⟨Research Paper⟩

Journal of the Korean
Statistical Society
Vol. 17, No. 1, 1988

L^1 Bandwidth Selection in Kernel Regression Function Estimation[+]

Myoungshic Jhun*

ABSTRACT

Kernel estimates of an unknown regression function are studied. Bandwidth selection rule minimizing integrated absolute error loss function is considered. Under some reasonable assumptions, it is shown that the optimal bandwidth is unique and can be computed by using bisection algorithm. Adaptive bandwidth selection rule is proposed.

1. Introduction

Let (X_i,Y_i), $1 \leqq i \leqq n$, be a random sample of size n from a bivariate distribution with a common joint density $f(x,y)$. The conditional mean or regression of Y on X is

$$\begin{aligned} r(x) &= E(Y \mid X=x) \\ &= \int y f(x,y) / f_x(x) dy \end{aligned} \tag{1.1}$$

where $f_X(x)$ is a marginal density of X. Nadaraya(1964) and Watson(1964) independently proposed nonparametric estimators of $r(x)$ based on the kernel method as introduced by Rosenblatt(1956) for density estimation. In specific the estimates have the form

*Department of Statistics, University of Michigan Ann Arbor, MI 48105 U.S.A.

[+] Research was supported in part by the National Science Foundation of U.S.A., under DMS-8702980.

$$r_n(x:h)=(nh)^{-1}\sum_{j=1}^{n} Y_j K((x-X_j)/h)/(nh)^{-1}\sum_{j=1}^{n} K((x-X_j)/h), \qquad (1.2)$$

where $K(\cdot)$ is a kernel function and h is a bandwidth such that $h \to 0$ as $n \to \infty$. As in kernel type density estimation (cf. Rosenblatt;1971) there is very little to choose between the various kernels. However, the choice of bandwidth is very crucial since a small h gives an estimator with a large variance, but a large h yields a large bias for the estimator.

Since the variance of $r_n(x;h)$ is inversely proportional to $f_x(x)$, integrated weighted squared error loss function

$$IWSE(h) = \int (r_n(x:h) - r(x))^2 f_x^2(x)dx$$

has been considered as a reasonable measure of the performance of the estimator. By noticing the expansion of

$$E\{IWSE(h)\} = A(nh)^{-1} + Bh^4 + o((nh)^{-1} + h^4)$$

where A and B are constants depending on underlying density f, Hall(1984) considered the optimization of h in a range $[c_1 n^{-1/5}, c_2 n^{-1/5}]$ for $c_2 > c_1 > 0$. Rice(1984) obtained asymptotic optimality in the same range in the fixed design setting. Hardle and Kelly(1987) considered cross-validatory choice for modified $r_n(x;h)$, which has nonrandom denominator $f(x)$, and obtained asymptotic optimal bandwidth for the IWSE criterion.

In this paper, we consider integrated absolute error loss

$$IAE(h) = \int |r_n(x:h) - r(x)| dx \qquad (1.3)$$

as a criterion for the selection of the bandwidth h. Even though it is an interesting criterion for the choice of the bandwidth h, mainly because of the mathematical tractability, theory for this criterion is quite slow to develop. The main result of this paper is that the optimal h minimizing the limit of normalized IAE(h) is unique and can be obtained by using so called bisection algorithm. The theory in this paper can be applied for integrated weighted absolute error

$$IWAE(h) = \int | r_n(x:h) - r(x) | f_x(x)dx,$$

with slight modification. Adaptive choice of the bandwidth is also proposed.

2. Some Preliminaries

Assumptions :

(A.1) $f(x,y) \leq C_1(1+x^2+y^2)^{-5/2}$ for some constant C_1.

(A.2) $r(x)$ and $f_x(x)$ are continuously differentiable up to second order.

(A.3) $K(x)$ is a symmetric density and $K(x) \leq C_2(1+x^2)^{-3}$ for some constant C_2.

Let *F* be a class of density functions satisfying (A.1) and (A.2), and

$$\alpha(x) = \{[r(.)f_x(.)]^{(2)}(x) - 1\}, \qquad (2.1)$$

$$\beta(x) = E[(Y-1)^2 \mid X = x] f_x(x), \qquad (2.2)$$

$$K_2 = \int_R x^2 K(x)\,dx,$$

$$\| K \| = \{\int_R K^2(x)\,dx\}^{1/2}.$$

For fixed h and x, under assumptions (A.1), (A.2), and (A.3), from Rosenblatt(1969), we have

$$r_n(x:h) - r(x) = (1/2)h^2 K_2 f_x(x)^{-1}\alpha(x) + f(x)^{-1} W_n + o_p(h^2) + O_p(n^{-1}h^{-1} + n^{-1/2}h^{3/2}) \qquad (2.3)$$

where W_n is asymptotically normal with mean zero and variance

$$(nh)^{-1}\|K\|^2\beta(x) \text{ as } n\rightarrow\infty \text{ provided } nh\rightarrow\infty \text{ as } n\rightarrow\infty$$

The rate at which the difference $\{r_n(x) - r(x)\}$ in (2.3) tends to zero as $n\rightarrow\infty$ is maximized to $O_p(n^{-2/5})$ if we set

$$h = c(f)n^{-1/5}$$

where c(f) is a positive constant depending on the underlying density function f(x,y). We aim at finding optimal c(f) for the integrated absolute error loss function(1.3).

For fixed $c > 0$ and $x \in R$, and $n \geq 1$, let

$$Zn(x,c) = n^{2/5}[r_n(x:cn^{-1/5}) - r(x)].$$

Proposition 1. Under the assumptions (A.1), (A.2), and (A.3), $Z_n(x,c)$ weakly converges to a normal random variable with mean

$$\mu(f:x,c)=(1/2)\,c^2K_2 f_x(x)^{-1}\alpha(x) \tag{2.4}$$

and variance

$$\sigma^2(f:x,c)=(1/c)\,\|K\|^2\,\beta(x) \tag{2.5}$$

Moreover, $|Z_n(x,c)|$, $n\geq1$, are uniformly integrable, for a. e. x.

Proof. The asymptotic normality of $Z_n(x,c)$ immediately follows from (2.3), and uniform integrability follows from the fact that

$$\int_{|z_n|>\alpha} |Z_n(x,c)|\,dP \leq (1/\alpha)\,E[Z_n(x,c)^2] < \infty. \blacksquare$$

3. Results

For convenience, we define following notations. Let Φ and ϕ denote the standard normal distribution function and density respectively. Let Z be a standard normal random variable, and $\Psi(y)=E|Z-y|$.

Then

$$\psi(y)=[2\Phi(y)-1]y+2\phi(y).$$

Note that

$$|\psi(y)-|y||\leq1$$
$$\psi'(y)=2\Phi(y)-1, \qquad \text{and}$$
$$\psi''(y)=2\phi(y)>0.$$

So, ψ is convex and its first two derivatives are bounded.

We consider the absolute error over an interval where the marginal density function $f_x(x)$ is bounded away from zero.

Suppose that

$$\min_{a\leq x\leq b} f_x(x)=\rho>0.$$

With $h=cn^{-1/5}$, where $c>0$ is fixed, let

$$J_n(c) = n^{2/5}\int_a^b | r_n(x, cn^{-1/5}) - r(x) | \, dx$$

and

$$\mathcal{U}_n(c) = E[J_n(c)].$$

Proposition 2. Under the assumptions (A.1), (A.2), and (A.3),

$$\lim_{n\to\infty} \mathcal{U}_n(c) = \int_a^b \sigma(f:x,c) \; \phi\left[\frac{\mu(f:x,c)}{\sigma(f:x,c)}\right] dx.$$

Proof. If Z denotes a standard normal random variable, then by Proposition 1,

$$Z_n(x,c) \overset{d}{\Rightarrow} \sigma(f:x,c)\{ Z + [\frac{\mu(f:x,c)}{\sigma(f:x,c)}]\}$$

and

$$\lim_{n\to\infty} E|Z_n(x,c)| = \sigma(f:x,c)\phi[\frac{\mu(f:x,c)}{\sigma(f:x,c)}]$$

as n→∞ for a.e. x. So, by using bounded convergence theorem, the result can be obtained easily. ■

Now, we will find the optimal c(f) which minimizes limit of the normalized IAE,

$$\lim_{n\to\infty} n^{2/5}\int_a^b | r_n(x, cn^{-1/5}) - r(x) | dx.$$

Let

$$\begin{aligned} H(f,c) &= \lim_{n\to\infty} \mathcal{U}_n(x,c) \\ &= \int_a^b \sigma(f:x,c)\phi[\frac{\mu(f:x,c)}{\sigma(f:x,c)}]dx \end{aligned}$$

Then, we have the following theorem, which minimizes the limit H(f,c).

Theorem. (minimizing the limit) For each $f \in F$, H(f,c) attains its minimum at a unique point c=c(f). Moreover, c(f) is the unique solution to the equation Δ(f,c)=0, where Δ is strictly increasing function, defined below in (2.6).

Proof. Now,

$$H(f,c)=\int_a^b \sigma(f:x,c)\,\psi\left[\frac{\mu(f:x,c)}{\sigma(f:x,c)}\right]dx$$

where $\mu(f{:}x,c)$ and $\sigma^2(f{:}x,c)$ are as in (2.4) and (2.5).

Since H(f,c)→∞ as c→0 or c→∞ and is continuous in c, it suffices to show that $(\partial/\partial c)$ H(f,c) vanishes only once. But,

$$(\partial/\partial c)H(f,c)=(-1/2)c^{-3/2}\|K\|\int_a^b \beta(x)^{1/2}\,\psi\left[\frac{\mu(f:x,c)}{\sigma(f:x,c)}\right]dx$$
$$+(5/4)cK_2\int_a^b \psi'\left[\frac{\mu(f:x,c)}{\sigma(f:x,c)}\right]f_x(x)^{-1}\,\alpha(x)dx$$

for all x and c.

Thus, we have

$$(\partial/\partial c)H(f,c)=c^{-2/3}\varDelta(f,c)$$

where

$$\varDelta(f,c)=(-1/2)\|K\|\int_a^b \beta(x)^{1/2}\,\psi\left[\frac{\mu(f:x,c)}{\sigma(f:x,c)}\right]dx \qquad (2.6)$$
$$+(5/4)c^{5/2}K_2\int_a^b \psi'\left[\frac{\mu(f:x,c)}{\sigma(f:x,c)}\right]f_x(x)^{-1}\,\alpha(x)dx$$

But,

$$\varDelta'(f,c)=(5/4)c^{3/2}K_2\int_a^b \psi'\left[\frac{\mu(f:x,c)}{\sigma(f:x,c)}\right]f_x(x)^{-1}\,\alpha(x)dx$$
$$+(5/4)c^4(K_2^2/\|K\|)\int_a^b \psi''\left[\frac{\mu(f:x,c)}{\sigma(f:x,c)}\right]f_x(x)^{-2}\,\alpha(x)\beta(x)^{-1/2}\,dx>0\,.$$

So, Δ(f,c) is an increasing function of c for each fixed f. Also Δ(f,c) converges to Δ(f, 0) 〈 0 as c tends to 0 and H(f,c) tends to infinity as c tends to infinity.

Therefore, for every f ∈ F, Δ(f,c) vanishes for exactly one value c=c(f) and, therefore, the same is true of H′(f,c). ■

Remark 1. By observing the facts $|\psi'(y)|\leq 1$ and $\psi(y)\geq\psi(0)$ for all y, we can find $c^-(f)$ and $c^+(f)$ such that

$$\varDelta[f,c^-(f)]\leq 0. \qquad (2.7)$$

and

$$\Delta[f, c^{+}(f)]\geq 0.$$

Thus, the minimizing value c(f) can be obtained numerically. Two inequalities (2.7) and (2.8) provide the starting point for an algorithm in which the function $\Delta[f,.]$ is evaluated at the midpoints of sequence of smaller and smaller intervals.

Remark 2. The procedures dicussed here can be used in the case of a regression analysis with a fixed (or a random) independent variable, i.e. $Y=r(x)+Z$ (or $Y=r(X)+Z$) with Z independent of X and mean zero. In fact, the class F is quite large enough to include many interesting density functions.

Remark 3. By giving a suitable metric ρ on the class F of density functions, which can handle both $\mu(f;x,c)$ and $\sigma^2(f;x,c)$ in (2.4) and (2.5), continuity of c(f) with respect to $f \in F$ can be considered. Now, we may have an initial estimate $f_{n,i}(x,y)$ of f(x,y) such that $\rho(f_{n,i},f)\rightarrow 0$ as $n\rightarrow\infty$ and then choose an adaptive bandwidth

$$h_a = c(f_{n,i})\, n^{-1/5}.$$

The idea is that the adaptive bandwidth h_a is not very sensitive to the initial estimate $f_{n,i}(x,y)$.

Acknowledgement

Thanks to referees for useful comments.

References

(1) Hall, P.(1984).Asymptotic Properties of Integrated Square Error and Cross-Validation for Kernel Estimation of a Regression Fuction,*Zeitschrift fur Wahrscheinlichkeits theorie,* 67, 175 ~196.

(2) Hardle, W. and Kelly, G.(1987).Nonparametric Kernel Regression Estimation-Optimal Choice of Bandwidth,*Statistics,* 18, 1, 21~35.

(3) Nadaraya, E.A.(1964).On Estimating Regression,*Theory of Probability and its Applications,* 9, 141~142.

(4) Rice, J.(1984).Bandwidth Choice for Nonparametric Regression,*The Annals of Statistics,* 12, 1215~1230.

(5) Rosenblatt, M.(1956). Remarks on some nonparametric estimators of a density function,*The Annals of Mathematical Statistics,* 27, 832~837.

(6) Rosenblatt, M.(1969). Conditional Probability Density and Regression Estimators, *Multivariate Analysis* Ⅱ, (ed. Krishnaiah) 25~31.

(7) Rosenblatt, M.(1971). Curve Estimates, *The Annals of Statistics* 1815~1842.

(8) Watson, G.S.(1964). Smooth Regression Analysis, *Sankhya Series* A, 26, 359~372.