응용통계연구 제 7권 1호
1994년 3월 pp. 149 - 158

An Adaptive Bandwidth Selection Algorithm in Nonparametric Regression

Kyung-Joon Cha[1], Seung-Woo Lee[2]

Abstract

Nonparametric regression technique using kernel estimator is an attractive alternative that has received some attention, recently. The kernel estimate depends on two quantities which have to be provided by the user: the kernel function and the bandwidth. However, the more difficult problem is how to find an appropriate bandwidth which controls the amount of smoothing (see Silverman, 1986). Thus, in practical situation, it is certainly desirable to determine an appropriate bandwidth in some automatic fashion. Thus, the problem is to find a data-driven or adaptive (i.e., depending only on the data and then directly computable in practice) bandwidth that performs reasonably well relative to the best theoretical bandwidth. In this paper, we introduce a relation between bias and variance of mean square error. Thus, we present a simple and effective algorithm for selecting local bandwidths in kernel regression.

1.Introduction

Bandwidth selection occupies an important role in the literature of nonparametric regression (cf. Marron, 1989 or Eubank, 1988). With few exceptions, the primary emphasis of this work has been on the selection of globally optimal bandwidth. However, Müller (1988) and Staniswalis (1989) showed that gains in estimator performance could be realized by optimizing the bandwidth locally rather than on global basis. Thus, in this paper, we present a simple and effective method for selecting local bandwidths in kernel regression.

Consider nonparametric regression to model the independent variable, y_i, by

$$y_i = m(t_i)+\varepsilon_i, \qquad i = 1,2,\cdots,n.$$

Here, the ε_i are independent, identically distrbuted random variables with zero mean and common variance σ^2, the t_i are the non-stochastic design points satisfying $0 \le t_1 < t_2 < \cdots < t_n \le 1$ and m is an unknown function. Without having to assume more about m than it satisfies such smoothness conditions, we may want to estimate $m(t)$ at some fixed argument t.

1) 133-791 Department of Mathematics, Hanyang University, Seongdong-Ku, Haengdang-Dong, 17, Seoul, Korea
2) 133-791 Department of Mathematics, Hanyang University, Seongdong-Ku, Haengdang-Dong, 17, Seoul, Korea

There are many interesting nonparametric estimators for $m(t)$. Examples of these can be found in Eubank (1988) and Gasser and Müller (1979). In particular, the class of kernel estimators of $m(t)$ proposed by Priestly and Chao (1972) is defined by

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

where t_i are equally spaced and $h>0$ is the bandwidth or window width. The function K is called a kernel function. It is assumed to be continuously differentiable, symmetric with support on [-1,1]. When it satisfies

$$\int_{-1}^{+1} z^j K(z)dz = \begin{cases} 1, & j = 0 \\ 0, & j = 1,2,\cdots,p-1 \\ k_p, & j = p, \end{cases} \quad (1)$$

it is called a kernel of order p.

To use $\widehat{m_h}(t)$ in practice, one requires to choose both h and K. Discussion of methods for selecting K can be found in Müller (1988). We will concentrate here on the problem of selecting h. The value used for h will be allowed to depend on the point of estimation t. The goal is to find a good choice of h for each value of t in the sense of making the mean squre error(mse) of estimation as small as possible.

There are several data adaptive local bandwidth selection techniques that have been proposed in the literature. Modifications of squared-error cross validation for consistent estimation of optimal local smoothing have been introduced by Hall and Schucany (1989). An alternative resampling approach that uses the bootstrap to estimate mse[$\widehat{m_h}(t)$] is described by Härdle and Bowman (1988). Two other approaches to estimate the mse that use pilot estimates of $m(t)$ have been studied by Müller (1985) and Staniswalis (1989). All of these algorithms involve a search for a local minimum of an estimated mse and require the specification of some other tuning parameter such as a global bandwidth for a pilot estimate of m. In contrast, the technique that is proposed in this paper does not require such initial value and there is no search required for minima of a cross-validation or estimated mse function.

2. Adaptive Bandwidths

The approach that will be proposed stems from simple asymptotic analysis. Now, by the Taylor expansion about t, it can be shown that if $m\in C^P[0,1]$, the expected value of $\widehat{m}$ at a fixed t is

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

where $m^{(j)}(t)$ is the j^{th} derivative of $m(t)$ and $z=(t-s)/h$.

For h sufficiently small, so that $[-1,1] \subset [\frac{t-1}{h}, \frac{t}{h}]$, the above expansion can be reduecd to

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

Hence, as results of (1), the asymptotic bias of $\widehat{m_h}(t)$ is

$$E[\widehat{m_h}(t)] - m(t) = \frac{(-1)^p}{p!} h^p m^{(p)}(t) k_p + o(h^p). \tag{2}$$

Also, by similar methods used in (2), the asymptotic variance of $\widehat{m_h}(t)$ is

$$\begin{aligned} var[\widehat{m_h}(t)] &= \frac{1}{n^2h^2} \sum_{i=1}^{n} K^2(\frac{t-t_i}{h}) var(y_i) \\ &= \frac{\sigma^2}{nh} \int_{(t-1)/h}^{t/h} K^2(z) dz + O(\frac{1}{n^2h^2}), \end{aligned}$$

where $z=(t-s)/h$. For sufficiently small h, this expression also reduces to

$$var[\widehat{m_h}(t)] = \frac{\sigma^2}{nh} \int_{-1}^{+1} K^2(z) dz + o(\frac{1}{nh}).$$

Therefore, the mean square error can be expressed as

$$mse[\widehat{m_h}(t)] = \frac{\sigma^2}{nh} Q + [\frac{h^p}{p!} m^{(p)}(t) k_p]^2 + o(\frac{1}{nh}) + o(h^{2p}), \tag{3}$$

where $Q = \int_{-1}^{+1} K^2(z) dz$. Hence, minimization of (3) with respect to h yields

$$h_t^* = \Big\{ \frac{\sigma^2 Q}{2pn(k_p m^{(p)}(t)/p!)^2} \Big\}^{1/(2p+1)} \tag{4}$$

if we ignore higher order terms. Therefore, by plugging (4) into (3), it can be easily shown that

$$var(h_t^*) = 2p\, bias^2(h_t^*), \tag{5}$$

again neglecting higher order terms.

The basic proposal here is to capatalize on the balance between variance and bias present ed in (5). Thus, for large n, we should have for any fixed h that

$$var[\widehat{m_h}(t)] \sim \frac{A}{nh} \tag{6}$$

$$bias^2[\widehat{m_h}(t)] \sim Bh^{2p} \tag{7}$$

for constants A and B. Thus, given several estimated values of the variance and bias, one can estimate $\hat{A}$ and $\hat{B}$ for A and B such as by least squares method and then solve (5) to find the adaptive bandwidth choice as

$$\hat{h_t} = \left\{ \frac{\hat{A}}{2pn\hat{B}} \right\}^{1/(2p+1)} \tag{8}$$

In other words, with a grid of h values without considering a boundary problem, both variance and bias^2 can be estimated. Then, given several estimated values of the variance and bias^2 fitting the relations (6) and (7), estimates $\hat{A}$ and $\hat{B}$ can be obtained.

3. Variance and Bias Estimators

3.1 Variance Estimator

In this section, we give a detailed description of our method for local bandwidths selection. It should be emphasized, however, that this is merely for simplicity and the approach extends directly to more general designs.

In order to implement the algorithm described in section 2, two important components should be considered, i.e., estimators of $\text{var}[\hat{m_h}(t)]$ and $\text{bias}[\hat{m_h}(t)]$. Let us first consider the estimator of variance. From section 2, the exact variance of $\hat{m_h}(t)$ is

$$\text{var}[\hat{m_h}(t)] = \frac{\sigma^2}{n^2h^2} \sum_{i=1}^{n} K^2\left(\frac{t-t_i}{h}\right).$$

Therefore, once a kernel K and a bandwidth h are selected, σ^2 is the only unknown quantity that needs to be estimated. Gasser, Sroka and Jennen-Steinmetz (1986) propose and Staniswalis (1989) uses, for an equispaced design,

$$\begin{aligned} \hat{\sigma^2} &= \frac{1}{6(n-2)} \sum_{i=1}^{n-2} [y_{i+2} - 2y_{i+1} + y_i]^2 \\ &= \frac{1}{6(n-2)} \sum_{i=2}^{n-1} [y_{i-1} - 2y_i + y_{i+1}]^2. \end{aligned} \tag{9}$$

Therefore, $\text{var}[\hat{m_h}(t)]$ can be easily estimated for given value of h by replacing σ^2 by $\hat{\sigma^2}$, namely

$$\text{var}(h) = \frac{\hat{\sigma^2}}{n^2h^2} \sum_{i=1}^{n} K^2\left(\frac{t-t_i}{h}\right).$$

3.2 Bias Estimator

In order to estimate $\text{bias}[\hat{m_h}(t)]$ directly from (2), one needs to estimate the p^{th} derivative of $m(t)$. Moreover, one wants to estimate

$$\frac{(-1)^p}{p!} h^p m^{(p)}(t) k_p,$$

where k_p is defined in (1).

Suppose that $K_p(z)$ is a kernel of order p. Hence as we have seen in section 2, by applying the standard Taylor's expansion and taking a few of leading terms, it can be shown that for a fixed h and t

$$E[\widehat{m_p}(t)] = m(t) + (-1)^p \frac{h^p m^{(p)}(t)}{p!} k_p + (-1)^{p+2} \frac{h^{p+2} m^{(p+2)}(t)}{(p+2)!} k_{p+2} + o(h^{p+2}), \quad (10)$$

where now we denote the corresponding kernel estimator by

$$\widehat{m_p}(t) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{h} K_p\left(\frac{t-t_i}{h}\right) y_i.$$

Let $K_{p+2}(z)$ be a kernel of order $p+2$. Then, it follows that

$$E[\widehat{m}_{p+2}(t)] = m(t) + (-1)^{p+2} \frac{h^{p+2} m^{(p+2)}(t)}{(p+2)!} k_{p+2} + o(h^{p+2}), \quad (11)$$

where $\widehat{m}_{p+2}(t) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{h} K_{p+2}\left(\frac{t-t_i}{h}\right) y_i$ and $k_{p+2} = \int_{-1}^{+1} z^{p+2} K_{p+2}(z) dz \neq 0$.

Now, let $K_b(z)$ be a new kernel such that $K_b(z) = K_p(z) - K_{p+2}(z)$ and $\widehat{\text{bias}}_a(h)$ a kernel estimator on this difference defined by

$$\widehat{\text{bias}}_a(h) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{h} K_b\left(\frac{t-t_i}{h}\right) y_i.$$

Then, $E[\widehat{\text{bias}}_a(h)]$ is obtained by taking the difference of two asymptotic expressions (10) and (11) and this difference gives

$$(-1)^p h^p m^{(p)}(t) overp! k_p + o(h^p)$$

by ignoring higher order terms. Hence, the leading term is the quantity which needs to be estimated.

4. Estimation of A and B

Up to this point, estimates of variance and bias have been developed. The next step is to estimate A and B from the expressions of (6) and (7).

Consider an estimator of $m(t)$. Then, for given h, the estimate of the exact variance is

$$\widehat{\text{var}}(h) = \frac{\widehat{\sigma^2}}{n^2 h^2} \sum_{i=1}^{n} K^2\left(\frac{t-t_i}{h}\right), \quad (12)$$

where $\widehat{\sigma^2}$ is defined by (9) and the estimate of bias is given by

$$\widehat{\text{bias}}_a(h) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{h}K_b\left(\frac{t-t_i}{h}\right)y_i. \tag{13}$$

In order to obtain estimators $\hat{A}$ and $\hat{B}$, one needs to evaluate (12) and (13) over a grid of predetermined fixed bandwidths, $h_1, h_2, \cdots, h_k$. Without taking into account any boundary problem at a fixed t, k variance estimates can be obtained from (12) and denoted by

$$(v_1, v_2, \cdots, v_k) = (\widehat{\text{var}}(h_1), \widehat{\text{var}}(h_2), \cdots, \widehat{\text{var}}(h_k)),$$

that is, $v_j = \dfrac{\hat{\sigma}^2}{n^2 h_j^2}\sum_{i=1}^{n}K^2\left(\dfrac{t-t_i}{h_j}\right)$. Also, by squaring the values of (13), we denote these k bias^2 estimates as

$$(b_1^2, b_2^2, \cdots, b_k^2) = (\widehat{\text{bias}}_a^2(h_1), \widehat{\text{bias}}_a^2(h_2), \cdots, \widehat{\text{bias}}_a^2(h_k)),$$

that is,

$$b_j^2 = \left\{\frac{1}{nh_j}\sum_{i=1}^{n}K_b\left(\frac{t-t_i}{h_j}\right)y_i\right\}^2, \qquad j = 1,2,\cdots,k.$$

Now, to derive the least squares estimates, let us consider (6) and (7). Thus, let v_j be modeled by

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

where ζ_j is independent and identically distributed normal random variable with mean 0 and variance σ^2, that is, an usual regression model. Hence, from the above equation, the least squares estimator of A can be found as

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

In a similar manner, let b_j^2 be fit to

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

where ξ_j is independent and identically distributd normal random variable with mean 0 and variance σ^2. Then, the least squares estimator of B is

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

Hence, equation (8) for adaptive bandwidth becomes

$$\hat{h_t} = \left\{ \frac{\hat{A}}{2pn\hat{B}} \right\}^{1/(2p+1)}$$

$$= \left\{ \frac{\sum_{j=1}^{k}(v_j/h_j)(\sum_{j=1}^{k}h_j^{4p})}{2p(\sum_{j=1}^{k}1/h_j^2)(\sum_{j=1}^{k}b_j^2 h_j^{2p})} \right\}^{1/(2p+1)} \quad (14)$$

It is clear that $\hat{h_t}$ balances bias^2 and var of mse and satisfies (5). Also, it is a completely data-driven bandwidth.

5. Numerical comparison between h_t^* and $\hat{h_t}$

In this section, some simulation studies are performed. We try to show how local bandwidth that is defined by equation (14) is performed compared with the true optimal bandwidth. In order to show a few different cases, random samples of size n=100, 200, 400 with σ=0.05, 0.1, 0.3 are used for simulations. The true function used is $m(t)=\sin(4t-1.3)$ and the true optimal bandwidths from (4) are calculated.

To simplify the simulation, p=2 case is considered. The Epanechnikov kernel $K(Z)=\frac{3}{4}(1-Z^2)$, $|Z|\le 1$, is used as a second order kernel, also the kernel $K(Z)=\frac{15}{32}(7Z^4-10Z^2+3)$, $|Z|\le 1$, which is found by Gasser and Müller (1979) and Gasser, Müller and Mammitzsch(1985) is used as the 4th order kernel. Table 1 through Table 4 show the asymptotically true optimal bandwidths from equation (4) and the estimated bandwidths from equation (14). These are obtained from $m(t)=\sin(4t-1.3)$ at t =0.489. Since the grid of bandwidths is one of critical tools for this method, several different grids of bandwidths are used to estimate the bandwidths. The maximum bandwidths are chosen arbitrarily but small enough to avoid a boundary problem. Also, the minimum bandwidth used is 0.06 that is large enough for a sufficient number of observations to be in the window. Then, seven equally divided bandwidths are used to get the estimated bandwidths.

As we can see from Table 1 through Table 4, this algorithm is stable even if σ increases up to 0.3. Also, it can be seen that the algorithm is more stable when the sample size increases. Another advantage of this algorithm is that it is not so sensative of the maximum bandwidth. Moreover, the estimation is not sensative of the maximum bandwidth when the sample size is small.

Table 1
Comparison of Local Optimal Bandwidths With True Optimal Bandwidth (Maximum bandwidth used is 0.32812)

σ	n	h_t^*	$\hat{h}_t$
0.05	100	0.082993	0.089043
	200	0.072249	0.077595
	400	0.062897	0.067592
0.1	100	0.109509	0.117506
	200	0.095334	0.102398
	400	0.082993	0.089195
0.2	100	0.144499	0.155245
	200	0.125793	0.135206
	400	0.109509	0.117741
0.3	100	0.169942	0.183069
	200	0.147943	0.159229
	400	0.128792	0.138576

Table 2
Comparison of Local Optimal Bandwidths With True Optimal Bandwidth (Maximum bandwidth used is 0.30)

σ	n	h_t^*	$\hat{h}_t$
0.05	100	0.082993	0.087996
	200	0.072249	0.076676
	400	0.062897	0.066773
0.1	100	0.109509	0.116147
	200	0.095334	0.101193
	400	0.082993	0.088118
0.2	100	0.144499	0.153583
	200	0.125793	0.133671
	400	0.109509	0.116344
0.3	100	0.169942	0.181380
	200	0.147943	0.157539
	400	0.128792	0.136981

Table 3
Comparison of Local Optimal Bandwidths With True Optimal Bandwidth (Maximum bandwidth used is 0.26)

σ	n	h_t^*	$\hat{h}_t$
0.05	100	0.082993	0.086672
	200	0.072249	0.075514
	400	0.062897	0.065753
0.1	100	0.109509	0.114477
	200	0.095334	0.099696
	400	0.082993	0.086784
0.2	100	0.144499	0.151805
	200	0.125793	0.131878
	400	0.109509	0.114651
0.3	100	0.169942	0.180130
	200	0.147943	0.155792
	400	0.128792	0.135131

Table 4
Comparison of Local Optimal Bandwidths With True Optimal Bandwidth (Maximum bandwidth used is 0.2179)

σ	n	h_t^*	$\hat{h}_t$
0.05	100	0.082993	0.085626
	200	0.072249	0.074481
	400	0.062897	0.064867
0.1	100	0.109509	0.113335
	200	0.095334	0.098440
	400	0.082993	0.085649
0.2	100	0.144499	0.151544
	200	0.125793	0.130779
	400	0.109509	0.113348
0.3	100	0.169942	0.181562
	200	0.147943	0.155634
	400	0.128792	0.133998

6. Conclusions

The proposed algorithm for selecting local bandwidths is practical and simple because it does not require an initial value. Also, similar adaptive approach may be developed for probability density estimation. We expect that the concern about boundary bias would not be so great in this setting even if the boundary problem still need to be considered.

However, the problem is the choice of a grid of bandwidths that are used to find variance and bias estimates. It would be further studied what grid of bandwidths should be used to estimate variance and bias by simulation study.

The main idea behind the proposed algorithm is to have some smoothing parameter prior to attempting to find a bandwidth that minimizes the mean squre error. Thus, the desirability should be obvious for an estimator that does not require a search for the minimum of a noisy curve such as typically encountered in cross validation or other resampling methods.

References

[1] Eubank, R. L. (1988), *Spline Smoothing and Nonparametric Regression,* Marcel Dekker, New York.

[2] Gasser, Th. and Müller, H. G. (1979), Kernel Estimation of Regression Function : In *S moothing Techniques for Curve Estimation,* Springer-Verlag.

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

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

[5] Hall, P. and Schucany, W. R. (1989), A Local Cross-Validation Algorithm, *Statistics and Probability Letters,* Vol. 8, 109-117.

[6] Hardle, W. and Bowman, A. W. (1988), Bootstraping in Nonparametric Regression: Local Adaptive Smoothing and Confidence Bands, *Journal of the American Statistical Association,* Vol. 83, 102-110.

[7] Marron, J. S. (1988), Automatic Smoothing Parameter Selection: *A Survey, Empirical Economics,* Vol. 13, 187-208.

[8] Muller, H. G. (1985), Empirical Bandwidth Choice for Nonparametric Kernel Regression by Means of Pilot Estimators, *Statistics and Decisions,* Vol. 2, 193-206.

[9] Muller, H. G. (1988), *Nonparametric Regression Analysis of Longitudinal Data,* SpringerVerlag.

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

[11] Silverman, B. W. (1986), *Density Estimation for Statistics and Data Analysis,* Chapman and Hall.

[12] Staniswalis, J. G. (1989), Local Bandwidth Selection for a Kernel Estimates, *Journal of the American Statistical Association,* Vol. 84, 284- 288.

비모수적 회귀선의 추정을 위한 bandwidth 선택 알고리즘

차경준[1], 이승우[2]

요약

커널 추정은 커널함수와 bandwidth에 의해서 결정이 된다. 그러나 평활의 정도를 조절하는 적절한 bandwidth를 찾는 것이 더욱 중요한 문제이다. 그러므로 이론적으로 최적의 bandwidth와 비교하여 실제자료에 잘 적용될 수 있는 적절한 bandwidth를 어떻게 찾느냐는 것이 문제가 된다. 본 논문에서는 평균제곱오차(mean square error)의 편의(bias)와 분산(variance)의 관계를 통하여 커널을 이용한 회귀선의 추정에 있어서 간단하고 효과적인 local bandwidth를 찾을 수 있는 알고리즘을 제안하였다.

1) (133-791) 서울특별시 성동구 행당동 17 한양대학교 자연과학대학 수학과
2) (133-791) 서울특별시 성동구 행당동 17 한양대학교 자연과학대학 수학과