Smoothing Parameter Selection For A New Regression Estimator For Non-negative Data

Baohua He

A Thesis

in

The Department

of

Mathematics and Statistics

Presented in Partial Fulfillment of the Requirements

for the Degree of Master of Science at

Concordia University

Montreal, Quebec, Canada

August 2009

Your file Votre référence
ISBN: 978-0-494-63095-2
Our file Notre référence
ISBN: 978-0-494-63095-2

Abstract

Smoothing Parameter Selection For A New Regression Estimator For Non-negative Data

Baohua He

In this thesis, the CV selection technique is applied into Chaubey, Laib and Sen (2008)'s estimator, which is a new regression estimation for nonnegative random variables. The estimator is based on a generalization of Hille's lemma and a perturbation idea. The first and second order MSE are derived. The ISE criteria for the optimal value of smoothing parameter is discussed and also calculated. The simulation results and the Graphical illustrations on the new estimator, comparing with Fan (1992, 2003)'s local kernel regression estimators are provided.

Acknowledgements

First of all, I would like to express my heartiest thanks to Prof. Xiaowen Zhou, who helped me with enrolling to the graduate program, based on which I commenced my life in Canada. What I regret is missing the chance to be his Master student aiming at the PhD orientation.

This work would not have been possible without the support of my advisor, Dr. Arusharka Sen, under whose insightful and professional supervision I chose this topic, and fulfilled this research. What's more, he generously dedicated 2^{nd} order MSE in my Thesis.

I particularly thank Prof. Chaubey, who is the chair of my examining committee, and also the chair of Department of Mathematics and Statistics. Am I lucky to have this authoritative statistician in my unforgettable study in Concordia; I still remember in multivariate class he described the basketball team in Chapel Hill.

I am indebted to Dr. D.Sen, the committee member during the defense. These 3 years, every time I consulted him, his response was always prompt and instructive. My final paper owes much to his thoughtful and helpful suggestions.

I have also benefited from many (maybe the most among students) discussions with Prof. Russell Steele at McGill, although he marked me B in his "Regression and ANOVA" class.

My keen appreciation goes to Dr. Murari Singh with his precious comments during the TA of him, and the oral defense. How can I forget our prayer in the coffee time?

I cannot end without thanking my family, Father Guangming He, Mother Xiuying Zhu, Sister Weijuan He and 5 uncles, little Nephew Ruijie Zheng and my infant Daughter Yijia He, on whose constant encouragement and love I have relied throughout all my time, no matter in China or overseas.

Finally, Special Gratitude is given to my girl friend Hua Yang , who walked with me when I was going through the darkest valley. The night is nearly over; the day is almost here and we are striving for our 50 annual covenant. So we fix our eyes not on what is seen, but on what is unseen. For what is seen is temporary, but what is unseen is eternal.

Dedication

Doxology

Oh, the depth of the riches of the wisdom and knowledge of God!
How unsearchable his judgments,
and unfathomable His ways!

"Who has known the mind of the Lord?
Or who has been his counselor?"

"Who has ever given to God,
that God should repay him?"
For from him and through him and to him are all things.
To him be the glory forever! Amen.

ὦβάθος πλούτου καὶ σοφίας καὶ γνώσεως θεοῦ:

ὡς ἀνεξεραύνητα τὰ κρίματα αὐτοῦ καὶ ἀνεξιχνίαστοι αἱ ὁδοὶ αὐτοῦ.

Τίς γὰρ ἔγνω νοῦν κυρίου;

ἢ τίς σύμβουλος αὐτοῦ ἐγένετο;

ἢ τίς προέδωκεν αὐτῷ,

καὶ ἀνταποδοθήσεται αὐτῷ;

ὅτι ἐξ αὐτοῦ καὶ δι' αὐτοῦ καὶ εἰς αὐτὸν τὰ πάντα: αὐτῷ ἡ δόξα εἰς τοὺς αἰῶνας: ἀμήν.

Table of Contents

1. Introduction

1.1 Parametric and Nonparametric Regression Approaches

Regression analysis is one of the most commonly used techniques in the statistics, which describes the relationship between dependent variable Y and explanatory variable(s) X. One might estimate the regression function m(.) in the model :

$$\Phi(Y) = m(X) + \varepsilon \qquad (1.1.1)$$

Where the ε is the residuals or errors.

In this thesis, we take:

$$\Phi(Y) = Y \qquad (1.1.2)$$

There are two approaches for the regression estimator: the Parametric Approach and Nonparametric one.

The task of Parametric approach is to determine the parameters and it has strict assumptions, for example: m(.) belongs to a specific parametric family with a set of all possible parameter values.

In the past century the parametric regression techniques have improved greatly and maturely. We may refer to Raymond.H.Myers (1990) and the references therein. Using the parametric regression techniques:

We may preliminarily evaluate the validity.

Then select the model (Cp, PRESS, R^2, stepwise…)

And check diagnostics (QQ, R-student, DFFITS, COVRATIO…)

And transform if necessary (Box-Cox, GLM…)

Based on a random sample, for another way, Nonparametric approach is quite simple to compute. It doesn't restrict the possible form of m(.) or is only with few assumptions about m(.).

The task of nonparametric model is to estimate the full regression curve.

1.2 The Kernel Estimators and Fan's Local Linear Estimator

In contrast to the Parametric approach, Nonparametric one has smooth and flexible form. Decades various nonparametric estimators of regression function m(.) have been proposed in the literature, among which the kernel regression estimators are widely used. We may refer to Chaubey, Laib and Sen (2008) and the references therein. We start from Nadaraya-Watson (1964)'s estimator, referring to Hardle,W.(1990) pp.127 :

$$\hat{m}(x) = \frac{n^{-1}\sum_{i=1}^{n} Y_i \; K\left(\frac{x-X_i}{h}\right)}{n^{-1}\sum_{i=1}^{n} K\left(\frac{x-X_i}{h}\right)} \quad (h \to 0 \text{ and } nh \to \infty.) \qquad (1.2.1)$$

Firstly, for numerator: $E\left[n^{-1}\sum_{i=1}^{n} Y_i \; K\left(\frac{x-X_i}{h}\right)\right] \to m(x)f(x)$ (1.2.2)

Similarly, for denominator: $E\left[n^{-1}\sum_{i=1}^{n} K\left(\frac{x-X_i}{h}\right)\right] \to f(x)$ (1.2.3)

Hence, $\hat{m}(x \to m(x)$ (1.2.4)

In Nadaraya-Watson's estimation, the estimator depends on a single smoothing parameter h and it may have boundary bias and fail to estimate the discontinuity at boundary.

Fan (1992, 2003) proposed a local regression, with different smoothing parameter h_n:
By minimizing

$$\sum_{i=1}^{n}\left(Y_i - a - b(x-X_i)\right)^2 K\left(\frac{x-X_i}{h_n}\right) \qquad (1.2.5)$$

The regression estimator can be obtained below:

$$\hat{m}(x) = \hat{a} = \frac{\sum_{i=1}^{n} Y_i \, w_i}{\sum_{i=1}^{n} w_i} \qquad (1.2.6)$$

Where

$$w_i = K\left(\frac{x - X_i}{h_n}\right)\left(S_{n,2} - (x - X_i)\, S_{n,1}\right) \tag{1.2.7}$$

With

K--- the kernel function, symmetric with zero mean and unit variance

h_n--- smoothing bandwidth.

And

$$S_{n,m} = \sum_{i=1}^{n} K\left(\frac{x - X_i}{h_n}\right) (x - X_i)^m \quad (m = 0,1,2) \tag{1.2.8}$$

In Fan's estimation, "Local" is so called, because:

when X_i is near x, $K(u)$ becomes large, $\hat{m}(x)$ can be more affected; and when X_i is far from x, $K(u)$ becomes small, $\hat{m}(x)$ can be less affected.

Here standard normal Kernel function is applied into the simulation:

$$K(u) = \frac{1}{\sqrt{2\pi}} \exp\left(\frac{-u^2}{2}\right) \tag{1.2.9}$$

The boundary problem is of great importance, but the Kernel Estimator may not provide admissible values of the regression function or its functionals at the boundaries, for regressions with restricted support. Any smoothing method will become less accurate near the boundary of the observation interval because fewer observations can be averaged, and thus variance or bias can be affected.

To alleviate this problem, now we propose Chaubey, Laib and Sen (2008)'s estimator, a new regression estimation for nonnegative random variables:

1.3 Chaubey, Laib and Sen's Estimator: A New Regression Estimator

1.3.1 Definition

Now, we motivate the introduction of the following estimator of $m(\cdot)$, that is

$$m_n(x) = \frac{n^{-1}\sum_{i=1}^n \phi(Y_i)Q_{x,v_n}(X_i)}{n^{-1}\sum_{i=1}^n Q_{x,v_n}(X_i)}, \tag{1.3.1.1}$$

where $Q_{x,v_n}(t) = \frac{1}{x}q_{v_n}(\frac{t}{x})$ is a density function on $[0,\infty)$ with mean x and variance $(xv_n)^2 \to 0$ as $n \to \infty$.

The above estimator, however, may not be defined at $x = 0$, except in cases where $m_n(0) = \lim_{x\to 0^+} m_n(x)$ exists. For instance, if $Q_{v_n,x}(\cdot)$ is a gamma density function with mean x and variance $(xv)_n^2$, defined for $x > 0$, by

$$Q_{x,v_n}(t) = \frac{1}{\beta_x^{\alpha_n}\ \Gamma(\alpha_n)} t^{\alpha_n-1}\ e^{-\alpha_n t/x}, \text{ where } \alpha_n = 1/v_n^2,\ \beta_x = v_n^2 x. \tag{1.3.1.2}$$

Then, the limit $m_n(0)$ may be computed as follows

$$\begin{aligned} m_n(0) &= \lim_{x\to 0^+} \frac{\sum_{i=1}^n \phi(Y_{[i]})X_{(i)}^{\alpha_n-1}\ e^{-\alpha_n X_{(i)}/x}}{\sum_{i=1}^n X_{(i)}^{\alpha_n-1}\ e^{-\alpha_n X_{(i)}/x}} \\ &= \lim_{x\to 0^+} \frac{\sum_{i=1}^n \phi(Y_{[i]})X_{(i)}^{\alpha_n-1}\ e^{-\alpha_n[X_{(i)}-X_{(1)}]/x}}{\sum_{i=1}^n X_{(i)}^{\alpha_n-1}e^{-\alpha_n[X_{(i)}-X_{(1)}]/x}} \\ &= \lim_{x\to 0^+} \frac{\phi(Y_{[1]})X_{(1)}^{\alpha_n-1} + \sum_{i=2}^n \phi(Y_{[i]})X_{(i)}^{\alpha_n-1}\ e^{-\alpha_n[X_{(i)}-X_{(1)}]/x}}{X_{(1)}^{\alpha_n-1} + \sum_{i=2}^n X_{(i)}^{\alpha_n-1}\ e^{-\alpha_n[X_{(i)}-X_{(1)}]/x}} \\ &= \phi(Y_{[1]}), \end{aligned}$$

where $X_{(i)}$ stands for the order statistic of X_i and $Y_{[i]}$ the corresponding concommitant, i.e., $Y_{[i]} = Y_j$ if $X_{(i)} = X_j$. However, in this case $m_n(0)$ does not consistently estimate $m(0)$.

To alleviate this situation, consider the following perturbed version of the above regression estimator:

$$\tilde{m}_n(x) := m_n(x+\epsilon_n) = \frac{n^{-1}\sum_{i=1}^n \phi(Y_i)Q_{x+\epsilon_n,v_n}(X_i)}{n^{-1}\sum_{i=1}^n Q_{x+\epsilon_n,v_n}(X_i)}, \quad x \geq 0, \tag{1.3.1.3}$$

where $Q_{x+\epsilon_n,v_n}(t) = \frac{1}{x+\epsilon_n}q_{v_n}(\frac{t}{x+\epsilon_n})$ and ϵ_n goes to 0 at an appropriate (sufficiently slow) rate as $n \to \infty$.

In this thesis, we focus on the special case where $Q_{v_n,x+\epsilon_n}(\cdot)$ is a gamma density function with mean $x+\epsilon_n$ and variance $v_n^2(x+\epsilon_n)^2$. Namely, for $x \geq 0$,

$$Q_{x+\epsilon_n,v_n}(t) = \frac{1}{\beta_{x+\epsilon_n}^{\alpha_n}\ \Gamma(\alpha_n)} t^{\alpha_n-1}\ e^{-\alpha_n t/(x+\epsilon_n)}, \text{ where } \alpha_n = 1/v_n^2,\ \beta_{x+\epsilon_n} = v_n^2(x+\epsilon_n). \tag{1.3.1.4}$$

Gamma density is naturally asymmetric to cope with discontinuity at $t = 0$.

1.3.2 The Properties of the New Estimators

1.3.2.1 Point-wise consistency

To obtain this property,the following generalization of the Hille's Lemmal is used:

Lemma A (Lemma 1, Chapter VII.1, Feller 1965). *Let h be any bounded and continuous function. Let $g_{x,n}(\cdot)$, $n = 1, 2, \ldots$ be a family of densities functions with mean $\mu_n(x)$ and variance $u_n^2(x)$ then we have as $\mu_n(x) \to x$ and $u_n(x) \to 0$*

$$\tilde{h}(x) = \int_{-\infty}^{\infty} h(t) g_{x,n}(t) dt \to h(x) \quad as \quad n \to \infty.$$

The convergence is uniform in every subinterval in which $u_n(x) \to 0$ and h is uniformly continuous

Apply the Hille's Lemma into both the denominator and numerator:

$$m_n(x) \quad = \quad \frac{n^{-1} \sum_{i=1}^{n} \phi(Y_i) Q_{x,v_n}(X_i)}{n^{-1} \sum_{i=1}^{n} Q_{x,v_n}(X_i)}, \tag{1.3.2.1}$$

Firstly, for numerator:

$$\mathsf{E}[\ n^{-1} \sum_{i=1}^{n} \phi(Y_i) g_{x,n}(X_i)\ ]$$

$$= \mathsf{E}[\ \phi(Y) g_{x,n}(X)\ ]$$

$$= \int m(t)\, g_{x,n}(t)\, f(t)\, dt$$

By the Lamma, $\int m(t) f(t)\, g_{x,n}(t) dt \to m(x) \mathrm{f(x)}$ (1.3.2.2)

Similarly, for denominator:

$$\mathsf{E}[\ n^{-1} \sum_{i=1}^{n} g_{x,n}(X_i)\ ]$$

$$= \mathsf{E}[\ g_{x,n}(X)\ ]$$

$$= \int g_{x,n}(t)\, f(t)\, dt$$

$$= \int f(t) g_{x,n}(t) dt$$

By Hille's Lamma, it converges $\mathrm{f(x)}$ (1.3.2.3)

Hence, $m_n(x) \quad = \quad \frac{n^{-1} \sum_{i=1}^{n} \phi(Y_i) g_{x,n}(X_i)}{n^{-1} \sum_{i=1}^{n} g_{x,n}(X_i)} \to \quad m(x)$ (1.3.2.4)

1.3.2.2 Uniform strong consistency

Theorems 1 below deals with the uniform consistency of the estimator $\tilde{m}_n(\cdot)$.

Theorem 1 *Under the suitable conditions,we have, (refering to Chaubey and Sen(2008)):*

$$\sup_{x\in[a,b]} |\tilde{m}_n(x) - m(x)| = 0 \ a.s. \quad as \quad n \to \infty.$$

1.3.2.3 Asymptotic Normality

Theorem 2 below delas with asymptotic normality for $\tilde{m}_n(\cdot)$.

Theorem 2 *Under the suitable conditions,we have, (refering to Chaubey and Sen(2008)):*

$$\sqrt{nv_n}(\tilde{m}_n(x) - m(x)) \xrightarrow{\mathcal{D}} \mathcal{N}\left(0, \quad \sigma^2(x)\right).$$

$$where\ \sigma^2(x) = \frac{1}{2\sqrt{\pi}} \frac{W_2(x) - m^2(x)}{xf(x)},$$

$$W_2\ (\ x\ )=\ E[\ Y^2\ |\ X\ =\ x\ \ ]$$

2 MSE (Mean Square Error) of the New Regression Estimator

2.1 Some Results leading to MSE

We have $m_n^+(x) = r_n^+(x)/f_n^+(x)$, $x \geq 0$, where

$$r_n^+(x) = n^{-1}\sum_{i=1}^{n} Y_i Q_{x+\epsilon_n,v_n}(X_i), \quad f_n^+(x) = n^{-1}\sum_{i=1}^{n} Q_{x+\epsilon_n,v_n}(X_i),$$

and $Q_{x,v}(t) = (1/x)q_v(t/x)$ for $x > 0, t > 0$, where

$$q_v(t) = \frac{t^{(1/v^2)-1}\exp(-t/v^2)}{(v^2)^{1/v^2}\Gamma(1/v^2)}, \; t > 0,$$

is the Gamma $(\alpha = (1/v^2), \beta = v^2)$ density.

Under the suitable conditions,we have, (refering to Chaubey, Sen and Sen(2007)):

$$O(v_n^{-1}) = \frac{I_2(q)}{v_n} = I_2(q) = \lim_{v\to 0} v \int_0^\infty (q_v(t))^2 dt = 1/\sqrt{4\pi} \text{ exists;}$$

$$\int Q := E[\,Q_{x,v_n}(X_i)] = \int Q_{x,v_n}(u)\, f(u,t)\, du = f(x) + f'(x)\cdot \epsilon_n + o(\epsilon_n)$$

$$\int Q\cdot Y := E[\,Q_{x,v_n}(X_i) Y_i] = \int Q_{x,v_n}(u)\, u f(u)\, du = r(x) + r'(x)\cdot \epsilon_n + o(\epsilon_n)$$

$$\int Q^2 := E\{[\,Q_{x,v_n}(X_i)\,]^2\} = \int Q^2_{x,v_n}(u)\, f(u)\, du = \frac{O(v_n^{1})}{x+\epsilon_n}[\,f(x) + f'(x)\cdot \epsilon_n + o(\epsilon_n)]$$

$$\int Q^2\cdot Y := E\{[Q_{x,v_n}(X_i)]^2\} = \int Q^2_{x,v_n}(u)\, u f(u)\, du = \frac{O(v_n^{-1})}{x+\epsilon_n}[\,r(x) + r'(x)\cdot \epsilon_n + o(\epsilon_n)]$$

$$\int Q^2\cdot Y^2 := E\{[Q_{x,v_n}(X_i) Y_i]^2\} = \int Q^2_{x,v_n}(u)\, u^2 f(u)\, du$$

$$= \frac{f(x)\,S^2(x) + [f(x)\,S^2(x)]'\cdot \epsilon_n + o(\epsilon_n)}{x+\epsilon_n}\cdot O(v_n^{-1})$$

where $S^2(u) = E(Y_1^2 | X_1 = t)$

2.2 Estimation for $f_n^+(x)$

2.2.1 Expectation for $f_n^+(x)$

$$E\ [f_n^+(x)] = E[\ n^{-1}\sum_{i=1}^{n} Q_{x,v_n}(X_i)\ ]$$

$$= E[\ Q(X)\ ]$$

$$= \int Q_{x,v_n}(u)\ f(u)\,du$$

$$= \int \frac{q_{x,v_n}[u/(x+\epsilon_n)]}{x+\epsilon_n} \cdot f(u)\,du \qquad (\text{let } y= u/(x+\epsilon_n)\ )$$

$$= \int \frac{q_{x,v_n}[u/(x+\epsilon_n)]}{x+\epsilon_n} \cdot f[y\cdot(x+\epsilon_n]\ d[y\cdot(x+\epsilon_n]$$

$$= \int q_{x,v_n}(y)\cdot f[y\cdot(x+\epsilon_n]\ dy$$

(expand in x by Taylor's expansion, and let $x_o = u-x$)

$$= \int q_{x,v_n}(y)\,[f(x)+f'(x)\cdot x_o + o(x_o)]\ dy$$

$$= f(x)\int q + f'(x)\int q\cdot x_o + \int q\cdot o(x_o)$$

$$= f(x)+f'(x)\cdot \epsilon_n + o(\epsilon_n)$$

2.2.2 Variance for $f_n^+(x)$

$$Var\ [f_n^+(x)] = Var[\ n^{-1} \sum_{i=1}^n Q_{x,v_n}(X_i)\ ]$$

$$= n^{-2}\, Var[\ \sum_{i=1}^n Q_{x,v_n}(X\ )\ ]$$

$$= n^{-1}\, Var[\ Q\,(X\ )\ ]$$

$$= n^{-1}\, \{E\,[\ Q\,(X\ )\ ]^2\ -\ E^2[\ Q(X\ )\ ]\ \}$$

$$= n^{-1}\ [\ \int Q^2\ -\ (\int Q\ )^2\ ]$$

$$= n^{-1}\ \{\ \frac{O\,(v_n^{-1})}{x + \epsilon_n}\,[\,f(x) + f'(x)\cdot \epsilon_n + o(\epsilon_n)\,]$$

$$-\ [\,f(x) + f'(x)\cdot \epsilon_n + o(\epsilon_n)]^2\ \}$$

2.3 Estimation for $r_n^+(x)$

2.3.1 Expectation for $r_n^+(x)$

$$E\,[r_n^+(x)] = E[\,n^{-1}\sum_{i=1}^{n} Q_{x,v_n}(X_i)\cdot Y_i\,]$$

$$= E[\,Q(X)\cdot Y\,]$$

$$= \int Q\,Y$$

$$= \int Q_{x,v_n}(u)\;r\,(u)\,du$$

$$= \int \frac{q_{x,v_n}[u/(x+\epsilon_n)]}{x+\epsilon_n}\cdot\;r\,(u)\,du \qquad (\text{let } y = u/(x+\epsilon_n)\ )$$

$$= \int \frac{q_{x,v_n}[u/(x+\epsilon_n)]}{x+\epsilon_n}\cdot\;r[y\cdot(x+\epsilon_n]\;\;d[y\cdot(x+\epsilon_n]$$

$$= \int q_{x,v_n}(y)\cdot\;r[y\cdot(x+\epsilon_n]\;\;dy$$

(expand in x by Taylor's expansion,
and let $x_o = u - x$)

$$= \int q_{x,v_n}(y)\,[r\,(x)+r'(x)\cdot x_o + o(x_o\,)]\;\;dy$$

$$= r\,(x)\int q + r'(x)\int q\cdot\,x_o + \int q\cdot\;o(\,x_o\,)$$

$$= r\,(x) + r'(x)\cdot\epsilon_n + o(\,\epsilon_n)$$

2.3.2 Variance for $r_n^+(x)$

$$Var\,[r_n^+(x)] = Var[\; n^{-1} \sum_{i=1}^n Q_{x,v_n}(X_i) \cdot Y_i \;]$$

$$= n^{-2}\, Var[\; \sum_{i=1}^n Q_{x,v_n}(X_i) \cdot Y_i \;]$$

$$= n^{-1}\, Var[\; Q(X) \cdot Y \;]$$

$$= n^{-1} \{ E[\; Q(X) \cdot Y \;]^2 - E^2[\; Q(X) \cdot Y \;] \}$$

$$= n^{-1} \; [\; \int Q^2 \cdot Y^2 - (\int Q \cdot Y)^2 \;]$$

$$= n^{-1} \{ \; \frac{f(x)\, S^2(x) + [f(x)\, S^2(x)]' \cdot \epsilon_n + o(\epsilon_n)}{(x + \epsilon_n) \cdot v_n} \cdot I_2(q) -$$

$$[\; r(x) + r'(x) \cdot \epsilon_n + o(\epsilon_n) \;]^2 \quad \}$$

$$\textit{where} \quad S^2(u) = \int t^2 f(t \mid u)\, dt$$

2.4 1st order Mean Squares Error (MSE) of the regression estimator

(" 1st order " is so called, because by Taylor's expansion, the final result is only in first order, e.g. $f'(x)$.)

$$
\begin{aligned}
MSE[m_n^+(x)] &= E[(m_n^+(x) - m(x))]^2 \\
&= E[(r_n^+(x)/f_n^+(x) - m(x)]^2 \\
&= E\{[(r_n^+(x)/f_n^+(x) - m(x)][f_n^+(x)/f(x) + 1 - f_n^+(x)/f(x)]\}^2 \\
&= E\{[(r_n^+(x) - m(x)\cdot f_n^+(x)]/f_n^+(x)\cdot[f_n^+(x)/f(x) + 1 - f_n^+(x)/f(x)]\}^2 \\
&= E\{[(r_n^+(x) - m(x)\cdot f_n^+(x)]/f_n^+(x)\cdot[f_n^+(x)/f(x)] \\
&\quad \underbrace{+[(r_n^+(x) - m(x)\cdot f_n^+(x)]/f_n^+(x)\cdot[1 - f_n^+(x)/f(x)]\}^2}_{=0} \\
&\approx E\{[(r_n^+(x) - m(x)\cdot f_n^+(x)]/f(x)\}^2 \\
&= f(x)^{-2}\cdot E[r_n^+(x) - m(x)\cdot f_n^+(x)]^2 \\
&= f^{-2}(x)\cdot\{Var[r_n^+(x) - m(x)\cdot f_n^+(x)] + E^2[r_n^+(x) - m(x)\cdot f_n^+(x)]\} \\
&= f^{-2}(x)\cdot\{Var[r_n^+(x) - m(x)\cdot f_n^+(x)]\} + f^{-2}(x)\cdot E^2[r_n^+(x) - m(x)\cdot f_n^+(x)] \\
&= f^{-2}(x)\cdot\{Var[r_n^+(x) - m(x)\cdot f_n^+(x)]\} + f^{-2}(x)\cdot E^2[r_n^+(x) - m(x)\cdot f_n^+(x)] \\
&= f^{-2}(x)\cdot\{Var[r_n^+(x) - m(x)\cdot f_n^+(x)]\} + f^{-2}(x)\cdot\{E[r_n^+(x)] - m(x)\cdot E[f_n^+(x)]\}^2 \\
&= f^{-2}(x)\cdot\{Var[n^{-1}\textstyle\sum_{i=1}^n Y_i\cdot Q_{x,v_n}(X_i) - m(x)\cdot n^{-1}\sum_{i=1}^n Q_{x,v_n}(X_i)]\} \\
&\quad + f^{-2}(x)\cdot\{[r(x)+r'(x)\cdot\epsilon_n + o(\epsilon_n)] - m(x)\cdot[f(x)+f'(x)\cdot\epsilon_n + o(\epsilon_n)\}^2 \\
&= f^{-2}(x)\cdot n^{-2}\cdot Var\{\textstyle\sum_{i=1}^n Q_{x,v_n}(X_i)\cdot[Y_i - m(x)]\} \\
&\quad + f^{-2}(x)\cdot\{[r(x)+r'(x)\cdot\epsilon_n + o(\epsilon_n)] - [m(x)f(x) + m(x)f'(x)\cdot\epsilon_n + m(x)o(\epsilon_n)]\}^2 \\
&= f^{-2}(x)\cdot n^{-2}\cdot Var\{\textstyle\sum_{i=1}^n Q_{x,v_n}(X_i)\cdot[Y_i - m(x)]\} \\
&\quad + f^{-2}(x)\cdot\{m(x)f(x)]'\cdot\epsilon_n + o(\epsilon_n) - m(x)f'(x)\cdot\epsilon_n - m(x)o(\epsilon_n)\}^2 \\
&= f^{-2}(x)\cdot n^{-2}\cdot n\,Var\{Q_{x,v_n}(X)\cdot[Y - m(x)]\} \\
&\quad + f^{-2}(x)\cdot\{[m'(x)f(x)\cdot\epsilon_n + o(\epsilon_n) - m(x)o(\epsilon_n)\}^2 \\
&= f^{-2}(x)\cdot n^{-1}\cdot\{E[Q_{x,v_n}(X)\cdot(Y - m(x))]^2 - E^2[Q_{x,v_n}(X)\cdot(Y - m(x))]\} \\
&\quad + f^{-2}(x)\cdot\{[m'(x)f(x)\cdot\epsilon_n + o(\epsilon_n) - m(x)o(\epsilon_n)\}^2
\end{aligned}
$$

$MSE[m_n^1(x)]$(continue)

$$= f^{-2}(x) \cdot n^{-1}\{ [\iint Q^2 \cdot Y^2 - 2m(x)) \iint Q^2 \cdot Y + m^2(x)) \int Q^2] - [\iint Q \cdot Y - m(x)) \iint Q]^2 \}$$

$$+ f^{-2}(x) \cdot \{ m'(x) f(x) \cdot \epsilon_n + o(\epsilon_n) - m(x) o(\epsilon_n) \}^2$$

$$= f^{-2}(x) \cdot n^{-1} \{ \frac{f(x) S^2(x) + [f(x) S^2(x)]' \cdot \epsilon_n + o(\epsilon_n)}{(x+\epsilon_n) \cdot v_n}$$

$$- 2m(x) \cdot \frac{O(v_n^{-1})[r(x) + r'(x) \cdot \epsilon_n + o(\epsilon_n)]}{x+\epsilon_n}$$

$$+ m^2(x) \cdot \frac{O(v_n^{-1})}{x+\epsilon_n} \}$$

$$- f^{-2}(x) \cdot n^{-1} \{ r(x) + r'(x) \cdot \epsilon_n + o(\epsilon_n) - m(x) \cdot [f(x) + f'(x) \cdot \epsilon_n + o(\epsilon_n)]^2 \}$$

$$+ f^{-2}(x) \cdot \{ m'(x) f(x) \cdot \epsilon_n + o(\epsilon_n) - m(x) o(\epsilon_n) \}^2$$

$$= \frac{I_2(q)}{f^2(x) n (x+\epsilon_n) v_n} \{ [f(x) S^2(x) + [f(x) S^2(x)]' \cdot \epsilon_n + o(\epsilon_n)]$$
$$-2m(x) \cdot [m(x) f(x) + (m(x) f(x))' \cdot \epsilon_n + o(\epsilon_n)]$$
$$+ m^2(x) \}$$

$$- f^{-2}(x) \quad n^{-1} \cdot \{ m'(x) f(x) \cdot \epsilon_n + o(\epsilon_n) - m(x) o(\epsilon_n) \}$$

$$+ f^{-2}(x) \cdot \{ m'(x) f(x) \cdot \epsilon_n + o(\epsilon_n) - m(x) o(\epsilon_n) \}$$

2.5 2nd order Mean Squares Error (MSE) of the regression estimator

("**2nd order**" is so called, because by Taylor's expansion,
the final result is kept in first order, e.g. $f''(x)$.)

We have $m_n^+(x) = r_n^+(x)/f_n^+(x)$, $x \geq 0$, where

$$r_n^+(x) = n^{-1}\sum_{i=1}^{n} Y_i Q_{x+\varepsilon_n,v_n}(X_i), \quad f_n^+(x) = n^{-1}\sum_{i=1}^{n} Q_{x+\varepsilon_n,v_n}(X_i),$$

and $Q_{x,v}(t) = (1/x)q_v(t/x)$ for $x > 0, t > 0$, where

$$q_v(t) = \frac{t^{(1/v^2)-1}\exp(-t/v^2)}{(v^2)^{1/v^2}\Gamma(1/v^2)}, \ t > 0,$$

is the Gamma $(\alpha = (1/v^2), \beta = v^2)$ density.

Below we determine the optimal (in the sense of minimizing the mean squared error (MSE) of $m_n^+(x)$) rates of convergence of $v_n \to 0, \varepsilon_n \to 0$ (which is necessary for $m_n^+(x)$ to be a consistent estimator of $m(x)$) as $n \to \infty$.

Consider

$$\begin{aligned} & m_n^+(x) - m(x) \\ = & (r_n^+(x)/f_n^+(x)) - m(x) \\ = & (1/f_n^+(x))(r_n^+(x) - m(x)f(x)) - (m(x)/f_n^+(x))(f_n^+(x) - f(x)). \end{aligned} \qquad (3.5.1)$$

Thus we may approximate

$$\text{MSE}[m_n^+(x)] := E(m_n^+(x) - m(x))^2 \approx \frac{1}{(f(x))^2}E(r_n^+(x) - m(x)f(x))^2 + \frac{m(x)}{(f(x))^2}E(f_n^+(x) - f(x))^2,$$

ignoring the product term $-2(m(x)/f^2(x))E[(r_n^+(x) - m(x)f(x))(f_n^+(x) - f(x))]$, because by the Cauchy-Schwartz inequality

$$\begin{aligned} |E[(r_n^+(x) - m(x)f(x))(f_n^+(x) - f(x))]| & \leq \sqrt{E(r_n^+(x) - m(x)f(x))^2}\sqrt{E(f_n^+(x) - f(x))^2} \\ & \leq \max\{E[(r_n^+(x) - m(x)f(x))]^2, E[(f_n^+(x) - f(x))]^2\}. \end{aligned}$$

Now

$$E(r_n^+(x) - m(x)f(x))^2 = \text{var}(r_n^+(x)) + (\text{bias}(r_n^+(x))^2,$$

and

$$\begin{aligned}
\mathrm{var}(r_n^+(x)) &= n^{-1}E(Y_1^2Q^2_{x+\varepsilon_n,v_n}(X_1)) - n^{-1}[E(Y_1Q_{x+\varepsilon_n,v_n}(X_1))]^2 \\
&\approx n^{-1}E(Y_1^2Q^2_{x+\varepsilon_n,v_n}(X_1)) \\
&= n^{-1}(x+\varepsilon_n)^{-2}\int_0^\infty m_2(t)q^2_{v_n}(t/(x+\varepsilon_n))f(t)dt, \text{ where } m_2(t) = E(Y_1^2|X_1=t), \\
&= n^{-1}(x+\varepsilon_n)^{-1}\int_0^\infty m_2(t(x+\varepsilon_n))f(t(x+\varepsilon_n))q^2_{v_n}(t)dt \\
&\approx \begin{cases} (nv_nx)^{-1}m_2(x)f(x) & \text{if } x>0, \\ (nv_n\varepsilon_n)^{-1}m_2(0)f(0) & \text{if } x=0, \end{cases} \qquad (3.5.2)
\end{aligned}$$

since

$$\begin{aligned}
q^2_{v_n}(t) &= \frac{(v_n^2/2)^{(2/v_n^2)-1}\Gamma((2/v_n^2)-1)}{(v_n^2)^{(2/v_n^2)}\Gamma^2(1/v_n^2)} \frac{t^{(2/v_n^2)-2}\exp(-2t/v^2)}{(v_n^2/2)^{(2/v_n^2)-1}\Gamma((2/v_n^2)-1)} \\
&= O(v_n^{-1})(\text{Gamma } (\alpha=(2/v_n^2)-1, \beta=v_n^2/2 \text{ density}),
\end{aligned}$$

so that, provided $m_2(\cdot)$, $f(\cdot)$ are assumed to be continuous,

$\int_0^\infty m_2(t(x+\varepsilon_n))f(t(x+\varepsilon_n))\text{Gamma } (t|\alpha=(2/v_n^2)-1, \beta=v_n^2/2)dt \to m_2(x)f(x)$ as $n\to\infty$.

Further,

$$\begin{aligned}
\mathrm{Bias}(r_n^+(x)) &= E(Y_1Q_{x+\varepsilon_n,v_n}(X_1)) - m(x)f(x) \\
&= (x+\varepsilon_n)^{-1}\int_0^\infty m(t)q_{v_n}(t/(x+\varepsilon_n))f(t)dt - m(x)f(x) \\
&= \int_0^\infty [r(t(x+\varepsilon_n)) - r(x)]q_{v_n}(t)dt, \text{ where } r(x)=m(x)f(x), \\
&= \int_0^\infty [(x(t-1)+\varepsilon_nt)r'(x) + (1/2)(x(t-1)+\varepsilon_nt)^2r''(x)]q_{v_n}(t)dt + \theta_n(x), \\
&\quad \text{by Taylor's expansion, with 3rd and higher order terms denoted by } \theta_n(x), \\
&= xr'(x)\int_0^\infty (t-1)q_{v_n}(t)dt + \varepsilon_nr'(x)\int_0^\infty tq_{v_n}(t)dt + (1/2)x^2r''(x)\int_0^\infty (t-1)^2q_{v_n}(t)dt \\
&\quad + xr''(x)\varepsilon_n\int_0^\infty t(t-1)q_{v_n}(t)dt + (1/2)\varepsilon_n^2r''(x)\int_0^\infty t^2q_{v_n}(t)dt + \theta_n(x) \\
&= \varepsilon_nr'(x) + (1/2)v_n^2x^2r''(x) + \varepsilon_nv_n^2xr''(x) + (1/2)\varepsilon_n^2(1+v_n^2)r''(x) + \theta_n(x) \\
&\approx \varepsilon_nr'(x) + (1/2)v_n^2x^2r''(x), \qquad (3.5.3)
\end{aligned}$$

using the facts that

$$\int_0^\infty tq_{v_n}(t)dt = 1, \ \int_0^\infty t^2q_{v_n}(t)dt = 1+v_n^2,$$

$$\int_0^\infty (t-1)^2q_{v_n}(t)dt = \int_0^\infty t(t-1)q_{v_n}(t)dt = v_n^2,$$

and ignoring higher order terms.

From Eq. (3.5.2) and (3.5.3) we get

$$\text{MSE}[r_n^+(x)] \approx \begin{cases} (nv_n x)^{-1} m_2(x) f(x) + [\varepsilon_n r'(x) + (1/2) v_n^2 x^2 r''(x)]^2 & \text{if } x > 0, \\ (nv_n \varepsilon_n)^{-1} m_2(0) f(0) + \varepsilon_n^2 (r'(0))^2 & \text{if } x = 0. \end{cases} \quad (3.5.4)$$

Similarily,we get

$$\text{MSE}[f_n^+(x)] \approx \begin{cases} (nv_n x)^{-1} f(x) + [\varepsilon_n f'(x) + [(1/2) v_n^2 x^2 f''(x)]^2 & \text{if } x > 0, \\ (nv_n \varepsilon_n)^{-1} f(0) + \varepsilon_n^2 (f'(0))^2 & \text{if } x = 0. \end{cases} \quad (3.5.5)$$

Finally,from Eq. (3.5.1), (3.5.4) and (3.5.5) we get

$$\text{MSE}[m_n^+(x)] \approx \begin{cases} \frac{1}{(f(x))^2}\{(nv_n x)^{-1} m_2(x) f(x) + [\varepsilon_n r'(x) + (1/2) v_n^2 x^2 r''(x)]^2\} \\ \quad + \frac{m(x)}{(f(x))^2}\{(nv_n x)^{-1} f(x) + [\varepsilon_n f'(x) + (1/2) v_n^2 x^2 f''(x)]^2\} & \text{if } x > 0, \\ \frac{1}{(f(x))^2}\{ (nv_n \varepsilon_n)^{-1} m_2(0) f(0) + \varepsilon_n^2 (r'(0))^2 \} \\ \quad + \frac{m(0)}{(f(x))^2}\{ (nv_n \varepsilon_n)^{-1} f(0) + \varepsilon_n^2 (f'(0))^2 \} & \text{if } x = 0. \end{cases} \quad (3.5.6)$$

Thus when $x > 0$, (3.5.6) shows that the optimal choice of ε_n is $\varepsilon_n = 0$, which gives the optimal choice of v_n to be $v_n = O(n^{-1/5})$ and the optimal order of MSE$[m_n^+(x)]$ is then the usual $O(n^{-4/5})$. Note that setting $\varepsilon_n = O(v_n^2)$ also leads to the same optimum.

On the other hand when $x = 0$, there is no optimal choice for $v_n > 0$ while that for ε_n is $\varepsilon_n = O((nv_n)^{-1/3})$, so that we must have $nv_n \to \infty$ for consistency. Setting $\varepsilon_n = v_n^2$ as above leads to $v_n = O(n^{-1/7})$ and MSE $(x) = O(n^{-4/7})$ which is suboptimal. Note, however, that setting $v_n = O(\varepsilon_n^{-1/2})$ leads to $\varepsilon_n = O(n^{-2/5})$, so that the order of MSE becomes the usual $n^{-4/5}$, but v_n becomes $O(n^{1/5})$ which means in this case $v_n \to \infty$!

3 Simulation and Results

3.1 CV Methods and ISE Criterion

CV(Cross-Validation) approach is a useful method to optimize the parameters v_n and ϵ_n. There are two forms of cross-validation : *Maximum Likelihood CV* and *Least-Squared CV*. Here, we focus our attention on the Least-Squared CV.

Below we describe optimal choice of (v_n, ϵ_n) by minimizing CV.

The cross-validation methods are adapted from Scott (1992) and Wand and Jones (1995).

Here define the ***leave one out estimate*** CV:

$$CV\ (v_n, \epsilon_n) \quad = \quad n^{-1} \sum_{j=1}^{n} [\ Y_j \ - \ m^{+}(X_j)]^2 \tag{3.1.1}$$

where

$$m^{+}(X_j\ ) \quad = \quad \frac{n^{-1} \sum_{i=1}^{n} Y_i\ Q_{x,v_n}(X_i)}{n^{-1} \sum_{i=1}^{n} Q_{x,v_n}(X_i)} \quad (\ j \neq i\ ) \tag{3.1.2}$$

and

$$Y \quad = \quad \mathrm{m}(X) + \epsilon \tag{3.1.3}$$

The v_n and ϵ_n minimizing this function are

$$(\hat{v}, \hat{\epsilon})_{CV} = \arg \min_{v_n, \epsilon_n} CV(v_n, \epsilon_n)$$

Scott and Terrell (1987) call the function an ***Unbiased Cross-Validation*** criterion.

Now we obtain the optimal regressor at $(\hat{v}, \hat{\epsilon})$:

$$m^{+}_{(\hat{v}, \hat{\epsilon})}(x) \quad = \quad \frac{n^{-1} \sum_{i=1}^{n} Y_i Q(\hat{v}, \hat{\epsilon})(X_i)}{n^{-1} \sum_{i=1}^{n} Q_{(\hat{v}, \hat{\epsilon})}(X_i)} \tag{3.1.4}$$

Consider a distance measure between $m^{+}_{n}(x)$ and $m(x)$, the ***Integrated Squared Error (ISE)*** is defined as

$$ISE(v_n, \epsilon_n) \quad = \quad \int_0^{\infty} [m^{\,+}(x) - m(x)]^2 f(x)\, dx \tag{3.1.5}$$

Replace $m^{+}(x)$ with $m^{+}_{(\hat{v}, \hat{\epsilon})}(x)$, thus we obtain:

$$ISE_{\ (\hat{v}, \hat{\epsilon})} \quad = \quad \int_0^{\infty} [\, m^{+}_{(\hat{v}, \hat{\epsilon})}(x) - m(x)]^2 f(x)\, dx \tag{3.1.6}$$

3.2 Simulation steps

Step1: chose the different values of X, ε and underlying functions:

Y=X+ 2*exp(-16*(X^2)) +ε	**X:exp(1) ε:norm(0,0.7^2)**	**X:exp(1) ε:doubleExp**	**X:weibull ε:norm(0,0.7^2)**	**X: weibull ε:doubleExp**
Y=sin(2*X)+ 2*exp(-16*(X^2)) +ε	**X:exp(1) ε:norm(0,0.5^2)**	**X:exp(1) ε:doubleExp**	**X:weibull ε:norm(0,0.5^2)**	**X: weibull ε:doubleExp**

Step2: generate random sample (size n=100,200,500)

Step3: set list of possible value e_n and v_n, (and h_n for Fan's Local Kernel Estimator)

Step4: determine the optimal parameters above by minimizing *UCV*

Step5: Calculating the mean sample ISE with the optimal parameters

Step6: Table and graph Simulation Illustrations

Following are the results and illustrations:

Table 1: Optimal ISE simulation results for Y=X+2*exp(-16*(X^2))+ε

Y=X+ 2*exp(-16*(X^2)) +ε	X:exp(1) ε:norm(0,0.7^2)	X:exp(1) ε:doubleExp	X:weibull ε:norm(0,0.7^2)	X: weibull ε:doubleExp
n=100	0.03584061 (0.03913404)	0.1958467 (0.2006739)	0.02308521 (0.02178451)	0.1190155 (0.1110448)
n=200	0.02258581 (0.02616064)	0.1168053 (0.1304714)	0.01371159 (0.01323952)	0.07219315 (0.07009928)
n=500	0.01231901 (0.01350157)	0.05675914 (0.0675265)	0.007551803 (0.00613061)	0.03658063 (0.03428299)

(the value outside parenthesis *is Chaubey and Sen's Estimator*, and inside one is Fan's Local Kernel Estimator)

Table2: Optimal ISE simulation results for Y=sin(2*X)+2*exp(-16*(X^2)) +ε

Y=sin(2*X)+ 2*exp(-16*(X^2)) +ε	X:exp(1) ε:norm(0,0.7^2)	X:exp(1) ε:doubleExp	X:weibull ε:norm(0,0.7^2)	X: weibull ε:doubleExp
n=100	0.01544821 (0.01495103)	0.199257 (0.1838041)	0.008892057 (0.007312179)	0.1377363 (0.1037686)
n=200	0.009834903 (0.009768867)	0.1283950 (0.1201265)	0.005304467 (0.004416901)	0.08062876 (0.06583907)
n=500	0.006080469 (0.005155077)	0.06577962 (0.05971275)	0.002839715 (0.0020642)	0.04228712 (0.03178110)

(the value outside parenthesis *is Chaubey and Sen's Estimator*, and inside one is Fan's Local Kernel Estimator)

Table 3: Optimal (e,v) simulation results for Y=X+2*exp(-16*(X^2))+ε

Y=X+ 2*exp(-16*(X^2)) +ε	X:exp(1) ε:norm(0,0.7^2)	X:exp(1) ε:doubleExp	X:weibull ε:norm(0,0.7^2)	X: weibull ε:doubleExp
n=100	(0.03099, 0.141225)	(0.06812, 0.207325)	(0.03456, 0.140175)	(0.0973, 0.210075)
n=200	(0.0267, 0.123175)	(0.05586667, 0.1796667)	(0.0225, 0.1215)	(0.07636667, 0.1910833)
n=500	(0.0228, 0.10375)	(0.0386, 0.149)	(0.01365, 0.09775)	(0.0538, 0.161)

Table4: Optimal (e,v) simulation results for Y=sin(2*X)+2*exp(-16*(X^2)) +ε

Y=sin(2*X)+ 2*exp(-16*(X^2)) +ε	X:exp(1) ε:norm(0,0.5^2)	X:exp(1) ε:doubleExp	X:weibull ε:norm(0,0.5^2)	X: weibull ε:doubleExp
n=100	(0.02508, 0.09725)	(0.07448, 0.207325)	(0.02316, 0.0839)	(0.10506, 0.1755)
n=200	(0.0169, 0.0785)	(0.06533333, 0.1816667)	(0.0171, 0.07425)	(0.08826667, 0.155)
n=500	(0.0207, 0.0665)	(0.0419, 0.1345)	(0.0139, 0.068)	(0.0646, 0.145)

Table 5: Optimal h simulation results for Y=X+2*exp(-16*(X^2))+ε

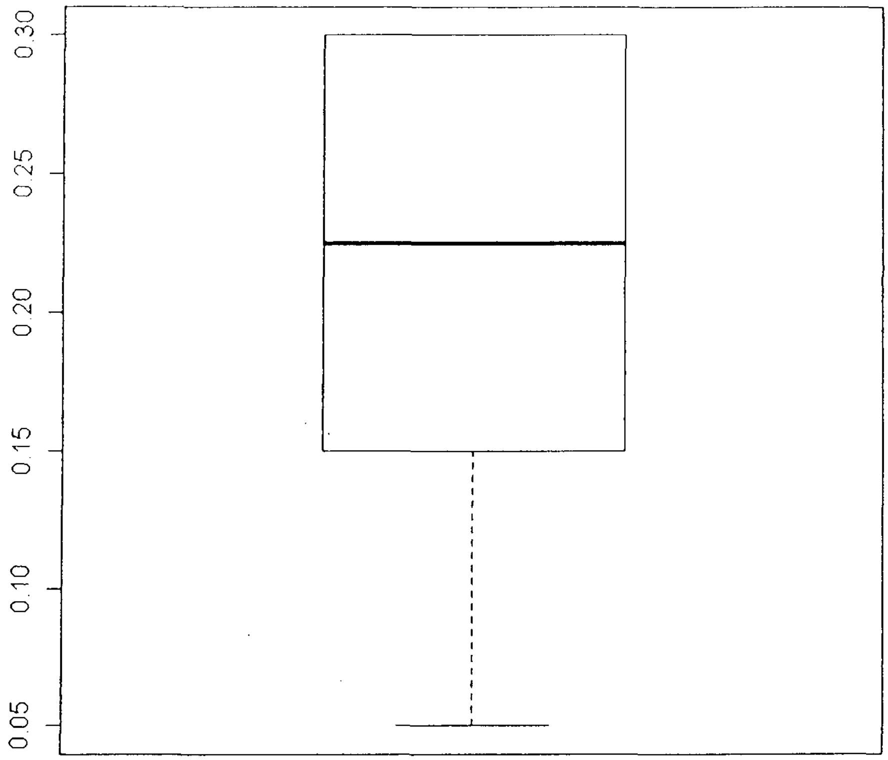
Y=X+ 2*exp(-16*(X^2)) +ε	X:exp(1) ε:norm(0,0.7^2)	X:exp(1) ε:doubleExp	X:weibull ε:norm(0,0.7^2)	X: weibull ε:doubleExp
n=100	0.197125	0.25841	0.191595	0.329015
n=200	0.173285	0.2477417	0.1513833	0.309483
n=500	0.131000	0.2174	0.118050	0.23325

Table6: Optimal h simulation results for Y=sin(2*X)+2*exp(-16*(X^2)) +ε

Y=sin(2*X)+ 2*exp(-16*(X^2)) +ε	X:exp(1) ε:norm(0,0.5^2)	X:exp(1) ε:doubleExp	X:weibull ε:norm(0,0.5^2)	X: weibull ε:doubleExp
n=100	0.161795	0.3192025	0.135535	0.34415
n=200	0.1337667	0.264075	0.1148667	0.2800667
n=500	0.1092	0.222900	0.10876	0.2428

From Table 1 to 6, we can see that

- **As the size n increases , the optimal *(e,v)* and *h* decreases.**
- **the optimal *(e,v)* between Y=X+2*exp(-16*(X^2))+ε and Y=sin(2*X)+2*exp(-16*(X^2)) +ε are close (the same to *h*)**
- **X:weibull and ε:norm obtain the best minimum comparing with others**



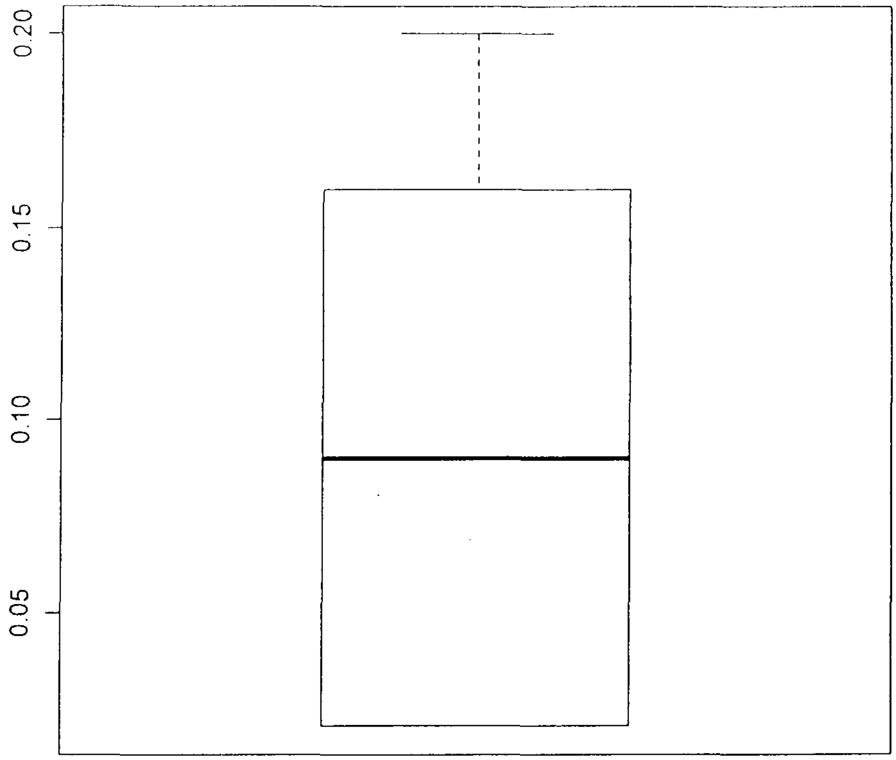
Plot *1:* boxplot of *e* simulation results

Y=X+2*exp(-16*(X^2))+ε

X: weibull

ε:doubleExp

n=100



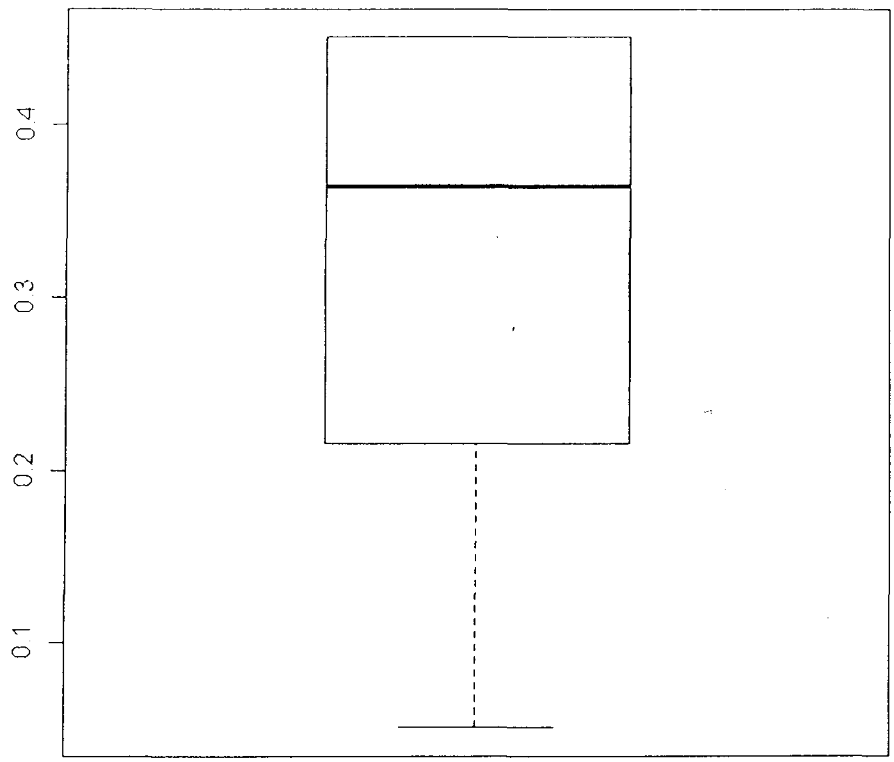
Plot *2:* boxplot of *v* simulation results

Y=X+2*exp(-16*(X^2))+ε

X: weibull

ε:doubleExp

n=100



Plot *3:* boxplot of *h* simulation results

Y=X+2*exp(-16*(X^2))+ε

X: weibull

ε:doubleExp

n=100

3.3 Scatterplot and Regression Estimators for Simulation IID Data

3.3.1 Simulation Summarizations

Comparing the graphics of the *Chaubey , Laib and Sen's Estimator and Fan's Local Estimator*, we have the following conclusions:

- The graphics of the *Fan's Local Estimators* are close to *Chaubey , Laib and Sen's Estimators*
- In small sample size(e.g.n=100), the graphics of the *Fan's Local Estimators* are closer than *Chaubey , Laib and Sen's Estimator.*
- In large sample size(e.g.n=500), the graphics of the *Chaubey ,Laib and Sen's Estimator* are closer than *Fan's Local Estimators.*
- As the sample size increases, the *Chaubey , Laib and Sen's Estimators* are much closer while *Fan's Local Estimators* are little closer.

3.3.2 Simulation Illustrations

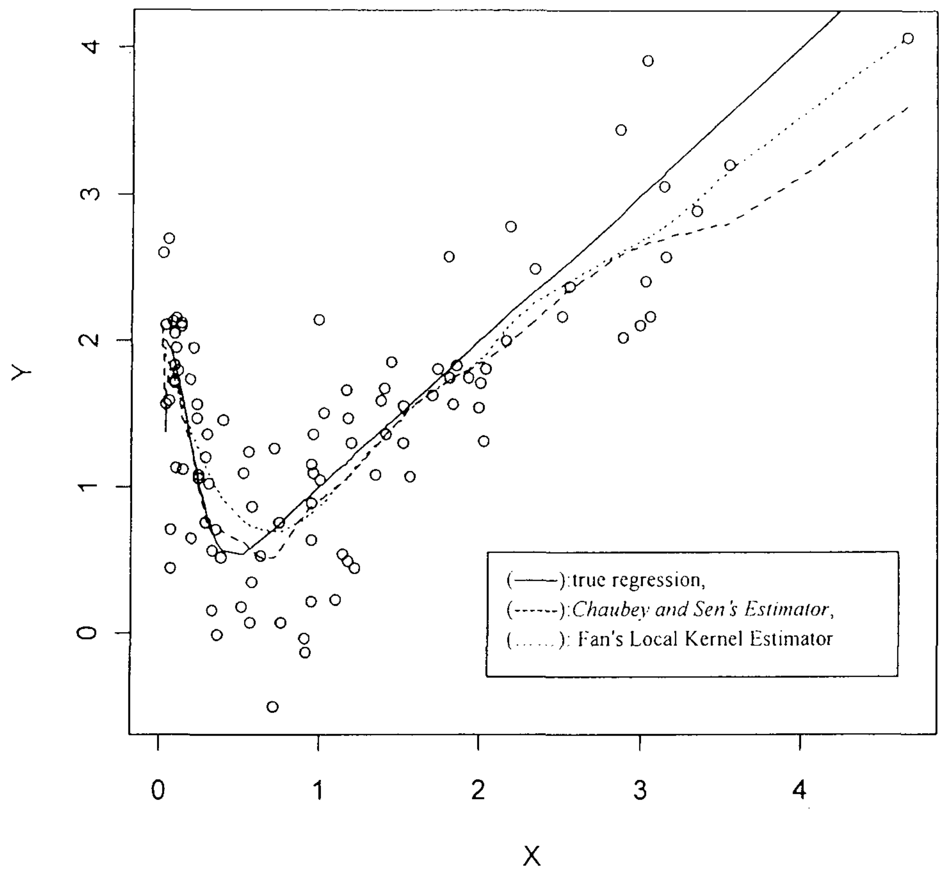


Figure 1(1): scatterplot and regression estimators for simulation IID data with:

Y=2*X+2*exp(-16*(X^2)) +ε

X:exp(1)

ε:norm(0,0.7^2)

n=100

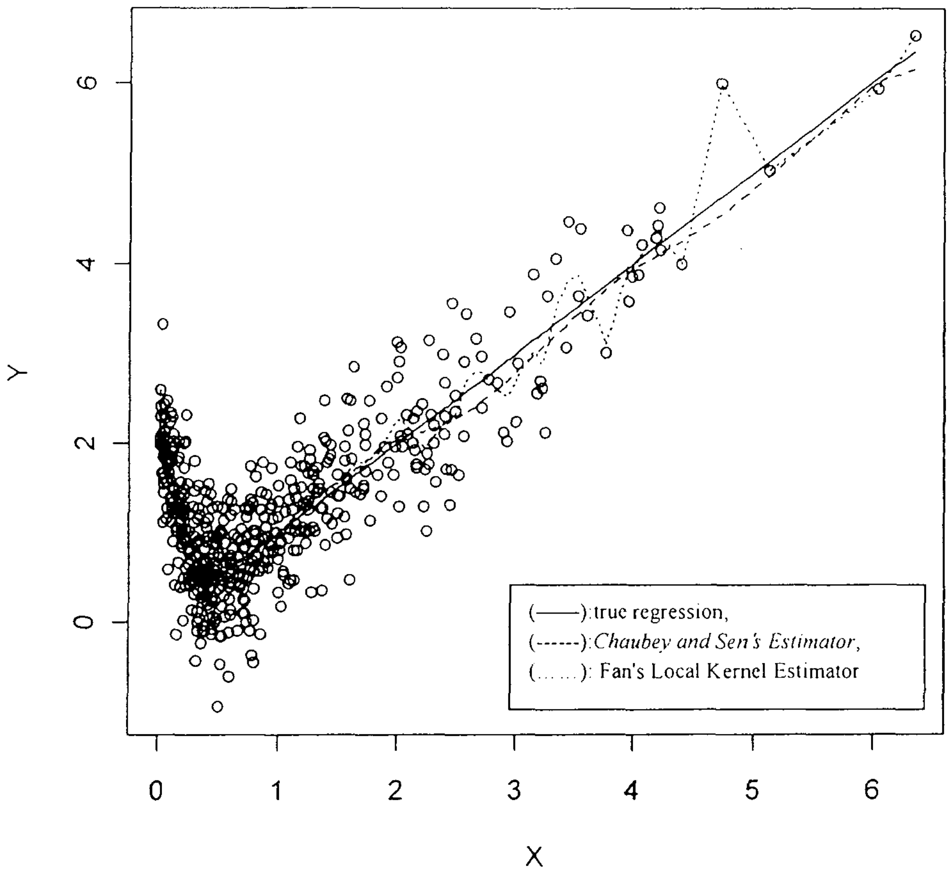


Figure 1(2): scatterplot and regression estimators for simulation IID data with:

Y= X+2*exp(-16*(X^2)) +ε

X:exp(1)

ε:norm(0,0.7^2)

n=500

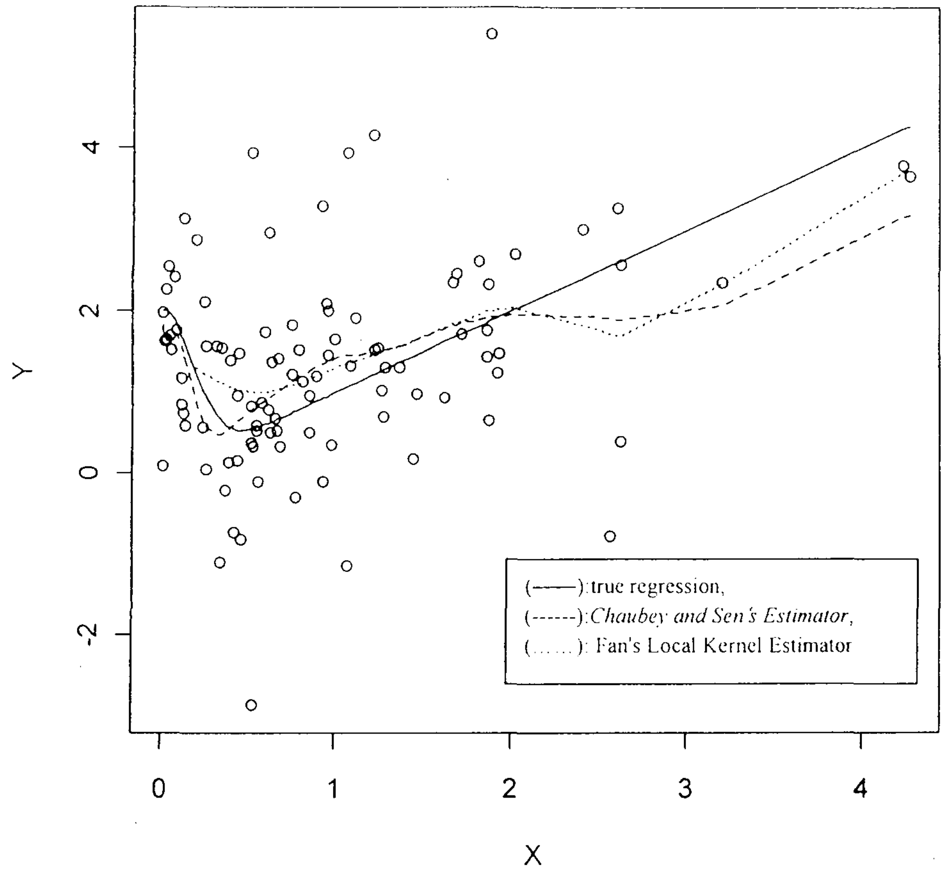


Figure 5(1): scatterplot and regression estimators for simulation IID data with:

Y=X+2*exp(-16*(X^2)) +ε

X:exp(1)

ε:doubleExp

n=100

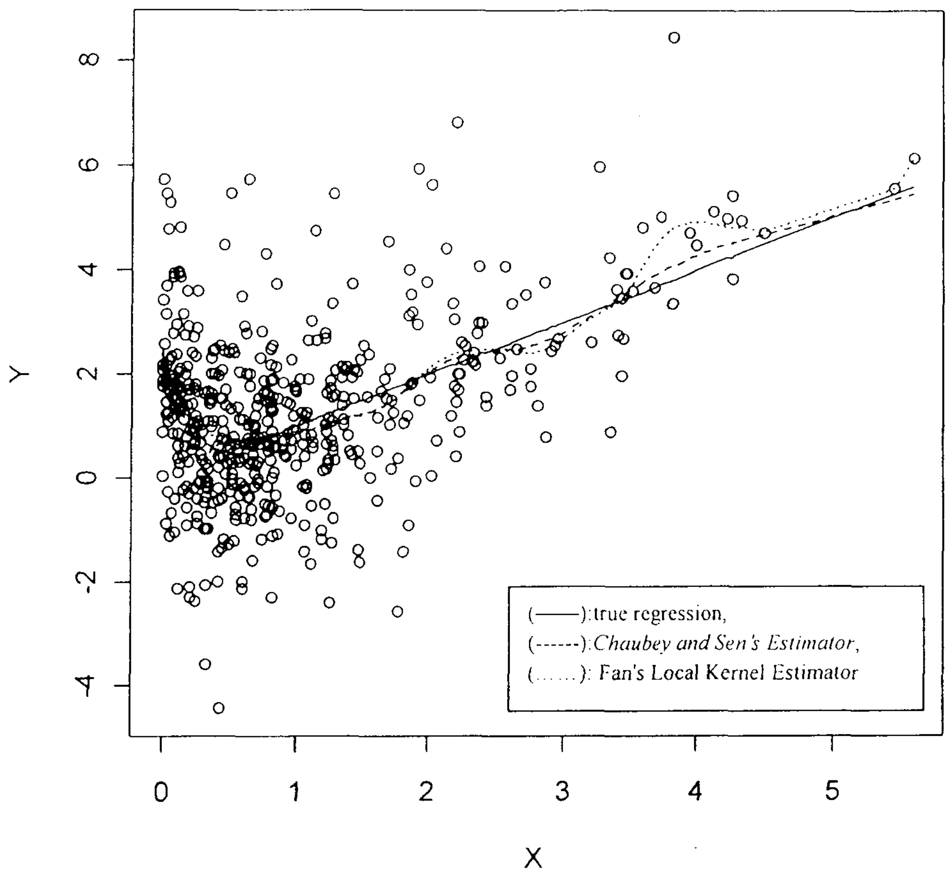


Figure 2(2): scatterplot and regression estimators for simulation IID data with:

Y=X+2*exp(-16*(X^2)) +ε

X:exp(1)

ε:doubleExp

n=500

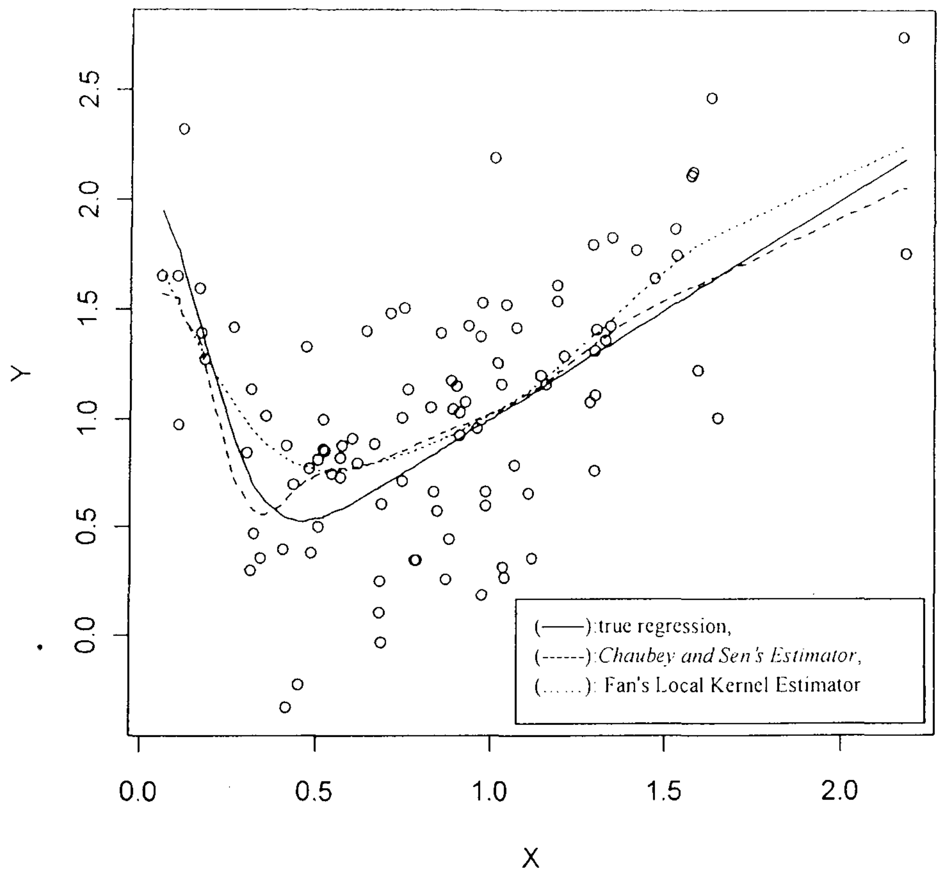


Figure 3(1): scatterplot and regression estimators for simulation IID data with:

Y=X+2*exp(-16*(X^2)) +ε

X:weibul

ε:norm(0,0.7^2)

n=100

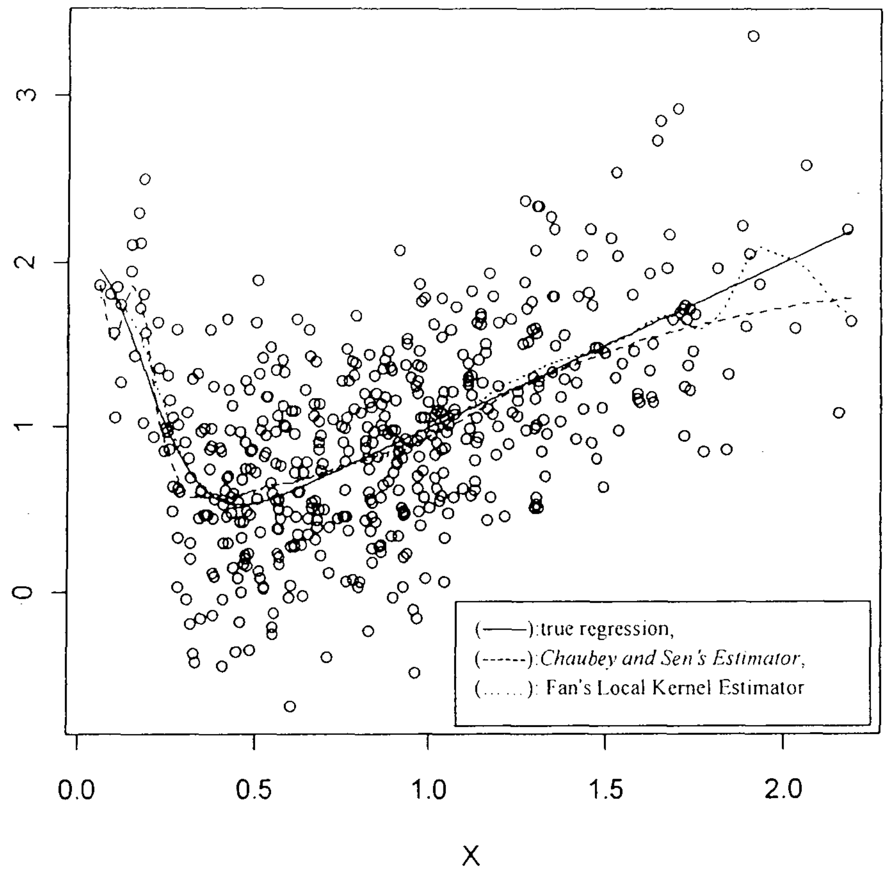


Figure 3(2): scatterplot and regression estimators for simulation IID data with:

Y=X+2*exp(-16*(X^2)) +ε

X:weibul

ε:norm(0,0.7^2)

n=500

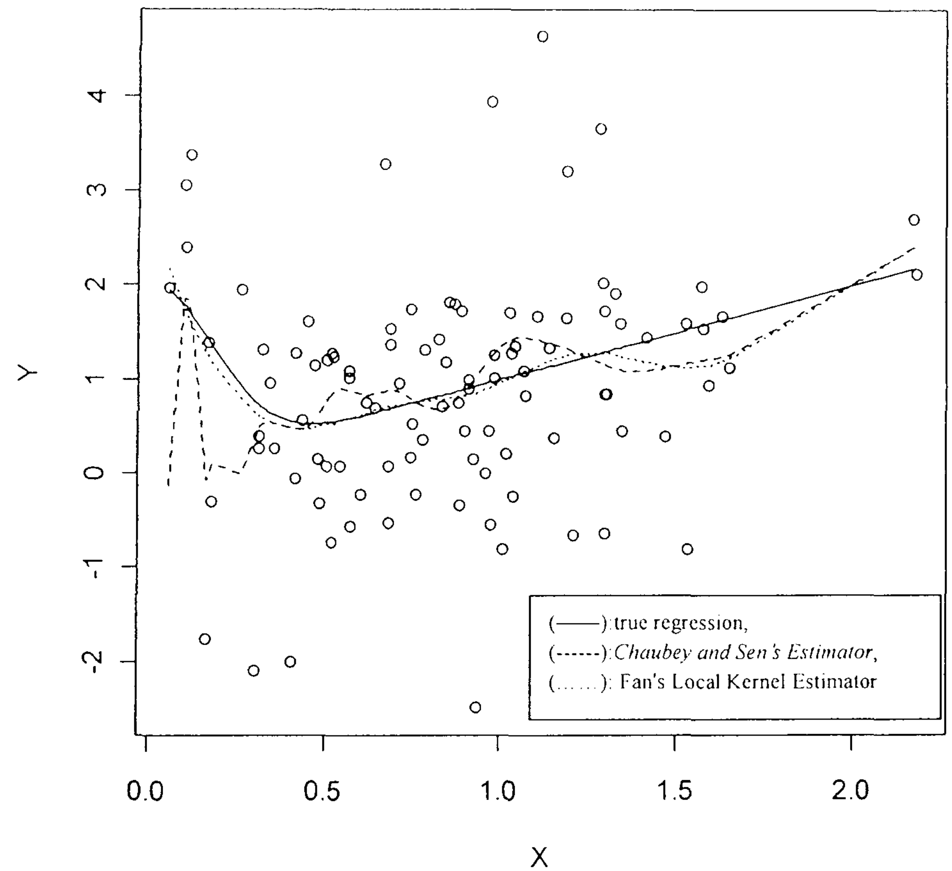


Figure 4(2): scatterplot and regression estimators for simulation IID data with:

Y=X+2*exp(-16*(X^2)) +ε

X: weibull(2,1)

ε:doubleExp

n=100

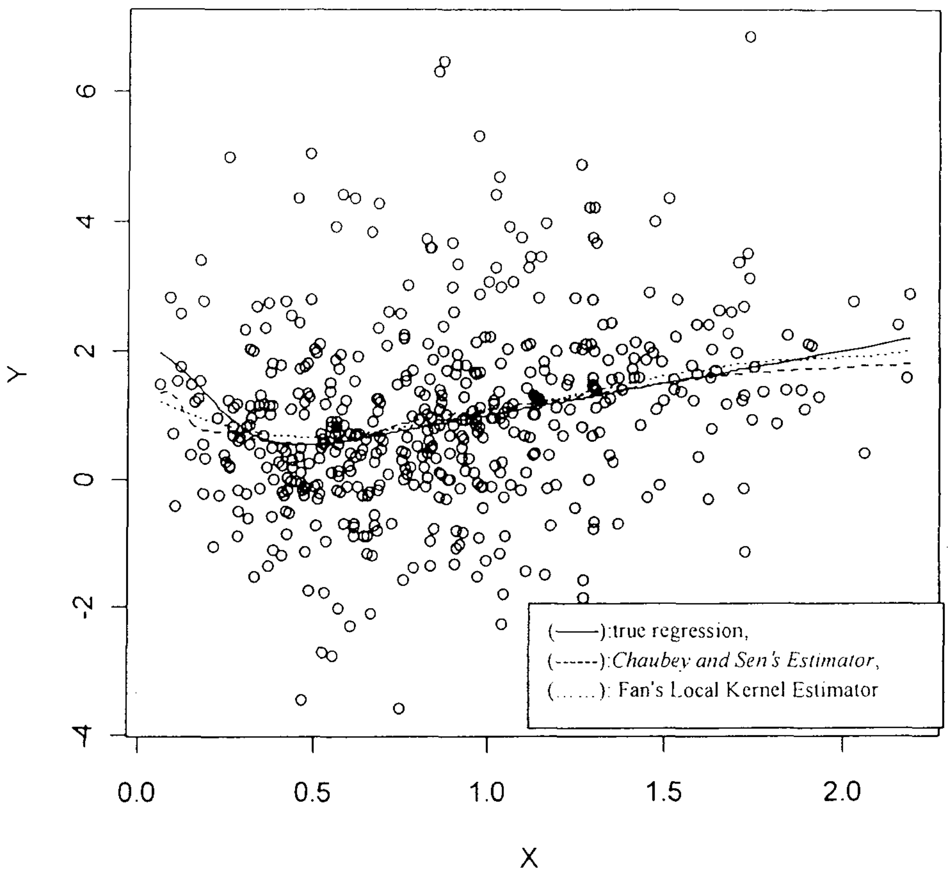


Figure 4(2): scatterplot and regression estimators for simulation IID data with:

Y=X+2*exp(-16*(X^2)) +ε

X: weibull(2,1)

ε:doubleExp

n=500

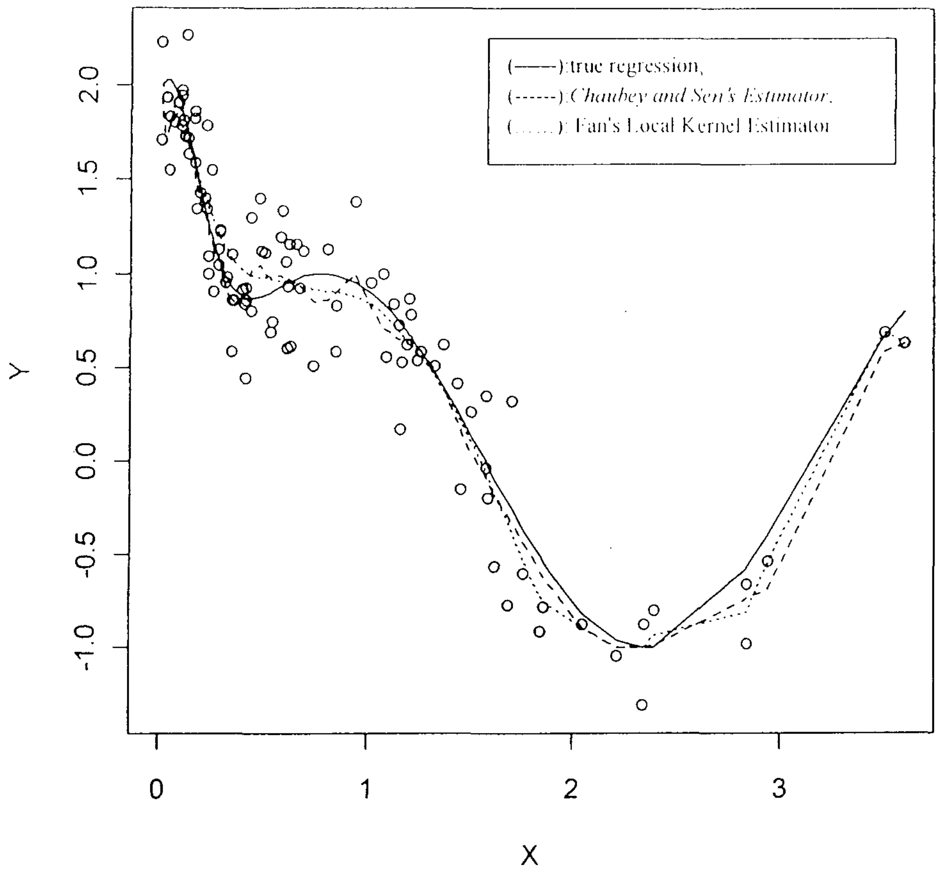


Figure 5(1): scatterplot and regression estimators for simulation IID data with:

Y=sin(2*X)+2*exp(-16*(X^2)) +ε

X:exp(1)

ε:norm(0,0.5^2)

n=100

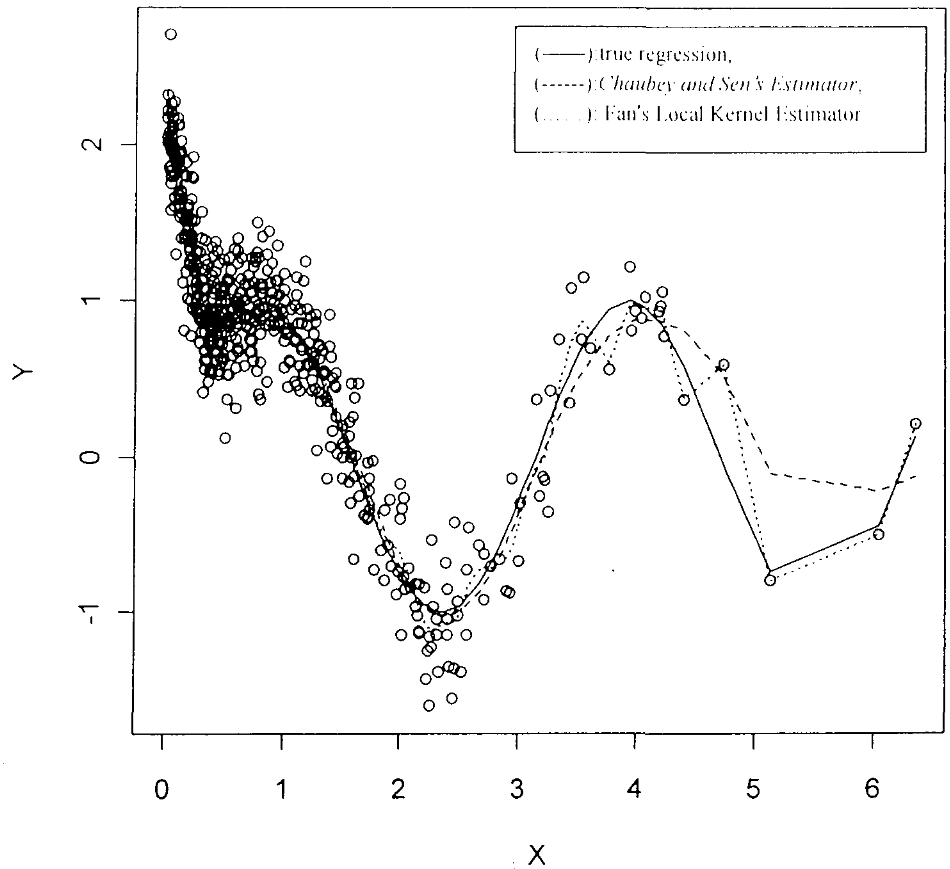


Figure 5(2): scatterplot and regression estimators for simulation IID data with:

Y=sin(2*X)+2*exp(-16*(X^2)) +ε

X:exp(1)

ε:norm(0,0.5^2)

n=500

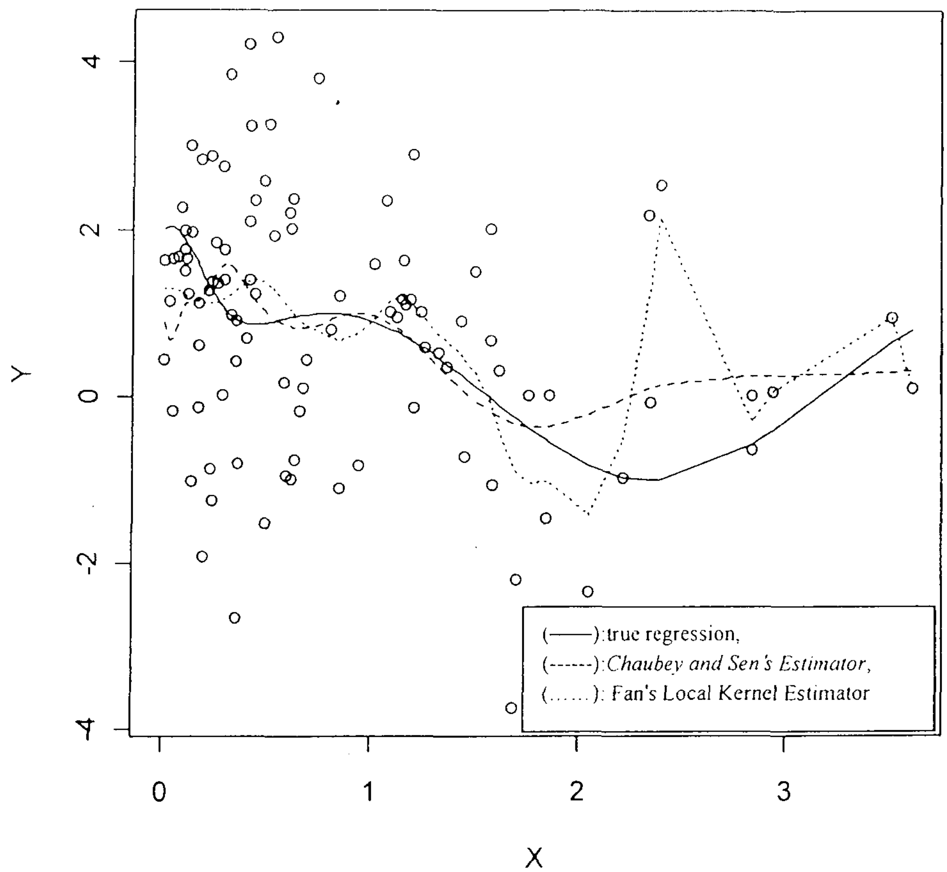


Figure 6(1): scatterplot and regression estimators for simulation IID data with:

Y=sin(2*X)+2*exp(-16*(X^2)) +ε

X:exp(1)

ε:doubleExp

n=100

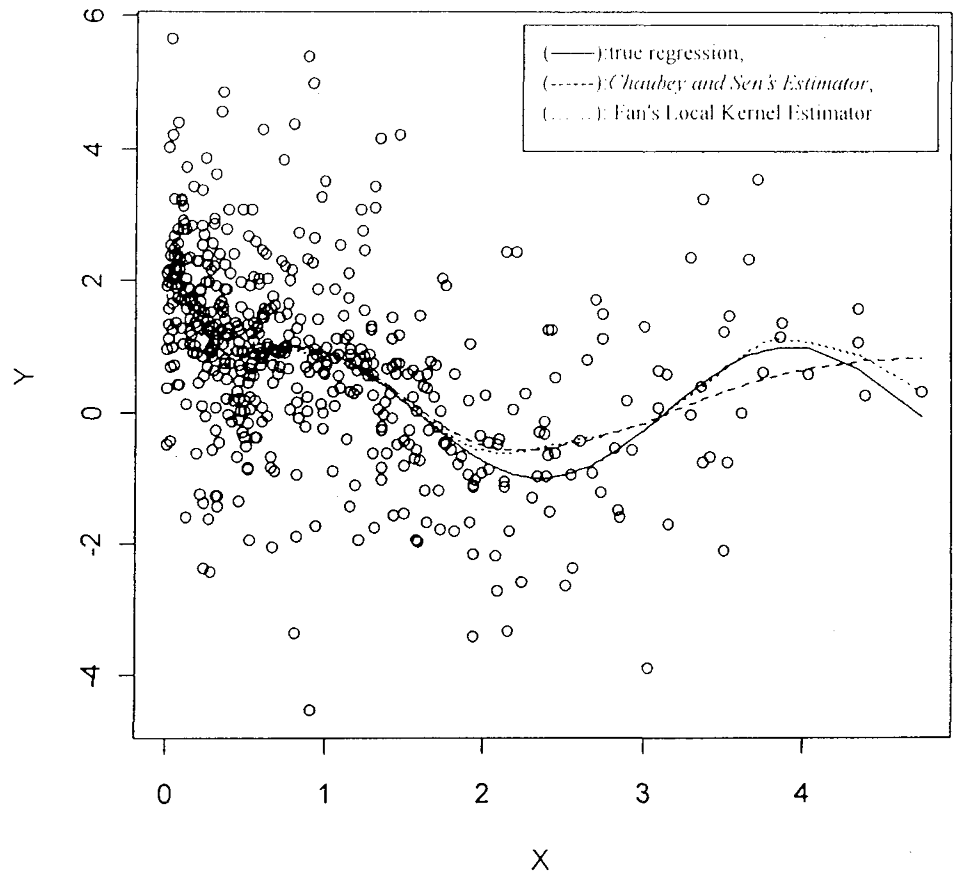


Figure 6(2): scatterplot and regression estimators for simulation IID data with:

Y=sin(2*X)+2*exp(-16*(X^2)) +ε

X:exp(1)

ε:doubleExp

n=500

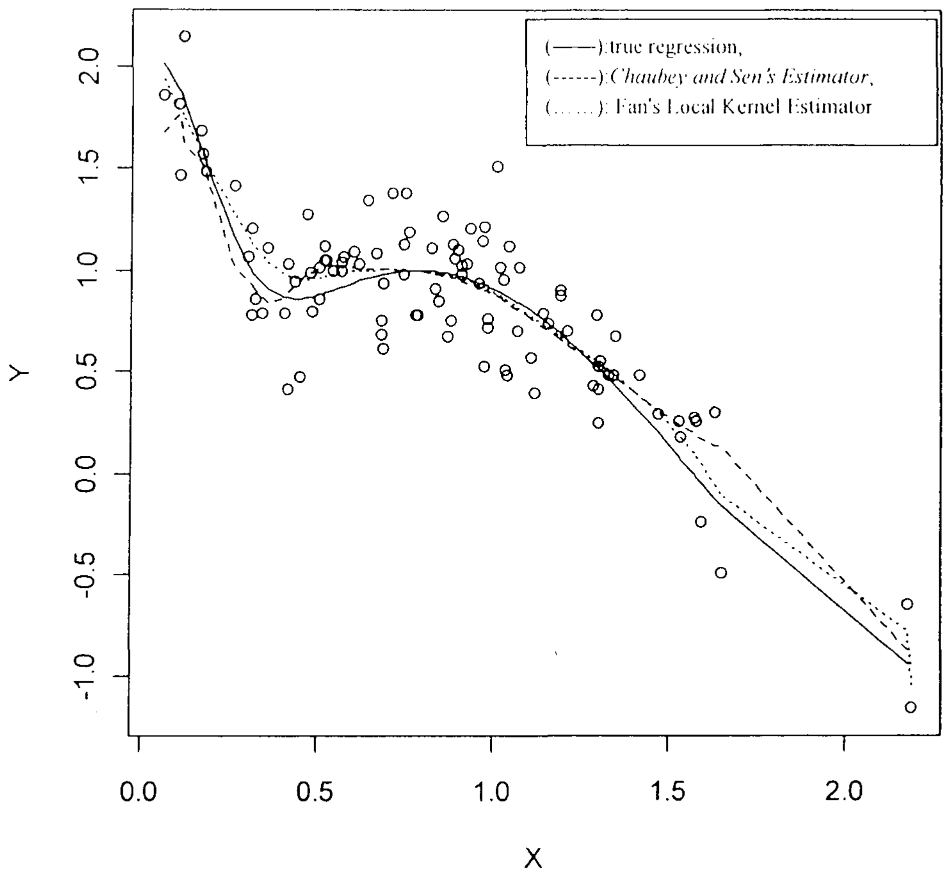


Figure 7(1): scatterplot and regression estimators for simulation IID data with:

Y=sin(2*X)+2*exp(-16*(X^2)) +ε

X:weibul

ε:norm(0,0.5^2)

n=100

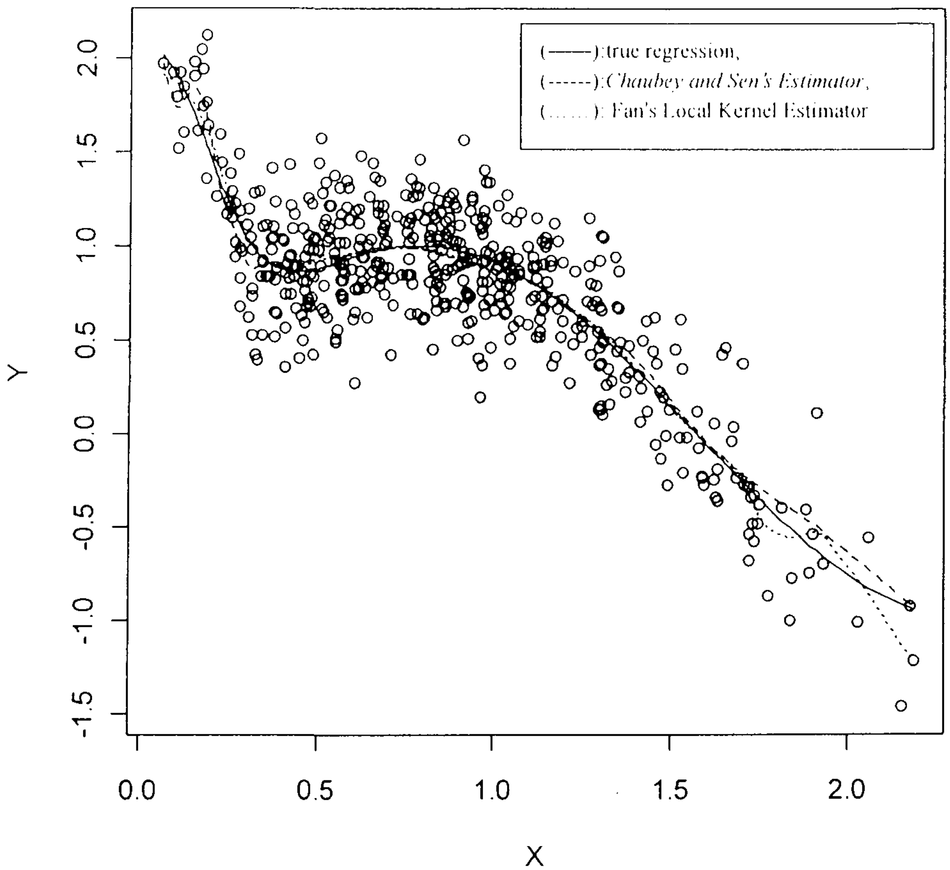


Figure 7(2): scatterplot and regression estimators for simulation IID data with:

Y=sin(2*X)+2*exp(-16*(X^2)) +ε

X:weibul

ε:norm(0,0.5^2)

n=500

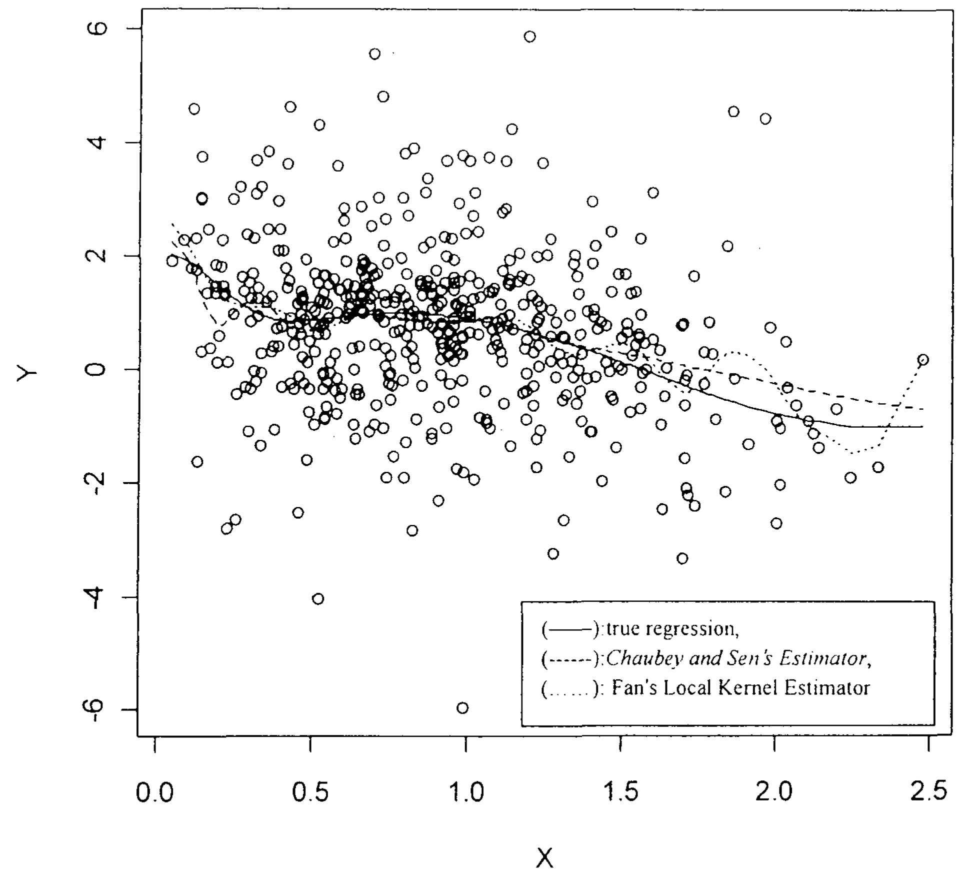


Figure 8(2): scatterplot and regression estimators for simulation IID data with:

Y=sin(2*X)+2*exp(-16*(X^2)) +ε

X: weibull

ε:doubleExp

n=500

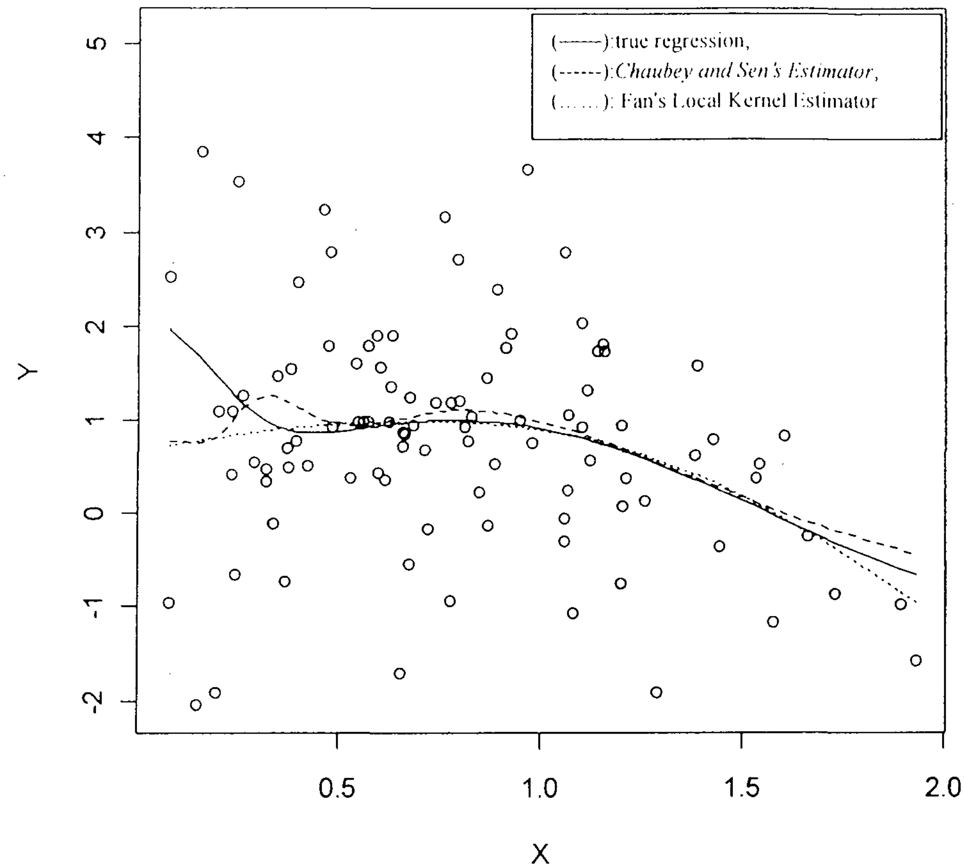


Figure 8(1): scatterplot and regression estimators for simulation IID data with:

Y=sin(2*X)+2*exp(-16*(X^2)) +ε

X: weibull

ε:doubleExp

n=100

3.4 Application: Regression Estimators for Real Data

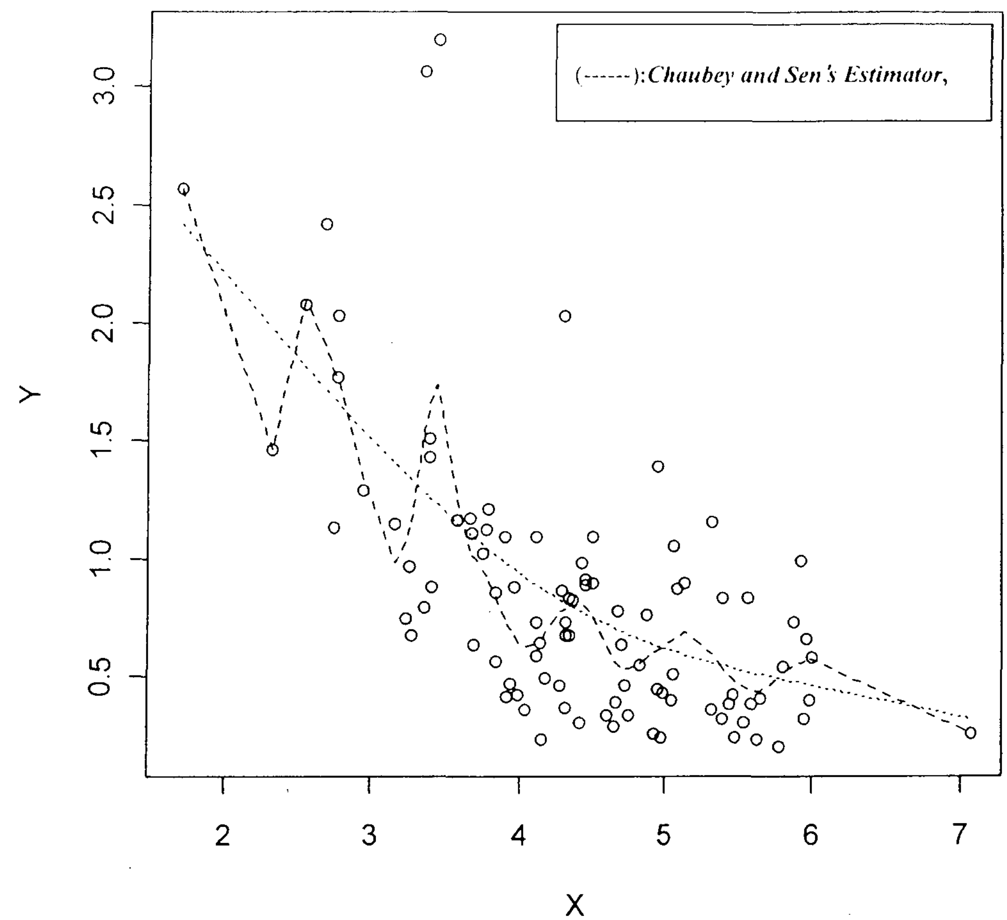


Figure 9: Regression estimators for "hardwood growth data":

(Data set source is from the Chaubey, Laib and Sen's paper, 2008:
initial height (*X*) versus 5-year height-growth (*Y*)*)*

The sample size n=94

The minimizing CV for Chaubey and Sen's Estimator is 0.2144442

while the optional (e,v) is (0, 0.025);

The minimizing CV for Fan's Local Kernel Estimator is 0.2050526

while the optional h is 0.95

4 Future Study

4.1 Validate and Improve the New Smooth Regression Estimator

Before and after using the new regression estimator, we may preliminarily evaluate the validity by checking the diagnostics (outliers and influential observations....)

Many parametric regression techniques such as: PRESS and Cook's distance can be taken advantage.

4.2 Generalize the d-dimensional case

We briefly discuss a generalization of our result to the d-dimensional case. For $d \geq 1$, denote by $\mathbf{X}_i = (X_{i1}, \ldots, X_{id})$ a d-dimensional vector random variable defined on $\mathbb{R}^{+d}$. Let $\mathbf{x} = (x_1, \ldots, x_d) \in \mathbb{R}^{+d}$ and $\epsilon_n = (\epsilon_{1n}, \ldots, \epsilon_{dn})$ such that for any $1 \leq i \leq d$, $\epsilon_{in} \to 0$. Then for any $\mathbf{t} \in \mathbb{R}^{+d}$, the density function defined in (1.3.1.2) takes the form

$$Q_{\mathbf{x}+\epsilon_n,v}(\mathbf{t}) = \frac{1}{(\prod_{i=1}^d \beta_{x_i+\epsilon_{in}})^\alpha \, (\Gamma(\alpha))^d} \left(\prod_{i=1}^d t_i \right)^{\alpha-1} e^{-\alpha \sum_{i=1}^d \frac{t_i}{x_i+\epsilon_{in}}}, \tag{5.2.1}$$

where $\alpha := \alpha_n = 1/v^2, \qquad \beta_{x_i+\epsilon_{in}} = v^2(x_i + \epsilon_{in})$ and $v := v_n$.

Let $\mathbf{Z}_i = (\mathbf{X}_i, Y_i)_{i \in \mathbb{N}}$ be a $\mathbb{R}^{+d} \times \mathbb{R}^+$-valued strictly stationary ergodic sequence. Let ϕ be a Borelian function of $\mathbb{R}^+$ into $\mathbb{R}$. We estimate then $m(\cdot)$ by

$$\tilde{m}_n(\mathbf{x}) = \frac{n^{-1} \sum_{i=1}^n \phi(Y_i) Q_{\mathbf{x}+\epsilon_n,v}(\mathbf{X}_i)}{n^{-1} \sum_{i=1}^n Q_{\mathbf{x}+\epsilon_n,v_n}(\mathbf{X}_i)}. \tag{5.2.2}$$

We need study such generalization of our result: for example, the properties and the applicaion.

4.3 Comparison:More Smooth Regression Estimators and Selectors

Here we consider only Unbiased Cross Validation(UCV),In a futurer researchwe shall consider also Biased Cross Validation(BCV).

In this thesis we compare only Local Kernel smoother, in a futurer researchwe shall consider other Kernel smoother, also various noparametric approaches such as PPT, k-NN and Spline.

Bibliography

[1] Chaubey,Y.P.,Sen,P.K. (1996).On Smooth Estimation of Survival and Density Functions. *Statistics and Decisions* 14,1-22.

[2] Chaubey,Y.P.,Sen,P.K. (1998).On Smooth Estimation of Hazard and Cumulative Hazard Functions. *Frontiers in Probability and Statistics*(eds. S.P.Mukherjee *et al.*) Narosa ,New Delhi,pp.92-100.

[3] Chaubey,Y.P.,Sen,P.K. (1999).On Smooth Estimation of Mean Residual Life. *Journal of Statistical Planning and Inference*,75,223-236.

[4] Chaubey,Y.P., Sen,A., and Sen,P.K.(2007). A New Smooth Density Estimator for Non-negative Random Variables.

[5] Chaubey,Y.P., Laib,N., and Sen,A.(2008). A Smooth Estimator of Regression Function for Non-negative *Dependent* Random Variables.

[6] Härdle, W. (1990). *Applied nonparametric regression.* Cambridge University Press.

[7] Härdle,W.(1991). *Smoothing techniques: with implementation in S.* Springer-Verlag, New York .

[8] Jianqing.Fan and I.Gijbels (2003).Local polynomial modelling and its applications. Boca Raton : CRC Press.

[9] Jianqing Fan and I. Gijbels.(1992). Variable Bandwidth and Local Linear Regression Smoothers : The Annals of Statistics, Vol.20,No.4,pp.2008-2036

[10] Ming-YenCheng,JianqingFan,J.S.Marron(1997).On Automatic Boundary Corrections: The Annals of Statistics, Vol.25,No.4,pp.1691-1708

[11] M.P.Wand,M.C.Jones (1995). Kernel Smoothing. Chapman & Hall

[12] Raymond H. Myers (1990). Classical and modern regression with applications Australia: Duxbury Thomson Learning