J. Statist. Comput. Simul., Vol. 37, pp. 37–44
Reprints available directly from the publisher
Photocopying permitted by license only

BOOTSTRAP SELECTION OF BANDWIDTH AND CONFIDENCE BANDS FOR NONPARAMETRIC REGRESSION

JULIAN J. FARAWAY

1444, Mason Hall, Department of Statistics, University of Michigan, Ann Arbor, Michigan 48109, USA

(*Received 2 June 1989; in final form 24 May 1990*)

A bootstrap method is developed to estimate the average squared error of a kernel based nonparametric regression estimator for a given bandwidth. This estimated average squared error is then minimised over the bandwidth to produce a regression estimate. Locally adaptive smoothing and simultaneous confidence bands may be obtained from this bootstrap method.

KEY WORDS: Nonparametric regression, bootstrap, bandwith selection, confidence bands.

1. INTRODUCTION

Let $x_1,\ldots,x_n$ be design points and let $e_1,\ldots,e_n$ be independent and identically distributed with unknown distribution F that has zero mean and variance $\sigma^2<\infty$. We observe $Y_1,\ldots,Y_n$ where

$$Y_i=r(x_i)+e_i \qquad i=1,\ldots,n.$$

The problem of interest is the estimation of the regression function $r(x)$ and the formation of simultaneous confidence bands for $r(x)$. Consider kernel regression estimates of the form

$$\hat{r}_h(x)=(nh)^{-1}\sum_{i=1}^{n} Y_iK((x-x_i)/h), \tag{1.1}$$

where K is a kernel function and h is the bandwidth. Appropriate selection of h is essential to the estimation of $r(x)$. The efficacy of the estimate may be judged according to various criteria among these being the average absolute error and the average squared error which are respectively

$$M_p(h)=n^{-1}\sum_{i=1}^{n}\left|\hat{r}_h(x_i)-r(x_i)\right|^p \qquad p=1,2.$$

For now we consider only $M_2(h)$, although the methods developed are simply extensible to $M_1(h)$. Various methods of bandwidth selection have been proposed among these being cross validation where h is chosen to minimise $MC(h)=n^{-1}\sum_{i=1}^{n}(Y_i-\hat{r}_{h,-i}(x_i))^2$ where $\hat{r}_{h,-i}$ is the regression estimate with ith observation omitted. Härdle, Hall and Marron (1988) discuss this and other crossvalidation

based bandwidth selection methods. It appears that there is not a great deal of difference between these methods so the version described above may be taken as representative. In Section 2, we describe a bootstrap method for estimating $M_2(h)$ and explain why this is not as straightforward as it might first appear. The method is an extension of work on bootstrap bandwidth selection for density estimates in Faraway and Jhun (1988). In contrast, Hall (1990) proposes a bootstrap method based on taking smaller resample sizes. We also describe the construction of simultaneous confidence bands for $r(x)$ and local adaptive smoothing where the bandwidth is allowed to vary with x. Construction of confidence bands is the major advantage of the bootstrap method. See Hall and Titterington (1988) for an alternative approach to the construction of confidence bands. We demonstrate that the bootstrapped process has the same limiting distribution as the actual regression estimation process. Härdle and Bowman (1988) also describe a bootstrap method of estimating $M_2(h)$. Their approach differs from ours in that they require explicit estimation of the second derivative of $r(x)$. They also discuss pointwise confidence bands and local adaptive smoothing as we do here. In Section 3, we conduct a simulation study to illustrate the application of our methods and to test their efficacy. The main conclusions of the study are that the bootstrap method of bandwidth selection is superior to crossvalidation, that the locally adaptive smoother is even more effective and that reasonably accurate simultaneous confidence bands may be obtained. In Section 4, we discuss the limitations of this work and describe possible extensions.

2. THE BOOTSTRAP METHOD

The prescription for the bootstrap method is as follows: First some residuals are required, which means that some initial estimate of $r(x)$ is needed. Cross-validation is a good choice here since it is simple to compute. Define the ith residual by

$$\tilde{e}_i = Y_i - \hat{r}_{h_0}(x_i) \qquad i = 1, \ldots, n,$$

where h_0 is the bandwidth selected by the crossvalidation method. Now the residuals must be recentred:

$$\hat{e}_i = \tilde{e}_i - n^{-1} \sum_{j=1}^{n} \tilde{e}_j \qquad i = 1, \ldots, n.$$

Since $\operatorname{var} \hat{e}_i < \sigma^2$ it is desirable to reinflate the residuals. This was done in the context of linear regression by Stine (1986) and, in this case, simulation shows it to be worthwhile. The reinflated residuals are

$$\hat{e}_i \to \hat{e}_i / \xi^{1/2}, \text{ where } \xi = \operatorname{var} \hat{e}_i / \sigma^2.$$

Direct calculation shows that $\xi = (1 - K(0)/nh)^2$ is an adequate approximation here. We then draw resampled residuals $e_1^*, \ldots, e_n^*$ from $\hat{e}_1, \ldots, \hat{e}_n$. Resampled Y^* are then constructed by

$$Y_i^* = \hat{r}_{h_0}(x_i) + e_i^* \qquad i = 1, \ldots, n.$$

For a given bandwidth h, an $\hat{r}_h^*(x)$ may be calculated from the bootstrap sample, $\underset{\sim}{Y}^*$. We repeat this procedure B times to obtain $\hat{r}_{jh}^*(x)$, $j = 1, \ldots, B$.

$M_2(h)$ may be decomposed into components of bias and variance. The variance may be adequately estimated by

$$(nB)^{-1} \sum_{i=1}^{n} \sum_{j=1}^{B} (\hat{r}_{jh}^*(x_i) - \bar{r}_h^*(x_i))^2,$$

where

$$\bar{r}_h^*(x_i) = B^{-1} \sum_{j=1}^{B} \hat{r}_{jh}^*(x_i).$$

The difference $\hat{r}_h(x_i) - \bar{r}_h^*(x_i)$, where $\hat{r}_h(x_i)$ is the estimator calculated from the given bandwidth h and the original data, fails as an estimate of the bias at x_i. To see this note that

$$E[\hat{r}_h(x_i) - \bar{r}_h^*(x_i)] = nh^{-1} \sum_{j=1}^{n} K((x_i - x_j)/h)\,(r(x_j) - E\hat{r}_{h_0}(x_j)),$$

which is a smoothing of the expected bias of the crossvalidated estimate. This will not work since as h increases, this estimate will tend to decrease, contrary to the known behaviour of the bias as h increases. This effect is also discussed in the context of quantile estimation in Padgett and Thombs (1986). The effect may be avoided by not having both terms in the bias estimate depend on h. Thus $\hat{r}_{h_0}(x_i) - \bar{r}_h^*(x_i)$ is a consistent estimate of the bias at x_i (see Appendix for proof) and may be combined with the estimate of the variance above to form a bootstrap estimate of $M_2(h)$:

$$MB(h) = n^{-1} \sum_{i=1}^{n} \left(B^{-1} \sum_{j=1}^{B} (\hat{r}_{jh}^*(x_i) - \bar{r}_h^*(x_i))^2 + (\hat{r}_{h_0}(x_i) - \bar{r}_h^*(x_i))^2 \right)$$

$$= (nB)^{-1} \sum_{i=1}^{n} \sum_{j=1}^{B} (\hat{r}_{jh}^*(x_i) - \hat{r}_{h_0}(x_i))^2.$$

The bootstrap choice of bandwidth will then be h_b where h_b minimises $MB(h)$ over h. Note that we may iterate the method, that is use h_b for the new initial estimate of $r(x)$ and apply the bootstrap method again.

2.1 Simultaneous Confidence Bands

Simultaneous confidence bands for $r(x)$ may be constructed in the following manner:

Let $\bar{b}_j = \max_{1 \leq i \leq n}[\hat{r}^*_{jh_b}(x_i) - \hat{r}_{h_0}(x_i)]$ and $\underline{b}_j = \max_{1 \leq i \leq n}[\hat{r}_{h_0}(x_i) - \hat{r}^*_{jh_b}(x_i)]$. $100(1-\alpha)\%$ confidence bands may then be constructed as $[\hat{r}_{h_b}(x) + \bar{b}_{(\alpha)}, \hat{r}_{h_b}(x) - \underline{b}_{(\alpha)}]$ where $\bar{b}_{(\alpha)}$ and $\underline{b}_{(\alpha)}$ are the appropriate sample percentile of the $\bar{b}$'s and $\underline{b}$'s respectively.

Note that these confidence bands are not simultaneously accurate for the entire domain of x but just at the design points. Of course, if the design points are suitably dense then the band will be approximately accurate for all x.

2.2 *Locally Adaptive Smoothing*

Locally adaptive smoothing may be implemented by using the bootstrap method to select a bandwidth for smoothing, h_i, at each of the design points instead of just averaging the average squared error over all the design points. Define the locally adaptive estimate as

$$\hat{r}_s(x) = (nh_i)^{-1} \sum_{i=1}^{n} Y_i K((x - x_i)/h_i).$$

It is reasonable to expect that the optimal local bandwidths, h_i, should be somewhat smooth as a function of x. If the number of bootstrap samples, B, is not sufficiently large, some irregularity may be expected in the h_i's, so it seems proper to smooth the estimated optimal local bandwidths. Some experimentation reveals that h_b is a suitable choice for the bandwidth to make this smooth.

2.3 *Theory*

This modified bootstrap procedure may be justified by showing that the bootstrapped process has the same limiting distribution as the process representing the actual regression estimate.

Let $x_i = i/n$ and let $r(x)$ be periodic with period 1 to avoid any problems with edge effects. Let $r(x)$ be twice differentiable and $r''(x)$ be continuous. Let the kernel K be a probability density with bounded support and $\int u^2 K(u)\, du = 1$. Known results (see, for example, Gasser and Müller (1979)) concerning $\hat{r}_h(x)$ are

$$E\hat{r}_h(x) - r(x) = (h^2/2) r''(x) + O(h^2),$$

$$\operatorname{var}(\hat{r}_h(x)) = (nh)^{-1} \sigma^2 \int K^2(x)\, dx + O((nh)^{-1}).$$

The rate at which the average squared error tends to zero is maximised when $h = cn^{-1/5}$ where c is some constant depending on r, f and K. For fixed $c \geqq 0$, $n \geqq 1$ and x consider the process

$$Z_n(x, c) = n^{2/5}[\hat{r}_h(x) - r(x)].$$

Write $\mu(x) = r''(x)/2$ and $\tau^2 = \sigma^2 \int K^2(x)\, dx$. Then

$$Z_n(x, c) \overset{\mathcal{D}}{\Rightarrow} N(c^2 \mu(x), \tau^2/c).$$

Let $\hat{r}_{h_o}$ be our initial density estimate, such that $\int_{-\infty}^{\infty} |\hat{r}''_{h_0}(x) - r''(x)|\, dx \to 0$. To ensure that this condition holds, we must take $nh_0^5 \to \infty$ while $h_0 \to 0$, $n \to \infty$. See Gasser and Müller (1984) for details.

Define the bootstrapped process as

$$Z_n^*(x, c) = n^{2/5}[\hat{r}_h^*(x) - \hat{r}_{h_0}(x)], \text{ where } h = cn^{-1/5}.$$

THEOREM *Let c be fixed and positive and let x vary over $[0, 1]$. For almost all sample sequences $Y_1, Y_2, \ldots$, the distribution of $Z_n^*(x, c)$ converges weakly to $N(c^2\mu(x), \tau^2/c)$ as $n \to \infty$.*

Proof See Appendix.

3. SIMULATION STUDY

We consider the following four regression functions $r(x)$:

i) $r(x) = \sin(2\pi x)$

ii) $r(x) = \sin(2\pi(1-x)^2)$

iii) $r(x) = (1 - 4(x - \frac{1}{2})^2)^2$

iv) $r(x) = \sin(4\pi x)$.

Three error distributions are considered: Standard normal, exponential (shifted to have mean 0) and a t with 3 degrees of freedom (scaled to have variance 1) and two sample sizes 100 and 400. We use a uniform kernel which facilitates rapid computation and means that only integer bandwidths need to be considered. Of course, other kernels might produce superior results, but the uniform kernel will suffice for a relative comparison of performance. We implement the local adaptive smoothing described above, trying both smoothing the local bandwidths using h_b as the bandwidth for this smoothing (LADs) and not smoothing the bandwidths (LADu). For each sample the value of h, h_a that minimises $M_2(h)$ is computed. We give the performance relative to this optimal choice of bandwidth as the ratio of the estimated expected average squared errors. We estimate the actual level of the nominally 80% simultaneous confidence bands computed for both the global and the locally smoothed bootstrap regression estimates by recording the percentage of the bands that actually cover the true regression curve. Results were based on 2000 replications and 200 bootstrap samples. Various numerical techniques to increase the accuracy of the bootstrap were tried with little discernible effect. See Hinkley (1988) for a description of these techniques.

See Table 1 for results. The estimated standard error of the ratios is no more than 5% of the given values and the estimated error in the coverage estimates is around 1%. In every case, the bootstrap method outperforms crossvalidation as a method of bandwidth choice. Curiously, the relative gap widens as we move from sample size 100 to 400. It is clear that it is necessary to smooth the bandwidths for local adaptive smoothing and if this is done the results are uniformly superior to the global smoothing methods. The computation of the actual level of the confidence bands is subject to two kinds of simulation error—that due to the

Table 1 $n=100$

Error distribution	*Regression function*	*Relative performance*				*Confidence level*		
		cv	*boot*	*LADs*	*LADu*	*global*	*locals*	*localu*
Normal	$2\sin(2\pi x)$	1.48	1.37	1.20	1.35	83.85	86.30	85.75
	$\sin(2\pi(1-x)^2)$	1.49	1.38	1.19	1.40	69.65	66.75	68.55
	$(1-4x^2)^2$	1.80	1.65	1.34	1.69	76.80	74.45	74.75
	$\sin(4\pi x)$	1.37	1.27	1.19	1.32	81.35	82.85	82.55
Exponential	$2\sin(2\pi x)$	1.49	1.37	1.21	1.37	79.55	82.35	78.90
	$\sin(2\pi(1-x)^2)$	1.54	1.43	1.23	1.47	66.55	62.90	62.05
	$(1-4x^2)^2$	1.80	1.65	1.37	1.70	75.65	72.95	72.30
	$\sin(4\pi x)$	1.34	1.25	1.19	1.30	74.60	77.85	75.25
t with 3df's	$2\sin(2\pi x)$	1.38	1.28	1.18	1.31	76.50	80.70	80.00
	$\sin(2\pi(1-x)^2)$	1.46	1.36	1.22	1.43	67.60	64.25	64.70
	$(1-4x^2)^2$	1.77	1.63	1.36	1.69	76.10	73.30	72.30
	$\sin(4\pi x)$	1.33	1.25	1.20	1.31	72.70	76.05	73.35

$n=400$

Error distribution	*Regression function*	*Relative performance*				*Confidence level*		
		cv	*boot*	*LADs*	*LADu*	*global*	*locals*	*localu*
Normal	$2\sin(2\pi x)$	1.38	1.26	1.17	1.30	79.75	87.00	89.05
	$\sin(2\pi(1-x)^2)$	1.34	1.22	1.12	1.31	76.80	65.05	73.80
	$(1-4x^2)^2$	1.59	1.42	1.22	1.46	84.60	84.15	85.30
	$\sin(4\pi x)$	1.30	1.20	1.10	1.23	86.60	88.95	90.75
Exponential	$2\sin(2\pi x)$	1.38	1.26	1.16	1.30	69.75	78.55	79.40
	$\sin(2\pi(1-x)^2)$	1.34	1.23	1.14	1.32	73.65	62.90	67.70
	$(1-4x^2)^2$	1.63	1.44	1.23	1.49	82.35	83.35	82.75
	$\sin(4\pi x)$	1.28	1.19	1.10	1.22	79.75	85.10	83.70
t with 3df's	$2\sin(2\pi x)$	1.34	1.25	1.16	1.29	75.00	80.35	81.45
	$\sin(2\pi(1-x)^2)$	1.32	1.24	1.14	1.32	69.55	61.30	64.50
	$(1-4x^2)^2$	1.57	1.43	1.23	1.47	79.45	81.85	81.85
	$\sin(4\pi x)$	1.28	1.20	1.12	1.22	80.35	83.40	83.25

bootstrap sample size and that due to the overall number of replications. Added to this is the error due to evaluating the bands only on a grid of values. Thus it is not possible to assess the accuracy of the confidence bands very carefully, but they do seem to be around the desired 80%. The confidence levels generally increase from sample size 100 to 400. The results are remarkably similar for the three error distributions.

A negative correlation was observed between h_b and h_a larger in magnitude than that between h_0 and h_a. The observed variances of the bandwidths were smallest for h_a and largest for h_0 with h_b falling in between.

4. DISCUSSION

Although the fact that the bootstrap selection of bandwidth appears to be superior to crossvalidation is interesting, the construction of simultaneous confidence bands

is the most important objective. The simulation results show that the method proposed appears to be at least approximately correct but a rigorous theoretical exposition would be required before these bands might be used with complete confidence.

One shortcoming of the method is that it relies on the assumption of constant variance in the errors. The bandwidth selection method might not be very sensitive to this but the confidence bands certainly would be. If the form of the deviation from constant variance were known or could be estimated then appropriate adjustments could be made.

In practice, we may have unequally spaced design points and some allowance will have to be made for edge effects.

The confidence bands for local adaptive estimates have constant width but some advantage may be gained by allowing width to vary according to the bandwidth.

Acknowledgements

The author thanks a referee for improvements in both the content and clarity of this paper.

References

Faraway, J. and Jhun, M. (1990). Bootstrap choice of bandwidth for density estimation. *JASA* **85**.

Gasser, T. and Müller, H. (1979). Kernel estimation of regression functions. In: *Smoothing Techniques for Curve Estimation. Springer Lecture notes* **757**, 23–68.

Gasser, T. and Müller, H. (1984). Estimating regression functions and their derivatives by the kernel method. *Scan. J. Stat.* 171–185.

Hall, P. (1990). Using the bootstrap to estimate mean squared error and select smoothing parameter in nonparametric problems. *J. Multi. Anal.* **32**, 177–203.

Hall, P. and Titterington, D. M. (1988). On confidence bands in nonparametric density estimation and regression. *J. Multi. Anal.* **27**, 228–254.

Härdle, W. and Bowman, A. (1988). Bootstrapping in nonparametric regression: Local adaptive smoothing and confidence bands. *JASA* **83**, 102–110.

Härdle, W., Hall, P. and Marron, J. (1988). How far are automatically chosen regression smoothing parameters from their optimum. *JASA* **83**, 86–95.

Hinkley, D. (1988). Bootstrap methods. *JRSSB* **50**, 321–337.

Padgett, W. and Thombs, L. (1986). Smooth nonparametric quantile estimation under censoring: Simulations and bootstrap methods. *Comm. Stat. B.* **15**, 1003–1025.

Stine, R. (1985). Bootstrap prediction intervals for regression. *JASA* **80**, 1026–1031.

APPENDIX

Proof of Theorem

First we show that

$$h^{-2}[E^*\hat{r}_h^*(x) - \hat{r}_{h_0}(x)] \to \mu(x).$$

To see this, note that

$$E^*\hat{r}_h^*(x) - \hat{r}_{h_0}(x) = \sum_{i=1}^{n} (nh)^{-1} K((x - x_i)/h)\hat{r}_{h_0}(x_i) - \hat{r}_{h_0}(x)$$

$$\approx \int_0^1 h^{-1} K((x-z)/h)\hat{r}_{h_0}(z)\,dz - \hat{r}_{h_0}(x) = \int_0^1 [K(u)\hat{r}_{h_0}(x+hu) - \hat{r}_{h_0}(x)]\,du \to h^{-2}\mu(x).$$

Also

$$\sum_{i=1}^{n} P[|(nh)^{-1} Y_i^* K((x-x_i)/h)|\varepsilon] \leqq \varepsilon^{-2} \sum_{i=1}^{n} E|(nh)^{-1} Y_i^* K((x-x_i)/h)|^2 \to 0 \qquad (\dagger)$$

and

$$\sum_{i=1}^{n} E[(nh)^{-1} Y_i^* K((x-x_i)/h) - E(nh)^{-1} Y_i^* K((x-x_i)/h)]^2$$

$$= \sum_{i=1}^{n} E[(nh)^{-1} \varepsilon_i^* K((x-x_i)/h)]^2 = \sigma^2 (nh)^{-1} \sum_{i=1}^{n} (nh)^{-1} K^2((x-x_i)/h)$$

$$\approx \sigma^2 (nh)^{-1} \int_0^1 h^{-1} K^2((x-z)/h)\,dz \to (nh)^{-1}\tau^2. \qquad (\dagger\dagger)$$

Since (†) and (††) hold, by a version of the CLT the theorem is proven.