한국통계학회논문집 제2권 1호
1995년 4월 pp. 99-111

Adaptive Kernel Density Estimation[1)]

Faraway,Julian.[2)] and Jhun,Myoungshic[3)]

Abstract

It is shown that the adaptive kernel methods can potentially produce superior density estimates to the fixed one. In using the adaptive estimates, problems pertain to the initial choice of the estimate can be solved by iteration. Also, simultaneous recommended for variety of distributions. Some data-based method for the choice of the parameters are suggested based on simulation study.

1. Introduction

Given data $x_1, x_2, \cdots, x_n$ we wish to estimate the unknown underlying density f which gave rise to these observations. Fixed kernel estimates of the form

$$\hat{f}(x) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{h}K\left(\frac{x-x_i}{h}\right)$$

have been extensively investigated in the literature. The idea of varying the bandwidth h to be larger in regions of data sparsity and smaller in regions where the data is plentiful has been proposed in the hope of producing less ragged estimates of the density. See Breiman, Meisel and Purcell(1977) for an early demonstration of this.

Our adaptive kernel density estimates will take the form

$$\hat{f}(x) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{h_i}K\left(\frac{x-x_i}{h_i}\right)$$

where $h_i = h\hat{f_0}(x_i)^{-\alpha}$ with $\hat{f_0}(x_i)$ being some initial estimate of the density at x_i chosen, for example, by a fixed kernel density estimate. h is the smoothing parameter and represents the overall amount of smoothing and α is the sensitivity parameter representing the degree of adaptation to sparsity. Proper selection of h and α is key to getting good kernel density estimates. Breiman et al. chose α=1. Abramson(1982) considered this choice α is too

1) This paper was supported (in part) by NON DIRECTED RESEARCH FUND, Korea Research Foundation, 1993.
2) Department of Statistics, University of Michgan, Ann Arbor, U.S.A.
3) Department of Statistics, Korea University, Seoul, 136-701, KOREA

large and propose $\alpha=1/2$ on the grounds that this particular choice of α is eliminated the leading bias term in the expansion of MSE at 0. Silverman(1985) suggests $\alpha=1/5$ for heuristic reasons based on the expansion of the mean integrated squared error(MISE).

We consider the problem of selecting the initial estimate of the density, α and h. We show, by empirical methods, that the problem of choosing the initial estimate of the density may be eliminated by iteration. We give empirical evidence to show that $\alpha=1/2$ may not be the best choice. We then investigate some data-dependent methods (crossvalidation and bootstrap) for selecting h and do a simulation study to show how they compare with fixed kernel estimates. Finally, we investigate the inter-relationship between α and h and how both may be simultaneously selected. We have used the L2 norm as our criterion, but the methods may, in most part, be extended to the L1 norm.

2. General set-up for simulation results

At this stage we give some common details of the numerical work which was used to obtain the results following:

The distributions we test our procedures on are

1) Standard Normal : N(0,1)
2) Bimodal Normal : $\frac{1}{2}N(-1,\frac{1}{4}) + \frac{1}{2}N(1,4)$
3) Contaminated Normal : $\frac{1}{2}N(0,4) + \frac{1}{2}N(0,\frac{1}{4})$
4) Standard Lognormal
5) Cauchy
6) Beta(2,2).

So we have short-tailed, long-tailed and asymmetric densities as test distributions in order to determine how the various procedures will perform in a wide variety of situations. Some rescaling and shifting was necessary for some of the distributions to get them fit our range comfortably.

The kernel used was the Epanechikov kernel

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

The simplicity of this kernel speeds computation.

The numerical integration necessary to compute the integrated squared errors

(ISE $= \int(f-\hat{f})^2$) was carried out on a grid of 100 points from -5 to 5. ISE's were

calculated at 20 values of bandwidth chosen evenly spaced on a log scale wide enough to include almost all conceivable choices of bandwidth generated by the bandwidth selection procedures for the given test distribution. Appropriate ranges were chosen using a pilot study. Uniform random numbers were obtained using a multiplicative congruential random number generator. Random numbers from the required distributions were then obtained using standard algorithms such as those found in Devroye(1987). Computations were carried out on a SUN 3/160. MISE's given are 10^{-2} x MISE.

3. Choice of the initial density estimate and α

In order to get our adaptive kernel estimates we need to compute

$$h_i = h\hat{f}_0(x_i)^{-\alpha}$$

We could obtain an initial estimate of the density f by using a fixed kernel density estimate using bandwidth h. However, we are introducing some undesirable dependency on the choice of h. It might be noted that the adaptive kernel estimate we make is not highly sensitive to this choice h, nevertheless, if this unpleasantness may be avoided, it would seem beneficial to do so. For this reason we propose the following iterative procedure.

Use a fixed kernel estimate based on any h:

1) Compute $\hat{f}(x_1), \cdots, \hat{f}(x_n)$.

2) Set $\tilde{f} = \hat{f}$ at x_i.

3) Repeat until convergence.

It is hard to prove the convergence of this iteration rigorously, but a heuristic proof is possible. For a wide variety of situations this procedure was tested and in every case convergence was swift and sure. We demonstrate this convergence in a few instances below. Hence, the initial choice of h is immaterial.

In order to investigate the appropriate choice of α we plot the ISE's for the adaptive kernel estimates at 4 stages of the iteration against log(h) for a sample of size 50 generated from N(0,1). The results for α=0.5 and α=0.2 for a variety of initial choices of bandwidth(init) are shown in figure 1. The solid line indicates the adaptive kernel estimate after one step of the iteration, the dashed lines after two, three and four iterations. Note that no matter what the initial choice is, the ISE's converge to the same shape. Note also that when α=0.5 the estimates become generally worse as another iteration is done and that the shape of the curve becomes non-convex whereas no such undesirable behavior occurs for α=0.2. Figure 2

shows the results of the same procedures applied to a sample from the contaminated normal distribution $\frac{1}{2}N(0,4) + \frac{1}{2}N(0,\frac{1}{4})$. As one can see, the picture remains very similar.

In conclusion one can say that if one wishes to avoid arbitrary selection of the pilot estimate, iteration is an appropriate method. Given the proposed alternatives of α=0.2 and α =0.5 then possibly it seems preferable to choose α=0.2, but we can make no clear decision here on the basis of just these plots. However, we will discuss further the choice of α and how it relates to h and the underlying true density in section 7.

4. Choice of h

Crossvalidation is presently the most popular method of data-dependent bandwidth selection. Faraway & Jhun(1990) describe a bootstrap based choice of bandwidth. We employ both these methods in comparing the performance of fixed and adaptive kernel density estimates.

Crossvalidation : The crossvalidation criterion will be

$$CV = \int \hat{f}(x)^2 dx - (2/n)\sum_{i=1}^{n} \hat{f}_{-i}(x)$$

where $\hat{f}_{-i}$ is the density estimate based on all the data except x_i.

The crossvalidated choice of bandwidth is that value of h which minimizes CV(h). See Silverman(1985) for details of this in the context of adaptive kernel density estimation.

Bootstrap : We use a smoothed bootstrap estimate of the MISE to select the bandwidth, because the unsmoothed bootstrap does not work. If we resample $X_1^*, \cdots, X_n^*$ from the empirical distribution F_n and get estimates of the density $\hat{f}_j^*(x)$ for $j = 1, \cdots, B$ where B is the number of bootstrap samples, then our bootstrapped estimate of the MISE is

$$BS(h) = (1/B)\sum_{j=0}^{B} \int [\hat{f}_j^*(x) - \bar{f}_B(x)]^2 dx$$

where $\bar{f}_B(x) = (1/B)\sum_{j=1}^{B} \hat{f}_j^*(x)$. However, BS(h) is found to be decreasing in h because it estimates only the variance component of the MISE. This is true for both fixed and adaptive kernel estimates.

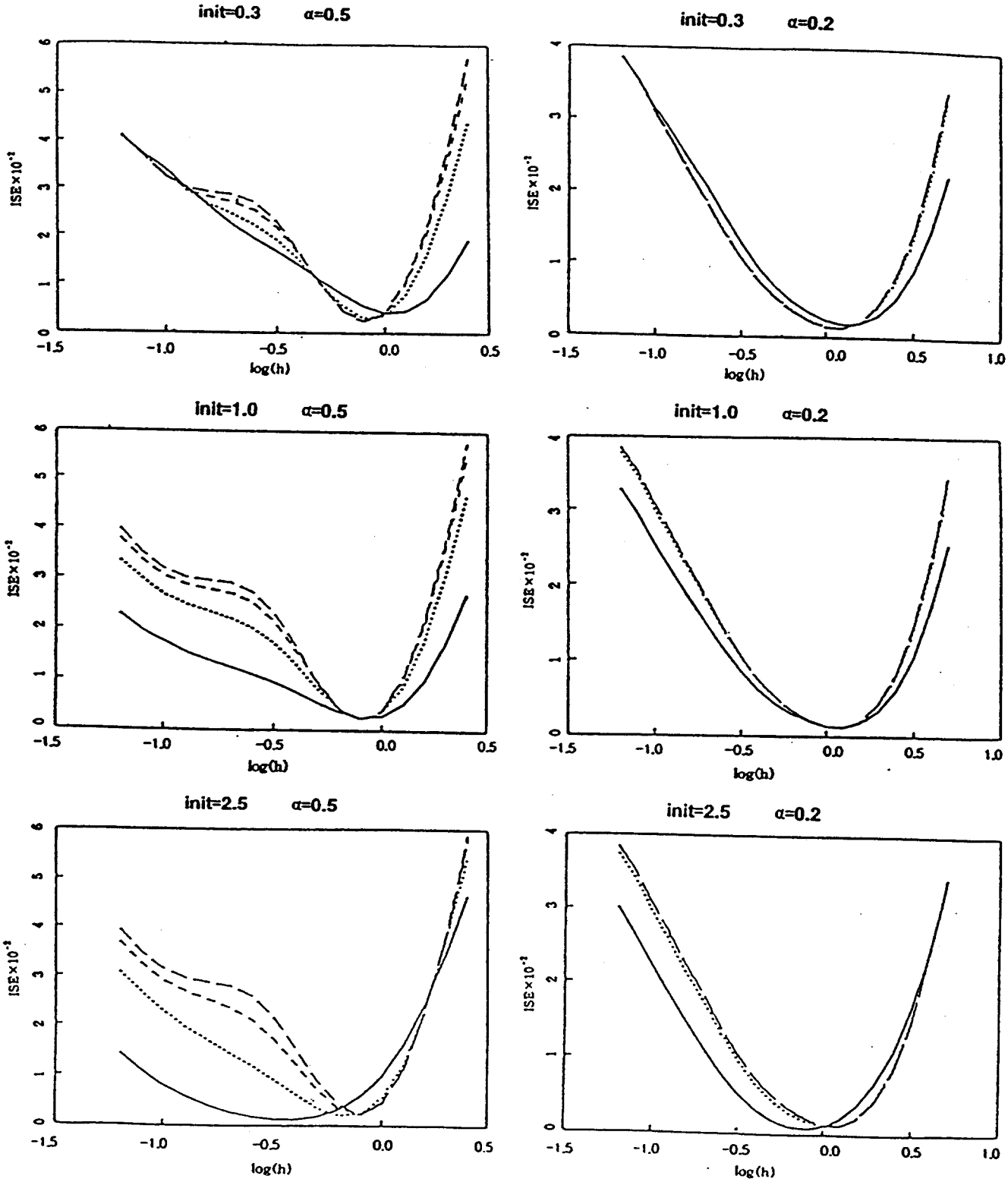

Figure 1 : Iteration for the choice of α (N(0,1), n=50)

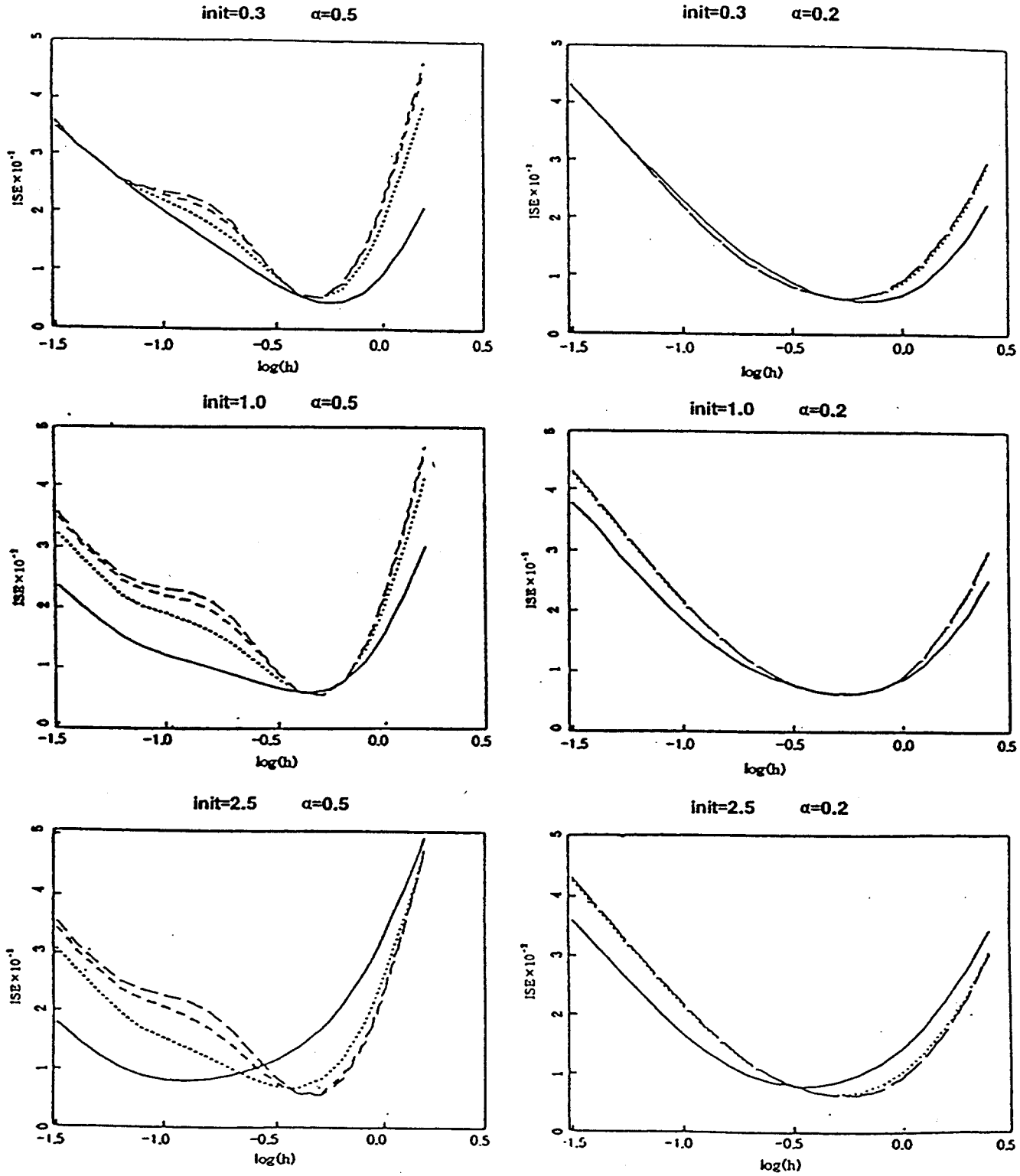

Figure 2 : Iteration for the choice of α ($\frac{1}{2}$N(0,4) + $\frac{1}{2}$N $(0, \frac{1}{4})$, n=50)

For this reason we use the smoothed bootstrap. We add a random amount $h_j'\epsilon$ to each X_j^* where ϵ is randomly distributed with the kernel K. h_j' are obtained from

$$h_j' = h' \hat{f}(x_j)^{-\alpha}$$

for adaptive kernel estimates where h' is some initial choice for the bandwidth or $h_j' = h'$ for fixed kernel estimates where h' is again some initial choice for the bandwidth. So each X_j^* is changed thus

$$X_j^* \rightarrow X_j^* + h_j'\epsilon.$$

We now construct $\hat{f}_j^*(x)$ from these new X^*'s and construct $BS(h,h')$

$$BS(h,h') = (1/B)\sum_{j=1}^{B}\int[\hat{f}_j^*(x) - \bar{f}_B(x)]^2 dx.$$

Our bootstrap choice of h is made by minimizing $BS(h,h')$ over h. In our simulation study we use crossvalidation to select h' so this bootstrap method may be viewed as an attempt to improve on the crossvalidated choice of bandwidth.

5. Simulation study

We have results for sample size 50 for both α=0.5 and α=0.2. 100 replications were used in each case. B=50 bootstrap samples were used. Of course this is on the low side but these methods are computationally expensive and some economy was required in order to complete the simulation in a reasonable time on the equipment available. Furthermore it seems reasonable to assume that any increase in the number of bootstrap samples taken would only serve to improve the performance of this estimator.

In our tables(1-4) of results we see adaptive kernel and fixed kernel estimates compared alongside each other. The same data was used for both. By 'fixed choice' we mean that fixed choice of bandwidth that minimizes the MISE over all the samples taken. By 'ISE choice' we mean the bandwidth is chosen so as to minimize the ISE for the given sample given knowledge of the true density. The sample mean of the ISE's corresponding to these best

choices is given along with an estimated standard error. The sample mean and SD of the bandwidth chosen is also given. So the ISE choice estimate is the best one could possibly do using a kernel density estimate. We present the corresponding statistics for the crossvalidated choice of bandwidth and the bootatrap choice of bandwidth.

We also give in table 4 the sample comparisons. In practice we would have some data and would ask "Which method is the best for this data?" So for each sample we check the ISE's for each of the methods and compare

1) For fixed kernel, the ISE's of the crossvalidated and bootstrap method.(cv/bs)

2) For adaptive kernel for some, the ISE's of the crossvalidated and bootstrap method.(cv/bs)

3) For the ISE choice, the ISE's of the fixed kernel and adaptive kernel method for some α (fix/adp).

What is shown in the table is the percentage of samples where a particular method had a smaller ISE than the other. Note that the percentages do not all sum to 100% because we calculate the ISE at only 20 bandwidths and so the remainder represents the percentage of samples where both methods made the same choice of bandwidth.

Table 1. Fixed Kernel Density Estimate (n=50)

Distribution	Fixed choice			ISE choice				Crossvalidation				Bootstrap			
	mise	se	h	mise	se	h	sd	mise	se	h	sd	mise	se	h	sd
normal	0.79	0.06	1.11	0.73	0.06	1.04	0.19	1.52	0.13	1.06	0.41	1.22	0.09	1.11	0.32
binomial	1.86	0.10	0.64	1.77	0.10	0.62	0.10	2.50	0.15	0.70	0.22	2.48	0.15	0.79	0.23
cont normal	1.45	0.09	0.79	1.39	0.08	0.78	0.12	2.21	0.15	0.76	0.30	1.95	0.12	0.86	0.28
lognormal	3.29	0.13	0.43	3.17	0.12	0.46	0.09	4.08	0.17	0.48	0.19	3.89	0.15	0.57	0.19
cauchy	5.78	0.34	0.25	5.51	0.34	0.24	0.04	7.57	0.44	0.28	0.11	7.20	0.42	0.31	0.10
beta	0.87	0.06	1.20	0.78	0.05	1.17	0.25	1.39	0.11	1.09	0.38	1.16	0.09	1.18	0.28

Table 2. Adaptive Kernel Density Estimate (n=50, α=0.2)

Distribution	Fixed choice			ISE choice				Crossvalidation				Bootstrap			
	mise	se	h	mise	se	h	sd	mise	se	h	sd	mise	se	h	sd
normal	0.71	0.06	0.90	0.64	0.06	0.90	0.15	1.39	0.12	0.92	0.32	1.15	0.10	0.95	0.25
binomial	1.82	0.11	0.55	1.73	0.10	0.52	0.08	2.65	0.19	0.56	0.17	2.51	0.16	0.61	0.17
cont normal	1.29	0.08	0.67	1.21	0.08	0.67	0.11	2.06	0.14	0.65	0.24	1.81	0.11	0.72	0.22
lognormal	3.21	0.13	0.41	3.06	0.13	0.40	0.09	4.40	0.24	0.40	0.18	4.03	0.19	0.46	0.17
cauchy	5.11	0.33	0.27	4.81	0.33	0.25	0.05	7.12	0.46	0.29	0.11	6.70	0.44	0.31	0.09
beta	1.00	0.06	1.00	0.86	0.05	0.96	0.24	1.48	0.11	0.92	0.30	1.29	0.09	0.97	0.23

Table 3. Adaptive Kernel Density Estimate (n=50, α=0.5)

Distribution	Fixed choice			ISE choice				Crossvalidation				Bootstrap			
	mise	se	h	mise	se	h	sd	mise	se	h	sd	mise	se	h	sd
normal	0.74	0.05	0.82	0.60	0.05	0.72	0.16	1.33	0.11	0.78	0.22	1.13	0.10	0.78	0.16
binomial	1.91	0.13	0.41	1.78	0.12	0.39	0.06	2.62	0.19	0.41	0.10	2.45	0.17	0.43	0.10
cont normal	1.04	0.07	0.61	0.95	0.07	0.59	0.10	1.82	0.13	0.57	0.18	1.61	0.11	0.61	0.16
lognormal	3.41	0.15	0.30	3.09	0.13	0.32	0.10	4.55	0.22	0.34	0.17	4.22	0.19	0.38	0.16
cauchy	4.38	0.31	0.27	3.97	0.30	0.26	0.06	6.58	0.46	0.30	0.11	6.05	0.44	0.32	0.09
beta	1.44	0.06	0.90	1.08	0.06	0.72	0.25	1.81	0.10	0.78	0.24	1.76	0.10	0.78	0.20

Table 4. Sample comparison (%)

Distribution	Fixed		α=0.2		α=0.5		α=0.2		α=0.5	
	cv	bs	cv	bs	cv	bs	fix	adp	fix	adp
normal	22	52	18	52	5	56	11	89	41	59
bimodal	49	39	35	32	14	32	22	78	38	62
cont normal	26	44	26	45	12	36	2	98	10	90
lognormal	41	46	35	48	23	47	25	75	46	54
cauchy	33	40	30	43	29	42	0	100	3	97
beta	20	63	20	58	21	36	80	20	82	18

6. Discussion of results

Comparison of fixed with adaptive kernel estimates : For α=0.2 we see that in terms of MISE of the fixed and ISE choices of bandwidth and sample comparisons the adaptive is superior to the fixed kernel in all cases except the Beta distribution. This is not surprising since Beta(2,2), which we have used, has no tails. Given that our reason for proposing the adaptive kernel estimate is to adapt for regions where data is sparse it is understandable that this method should not work so well for this density. In making the same comparisons for α =0.5 the superiority of the adaptive kernel is not so clear cut as one can see that the fixed kernel outperforms the adaptive in some instances for distributions other than Beta.

However, when we consider the practical situation where the bandwidth must be chosen without knowledge of the true density, we may be using some automatic choice of bandwidth like crossvalidation or bootstrap. Here the results are mixed and there is no clear edge for the adaptive kernel method. This indicates the problem of selecting the bandwidth for an adaptive kernel density estimate is more difficult than one for the fixed kernel.

Comparison of Crossvalidation and Bootstrap : For the adaptive kernel we see that bootstrap is superior to crossvalidation in terms of MISE and sample comparisons almost across the board. For the fixed kernel the difference between the two methods is less pronounced but there is a clear edge to the bootstrap method. Note that bootstrap tends to

choose wider bandwidths than crossvalidation or the ideal ISE choice. This tendency to slightly oversmooth may not be unwelcome in practice.

Comparison of α=0.2 with α=0.5 : We can give no clearer preference toward either choice. It is apparent that the best choice of α is dependent on the underlying density and so we investigate this problem in the next section.

7. Simultaneous choice of the smoothing parameters

In order to investigate the behavior of the estimator over simultaneously varying choices of α and the bandwidth h, we did a further simulation study. This time we computed the estimate over 20×20 grid of α×h. 100 replications were made for sample of size 50 and 100. The same test distributions were used. We display our results in two ways.

The results are displyed in tables 5 and 6. For 'Fixed choice' we give the fixed values of α and h which minimizes the MISE over all replications. For the 'ISE choice', the ISE is minimized over α and h for each sample and statistics for these choices are given. In the column marked 'corr' we give the correlation between the choice of α and h.

Table 5. Adaptive Kernel Density Estimates (n=50)

Distribution	Fixed Choice			ISE choice						
	mise	α	h	mise	se	α	sd	h	sd	corr
normal	0.63	0.20	0.90	0.46	0.05	0.30	0.25	0.90	0.30	-0.85
bimodal	1.91	0.30	0.50	1.71	0.09	0.23	0.47	0.58	0.27	-0.95
cont normal	0.88	0.70	0.55	0.62	0.05	0.69	0.24	0.53	0.13	-0.83
lognormal	2.93	0.10	0.41	2.42	0.12	0.33	0.46	0.39	0.18	-0.79
cauchy	3.06	0.70	0.35	2.04	0.18	0.75	0.25	0.33	0.06	0.20
beta	0.83	-0.60	2.23	0.72	0.06	-0.44	0.50	2.11	1.02	-0.93

Table 6. Adaptive Kernel Density Estimates (n=100)

Distribution	Fixed Choice			ISE choice						
	mise	α	h	mise	se	α	sd	h	sd	corr
normal	0.47	0.30	0.82	0.32	0.03	0.30	0.26	0.79	0.28	-0.91
bimodal	1.04	0.30	0.45	0.89	0.05	0.36	0.36	0.44	0.17	0.91
cont normal	0.59	0.60	0.55	0.41	0.03	0.64	0.21	0.51	0.12	-0.80
lognormal	1.95	0.20	0.33	1.68	0.08	0.30	0.41	0.32	0.12	-0.65
cauchy	1.52	0.60	0.33	1.03	0.08	0.64	0.21	0.31	0.05	0.38
beta	0.53	-0.50	1.65	0.44	0.03	-0.40	0.54	1.76	0.95	-0.92

Discussion : The optimal choice of α varies greatly. It is large for the contaminated normal and cauchy, our long tailed distributions. It is quite negative for the beta. This most interesting since negative choices for α would not seem reasonable at first glance. However, with a little thought, one can see how this is reasonable for the beta(2,2). One might expect the same effect for a uniform distribution. Note also that the sample standard deviations on the choice of α are relatively large. These observations lead us to propose a databased method of simultaneous selection of α and bandwidth. A marked negative correlation is shown between choice of α and bandwidth except in the case of Cauchy. Of course we know that α and the bandwidth are linked in some non-simple way but it is curious that the Cauchy should behave in so different manner. The MISE's of the ISE choices show that, if we can only choose α and the bandwidth well, we may be able to obtain superior estimates.

Data-based choice of the smoothing parameters : We may use crossvalidation or the bootstrap method proposed earlier but the major difficulty is that previously we considered searching over a grid of 20 values of the bandwidth, now we introduce an additional dimension of α. We must be wary of search methods which depend on convexity but evaluation at all points on the grid may be prohibitively expensive. For this reason we restricted ourselves to a 10×8 grid of $\alpha\times h$ chosen appropriately for each distribution. A pilot study showed that crossvalidation performed rather poorly in simutaneously selecting h and α. This might have been expected since the crossvalidated choice of just h is rather variable and adding the extra dimension of α just proves to be too much.

However, our bootstrap method requires an initial choice of the parameters. Previously we used the crossvalidated choice but since this proved to be rather poor, the bootstrap choice based on these initial values, although an improvement, was not so good either. Iteration of the bootstrap choice would be advisable but rather expensive computationally. Therefore we used the bootstrap choice from the fixed kernel method as our initial choice of h and so $\alpha=0$ was our initial choice of α .

We were able to do 100 replications. We give simple comparisons for the bootstrap fixed and adaptive versions, otherwise the layout is as before.

Table 7. Bootstrap choice of smoothing parameters (n=50)

Distribution	Fixed estimates				Adaptive estimates						
	mise	se	h	sd	mise	se	α	sd	h	sd	corr
normal	1.22	0.09	1.11	0.32	1.15	0.08	0.19	0.25	1.05	0.30	-0.71
bimodal	2.48	0.15	0.79	0.23	2.52	0.14	-0.03	0.26	0.81	0.24	-0.70
cont normal	1.95	0.12	0.86	0.28	1.51	0.12	0.46	0.21	0.68	0.20	-0.39
lognormal	3.89	0.15	0.57	0.19	3.97	0.16	0.16	0.23	0.61	0.18	-0.50
cauchy	7.20	0.42	0.31	0.10	5.11	0.34	0.53	0.26	0.32	0.08	0.25
beta	1.16	0.09	1.18	0.28	1.14	0.10	-0.10	0.34	1.48	0.56	-0.87

Table 8. Sample comparison (%)

Distribution	Bootstrap		ISE	
	fix	adp	fix	adp
normal	36	64	12	88
bimodal	53	47	7	93
cont normal	22	78	0	99
lognormal	65	35	7	93
cauchy	13	87	0	100
beta	44	56	4	96

Discussion : In comparing the bootstrap fixed and adaptive results we see that the adaptive method does better for these distributions, about the same for two and worse for one other. Some explanation for this behavior may be found in studying the ISE choices in the fixed and adaptive cases. Where (given omniscent choice of α) the adaptive kernel method provides a big increase in performance we can expect the bootstrap adaptive method do well. Otherwise, when no great improvement is possible, adding data-dependent choice of α serves mostly to add extra variability to the problem and hence the moderate performance.

8. Conclusion

The use of adaptive kernel methods can potentially produce superior estimates of the density. Realizing this potential is problematic. One must select both the bandwidth h and sensitivity parameter α. No fixed choice of α can be recommended. If data-dependent methods are to be used in selecting parameters, bootstrap is superior to crossvalidation. In practice there can be no guarantee that adaptive kernel will do better than fixed kernel although there is some indication that in general one can expect superior performance. We have only studied the univariate case. In higher dimensions adaptive kernel methods show a more distinct superiority to the fixed kernel method. In this situation our methods will be more useful.

References

[1] Abramson,I.S.(1982). On bandwidth variation in Kernel estimates - a square root law, *Annals of Mathematical Staistics*, Vol. 10, 1217-1223.

[2] Breiman, L., Meisel, W., Purcell, E.(1977). Variable Kernel Estimates of Multivariate Densities. *Technometrics,* Vol. 19, 135-144

[3] Devroye,L.(1987). *Non-uniform random variate generation*, Springer Verlag

[4] Faraway,J.,Jhun,M.(1990). Bootstrap Choice of Bandwidth for Density Estimation, *Journal of the American Statistical Association*, Vol. 85, 1119-1122

[5] Silverman,B.W.(1985). *Density Estimation for Statistics and Data Analysis*, Chapman Hall, London.