR and LAD estimators for eight error distributions.

From the main text,

$$\mathrm{R}_{\mathrm{LowLSR}}(f) = 3\sigma^2 g(0)^2, \quad \mathrm{R}_{\mathrm{LAD}}(f) = \frac{3g(0)^2}{4f(0)^2}.$$

Example 1: Normal distribution. The error density function is $f(\epsilon) = \frac{1}{\sqrt{2\pi}}\exp(-\epsilon^2/2)$, which implies

$$f(0) = \frac{1}{\sqrt{2\pi}}, \quad g(0) = 2\int_{-\infty}^{\infty}\frac{1}{2\pi}e^{-\epsilon^2}\mathrm{d}\epsilon = \frac{1}{\sqrt{\pi}}.$$

Due to $\sigma^2 = 1$, we have

$$\mathrm{R}_{\mathrm{LowLSR}}(f) = 3/\pi \approx 0.95, \quad \mathrm{R}_{\mathrm{LAD}}(f) = 1.50.$$

In other words, the RLowLAD estimator is almost as efficient as the LowLSR estimator for the normal distribution.

Example 2: Mixed normal distribution. The error density function is

$$f(\epsilon;\alpha,\sigma_0) = (1-\alpha)\frac{1}{\sqrt{2\pi}}e^{-\epsilon^2/2} + \alpha\frac{1}{\sqrt{2\pi}\sigma_0}e^{-\epsilon^2/(2\sigma_0^2)}$$

with $0 < \alpha \leq 1/2$ and $\sigma_0 > 1$, which implies

$$\begin{aligned}
f(0;\alpha,\sigma_0) =&(1-\alpha)\frac{1}{\sqrt{2\pi}} + \alpha\frac{1}{\sqrt{2\pi}\sigma_0},\\
g(0;\alpha,\sigma_0) =&2\int_{-\infty}^{\infty} f^2(\epsilon)\mathrm{d}\epsilon\\
=&2\left\{(1-\alpha)^2\int_{-\infty}^{\infty}\frac{1}{2\pi}\mathrm{e}^{-\epsilon^2}\mathrm{d}\epsilon + 2\alpha(1-\alpha)\int_{-\infty}^{\infty}\frac{1}{2\pi\sigma_0}\mathrm{e}^{-(\frac{\epsilon^2}{2}+\frac{\epsilon^2}{2\sigma_1^2})}\mathrm{d}\epsilon\right.\\
&\left.+\alpha^2\int_{-\infty}^{\infty}\frac{1}{2\pi\sigma_0^2}\mathrm{e}^{-\frac{x^2}{\sigma_0^2}}\mathrm{d}\epsilon\right\}\\
=&2\left\{(1-\alpha)^2\frac{1}{2\sqrt{\pi}} + 2\alpha(1-\alpha)\frac{1}{\sqrt{2\pi}\sqrt{1+\sigma_0^2}} + \alpha^2\frac{1}{2\sqrt{\pi}\sigma_0}\right\},\\
\mathrm{Var}(\epsilon_i) =&(1-\alpha)+\alpha\sigma_0^2 \triangleq \sigma^2.
\end{aligned}$$

Thus,

$$\begin{aligned}
\mathrm{R}_{\mathrm{LowLSR}}(\alpha,\sigma_0) =&12\left\{(1-\alpha)+\alpha\sigma_0^2\right\}\left\{(1-\alpha)^2\frac{1}{2\sqrt{\pi}} + 2\alpha(1-\alpha)\frac{1}{\sqrt{2\pi}\sqrt{1+\sigma_0^2}} + \alpha^2\frac{1}{2\sqrt{\pi}\sigma_0}\right\}^2,\\
\mathrm{R}_{\mathrm{LAD}}(\alpha,\sigma_0) =&\frac{3\left\{(1-\alpha)^2\frac{1}{2\sqrt{\pi}} + 2\alpha(1-\alpha)\frac{1}{\sqrt{2\pi}\sqrt{1+\sigma_0^2}} + \alpha^2\frac{1}{2\sqrt{\pi}\sigma_0}\right\}^2}{\left\{(1-\alpha)\frac{1}{\sqrt{2\pi}} + \alpha\frac{1}{\sqrt{2\pi}\sigma_0}\right\}^2}.
\end{aligned}$$

In particular,

$$\begin{aligned}
\mathrm{R}_{\mathrm{LowLSR}}(0.1,3) \approx&1.80, \quad \mathrm{R}_{\mathrm{LowLSR}}(0.1,10) \approx 10.90,\\
\mathrm{R}_{\mathrm{LAD}}(0.1,3) \approx&1.38, \quad \mathrm{R}_{\mathrm{LAD}}(0.1,10) \approx 1.27.
\end{aligned}$$

Example 3: t distribution. The error density function is

$$f(\epsilon;\nu) = \frac{\Gamma((\nu+1)/2)}{\sqrt{\nu\pi}\Gamma(\nu/2)}\left(1+\frac{\epsilon^2}{\nu}\right)^{-(\nu+1)/2)}$$

with the degree of freedom $\nu > 2$, which implies

$$\begin{aligned}
f(0) =&\frac{\Gamma((\nu+1)/2)}{\sqrt{\nu\pi}\Gamma(\nu/2)},\\
g(0) =&2\int_{-\infty}^{\infty}\frac{1}{\nu\pi}\left(\frac{\Gamma((\nu+1)/2)}{\Gamma(\nu/2)}\right)^2\left(1+\frac{\epsilon^2}{\nu}\right)^{-(\nu+1)}\mathrm{d}\epsilon = \frac{2}{\sqrt{\nu\pi}}\left(\frac{\Gamma((\nu+1)/2)}{\Gamma(\nu/2)}\right)^2\frac{\Gamma(\nu+1/2)}{\Gamma(\nu+1)}.
\end{aligned}$$

Due to $\sigma^2 = \nu/(\nu-2)$, we have

$$\begin{aligned}
\mathrm{R}_{\mathrm{LowLSR}}(\nu) =&\frac{12}{(\nu-2)\pi}\left(\frac{\Gamma((\nu+1)/2)}{\Gamma(\nu/2)}\right)^4\left(\frac{\Gamma(\nu+1/2)}{\Gamma(\nu+1)}\right)^2,\\
\mathrm{R}_{\mathrm{LAD}}(\nu) =&3\left(\frac{\Gamma((\nu+1)/2)}{\Gamma(\nu/2)}\right)^2\left(\frac{\Gamma(\nu+1/2)}{\Gamma(\nu+1)}\right)^2.
\end{aligned}$$

For $\nu = 3$,

$$\mathrm{R}_{\mathrm{LowLSR}}(3) = 75/(4\pi^2) \approx 1.90, \quad \mathrm{R}_{\mathrm{LAD}}(3) = 75/64 \approx 1.17.$$

Example 4: Laplace (double exponential) distribution. The error density function is $f(\epsilon) = \frac{1}{2}\exp(-|\epsilon|)$, which implies

$$f(0) = \frac{1}{2}, \quad g(0) = 2\int_{-\infty}^{\infty} \frac{1}{4} e^{-2|\epsilon|} \mathrm{d}\epsilon = \frac{1}{2}.$$

Due to $\sigma^2 = 2$, we have

$$\mathrm{R}_{\mathrm{LowLSR}}(f) = 1.50, \quad \mathrm{R}_{\mathrm{LAD}}(f) = 0.75.$$

Example 5: Logistic distribution. The error density function is $f(\epsilon) = \exp(\epsilon)/(\exp(\epsilon) + 1)^2$, which implies

$$f(0) = \frac{1}{4}, \quad g(0) = 2\int_{-\infty}^{\infty} \frac{e^{2\epsilon}}{(\exp(\epsilon)+1)^4} \mathrm{d}\epsilon = \frac{1}{3}.$$

Due to $\sigma^2 = \pi^2/3$, we have

$$\mathrm{R}_{\mathrm{LowLSR}}(f) = \pi^2/9 \approx 1.10, \quad \mathrm{R}_{\mathrm{LAD}}(f) = 4/3 \approx 1.33.$$

Example 6: Cauchy distribution. The error density function is $f(\epsilon) = 1/(\pi(1+\epsilon^2))$, which implies

$$f(0) = \frac{1}{\pi}, \quad g(0) = \frac{1}{\pi}, \quad \mathrm{Var}(\epsilon) = \infty.$$

Thus,

$$\mathrm{R}_{\mathrm{LowLSR}}(f) = \infty, \quad \mathrm{R}_{\mathrm{LAD}}(f) = 0.75.$$

Example 7: Mixed double Gamma distribution. The error density function is

$$f(\epsilon; \alpha, k) = (1-\alpha)\frac{1}{2}e^{-|\epsilon|} + \alpha \frac{1}{2\Gamma(k+1)}|\epsilon|^k e^{-|\epsilon|}$$

with parameter $k > 0$ and the mixed ratio α, which implies

$$\begin{aligned}
f(0;\alpha,k) =& \frac{1-\alpha}{2} + \frac{\alpha}{2\Gamma(k+1)}, \\
g(0;\alpha,k) =& 2\int_{-\infty}^{\infty} f^2(\epsilon;\alpha,k)\mathrm{d}\epsilon \\
=& \int_{-\infty}^{\infty} \frac{(1-\alpha)^2}{2} e^{-2|\epsilon|} + \frac{(1-\alpha)\alpha}{\Gamma(k+1)}|\epsilon|^k e^{-2|\epsilon|} + \frac{\alpha^2}{2\Gamma(k+1)^2}|\epsilon|^{2k} e^{-2|\epsilon|}\mathrm{d}\epsilon \\
=& \frac{(1-\alpha)^2}{2} + \frac{\alpha(1-\alpha)}{2^k} + \frac{\alpha^2\Gamma(2k+1)}{2^{2k+1}\Gamma^2(k+1)}, \\
\mathrm{Var}(\epsilon|\alpha,k) =& 1 + \alpha k.
\end{aligned}$$

Thus,

$$\mathrm{R}_{\mathrm{LowLSR}}(\alpha,k) = 3(1+\alpha k)\left\{\frac{(1-\alpha)^2}{2}+\frac{\alpha(1-\alpha)}{2^k}+\frac{\alpha^2\Gamma(2k+1)}{2^{2k+1}\Gamma^2(k+1)}\right\},$$

$$\mathrm{R}_{\mathrm{LAD}}(\alpha,k) = 3\left\{\frac{(1-\alpha)^2}{2}+\frac{\alpha(1-\alpha)}{2^k}+\frac{\alpha^2\Gamma(2k+1)}{2^{2k+1}\Gamma^2(k+1)}\right\}^2 \Bigg/ \left\{4\left(\frac{1-\alpha}{2}+\frac{\alpha}{2\Gamma(k+1)}\right)^2\right\}.$$

In particular,

$$\mathrm{R}_{\mathrm{LowLSR}}(0.1,3) \approx 1.63, \quad \mathrm{R}_{\mathrm{LowLSR}}(0.1,10) \approx 2.44,$$
$$\mathrm{R}_{\mathrm{LAD}}(0.1,3) \approx 0.68, \quad \mathrm{R}_{\mathrm{LAD}}(0.1,10) \approx 0.68.$$

Example 8: Bimodal distribution (mixed normal distribution with different locations). The error density function is

$$f(\epsilon;\mu) = 0.5\frac{1}{\sqrt{2\pi}}e^{-(\epsilon-\mu)^2/2} + 0.5\frac{1}{\sqrt{2\pi}}e^{-(\epsilon+\mu)^2/2}$$

with $\mu > 0$, which implies

$$f(0;\mu) = \frac{e^{-\mu^2}}{\sqrt{2\pi}}, \quad g(0;\mu) = \frac{1+e^{-\mu^2}}{2\sqrt{\pi}}, \quad \sigma^2 = 1+\mu^2.$$

Thus,

$$\mathrm{R}_{\mathrm{LowLSR}}(\mu) = 3(1+\mu^2)\frac{1+e^{-\mu^2}}{2\sqrt{\pi}}, \quad \mathrm{R}_{\mathrm{LAD}}(\mu) = 3\left(\frac{1+e^{-\mu^2}}{2\sqrt{\pi}}\right)^2 \Bigg/ \left(4\left(\frac{e^{-\mu^2}}{\sqrt{2\pi}}\right)^2\right).$$

In particular,

$$\mathrm{R}_{\mathrm{LowLSR}}(1) \approx 0.89, \quad \mathrm{R}_{\mathrm{LowLSR}}(3) \approx 2.39,$$
$$\mathrm{R}_{\mathrm{LAD}}(1) \approx 5.18, \quad \mathrm{R}_{\mathrm{LAD}}(3) \approx 2.46\times 10^7.$$

Variance Ratio Functions for Three Error Distributions

To further illustrate the trade-off between the sharp-peak and heavy-tailed errors, we consider three of the above examples: Examples 2, 3, and 8.

For the mixed normal distribution, we list in Table 4 the critical value of σ_0 such that the LowLAD (or RLowLAD) and LowLSR estimators have the same variance. When σ_0 is smaller than the critical value, the LowLAD (or RLowLAD) estimator is more efficient than the LowLSR estimator. The overall comparison curve is given in Figure 12. Since LowLAD and RLowLAD have a close relationship, we consider only RLowLAD in the following comparisons. Another critical σ_0 curve comparing RLowLAD and LAD is provided in Figure 13. When σ_0 is larger than the critical value, the RLowLAD estimator is more efficient than the LAD estimator.

For the $t(\nu)$ distribution, the variance ratio function between LowLSR and RLowLAD is shown in Figure 14. We see that the ratio between LowLSR and RLowLAD is greater

than 1 for small degrees of freedom (from 3 to 18). As ν increases, the ratio converges to $3/\pi \approx 0.95$. This phenomenon is expected because the T distribution converges to the normal distribution as $\nu \to \infty$, and the variance ratio for the normal distribution is 0.95. Figure 15 shows the ratio between LAD and RLowLAD, where we see that this ratio is greater than 1 for all degrees of freedom. As ν increases, the ratio converges to 1.50. This is expected from the stirling formula $\Gamma(\nu) = \sqrt{2\pi}e^{-\nu}\nu^{\nu-1/2}$, which implies $\mathrm{R}_{\mathrm{LAD}}(f) \to 1.50$ as $\nu \to \infty$.

For the bimodal distribution, the variance ratio function between LowLSR and RLowLAD is given in Figure 16. As μ increases, the ratio becomes smaller and achieves the minimum value 0.89 at point 1.12, and then becomes larger and tends to ∞. The variance ratio function between LAD and RLowLAD is shown in Figure 17. As μ increases, the ratio diverges fast to ∞.

Appendix F. One Key Difference between LS and LAD Methods

For the LS method, the LS estimator and the LowLSR estimator are asymptotically equivalent; while for the LAD method, the asymptotic variances of the LAD estimator and the LowLAD estimator are very different, although their asymptotic biases are the same. To understand this discrepancy, we show that the objective functions of the LS and LowLSR estimation are asymptotically equivalent while those of the LAD and LowLAD estimation are not. Note that the objective function of the LowLSR estimation is equivalent to

$$
\begin{aligned}
&4\sum_{j=1}^{k}\left(\tilde{Y}_{ij}^{(1)} - \alpha_{i1}d_j - \alpha_{i3}d_j^3\right)^2 \\
&= \sum_{j=1}^{k}\left(Y_{i+j} - Y_{i-j} - 2\alpha_{i1}d_j - 2\alpha_{i3}d_j^3\right)^2 \\
&= \sum_{j=1}^{k}\left(Y_{i+j} - Y_{i-j} - (\alpha_{i0} - \alpha_{i0}) - \alpha_{i1}\left(d_j - d_{-j}\right) - \alpha_{i2}\left(d_j^2 - d_{-j}^2\right) - \alpha_{i3}\left(d_j^3 - d_{-j}^3\right) - \alpha_{i4}\left(d_j^4 - d_{-j}^4\right)\right)^2 \\
&= \sum_{j=1}^{k}\left(Y_{i+j} - \alpha_{i0} - \alpha_{i1}d_j - \alpha_{i2}d_j^2 - \alpha_{i3}d_j^3 - \alpha_{i4}d_j^4\right)^2 \\
&+ \sum_{j=1}^{k}\left(Y_{i-j} - \alpha_{i0} - \alpha_{i1}d_{-j} - \alpha_{i2}d_{-j}^2 - \alpha_{i3}d_{-j}^3 - \alpha_{i4}d_{-j}^4\right)^2 \\
&-2\sum_{j=1}^{k}\left(Y_{i+j} - \alpha_{i0} - \alpha_{i1}d_j - \alpha_{i2}d_j^2 - \alpha_{i3}d_j^3 - \alpha_{i4}d_j^4\right)\left(Y_{i-j} - \alpha_{i0} - \alpha_{i1}d_{-j} - \alpha_{i2}d_{-j}^2 - \alpha_{i3}d_{-j}^3 - \alpha_{i4}d_{-j}^4\right),
\end{aligned}
$$

where $d_{-j} = -j/n$. It can be shown that the cross term in the last equality is a higher-order term and is negligible, and the first two terms constitute exactly the objective function of the least squares estimation. More specifically, we can show that

$$
\begin{aligned}
\hat{m}_{\mathrm{LS}}^{(1)}(x_i) - m^{(1)}(x_i) - \mathrm{Bias} &\doteq (0,1,0,0,0)\left(\sum_{j=-k}^{k} X_j X_j'\right)^{-1}\sum_{j=-k}^{k} X_j\epsilon_{i+j} \\
&= \sum_{j=-k}^{k} D_j\epsilon_{i+j}
\end{aligned}
$$

with $X_j = \left(1, d_j, d_j^2, d_j^3, d_j^4\right)^T$ and

$$
D_j = n\frac{j(75k^4 + 150k^3 - 75k + 25) - j^3\left(105k^2 + 105k - 35\right)}{k(8k^6 + 28k^5 + 14k^4 - 35k^3 - 28k^2 + 7k + 6)} = -D_{-j},
$$

and

$$\begin{aligned}
&\hat{m}^{(1)}_{\text{LowLSR}}(x_i) - m^{(1)}(x_i) - \text{Bias} \\
&\doteq (1,0)\left[\begin{pmatrix} 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \end{pmatrix}\left(\sum_{j=1}^k X_j X_j'\right)\begin{pmatrix} 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \end{pmatrix}^T\right]^{-1} \\
&\quad\cdot\begin{pmatrix} 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \end{pmatrix}\sum_{j=1}^k X_j \tfrac{\epsilon_{i+j}-\epsilon_{i-j}}{2} \\
&= \sum_{j=1}^k 2D_j \tfrac{\epsilon_{i+j}-\epsilon_{i-j}}{2} = \sum_{j=-k}^k D_j \epsilon_{i+j}.
\end{aligned}$$

The influence functions of $\hat{m}^{(1)}_{\text{LS}}(x_i)$ and $\hat{m}^{(1)}_{\text{LowLSR}}(x_i)$ are exactly the same, so the contribution of the cross term is null.

On the contrary, the objective function of the LowLAD estimator is equivalent to

$$\begin{aligned}
&2\sum_{j=1}^k \left|\tilde{Y}^{(1)}_{ij} - \beta_{i1}d_j - \beta_{i3}d_j^3\right| \\
&= \sum_{j=1}^k \left|Y_{i+j} - Y_{i-j} - (\beta_{i0}-\beta_{i0}) - \beta_{i1}(d_j - d_{-j}) - \beta_{i2}\left(d_j^2 - d_{-j}^2\right) - \beta_{i3}\left(d_j^3 - d_{-j}^3\right) - \beta_{i4}\left(d_j^4 - d_{-j}^4\right)\right| \\
&= \sum_{j=1}^k \left|Y_{i+j} - \beta_{i0} - \beta_{i1}d_j - \beta_{i2}d_j^2 - \beta_{i3}d_j^3 - \beta_{i4}d_j^4\right| + \sum_{j=1}^k \left|Y_{i-j} - \beta_{i0} - \beta_{i1}d_{-j} - \beta_{i2}d_{-j}^2 - \beta_{i3}d_{-j}^3 - \beta_{i4}d_{-j}^4\right| \\
&\quad + \text{extra term},
\end{aligned}$$

where the extra term is equal to

$$\begin{aligned}
&-2\max\Bigg(\text{sign}\left(Y_{i+j} - \beta_{i0} - \beta_{i1}d_j - \beta_{i2}d_j^2 - \beta_{i3}d_j^3 - \beta_{i4}d_j^4\right) \\
&\cdot\text{sign}\left(Y_{i-j} - \beta_{i0} - \beta_{i1}d_{-j} - \beta_{i2}d_{-j}^2 - \beta_{i3}d_{-j}^3 - \beta_{i4}d_{-j}^4\right), 0\Bigg) \\
&\cdot\min\left(\left|Y_{i+j} - \beta_{i0} - \beta_{i1}d_j - \beta_{i2}d_j^2 - \beta_{i3}d_j^3 - \beta_{i4}d_j^4\right|, \left|Y_{i-j} - \beta_{i0} - \beta_{i1}d_{-j} - \beta_{i2}d_{-j}^2 - \beta_{i3}d_{-j}^3 - \beta_{i4}d_{-j}^4\right|\right),
\end{aligned}$$

and cannot be neglected, where $sign(x) = 1$ if $x > 0$ and $sign(x) = -1$ if $x < 0$. More specifically, we can show that

$$\begin{aligned}
&\hat{m}^{(1)}_{\text{LAD}}(x_i) - m^{(1)}(x_i) - \text{Bias} \\
&\doteq (0,1,0,0,0)\left(2f(0)\sum_{j=-k}^k X_j X_j'\right)^{-1}\sum_{j=-k}^k X_j \text{sign}(\epsilon_{i+j}) \\
&= \tfrac{1}{2f(0)}\sum_{j=-k}^k D_j \text{sign}(\epsilon_{i+j}) = \tfrac{1}{f(0)}\sum_{j=1}^k D_j \tfrac{\text{sign}(\epsilon_{i+j})-\text{sign}(\epsilon_{i-j})}{2},
\end{aligned}$$

and

$$\begin{aligned}
&\hat{m}^{(1)}_{\text{LowLAD}}(x_i) - m^{(1)}(x_i) - \text{Bias} \\
&\doteq (1,0)\left[2g(0)\begin{pmatrix} 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \end{pmatrix}\left(\sum_{j=1}^k X_j X_j'\right)\begin{pmatrix} 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \end{pmatrix}^T\right]^{-1} \\
&\quad\cdot\begin{pmatrix} 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \end{pmatrix}\sum_{j=1}^k X_j \text{sign}\left(\tfrac{\epsilon_{i+j}-\epsilon_{i-j}}{2}\right) \\
&= \tfrac{1}{g(0)}\sum_{j=1}^k D_j \text{sign}\left(\tfrac{\epsilon_{i+j}-\epsilon_{i-j}}{2}\right).
\end{aligned}$$

So the contribution of the extra term to the influence function is

$$D_j\left[\frac{\text{sign}\left(\frac{\epsilon_{i+j}-\epsilon_{i-j}}{2}\right)}{g(0)} - \frac{\text{sign}(\epsilon_{i+j})-\text{sign}(\epsilon_{i-j})}{2f(0)}\right] = \frac{D_j}{g(0)}\begin{cases} \text{sign}(\epsilon_{i+j}-\epsilon_{i-j}), \text{ if } \epsilon_{i+j}\epsilon_{i-j} > 0, \\ \left(1 - \frac{g(0)}{f(0)}\right)\text{sign}(\epsilon_{i+j}), \text{ if } \epsilon_{i+j}\epsilon_{i-j} < 0, \end{cases}$$

which is not null, where we neglect the event that $\epsilon_{i+j}\epsilon_{i-j} = 0$ because the probability of such an event is zero.

References

G. Boente and D. Rodriguez. Robust estimators of high order derivatives of regression functions. *Statistics & Probability Letters*, 76(13):1335–1344, 2006.

L.D. Brown and M. Levine. Variance estimation in nonparametric regression via the difference sequence method. *The Annals of Statistics*, 35(5):2219–2232, 2007.

J.L.O. Cabrera. locpol: Kernel local polynomial regression. R package version 0.6-0, 2012. URL `http://mirrors.ustc.edu.cn/CRAN/web/packages/locpol/index.html`.

R. Charnigo, M. Francoeur, M.P. Mengüç, A. Brock, M. Leichter, and C. Srinivasan. Derivatives of scattering profiles: tools for nanoparticle characterization. *Journal of the Optical Society of America*, 24(9):2578–2589, 2007.

R. Charnigo, M. Francoeur, M.P. Mengüç, B. Hall, and C. Srinivasan. Estimating quantitative features of nanoparticles using multiple derivatives of scattering profiles. *Journal of Quantitative Spectroscopy and Radiative Transfer*, 112(8):1369–1382, 2011a.

R. Charnigo, B. Hall, and C. Srinivasan. A generalized C_p criterion for derivative estimation. *Technometrics*, 53(3):238–253, 2011b.

P. Chaudhuri and J.S. Marron. SiZer for exploration of structures in curves. *Journal of the American Statistical Association*, 94(447):807–823, 1999.

W.S. Cleveland. Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association*, 74(368):829–836, 1979.

K. De Brabanter, J. De Brabanter, B. De Moor, and I. Gijbels. Derivative estimation with local polynomial fitting. *Journal of Machine Learning Research*, 14(1):281–301, 2013.

M. Delecroix and A.C. Rosa. Nonparametric estimation of a regression function and its derivatives under an ergodic hypothesis. *Journal of Nonparametric Statistics*, 6(4):367–382, 1996.

N.R. Draper and H. Smith. *Applied Regression Analysis*. Wiley and Sons, New York, 2nd edition, 1981.

J. Fan and I. Gijbels. *Local Polynomial Modelling and Its Applications*. Chapman & Hall, London, 1996.

J. Fan and P. Hall. On curve estimation by minimizing mean absolute deviation and its implications. *The Annals of Statistics*, 22(2):867–885, 1994.

S. Ghosal, A. Sen, and A.W. vander Vaart. Testing monotonicity of regression. *The Annals of Statistics*, 28(4):1054–1082, 2000.

I. Gijbels and A.C. Goderniaux. Data-driven discontinuity detection in derivatives of a regression function. *Communications in Statistics-Theory and Methods*, 33(4):851–871, 2005.

B. Hall. *Nonparametric Estimation of Derivatives with Applications*. PhD thesis, University of Kentucky, Lexington, Kentucky, 2010.

W. Härdle. *Applied Nonparametric Regression*. Cambridge University Press, Cambridge, 1990.

W. Härdle and T. Gasser. Robust non-parametric function fitting. *Journal of the Royal Statistical Society, Series B*, 46(1):42–51, 1984.

W. Härdle and T. Gasser. On robust kernel estimation of derivatives of regression functions. *Scandinavian Journal of Statistics*, 12(3):233–240, 1985.

P.J. Huber and E.M. Ronchetti. *Robust Statistics*. John Wiley & Sons, Inc., 2009.

B. Kai, R. Li, and H. Zou. Local composite quantile regression smoothing: an efficient and sate alternative to local polynomial regression. *Journal of the Royal Statistical Society, Series B*, 72(1):49–69, 2010.

G.D. Knott. *Interpolating Cubic Splines*. Spring, 1st edition, 2000.

R. Koenker. A note on l-estimation for linear models. *Statistics & Probability Letters*, 2(6): 323–325, 1984.

R. Koenker. *Quantile Regression*. Cambridge University Press, New York, 2005.

R. Koenker and G. Bassett. Regression quantiles. *Econometrica*, 46(1):33–50, 1978.

X.R. Li and V.P. Jilkov. Survey of maneuvering target tracking. Part I: Dynamic models. *IEEE Transations on Aerospace and Electronic Systems*, 39(4):1333–1364, 2003.

X.R. Li and V.P. Jilkov. Survey of maneuvering target tracking. Part II: Motion models of ballistic and space targets. *IEEE Transations on Aerospace and Electronic Systems*, 46 (1):96–119, 2010.

Y. Liu and K. De Brabanter. Derivative estimation in random design. 32nd Conference on Neural Information Processing Systems (NeurIPS 2018), Montréal, Canada., 2018.

I. Matyasovszky. Detecting abrupt climate changes on different time scales. *Theoretical and Applied Climatology*, 105(3-4):445–454, 2011.

H.G. Müller. *Nonparametric Regression Analysis of Longitudinal Data*. Springer, New York, 1988.

H.G. Müller, U. Stadtmüller, and T. Schmitt. Bandwidth choice and confidence intervals for derivatives of noisy data. *Biometrika*, 74(4):743–749, 1987.

J. Newell, J. Einbeck, N. Madden, and K. McMillan. Model free endurance markers based on the second derivative of blood lactate curves. In *Proceedings of the 20th International Workshop on Statistical Modelling*, pages 357–364, Sydney, 2005.

F. Osorio. L1pack: Routines for $L1$ estimation. R package version 0.3, 2015. URL `http://www.ies.ucv.cl/l1pack/`.

A. Pakes and D. Pollard. Simulation and the asymptotics of optimization estimators. *Econometrica*, 57(5):1027–1057, 1989.

C. Park and K.H Kang. SiZer analysis for the comparison of regression curves. *Computational Statistics & Data Analysis*, 52(8):3954–3970, 2008.

J. Porter and P. Yu. Regression discontinuity designs with unknown discontinuity points: Testing and estimation. *Journal of Econometrics*, 189(1):132–147, 2015.

J. Ramsay and B. Ripley. pspline: Penalized smoothing splines. R package version 1.0-16, 2013. URL `http://mirrors.ustc.edu.cn/CRAN/web/packages/pspline/index.html`.

J.O. Ramsay and B.W. Silverman. *Applied Functional Data Analysis: Methods and Case Studies*. Springer, New York, 2002.

D. Ruppert and M.P. Wand. Multivariate locally weighted least squares regression. *The Annals of Statistics*, 22(3):1346–1370, 1994.

C.J. Stone. Additive regression and other nonparametric models. *The Annals of Statistics*, 13(2):689–705, 1985.

P.S. Swain, K. Stevenson, A. Leary, L.F. Montano-Gutierrez, I.B.N. Clark, J. Vogel, and T. Pilizota. Inferring time derivatives including cell growth rates using gaussian process. *Nature Communications*, 7: 13766, 2016.

G. Wahba and Y. Wang. When is the optimal regularization parameter insensitive to the choice of the loss function? *Communications in Statistics-Theory and Methods*, 19(5): 1685–1700, 1990.

F.T. Wang and D.W. Scott. The L_1 method for robust nonparametric regression. *Journal of the American Statistical Association*, 89(425):65–76, 1994.

W.W. Wang and L. Lin. Derivative estimation based on difference sequence via locally weighted least squares regression. *Journal of Machine Learning Research*, 16:2617–2641, 2015.

W.W. Wang and P. Yu. Asymptotically optimal differenced estimators of error variance in nonparametric regression. *Computational Statistics & Data Analysis*, 105:125–143, 2017.

A.H. Welsh. Robust estimation of smooth regression and spread functions and their derivatives. *Statistica Sinica*, 6(2):347–366, 1996.

Z. Zhao and Z. Xiao. Efficient regression via optimally combining quantile information. *Econometric Theory*, 30(6):1272–1314, 2014.

S. Zhou and D.A. Wolfe. On derivative estimation in spline regression. *Statistica Sinica*, 10 (1):93–108, 2000.

H. Zou and M. Yuan. Composite quantile regression and the oracle model selection theory. *The Annals of Statistics*, 36(3):1108–1126, 2008.

			LowLAD	**RLowLAD**	**LowLSR**	**LAD**	**LowLAD**	**RLowLAD**	**LowLSR**	**LAD**
A	f	σ	$n = 100$	$n = 100$	$n = 100$	$n = 100$	$n = 500$	$n = 500$	$n = 500$	$n = 500$
						$\sigma_0 = 1$				
1	0.5	0.1	0.41(0.10)	0.34(0.10)	0.83(0.32)	0.37(0.10)	0.18(0.04)	0.15(0.04)	0.40(0.12)	0.17(0.05)
		0.5	1.79(0.38)	1.49(0.43)	1.51(0.45)	1.76(0.46)	0.82(0.15)	0.67(0.19)	0.70(0.19)	0.81(0.22)
	1	0.1	0.41(0.10)	0.34(0.10)	0.84(0.32)	0.37(0.10)	0.19(0.04)	0.15(0.04)	0.40(0.12)	0.17(0.05)
		0.5	1.78(0.37)	1.49(0.41)	1.51(0.43)	1.74(0.45)	0.82(0.15)	0.68(0.18)	0.70(0.19)	0.81(0.21)
10	0.5	0.1	0.41(0.09)	0.35(0.10)	0.85(0.33)	0.38(0.10)	0.18(0.03)	0.15(0.04)	0.39(0.12)	0.17(0.04)
		0.5	1.78(0.36)	1.47(0.40)	1.48(0.42)	1.75(0.45)	0.79(0.15)	0.64(0.18)	0.66(0.18)	0.75(0.21)
	1	0.1	0.45(0.09)	0.38(0.10)	0.85(0.31)	0.41(0.10)	0.24(0.04)	0.22(0.04)	0.43(0.12)	0.23(0.05)
		0.5	1.79(0.36)	1.50(0.40)	1.53(0.43)	1.76(0.45)	0.82(0.15)	0.67(0.18)	0.68(0.18)	0.80(0.21)
						$\sigma_0 = 10$				
1	0.5	0.1	0.46(0.16)	0.39(0.14)	7.79(3.15)	0.39(0.11)	0.19(0.04)	0.15(0.04)	3.78(1.17)	0.17(0.04)
		0.5	2.11(0.51)	1.76(0.53)	7.88(3.04)	1.89(0.50)	0.93(0.19)	0.76(0.22)	3.84(1.17)	0.85(0.22)
	1	0.1	0.45(0.14)	0.38(0.14)	7.93(3.33)	0.39(0.11)	0.19(0.04)	0.15(0.04)	3.77(1.13)	0.17(0.04)
		0.5	2.15(0.49)	1.81(0.53)	8.00(3.16)	1.93(0.50)	0.92(0.18)	0.75(0.21)	3.81(1.13)	0.83(0.22)
10	0.5	0.1	0.46(0.14)	0.38(0.13)	8.03(3.38)	0.39(0.11)	0.18(0.04)	0.15(0.05)	3.66(1.26)	0.17(0.04)
		0.5	2,12(0.50)	1.77(0.52)	8.05(3.24)	1.89(0.50)	0.94(0.17)	0.77(0.20)	3.94(1.16)	0.85(0.23)
	1	0.1	0.49(0.13)	0.42(0.13)	7.95(3.09)	0.42(0.10)	0.25(0.04)	0.22(0.04)	3.84(1.13)	0.23(0.04)
		0.5	2.17(0.52)	1.82(0.54)	8.07(3.16)	1.92(0.51)	0.94(0.18)	0.77(0.22)	3.80(1.12)	0.86(0.23)

Table 3: Simulation comparison among LowLAD, RLowLAD, LowLSR and LAD.

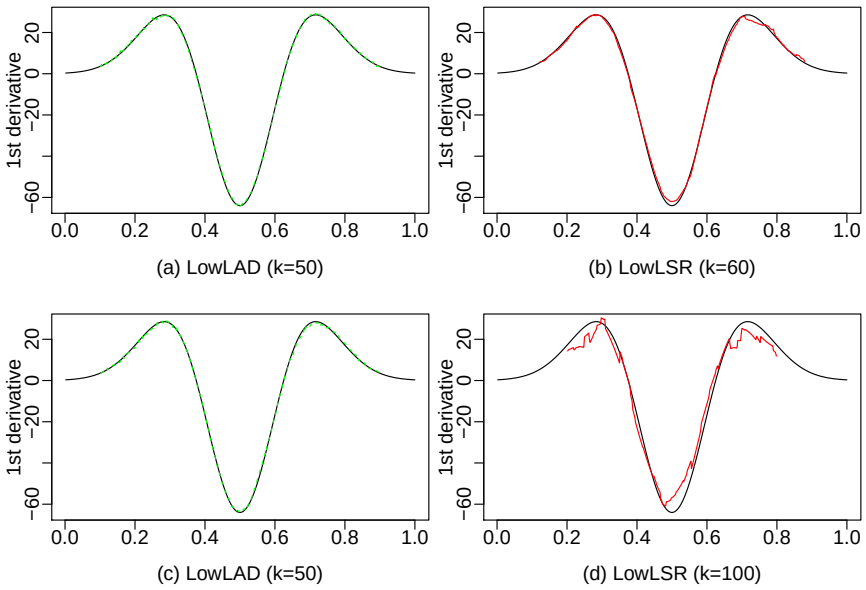

Figure 5: (a)-(d) The true second-order derivative function (bold line), LowLAD (green line) and LowLSR estimators (red line) based on the simulated data set from Figure 3.

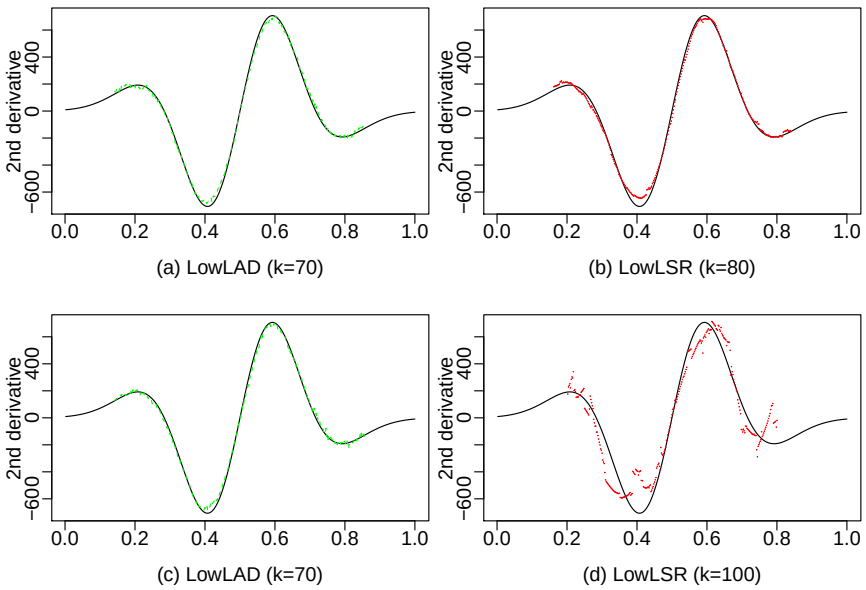

Figure 6: (a)-(d) The true second-order derivative function (bold line), LowLAD (green line) and LowLSR estimators (red line) based on the simulated data set from Figure 4.

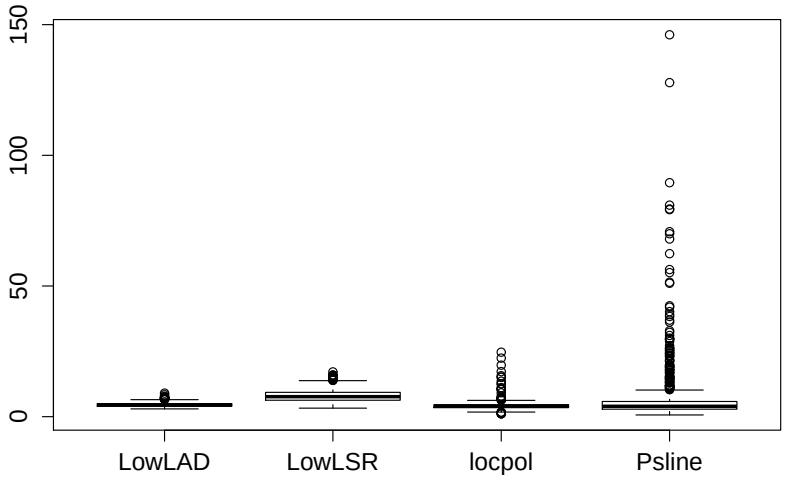

Figure 7: Boxplot of four estimators for the function m_4 with $\epsilon \sim 95\%N(0, 0.1^2) + 5\%N(0, 1^2)$.

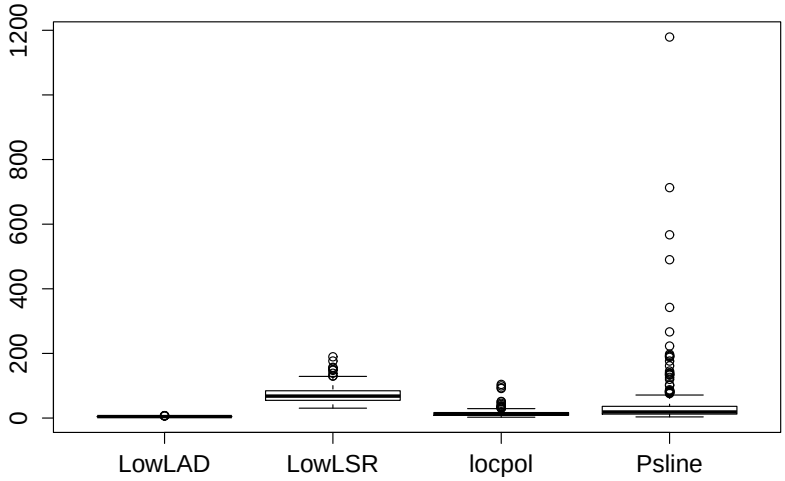

Figure 8: Boxplot of four estimators for the function m_4 with $\epsilon \sim 95\%N(0, 0.1^2) + 5\%N(0, 10^2)$.

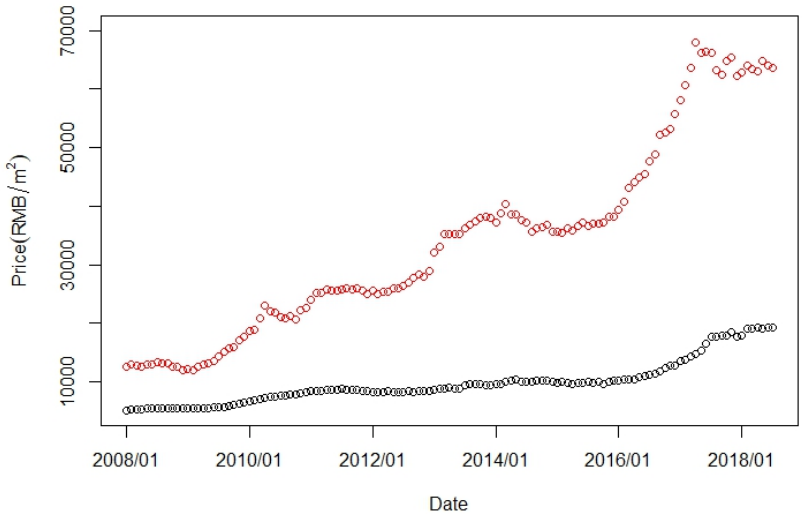

Figure 9: Black and green points denote the house prices in Beijing and Jinan, respectively.

α	0.001	0.002	0.003	0.004	0.005	0.006	0.007	0.008	0.009
σ_0^{LowLAD}	24.04	17.10	14.04	12.22	10.99	10.09	9.38	8.82	8.36
$\sigma_0^{\text{RLowLAD}}$	7.19	5.27	4.44	3.95	3.62	3.38	3.20	3.05	2.93
α	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
σ_0^{LowLAD}	7.96	5.88	4.99	4.47	4.12	3.87	3.69	3.54	3.42
$\sigma_0^{\text{RLowLAD}}$	2.83	2.29	2.06	1.92	1.83	1.76	1.71	1.67	1.64
α	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45	0.50
σ_0^{LowLAD}	3.32	3.01	2.86	2.79	2.77	2.78	2.83	2.92	3.05
$\sigma_0^{\text{RLowLAD}}$	1.61	1.52	1.48	1.45	1.43	1.42	1.42	1.42	1.43

Table 4: The critical values of σ_0 that equate the variances of the LowLAD (RLowLAD) and LowLSR derivative estimators with different contaminations.

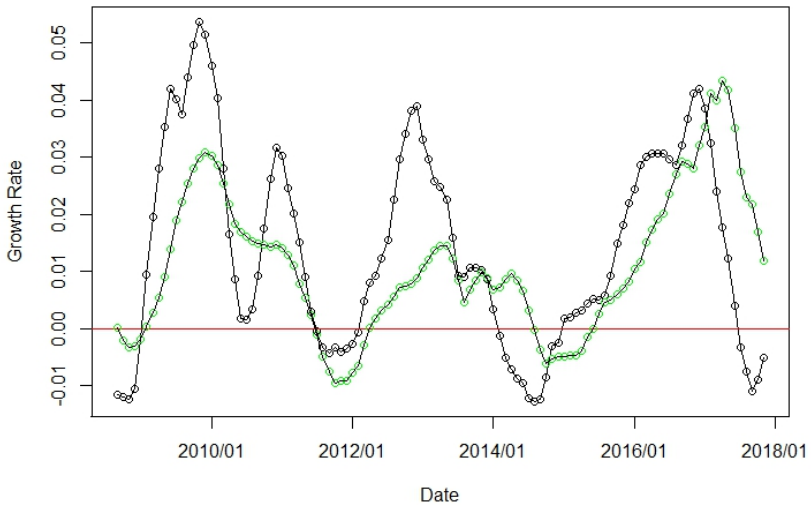

Figure 10: Black and green curves denote the relative growth rates for Beijing and Jinan, respectively. Relative growth rate is defined as $RLowLAD/Price$.

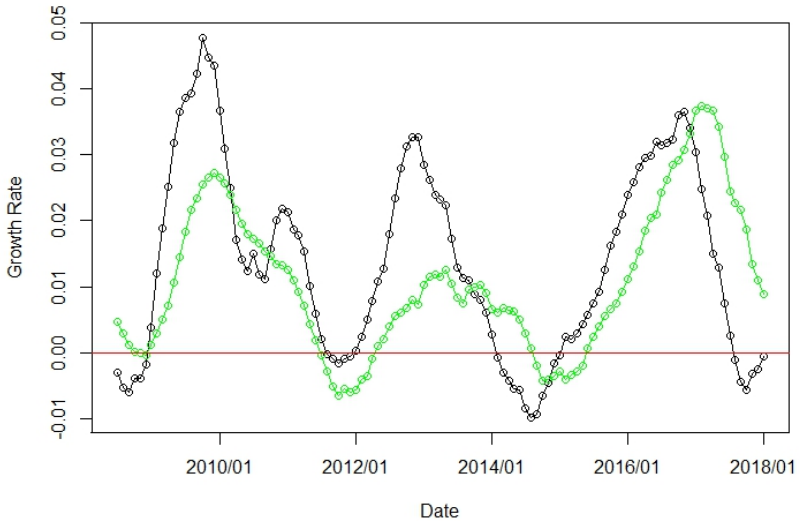

Figure 11: Black and green curves denote the relative growth rates based on the lower-order RLowLAD estimator.

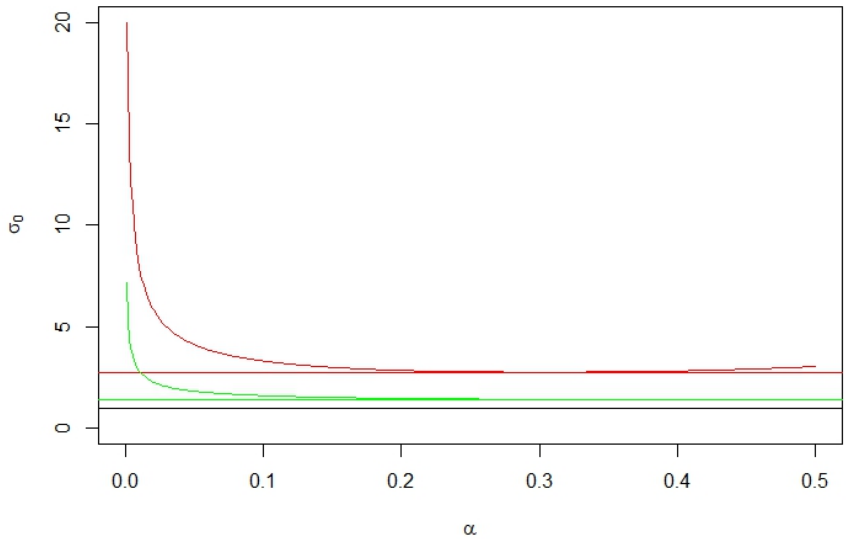

Figure 12: The red line is the critical σ_0 curve between LowLSR and LowLAD with $\epsilon_i \sim (1-\alpha)N(0,1) + \alpha N(0,\sigma_0^2)$, and the red horizontal line is $\sigma_0 = 2.77$; The green line is the critical σ_0 curve between LowLSR and RLowLAD, and the green horizontal line is $\sigma_0 = 1.42$; the black line is $\sigma_0 = 1$.

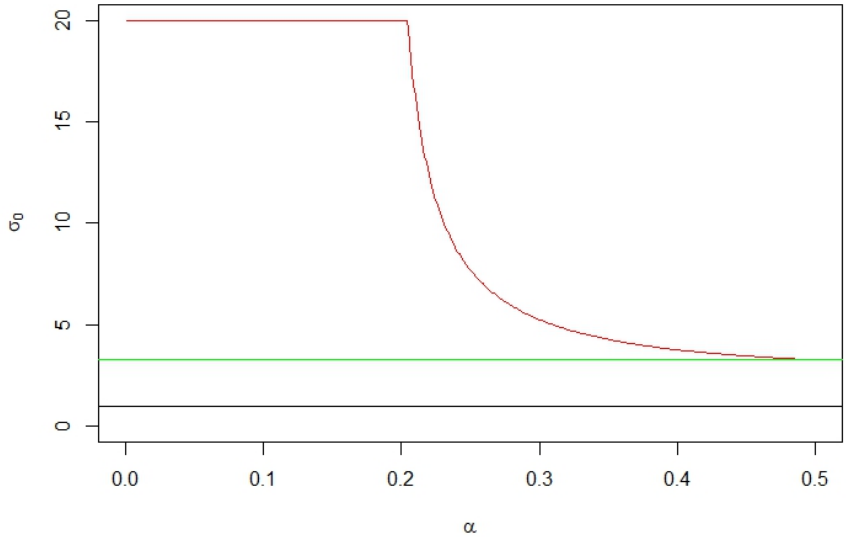

Figure 13: The red line is the critical σ_0 curve between RLowLAD and LAD with the same error distribution as in Figure 12, where the ratio larger than 20 is truncated at 20, the green horizontal line is $\sigma_0 = 3.28$, and the black line is $\sigma_0 = 1$.

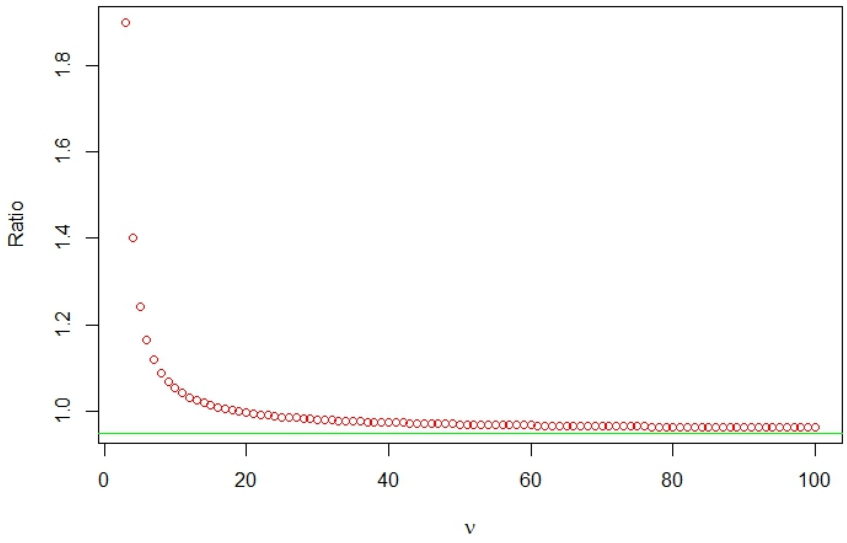

Figure 14: The red point-curve is the variance ratio function between LowLSR and RLowLAD for $t(\nu)$ with different ν's; the green horizontal line is $Ratio = 0.95$.

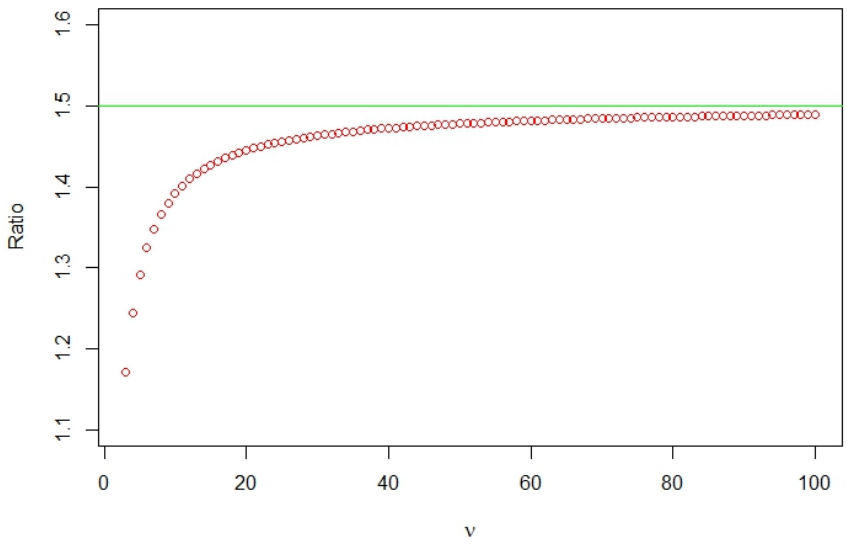

Figure 15: The red point-curve is the variance ratio function between LAD and RLowLAD estimators for (ν) with different ν's; and the green horizontal line is $Ratio = 1.50$.

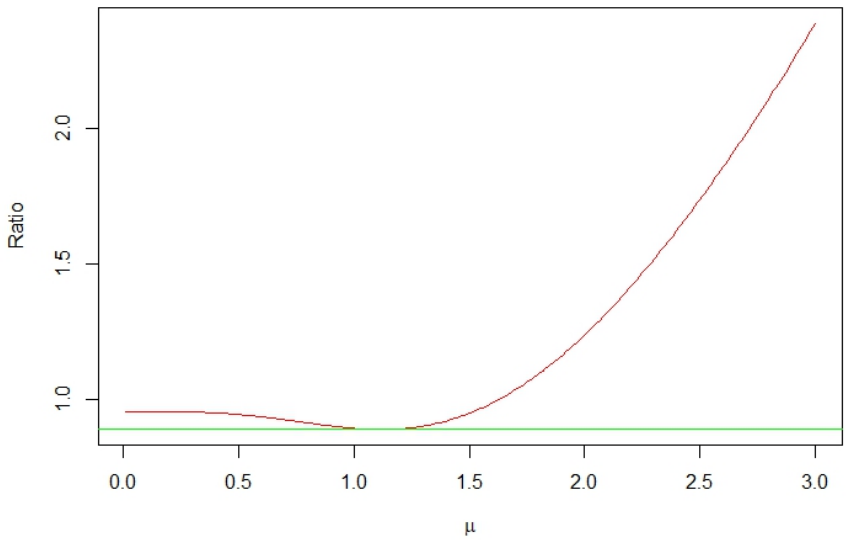

Figure 16: The red curve is the variance ratio function between LowLSR and RLowLAD for $\epsilon_i \sim 0.5N(\mu, 1) + 0.5N(-\mu, 1)$ with different μ's; the green horizontal line is $Ratio = 0.89$.

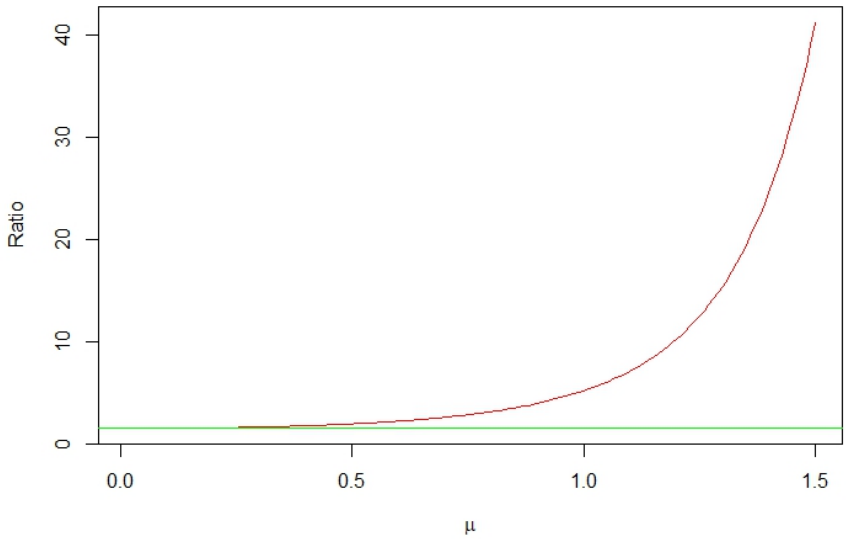

Figure 17: The red curve is the variance ratio function between RLowLAD and LAD for the same error distribution as in Figure 16; the green horizontal line is $Ratio = 1.50$.