The Korean Communications
in Statistics Vol. 4, No. 2, 1997
pp. 453 - 463

Local Bandwidth Selection for Nonparametric Regression[1]

Seung-Woo Lee[2], Kyung-Joon Cha[3]

Abstract

Nonparametric kernel regression has recently gained widespread acceptance as an attractive method for the nonparametric estimation of the mean function from noisy regression data. Also, the practical implementation of kernel method is enhanced by the availability of reliable rule for automatic selection of the bandwidth.

In this article, we propose a method for automatic selection of the bandwidth that minimizes the asymptotic mean square error. Then, the estimated bandwidth by the proposed method is compared with the theoretical optimal bandwidth and a bandwidth by plug-in method. Simulation study is performed and shows satisfactory behavior of the proposed method.

1. Introduction

Recently, the field of nonparametrics has expended its appeal with an array of new tools for statistical analysis. Without making specific distributional assumptions, these new tools offer sophisticated substitute to traditional parametric models for exploring large amounts of scattered data. As one of those tools, nonparametric kernel regression has become a conspicuous statistical research issue. Now, development in computation and the fast computational facilities available to statisticians have had an important consequence upon statistical study, and in particular the advance of nonparametric data analysis procedures. The merits of this approach include brief in terms of interpretability and mathematical analysis.

As with any nonparametric regression procedure, an important choice to be made is the amount of averaging performed to obtain the regression estimate. For a kernel-type estimator, this is controlled by a parameter usually referred to as the bandwidth. When a single bandwidth is used for the whole range of the data, it is generally called a global bandwidth kernel estimator. Otherwise, if the estimated bandwidth depends on a point of estimation, the estimator is called a local bandwidth kernel estimator. Bandwidths that are too small produce estimates that are too wiggly, tending toward interpolation of the data, and bandwidths that

[1] This research is partially supported by the Research Institute for Natural Sciences at Hanyang University, 1997.

[2] Assistant Professor, Department of Applied Statistics, Seo Kyeong University, Seoul 133-704, Korea

[3] Assistant Professor, Department of Mathematics, Hanyang University, Seoul 133-791, Korea

are too large smooth out features in the true mean function. A data-driven bandwidth selector that estimates the correct amount of smoothing is very useful for the analyst.

The main purpose of this paper is to develop a local adaptive bandwidth selection method. To show practical performance of the proposed method, we compare the developed method with plug-in type bandwidth selection method which involves estimation of unknown functional that appears in formulas for the asymptotically optimal bandwidth.

Let $m(t)$, $t \in [0,1]$, be an unknown regression function with p continuous derivatives. Observations y_i, $i = 1, 2, \cdots, n$ of m have been made, which are contaminated with error and are of the form

$$y_i = m(t_i) + \varepsilon_i .$$

The ε_i are independent and identically distributed, their distribution satisfy $E(\varepsilon_i) = 0$ and $E(\varepsilon_i^2) = \sigma^2 < \infty$. The simplifying assumption that $0 = t_1 < t_2 < \cdots < t_n = 1$ are equally spaced in [0,1] is made. The objective is to estimate $m(t)$ formed observations.

A computationally simpler method of nonparametric curve fitting uses kernel estimates. Here the class of kernel estimators of $m(t)$ proposed by Priestley and Chao (1972) is adapted and defined by

$$\widehat{m}(t;h) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{h} K\left(\frac{t-t_i}{h}\right) y_i ,$$

where K is a kernel function and h is the bandwidth. The kernel function, K, is assumed to be continuously differentiable and of order p in the sense that

$$\int_{-1}^{1} z^j K(z)dz = \begin{cases} 1 & j = 0, \\ 0 & j = 1, \cdots, p-1, \\ k_p \neq 0 & j = p. \end{cases}$$

2. Local Bandwidth Selection

There are several local bandwidth selection methods studied in the literature. However, recently, Brockmann, Gasser and Herrmann (1993) showed that theoretical and practical advantages of local bandwidth can be obtained when the optimal bandwidth is estimated from the data.

The proposed approach is based on a simple asymptotic relation between variance and bias^2 which is introduced by Cha and Lee (1995). Schucany (1995) proposed a similar method to find a local bandwidth via the bias and variance decomposition of the mean square error. In fact, he used two term bias expansion and adopted polynomial regression model to estimate bias itself instead of bias^2, moreover he even used a higher order polynomial regression model by multiplying two term bias expansion and its derivative to estimate approximately linear region of the true curve where its second derivative may vanish. Then, he solved the variance and bias decomposition using Newton's method to find the estimated bandwidth which clearly does not guarantee the relation (3).

For the proposed method, the optimal is taken to be in the sense of minimizing the mean square error(mse)

$$\mathrm{mse}[\ \widehat{m}(t;h)] = \mathrm{var}(\ h) + \mathrm{bias}^2(\ h), \tag{1}$$

where var(h) = var[$\widehat{m}(t;h)$] and bias(h) = E[$\widehat{m}(t;h)$] - $m(t)$.

Suppose that $m(t)$ is p times continuously differentiable function on the unit interval. Then, the expected value of $\widehat{m}(t;h)$ at a fixed t is

$$E[\ \widehat{m}(t;h)] = \int_{(t-1)/h}^{t/h} K(z)[m(t) - zhm^{(1)}(t) + \frac{(zh)^2}{2!} m^{(2)}(t) + \cdots$$
$$+ (-1)^p \frac{(zh)^p}{p!} m^{(p)}(t)]dz + o(h^p),$$

where $m^{(j)}(t)$ is the jth derivative of $m(t)$ and $z = \frac{t-s}{h}$. Hence, $[-1,1] \subset [\frac{t-1}{h}, \frac{t}{h}]$ for sufficiently small h, the above expectation becomes

$$E[\ \widehat{m}(t;h)] = \int_{-1}^{1} K(z)[m(t) - zhm^{(1)}(t) + \frac{(zh)^2}{2!} m^{(2)}(t) + \cdots$$
$$+ (-1)^p \frac{(zh)^p}{p!} m^{(p)}(t)]dz + o(h^p)$$
$$= m(t) + \frac{(-1)^p}{p!} h^p m^{(p)}(t) \int_{-1}^{1} z^p K(z)dz + o(h^p)$$

by assumptions of the kernel function of order p. Hence, when $nh \rightarrow \infty$ as $n \rightarrow \infty$ and $h \rightarrow 0$, the asymptotic bias of $\widehat{m}(t;h)$, bias(h), is

$$\mathrm{bias}\,(h) = \frac{(-1)^p}{p!} h^p m^{(p)}(t)\ k_p\ +\ \mathrm{o}(\ h^p),$$

where $k_p = \int_{-1}^{1} z^p K(z)dz$ and the asymptotic variance of $\widehat{m}(t;h)$, var(h), is

$$\mathrm{var}\,(h) = \frac{\sigma^2}{nh} \int_{-1}^{1} K^2(z)\,dz\ +\ \mathrm{o}(\frac{1}{nh})$$

for equally spaced t_i's (see Gasser and Müller, 1979, for details). Hence, by ignoring the vanishing terms, the asymptotically optimal bandwidth which minimizes the asymptotic mse [$\widehat{m}(t;h)$] is

$$h_{opt} = \left\{ \frac{\sigma^2 Q}{2pn(k_p m^{(p)}(t)/p!)^2} \right\}^{1/(2p+1)}, \tag{2}$$

where $Q=\int_{-1}^{1}K^2(z)dz$. Then, it can be easily shown that by substituting (2) into (1), we obtain

$$\mathrm{var}(h_{opt})=2p\,\mathrm{bias}^2(h_{opt}), \tag{3}$$

that is, the optimal bandwidth which minimizes mse balances variance and $2p\mathrm{bias}^2$. In other words, the intersection of variance and $2p\mathrm{bias}^2$ gives the optimal bandwidth.

For a fixed h, now let us express $\mathrm{var}(h)$ and $\mathrm{bias}^2(h)$ as

$$\mathrm{var}(h)=\frac{A}{nh}+o(\frac{1}{nh}) \text{ and } \mathrm{bias}^2(h)=Bh^{2p}+o(h^{2p}) \tag{4}$$

for constants A and B.

As is often the case, it is usual that we can consider the remainder term as error term. Hence, (4) can be considered having the structure of regression models in power of h. Thus, for fixed h's, we may apply the least squares method to estimate A and B, and if we can get $\hat{A}$ and $\hat{B}$ for A and B, the estimated adaptive local bandwidth becomes

$$\hat{h}_t=\left\{\frac{\hat{A}}{2pn\hat{B}}\right\}^{1/(2p+1)}. \tag{5}$$

Now, let $\mathrm{var}(h)$ be modeled by, for fixed h_j's,

$$v_j=\frac{A}{nh_j}+\zeta_j, \quad j=1,2,\cdots,k, \tag{6}$$

where ζ_j are error terms. In order to estimate A, let us look at the $\mathrm{var}(h)$. It is easy to see that the exact variance of $\widehat{m}(t;h)$ is

$$\frac{\sigma^2}{n^2h^2}\sum_{i=1}^{n}K^2(\frac{t-t_i}{h}),$$

thus we need only to estimate σ^2 to estimate variance. Hence, we can adopt v_j as

$$v_j=\frac{\hat{\sigma}^2}{n^2h_j^2}\sum_{i=1}^{n}K^2(\frac{t-t_i}{h_j})$$

for fixed h_j's.

Then, (6) yields the least squares estimator of A as

$$\hat{A}=\frac{n\sum_{j=1}^{k}(v_j/h_j)}{(\sum_{j=1}^{k}1/h_j^2)}.$$

Similarly, let $\mathrm{bias}^2(h)$ be modeled by, for fixed h_j's,

$$b_j^2=Bh_j^{2p}+\xi_j, \quad j=1,2,\cdots,k, \tag{7}$$

where ξ_j are error terms. In order to estimate B, let us look at

$$\frac{1}{nh}\sum_{i=1}^{n} K_D\left(\frac{t-t_i}{h}\right)y_i,$$

where $K_D(z) = K_p(z) - K_{p+2}(z)$ with pth and (p+2)th order kernels. Here, we used $K_D(z)$ since the expected value of $\frac{1}{nh}\sum_{i=1}^{n} K_D(\frac{t-t_i}{h})y_i$ is

$$\begin{aligned}
&E\left[\frac{1}{nh}\sum_{i=1}^{n} K_D\left(\frac{t-t_i}{h}\right)y_i\right] \\
&\quad = E\left\{\frac{1}{nh}\sum_{i=1}^{n} K_p\left(\frac{t-t_i}{h}\right)y_i - \frac{1}{nh}\sum_{i=1}^{n} K_{p+2}\left(\frac{t-t_i}{h}\right)y_i\right\} \\
&\quad = \left\{m(t) + \frac{(-1)^p}{p!}h^p m^{(p)}(t)k_p + \frac{(-1)^{p+2}}{(p+2)!}h^{p+2}m^{(p+2)}(t)k_{p+2} + o(h^{p+2})\right\} \\
&\qquad - \left\{m(t) + \frac{(-1)^{p+2}}{(p+2)!}h^{p+2}m^{(p+2)}(t)k'_{p+2} + o(h^{p+2})\right\},
\end{aligned}$$

where $k'_{p+2} = \int_{-1}^{1} z^{p+2}K_{p+2}(z)dz$. Thus, taking the difference of two asymptotic expressions and ignoring the vanishing terms leave the leading term $\frac{(-1)^p}{p!}h^p m^{(p)}(t)k_p$. Hence, we can adopt b_j as

$$b_j = \frac{1}{nh_j}\sum_{i=1}^{n} K_D\left(\frac{t-t_i}{h_j}\right)y_i.$$

for fixed h_j's. As we showed, the expected value of b_j is

$$E(b_j) = \frac{(-1)^p}{p!}h_j^p m^{(p)}(t)k_t$$

by taking only leading term, thus

$$\begin{aligned}
\{E(b_j)\}^2 &= \left\{\frac{(-1)^p}{p!}h_j^p m^{(p)}(t)k_p\right\}^2 \\
&= \left\{\frac{1}{p!}m^{(p)}(t)k_p\right\}^2 h_j^{2p} \\
&= Bh_j^{2p}
\end{aligned}$$

which need to be estimated. Also, from (7),

$$E(b_j^2) = E(Bh_j^{2p} + \xi_j) = Bh_j^{2p}.$$

Thus, we can use (7) to find the least squares estimator of B and this yields

$$\hat{B} = \frac{\sum_{j=1}^{k} b_j^2 h_j^{2p}}{\sum_{j=1}^{k} h_j^{4p}}.$$

Hence, we can get the estimated local bandwidth, $\hat{h}_t$, defined by (5).

Therefore, with predetermined fixed values of h's, both v_j and b_j^2 can be obtained. And then, given several estimated values of v_j and b_j^2 fitting the relations (6) and (7), estimates $\hat{A}$ and $\hat{B}$ can be obtained. It should be noticed that the data-driven bandwidth $\hat{h}_t$ satisfies (3), i.e., balances variance and bias^2.

Now, let us consider another way to estimate the unknown $m^{(p)}(t)$ in (2). One way to estimate $m^{(p)}(t)$ is the plug-in method which was introduced by Gasser, Müller and Mammitzsch (1985) and Müller (1985).

In problems of nonparametric curve estimation, the optimal amount of smoothing depends on unknown characteristics of the curve, such as derivatives of the curve at the point of estimation. One way of estimating those characteristics is to construct one or more preliminary curve estimators, compute derivative estimators from those, and substitute back into (2), it is called the plug-in approach to local adaptive bandwidth selection.

Müller (1985) and Staniswalis (1989) showed that an empirical approach to selecting the amount of smoothing is to employ pilot estimators to approximate those derivatives, and substitute the approximate values into an analytical formula for the desired local bandwidth such as (2). Gasser, Kneip and Köhler (1991) studied an iterative plug-in approach. Recently, Fan and Gijbels (1995) proposed a method which combines both plug-in and cross-validation for local least squares regression.

Woodroofe (1970) introduced a version of the plug-in selector to use the data to choose the bandwidth of a kernel density estimator. Also, Sheather (1986) developed a method so-called "solve-the-equation" to overcome this difficulty. The basic idea is to substitute estimates into an asymptotic representation of the optimal bandwidth. Such methods have been slowly gain acceptance because care must by taken concerning which estimates are plugged in.

Now, for plug-in type estimator, let us consider the case of $p=2$ because it is simple but important. When $p=2$, the asymptotically optimal bandwidth, h_{opt} is

$$h_{opt} = \left\{ \frac{\sigma^2 Q}{4n(k_2 m^{(2)}(t)/2!)^2} \right\}^{1/5}. \tag{8}$$

However, (8) still contains unknown $m^{(2)}(t)$, thus plug-in type estimator adapts $\hat{m}^{(2)}(t;h^*)$ as

$$\hat{m}^{(2)}(t;h^*) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{h^{*3}} K^*\left(\frac{t-t_i}{h^*}\right) y_i, \tag{9}$$

where $K^*(z) = \frac{105}{16}(-5z^4+6z^2-1)$, $|z| \leq 1$, the optimal kernel of order (2,4) from Gasser, Müller and Mammitzsch (1985). Thus, from (9), the optimal bandwidth that minimizes the asymptotic mean square error is

$$h^*_{opt} = \left\{ \frac{5\, \hat{\sigma}^2 Q^*}{4n(k^*_4\ m^{(4)}(t)/4!)^2} \right\}^{1/9},$$

where $Q^* = \int_{-1}^{1} K^{*2}(z) dz$ and $k^*_4 = \int_{-1}^{1} z^4\, K^*(z) dz$.

Therefore, the estimated plug-in type bandwidth can be expressed as .

$$\hat{h}_{plug} = \left\{ \frac{\hat{\sigma}^2 Q}{4n(k_2 \ \widehat{m}^{(2)}(t; h^*_{opt})/2!)^2} \right\}^{1/5} . \qquad (10)$$

3. Finite Sample Simulation

We conduct a simulation study to evaluate and compare the proposed approach with the bandwidth selectors described in Section 2.

To estimate A defined in (6), we only need to estimate σ^2. For this purpose, we use the estimator proposed by Gasser, Sroka and Jennen-Steinmetz (1986) and used by Stainswalis (1989) which has the form

$$\hat{\sigma}^2 = \frac{1}{6(n-2)} \sum_{i=2}^{n-1} [y_{i-1} - 2y_i + y_{i+1}]^2 .$$

The t_i's are generated from the uniform distribution on [0,1]. The sample size is taken as $n=50$ and $n=100$ for comparisons of numerical performance based on sample sizes. Normal errors $\sigma=0.2$ is used throughout. The true regression function is

$$m(t) = -1.2\sin(3.5\pi t + 2.5) + \sin(1.8\pi t - 2) \qquad (11)$$

with an equidistant design on [0,1].

A simulation is undertaken to compare the optimal bandwidth given in (2), the local adaptive bandwidth given in (5) and the plug-in bandwidth from (10).

In order to simplify the simulation study, the case of p=2 is considered. For $K_D(z)$ used to get $\hat{h}_t$, the Epanechnikov kernel, $K(z) = \frac{3}{4}(1-z^2)$, $|z| \leq 1$, is used as a second order kernel, also the kernel $K(z) = \frac{15}{32}(7z^4 - 10z^2 + 3)$, $|z| \leq 1$, which is found by Gasser and Müller (1979) and Gasser, Müller and Mammitzsch (1985), is used as the 4th order kernel. For the predetermined grid of bandwidths that are used to estimate $\hat{A}$ and $\hat{B}$, seven equally spaced bandwidths between 0.06 and 0.3 are used. These are obtained through simulations, also simulation results reveals that small changes in predetermined minimum and maximim bandwidths do not severely affect the overall performance of $\hat{h}_t$. To make comparisons, the true $m^{(2)}(t)$ and $m^{(4)}(t)$ that can be obtained from (11) are used for h_{opt} and h^*_{opt}, respectively,

Table 1 and Table 2 give the results for the asymptotic true optimal bandwidth from equation (8), the local adaptive bandwidth given in (5) and the plug-in type bandwidth from (10). We can clearly see that $\hat{h}_t$ closely estimates the true optimal bandwidth and is competitive with plug-in estimator. In fact, it can be realized that $\hat{h}_t$ is more stable than $\hat{h}_{plug}$ when the sample size small. For overall, $\hat{h}_t$ shows better performance than $\hat{h}_{plug}$.

Table 1

($n=50$, $\sigma=0.2$)

t	h_{opt}	$\hat{h}_t$	$\hat{h}_{plug}$
0.02626	0.08084	0.08831	0.08303
0.10707	0.07172	0.06133	0.06088
0.22828	0.05384	0.05866	0.04333
0.24848	0.05529	0.05237	0.04274
0.30909	0.06966	0.04749	0.04467
0.45050	0.06392	0.06972	0.04311
0.47070	0.06138	0.05993	0.04193
0.51111	0.06087	0.05410	0.04192
0.69292	0.07102	0.06877	0.04635
0.71313	0.06531	0.06976	0.04490
0.81414	0.06467	0.06803	0.04438
0.95555	0.06766	0.06147	0.05564

Table 2

($n=100$, $\sigma=0.2$)

t	h_{opt}	$\hat{h}_t$	$\hat{h}_{plug}$
0.02311	0.06802	0.05083	0.05343
0.10351	0.06427	0.05159	0.07032
0.12361	0.05613	0.05633	0.06020
0.29447	0.05575	0.05249	0.03817
0.30452	0.05890	0.05072	0.04255
0.31457	0.06305	0.04930	0.06751
0.42512	0.06133	0.07746	0.05163
0.55577	0.05881	0.05211	0.04537
0.67638	0.06869	0.06619	0.05165
0.75678	0.05263	0.05388	0.03348
0.82713	0.05902	0.05039	0.03223
0.93768	0.06733	0.07383	0.05660

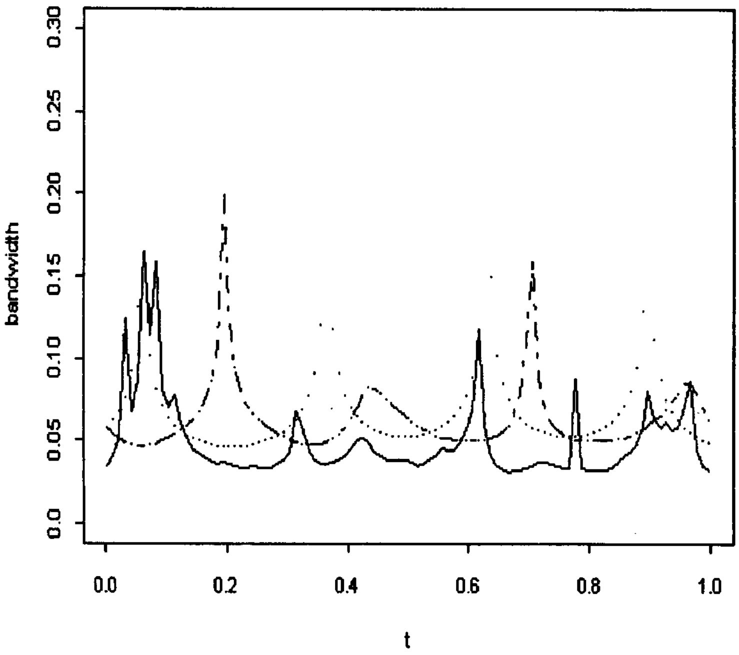

Figure 1 : Overlaying estimated local bandwidth plots of h_{opt} (dotted line), $\hat{h}_t$ (dashed line), $\hat{h}_{plug}$ (solid line) with $n=100$, $\sigma=0.2$

For comparison, the overlay plots of h_{opt}, $\hat{h}_t$, and $\hat{h}_{plug}$ are shown in Figure 1. It can be seen that an appropriate pattern of local smoothing has been achieved. Some small discrepancy between h_{opt}, $\hat{h}_t$, and $\hat{h}_{plug}$ is caused by random error. That is, there is a tendency for $\hat{h}_t$ to follow the asymptotically optimal bandwidths and the plug-in bandwidths. This result is also supported by Figure 2.

Since the goal of nonparametric regression is to fit an entire curve that shows relations between independent and dependent variables, it is important to investigate the finite sample behavior of the whole estimated curve.

Figure 2 shows the overlay plots of the true and estimated curve with observations. We can see that the estimated curve is well-fitted to the true curve. That is, the proposed method detects the bimodality of the true curve and very sharp slope of both ends, hence it is stable for a finite sample. This simulation demonstrates that the proposed method is stable and the estimated curve is close to the true curve for even small sample as well as large sample size.

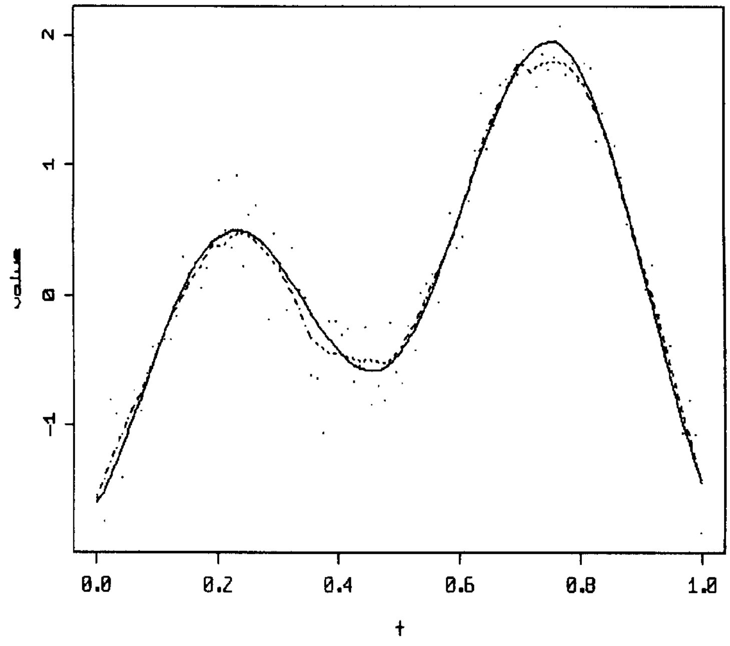

Figure 2 : Overlaying plots of estimated curve with $\hat{h}_t$ (dotted line), observations (points) & true function (solid line) with $n=100$, $\sigma=0.2$

4. Concluding Remarks

The merit of the proposed approach is that it does not require an initial value to estimate the optimal bandwidth, also it could be applied to the global bandwidth selection method as well as nonparametric density estimation. Since the proposed method tries only to find the intersection of variance and squared bias, i.e., find the point of balancing variance and squared bias, using variance and bias estimators, we need not search for a minimum of mean square error from noisy curve. Thus, it is simple and practical.

We expect that the concern about the boundary bias would not be so great, however the boundary problem still need to take account.

References

[1] Brockmann, M., Gasser, Th. and Herrmann, E. (1993). Locally Adaptive Bandwidth Choice for Kernel Regression Estimators, *Journal of the American Statistical Association*, Vol. 88, No. 424, 1302-1309.

[2] Cha, K. J. and Lee, S. W. (1995). Nonparametric Regression Curve Estimation by Locally Data-Dependent Bandwidths, *Proceedings of the Autumn Conference*, 40-47.

[3] Fan, J. and Gijbels, I. (1995). Data-Driven Bandwidth Selection in Local Polynomial Fitting: Variable Bandwidth and Spatial Adaption, *Journal of the Royal Statistical Society*, B57, 371-394.

[4] Gasser, Th., Kneip, A. and Köhler (1991). A Fast and Flexible Method for Automatic Smoothing, *Journal of the American Statistical Association*, Vol.86, No.415, 643-652.

[5] Gasser, Th. and Müller, H. G. (1979). *Kernel Estimation of Regression Function: In Smoothing Techniques for Curve Estimation* (Gasser, Th. and Rosenblatt, M. eds), 23-68, Heidelberg, Springer-Verlag.

[6] Gasser, Th., Müller, H. G. and Mammitzsch, V. (1985). Kernels for Nonparametric Curve Estimation, *Journal of the Royal Statistical Society*, B47, 238-252.

[7] Gasser, Th., Sroka, L. and Jennen-Steinmetz, C. (1986). Residual Variance and Residual Pattern in Nonlinear Regression, *Biometrika*, Vol.73, 625-633.

[8] Müller, H. G. (1985). Empirical Bandwidth Choice for Nonparametric Kernel Regression by Means of Pilot Estimators, *Statistics and Decisions*, Supplement Issue, No.2, 193-206.

[9] Priestley, M. B. and Chao, M. T. (1972). Nonparametric Function Fitting, *Journal of the Royal Statistical Society*, B34, 358-392.

[10] Schucany, W. R. (1995). Adaptive Bandwidth Choice for kernel Regression, *Journal of the American Statistical Association*, Vol.90, No.430, 535-540.

[11] Sheather, S. J. (1986). An Improved Data-Based Algorithm for Choosing The Window Width When Estimating The Density at A Point, *Computational Statistics and Data Analysis*, Vol.4, 61-65.

[12] Staniswalis, J. G. (1989). Local Bandwidth Selection for a Kernel Estimates, *Journal of the American Statistical Association,* Vol.84, No.405, 284- 288.

[13] Woodroofe, M. (1970). On Choosing a Delta-Sequence, *Annals of Mathematical Statistics*, Vol. 41, 1665-1671.