Journal of the Korean Statistical Society (2001), **30**: *1, pp 21-30*

A Study on Bandwith Selection Based on ASE for Nonparametric Regression Estimator

Tae Yoon Kim[1]

ABSTRACT

Suppose we observe a set of data $(X_1, Y_1), \cdots, (X_n, Y_n)$ and use the Nadaraya-Watson regression estimator to estimate $m(x) = E(Y|X = x)$. In this article bandwidth selection problem for the Nadaraya-Watson regression estimator is investigated. In particular cross validation method based on average square error (ASE) is considered. Theoretical results here include a central limit theorem that quantifies convergence rates of the bandwidth selector.

Keywords: Bandwidth selection, Kernel regression,ASE

1. Introduction

Let $(X_1, Y_1), \ldots, (X_n, Y_n)$ be independent identically distributed R^2-valued random vectors with Y real valued. Consider the problem of estimating the regression function,

$$m(x) = E(Y|X = x)$$

using $(X_1, Y_1), \ldots, (X_n, Y_n)$. To estimate $m(x)$ kernel estimators introduced by Nadaraya and Watson are considered:

$$\hat{m}_h(x) = \frac{\sum_{i=1}^n K(\frac{x-X_i}{h})Y_i}{\sum_{i=1}^n K(\frac{x-X_i}{h})}$$

where $K : R \to R$ is a kernel function and $h = h(n) \in R^+$ is the bandwidth (i.e, smoothing parameter). One of the crucial points in applying $\hat{m}_h$ is the choice of the bandwidth h. In this paper cross validation rule, an ASE based bandwidth rule, is investigated. The cross validation rule basically attempts to estimate ASE given by

$$d_A(h) = n^{-1} \sum_{j=1}^{n} [\hat{m}_h(X_j) - m(X_j)]^2 w(X_j)$$

[1]Department of Statistics, Keimyung University, Taegu, 704-701, Korea.

and its minimizer $\hat{h}_0$ by finding $\hat{h}$, the minimizer of

$$CV(h) = n^{-1}\sum_{j=1}^{n}[Y_j - \hat{m}_{j,h}(X_j)]^2 w(X_j)$$

where $\hat{m}_{j,h}(X_j)$ is a "leave one out" version of $\hat{m}$; that is, the observation (X_j, Y_j) is left out in constructing $\hat{m}_j$. The weight function w is introduced for elimination of boundary effects.

Härdle, Hall and Marron (1988) established the convergence rate for the cross validation rule for Priestley and Chao regression estimator given by

$$\hat{m}_h(x) = n^{-1}h^{-1}\sum_{i=1}^{n} K(\frac{x - x_i}{h})Y_i.$$

In their setting, it is assumed that $x_1, \ldots, x_n$ are equally spaced design points on unit interval. Since Härdle *et al.* (1988) little work has been done about ASE based bandwidth selectors for kernel regression estimator in various settings, though their result is restricted to a very simple case. Behind this is not that kernel regression estimator has limited use but that analyzing bandwidth selector in more general setting usually faces difficulty from analytical point of view. For example, behaviour of ASE has been relatively unknown compared to integrated square error (ISE) (see Kim (1997) for a related result).

In this paper we extends their result by establishing convergence rates for more general Nadaraya-Watson regression estimator with random design points. Note that asymptotic optimality of the cross validation rule for Nadaraya-Watson regression estimator is verified by Härdle and Marron (1985).

2. Asymptotic results

To handle technical difficulty from the random denominator of the Nadaraya-Watson estimator $\hat{m}_h$ we will consider the following distances (see, e.g. Härdle and Marron (1985));

$$d_A^*(h) = n^{-1}\sum_{j=1}^{n}[\hat{m}_h(X_j) - m(X_j)]^2 \hat{f}_h^2(X_j) f^{-2}(X_j) w(X_j) = n^{-1}\sum_{j=1}^{n}[\hat{m}_h^*(X_j)]^2$$

and $d_M^*(h) = Ed_A^*(h)$. Indeed one may write

$$\hat{m}_h^*(x) = (\hat{m}(x) - m(x))\hat{f}_h(x) f(x)^{-1}$$

where

$$\hat{f}_h(x) = (nh)^{-1} \sum_{i=1}^{n} K(\frac{x - X_i}{h}).$$

Now let $\hat{h}_0^*$ and h_0^* be the minimizer of $d_A^*(h)$ and $d_M^*(h)$ respectively. If m'' is uniformly continuous, then under the assumption that all the moments of ϵ_i exist, $d_A(h)$, $d_A^*(h)$, and $d_M^*(h)$ are approximately

$$d_m^*(h) = n^{-1}h^{-1}\sigma^2 \int f^{-1}w \int K^2 + h^4(\int u^2 K/2)^2 \int (m'')^2 f^{-1}w,$$

in the sense that

$$\sup_{h \in H_n} \left(\left| \frac{d_A(h) - d_m^*(h)}{d_m^*(h)} \right| + \left| \frac{d_A^*(h) - d_m^*(h)}{d_m^*(h)} \right| + \left| \frac{d_M^*(h) - d_m^*(h)}{d_m^*(h)} \right| \right) \to 0 \qquad (2.1)$$

in probability as $n \to \infty$, where $H_n = [n^{-1+\delta}, n^{-\delta}]$, for arbitrary small $\delta > 0$ (see Marron and Härdle (1986)). A consequence of (2.1) is that $\hat{h}_0$ and h_0^* are each roughly equal to the unique minimizer of d_m^*, $h_m^* = c_0 n^{-1/5}$ where

$$c_0 = \left[\sigma^2 \int wf^{-1} \int K^2 / (\int u^2 K)^2 \int (m'')^2 wf^{-1} \right]^{1/5}; \qquad (2.2)$$

that is,

$$\hat{h}_0/h_m^*, \hat{h}_0^*/h_m^*, h_0^*/h_m^* \to 1 \qquad (2.3)$$

in probability. In addition it has been proved by Härdle and Marron (1985)

$$\hat{h}/h_m^* \to 1. \qquad (2.4)$$

Major objective of this article is to study how fast the convergence in (2.3) and (2.4) occurs. Now assumptions for Theorems 1 and 2 are summarized. (a) The errors, ϵ_i are iid with mean 0 and all other moments finite. (b) The kernel function, K, is a symmetric, compactly supported probability density with a Hölder continuous second derivative. (c) The regression function m has a uniformly continuous, integrable second derivative.

Theorem 1. *Under the preceding assumptions*

$$n^{3/10}(\hat{h} - \hat{h}_0) \to N(0, \sigma_1^2) \qquad (2.5)$$

in distribution.

Note that in (2.2), (2.3) and (2.4), all of $\hat{h}$, $\hat{h}_0$, $\hat{h}_0^*$, h_0^* and h_m^* are tending to zero at the rate $n^{-1/5}$ and hence (2.5) says that the relative difference between $\hat{h}$ and $\hat{h}_0^*$ is of the very slow order $n^{-1/10}$. It is also shown in the following theorem that the difference between $\hat{h}_0$ and h_0^* is of the same order.

Theorem 2. *Under the preceding assumptions*

$$n^{3/10}(\hat{h}_0 - h_0^*) \to N(0, \sigma_2^2)$$

in distribution.

Similar results have been established for Priestly and Chao estimator by Hall *et al.* (1988), but with the equally spaced design points on the unit interval (i.e. $x_i = i/n$ on [0,1]). This setting hardly justifies various situations that may occur in reality. Our results extend their result to more general settings. As a result some remarks made by Hall *et al.* still remain valid while other remarks need to be be changed. For example, their remark that the bandwidth selector suffers the excruciatingly slow speed still holds as long as $\hat{h}_0$ is sought but the remark that extension to the multivariate X_i is statriaghtforward seems to be inappropriate. In fact it is shown by Kim (1997) that random denominator in Nadaraya-Watson may cause some problem to quadratic errors.

3. Proofs

Most of steps taken in the proof below are those taken by Härdle *et al.* (1988) but adjustments are to be made to handle random design points. Now we define the following quantities for later use.

$$\bar{d}_A(h) = n^{-1} \sum_{j=1}^{n} [\hat{m}_{h,j}(X_j) - m(X_j)]^2 w(X_j)$$

$$\bar{d}_A^*(h) = n^{-1} \sum_{j=1}^{n} [\hat{m}_{h,j}(X_j) - m(X_j)]^2 \hat{f}_{h,j}(X_j)^2 f(X_j)^{-2} w(X_j).$$

The proof of Theorem 2 is based on the expansion

$$\begin{aligned} 0 = d_A'(\hat{h}_0) &= d_M^{*\prime}(\hat{h}_0) + d_A'(\hat{h}_0) - d_A^{*\prime}(\hat{h}_0) + d_A^{*\prime}(\hat{h}_0) - d_M^{*\prime}(\hat{h}_0) \\ &= (\hat{h}_0 - h_0^*) d_M^{*\prime\prime}(h_1) + d_A'(\hat{h}_0) - d_A^{*\prime}(\hat{h}_0) + D'(\hat{h}_0), \end{aligned} \tag{3.1}$$

where h_1 is between $\hat{h}_0$ and h_0^*, and where

$$D(h) = d_A^*(h) - d_M^*(h)$$

and where D', d_A', $d_A^{*\prime}$, and $d_M^{*\prime}$ denote the derivatives with respect to h of D d_A, d_A^* and d_M respectively.

For the proof of Theorem 1, note that

$$CV(h) = \bar{d}_A(h) + \delta(h) - n^{-1}\sum_{j=1}^{n}[m(X_j) - Y_j]^2 w(X_j)$$

where

$$\delta(h) = 2n^{-1}\sum_{j=1}^{n}(\hat{m}_{h,j}(X_j) - m(X_j))(m(X_j) - Y_j)w(X_j).$$

Now write

$$CV(h) = \bar{d}_A(h) - \bar{d}_A^*(h) + \bar{d}_A^*(h) - d_M^*(h) + d_M^*(h)$$

$$+\delta(h) - \delta^*(h) + \delta^*(h) - n^{-1}\sum_{j=1}^{n}[m(X_j) - Y_j]^2 w(X_j)$$

where

$$\delta^*(h) = 2n^{-1}\sum_{j=1}^{n}(\hat{m}_{h,j}(X_j) - m(X_j))(m(X_j) - Y_j)\hat{f}_{h,j}(X_j)f(X_j)^{-1}w(X_j).$$

Then the proof of Theorem 1 uses the following expansion

$$0 = CV'(\hat{h}) = \bar{d}_A'(\hat{h}) - \bar{d}_A^{*\prime}(\hat{h}) + \bar{d}_A^{*\prime}(\hat{h}) - d_M^{*\prime}(\hat{h}) + d_M^{*\prime}(\hat{h}) + \delta'(\hat{h}) - \delta^{*\prime}(\hat{h}) + \delta^{*\prime}(\hat{h}). \tag{3.2}$$

To analyze expressions (3.1) and (3.2), we need the following lemmas. Notation used there includes

$$r_n(h) = n^{-1}h^{-1} + h^4.$$

Lemma 1. *For $l = 1, 2, \ldots$, there is a constant c_4, so that*

$$\sup_{h\in H_n} E|r_n(h)^{-1}h^{1/2}D'(h)|^{2l} \le c_4 \tag{3.3}$$

and

$$\sup_{h\in H_n} E|r_n(h)^{-1}h^{1/2}\delta_2'(h)|^{2l} \le c_4. \tag{3.4}$$

Furthermore, there is an $\eta_1 > 0$ *and a constant* c_5 *so that*

$$E|r_n(h)^{-1}h^{1/2}[D'(h) - D'(h')]|^{2l} \leq c_5(h^{-1}|h-h'|)^{\eta_1 l} \tag{3.5}$$

$$E|r_n(h)^{-1}h^{1/2}[\delta^{*\prime}(h) - \delta^{*\prime}(h')]|^{2l} \leq c_5(h^{-1}|h-h'|)^{\eta_1 l} \tag{3.6}$$

whenever $h, h' \in H_n$, *with* $h \leq h'$ *and* $|h^{-1}(h-h')| \leq 1$.

Proof: An application of Lemma 2 of Kim and Cox (1997) yields the desired result. In fact Lemma 2 is established for dependent variables, which can be easily adapted to the iid case. □

Lemma 2. *For any* $\eta_2 \in (0, 1/10)$,

$$\sup_{h\in H_n} \{r_n(h)^{-1}h^{1/2}[|D'(h)| + |\delta^{*\prime}(h)|]\} = O_p(n^{\eta_2}). \tag{3.7}$$

Furthermore, if $h_1 n^{1/5}$ *tends to a constant, then*

$$\sup_{|h-h_1|\leq n^{-1/5-\eta_2}} r_n(h)^{-1}h^{1/2}[|D'(h) - D'(h_1)| + |\delta^{*\prime}(h) - \delta^{*\prime}(h_1)|] = o_p(1). \tag{3.8}$$

Proof: Basically (3.7) and (3.8) follows from Lemma 1 above. See the proof of Lemma 2 of Härdle *et al.* (1988) for its detailed verification. □

Lemma 3. *For any* $0 < an^{-1/5} < h < bn^{-1/5} < \infty$

$$\sup_h |d'_A(h) - d^{*\prime}_A(h)| = o_p(n^{-7/10})$$

Proof: Note that

$$d_A(h) - d^*_A(h) = n^{-1}\sum_i \Big[-2\hat{m}^{*2}_h(X_i)(1 - f(X_i)/\hat{f}_h(X_i))w(X_i)$$

$$+\hat{m}^{*2}_h(X_i)(1 - f(X_i)/\hat{f}_h(X_i))^2 w(X_i)\Big] = S_1(h) + S_2(h).$$

where

$$S_1(h) = -2n^{-1}\sum_i \hat{m}^{*2}_h(X_i)(1 - f(X_i)/\hat{f}_h(X_i))w(X_i)$$

$$S_2(h) = n^{-1}\sum_i \hat{m}^{*2}_h(X_i)(1 - f(X_i)/\hat{f}_h(X_i))^2 w(X_i).$$

Consider $S_1(h)$ first. It is easy to see that

$$dS_1(h)/dh = -2n^{-1}\sum_i \Big\{[\hat{m}^{*2}_h(X_i)w(X_i) - d^*_M(h) + d^*_M(h)]'[1 - f(X_i)/\hat{f}_h(X_i)]$$

$$+ [\hat{m}_h^{*2}(X_i)w(X_i) - d_M^*(h) + d_M^* h][1 - f(X_i)/\hat{f}_h(X_i)]'\Big\}.$$

Then the above expression is less than

$$2\sup_x |(1 - f(x)/\hat{f}_h(x))w(x)||D(h)' + d_M^*(h)'|$$

$$+ \sup |(1 - f(x)/\hat{f}_h(x))w(x))'||D(h) + d_M^*(h)|. \tag{3.9}$$

Now we have that the first term of (3.9) is, for $0 < \eta_2 < 1/10$,

$$O_p(n^{-7/10+\eta_2-2/5} + n^{-3/5-2/5}) = o_p(n^{-7/10}), \tag{3.10}$$

and the second term is

$$O_p(n^{-1/5-9/10} + n^{-1/5-4/5}) = o_p(n^{-7/10}). \tag{3.11}$$

In (3.10) we used the uniform strong consistency of $\hat{f}_h$ to f on the compact set of x i.e., $\sup_{x\in C} |\hat{f}_h(x) - f(x)| = O_p(n^{-2/5})$ and (3.7). To verify (3.11), it is easy to check

$$|w(x)[1 - f(x)/\hat{f}_h(x)]'| = |f(x)w(x)[\hat{f}_h(x)]'\hat{f}_h^{-2}(x)| \le c|[\hat{f}_h(x)]'|$$

$$= |-h^{-1}(nh)^{-1}\sum_i K(\frac{x - X_i}{h}) + h^{-1}(nh)^{-1}\sum_i L(\frac{x - X_i}{h})| = O_p(n^{-1/5})$$

where $L(u) = -uK'(u)$ which satisfies the usual conditions for the kernel function K. Further it can be found that if $h \sim cn^{-1/5}$ for some constant $c > 0$ then $d_M^*(h) = O(n^{-4/5})$ and $|d_A^*(h) - d_M^*(h)| = O_p(n^{-9/10})$ (see e.g. Hall (1984)). Similar argument can be applied to $S_2(h)$. □

Lemma 4. *For some* $\epsilon > 0$, $|\hat{h}_0 - h_0^*| + |\hat{h} - h_0^*| = O_p(n^{-1/5-\epsilon})$.

Proof: First remember that $\hat{h}_0^*$ is the minimizer of d_A^*. Now by (2.3) and (3.7)

$$d_A^{*\prime}(\hat{h}_0) = d_A^{*\prime}(\hat{h}_0) - d_A^{*\prime}(\hat{h}_0^*) = d_M^{*\prime}(\hat{h}_0) - d_M^{*\prime}(\hat{h}_0^*) + O_p(n^{-7/10+\eta_2}).$$

But by Lemma 3

$$d_A^{*\prime}(\hat{h}_0) = d_A^{*\prime}(\hat{h}_0) - d_A'(\hat{h}_0) = o_p(n^{-7/10}).$$

Thus

$$d_M^{*\prime}(\hat{h}_0) - d_M^{*\prime}(\hat{h}_0^*) + O_p(n^{-7/10+\eta_2}) = o_p(n^{-7/10}).$$

Then $d_M^{*\prime}(\hat{h}_0) - d_M^{*\prime}(\hat{h}_0^*) = (\hat{h}_0 - \hat{h}_0^*)d_M''(h_1)$ where h_1 is between $\hat{h}_0$ and $\hat{h}_0^*$. So letting $\epsilon = -\eta_2 + 1/10$, $|\hat{h}_0 - \hat{h}_0^*| = O_p(n^{-1/5-\epsilon})$ holds by (3.12) below. □

Lemma 5. *For any* $0 < an^{-1/5} < h < bn^{-1/5} < \infty$

$$\sup_h |\delta^{*\prime}(h) - \delta'(h)| + |\bar{d}'_A(h) - \bar{d}^{*\prime}_A(h)| = o_p(n^{-7/10}).$$

Proof is done in a very similar fashion as in the proof of Lemma 3. □

Lemma 6. *For any* $0 < an^{-1/5} < h < bn^{-1/5} < \infty$,

$$\sup_h |\bar{d}^{*\prime}_A(h) - d^{*\prime}_M(h)| = D'(h) + o_p(n^{-7/10}).$$

Proof: Note that

$$\bar{d}^{*\prime}_A(h) - d^{*\prime}_M(h) = \bar{d}^{*\prime}_A(h) - d^{*\prime}_A(h) + D'(h).$$

By (2.3), it suffices to show that

$$|\bar{d}^{*\prime}_A(h) - d^{*\prime}_A(h)| = o_p(n^{-7/10})$$

which could be verified as in the proof of Lemma 3 of Härdle and Marron (1985).

Lemma 7.

$$n^{7/10} \begin{bmatrix} D'(h_0^*) \\ \delta^{*\prime}(h_0^*) \end{bmatrix} \to N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_3^2 & \sigma_{34} \\ \sigma_{34} & \sigma_4^2 \end{bmatrix} \right)$$

in distribution, where (letting $*$ *denote convolution)*

$$\sigma_3^2 = 8c_0^{-3}\sigma^4[\int w^2 \int (K*K - K*L)^2] + 4c_0^2\sigma^2[\int u^2K]^2[\int (m'')^2 f^{-4} w^2].$$

$$\sigma_4^2 = 8c_0^{-3}\sigma^4[\int w^2 \int (K - L)^2] + 4c_0^2\sigma^2[\int u^2K]^2[\int (m'')^2 f^{-4} w^2].$$

and

$$\sigma_{34} = -8c_o^{-3}\sigma^4[\int w^2 \int (K*K - K*L)(K - L)] - 4c_0^2\sigma^2[\int u^2K]^2[\int (m'')^2 f^{-4} w^2].$$

Proof: This proof is almost identical to the proof of Lemma 4 of Härdle, *et al.* (1988). Details are omitted.

To finish the proof of Theorem 2, note first that

$$n^{2/5} d''_M(h_1) \to c_3 \tag{3.12}$$

where $c_3 = (2/c_0^3)\sigma^2[\int K^2][\int wf^{-2} + 3c_0^2[\int u^2K]^2[\int (m'')^2 w/f^{-2}]$. It follows from Lemma 4 and (3.8) that $D'(\hat{h}_0) = D'(h_0^*) + o_p(n^{-7/10})$. Hence by Lemma 7, $n^{7/10}D'(\hat{h}_0) \to N(0, \sigma_3^2)$. Applying Lemmas 2-4 and (3.12) to (3.1), we have

$$\begin{aligned} 0 &= d_M^{*\prime}(\hat{h}_0) + D'(h_0^*) + o_p(n^{-7/10}) \\ &= (\hat{h}_0 - h_0^*)c_3 n^{-2/5} + D'(h_0^*) + o_p(n^{-7/10}), \end{aligned} \tag{3.13}$$

from which it follows that $n^{3/10}(\hat{h}_0 - h_0^*) \to N(0, \sigma_2^2)$ where $\sigma_2^2 = \sigma_3^2/c_3^2$.

The proof of Theorem 1 takes slightly more work than the proof of Theorem 2. For $h \in [an^{-1/5}, bn^{-1/5}]$ (where a and b are arbitrary constants), (3.2) and Lemmas 2, 4, 5 and 6 give

$$0 = d_M^{*\prime}(h_0^*) + \delta^{*\prime}(h_0^*) + D'(h_0^*) + o_p(n^{-7/10}). \tag{3.14}$$

Working on (3.14) as in (3.1) and (3.13) gives

$$0 = (\hat{h} - h_0^*)c_3 n^{-2/5} + \delta^{*\prime}(h_0^*) + D'(h_0^*) + o_p(n^{-7/10})$$

which after subtracting (3.13) yields

$$-\delta^{*\prime}(h_0^*) = (\hat{h} - \hat{h}_0)c_3 n^{-2/5} + o_p(n^{-7/10}).$$

Hence by Lemma 7, $n^{3/10}(\hat{h} - \hat{h}_0) \to N(0, \sigma_1^2)$, where $\sigma_1^2 = \sigma_4^2/c_3^2$. □

ACKNOWLEDGEMENTS

The author dedicates this work to the Professor JONGBIN KIM for his 30-year service at Yonsei University. This paper was accomplished with research fund provided by Korea Research Foundation, support for 1997 faculty research abroad.

REFERENCES

Marron, J. S.and Härdle W. (1986) Random approximations to some measures of accuracy in nonparametric curve estimation. *Journal of Multivariate Analysis* **20** 91-113.

Härdle, W., Hall, P. and Marron, J.S. (1988) How far are automatically chosen regression smoothing parameters from their optimum ? *Journal of American Statistical Society* **83** 86-98.

Härdle, W. and Marron, J.S. (1985) Optimal bandwidth selection in nonparametric regression function estimation. *The Annals of Statistics* **13** 1465-1481.

Kim, T.Y and D. Cox (1997) A study on bandwidth selection in density estimation under dependence. *Journal of Multivariate Analysis* **62** 190-203

Kim, T. Y. (1997) Central limit theorem for quadratic errors of nonparametric estimators *Journal of Statistical Planning and Inference* **64** 193-204.