Journal of the Korean
Statistical Society
Vol. 22, No. 2, 1993

Nonparametric Kernel Regression Function Estimation with Bootstrap Method †

Daehak Kim[1]

ABSTRACT

In recent years, kernel type estimates are abundant. In this paper, we propose a bandwidth selection method for kernel regression of fixed design based on bootstrap procedure. Mathematical properties of proposed bootstrap-based bandwidth selection method are discussed. Performance of the proposed method for small sample case is compared with that of cross-validation method via a simulation study.

KEYWORDS : Kernel regression, bandwidth, fixed design, bootstrap, consistency.

1. INTRODUCTION

Let $Y_1, Y_2, \cdots, Y_n$ be observations on the unknown regression function $\theta(\cdot)$ with a model

$$Y_i = \theta(x_i) + \epsilon_i, \quad i = 1, \cdots, n \tag{1.1}$$

where ϵ_i are iid from unknown F with mean 0 and finite variance σ^2 and x_i are fixed design points. Without loss of generality, We assume the regression function θ is defined on the closed interval $[0, 1]$ and x_i are equally spaced. Among the

[1] Department of Statistics, Pusan University of Foreign Studies, Nam, Pusan 608-738, Korea
† This paper was supported in part by NON DIRECTED RESEARCH FUND, Korea Research Foundation, 1991

nonparametric estimates of regression function, we consider the kernel estimate of regression function of the form

$$\hat{\theta}_n(x : h) = \sum_{i=1}^{n} \frac{K((x - x_i)/h)}{\sum_{j=1}^{n} K((x - x_j)/h)} Y_i \tag{1.2}$$

where kernel $K(\cdot)$ is symmetric probability density function and bandwidth h represents over all amount of smoothing which depends on n and tends to 0 as $n \to \infty$ but $nh \to \infty$. This kernel estimator was independently proposed by Nadaraya (1964) and Watson (1964).

Correct choice of bandwidth in kernel estimation is a crucial problem. (See Silverman (1985)) Cross-validation method for selecting the bandwidth is known as one of the most widely used data dependent approach.

In this paper, we propose another bandwidth selection method based on bootstrap method for $\hat{\theta}_n(x : h)$. We will prove the consistency of bootstrap method in kernel regression. Then, we estimate the quantity,

$$E \int |\hat{\theta}_n(x : h) - \theta(x)|^p dx \tag{1.3}$$

for p=1,2 using bootstrap method and find data dependent bandwidth h which minimise the bootstrap estimate of (1.3). To study and compare the performance of proposed bandwidth selection method with that of cross-validation method, a Monte Carlo simulation study is conducted.

2. CONSISTENCY OF BOOTSTRAP METHOD

In this section, asymptotics of the kernel regression estimate is considered. Then, consistency of bootstrap method is shown. Throughout this paper, we assume the following reasonably mild conditions.

[Assumptions]

A.1 The bandwidth $h \to 0$ and $nh \to \infty$ as $n \to \infty$.

A.2 The kernel K is symmetric with finite support $[-A, A]$

A.3 The regression function $\theta(x)$ is continuous and twice differentiable.

2.1 Asymptotic behaviours

Asymptotic optimal bandwidth for (1.2) is $h = cn^{-1/5}$ in $IMSE$ sense where c is a positive constant depending on θ. At this point, we will consider the normalized process

$$Z_n(x:c) = n^{2/5}(\hat{\theta}_n(x : cn^{-1/5}) - \theta(x)) \tag{2.1}$$

Then we have the following proposition

Proposition 1. For fixed $c > 0$ and $x \in [0,1]$ and $n \geq 1$ and with A1,A2, and A3, $Z_n(x:c) \xrightarrow{d} N(c^2\mu(x),\ \sigma^2 \int_{-A}^{A} K^2(t)dt/c)$ as $n \to \infty$, a.e. x where $\mu(x) = \frac{1}{2}\theta''(x)\int_{-A}^{A} t^2K(t)dt$.

Proof. The asymptotic normality of $Z_n(x : c)$ were discussed by Rosenblatt(1969).

2.2 Bootstrap procedure

Let $\hat{\theta}_n(x)$ be some initial estimate of $\theta(x)$ with a given data set $Y_1, Y_2, \cdots, Y_n$. From this initial estimate, we can get an estimated residuals $\tilde{\epsilon}_i$ by $\tilde{\epsilon}_i = Y_i - \hat{\theta}_n(x_i), i = 1, 2, \cdots, n$ and let $\hat{\epsilon}_i$ be centered residual of $\tilde{\epsilon}_i$. That is,

$$\hat{\epsilon}_i = \tilde{\epsilon}_i - \frac{1}{n}\sum_{i=1}^{n} \tilde{\epsilon}_i, \quad i = 1, 2, \cdots, n \tag{2.2}$$

Freedman (1981) discussed the effect of centering of estimated residuals $\tilde{\epsilon}_i$ in bootstrapping of linear regression. We use the centering of residuals for the nonparametric kernel regression.

Now let $\hat{F}_n$ be empirical distribution function of centered residuals $\hat{\epsilon}_i$ and let $\hat{\epsilon}_1^*, \hat{\epsilon}_2^*, \cdots, \hat{\epsilon}_n^*$ be i.i.d. sample from the $\hat{F}_n$. From these conditionally independent random variable $\hat{\epsilon}_i^*$, we can get a new data set Y_i^* by

$$Y_i^* = \hat{\theta}_n(x_i) + \hat{\epsilon}_i^*, i = 1, 2, \cdots, n \tag{2.3}$$

2.3 Consistency

In kernel regression function estimation, the consistency of bootstrap method will be achieved by showing that the bootstrap distribution of $Z_n(x : c)$ has the same limiting distribution as the actual distribution of $Z_n(x; c)$.

Let

$$S(\theta) = [\hat{\theta}_n(x) : \int |(\hat{\theta}_n''(x) - \theta''(x)|dx \to 0, \int t^2 d\hat{F}_n(t) \to \sigma^2]$$

Define the bootstrap regression estimate of the form (1.2) by

$$\hat{\theta}_n^*(x:h) = \sum_{i=1}^{n} \frac{K((x-x_i)/h)}{\sum_{j=1}^{n} K((x-x_j)/h)} Y_i^* \tag{2.4}$$

and consider the bootstrap process $Z_n^*(x:c)$ of $Z_n(x:c)$ where

$$Z_n^*(x:c) = n^{2/5}(\hat{\theta}_n^*(x:cn^{-1/5}) - \hat{\theta}_n(x)) \tag{2.5}$$

From now on, denote E^* for conditional expectation under $\hat{F}_n$. We will consider the asymptotic behaviours of (2.5).

Lemma. If $\hat{\theta}_n(x) \in S(\theta)$, and with A.1, A.2, and A.3, $\lim_{n\to\infty} h^{-2}E^*(\hat{\theta}_n^*(x:h) - \hat{\theta}_n(x)) = K_2\theta''(x)/2$ where $K_2 = \int_{-A}^{A} t^2K(t)dt$.

Proof. By using inegral form of remainder term in taylor series expansion of θ and the definition of integral, we can easily get the results.

Then, we have the following theorem which says that the bootstrap process $Z_n^*(x:c)$ has the same limiting distribution as the actual process $Z_n(x:c)$.

Theorem. For fixed $c > 0$ and with A.1, A.2, and A.3,

$$Z_n^*(x:c) \xrightarrow{\mathrm{d}} N(c^2\mu(x), \sigma^2 \int_{-A}^{A} K^2(t)dt/c) \text{ as } n \to \infty, \text{ a.e.}x.$$

Proof. By using lemma, we get the mean and variance of (2.5). It remains only to prove the aymptotic normality of (2.5). Araujo and Gine(1980) proved the asymptotic normality of sum of a triangular array of row-wise independent random variables. The conditions for the astmptotic normality of (2.5) can be easily checked by using markov inequality, A.1, and lemma.

This consistency of bootstrap method can be used to the selection of data driven bandwidth. Let c_p^θ and c_p^b be the optimal choices minimising the limits, for p=1,2

$$\lim_{n\to\infty} n^{2/5}E\int_I |\hat{\theta}_n(x:cn^{-1/5}) - \theta(x)|^p dx$$

and

$$\lim_{n\to\infty} n^{2/5}E^*\int_I |\hat{\theta}_n^*(x:cn^{-1/5}) - \hat{\theta}_n(x)|^p dx$$

respectively, where I is the set which the regression function $\theta(\cdot)$ is defined, then we have the following corollary from the Theorem.

Corollary 1. For p =1,2

$$\lim_{n\to\infty} n^{2/5}E\int_I |\hat{\theta}_n(x:c_p^\theta n^{-1/5}) - \theta(x)|^p dx = \lim_{n\to\infty} n^{2/5}E\int_I |\hat{\theta}_n(x:c_p^b n^{-1/5}) - \theta(x)|^p dx.$$

3. SIMULATION STUDY

In this section, performance of bootstrap based bandwidth selection method for fixed sample size is compared with that of existing cross-validation method thru Monte Carlo simulation. Before simulations, we describe the two bandwidth selction method, cross-validation and bootstrap respectively with mathematical formula for the purpose of understanding of the two methods.

3.1 Cross-validation.

Wong (1983) firstly suggested the cross-validation method for bandwidth selection. The cross-validation method is to select the bandwidth which minimises

$$CV(h) = \frac{1}{n}\sum_{i=1}^{n}\{Y_i - \widehat{\theta}_{n,-i}(x_i : h)\}^2 \tag{3.1}$$

where $\widehat{\theta}_{n,-i}(x : h) = \sum_{j\neq i}^{n} \frac{K((x-x_j)/h)}{\sum_{j\neq i}^{n} K((x-x_j)/h)} Y_j$. $CV(h)$ measures the average ability of $\widehat{\theta}_{n,-i}(x_i : h)$ to predict the new observation Y_i.

3.2 Bootstrap method

The bootstrap method is to estimate $IMSE$ from the given sample and then find the bandwidth minimising the bootstrap estimate of $IMSE$. The straightforward approach would be to resample the estimated and centered residuals. But it requires the initial regression function estimate $\widehat{\theta}_n(x : h_0)$ of $\theta(x)$ for the residual estimation. So we take cross-validatory bandwidth h_0 as an objective initial estimate of bandwidth h.

Let $\widehat{\epsilon}_1^*, \widehat{\epsilon}_2^*, \cdots, \widehat{\epsilon}_n^*$ be i.i.d. samples from the $\widehat{F}_n$ where $\widehat{F}_n$ is empirical distribution function of estimated and centered residuals from the initial regression estimates. With these residuals, we obtain the resampled Y_i^* and construct the bootstrapped regression function estimate

$$\widehat{\theta}_{n:j}^*(x : h) = \sum_{i=1}^{n} \frac{K((x-x_i)/h)}{\sum_{i=1}^{n} K((x-x_i)/h)} Y_i^* \text{ for } j = 1, 2, \cdots, B$$

where Y_i^* is the same as in (2.3). Then, bootstrap estmates of $IMSE$ would be

$$\widehat{IMSE} = \frac{1}{B}\sum_{j=1}^{B}\int [\widehat{\theta}_{n:j}^*(x : h) - \widehat{\theta}_n(x : h_0)]^2 dx. \tag{3.2}$$

We can get the bandwidth minimising this estimate.

3.3 Simulation results.

In order to compare the performance of bootstrap method with existing cross-validation method for the bandwidth selection, Monte Carlo simulation were carried out.

Our test regression function considered were the following three different types of functions.

$$\theta_1(x) = sin(4\pi x) \text{ (periodic function)}$$
$$\theta_2(x) = 2x \text{ (linear function)}$$
$$\theta_3(x) = 32x(x-0.5)(x-1) \text{ (3^{rd} order polynomial)}$$

For reasons of computational efficiency, we used the Epanechnikov kernel (1969)

$$K(x) = \begin{cases} \frac{3}{4}(1-x^2) & \text{for } |x| \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

These curves were computed on [0,1] with two sample size n=50 and n=100. Errors were generated from the standard normal distribution. All computation was carried out by micro VAX system and random numbers were generated thru subroutine RNNOR and RNUND in IMSL.

$B = 50$ bootstrap sample were used. We had 100 replications for each test function. The numerical integration necessary to compute the relevant integrated squared error was done using an evenly spaced grid of 100 point from 0 to 1 using Simpson's method.

In Table 1, we calculated the $IMSE$ and the bandwidth h for each sample size. By best ISE we mean that the bandwidth is chosen for any given sample to minimise integrated squared error where we presume the full knowledge of the true underlying regression function. Hence this is the best bandwidth could possibly achieved with a kernel regression estimator given complete knowledge.

Sample correlation coefficients between the bandwidth chosen by each method and ISE based bandwidth are calculated respectively in Table 2. The correlation coefficients between ISEs and the estimate of ISEs which are the minimum value of the two method at each replications are also derived respectively.

Finally, we give some statistics in the columns marked by B/CV of the Table 2. These statistics are the percentage of samples where the bootstrap method produced a lower ISE followed by percentage of samples where cross-validation was superior.

4. CONCLUSION

The bootstrap method for the choice of bandwidth generally provides an improvement in performance where cross-validatory bandwidth is used as an objective

initial bandwidth. For all the three different functions considered and for the two sample sizes, the bootstrap based bandwidth choice yields smaller $IMSE$ than cross-validatory bandwidth. In examining the performance for individual samples, in all cases the booostrap method has smaller ISE, a great percentage of the time as shown in the Table 2 remarked by "comparison".

We proposed the bootstrap method for selecting the data driven bandwidth for the fixed design kernel regression function estimation. Where computational consideration permit, bootstrap compares favorably with cross-validation for these small sample size.

REFERENCES

(1) Araujo, A. and Gine, E. (1980). *The central limit theorem for real and banach valued random variable.* New York; Jhon Wiley and Sons.

(2) Beran, R. (1985). Bootstrap Methods in Statistics. *Jber. Math. Verein.*, 86, 14-30.

(3) Epanechnikov, V.A. (1969). Nonparametric estimation of multidimensional probability density. *Theory of Probability and its Applications*, 14, 153-158.

(4) Freedman, D.A (1981). Bootstrapping regression models. *Annals of Statistics*, 9, 1218-1228.

(5) Nadaraya, E.A. (1964). On estimating regression. *Theory of Probability and its Applications*, 9, 141-142.

(6) Rosenblatt,M.(1969). Conditional probability density and regression estimator. In, *Multivariate analysis II*, ed.,Krishnaiah,25-31.

(7) Silverman, B.W. (1985). *Density estimation for statistics and data analysis.* Chapman and hall , London.

(8) Watson, G.S. (1964). Smooth regression Analysis. *Shankya, Ser A.*, 26, 359-372.

(9) Wong, W.H. (1983). On the consistency of cross-validation in kernel nonparametric regression. *Annals of Statistics*, 11, 1136-1141.

Table 1 Comparison of Methods

(a). Best ISE

	n=50				n=100			
θ	$IMSE$	se	h	sd	$IMSE$	se	h	sd
1	.12790	.05849	.1178	.0194	.07543	.03094	.1024	.01686
2	.05356	.04636	.3349	.09204	.02651	.01957	.2938	.08043
3	.10811	.04805	.1529	.02955	.06843	.02811	.1267	.02587

(b). Cross-validation

	n=50				n=100			
θ	$IMSE$	se	h	sd	$IMSE$	se	h	sd
1	.14847	.06262	.1353	.03055	.09095	.03652	.1020	.02701
2	.07332	.05388	.3374	.09676	.03605	.02398	.2941	.08035
3	.12567	.05373	.1586	.03736	.07842	.03048	.1356	.03243

(c). Bootstrap

	n=50				n=100			
θ	$IMSE$	se	h	sd	$IMSE$	se	h	sd
1	.14654	.05946	.1390	.02567	.08726	.03316	.1076	.02259
2	.06899	.05077	.3444	.07278	.03529	.02294	.2939	.06857
3	.12154	.05250	.1640	.02713	.07742	.03062	.1474	.02541

Table 2 Correlation and Sample comparison

(a) For $n = 50$

function	Bandwidth		ISE		Comparison
	CV	$BOOT$	CV	$BOOT$	B/CV
1	-.3153	-.3842	.8330	.9058	51 / 33
2	-.2581	-.3929	.8859	.8830	50 / 42
3	-.3929	-.4468	.9340	.9421	50 / 40

(b) For $n = 100$

function	Bandwidth		ISE		Comparison
	CV	$BOOT$	CV	$BOOT$	B/CV
1	-.4896	-.4713	.9481	.9523	53 / 36
2	-.2841	-.3667	.8823	.9239	59 / 33
3	-.3940	-.4547	.9476	.9738	66 / 23