APPLICATIONES MATHEMATICAE
22,1 (1993), pp. 123–138

L. GAJEK (Warszawa and Łódź)
A. LENIC (Łódź)

AN APPROXIMATE NECESSARY CONDITION FOR THE OPTIMAL BANDWIDTH SELECTOR IN KERNEL DENSITY ESTIMATION

Abstract. An approximate necessary condition for the optimal bandwidth choice is derived. This condition is used to construct an iterative bandwidth selector. The algorithm is based on resampling and step-wise fitting the bandwidth to the density estimator from the previous iteration. Examples show fast convergence of the algorithm to the bandwidth value which is surprisingly close to the optimal one no matter what is the initial knowledge on the unknown density.

1. Introduction. Let $X_1, \dots, X_n$ be i.i.d. observations from an unknown density f. To describe the quality of a given estimator $\widehat{f}$ of f one can use the *integrated square error* (ISE) of $\widehat{f}$:

$$\mathrm{ISE}(\widehat{f}) = \int [\widehat{f}(x) - f(x)]^2 \, dx \,,$$

and the *mean integrated square error* (MISE) of $\widehat{f}$:

$$\mathrm{MISE}(\widehat{f}) = E[\mathrm{ISE}(\widehat{f})] \,.$$

Having k samples from f, one can evaluate MISE of $\widehat{f}$ by the *average integrated square error* (AISE) defined as follows:

$$\mathrm{AISE}(\widehat{f}) = \frac{1}{k} \sum_{i=1}^{k} \mathrm{ISE}(\widehat{f}_i) \,,$$

where $\widehat{f}_i$ denotes the estimate $\widehat{f}$ of f obtained from the ith sample.

1991 *Mathematics Subject Classification*: Primary 62G07, 62G09; Secondary 65D10.
Key words and phrases: kernel density estimation, bandwidth selection, resampling.

Throughout the paper we shall consider the kernel estimator

$$f(x;h) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right), \tag{1}$$

where K is a kernel function and $h > 0$ is the bandwidth (see Rosenblatt (1956) and Parzen (1962)). A proper choice of h is crucial for the precision of estimation (see e.g. Silverman (1986) for a discussion of the problem). A survey of existing selection methods for h can be found in Silverman (1986) and Härdle, Hall and Marron (1988).

If the MISE of the kernel estimator (1) is of main interest, the optimal bandwidth $\widehat{h}$ can be obtained by solving the following minimization problem: Find $\widehat{h} > 0$ such that

$$E\int [f(x;\widehat{h}) - f(x)]^2\,dx = \min_{h>0}\left\{E\int [f(x;h) - f(x)]^2\,dx\right\}. \tag{2}$$

To solve the above problem one needs to know f which is unknown. However, $f(x;\widehat{h})$ itself is by (2) the best possible approximation of f, so assuming for the moment that $f(x;\widehat{h})$ is known, one might instead find a solution of

$$E^*\int [f^*(x;h) - f(x;\widehat{h})]^2\,dx \to \min_h!, \tag{3}$$

where $f^*(x;h)$ is the kernel estimator (1) based on samples $X_1^*, \ldots, X_n^*$ from the density $f(x;\widehat{h})$ and E^* is taken with respect to $X_1^*, \ldots, X_n^*$. Since $f(x;\widehat{h})$ approximates f, the solution of (3) should be close to $\widehat{h}$. Thus we arrive at the following approximate necessary condition for the optimal bandwidth selector $\widehat{h}$:

$$E^*\int [f^*(x;\widehat{h}) - f(x;\widehat{h})]^2\,dx \approx \min_{h>0}\left\{E^*\int [f^*(x;h) - f(x;\widehat{h})]^2\,dx\right\}. \tag{4}$$

Let $h^* = h^*(X_1, \ldots, X_n)$ be a value of $\widehat{h}$ for which (4) holds with the exact equality sign $=$ being put in place of $\approx$, i.e.

$$E^*\int [f^*(x;h^*) - f(x;h^*)]^2\,dx = \min_{h>0}\left\{E^*\int [f^*(x;h) - f(x;h^*)]^2\,dx\right\}.$$

In the paper we investigate a self-learning algorithm which detects h^*. In Section 2 we describe the algorithm and discuss some of its properties.

In Section 3 we present how the algorithm works for the data from normal, beta$(3,5)$, Cauchy, and a bimodal mixture of normal distributions. It turns out that the choice of the initial point of the algorithm is of minor importance influencing only the number of iterations. The resulting bandwidth selector is usually very close to the minimizer $\widehat{h}$ of (2) which is optimal provided the density f is known. Finally, we have applied the algorithm to the eruption lengths of 107 eruptions of the Old Faithful geyser. The resulting bandwidth $h = 0.21$ is surprisingly close to Silverman's subjective choice $h = 0.25$ (comp. Silverman (1986)).

2. The self-learning algorithm. Let $X_1, \dots, X_n$ be an i.i.d. sample from an unknown density f. Let f_0 be some initial density, depending on the sample $X_1, \dots, X_n$ or not. Let ε be a positive real.

1° Put $i := 0$, $h_0 := 0$.

2° Generate m samples $X_{1j}^*, \dots, X_{nj}^*$, $j = 1, \dots, m$, of size n from f_i.

3° Find $h_{i+1} > 0$ such that

$$\frac{1}{m}\sum_{j=1}^{m} \int_{\mathbb{R}} [f_i(x) - f_j^*(x; h_{i+1})]^2 \, dx = \min_{h>0} \left\{ \frac{1}{m} \sum_{j=1}^{m} \int [f_i(x) - f_j^*(x; h)]^2 \, dx \right\},$$

where

$$f_j^*(x; h) := \frac{1}{nh} \sum_{k=1}^{n} K\left(\frac{x - X_{kj}^*}{h} \right).$$

4° Put $i := i + 1$ and

$$f_i(x) := \frac{1}{nh_i} \sum_{k=1}^{n} K\left(\frac{x - X_k}{h_i} \right).$$

5° If $|h_i - h_{i-1}| < \varepsilon$, then stop; otherwise go to 2°.

It is easy to see that when m is large and ε small, the algorithm stops near the point

$$h^* = h^*(X_1, \dots, X_n)$$

such that

$$E^* \int [f(x; h^*) - f^*(x; h^*)]^2 \, dx \le E^* \int [f(x; h^*) - f^*(x; h)]^2 \, dx \,, \tag{5}$$

where $f^*(x; h)$ is the estimator (1) based on the sample $X_1^*, \dots, X_n^*$ from the density $f(x; h^*)$ and the expectation operator E^* is taken with respect to this sample. Let us notice that h^* differs from the bandwidth h_T proposed by Taylor (1989), which was to minimize

$$E^* \int [f(x; h) - f^*(x; h)]^2 \, dx \,.$$

A related bootstrap method of selecting the bandwidth was considered by Faraway and Jhun (1990). However, instead of finding h^* defined by (5), they have proposed a one- or two-step bootstrap procedure which, in fact, strongly depends on the initial choice of the bandwidth.

A survey of existing bootstrap methods and their applications can be found in Léger, Politis and Romano (1992).

3. Simulation results. Samples of size $n = 50$ are taken from normal $N(0, 1)$, bimodal mixture of normals $0.5N(-1.5; 0.25) + 0.5N(1.5; 0.25)$, beta$(3, 5)$ and standard Cauchy distributions, respectively. We have applied

the random number generator ULTRA combined with standard procedures. The bootstrap samples from kernel estimators were generated following Silverman (1986, p. 143). The Epanechnikov kernel is used in all examples except for the last one where the Gaussian kernel is applied.

In Table 1 the successive iterations of the bandwidth and the corresponding values of ISE for the self-learning algorithm are shown for the normal $N(0,1)$ distribution when the initial value of bandwidth $h_0 = 1$ is chosen. For $h_0 = 0.2$ the results are given in Table 2. The corresponding kernel density estimators are shown in Figures 1–10. Here and further on, the number of bootstrap samples is $m = 20$. The iterations were proceeded as long as the sequence h_i was monotone. As can be seen, the final values of the bandwidth are for both cases quite close to each other.

TABLE 1

h_i	ISE(h_i)
1	$222657 \cdot 10^{-7}$
0.71	$675786 \cdot 10^{-8}$
0.57	$338645 \cdot 10^{-8}$
0.54	$309890 \cdot 10^{-8}$
0.51	$296933 \cdot 10^{-8}$
0.50	$295942 \cdot 10^{-8}$

TABLE 2

h_i	ISE(h_i)
0.2	$197113 \cdot 10^{-7}$
0.29	$793087 \cdot 10^{-8}$
0.42	$339964 \cdot 10^{-8}$
0.48	$298620 \cdot 10^{-8}$

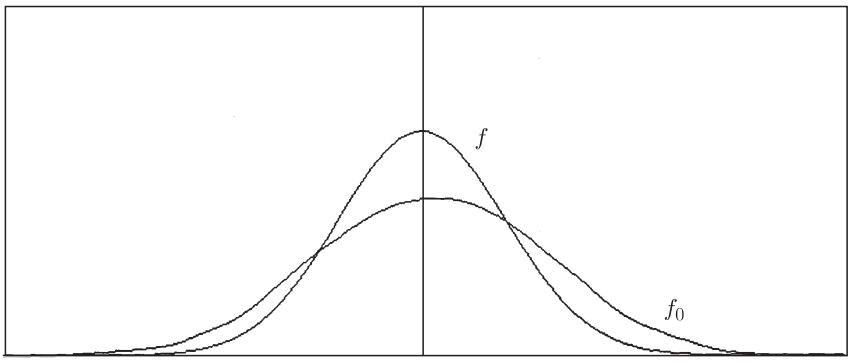

Fig 1. Estimation of the normal $N(0,1)$ density f. The initial density f_0 is the kernel estimator with $h_0 = 1$

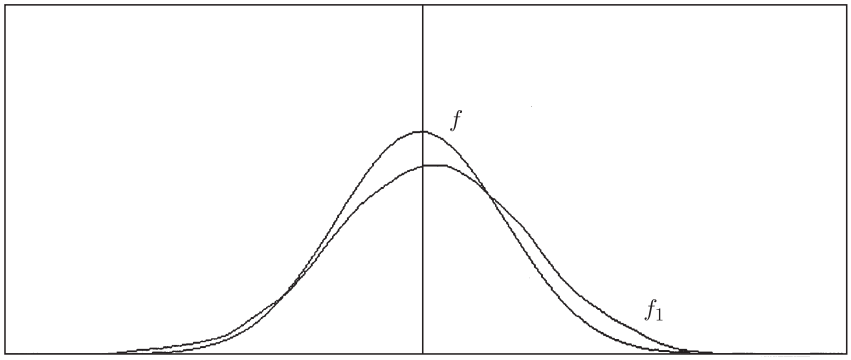

Fig 2. Estimation of the normal $N(0,1)$ density f: the first iteration f_1 of the algorithm ($h_1 = 0.71$)

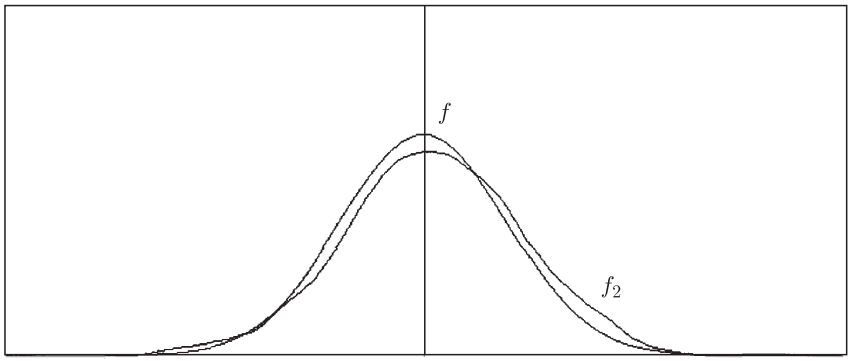

Fig 3. Estimation of the normal $N(0,1)$ density f: the second iteration f_2 of the algorithm ($h_2 = 0.57$)

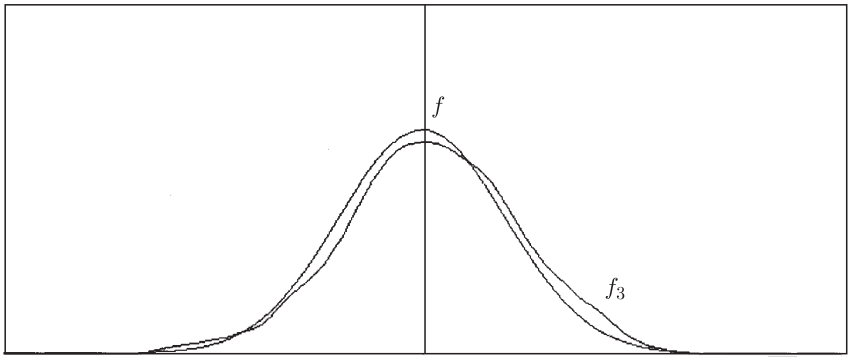

Fig 4. Estimation of the normal $N(0,1)$ density f: the third iteration f_3 of the algorithm ($h_3 = 0.54$)

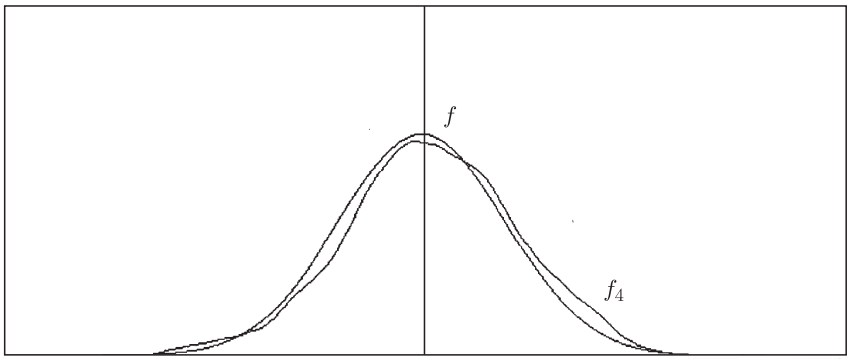

Fig 5. Estimation of the normal $N(0,1)$ density f: the fourth iteration f_4 of the algorithm ($h_4 = 0.51$)

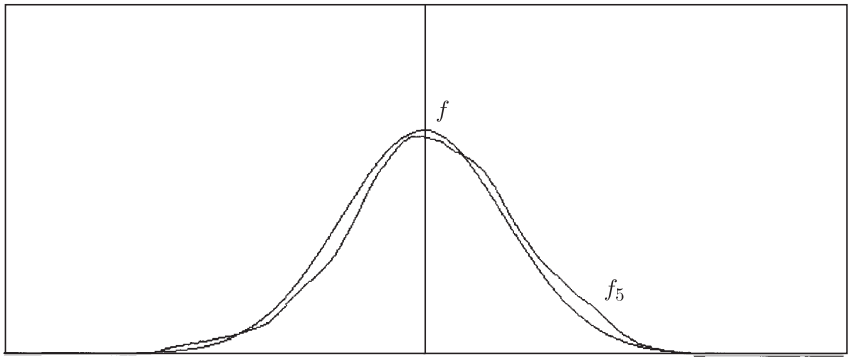

Fig 6. Estimation of the normal $N(0,1)$ density f: the final iteration f_5 of the algorithm ($h_5 = 0.50$)

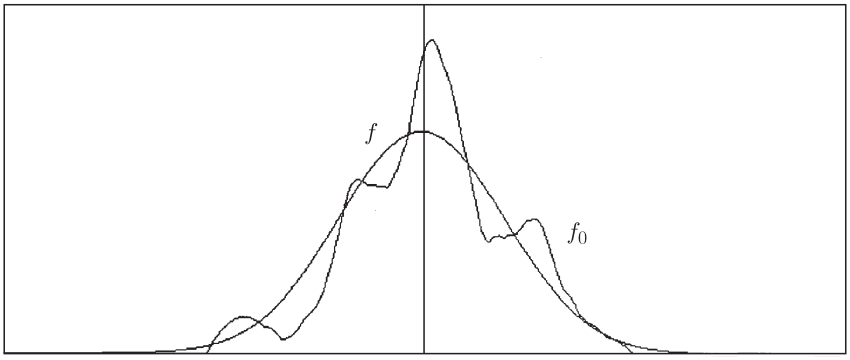

Fig 7. Estimation of the normal $N(0,1)$ density f. The initial density f_0 is the kernel estimator with $h_0 = 2$.

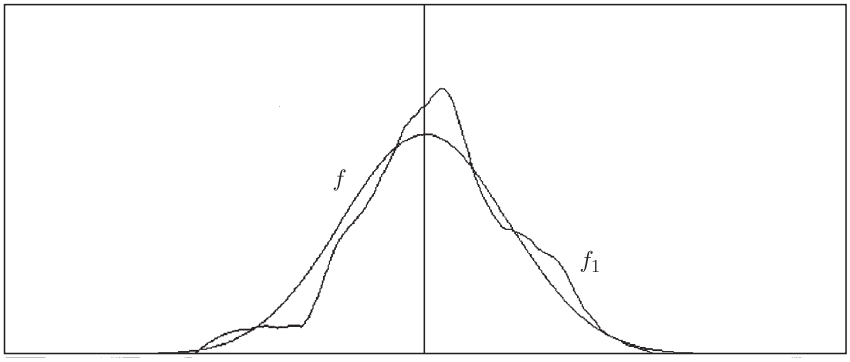

Fig 8. Estimation of the normal $N(0,1)$ density f: the first iteration f_1 of the algorithm ($h_1 = 0.29$)

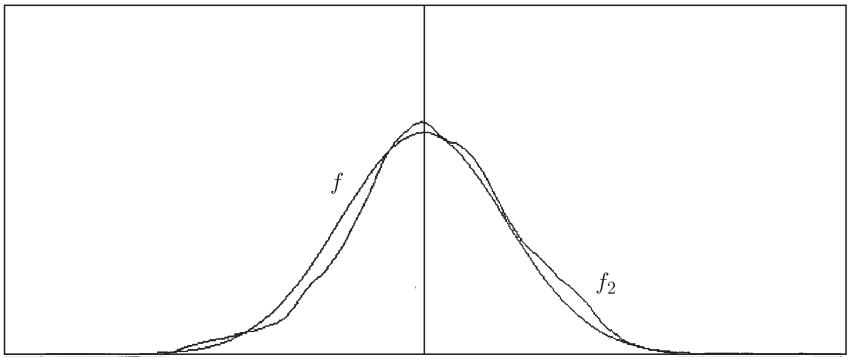

Fig 9. Estimation of the normal $N(0,1)$ density f: the second iteration f_2 of the algorithm ($h_2 = 0.42$)

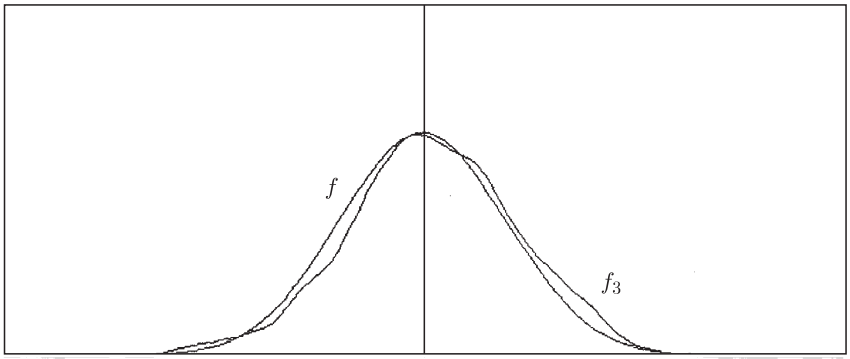

Fig 10. Estimation of the normal $N(0,1)$ density f: the final iteration f_3 of the algorithm ($h_3 = 0.48$)

An analogous comparison is provided for the bimodal mixture of normal distributions when the initial value $h_0 = 1$ (Table 3, Figures 11–17), and when $h_0 = 0.1$ (Table 4, Figures 18–20). As in the case of normal distribution, the algorithm leads to a relatively stable choice of bandwidth in a few iterations.

TABLE 3

h_i	ISE(h_i)
1	$111201 \cdot 10^{-6}$
0.76	$732165 \cdot 10^{-7}$
0.65	$537741 \cdot 10^{-7}$
0.50	$301820 \cdot 10^{-7}$
0.43	$217184 \cdot 10^{-7}$
0.39	$179538 \cdot 10^{-7}$
0.36	$157337 \cdot 10^{-7}$

TABLE 4

h_i	ISE(h_i)
1	$111201 \cdot 10^{-6}$
0.1	$323469 \cdot 10^{-7}$
0.25	$128924 \cdot 10^{-7}$
0.33	$140739 \cdot 10^{-7}$

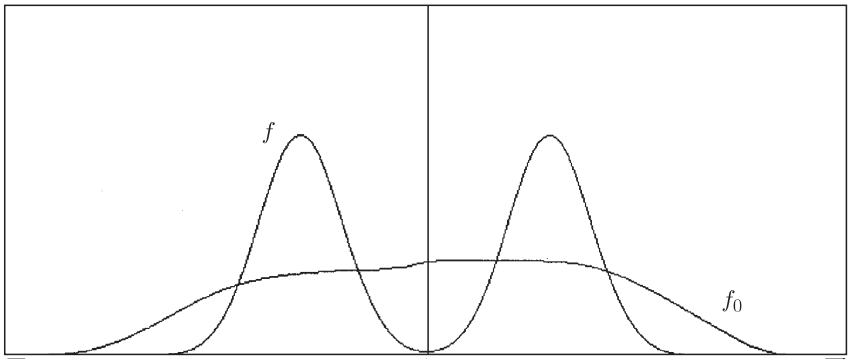

Fig 11. Estimation of the bimodal mixture f of normal distributions. The initial density f_0 is the kernel estimator with $h_0 = 1$.

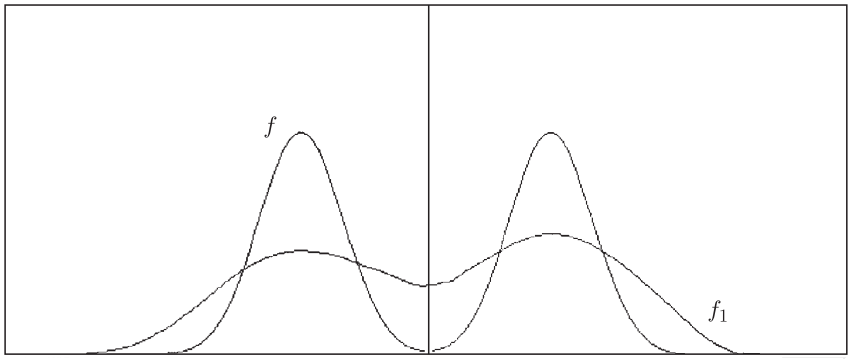

Fig 12. Estimation of the bimodal mixture f of normal distributions: the first iteration f_1 of the algorithm ($h_1 = 0.76$)

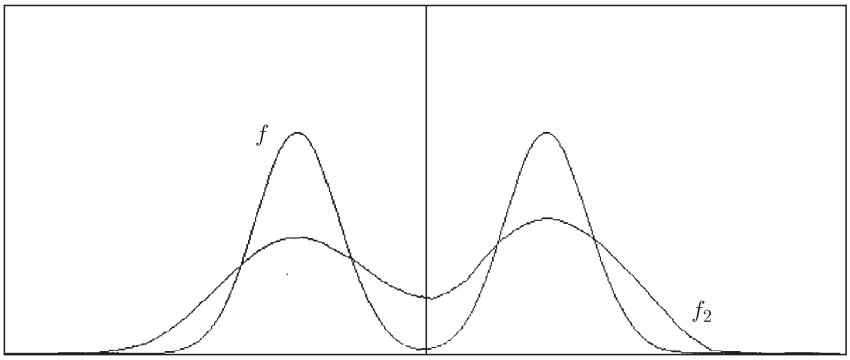

Fig 13. Estimation of the bimodal mixture f of normal distributions: the second iteration f_2 of the algorithm ($h_2 = 0.65$)

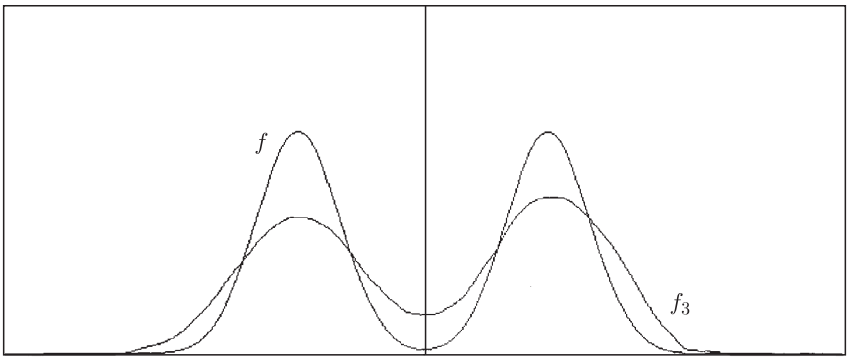

Fig 14. Estimation of the bimodal mixture f of normal distributions: the third iteration f_3 of the algorithm ($h_3 = 0.50$)

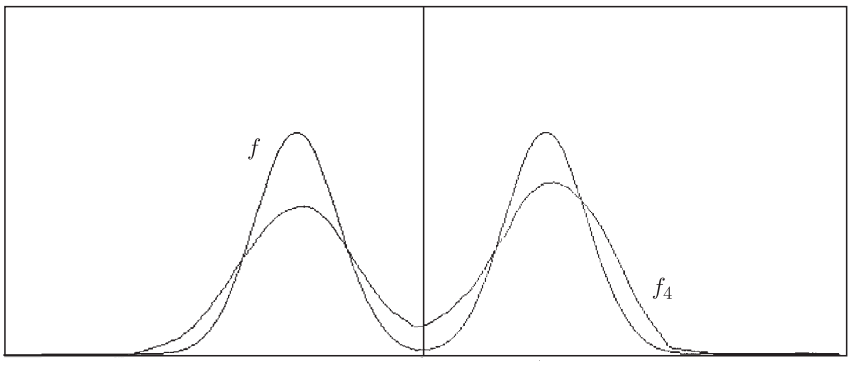

Fig 15. Estimation of the bimodal mixture f of normal distributions: the fourth iteration f_4 of the algorithm ($h_4 = 0.43$)

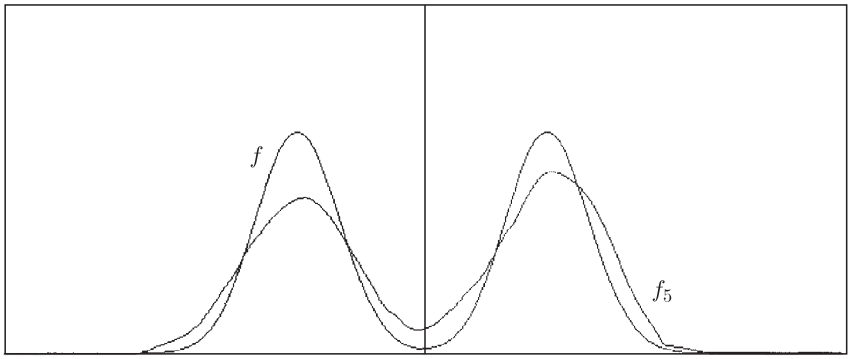

Fig 16. Estimation of the bimodal mixture f of normal distributions: the fifth iteration f_5 of the algorithm ($h_5 = 0.39$)

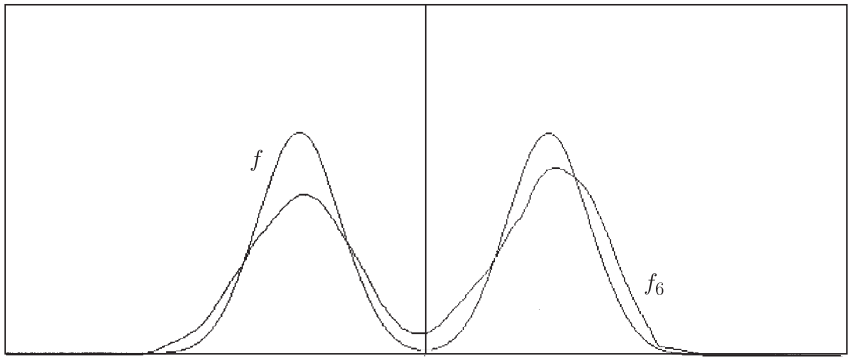

Fig 17. Estimation of the bimodal mixture f of normal distributions: the sixth iteration f_6 of the algorithm ($h_6 = 0.36$)

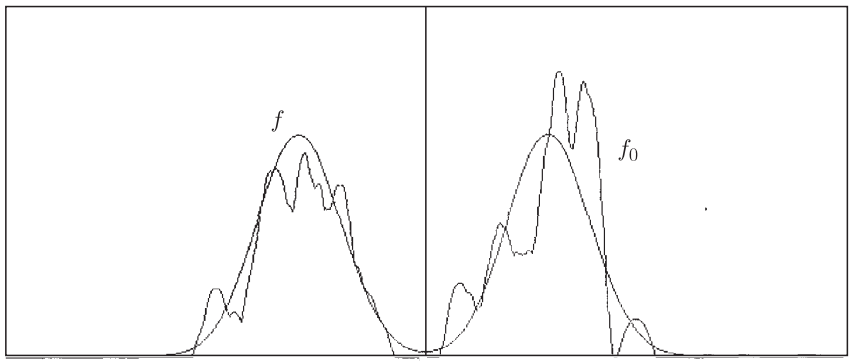

Fig 18. Estimation of the bimodal mixture f of normal distributions. The initial density f_0 is the kernel estimator with $h_0 = 0.1$.

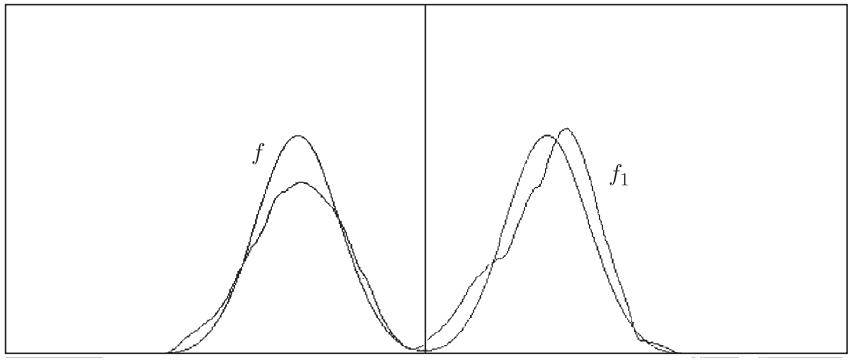

Fig 19. Estimation of the bimodal mixture f of normal distributions: the first iteration f_1 of the algorithm ($h_1 = 0.25$)

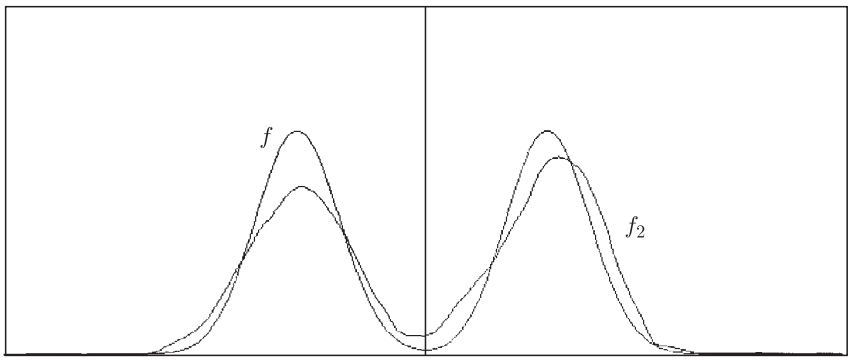

Fig 20. Estimation of the bimodal mixture f of normal distributions: the final iteration f_2 of the algorithm ($h_2 = 0.33$)

For the beta$(3, 5)$ distribution we have started the algorithm with the initial density f_0 which is chosen to be uniform, according to some "prior" knowledge (which is purposely very far from the truth). Though the data are not taken into account at the first step at all, the algorithm leads in a few steps (see Table 5) to a very precise kernel estimator (Figures 21 and 22).

TABLE 5

h_i	ISE(h_i)
1	$131453 \cdot 10^{-5}$
0.99	$131058 \cdot 10^{-5}$
0.53	$979875 \cdot 10^{-6}$
0.29	$493590 \cdot 10^{-6}$
0.17	$137402 \cdot 10^{-6}$
0.11	$492820 \cdot 10^{-7}$
0.08	$518094 \cdot 10^{-7}$

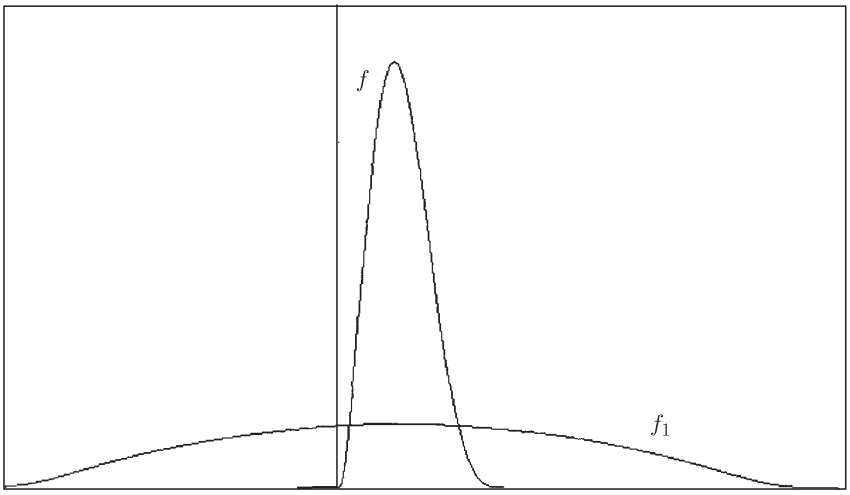

Fig 21. Estimation of the beta$(3,5)$ distribution f when the initial density f_0 is uniform: the first iteration f_1 of the algorithm ($h_1 = 1$)

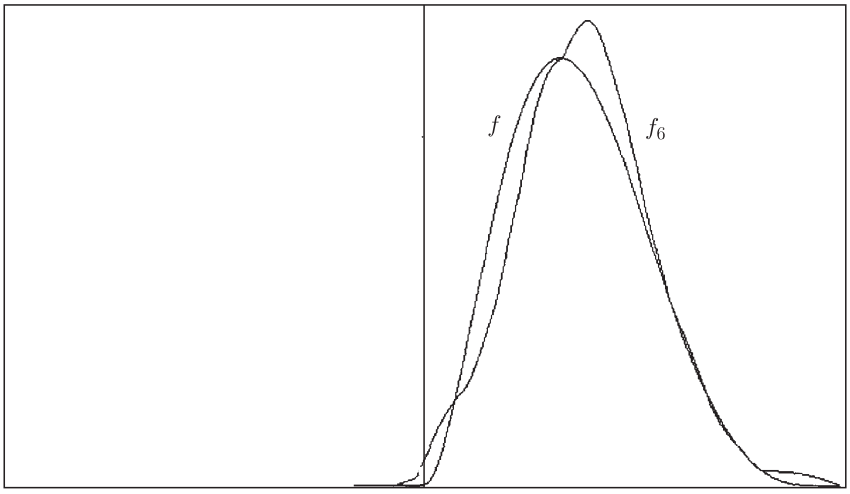

Fig 22. Estimation of the beta$(3,5)$ distribution f: the final iteration f_6 of the algorithm ($h_6 = 0.08$)

Whether the moment assumptions are important for the convergence of the algorithm was investigated in the case of the Cauchy distribution. As in the previous cases, the algorithm needs only a few iterations to find a stable bandwidth value both for $h_0 = 1$ (Table 6 and Figures 23 and 24) and for $h_0 = 0.1$ (Table 7 and Figures 25 and 26).

TABLE 6

h_i	ISE(h_i)
1	$133317 \cdot 10^{-7}$
0.88	$115277 \cdot 10^{-7}$
0.76	$108783 \cdot 10^{-7}$
0.67	$114030 \cdot 10^{-7}$
0.63	$119358 \cdot 10^{-7}$

TABLE 7

h_i	ISE(h_i)
0.1	$822912 \cdot 10^{-7}$
0.15	$515694 \cdot 10^{-7}$
0.21	$344089 \cdot 10^{-7}$
0.35	$208800 \cdot 10^{-7}$
0.50	$148847 \cdot 10^{-7}$
0.58	$128417 \cdot 10^{-7}$
0.59	$126408 \cdot 10^{-7}$
0.61	$122654 \cdot 10^{-7}$
0.63	$119358 \cdot 10^{-7}$
0.66	$115205 \cdot 10^{-7}$

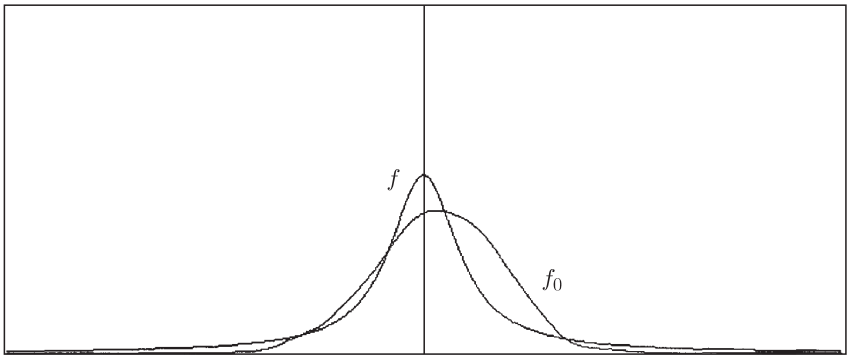

Fig 23. Estimation of the Cauchy density f. The initial density f_0 is the kernel estimator with $h_0 = 1$.

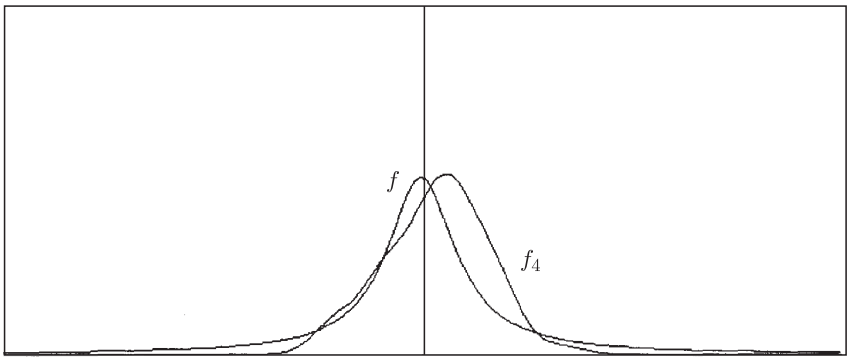

Fig 24. Estimation of the Cauchy density f: the final iteration f_4 of the algorithm ($h_4 = 0.63$)

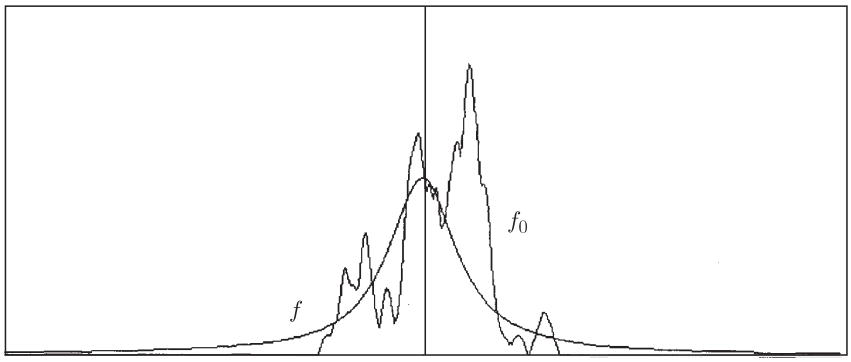

Fig 25. Estimation of the Cauchy density f. The initial density f_0 is the kernel estimator with $h_0 = 0.1$.

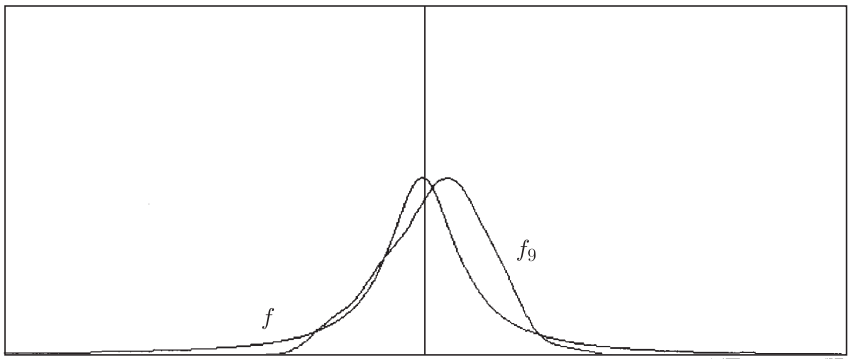

Fig 26. Estimation of the Cauchy density f: the final iteration f_9 of the algorithm ($h_9 = 0.66$)

In Table 8 we have compared the bandwidth values resulting from the self-learning algorithm with the optimal ones which were found from one hundred bootstrap samples, taken from the underlying densities. It shows that in each case the algorithm stops very close to the optimal bandwidth value $\widehat{h}$.

TABLE 8

Distribution	Self-learning algorithm	Optimal bandwidth
Standard normal	0.48–0.50	0.49
bimodal normal	0.36	0.29
Beta$(3, 5)$	0.08	0.10
Cauchy	0.63–0.66	0.59

Repeating the above experiments we have observed that the selected bandwidth $h^*(X_1, \ldots, X_n)$ differs from sample to sample by a small percentage. For larger sample sizes one should increase the size m of the bootstrap samples because the random fluctuation of the bootstrap AISE can influence the stability of the algorithm too much.

Finally, we have applied the self-learning algorithm to the eruption lengths of 107 eruptions of the Old Faithful geyser (Silverman (1986)). Following Silverman (1986), we have used the Gaussian kernel here. Surprisingly enough, the resulting bandwidth was $h = 0.21$ which is quite close to Silverman's subjective choice $h = 0.25$. Figure 27 shows the kernel estimator for $h_0 = 1$. The kernel estimator resulting from the self-learning algorithm is shown in Figure 28 while the one corresponding to $h = 0.25$ is given in Figure 29.

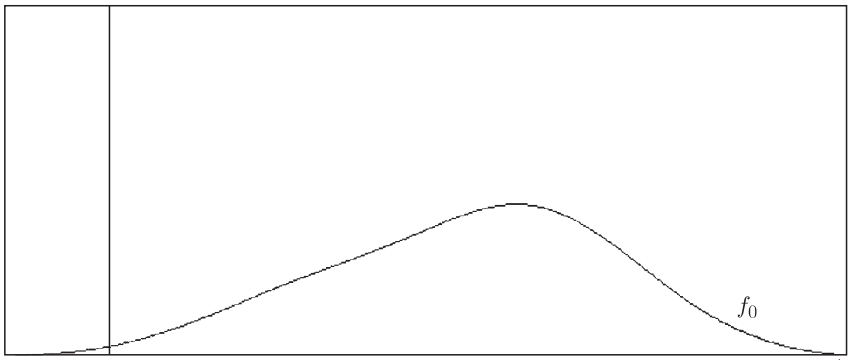

Fig 27. Estimation of the density of the eruption length of 107 eruptions of the Old Faithful geyser. The initial density f_0 is the kernel estimator with $h_0 = 1$.

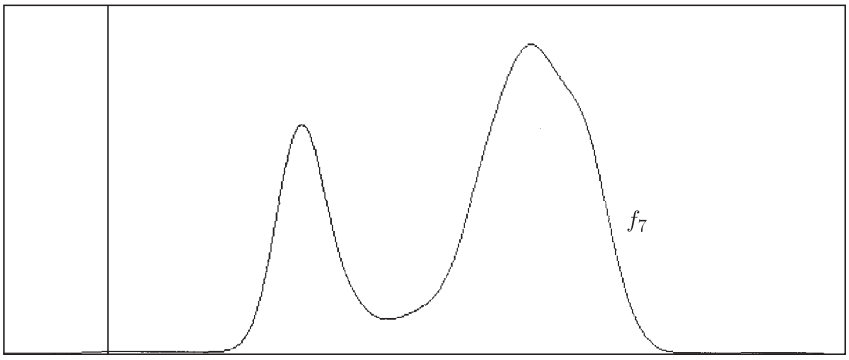

Fig 28. Estimation of the density of the eruption length of 107 eruptions of the Old Faithful geyser: the estimator f_7 resulting from the algorithm has bandwidth $h_7 = 0.25$.

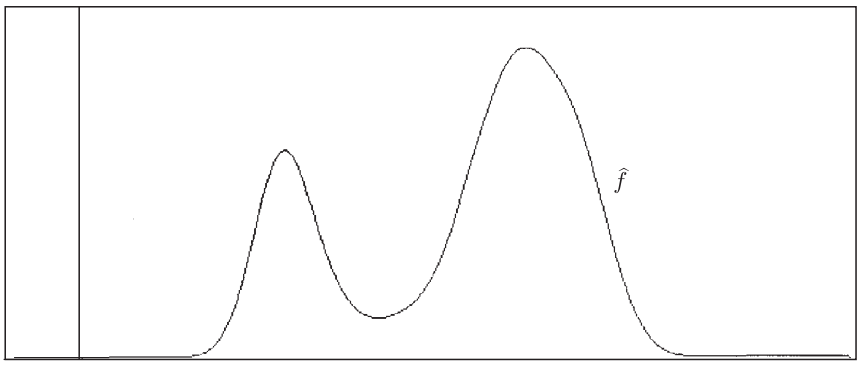

Fig 29. Estimation of the density of the eruption length of 107 eruptions of the Old Faithful geyser: $\widehat{f}$ is the kernel estimator with $h = 0.25$.

References

J. J. Faraway and M. Jhun (1990), *Bootstrap choice of bandwidth for density estimation*, J. Amer. Statist. Assoc. 85, 1119–1122.

W. Härdle, P. Hall and J. S. Marron (1988), *How far are automatically chosen regression smoothing parameters from their optimum? (with comments)*, ibid. 74, 105–131.

C. Léger, D. N. Politis and J. P. Romano (1992), *Bootstrap technology and applications*, Technometrics 43, 378–398.

E. Parzen (1962), *On estimation of a probability density function and mode*, Ann. Math. Statist. 33, 1065–1076.

M. Rosenblatt (1956), *Remarks on some nonparametric estimates of a density function*, ibid. 27, 832–837.

B. W. Silverman (1986), *Density Estimation for Statistics and Data Analysis*, Chapman and Hall, London.

C. C. Taylor (1989), *Bootstrap choice of the smoothing parameter in kernel density estimation*, Biometrika 76, 705–712.

INSTITUTE OF MATHEMATICS
POLISH ACADEMY OF SCIENCES
ŚNIADECKICH 8
00-950 WARSZAWA, POLAND

INSTITUTE OF MATHEMATICS
TECHNICAL UNIVERSITY OF ŁÓDŹ
AL. POLITECHNIKI 11
90-924 ŁÓDŹ, POLAND

Received on 24.6.1993;
revised version on 10.9.1993