International Journal of Statistical Sciences *ISSN 1683–5603*
Vol. 9(Special Issue), 2009, pp 217-231

On the Robust Modal Local Polynomial Regression

Weixing Song[1]**, Haiyan Wang and Weixin Yao**
Department of Statistics
Kansas State University
Manhattan, KS, 66502

[Received December 11, 2008; Accepted February 21, 2009]

Abstract

Modal local polynomial regression uses double kernel as the loss function to gain some robustness in the nonparametric regression. Current researches use the standard normal density function as the weight function to down-weigh the influences from the outliers. This paper extends the standard normal weight function to a general class weight functions. All the theoretical properties found by using normal weight function are preserved with general weight function without imposing additional conditions.

Keywords and Phrases: Robustness, Mean Squares Error, Local Polynomial, Nonparametric Smoothing, EM Algorithm.

AMS Classification: 62G08; 62G35; 62G99.

1 Introduction

Estimating the regression function $m(x) = E(Y|X)$ in the following classical regression model

$$Y = m(X) + \varepsilon \tag{1.1}$$

has been, and will continue to be, a long-standing interest among theoretical statisticians and practitioners as well. If the regression function $m(x)$ has a parametric form, that is, $m(x) = m(x;\beta)$, the research interest mainly focuses on the estimation of the parameters.

[1]The research of this author is supported by the USRG No. 2232 of Kansas State University

Least squares and maximum likelihood are the most commonly used estimation procedures. But some estimation procedures, like least squares, can be seriously affected by the outliers or some other unusual data points. To overcome these drawbacks, many remedial measures have been taken to make up this disadvantage, such as the M-estimation procedure, the minimum L_1-norm regression, the iteratively reweighted least squaress robust regression, the least median of squares regression, quantile regression and other robust regression procedures involving ranks or trimming one or several of the extreme squared deviations before applying the least squares criterion. For a comprehensive understanding of these topics, see Huber (1974), Härdle (1992), Chatterjee and Mächler (1995), Rousseeuw (1984), Rousseeuw and Croux (1993), Wilcox (1997), and the references therein. The discussion on rank related robust estimation procedure can be found in He and Liang (2000), Saleh and Shiraishi (1993), Chen and Saleh (2000) and the monographes by Koenker (2005), Saleh (2006). Also see Koul and Saleh (1993,1995), Alimoradi and Saleh (1999) for the extension of these methods to other important statistical models.

When $m(x)$ is unknown, some well known nonparametric estimation procedures, such as the Nadaraya-Watson estimator, Gasser-Müller estimator, local polynomial, spline and orthogonal series estimators can be used to estimate $m(x)$. Like the parametric case, the least squares ingredient embedded in all the above procedures contributes their lack of robustness. It is known that the M-type of regression estimators are natural candidates for obtaining robustness properties. The nonparametric M-estimators or local M-estimators for regression functions have been studied by many authors, see Cleveland, W.S. (1979), Tsybakov, A.B. (1986), Fan, Hu, and Truong (1994) and the references therein. Although M-estimators inherit many desirable properties from the local least squares regression, they are defined implicitly and numerical implementation requires an iterative scheme, hence computationally extensive. To overcome this drawback, Fan and Jiang (2000) proposed a one-step local M-estimator with a variable bandwidth. They showed that it shares the same computational expediency as the local least squares estimator, and possesses the same asymptotic performance as the local M-regression estimator when the initial estimator behaves well. Recently, Yao, Lindsay and Li (2008) (YLL) use a normal density as the outlier-resistent weight function to construct the M-estimator, called modal local estimator, for the regression function. Their method not only provides a new understanding of M-type estimation procedure, but also the computation burden is greatly eased by employing the EM algorithm. It is also worth to mention that their procedure does not use variable bandwidth. One interesting question related to their study is that if there exists any other choices of the outlier-resistent weight function, so that the resulting estimators have smaller mean squares errors. The current paper will try to answer this question.

The paper is organized as follows. The model local constant and local polynomial regression procedure will be introduced briefly in Section 2; and the modal local polynomial procedure with general outlier-resistent weight function will be constructed in Section 3, some theoretical results will be also presented there; Section 4 presents some criteria on selecting the optimal bandwidth and the weight function; the proofs of all main results are postponed to Section 5.

2 Modal Local Constant and Local Polynomial

Let $\{(X_i, Y_i) : i = 1, 2, \cdots, n\}$ be independent and identical copies from a population (X, Y) having joint density function $f(x, y)$. Modal regression method uses the mode of the conditional density function $f(y|x)$ to estimate the regression function $m(x) = E(Y|X = x)$. The relationship $f(y|x) = f(x, y)/f(x)$ implies that the modes of $f(y|x)$ are the same as the modes of $f(x, y)$ with x fixed. An estimation is needed if $f(x, y)$ is unknown. The commonly used nonparametric estimator of $f(x, y)$ is the kernel density estimator given as follows:

$$\hat{f}(x, y) = \frac{1}{n} \sum_{i=1}^{n} K_{h_1}(X_i - x) K_{h_2}(Y_i - y), \tag{2.1}$$

where $K_h(x) = K(x/h)/h$ with $K(x)$ being a symmetric density function, and h_1, h_2 being the bandwidths. Then an estimator of $m(x)$ will be defined as the mode of $\hat{f}$(x,y). Since $\hat{f}(x, y)$ is in fact a mixture distribution, a modified EM algorithm is employed to find the mode, see Li, et al.(2007).

The above procedure for estimation the regression function $m(x)$ is called the modal local constant method. For any given x, the mode of $\hat{f}(x, y)$, as a function of x, is estimated by a constant. In a similar fashion as is done to extend the Nadaraya-Watson estimator to local polynomial estimator in estimating the regression function, YLL extends the modal local constant regression to the modal local polynomial regression. In fact, if the $(p + 1)$-th derivative of $m(x)$ at the point x_0 exists, the unknown regression function $m(x)$ can be estimated locally by a polynomial of order p. This polynomial is fitted by maximizing the following formula

$$\sum_{i=1}^{n} K_{h_1}(X_i - x_0) \phi_{h_2}\Big(Y_i - \sum_{j=0}^{p} \beta_j (X_i - x_0)^j\Big) \tag{2.2}$$

over $\beta = (\beta_0, \beta_1, \cdots, \beta_p)$. Denote the solution to the above maximizing problem as $\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1, \cdots, \hat{\beta}_p)$. The regression function $m(x)$ at $x = x_0$ is estimated by $\hat{\beta}_0$. The kernel function $K(\cdot)$ is chosen to be standard normal. To simplify the M-step in the EM algorithm, they deliberately choose the weight function $\phi(\cdot)$ to be the standard normal density. Under some conditions, they obtained the expressions for the bias and variance of $\hat{\beta}_v$, $v = 0, 1, \cdots, p$. Normality of these estimators are also proven under a minimal set of regularity conditions. Theoretical results and simulation studies suggest that the above modal local polynomial procedure is no worse than local polynomial regression based on the mean squares error criterion. In some cases, the modal local polynomial can have smaller asymptotic mean squares error than the local polynomial regression by carefully choosing the bandwidth in the weight function. The result (c) in their Theorem 4.2 claims that if the error term ε has a normal distribution, the optimal modal local polynomial will be the same as the local polynomial regression.

Noticing that the kernel function $K(\cdot)$ and the weight function $\phi(\cdot)$ are both set to be standard normal in YLL's paper, we become interested in the performance of the modal local polynomial regression if the weight function is not restricted to the standard normal.

One may notice that the modal local polynomial procedure and the general modal local polynomial procedure are similar to the local likelihood and quantile regression procedures. The quantile regression is a robust regression procedure, the general framework is the following. Let $G(y|x)$ be the conditional distribution of Y given $X = x$. The conditional α-quantile function is $\xi_\alpha(x) = G^{-1}(\alpha|x)$. The quantile function and its derivatives are estimated by minimizing

$$\sum_{i=1}^{n} K_h(X_i - x_0) l_\alpha[Y_i - \beta_0 - \beta_1(X_i - x_0) - \cdots - \beta_p(X_i - x_0)^p]$$

where $l_\alpha(t) = |t| + (2\alpha - 1)t$. The resulting estimator for $\xi_\alpha^{(v)}(x_0)$ is simply $v!\hat{\beta}_v$. Now, suppose $\alpha = 0.5$ and G has a unimodal density function which is symmetric around 0, then one may realize that the resulting 50-th quantile estimator is also the estimator of the conditional expectation of Y given $X = x_0$. But as we know, the computational loads for finding the quantile estimator in some cases are formidable, while the modal local or the general modal local regression procedure is superior than the quantile regression in this regard.

3 General Modal Local Polynomial and Main Results

By loosening the restriction on the weight function, we will maximize the following formula

$$\sum_{i=1}^{n} K_{h_1}(X_i - x_0) W_{h_2}\Big(Y_i - \sum_{j=0}^{p} \beta_j (X_i - x_0)^j\Big) \tag{3.1}$$

over the range of $\beta = (\beta_0, \beta_1, \cdots, \beta_p)$, where $K(\cdot)$ and $W(\cdot)$ function are not necessarily to be standard normal. Still denote the solution to the maximizing problem (3.1) as $\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1, \cdots, \hat{\beta}_p)$. Correspondingly, $\hat{m}_v(x_0) = v!\hat{\beta}_v$ is the estimator of the v-th derivative of $m(x)$ at $x = x_0$.

The basic assumptions on the error term ε in the model (1.1) will be the same as in YLL, such as $E(\varepsilon) = 0$, ε has a continuous density function. In this paper we will assume the homoscedasticity of ε. The discussion on the heteroscedastic case will be straightforward. The marginal density function of X and ε will be denoted as $f(\cdot)$ and $g(\cdot)$, respectively.

The moments of the kernel function $K(\cdot)$ and $K^2(\cdot)$ are denoted respectively by

$$\mu_j = \int t^j K(t)dt, \quad \nu_j = \int t^j K^2(t)dt.$$

The following matrices and vectors are also used in the subsequent theorems and theoretical argument.

$$S = (\mu_{j+l})_{0\le j,l\le p}, \quad \tilde{S} = (\mu_{j+l+1})_{0\le j,l\le p}, \quad S^* = (\nu_{j+l})_{0\le j,l\le p},$$

$$c_p = (\mu_{p+1}, \cdots, \mu_{2p+1})^T, \quad \tilde{c}_p = (\mu_{p+2}, \cdots, \mu_{2p+2})^T,$$
$$e_{v+1} = (0, \cdots, 0, 1, 0, \cdots, 0)^T \text{ with 1 on the } (v+1)\text{-th position.}$$

The conditions on the regression function $m(\cdot)$, design density function $f(\cdot)$ and error term density function $g(\cdot)$ are identical to those of YLL. We list all those conditions below for the sake of completeness:

(m) $m(x)$ has continuous bounded $(p+2)$-th derivative.

(f) $f(x)$ has continuous and bounded first and second derivatives.

(g1) $g(x)$ is unimodal, symmetric around 0, and has variance σ^2.

(g2) $g(x)$ has bounded first and second derivatives.

About the kernel function $K(\cdot)$, the weight function $W(\cdot)$, and the bandwidth h_1, we shall assume

(k) $K(x)$ is symmetric around 0 and has finite $(2p+2)$-th moments. Also $\int t^p K^2(t)$ is finite.

(w) $W(\cdot)$ is twice differentiable with $W'(0) = 0, \;\; W''(0) \neq 0.$

(h1) $h_1 \to 0$, $nh_1^3 \to \infty$ as $n \to \infty$.

The asymptotic bias and variance of the general modal local polynomial estimator of $\hat{m}_v(x_0)$ are given by the following theorem.

Theorem 3.1. *Under the assumption (m), (f), (g1), g(2), (k), (w), and (h1), the asymptotic variance of $\hat{m}_v(x_0)$ is given by*

$$\mathit{Var}(\hat{m}_v(x_0)) = \frac{h_2 v!^2 G(h_2) e_{v+1}^T S^{-1} S^* S^{-1} e_{v+1}}{F^2(h_2) f(x_0) n h_1^{2v+1}} + o\Big(\frac{1}{n h_1^{2v+1}}\Big), \tag{3.2}$$

where $F(h) = \int W''(t) g(th) dt, \;\; G(h) = \int (W'(t))^2 g(th) dt$. The asymptotic bias when $p-v$ odd is given by

$$\mathit{Bias}(\hat{m}_v(x_0)) = \frac{e_{v+1}^T S^{-1} c_p v! m^{(p+1)}(x_0) h_1^{p-v+1}}{(p+1)!} + o(h_1^{p-v+1}), \tag{3.3}$$

and when $p - v$ even is given by

$$\mathit{Bias}(\hat{m}_v(x_0)) = \frac{e_{v+1}^T S^{-1} \tilde{c}_p v! h_1^{p-v+2} b_p(x_0)}{(p+2)!} + o(h_1^{p-v+2}), \tag{3.4}$$

where

$$b_p(x_0) = m^{(p+2)}(x_0) + \frac{(p+2) m^{(p+1)}(x_0) f'(x_0)}{f(x_0)}$$

The above theorem reproduces YLL's result for the general weight function W. One can see that the asymptotic bias of the general modal local polynomial is the same as that of local polynomial regression and the asymptotic variance is $h_2G(h_2)F(h_2)^{-2}/\sigma^2$ times that of the local polynomial regression. In YLL's case, one can seek h_2 such that $h_2G(h_2)F(h_2)^{-2}/\sigma^2 < 1$ to make the asymptotic mean squares error smaller than the local polynomial regression. Our general modal local polynomial method provides an additional flexibility, that is, we can select a proper weight function W to achieve the same effect.

Theorem 4.2 in YLL states the properties of $hG(h)/F^2(h)$ when the weight function is standard normal. The following Theorem shows that the same results still true for the general weight function W.

Theorem 3.2. *Let $F(h)$, $G(h)$ be the same as in Theorem 3.1. Suppose the condition (w) holds. Then we have*

(a). $\lim_{h\to\infty} hG(h)F^{-2}(h) = \sigma^2$*;*

(b). $\inf_h hG(h)F^{-2}(h) \le \sigma^2$.

YLL also shows that if W and g are both standard normal, then $hG(h)F^{-2}(h) > \sigma^2$ for any finite h and $\inf_h hG(h)F^{-2}(h) = \sigma^2$. Since h_2 can not be infinite in the real application, so as far as the mean squares errors be concerned, the performance of the modal local polynomial will be inferior to the local polynomial regression if W is chosen to be standard normal. In the general modal local polynomial set up, we are free to select the form of the weight function W subject to assumption (w). Are there any weight function W which can make $\inf_h hG(h)F^{-2}(h) < \sigma^2$? More discussion on this can be found in the next section.

The asymptotic normality of $\hat{m}^{(v)}(x_0)$ can also be proved in a similar fashion as in YLL. Only small modification is needed to accommodate the substitution of the general weight function for the standard normal weight function.

Theorem 3.3. *Under same assumptions of Theorem 3.2,*

$$\sqrt{nh_1}\Big[\hat{m}^{(v)}(x_0) - m^{(v)}(x_0) - h_1^{p+1}b(x_0) + O(h_1^{p+2})\Big] \Longrightarrow N(0, \sigma_v^2(x_0)).$$

where

$$b(x_0) = F(h_2)\beta_{p+1}f(x_0)e_{v+1}^T\Delta^{-1}c_p/h_2^2, \;\; \sigma_v^2(x_0) = e_{v+1}^T\Delta^{-1}\Lambda\Delta^{-1}e_{v+1}$$

and $\Lambda = G(h_2)f(x_0)S^*/(nh_1h_2^3)$.

The proof of above theorem follows a routine manner. For the sake of brevity, the details will not be given here, see Fan and Gijbels (1997) for a full reference.

4 Bandwidth and Weight Function Selection

As mentioned before, by selecting a proper h_2, YLL's modal local polynomial estimator will have smaller means square error than the local polynomial estimator, and more robust than

the local polynomial estimator if outliers present or the error distribution has heavy tails. While the general modal local polynomial estimator proposed in this paper has an additional flexibility of selecting the weight function besides h_2. This extra flexibility may result in a better estimator than that of modal local polynomial and local polynomial in some cases.

4.1 Optimal Bandwidth

The theoretical optimal local bandwidth for estimating $m^{(v)}(\cdot)$ is obtained by minimizing the mean squares error which is the sum of squared bias and variance of the estimator $m^{(v)}(\cdot)$. Theorem 3.1 provides us some results of the asymptotic bias and variance, so the asymptotic optimal local bandwidth can be obtained by minimizing the asymptotic mean squares error. Since the asymptotic means square error is similar to that of YLL except the definitions of the functions F and G, so the following theorem about the optimal local bandwidth can be obtained exactly same as that of YLL. We only state the theorems without any proof.

Theorem 4.1. *For the general modal local polynomial regression, the asymptotic optimal local bandwidth* (h_1, h_2) *by minimizing the asymptotic mean squares error based on Theorem 3.1 are*

$$\begin{aligned} h_{2,opt} &= \arg\min_h hG(h)F^{-2}(h), \\ h_{1,opt} &= \begin{cases} C_1(p,v)n^{-1/(2p+3)} & \text{if } p-v \text{ is odd,} \\ C_2(p,v)n^{-1/(2p+5)} & \text{if } p-v \text{ is even,} \end{cases} \end{aligned}$$

where

$$\begin{aligned} C_1(p,v) &= \left[\frac{(2v+1)v!^2(p+1)!^2 e_{v+1}^T S^{-1}S^*S^{-1}e_{v+1}h_{2,opt}G(h_{2,opt})}{n(2p-2v+2)f(x_0)(e_{v+1}^T S^{-1}c_p v! m^{(p+1)}(x_0))^2 F^2(h_{2,opt})}\right]^{1/(2p+3)}, \\ C_2(p,v) &= \left[\frac{(2v+1)v!^2(p+1)!^2 e_{v+1}^T S^{-1}S^*S^{-1}e_{v+1}h_{2,opt}G(h_{2,opt})}{n(2p-2v+2)f(x_0)(e_{v+1}^T S^{-1}\tilde{c}_p v! b_p(x_0))^2 F^2(h_{2,opt})}\right]^{1/(2p+3)}. \end{aligned}$$

In particular, for the local linear case, $p = 1$ *and* $v = 0$*, then*

$$C_1(1,0) = \left[\frac{\nu_0 h_{2,opt}G(h_{2,opt})}{nf(x_0)(m^{(2)}(x_0))^2 F^2(h_{2,opt})}\right]^{1/5}.$$

The global bandwidth is obtained by minimizing the mean integrated square error of the form

$$\int \Big([\text{Bias}(\hat{m}^{(v)}(x))]^2 + \text{Var}(\hat{m}^{(v)}(x))\Big) dK(x)$$

with the integrating positive measure K.

Theorem 4.2. *For the general modal local polynomial regression, the asymptotic optimal global bandwidth* (h_1, h_2) *by minimizing the asymptotic mean integrated square error based on Theorem 3.1 are*

$$\begin{aligned} h_{2,opt} &= \arg\min_h hG(h)F^{-2}(h), \\ h_{1,opt} &= \begin{cases} C_1(p,v)n^{-1/(2p+3)} & \text{if } p-v \text{ is odd,} \\ C_2(p,v)n^{-1/(2p+5)} & \text{if } p-v \text{ is even,} \end{cases} \end{aligned}$$

where

$$\begin{aligned} C_1(p,v) &= \left[\frac{(2v+1)v!^2(p+1)!^2 e_{v+1}^T S^{-1}S^*S^{-1}e_{v+1}h_{2,opt}G(h_{2,opt})I_1}{n(2p-2v+2)(e_{v+1}^T S^{-1}c_p v! m^{(p+1)}(x_0))^2 F^2(h_{2,opt})}\right]^{1/(2p+3)}, \\ C_2(p,v) &= \left[\frac{(2v+1)v!^2(p+1)!^2 e_{v+1}^T S^{-1}S^*S^{-1}e_{v+1}h_{2,opt}G(h_{2,opt})I_1}{n(2p-2v+2)(e_{v+1}^T S^{-1}\tilde{c}_p v! I_2 F^2(h_{2,opt})}\right]^{1/(2p+3)}, \\ I_1 &= \int 1/f(x)dK(x), \quad I_2 = \int b^2(x)dK(x). \end{aligned}$$

In particular, for the local linear case, $p = 1$ *and* $v = 0$*, then*

$$C_1(1,0) = \left[\frac{\nu_0 h_{2,opt}G(h_{2,opt})I_1}{n\mu_2^2 F^2(h_{2,opt})I_2}\right]^{1/5}$$

Since the asymptotic optimal bandwidths depend on some unknown quantities such as the error density g, the design density f and the second derivative of the regression function m, so we have to estimate them in order to find the plug-in type bandwidth. As YLL points out that there is no explicit solution for the optimal h_2, so one can use the grid search method. For these purposes, they get the residuals $\hat{\varepsilon}_i = y_i - \hat{m}(x_i)$ by fitting the data using some smoothing methods, such as, kernel, local polynomial etc.. If W is selected to be normal, YLL proposes two methods which do not involve any integration. Since W may not be normal in our case, so the first method does not work, but the second method is applicable. In fact, notice that

$$\begin{aligned} F(h) &= \int W''(t)g(th)dt = \frac{1}{h}E[W''(\varepsilon/h)], \\ G(h) &= \int (W'(t))^2 g(th)dt = \frac{1}{h}E[W'(\varepsilon/h)]^2, \end{aligned} \tag{4.1}$$

By approximating the above expectation by the sample mean, we can choose h_2 to minimize $h_2\hat{G}(h_2)/\hat{F}^2(h_2)$, where

$$\hat{F}(h) = \frac{1}{nh}\sum_{i=1}^n W''(\hat{\varepsilon}_i/h), \quad \hat{G}(h) = \frac{1}{nh}\sum_{i=1}^n [W'(\hat{\varepsilon}_i/h)]^2. \tag{4.2}$$

Now we can use the grid method suggested by YLL to look for the optimal h_2 and h_1 based on the Theorem 4.1 and 4.2.

4.2 Weight Function Selection

By Theorem 3.1, we know that the asymptotic bias of the general modal local polynomial estimators is the same as that of the local polynomial regression and YLL's modal local polynomial regression, while the asymptotic variance of the general modal local is different from that of the local polynomial regression in that there is an extra factor $h_2G(h_2)/F^2(h_2)$ in the general modal local polynomial case, it is also different from the asymptotic variance of the YLL's modal local polynomial in that the weight function in G and F functions is restricted to standard normal in YLL's case. So our general modal local polynomial not only equips with the ability to keep the robustness which is inherited from the YLL's idea, but also achieves a smaller mean squares error than the modal local polynomial regression procedure or the local polynomial regression procedure if the weight function W is properly chosen. The latter fact stands up as our unique innovative contribution comparing to the existing literatures.

As an example, recall the result (c) of Theorem 4.2 in YLL. If the error density is a normal distribution with mean 0 and variance σ^2, then for any finite h_2, $h_2G(h_2)/F^2(h_2) > \sigma^2$. This implies that YLL's modal local polynomial regression procedure is always inferior to the local polynomial regression procedure in that the former has a larger asymptotic mean squares error than the latter no matter how to choose h_2. Now if we let $W(x) = ca^{-x^2/2}$, where c is the normalizing constant to make $W(x)$ to be a density function. Also let $MSE_W(h)$ be the asymptotic mean squares error corresponding to above $W(x)$ with $h_2 = h$, let $MSE_\phi(h)$ be the asymptotic mean squares error corresponding to the standard normal density function with $h_2 = h$. Then for any finite $h > 0$ and σ^2, $\sigma^2 < MSE_W(h) < MSE_\phi(h)$ if $1 < a < e$. In fact, this example is only for the purpose of illustration, because $W(x) = ca^{-x^2/2}$ is also a normal density function with mean 0 and variance $1/\sqrt{\ln a}$. It is easy to see $MSE_W(h\sqrt{\ln a}) = MSE_\phi(h)$.

As another example, we consider the following Bisquare weight function

$$B(x) = \begin{cases} [1-(x/4.685)^2]^2 & |x| \le 4.685 \\ 0 & |x| > 4.685 \end{cases}$$

$B(x)$ is often used in the iterative reweighted least squares robust regression procedure. It is not twice differentiable, so it is not a candidate for $W(x)$. We can use the interpolation technique to make this function smoother such that the second derivative exists, in fact, the following modified version of $B(x)$, denoted by $W(x)$, satisfies the condition (w).

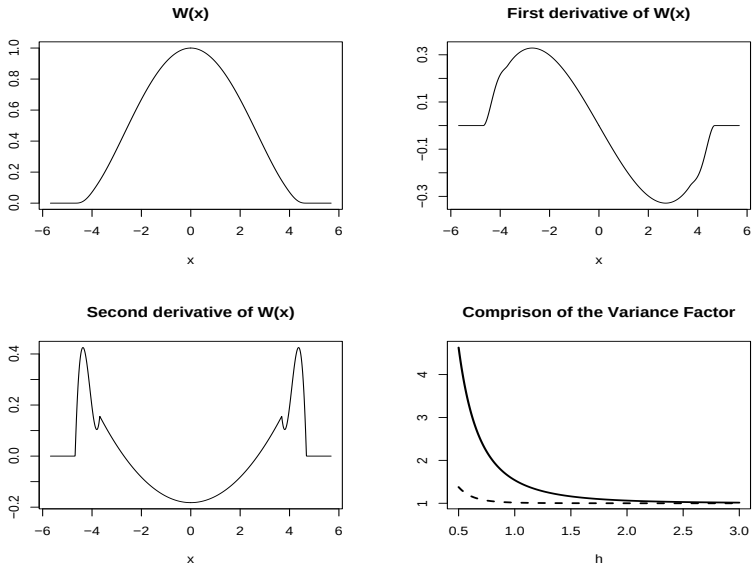
$$W(x) = \begin{cases} 0 & x < -4.865 \\ B_1(x) & -4.865 \le x < -4.865 + \eta \\ B(x) & -4.685 + \eta \le x < 4.685 - \eta \\ B_2(x) & 4.685 - \eta \le x < 4.685 \\ 0 & x \ge 4.685 \end{cases}$$

where $B_1(x)$ and $B_2(x)$ are polynomials of order 5 and $B_1(x) = B_2(-x)$ which coefficients

depend on η. For example, if $\eta = 1$, then

$$B_2(x) = -0.182x^5 + 3.724x^4 - 30.301x^3 + 122.811x^2 - 248.397x + 201.158.$$

The following figure shows the plots of $W(x)$ with $\eta = 1$, the first and the second derivative of $W(x)$, also the factor $hG(h)/F^2(h)$ for YLL's modal local polynomial and our General modal local polynomial. The solid curve is the variance factor from YLL, and the dashed line is the variance factor from our general modal local polynomial procedure. We can see that the factor in our case is uniformly smaller than that of YLL's case, in particular for smaller h, which implies that our method has smaller mean squares error.



We are also considering the existence of W, with condition (w), such that $MSE_W < \sigma^2$ for some finite h. If such W exists, then in the sense of mean squares error, our procedure will do a better job than the local polynomial regression procedure. Clearly such W function can not be the form of $W(x) = ca^{-x^2/2}$. In fact, the inequality $MSE_W < \sigma^2$ holds if and only if

$$E[W'(X/h)]^2 < [EW''(X/h)]^2,$$

where the expectations are taken under the normal distribution with mean 0 and variance σ^2. We cannot give a determined answer to this issue. All the functions we tried only support the converse inequality. This will be our future research topic.

5 Proofs of the Main Results

The proof of Theorem 3.1 is similar to that of Theorem 4.1 in YLL, only the difference is stated here.

Proof of Theorem 3.1. Let $\theta = (\beta_0, h_1\beta_1, \ldots, h_1^p\beta_p)^T$, $\hat{\theta} = (\hat{\beta}_0, h_1\hat{\beta}_1, \ldots, h_1^p\hat{\beta}_p)^T$,

$$\theta_0 = (m(x_0), h_1 m'(x_0), \ldots, h_1^p m^{(p)}(x_0)/p!)$$

and

$$Z_i = \left(1, \frac{X_i - x_0}{h_1}, \ldots, \frac{(X_i - x_0)^p}{h_1^p}\right).$$

Since $\hat{\beta}$ maximizes (3.1), $\hat{\theta}$ will maximize

$$l_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} K_{h_1}(X_i - x_0) W_{h_2}(Y_i - \theta^T Z_i).$$

Denote

$$\begin{aligned} W_n &\hat{=} \frac{\partial l_n(\theta_0)}{\partial \theta} = -\frac{1}{nh_2^2}\sum_{i=1}^{n} K_{h_1}(X_i - x_0) W'\left(\frac{Y_i - \theta^T Z_i}{h_2}\right) Z_i, \\ \Delta_n &\hat{=} \frac{\partial^2 l_n(\theta_0)}{\partial \theta^2} = \frac{1}{nh_2^3}\sum_{i=1}^{n} K_{h_1}(X_i - x_0) W''\left(\frac{Y_i - \theta^T Z_i}{h_2}\right) Z_i Z_i', \end{aligned}$$

where W', W'' are the first and second derivative of W.

Define

$$S_{j,k} = K_{h_1}(X - x_0)\left(\frac{X - x_0}{h_1}\right)^j W_{h_2}^{(k)}(Y_i - \theta_0' Z_i).$$

Then by changing variables in the integration, the expectation of $S_{j,k}$ is given by

$$ES_{j,k} = \int s^j K(s)\left[\int W^{(k)}(t) g(\theta_0^T s + th_2 - m(x_0 + sh_1))dt\right] f(x_0 + sh_1)ds,$$

where $W^{(k)}$ is the k-th derivative of W, g and f are the density functions of the error term and the design variable.

By the definition of θ_0 and the conditions (m), (g2), one has

$$\begin{aligned} &\theta_0^T s - m(x_0 + sh_1) = O(h_1^{p+1}), \\ &g(th_2 + \theta_0^T s - m(x_0 + sh_1)) = g(th_2) + O(h_1^{p+1}). \end{aligned} \tag{5.1}$$

So if k is even,

$$ES_{j,k} = [\mu_j f(x_0) + h_1\mu_{j+1} f'(x_0)] \int W^{(k)}(t) g(th_2)dt + O(h_1^2). \tag{5.2}$$

If k is odd, the integration in (5.2) vanishes, one needs to obtain some high order terms in the expansion (5.1). Let $b_p(s) = \beta_{p+1}(sh_1)^{p+1} + \beta_{p+2}(sh_1)^{p+2}$, then

$$g(\theta_0^T s + th_2 - m(x_0 + sh_1)) = g(th_2) - b_p(x)g'(th_2) + O(h_1^{p+3}).$$

Using this expansion, and let $H_k(h_2) = -\int W^{(k)}(t)g'(th_2)dt$, we have

$$ES_{j,k} \quad = \quad -H_k(h_2)h_1^{p+1}R(x_0, h_1),$$

where

$$R(x_0, h_1) = \mu_{j+p+1}\beta_{p+1}f(x_0) + h_1\mu_{j+p+2}(\beta_{p+1}f'(x_0) + \beta_{p+2}f(x_0)) + O(h_1^2).$$

Now, define

$$T_{j,k} = K_{h_1}^2(X - x_0)\left(\frac{X - x_0}{h_1}\right)^2\left[W_{h_2}^{(k)}(Y - \theta_0^T Z)\right]^2.$$

Then by changing variables in integration, one has

$$ET_{j,k} = \frac{1}{h_1 h_2}\left[f(x_0)\nu_j\int [W^{(k)}(t)]^2 g(th_2)dt + o(h_1)\right].$$

By the Taylor expansion of $l_n(\theta)$ at $\theta = \theta_0$, the maximizer $\hat{\theta}_n$ of $l_n(\theta)$satisfies

$$\hat{\theta}_n - \theta_0 = -\Delta_n^{-1}W_n + O_p(\|\hat{\theta}_n - \theta_0\|^2).$$

Note that

$$E(\Delta_n) = h_2^{-2}E(S_{i+j,2})_{0\le i,j\le p} = \frac{F(h_2)}{h_2^2}\left[f(x_0)S + f'(x_0)h_1\tilde{S}\right] + O(h_1^2)$$

and Var$(S_{j,k})/n = O(1/nh_1)$, easy to see that

$$\Delta_n = \frac{F(h_2)}{h_2^2}\left[f(x_0)S + f'(x_0)h_1\tilde{S}\right] + O_p\left(h_1^2 + \frac{1}{\sqrt{nh_1}}\right).$$

Also

$$EW_n = h_2^{-1}E(S_{i,1})_{0\le i\le p}$$

$$= h_2^{-1}H_1(h_2)h_1^{p+1}\left[\beta_{p+1}f(x_0)c_p + h_1(\beta_{p+1}f'(x_0) + \beta_{p+2}f(x_0))\tilde{c}_p\right] + O(h_1^{p+3}).$$

By the well known matrices result $(A + hB)^{-1} = A^{-1} - hA^{-1}BA^{-1} + O(h^2)$ and

$$\begin{aligned} H_1(h_2) &= -\int W'(t)g'(th_2)dt \\ &= \frac{1}{h_2}g(th_2)W'(t)\Big|_{-\infty}^{\infty} + \frac{1}{h_2}\int g(th_2)W''(t)dt = \frac{F(h_2)}{h_2}, \end{aligned}$$

the asymptotic expansion for the bias term will be

$$\text{Bias}(\hat{\theta}) = h_1^{p+1}\left[\beta_{p+1}S^{-1}c_p + h_1 B(x_0) + O\left(h_1^2 + \frac{1}{\sqrt{nh_1}}\right)\right],$$

where

$$B(x_0) = \frac{\beta_{p+1}f'(x_0) + \beta_{p+2}f(x_0)}{f(x_0)}S^{-1}\tilde{c}_p - \frac{f'(x_0)}{f(x_0)}\beta_{p+1}S^{-1}\tilde{S}S^{-1}c_p.$$

Recall that $\hat{\theta}_v = h_1^v\hat{m}^{(v)}/v!$ and the $(v+1)$-th element of $S^{-1}\tilde{c}_p$ and of $S^{-1}\tilde{S}S^{-1}c_p$ are 0 for $p-v$ odd, and the $(v+1)$-th element of $S^{-1}c_p$ and of $S^{-1}\tilde{S}S^{-1}c_p$ are 0 for $p-v$ even, we obtain the results of (3.3) and (3.4).

Notice that

$$\begin{aligned}\text{Var}(W_n) &= \frac{1}{nh_2^2}E(T_{i+j,1})_{0\le i,j\le p} + o(1/nh_1)\\ &= \frac{1}{nh_1h_2^3}f(x_0)G(h_2)S^* + o(1/nh_1).\end{aligned}$$

So the asymptotic expansion for the variance $\hat{\theta}$ is given by

$$\text{Var}(\hat{\theta}) = \Delta_n^{-1}\text{Var}(W_n)\Delta_n^{-1} = \frac{h_2G(h_2)}{nh_1f(x_0)}S^{-1}S^*S^{-1} + o(1/nh_1)$$

which implies the result (3.2). Other claims can be shown similarly. Also see Fan and Jiang (2000). □

Proof of Theorem 3.2. Since (b) is implied by (a), so we only have to show (a). By changing or variable,

$$\frac{hG(h)}{F^2(h)} = \frac{h\int(W'(t))^2g(th)dt}{\left(\int W''(t)g(th)dt\right)^2} = \frac{hG(h)}{F^2(h)} = \frac{\int(W'(t/h))^2(h/t)^2t^2g(t)dt}{\left(\int W''(t/h)g(t)dt\right)^2}.$$

From assumptions (w), (g1), let $h\to\infty$ on the right end of the above expression, one have

$$\frac{hG(h)}{F^2(h)} \to \frac{\int(W''(0))^2t^2g(t)dt}{\left(\int W''(0)g(t)dt\right)^2} = \int t^2g(t)dt = \sigma^2$$

which yield result (a). □

Proof of Theorem 3.3. The proof is similar to that of Theorem 4.3 in YLL, so we only states the differences for the sake of brevity. Using YLL's notation, it is then suffice to show that $nE|\xi_1|^3\to 0$, where

$$\xi_1 = -\frac{1}{\sqrt{nh_1}h_2^2}K\left(\frac{X_i - x_0}{h_1}\right)W'\left(\frac{Y_i - \theta^TZ_i}{h_2}\right)d^TZ_i,$$

d is an arbitrary $p+1$-dimension real vector. Easy to see $nE|\xi_1|^3$ is bounded above by

$$O_p(nn^{-3/2}h_1^{-3/2})\sum_{j=0}^{p} E\left|K^3\left(\frac{X_i-x_0}{h_1}\right)W'^3\left(\frac{Y_i-\theta^T Z_i}{h_2}\right)\left(\frac{X_i-x_0}{h_1}\right)^{3j}\right|,$$

by the boundedness of W', the summand in the above expression is of the order $O(h_1)$, hence $nE|\xi_1|^3 = O(1/\sqrt{nh_1})$ which tends to 0 by assumption (h1). □

References

[1] Alimoradi, S. and Saleh, A. K. Md. E. (1999). On some L-estimation in auto-regression models. Jour. Statist. Research. **32**, 85-111.

[2] Chatterjee, S. and Mächler, M. (1995). Robust regression: A weighted least squares approach. Communications in Statistics, Theory and Methods, **26**, 1381-1394.

[3] Chen, G. J., and Saleh, A. K. Md. E. (2000). Strong consistency and asymptotic normality of regression quantiles in linear models. Jour. Anhui University, Natural Sciences Edition. **24**, 1-17.

[4] Cleveland, W. S. (1979). Robust locally weighted regression and smoothing scatterplots. JASA, **74**, 829-836.

[5] Fan, J. Q., Gijbels, I. (1997). *Local Polynomial Modelling and Its Applications. Chapman & Hall/CRC*.

[6] Fan, J. Q., Hu, T., and Truong, Y. (1994). robust non-parametric function estimation. Scand. J. Statist. **21**, 433-446.

[7] Fan, J. Q. and Jiang, J. C. (2000). Variable bandwidth and one-step local M-estimator. Science in China (Series A), **43**(1), 65-81.

[8] Härdle, W. (1992). *Applied nonparametric regression*. Cambridge University Press: New York.

[9] He, X. M. and Liang, H. (2000). Quantile regression estimates for a class of linear and partially linear errors-in-variables models. Statistica Sinica, **10**(1), 129-140.

[10] Huber, P. (1974). Robust Statistics. Wiley, New York.

[11] Koenker, R. (2005). *Quantile Regression*. Cambridge University Press.

[12] Koul, H. L. and Saleh, A. K. Md. E. (1993). R-estimation of the parameters of autoregressive AR(p) models. Ann. of Statist. **21**, 534-551.

[13] Koul, H. L. and Saleh, A. K. Md. E. (1995). Autoregression quantiles and related rank-score process. Ann. of Statist. **23**(2), 670-689.

[14] Li, J., Ray, S., and Lindsay, B. G. (2007). A nonparametric statistical approach to clustering via mode identification. Journal of Machine Learning Research, **8**(8), 1687-1723.

[15] Rousseeuw, P. J. (1984). Least median squares regression. JASA, **79**, 871-880.

[16] Rousseeuw, P. J. and Croux, C. (1993). Alternatives to median absolute deviation. JASA, **88**, 1279-1281.

[17] Saleh, A. K. Md. E., Shiraishi, T. (1993). Robust estimation for the parameter of multiple-design multivariate linear model under general restriction. Journal of Nonparametric Statistics. **2**, 295-305.

[18] Saleh, A. K. Md. E.(2006). *Theory of Preliminary Test and Stein-Type Estimation with Applications.* Wiley & Sons Inc., New York.

[19] Tsybakov, A. B. (1986). Robust reconstruction of functions by the local-approximation method. Problems of Information Transmission. **22**, 133-146.

[20] Wilcox, R. (1997). *Introduction to the robust estimation and hypothesis testing*. Academic Press: San Diego.

[21] Yao, W. X., Lindsay, B. and Li, R. Z. (2008). Adaptive robust local polynomial regression. *Submitted*.