Local Modal Regression

Weixin Yao

Department of Statistics, Kansas State University, Manhattan, Kansas 66506, U.S.A.

wxyao@ksu.edu

Bruce G. Lindsay and Runze Li

Department of Statistics, The Pennsylvania State University, University Park Pennsylvania, 16802-2111, U.S.A.

bgl@psu.edu, rli@stat.psu.edu

Abstract

A local modal estimation procedure is proposed for the regression function in a nonparametric regression model. A distinguishing characteristic of the proposed procedure is that it introduces an additional tuning parameter that is automatically selected using the observed data in order to achieve both robustness and efficiency of the resulting estimate. We demonstrate both theoretically and empirically that the resulting estimator is more efficient than the ordinary local polynomial regression estimator in the presence of outliers or heavy tail error distribution (such as t-distribution). Furthermore, we show that the proposed procedure is as asymptotically efficient as the local polynomial regression estimator when there are no outliers and the error distribution is a Gaussian distribution. We propose an EM type algorithm for the proposed estimation procedure. A Monte Carlo simulation study is conducted to examine the finite sample performance of the proposed method. The simulation results confirm the theoretical findings. The proposed methodology is further illustrated via an analysis of a real data example.

Key words: Adaptive regression; Local polynomial regression; M-estimator; Modal regression; Robust nonparametric regression.

1 Introduction

Local polynomial regression has been popular in the literature due to its simplicity of computation and nice asymptotic properties (Fan and Gijbels, 1996). In the presence of outliers, the local M-estimator has been investigated by many authors. See Härdle and Gasser (1984); Tsybakov (1986); Härdle and Tsybakov (1988); Hall and Jones (1990); Fan, Hu, and Truong (1994); Fan and Jiang (2000); Jiang and Mack (2001), among others. As usual, a nonparametric M-type of regression will be more efficient than least-squares based nonparametric regression when there are outliers or the error distribution has a heavy tail. However, these methods lose some efficiency when there are no outliers or the error distribution is normal. Thus, it is desirable to develop a new local modeling procedure, which can achieve both robustness and efficiency by adapting to different types of error distributions.

In this paper, we propose local modal regression procedure. Sampling properties of the proposed estimation procedure are systematically studied. We show that the proposed estimator is more efficient than the ordinary least-squares based local polynomial regression estimator in the presence of outliers or heavy tail error distribution. Furthermore, the proposed estimator achieves a full asymptotic efficiency of the ordinary local polynomial regression estimator when there are no outliers and the error distribution is Gaussian distribution. We further develop a modal EM algorithm for the local modal regression. Thus, the proposed modal regression can be implemented easily in practice. We conduct a Monte Carlo simulation to assess the finite sample performance of the proposed procedure. The simulation results show that the proposed procedure is robust to outliers, and performs almost as well as the local likelihood regression estimator constructed by using the true error function. In other words, the proposed estimator is almost as efficient as an omniscient estimator.

The rest of this paper is organized as follows. In Section 2, we propose the local modal regression, develop the modal EM algorithm for the local modal regression estimator, and study the asymptotic properties of the resulting estimator. In Section 3, Monte Carlo simulation

study is conducted, and a real data example is used to illustrate the proposed methodology. Technical conditions and proofs are given in the Appendix.

2 Local Modal Regression Estimator

Suppose that $(x_1, y_1), \ldots, (x_n, y_n)$ are an independent and identically distributed random sample from

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

where $\mathrm{E}(\epsilon \mid X = x) = 0$, $\mathrm{var}(\epsilon \mid X = x) = \sigma^2(x)$, and $m(\cdot)$ is an unknown nonparametric smoothing function to be estimated. Local polynomial regression is to locally approximate $m(x) = \mathrm{E}(Y \mid X = x)$ by a polynomial function. That is, for x in a neighborhood of x_0, we approximate

$$m(x) \approx \sum_{j=0}^{p} \frac{m^{(j)}(x_0)}{j!}(x - x_0)^j \equiv \sum_{j=0}^{p} \beta_j (x - x_0)^j,$$

where $\beta_j = m^{(j)}(x_0)/j!$.

The local parameter $\boldsymbol{\theta} = (\beta_0, \ldots, \beta_p)$ is estimated by minimizing the following weighted least squares function

$$\sum_{i=1}^{n} K_h(x_i - x_0) \left\{ y_i - \sum_{j=0}^{p} \beta_j (x_i - x_0)^j \right\}^2, \tag{2.1}$$

where $K_h(t) = h^{-1}K(t/h)$, a rescaled kernel function of $K(t)$ with a bandwidth h. The properties of local polynomial regression have been well studied (see, for example, Fan and Gijbels, 1996). It is also well known that the least squares estimate is sensitive to outliers. In this section, we propose local modal regression to achieve both robustness and efficiency.

Our local modal regression estimation procedure is to *maximize* over $\boldsymbol{\theta} = (\beta_0, \ldots, \beta_p)$

$$\ell(\boldsymbol{\theta}) \equiv \frac{1}{n}\sum_{i=1}^{n} K_{h_1}(x_i - x_0)\phi_{h_2}\left(y_i - \sum_{j=0}^{p}\beta_j(x_i - x_0)^j\right), \tag{2.2}$$

where $\phi_{h_2}(t) = h_2^{-1}\phi(t/h_2)$ and $\phi(t)$ is a *kernel density function*. The choice of ϕ is not very crucial. For ease of computation, we use the standard normal density for $\phi(t)$ throughout this paper. See (2.5) below. It is well known that the choice of $K(\cdot)$ is not very important. In our examples, we will also use Gaussian kernel for $K(\cdot)$. The choices of the bandwidths h_1 and h_2 will be discussed later. Denote the maximizer of $\ell(\boldsymbol{\theta})$ to be $\hat{\boldsymbol{\theta}} = (\hat{\beta}_0, \cdots, \hat{\beta}_p)$. Then the estimator of the v-th derivative of $m(x)$, $m^{(v)}(x)$, will be

$$\hat{m}^{(v)}(x_0) = v!\hat{\beta}_v, \quad \text{for} \quad v = 0, \cdots, p. \tag{2.3}$$

We will refer to $\hat{\boldsymbol{\theta}}$ as the local modal regression (LMR) estimator . Specially, when $p = 1$ and $v = 0$, we refer to this method as local linear modal regression (LLMR). When $p = 0$, (2.2) reduces to

$$\frac{1}{n}\sum_{i=1}^{n} K_{h_1}(x_i - x_0)\phi_{h_2}(y_i - \beta_0), \tag{2.4}$$

which is a kernel density estimate of (X, Y) at (x_0, y_0) with $y_0 = \beta_0$. Hence, the resulting estimate $\hat{\beta}_0$, by maximizing (2.4), is indeed the mode of the kernel density estimate in the y direction given $X = x_0$ (Scott, 1992, §8.3.2). This is the reason why we call our method local modal regression. In this paper, we will mainly consider univariate X. The proposed estimate is applicable for multivariate X, but is practically less useful due to the "curse of dimensionality".

In general, it is known that the sample mode is inherently insensitive to outliers as an estimator for the population mode. The robustness of the proposed procedure can be further interpreted from the point of view of M-estimation. If we treat $-\phi_{h_2}(\cdot)$ as a loss function, the

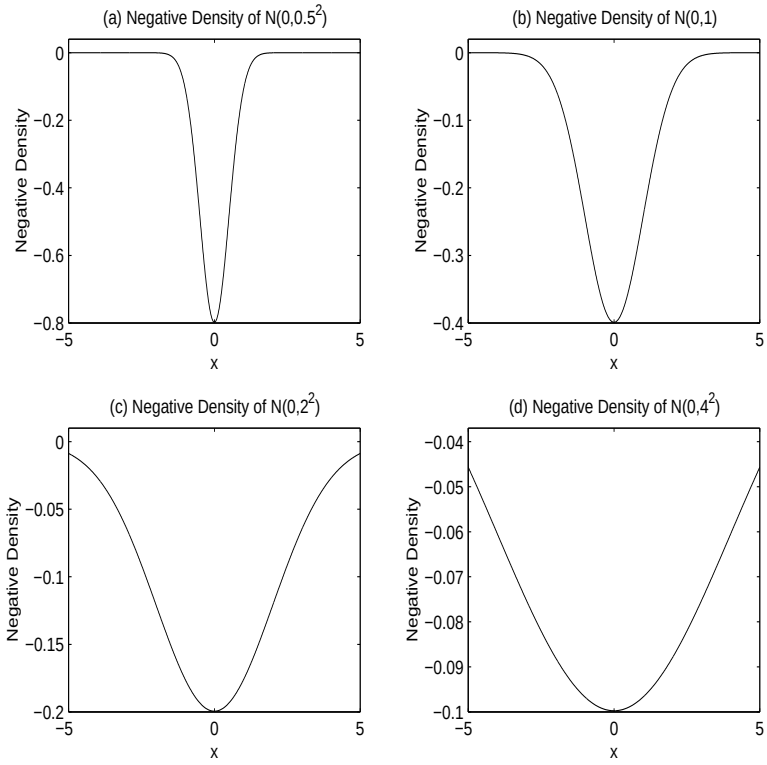

Figure 1: Plot of **Negative** Normal Densities.

resulting M-estimator is a local modal regression estimator. The bandwidth h_2 determines the degree of robustness of the estimator. Figure 1 provides insights into how the local modal regression estimator achieves the adaptive robustness. Note that h_2^2 corresponds to the variance in the normal density. From Figure 1, it can be seen that the *negative* normal density with small h_2, such as, $h_2 = 0.5$, looks like an outlier resistant loss function, while the shape of the negative normal density with large h_2, for example, $h_2 = 4$, is similar to the L_2-loss function. In practice, h_2 is selected by a data-driven method so that the resulting local estimate is adaptively robust. The issue of selection of both bandwidths h_1 and h_2 will be addressed later on.

2.1 Modal expectation-maximization algorithm

In this section, we extend the modal expectation-maximization (MEM) algorithm, proposed by Li, Ray, and Lindsay (2007), to maximize (2.2). Similar to an EM algorithm, the MEM algorithm also consists of two steps: E-step and M-step. Let $\boldsymbol{\theta}^{(0)} = (\beta_0^{(0)}, \ldots, \beta_p^{(0)})$ be the initial value and start with $k = 0$:

E-Step: In this step, we update $\pi(j \mid \boldsymbol{\theta}^{(k)})$ by

$$\pi(j \mid \boldsymbol{\theta}^{(k)}) = \frac{K_{h_1}(x_j - x_0)\phi_{h_2}(y_j - \sum_{l=0}^{p} \beta_l^{(k)}(x_j - x_0)^l)}{\sum_{i=1}^{n} \left\{ K_{h_1}(x_i - x_0)\phi_{h_2}(y_i - \sum_{l=0}^{p} \beta_l^{(k)}(x_i - x_0)^l) \right\}},$$

$$j = 1, \ldots, n.$$

M-Step: In this step, we update $\boldsymbol{\theta}^{(k+1)}$

$$\begin{aligned} \boldsymbol{\theta}^{(k+1)} &= \arg\max_{\boldsymbol{\theta}} \sum_{j=1}^{n} \left\{ \pi(j \mid \boldsymbol{\theta}^{(k)}) \log \phi_{h_2}(y_j - \sum_{l=0}^{p} \beta_l (x_j - x_0)^l) \right\} \\ &= (X^T W_k X)^{-1} X^T W_k Y \end{aligned} \tag{2.5}$$

since $\phi(\cdot)$ is the density function of a standard normal distribution. Here $X = (x_1^*, \ldots, x_n^*)^T$ with $x_i^* = (1, x_i - x_0, \cdots, (x_i - x_0)^p)^T$, W_k is an $n \times n$ diagonal matrix with diagonal elements $\pi(j \mid \boldsymbol{\theta}^{(k)})$s, and $Y = (y_1, \ldots, y_n)^T$.

The MEM algorithm requires one to iterate the E-step and the M-step until the algorithm converges. The ascending property of the proposed MEM algorithm can be established along the lines of Li, Ray, and Lindsay (2007). The closed form solution for $\boldsymbol{\theta}^{(k+1)}$ is one of the benefits of using normal density function $\phi_{h_2}(\cdot)$ in (2.2). If $h_2 \to \infty$, it can be seen in the E step that

$$\pi(j \mid \boldsymbol{\theta}^{(k)}) \to \frac{K_{h_1}(x_j - x_0)}{\sum_{i=1}^{n} K_{h_1}(x_i - x_0)} \propto K_{h_1}(x_j - x_0).$$

Thus, the LMR converges to the ordinary local polynomial regression (LPR). That is, the LPR is a limiting case of the LMR. This can also be roughly seen by the following approximation

$$\phi_{h_2}(x) = (\sqrt{2\pi}h_2)^{-1} \exp(-\frac{x^2}{2h_2^2}) \approx (\sqrt{2\pi}h_2)^{-1}(1 - \frac{x^2}{2h_2^2}).$$

(Note that this approximation only holds when h_2 is quite large.) This is another benefit of using the normal density $\phi_{h_2}(\cdot)$ for the LMR. This property makes LMR estimator achieve full asymptotic efficiency under the normal error distribution.

From the MEM algorithm, it can be seen that the major difference between the LPR and LMR lies in the E-step. The contribution of observation (x_i, y_i) to the LPR depends on the weight $K_h(x_i - x_0)$, which in turn depends on how close x_i is to x_0 only. On the other hand, the weight in the LMR depends on both how close x_i is to x_0 and how close y_i is to the regression curve. This weight scheme allows the LMR to downweight the observations further away from the regression curve to achieve adaptive robustness.

The reweighted least squares algorithm (IRWLS) can be also applied to our proposed local modal regression. When normal kernel is used for $\phi(\cdot)$, the reweighted least squares algorithm is actually equivalent to the proposed EM algorithm (but they are different if $\phi(\cdot)$ is not normal). In addition, IRWLS has been proved to have monotone and convergence property if $-\phi(x)/x$ is nonincreasing. But the proposed EM algorithm has been proved to have monotone property for any kernel density $\phi(\cdot)$. Note that $-\phi(x)/x$ is not nonincreasing if $\phi(x)$ has normal density. Therefore, the proposed EM algorithm provides a better explanation why the IRWLS is monotone for normal kernel density.

2.2 Theoretical properties

We first establish the convergence rate of the LMR estimator in the following theorem, whose proof can be found in the Appendix.

Theorem 2.1. *Under the regularity conditions (A1)—(A7) in the Appendix, with probability approaching to 1, there exists a consistent local maximizer* $\hat{\boldsymbol{\theta}} = (\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p)$ *of (2.2) such that*

$$\left| h_1^v \left\{ \hat{m}_v(x_0) - m^{(v)}(x_0) \right\} \right| = O_p\left((nh_1)^{-1/2} + h_1^{p+1} \right), \ v = 0, 1, \ldots, p,$$

where $\hat{m}_v(x_0) = v!\hat{\beta}_v$ *is the estimate of* $m^{(v)}(x_0)$ *and* $m^{(v)}(x_0)$ *is the* v^{th} *derivative of* $m(x)$ *at* x_0.

To derive the asymptotic bias and variance of the LMR estimator, we need the following notation. The moments of K and K^2 are denoted respectively by

$$\mu_j = \int t^j K(t)dt \qquad \text{and} \qquad \nu_j = \int t^j K^2(t)dt.$$

Let S, $\tilde{S}$, and S^* be $(p+1) \times (p+1)$ matrix with (j,l)-element μ_{j+l-2}, μ_{j+l-1}, and ν_{j+l-2}, respectively, and c_p and $\tilde{c}_p$ be $p \times 1$ vector with j-th element μ_{p+j} and μ_{p+j+1}, respectively. Furthermore, let $e_{v+1} = (0, \ldots, 0, 1, 0, \ldots, 0)^T$, a $p \times 1$ vector with 1 on the $(v+1)^{th}$ position.

Let

$$F(x, h_2) = \mathrm{E}\left\{ \phi''_{h_2}(\epsilon) \mid X = x \right\} \text{ and } G(x, h_2) = \mathrm{E}\left\{ \phi'_{h_2}(\epsilon)^2 \mid X = x \right\}, \tag{2.6}$$

where $\phi'_{h_2}(\epsilon)$ is the first derivative of $\phi_{h_2}(\epsilon)$ and $\phi''_{h_2}(\epsilon)$ is the second derivative of $\phi_{h_2}(\epsilon)$.

If ϵ and X are independent, then $F(x, h_2)$ and $G(x, h_2)$ are independent of x and we will use $F(h_2)$ and $G(h_2)$ to denote them respectively in this situation. Furthermore, denote the marginal density of X, i.e. the *design density*, by $f(\cdot)$.

Theorem 2.2. *Under the regularity conditions (A1)—(A7) in the Appendix, the asymptotic variance of* $\hat{m}_v(x_0)$*, given in Theorem 2.1, is given by*

$$var\{\hat{m}_v(x_0)\} = e_{v+1}^T S^{-1} S^* S^{-1} e_{v+1} \frac{v!^2}{f(x_0) n h_1^{1+2v}} G(x_0, h_2) F(x_0, h_2)^{-2} + o(n^{-1} h_1^{-1-2v}). \tag{2.7}$$

The asymptotic bias of $\hat{m}_v(x_0)$, denoted by $b_v(x_0)$, for $p-v$ odd is given by

$$b_v(x_0) = e_{v+1}^T S^{-1} c_p \frac{v!}{(p+1)!} m^{(p+1)}(x_0) h_1^{p+1-v} + o(h_1^{p+1-v}). \tag{2.8}$$

Furthermore, the asymptotic bias for $p-v$ even is

$$b_v(x_0) = e_{v+1}^T S^{-1} \tilde{c}_p \frac{v! h_1^{p+2-v}}{(p+2)!} \left\{ m^{(p+2)}(x_0) + (p+2) m^{(p+1)}(x_0) a(x_0) \right\} + o(h_1^{p+2-v}), \tag{2.9}$$

provided that $m^{(p+2)}(\cdot)$ are continuous in a neighborhood of x_0 and $nh_1^3 \to \infty$, where

$$a(x_0) = \frac{\frac{\partial F(x,h_2)}{\partial x}|_{x=x_0} f(x_0) + F(x_0,h_2) f'(x_0)}{F(x_0,h_2) f(x_0)}. \tag{2.10}$$

The proof of Theorem 2.2 is given in the Appendix. Based on (2.7) and the asymptotic variance of the LPR estimator given in Fan and Gijbels (1996), we can show that the ratio of the asymptotic variance of the LMR estimator to that of the LPR estimator is given by

$$R(x_0, h_2) \triangleq \frac{G(x_0,h_2) F^{-2}(x_0,h_2)}{\sigma^2(x_0)}. \tag{2.11}$$

The ratio $R(x_0, h_2)$ depends on x_0 and h_2 only, and it plays an important role in the discussion of relative efficiency in Section 2.5. Furthermore, the ideal choice of h_2 is

$$h_{2,opt} = \arg\min_{h_2} R(x_0,h_2) = \arg\min_{h_2} G(x_0,h_2) F^{-2}(x_0,h_2). \tag{2.12}$$

From (2.12), we can see that $h_{2,opt}$ dose not depend on n and only depends on the conditional error distribution of ϵ given X.

Based on (2.8), (2.9), and the asymptotic bias of the LPR estimator (Fan and Gijbels, 1996), we know that the LMR estimator and the LPR estimator have the same asymptotic bias when $p-v$ is odd. When $p-v$ is even, they are still the same provided that ϵ and X are

independent as $a(x_0)$ defined in (2.10) equals $f'(x_0)/f(x_0)$, but they are different if ϵ and X are not independent. Similar to the LPR, the second term in (2.9) often creates extra bias. Thus, it is preferable to use odd values of $p-v$ in practice. Thus, it is consistent with the selection order of p for the LPR (Fan and Gijbels, 1996). From now on, we will concentrate on the case when $p-v$ is odd.

Theorem 2.3. *Under the regularity conditions (A1)—(A7) in the Appendix, the estimate $\hat{m}_v(x_0)$, given in Theorem 2.1, has the following asymptotic distribution*

$$\frac{\hat{m}_v(x_0) - m^{(v)}(x_0) - b_v(x_0)}{\sqrt{var\{\hat{m}_v(x_0)\}}} \xrightarrow{L} N(0,1).$$

The proof of Theorem 2.3 is given in the Appendix.

2.3 Asymptotic bandwidth and relative efficiency

Note that the mean squared error (MSE) of the LMR estimator, $\hat{m}_v(x_0)$, is

$$b_v^2(x_0) + \text{var}\{\hat{m}_v(x_0)\}. \tag{2.13}$$

The asymptotic optimal bandwidth for odd $p-v$, that minimizes the MSE, is

$$h_{1,opt} = R(x_0, h_2)^{1/(2p+3)} h_{LPR}, \tag{2.14}$$

where h_{LPR} is the asymptotic optimal bandwidth for LPR (Fan and Gijbels, 1996),

$$h_{LPR} = C_{v,p}\left[\frac{\sigma^2(x_0)}{\{m^{(p+1)}(x_0)\}^2 f(x_0)}\right]^{1/(2p+3)} n^{-1/(2p+3)}, \tag{2.15}$$

with

$$C_{v,p}(K) = \left\{\frac{(p+1)!^2(2v+1)e_{v+1}^T S^{-1} S^* S^{-1} e_{v+1}}{2(p+1-v)(e_{v+1}^T S^{-1} c_p)^2}\right\}^{1/(2p+3)}.$$

The asymptotic relative efficiency (ARE) between the LMR estimator with $h_{1,opt}$ and $h_{2,opt}$ and the LPR estimator with h_{LPR} of $m^{(v)}(x_0)$ with order p is

$$\text{ARE} = \frac{\text{MSE(LPR)}}{\text{MSE(LMR)}} = R(x_0, h_{2,opt})^{-(2p-2v+2)/(2p+3)}. \tag{2.16}$$

From (2.16), we see that $R(x_0, h_2)$ completely determines the ARE for fixed p and v. Let us study the properties of $R(x, h_2)$ further.

Theorem 2.4. *Let $g_{\epsilon|x}(t)$ be the conditional density of ϵ given $X = x$. For $R(x, h_2)$ defined in (2.11), given any x, we have the following results.*

(a) $\lim_{h_2 \to \infty} R(x, h_2) = 1$ *and hence* $\inf_{h_2} R(x, h_2) \leq 1$;

(b) If $g_{\epsilon|x}(t)$ is a normal density, $R(x, h_2) > 1$ for any finite h_2 and $\inf_{h_2} R(x, h_2) = 1$.

(c) Assuming $g_{\epsilon|x}(t)$ has bounded third derivative, if $h_2 \to 0$, $R(x, h_2) \to \infty$.

The proof of Theorem 2.4 is given in the Appendix. From (a) and (2.16), one can see that the supremum (over h_2) of the relative efficiency between the LMR and LPR is larger than or equal to 1. Hence LMR works at least as well as the LPR for any error distribution. If there exists some h_2 such that $R(x, h_2) < 1$, then the LMR estimator has smaller asymptotic MSE than the LPR estimator.

As discussed in section 2.3, when $h_2 \to \infty$, the LMR converges to the LPR. The equation $\lim_{h_2 \to \infty} R(x, h_2) = 1$ of (a) confirms this result. It can be seen from (b) that when $\epsilon \sim N(0, 1)$, the optimal LMR (with $h_2 \to \infty$) is the same as LPR. This is the reason why LMR will not lose efficiency under normal distribution. From (c) one can see that the optimal h_2 should not be too small, which is quite different from the needed locality affect of h_1.

Table 1 lists the asymptotic relative efficiency between the LLMR estimator (LMR with $p = 1$ and $v = 0$), and the local linear regression (LLR) estimator for normal error distribution and some special error distributions that are generally used to evaluate the robustness of

a regression method. The normal mixture is used to mimic the outlier situation. This kind of mixture distribution is also called the contaminated normal distribution. The t-distributions with degrees of freedom from 3 to 5 are often used to represent heavy-tail distributions. From Table 1, one can see that the improvement of LLMR over LLR is substantial when there are outliers or the error distribution has heavy tails.

Table 1: Relative efficiency between the LLMR estimator and the LLR estimator

Error Distribution	Relative Efficiency
$N(0,1)$	1
$0.95N(0,1)+0.05N(0,3^2)$	1.1745
$0.95N(0,1)+0.05N(0,5^2)$	1.6801
t-distribution with df=5	1.1898
t-distribution with df=3	1.7169

3 Simulation Study and Application

In this section, we will conduct a Monte Carlo simulation to assess the performance of the proposed LLMR and compare it with LLR and some commonly used robust estimators. We first address how to select the bandwidths h_1 and h_2 in practice.

3.1 Bandwidth selection in practice

In our simulation setting, ϵ and X are independent. Thus, we need to estimate $F(h_2)$ and $G(h_2)$ defined in (2.6) in order to find the optimal bandwidth $h_{2,opt}$ based on (2.12). To this end, we first get an initial estimate of $m(x)$, denoted by $\hat{m}_I(x)$ and the residual $\hat{\epsilon}_i = y_i - \hat{m}_I(x_i)$, by fitting the data using any simple robust smoothing method, such as

LOWESS. Then we estimate $F(h_2)$ and $G(h_2)$ by

$$\hat{F}(h_2) = \frac{1}{n}\sum_{i=1}^{n} \phi''_{h_2}(\hat{\epsilon}_i) \quad \text{and} \quad \hat{G}(h_2) = \frac{1}{n}\sum_{i=1}^{n} \{\phi'_{h_2}(\hat{\epsilon}_i)\}^2,$$

respectively. Then $R(h_2)$ can be estimated by $\hat{R}(h_2) = \hat{G}(h_2)\hat{F}(h_2)^{-2}/\hat{\sigma}^2$, where $\hat{\sigma}$ is estimated based on the pilot estimates, $\hat{\epsilon}_1, \ldots, \hat{\epsilon}_n$, of the error term. Using the grid search method, we can easily find $\hat{h}_{2opt}$ to minimize $\hat{R}(h_2)$. (Note that $\hat{h}_{2opt}$ would not depend on x.) From Theorem 2.4(c), we know that the asymptotically optimal h_2 is never too small. Based on our empirical experience, the size of chosen h_2 is usually comparable to the standard deviation of the error distribution. Hence the possible grid points for h_2 can be: $h_2 = 0.5\hat{\sigma} \times 1.02^j, j = 0, \ldots, k$, for some fixed k (such as $k = 90$).

The asymptotically optimal bandwidth h_1 is much easier to estimate after finding $\hat{h}_{2opt}$. Based on the formula (2.14) in Section 2.5, the asymptotically optimal bandwidth for h_1 of LLMR is h_{LLR} multiplied by a factor $\{R(h_{2opt})\}^{1/5}$. After finding $\hat{h}_{2opt}$, we can estimate $\{R(h_{2opt})\}^{1/5}$ by $\{\hat{R}(\hat{h}_{2opt})\}^{1/5}$. We can then employ an existing bandwidth selector for LLR, such as the plug-in method (Ruppert, Sheather, and Wand, 1995). If the optimal bandwidth selected for LLR is $\hat{h}_{LLR}$, then h_1 is estimated by $\hat{h}_{1opt} = \{\hat{R}(\hat{h}_{2,opt})\}^{1/5}\hat{h}_{LLR}$.

When ϵ and X are independent, the relationship (2.14) also holds for the global optimal bandwidth that is obtained by minimizing weighted Mean Integrated Square Error $\int [b_v^2(x) + \mathrm{var}\{\hat{m}_v(x)\}]\, w(x)dx$, where $w \geq 0$ is some weight function, such as 1 or design density $f(x)$. Hence the above proposed way to find $\hat{h}_{1opt}$ also works for the global optimal bandwidth. For the simplicity of computation, we used the global optimal bandwidth for $\hat{h}_{LLR}$ and thus $\hat{h}_{1opt}$ for our examples in Section 3.2 and 3.3.

3.2 Simulation study

For comparison, we include in our simulation study the local likelihood regression (LLH) estimator (Tibshirani and Hastie, 1987) assuming the error distribution is known. Specifically, suppose the error distribution is $g(t)$, the LLH estimator finds $\hat{\boldsymbol{\theta}} = (\hat{\beta}_0, \hat{\beta}_1)$ by maximizing the following local likelihood

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} \sum_{i=1}^{n} K_h(x_i - x_0) \log\{g(y - \beta_0 - \beta_1(x_i - x_0))\} \,. \tag{3.1}$$

The estimate of regression function $m(x_0)$ is $\hat{m}(x_0) = \hat{\beta}_0$.

If the error density $g(t)$ is assumed to be known, the LLH estimator (3.1) is the most efficient estimator. However, in reality, we will seldom know the true error density. The LLH estimator is just used as a benchmark, omniscient estimator to check how well the LLMR estimator adapts to different true densities.

We generate the independent and identically distributed (i.i.d.) data $\{(x_i, y_i), i = 1, \ldots, n\}$ from the model $Y_i = 2\sin(2\pi X_i) + \epsilon_i$, where $X_i \sim U(0, 1)$. We consider the following three cases:

Case I: $\epsilon_i \sim N(0, 1)$.

Case II: $\epsilon_i \sim 0.95N(0, 1) + 0.05N(0, 5^2)$. The 5% data from $N(0, 5^2)$ are most likely to be outliers.

Case III: $\epsilon_i \sim t_3$.

We compared the following five estimators:

1. Local linear regression (LLR). We used the plug-in bandwidth (Ruppert, Sheather, and Wand, 1995).

2. Local ℓ_1 regression/median regression (LMED).

3. Local M estimator (LM) using Huber's function $\psi(x) = \max\{-c, \min(c, x)\}$. As in Fan and Jiang (2000), we take $c = 1.35\hat{\sigma}$, where $\hat{\sigma}$ is the estimated standard deviation of the error term by MAD estimator i.e.

$$\hat{\sigma} = 1.4826 \times \text{Median}(|\hat{\boldsymbol{\epsilon}} - \text{Median}(\hat{\boldsymbol{\epsilon}})|),$$

where $\hat{\boldsymbol{\epsilon}} = (\hat{\epsilon}_1, \ldots, \hat{\epsilon}_n)$ are the pilot estimates of the error term.

4. Local linear modal regression (LLMR) estimator (LMR with $p = 1$ and $v = 0$).

5. Local likelihood regression (LLH) using the true error density.

For comparison, in Table 2, we reported the relative efficiency between different estimators and the benchmark estimator LLH, where RE(LLMR) is the relative efficiency between the LLMR estimator and the LLH estimator. That is, RE(LLMR) is the ratio of MSE(LLH) to MSE(LLMR) (based on 50 equally spaced grid points from 0.05 to 0.95 and 500 replicates). The same notation applies to other methods.

From Table 2, it can be seen that for normal error, LLMR had a relative efficiency very close to 1 from the small sample size 50 to the large sample size 500. Notice that in Case I, we need not use a robust procedure and LLR should work the best in this case. Note that in this case LLR is the same as LLH. However the newly proposed method LLMR worked almost as well as LLR/LLH when the error distribution is exactly the normal distribution. Hence LLMR adapted to normal errors very well. In addition, we can see that LM lost about 8% efficiency for the small size 50 and lost about 5% efficiency for the large sample size 500. LMED lost more than 30% efficiency under normal error.

For contaminated normal error, LLMR still had a relative efficiency close to 1 and worked better than LM, especially for large sample sizes. Hence LLMR adapted to contaminated normal error distributions quite well. In this case, LLR lost more than 40% efficiency and LMED lost about 30% efficiency.

For t_3 error, it can be seen from Table 2 that LLMR also worked similarly to LLH and a little better than LM, especially for large sample sizes. Hence LLMR also adapted to t-distribution errors quite well. In this case, LLR lost more than 40% efficiency and LMED lost about 15% efficiency.

Table 2: Relative efficiency between different estimators and the LLH estimator

Error Distribution	n	RE(LLR)	RE(LMED)	RE(LM)	RE(LLMR)
$N(0,1)$	50	1	0.6676	0.9235	0.9979
	100	1	0.6757	0.9358	0.9992
	250	1	0.6753	0.9455	0.9997
	500	1	0.6818	0.9488	0.9998
$0.95N(0,1)+0.05N(0,5^2)$	50	0.6446	0.7314	0.9392	0.9127
	100	0.5828	0.7113	0.9298	0.9210
	250	0.5598	0.7222	0.9375	0.9675
	500	0.5691	0.7246	0.9402	0.9859
t_3	50	0.6948	0.8196	0.9809	0.9514
	100	0.6429	0.8386	0.9611	0.9350
	250	0.5462	0.8470	0.9428	0.9617
	500	0.5743	0.8497	0.9442	0.9747

3.3 An application

In this section, we illustrate the proposed methodology by analysis of the Education Expenditure Data (Chatterjee and Price, 1977). This data set consists of 50 observations from 50 states, one for each state. The two variables to be considered here are X, the number of residents per thousand residing in urban areas in 1970 and Y, the per capita expenditure on public education in a state, projected for 1975. For this example, one can easily identify the outlier. We use this example to show how the obvious outlier will affect the LLR fit and the LLMR fit.

Figure 2 is the scatter plot of original observations and the fitted regression curves by LLR and LLMR. From Figure 2, one can see that there is an extreme observation (outlier). This extreme observation is from Hawaii, which has very high per capita expenditure on public education with x value close to 500. This observation created the big difference between the two fitted curves around $x = 500$. The observations with x around 500 appear to go down in that area compared to the observations with x around 600. Thus the regression function should also go down when x moves from 600 to 500. The LLMR fit reflected this fact. (For this example, the robust estimators LMED and LM provided similar results to LLMR.) However the LLR fit went up in that area, due to the big impact of the extreme observation from Hawaii. In fact, this extreme observation received about a 10% weight in the LLR fit at $x = 500$, compared to nearly 0% weight in the LLMR fit. Hence, unlike local linear regression, local linear modal regression adapts to, and is thereby robust to, outliers.

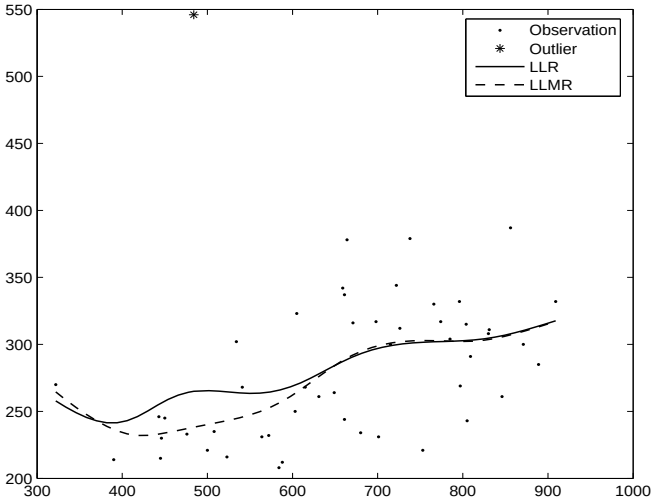

Figure 2: Plot of fitted regression curves for Education Expenditure Data. The star point is the extreme observation from Hawaii. The solid curve is the local linear regression (LLR) fit. The dash-dash curve is the local linear modal regression (LLMR) fit.

4 Discussion

In this paper, we proposed a local modal regression proceduce. It introduces an additional tuning parameter that is automatically selected using the observed data in order to achieve both robustness and efficiency of the resulting nonparametric regression estimator. Modal regression has been briefly discussed in Scott (1992, §8.3.2) without any detailed asymptotic results. Scott (1992, §8.3.2) used a constant β_0 to estimate the local mode as (2.4). Due to the advantage of local polynomial regression over the local constant regression, we extended the local constant structure to local polynomial structure and provided a systematic study of the asymptotic results of the local modal regression estimator. As a measure of center, the modal regression uses the "most likely" conditional values rather than the conditional average. When the conditional density is symmetric, these two criteria match. However, as Scott (1992, §8.3.2) stated that modal regression, besides the robustness, can explore more complicated data structure when there are multiple local modes. Hence local modal regression may be applied to mixture of regression (Goldfeld and Quandt, 1976; Fruhwirth-Schnatter, 2001; Rossi, Allenby, and McCulloch, 2005; Green and Richardson, 2002) and "change point" problem (Lai, 2001; Bai and Perron, 2003; Goldenshluger, Tsbakov, and Zeevi, 2006). These require further research.

Chu, et al. (1998) also used the Gaussian kernel as the outlier-resistent function in their proposed local constant M-smoother for image processing. However, they let $h_2 \to 0$ and aimed at edge-preserving smoothing when there is jump in the regression curves. In this paper, the goal was different; we sought to provide an adaptive robust regression estimate for the smooth regression function $m(x)$ by adaptively choosing h_2. In addition, we proved that for regression estimate, the optimal h_2 does not depend on n and should not be too small.

In addition, note that the local modal regression does not estimate the mean function in general. It requires the assumption $\mathrm{E}(\phi'_{h_2}(\epsilon) \mid X = x) = 0$, which holds if the error density

is symmetric about 0. If the above assumption about the error density does not hold, the proposed estimate is actually estimating the function

$$\tilde{m}(x) = \arg\max_{m} \mathrm{E}[\phi_{h_2}(y-m) \mid X = x],$$

which converges to the mode $\mathrm{E}(Y \mid X = x)$ if $h_2 \to 0$ and the bias depends on h_2. For the general error distribution and fixed h_2, all the asymptotics provided in this paper still apply if we replace the mean function $m(x)$ by $\tilde{m}(x)$.

APPENDIX: PROOFS

The following technical conditions are imposed in this section.

Technical Conditions:

(A1) $m(x)$ has continuous $(p+1)^{th}$ derivative at the point x_0.

(A2) $f(x)$ has continuous first derivative at the point x_0 and $f(x_0) > 0$.

(A3) $F(x, h_2)$ and $G(x, h_2)$ are continuous with respect to x at the point x_0, where $F(x, h_2)$ and $G(x, h_2)$ are defined in (2.6).

(A4) $K(\cdot)$ is a symmetric (about 0) probability density with compact support $[-1, 1]$.

(A5) $F(x_0, h_2) < 0$ for any $h_2 > 0$.

(A6) $\mathrm{E}(\phi'_{h_2}(\epsilon) \mid X = x) = 0$ and $\mathrm{E}(\phi''_{h_2}(\epsilon)^2 \mid X = x)$, $E(|\phi'_{h_2}(\epsilon)^3| \mid X = x)$, and $\mathrm{E}(\phi'''_{h_2}(\epsilon) \mid X = x)$ are continuous with respect to x at the point x_0.

(A7) The bandwidth h_1 tends to 0 such that $nh_1 \to \infty$ and the bandwidth h_2 is a constant and does not depend on n.

The above conditions are not the weakest possible conditions, but they are imposed to facilitate the proofs. For example, the compact support restriction on $K(\cdot)$ is not essential and can be removed if we put restriction on the tail of $K(\cdot)$. The condition (A5) ensures that there exists a local maximizer of (2.2). In addition, although h_1 is assumed to go to zero when $n \to \infty$, h_2 is assumed to be a fixed constant and its optimal values only depend on the error density not n. The condition $\mathrm{E}(\phi'_{h_2}(\epsilon) \mid X = x) = 0$ ensures the proposed estimate is consistent and it is satisfied if the error density is symmetric about 0. However, we don't require the error distribution to be symmetric about 0. If the assumption $\mathrm{E}(\phi'_{h_2}(\epsilon) \mid X = x) = 0$ doesn't hold, the proposed estimate is actually estimating the function

$$\tilde{m}(x) = \arg\max_m \mathrm{E}[\phi_{h_2}(y - m) \mid X = x].$$

Denote $X_i^* = \{1, (X_i - x_0)/h_1, \ldots, (X_i - x_0)^p/h_1^p\}^T$, $H = \mathrm{diag}\{1, h_1, \ldots, h_1^p\}$, $\boldsymbol{\theta} = (\beta_0, \beta_1, \ldots, \beta_p)^T$, $\boldsymbol{\theta}^* = H\boldsymbol{\theta}$, $R(X_i) = m(X_i) - \sum_{j=0}^p \beta_j (X_i - x_0)^j$, and $K_i = K_{h_1}(x_i - x_0)$, where $\beta_j = m^{(j)}(x_0)/j!, j = 0, 1, \ldots, p$. The following lemmas are needed for our technical proofs.

Lemma A.1. *Assume that the conditions A1-A6 hold. We have*

$$\frac{1}{n}\sum_{i=1}^n K_i \phi''_{h_2}(\epsilon_i) \left(\frac{X_i - x_0}{h_1}\right)^j = F(x_0, h_2) f(x_0) \mu_j + o_p(1) \tag{A.1}$$

and

$$\frac{1}{n}\sum_{i=1}^n K_i \phi''_{h_2}(\epsilon_i) R(X_i) \left(\frac{X_i - x_0}{h_1}\right)^j = h_1^{p+1} F(x_0, h_2) f(x_0) \mu_{j+p+1} \frac{m^{(p+1)}(x_0)}{(p+1)!} + o_p(h_1^{p+1}). \tag{A.2}$$

Proof. We shall prove (A.1), since (A.2) can be shown by the same arguments. Denote $T_n = n^{-1}\sum_{i=1}^n K_i \phi''_{h_2}(\epsilon_i)(\frac{X_i - x_0}{h_1})^j$. In the same lines of arguments as in Lemma 5.1 of (Fan

and Jiang, 2000), we have

$$\mathrm{E}(T_n) = F(x_0, h_2) f(x_0) \mu_j + o(1),$$

and

$$\mathrm{var}(T_n) \leq \frac{1}{n^2} \sum_{i=1}^{n} E \left\{ K_i^2 \mathrm{E}(\phi''_{h_2}(\epsilon_i)^2 \mid X_i) \left(\frac{X_i - x_0}{h_1} \right)^{2j} \right\} = \frac{1}{nh_1} \nu_{2j} H(x_0, h_2) f(x_0)(1 + o(1))$$

Based on the result $T_n = E(T_n) + O_p(\sqrt{\mathrm{var}(T_n)})$ and the assumption $nh_1 \to \infty$, it follows that $T_n = F(x_0, h_2) f(x_0) \mu_j + o_p(1)$.

Proof of Theorem 2.1. Denote $\alpha_n = (nh_1)^{-1/2} + h_1^{p+1}$. It is sufficient to show that for any given $\eta > 0$, there exists a large constant c such that

$$P\{ \sup_{|\mu\|=c} \ell(\boldsymbol{\theta}^* + \alpha_n \mu) < \ell(\boldsymbol{\theta}^*) \} \geq 1 - \eta, \tag{A.3}$$

where $\ell(\boldsymbol{\theta})$ is defined in (2.2).

By using Taylor expansion, it follows that

$$\begin{aligned}
\ell(\boldsymbol{\theta}^* + \alpha_n \mu) - \ell(\boldsymbol{\theta}^*) =& \frac{1}{n} \sum_{i=1}^{n} K_i \left\{ \phi_{h_2}(\epsilon_i + R(X_i) + \alpha_n \mu^T X_i^*) - \phi_{h_2}(\epsilon_i + R(X_i)) \right\} \\
=& \frac{1}{n} \sum_{i=1}^{n} K_i \left\{ \phi'_{h_2}(\epsilon_i + R(X_i)) \alpha_n \mu^T X_i^* + \frac{1}{2} \phi''_{h_2}(\epsilon_i + R(X_i)) \alpha_n^2 (\mu^T X_i^*)^2 \right. \\
& \left. - \frac{1}{6} \phi'''_{h_2}(z_i) \alpha_n^3 (\mu^T X_i^*)^3) \right\} \\
\triangleq & I_1 + I_2 + I_3,
\end{aligned} \tag{A.4}$$

where z_i is between $\epsilon_i + R(X_i)$ and $\epsilon_i + R(X_i) + \alpha_n \mu^T X_i^*$.

By directly calculating the mean and variance, we obtain

$$\mathrm{E}(I_1) = \alpha_n h_1^{p+1} \mathrm{E}\left\{K_i \mathrm{E}(\phi''_{h_2}(\epsilon_i) \mid X_i)\mu^T X_i^*\right\}(1+o(1)) = O(\alpha_n h_1^{p+1} c);$$
$$\mathrm{var}(I_1) = n^{-1}\alpha_n^2 \mathrm{var}[K_i \phi'_{h_2}\{\epsilon_i + R(X_i)\}(\mu^T X_i^*)] = O(\alpha_n^2 (nh_1)^{-1} c^2).$$

Hence

$$I_1 = O(\alpha_n h_1^{p+1} c) + \alpha_n c O_p((nh_1)^{-1/2}) = O_p(c\alpha_n^2).$$

Similarly,

$$I_3 = \frac{1}{n}\sum_{i=1}^{n}\left\{-\frac{1}{6}K_i \phi'''_{h_2}(z_i)\alpha_n^3 (\mu^T X_i^*)^3\right\} = O_p(\alpha_n^3).$$

From Lemma A.1, it follows that

$$I_2 = \frac{1}{n}\sum_{i=1}^{n}\left\{\frac{1}{2}K_i \phi''_{h_2}(\epsilon_i + R(X_i))\alpha_n^2 \mu^T X_i^* X_i^{*T}\mu\right\} = \alpha_n^2 F(x_0, h_2) f(x_0)\mu^T S\mu(1+o_p(1)).$$

Noticing that S is a positive matrix, $\|\mu\| = c$, and $F(x_0, h_2) < 0$, we can choose c large enough such that I_2 dominates both I_1 and I_3 with probability at least $1-\eta$. Thus (A.3) holds. Hence with probability approaching 1 (wpa1), there exists a local maximizer $\hat{\boldsymbol{\theta}}^*$ such that $||\hat{\boldsymbol{\theta}}^* - \boldsymbol{\theta}^*|| \leq \alpha_n c$, where $\alpha_n = (nh_1)^{-1/2} + h_1^{p+1}$. Based on the definition of $\boldsymbol{\theta}^*$, we can get, wpa1, $\left|h_1^v\left\{\hat{m}_v(x_0) - m^{(v)}(x_0)\right\}\right| = O_p\left((nh_1)^{-1/2} + h_1^{p+1}\right)$. □

Define

$$W_n = \sum_{i=1}^{n} X_i^* K_i \phi'_{h_2}(\epsilon_i). \tag{A.5}$$

We have the following asymptotic representation.

Lemma A.2. *Under conditions (A1)—(A6), it follows that*

$$\hat{\boldsymbol{\theta}}^* - \boldsymbol{\theta}^* = h_1^{p+1}\frac{m^{(p+1)}(x_0)}{(p+1)!}S^{-1}c_p(1+o_p(1)) + \frac{S^{-1}W_n}{nF(x_0,h_2)f(x_0)}(1+o_p(1)). \tag{A.6}$$

Proof. Let

$$\hat{\gamma}_i = R(X_i) - \sum_{j=0}^{p} (\hat{\beta}_j - \beta_j)(X_i - x_0)^j = R(X_i) - (\hat{\boldsymbol{\theta}}^* - \boldsymbol{\theta}^*)^T X^*.$$

Then

$$Y_i - \sum_{j=0}^{p} \hat{\beta}_j (X_i - x_0)^j = \epsilon_i + \hat{\gamma}_i.$$

The solution $\hat{\boldsymbol{\theta}}^*$ satisfies the equation

$$\sum_{i=1}^{n} X_i^* K_i \phi'_{h_2}(\epsilon_i + \hat{\gamma}_i) = \sum_{i=1}^{n} X_i^* K_i \left\{ \phi'_{h_2}(\epsilon_i) + \phi''_{h_2}(\epsilon_i)\hat{\gamma}_i + \frac{1}{2}\phi'''_{h_2}(\epsilon^*)\hat{\gamma}^2 \right\} = 0, \tag{A.7}$$

where ϵ^* is between ϵ_i and $\epsilon_i + \hat{\gamma}_i$. Note that the second term on the left hand side of (A.7) is

$$\sum_{i=1}^{n} K_i \phi''_{h_2}(\epsilon_i) R(X_i) X_i^* - \sum_{i=1}^{n} K_i \phi''_{h_2}(\epsilon_i) X_i^* X_i^{*\prime} (\hat{\boldsymbol{\theta}}^* - \boldsymbol{\theta}^*) \equiv J_1 + J_2. \tag{A.8}$$

Applying Lemma A.1, we obtain

$$J_1 = n h_1^{p+1} F(x_0, h_2) f(x_0) c_p \frac{m^{(p+1)}(x_0)}{(p+1)!} + o_p(n h_1^{p+1}),$$

and

$$J_2 = -nF(x_0, h_2) f(x_0) S (1 + o_p(1))(\hat{\boldsymbol{\theta}}^* - \boldsymbol{\theta}^*).$$

From the Theorem 2.1, we know $||\hat{\boldsymbol{\theta}}^* - \boldsymbol{\theta}^*|| = O_p\{h_1^{p+1} + (nh_1)^{-1/2}\}$, hence

$$\begin{aligned} \sup_{i:|X_i - x_0|/h_1 \le 1} |\hat{\gamma}_i| &\le \sup_{i:|X_i - x_0|/h_1 \le 1} |R(X_i)| + (\hat{\boldsymbol{\theta}}^* - \boldsymbol{\theta}^*)^T X^* \\ &= O_p(h_1^{p+1} + ||\hat{\boldsymbol{\theta}}^* - \boldsymbol{\theta}^*||) = O_p(||\hat{\boldsymbol{\theta}}^* - \boldsymbol{\theta}^*||) = o_p(1), \end{aligned} \tag{A.9}$$

and

$$\sup_{i:|X_i-x_0|/h_1\leq 1} |\hat{\gamma}_i^2| = o_p(1)O_p(||\hat{\boldsymbol{\theta}}^* - \boldsymbol{\theta}^*||) = o_p(||\hat{\boldsymbol{\theta}}^* - \boldsymbol{\theta}^*||).$$

Also similar to the proof of Lemma A.1, we have

$$\begin{aligned}\mathrm{E}\left\{K_i(X_i - x_0)^j/h_1^j\right\} &= \int \frac{1}{h_1}\left\{K\left(\frac{x-x_0}{h_1}\right)\right\}\left(\frac{x-x_0}{h_1}\right)^j f(x)dx \\ &= \mu_j f(x_0) + o(1). \end{aligned} \tag{A.10}$$

Based on (A.9), (A.10), and condition (A6),

$$\begin{aligned}\mathrm{E}\left\{\sum_{i=1}^n K_i\hat{\gamma}_i^2\phi_{h_2}'''(\epsilon^*)(X_i - x_0)^j/h_1^j\right\} &= o_p(||\hat{\boldsymbol{\theta}}^* - \boldsymbol{\theta}^*||)\sum_{i=1}^n \mathrm{E}\left\{K_i(X_i - x_0)^j/h_1^j\right\} \\ &= o_p(n||\hat{\boldsymbol{\theta}}^* - \boldsymbol{\theta}^*||) = o_p(J_2)\end{aligned}$$

and

$$\begin{aligned}&\mathrm{var}\left\{\sum_{i=1}^n K_i\hat{\gamma}_i^2\phi_{h_2}'''(\epsilon^*)(X_i - x_0)^j/h_1^j\right\} \\ &= o_p(n||\hat{\boldsymbol{\theta}}^* - \boldsymbol{\theta}^*||^2)\int \frac{1}{h_1^2}\left\{K\left(\frac{x-x_0}{h_1}\right)\right\}^2\left(\frac{x-x_0}{h_1}\right)^{2j} f(x)dx \\ &= o_p(n^2||\hat{\boldsymbol{\theta}}^* - \boldsymbol{\theta}^*||^2(nh_1)^{-1}).\end{aligned}$$

Hence for the third term on the left-hand side of (A.7),

$$\sum_{i=1}^n K_i\hat{\gamma}_i^2 X_i^*\phi_{h_2}'''(\epsilon^*) = o_p(J_2) + \sqrt{o_p(n^2||\hat{\boldsymbol{\theta}}^* - \boldsymbol{\theta}^*||^2(nh_1)^{-1})} = o_p(J_2).$$

Then, it follows from (A.5) and (A.7) that

$$\hat{\boldsymbol{\theta}}^* - \boldsymbol{\theta}^* = \{F(x_0,h_2)f(x_0)S\}^{-1} h_1^{p+1} F(x_0,h_2)f(x_0)c_p \frac{m^{(p+1)}(x_0)}{(p+1)!}(1+o_p(1)) + \frac{S^{-1}W_n}{nF(x_0,h_2)f(x_0)}(1+o_p(1)),$$

which is (A.6).

Proof of Theorem 2.2. Based on (A.5) and the condition (A6), we can easily get $\mathrm{E}(W_n) = 0$. Similar to the proof in Lemma A.1, we have

$$E\left\{K_i^2 \phi'_{h_2}(\epsilon_i)^2 (X_i - x_0)^j / h_1^j\right\} = h_1^{-1}\nu_j G(x_0,h_2)f(x_0)\{1+o(1)\}.$$

So

$$\mathrm{cov}(W_n) = nh_1^{-1}G(x_0,h_2)f(x_0)S^*(1+o(1)). \tag{A.11}$$

Based on the result (A.6), the asymptotic bias $b_v(x_0)$ and variance of $\hat{m}_v(x_0)$ are naturally given by

$$b_v(x_0) = h_1^{p+1-v}\frac{v!}{(p+1)!}m^{(p+1)}(x_0)e_{v+1}^T S^{-1}c_p + o(h_1^{p+1-v})$$

and

$$\mathrm{var}\{\hat{m}_v(x_0)\} = \frac{v!^2}{nh_1^{1+2v}f(x_0)}G(x_0,h_2)F(x_0,h_2)^{-2}e_{v+1}^T S^{-1}S^*S^{-1}e_{v+1} + o\left(\frac{1}{nh_1^{1+2v}}\right).$$

By simple calculation, we can know the $(v+1)^{th}$ element of $S^{-1}c_p$ is zero for $p-v$ even. So we need higher order expansion of asymptotic bias for $p-v$ even. Following the similar arguments as Lemma A.1, if $nh_1^3 \to \infty$, we can easily prove

$$J_1 = nh_1^{p+1}\left[F(x_0,h_2)f(x_0)c_p\frac{m^{(p+1)}(x_0)}{(p+1)!}\right.$$
$$\left.+h_1\tilde{c}_p\left\{(Ff)'(x_0)\frac{m^{(p+1)}(x_0)}{(p+1)!}+F(x_0,h_2)f(x_0)\frac{m^{(p+2)}(x_0)}{(p+2)!}\right\}\right]\{1+o_p(1)\},$$
$$J_2 = -n\left\{F(x_0,h_2)f(x_0)S+h_1\tilde{S}(Ff)'(x_0)\right\}\{1+o_p(1)\}(\hat{\boldsymbol{\theta}}^*-\boldsymbol{\theta}^*),$$

where J_1 and J_2 is defined in (A.8) and $(Ff)'(x_0) = \frac{\partial F(x,h_2)}{\partial x}|_{x=x_0}f(x_0) + F(x_0,h_2)f'(x_0)$.

Then, it follows from (A.7) that

$$\begin{aligned}\hat{\boldsymbol{\theta}}^*-\boldsymbol{\theta}^* &= h_1^{p+1}\left\{F(x_0,h_2)f(x_0)S+h_1\tilde{S}(Ff)'(x_0)\right\}^{-1}\left[F(x_0,h_2)f(x_0)c_p\frac{m^{(p+1)}(x_0)}{(p+1)!}\right.\\
&\left.+h_1\tilde{c}_p\times\left\{(Ff)'(x_0)\frac{m^{(p+1)}(x_0)}{(p+1)!}+F(x_0,h_2)f(x_0)\frac{m^{(p+2)}(x_0)}{(p+2)!}\right\}\right]\{1+o_p(1)\}\\
&+\frac{f^{-1}(x_0)S^{-1}W_n}{nF(x_0,h_2)}\{1+o_p(1)\}\\
&= h_1^{p+1}\left\{\frac{m^{(p+1)}(x_0)}{(p+1)!}S^{-1}c_p+h_1b^*(x_0)\right\}(1+o_p(1))+\frac{S^{-1}W_n}{nF(x_0,h_2)f(x_0)}(1+o_p(1)),\end{aligned}$$

where

$$\begin{aligned}b^*(x_0) &= F^{-1}(x_0,h_2)f^{-1}(x_0)S^{-1}\tilde{c}_p\left\{(Ff)'(x_0)\frac{m^{(p+1)}(x_0)}{(p+1)!}+F(x_0,h_2)f(x_0)\frac{m^{(p+2)}(x_0)}{(p+2)!}\right\}\\
&+F^{-1}(x_0,h_2)f^{-1}(x_0)(Ff)'(x_0)\frac{m^{(p+1)}(x_0)}{(p+1)!}S^{-1}\tilde{S}S^{-1}c_p \quad .\end{aligned}$$

For $p-v$ even, since the $(v+1)^{th}$ element of $S^{-1}c_p$ and $S^{-1}\tilde{S}S^{-1}c_p$ is zero (see Fan and Gijbels, 1996 for more detail), the asymptotic bias $b_v(x_0)$ of $\hat{m}_v(x_0)$ are naturally given by

$$\begin{aligned}b_v(x_0) =& e_{v+1}^TS^{-1}\tilde{c}_p\frac{v!}{(p+2)!}\left\{m^{(p+2)}(x_0)+(p+2)m^{(p+1)}(x_0)\frac{(Ff)'(x_0)}{F(x_0,h_2)f(x_0)}\right\}h_1^{p+2-v}\\
&+o(h_1^{p+2-v}).\end{aligned}$$

Proof of Theorem 2.3. It is sufficient to show that

$$W_n^* \equiv \sqrt{h_1/n} W_n \xrightarrow{L} N(0, D) \tag{A.12}$$

where $D = G(x_0, h_2) f(x_0) S^*$, because using Slutsky's theorem , it follows from (A.6), (A.12), and Theorem 2.2 that

$$\frac{\hat{m}_v(x_0) - m^{(v)}(x_0) - b_v(x_0)}{\sqrt{\text{var}\{\hat{m}_v(x_0)\}}} \xrightarrow{L} N(0,1).$$

Next we show (A.12). For any unit vector $d \in \mathbb{R}^{p+1}$, we prove

$$\{d^T \text{cov}(W_n^*) d\}^{-\frac{1}{2}} \{d^T W_n^* - d^T E(W_n^*)\} \xrightarrow{L} N(0,1).$$

Let

$$\xi_i = \sqrt{h_1/n} K_i \phi'_{h_2}(\epsilon_i) d^T X_i^*.$$

Then $d^T W_n^* = \sum_{i=1}^n \xi_i$. We check the Lyapunov's condition. Based on (A.11), we can get $\text{cov}(W_n^*) = G(x_0, h_2) f(x_0) S^* (1+o(1))$ and $\text{var}(d^T W_n^* d) = d^T \text{cov}(W_n^*) d = G(x_0, h_2) f(x_0) d^T S^* d (1+ o(1))$. So we only need to prove $nE|\xi_1|^3 \to 0$. Noticing that $(d' X_i)^2 \le ||d||^2 ||X_i||^2, \phi'(\cdot)$ is bounded, and $K(\cdot)$ has compact support,

$$\begin{aligned} nE|\xi|^3 &\le O(nn^{-3/2} h_1^{3/2}) \sum_{j=0}^{p} E \left| K_1^3 \phi'_{h_2}(\epsilon_1)^3 \left(\frac{X_1 - x_0}{h_1} \right)^{3j} \right| \\ &= O(n^{-1/2} h_1^{3/2}) \sum_{j=0}^{p} O(h_1^{-2}) = O((nh_1)^{-1/2}) \to 0 \end{aligned}$$

So the asymptotic normality for W_n^* holds with covariance matrix $G(x_0, h_2) f(x_0) S^*$.

Proof of Theorem 2.4

(a) Note that

$$\phi''_{h_2}(t) = h_2^{-3}(\frac{t^2}{h_2^2} - 1)\phi(t/h_2) \text{ and } \phi'_{h_2}(t) = -\frac{t}{h_2^3}\phi(t/h_2).$$

Based on (2.6), when $h_2 \to \infty$, we have

$$\begin{aligned} h_2^3 F(x, h_2) &= \int (t^2/h_2^2 - 1)\phi(t/h_2) g_{\epsilon|x}(t) dt \to -\phi(0) \\ h_2^6 G(x, h_2) &= \int t^2 \phi^2(t/h_2) g_{\epsilon|x}(t) dt \to \phi^2(0)\sigma^2(x). \end{aligned}$$

Then $G(x, h_2)F^{-2}(x, h_2) = h_2^6 G(x, h_2)/(h_2^3 F(x, h_2))^2 \to \sigma^2(x)$, when $h_2 \to \infty$. So for any x

$$\lim_{h_2 \to \infty} R(x, h_2) = \lim_{h_2 \to \infty} G(x, h_2)F^{-2}(x, h_2)\sigma^{-2}(x) = 1. \tag{A.13}$$

From (A.13), we can get $\inf_{h_2} R(x, h_2) \leq \lim_{h_2 \to \infty} R(x, h_2) = 1$ for any x.

(b) Suppose $g_{\epsilon|x}(t)$ is the density function of $N(0, \sigma^2(x))$. By some simple calculations, we can get

$$F(x, h_2) = h_2^{-3} \int (t^2/h_2^2 - 1)\phi(t/h_2) g_{\epsilon|x}(t) dt = -\frac{1}{\sqrt{2\pi}}(\sigma^2(x) + h_2^2)^{-3/2} \tag{A.14}$$

$$G(x, h_2) = h_2^{-6} \int t^2 \phi^2(t/h_2) g_{\epsilon|x}(t) dt = \frac{\sigma^2(x)}{2\pi h_2^3}(2\sigma^2(x) + h_2^2)^{-3/2} \tag{A.15}$$

Hence

$$R(x, h_2) = G(x, h_2)F(x, h_2)^{-2}\sigma^{-2}(x) = \left(\frac{h_2^4 + 2\sigma^2(x)h_2^2 + \sigma^4(x)}{h_2^4 + 2\sigma^2(x)h_2^2}\right)^{3/2} > 1$$

From (a), we can get $\inf_{h_2} R(x, h_2) = 1$, for any x.

(c) Suppose that $h_2 \to 0$, then

$$
\begin{aligned}
F(x,h_2) &= h_2^{-3}\int (t^2/h_2^2-1)\phi(t/h_2)g_{\epsilon|x}(t)dt \\
&= h_2^{-2}\int (t^2-1)\phi(t)\left\{g_{\epsilon|x}(0)+g'_{\epsilon|x}(0)th_2+\frac{1}{2}g''_{\epsilon|x}(0)t^2h_2^2+o(h_2^2)t^3\right\}dt \\
&= \frac{1}{2}g''_{\epsilon|x}(0)\int (t^4-t^2)\phi(t)dt+o(h_2^2) \\
&= g''_{\epsilon|x}(0)+o(1)\,; \\
h_2^3G(x,h_2) &= h_2^{-3}\int t^2\phi^2(t/h_2)g_{\epsilon|x}(t)dt \\
&= \int t^2\phi^2(t)g_{\epsilon|x}(th_2)dt \\
&= \int t^2\phi^2(t)\left\{g_{\epsilon|x}(0)+o(1)t\right\}dt \\
&= g_{\epsilon|x}(0)\nu_2+o(1)\,.
\end{aligned}
$$

So we can easily see that $R(x,h_2)=G(x,h_2)F(x,h_2)^{-2}\sigma^{-2}(x)\to\infty$ as $h_2\to 0$. □

Acknowledgements

The authors are grateful to the editors and the referees for insightful comments on the article. Bruce G. Lindsay's research was supported by a NSF grant CCF 0936948. Runze Li's research was supported by National Natural Science Foundation of China grant 11028103, NSF grant DMS 0348869, and National Institute on Drug Abuse (NIDA) grant P50-DA10075. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIDA or the NIH.

References

Bai, J. and Perron, P. (2003). Computation and Analysis of Multiple Structural Change Models. *Journal of Applied Econometrics*, 18, 1-22.

Chatterjee, S. and Price, B. (1977). *Regression Analysis by Example.* John Wiley and Sons, New York.

Chu, C. K., Glad, I., Godtliebsen, F., and Marron, J. S. (1998). Edge-Preserving Smoothers for Image Processing (with discussion). *J. Amer. Statist. Assoc.*, 93, 526-556.

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

Fan, J., Hu, T., and Truong, Y. (1994). Robust Non-Parametric Function Estimation. *Scand. J. Statist.*, 21, 433-446.

Fan, J. and Jiang, J. (2000). Variable Bandwidth and One-Step Local M-Estimator. *Sci. China Ser. A*, 43, 65-81.

Frühwirth-Schnatter, S. (2001), Markov Chain Monte Carlo Estimation of Classical and Dynamic Switching and Mixture Models. *J. Amer. Statist. Assoc.*, 96, 194-209.

Goldenshluger, A., Tsbakov, A., and Zeevi, A. (2006). Optimal Change-Point Estimation from Indirect Observations. *Ann. Statist.*, 34, 350-372.

Goldfeld, S. M. and Quandt, R. E. (1976). A Markov Model for Switching Regression. *Journal of Econometrics*, 1, 3-16.

Green, P. J. and Richardson, S. (2002). Hidden Markov Models and Disease Mapping. *J. Amer. Statist. Assoc.*, 97, 1055-1070.

Hall, P. and Jones, M. C. (1990). Adaptive M-Estimation in Nonparametric Regression. *Ann. Statist.*, 18, 1712-1728.

Härdle, W. and Gasser, T. (1984). Robust Nonparametric Function Fitting. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 46, 42-51.

Härdle, W. and Tsybakov, A. B. (1988). Robust Nonparametric Regression with Simultaneous Scale Curve Estimation. *Ann. Statist.*, 16, 120-135.

Jiang, J. and Mack, Y. (2001). Robust Local Polynomial Regression for Dependent Data. *Statist. Sinica*, 11, 705-722.

Lai, T. L. (2001). Sequential Analysis: Some Classical Problems and New Challenges. *Statist. Sinica*, 11, 303-408.

Li, J., Ray, S., and Lindsay, B. G. (2007). A Nonparametric Statistical Approach to Clustering via Mode Identification. *J. Mach. Learn. Res.*, 8(8), 1687-1723.

Rossi, P. E., Allenby, G. M., and McCulloch, R. (2005). *Bayesian Statistics and Marketing.* Chichester: Wiley.

Ruppert, D., Sheather S. J., and Wand M. P. (1995). An Effective Bandwidth Selector for Local Least Squares Regression. *J. Amer. Statist. Assoc.*, 90, 1257-1270.

Scott, D. W. (1992). *Multivariate Density Estimation: Theory, Practice and Visualization.* New York: Wiley.

Tibshirani, R. J. and Hastie, T. J. (1987). Local Likelihood Estimation. *J. Amer. Statist. Assoc.*, 82, 559-567.

Tsybakov, A. B. (1986). Robust Reconstruction of Functions by the Local Approximation Method. *Problems in Information Transmission*, 22, 133-146.