Journal of Official Statistics
Vol. 2, No. 2, 1986, pp. 193–202

Some Difficulties Involving Nonparametric Estimation of a Density Function[1]

Yadolah Dodge[2]

Abstract: The object of this study is do demonstrate some of the difficulties involving the nonparametric estimation of a density function. Via an extensive simulation study it is shown that a) the optimal asymptotic value of the scale factor h in Parzen's method of estimation is in many cases significantly far from the optimal simulated ones under five different distributions, b) the optimal asymptotic value of the scale factor holds for the normal distribution, c) the value of the scale factor changes substantially at different points of a distribution, d) by increasing the sample size the value of the scale factor decreases slowly, e) there is no difference in using $L1$ or $L2$ norm in determining the optimal value of the scale factor.

Key words: Nonparametric estimation; density estimation; kernel density estimation; simulation.

1. Introduction and Summary

There are problems in estimating an $f(x)$ when one has a sequence of independent identically distributed random variables $X_1, X_2, \ldots, X_n$ with an unknown common probability density function $f(x)$. Parzen (1962) gave an effective method in which one selects a kernel function $K(x) \geq 0$ such that:

$$\int K(x)\, dx = 1.$$

After selecting a $K(x)$, one can estimate the density function by:

$$\hat{f}(x) = 1/n \sum_j 1/h\, K[(x - x_j)/h]. \qquad (1.1)$$

However, choosing the scale factor h for a given kernel is not at all easy. There are many asymptotic results on how h should be selected in order to obtain the best estimate of the density. As suggested by Parzen, h must be a function of n so that as n tends to infinity h tends to 0 and $nh \to \infty$. But it is obvious that the optimal h depends on $f(x)$.

The purpose of this paper is to study the goodness of the scale factor h in practical situations and, in particular, to find out:

a) how the optimal asymptotic value of h varies as a function of x in different densities in comparison with practical situation,

b) on the basis of $L1$ and $L2$ norms, to find the optimal value of h for which the expected distance between $\hat{f}$ and f is minimized, and

[1] This research is supported by the Swiss National Fund, project number 2.843–0.80.
[2] Groupe d'informatique et de statistique, Université de Neuchâtel, Switzerland.
The authors would like to thank the referees and the associate editor for their helpful comments.

c) the effect of sample size on the optimal asymptotic value of h in contrast with simulated ones ?

To answer these questions, an extensive simulation study was carried out. Using five distributions, namely : normal, gamma, two mixed normal and Cauchy. For each distribution at, say, point x_k the values of $L1$ and $L2$ norms were calculated for various values of h and finally the results were plotted as $(h, L1)$ and $(h, L2)$. For each x_k, the experiment was repeated 200 times and we obtained the mean value of h at the point where minimum values of $L1$ and $L2$ occurred. The effect of sample size on the asymptotic and simulated h were carried out only on the normal distribution. For other distributions, a sample size of 100 was selected. A number of points x_k was chosen according to each distribution.

It is shown that the optimal asymptotic value of h in most cases is significantly far from the simulated ones, and this value changes dramatically at different points of each distribution. Moreover, both $L1$ and $L2$ norms show exactly the same h at their minimums. We show this last result only for the normal distribution.

2. Asymptotic Results

The estimation of a probability density function dates back to Karl Pearson (1902 a, b) who attempted to estimate the probability density function by computing the sample moments.

For the historical developments and a bibliography on the topic see Wegman (1972), Tapia and Thompson (1978), Wertz and Schneider (1979), and Bean and Tsokos (1980). Rosenblatt's (1956) initial paper seems to be a pioneering paper to extend the concept of histogram as a method for density estimation into modern nonparametric density estimation.

A nonparametric density estimation is a procedure which selects a function f_n within the set of real valued functions to estimate a density function without imposing any restrictons regarding a parametrized subset.

Rosenblatt (1956) proposed an estimator $f_n(x)$ of the form:

$$f_n(x) = (F_n(x+h) - F_n(x-h))/2h$$

where

$$F_n(x) = (\text{number of observations} \leq x)/n.$$

Parzen (1962) considered the general setting of Rosenblatt's estimate. He introduced a kernel estimate as given in (1.1) assuming:

$$\int_{-\infty}^{\infty} |K(u)|\, du < \infty ,$$

$$\sup_{-\infty<u<\infty} |K(u)| < \infty ,$$

$$\lim_{|u|\to\infty} |u\,K(u)| = 0 ,$$

and

$$K(u) \geq 0 , \int_{-\infty}^{\infty} K(u)\, du = 1.$$

Thus Rosenblatt's estimate corresponds to $K(x) = 1/2h$ for $|x|<h$ and 0 elsewhere. Given f and using Taylor expansions Parzen found h which minimizes the asymptotic mean square error at point x as :

$$h = \alpha(K)\,(f(x)/f''(x))^{1/5} n^{-1/5} .$$

Choosing

$$K(y) = 15/16\,(1-y^2)^2 , |y| \leq 1$$

we have $\alpha(K) = 2.0362$ and

$$h = 2.0362\,(f(x)/f''(x)^2)^{1/5}\, n^{-1/5} . \qquad (1.2)$$

In this paper we attempt to compare the exact optimal value of h with the optimal asymptotic one. It is well known (see for instance Anderson (1969)) that the shape of the kernel K is of secondary importance as far as estimation is concerned.

3. The Monte Carlo Study

Five different parameter families are employed in the simulation study to yield the distributions : normal, gamma, Cauchy and two mixed normal. Of these distributions, three are symmetric and one asymmetric. The Cauchy distribution is a long tailed distribution which produces outliers and the gamma distribution, on the other hand, is an asymmetric distribution. The two mixed normal distributions are bimodal, one is with equal weight (symmetric) and the other with unequal weight (asymmetric).

Random samples were generated from each distribution by applying the probability integral transformation

$$X_i = F^{-1}(U_i)$$

to independent uniform (0,1) variates U_i. The simulated random samples were generated on VAX/VMS version 2.4 Neuchâtel University in Switzerland.

The sample size was fixed at 100 observations for each distribution at each considered point x, except for the normal distribution in which we studied the effect of the sample size on the scale factor h. Each experiment was repeated 200 times independently at each point x. At each point we estimated $\hat{f}$ by f from formula (1.1) and for a given point we incremented the value of simulated h (denoted by h_s) from 0.05 to an appropriate value so that the minimum of $L1$ and $L2$ occurred in that range. The optimal asymptotic value of h (denoted by h_a) is obtained by (1.2) (see Tapia and Thompson p. 59 formula 138).

The measures of error were Mean Square Error (MSE) and Mean Absolute Error (MAE) known as pointwise $L2$ and $L1$ norms respectively. In each experiment the value of f is calculated, then the deviations are averaged over the 200 replications, i.e.,

$$L1 = 1/200 \sum_{i=1}^{200} |f - \hat{f}|$$

and

$$L2 = 1/200 \sum_{i=1}^{200} (f - \hat{f})^2 .$$

To measure the effect of the sample size on the optimal h_s, we generated random samples of the sizes 10, 20, 50, 100, 200, 400, 600, 800, 1 000, and 2 000 at point $x = 0$ (mean) of the normal distribution with variance 1, repeating the experiment 200 times. For each experiment we recorded the values of h_s and h_a along with the values of $L1$ and $L2$ on a given point x.

We shall begin the discussion on our findings with the normal distribution.

3.1. Normal Distribution With Mean 0 and Variance 1

A total of 14 points were chosen for the normal distribution. Table 3.1 gives the values of x, the minimum values of $L1$ and $L2$ and the corresponding values of h_s and h_a. As can be seen from this table, except for the inflexion point (x=1) there is no significant difference between the h_s and h_a at all other points. Notice that because the normal distribution is symmetric about 0, we only took points on the right side of the mean. As we move from the point 0 to 3.05 (over three standard deviations) the minimum values of $L1$ and $L2$ decrease. The average values of h_s and h_a are 1.38 respectively. It seems that 1 is a reasonable value to be chosen for h given a normal distribution and a sample size of 100.

Table 3.1. Minimum values of L1 and L2, asymptotic and simulated values of h at different points for normal (0,1) distribution

Point x	$L1$	$L2$	simulated h	asymptotic h
0.00	0.03567	0.00182	0.95	0.97
0.05	0.03425	0.00184	0.95	0.98
0.65	0.02708	0.00113	1.25	1.27
0.85	0.01980	0.00063	1.50	1.75
0.95	0.01589	0.00041	1.75	2.71
1.00	0.01567	0.00038	1.90	∞
1.05	0.01212	0.00023	2.15	2.71
1.15	0.00889	0.00013	2.80	1.75
1.75	0.01834	0.00057	0.80	0.99
1.95	0.01740	0.00050	1.05	0.94
2.00	0.01663	0.00047	0.95	0.93
2.15	0.01500	0.00036	0.90	0.92
2.45	0.01106	0.00020	1.10	0.93
3.05	0.00481	0.00003	0.60	1.06
		Total	18.65	17.91
		Mean	1.33	1.38

Table 3.2. Minimum values of L1 and L2, asymptotic and simulated values h at the mean point for normal (0,1) distribution with different sample sizes

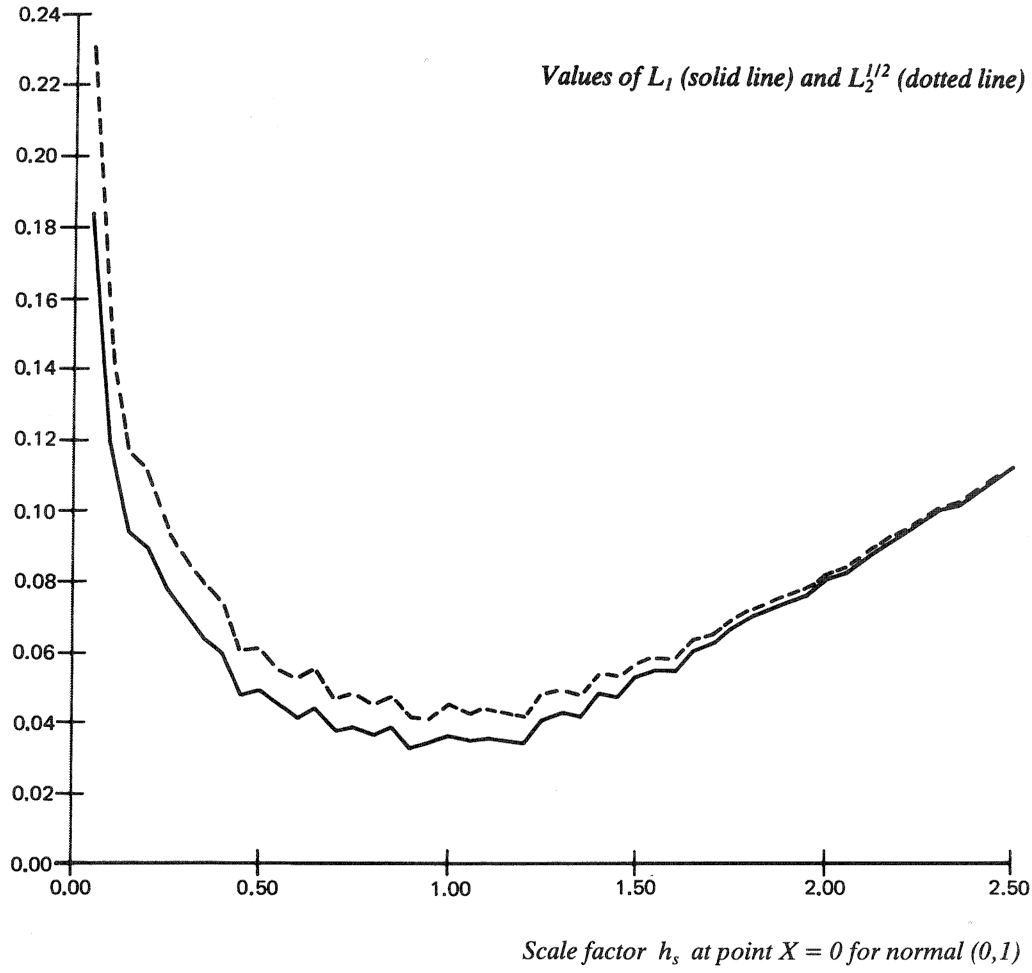
Sample size	$L1$	$L2$	h_s	h_a
10	0.06920	0.00717	1.75	1.54
30	0.05172	0.00382	1.30	1.24
50	0.04542	0.00296	1.30	1.12
100	0.03567	0.00182	0.95	0.97
200	0.02685	0.00118	0.85	0.85
400	0.02128	0.00073	0.80	0.74
600	0.01977	0.00060	0.70	0.68
800	0.01780	0.00047	0.70	0.64
1000	0.01639	0.00040	0.65	0.61
2000	0.01171	0.00021	0.50	0.54

The values of L1, L2 and simulated h were obtained averaging over 200 replications.

To see the effect of the sample size on h, independent random samples of sizes 10, 30, 50, 100, 200, 400, 600, 800, 1 000 and 2 000 were generated from the normal distribution. As before, each experiment was repeated 200 times. Table 3.2 gives the sample sizes, values of $L1$ and $L2$, and h_a and h_s at point $x = 0$.

It can be seen that even when the sample size is 10, there is no significant difference between h_a and h_s. The values of h_s range from 1.75 for sample size 10 to 0.50 for the sample sizes which confirms the slow rate of decrease of h in (1.2). Another interesting result is that the minimum values of $L1$ and $L2$ norms are attained for the same value of h in the simulation study. Figure 3.1 shows the range of h_s at point $x = 0$ for the sample size of 100. Also note that the values of $L1$ and $L2$ agree for almost every value of h_s. (The dotted line is the square root of $L2$ and the solid line is the value of $L1$). We plotted this graph to demonstrate how the optimal h_s is attained at a given point.

Fig. 3.1. Variation of L_1 and $L_2^{1/2}$ expected errors at the mode of the standard normal distribution as a function of h.
Sample size: 100.



3.2. *Gamma distribution*

For the gamma distribution we considered :

$$f(x) = x\, e^{-x}$$

In case 1 a total of 18 points were chosen on the x axis. In this case, only for the first five points, namely $x = 0.05, 0.65, 0.95, 1.0$, and 1.05 the optimal asymptotic and the simulated h show similarities. As x increases above the mean, h_s also increases so that at point $x = 4.0$ the value of h_s reaches 11.70.

The optimal asymptotic value of h as can be seen from Table 3.3. first increases up to the point $x = 2.0$ and then starts to decrease, so that at point 4.0 this value falls to 1.80. Note that the gamma distribution is asymmetric and its tail produces some outliers. For this reason, we have studied the variation of h on a wide range of x's. In this case the great variability of h_s should be a warning in using the fixed h when we have a long tailed distribution (h_s changes from 0.20 to 11.70 as x changes from 0.05 to 4.0). The mean value of h_s is 2.63 and for h_a is 1.94, but such values may not be of practical use due to great variations within the range of x for both h_s and h_a.

Table 3.3. Minimum values of L2, asymptotic and simulated values of h at different points for gamma (2,1) distribution

Point x	$L2$	simulated h	asymptotic h
0.05	0.00148	0.20	0.34
0.65	0.00252	0.70	0.75
0.95	0.00214	0.90	0.95
1.00	0.00172	0.85	0.99
1.05	0.00193	1.05	1.03
1.15	0.00173	0.90	1.12
1.45	0.00103	1.20	1.48
1.65	0.00071	1.45	1.90
1.75	0.00058	1.65	2.24
1.95	0.00039	2.00	4.54
2.00	0.00036	2.05	∞
2.05	0.00031	2.35	4.67
2.15	0.00017	2.35	3.10
2.35	0.00012	2.55	2.34
2.45	0.00009	2.80	2.18
3.05	0.00001	5.25	1.83
3.45	0.00000	7.35	1.78
4.00	0.00000	11.70	1.80
	Total	47.3	33.04
	Mean	2.63	1.94

3.3. Cauchy Distribution

The Cauchy distribution is a long tailed distribution with no mean or variance. It is symmetric about 0 and for this reason we took a total of 13 points on the right side of $x = 0$ (including 0). As can be seen from Table 3.4, h_s increases as x increases, while h_a at first increases and then decreases on the same range of x. At point $x = 2.05$ where one expects some outliers the value of h_s reaches 12.95, but its corresponding h_a is only 1.51. The mean value of h_s is 3.69 and for h_a is 1.49. Such a difference even in the mean value shows how **poor the asymptotic optimal value of h is when we have a symmetric but flat tailed distribution. The variation of h_s on the range of x (0 to 2.05) is from 0.80 to 12.95 and h_a varies on the same range of x from 0.77 to 1.51.**

3.4. Normal Mixed Distribution

Two cases were considered :

1) 0.25N(0,1) + 0.75N(4,1) and
2) 0.50N(0,1) + 0.50N(3,1).

In case 1 we have a bimodal asymmetric distribution and in case 2 we have a bimodal symmetric distribution. For case 1 we took a total of 10 points on the x axis ranging from 3.0 to 6.0 (with the exception of $x = 0$ at which we showed the variation of $L1$ and $L2$ versus h_s.) In the situation with two modes, h_s and h_a agree in most cases, except for the inflexion point and the first two points $x = 3.0$ and $x = 3.05$. Table 3.5 shows the values of x , minimum values of $L1$, $L2$, h_s and h_a respectively. The average value of h_s is 1.7 and for h_a is 1.9. The difference is due to the first two points.

Table 3.4. Minimum values of L2, asymptotic and simulated values of h at different points for Cauchy distribution

Point x	$L2$	simulated h	asymptotic h
0.00	0.00226	0.80	0.77
0.05	0.00223	0.90	0.78
0.51	0.00096	1.40	1.78
0.55	0.00086	1.30	2.61
0.61	0.00064	1.40	2.51
0.65	0.00060	1.65	1.86
0.71	0.00042	2.00	1.51
0.95	0.00013	2.80	1.19
1.05	0.00008	3.50	1.16
1.15	0.00006	4.25	1.16
1.45	0.00002	6.65	1.23
1.65	0.00001	8.40	1.31
2.05	0.00000	12.95	1.51
	Total	48	19.38
	Mean	3.69	1.49

Table 3.5. Minimum values of L2, asymptotic and simulated values of h at different points for 0.25 N (0,1)+0.75N (4,1) distribution

Point x	$L2$	simulated h	asymptotic h
0.0	0.00048	1.25	1.29
3.0	0.00025	2.35	3.82
3.05	0.00035	2.25	3.58
3.50	0.00108	1.20	1.19
4.00	0.00115	1.20	1.03
4.15	0.00126	1.10	1.04
4.80	0.00059	1.60	1.66
4.90	0.00052	1.80	2.17
5.0	0.00032	2.20	∞
5.10	0.00021	2.45	2.17
	Total	18.65	18.94
	Mean	1.7	1.9

Table 3.6. Minimum values of L2, asymptotic and simulated values of h at different points for 0.5 N (0,1)+0.5N (3,1) distribution.

Point x	$L2$	simulated h	asymptotic h
0.0	0.00080	1.30	1.16
1.50	0.00074	1.35	1.12
1.65	0.00076	1.15	1.14
2.00	0.00002	4.55	1.51
2.05	0.00003	4.45	1.65
2.20	0.00008	3.50	2.94
2.25	0.00014	3.05	7.83
2.35	0.00030	2.40	2.15
2.75	0.00085	1.65	1.25
3.00	0.00088	1.45	1.16
3.90	0.00034	2.00	2.41
4.00	0.00023	2.75	8.41
4.05	0.00018	3.05	3.02
5.00	0.00019	0.95	1.07
6.00	0.00002	0.95	1.19
	Total	34.65	38.01
	Mean	2.31	2.5

In case 2 the difference starts from point $x = 2.0$, where h_s is 4.55 and h_a is 1.51. At point $x = 2.25$ the situation reverses so that h_a becomes 7.83 while h_s remains at 3.05. This matter will be repeated at point $x = 4.0$ where h_s becomes 2.75 and h_a reaches 8.41. The average value of h_s is 2.31 and h_a is 2.5, they do not significantly differ from each other. These variations are presented in Table 3.6.

4. Concluding Remarks

The poor performance of the optimal asymptotic value of h was demonstrated in an extensive simulation study for Cauchy (long tailed and symmetric) and gamma (long tailed but asymmetric) distributions. The optimal asymptotic h performed in accordance with the simulated h under the normal distribution. The ∞ values in the tables are artifacts of both the truncation of Taylor series used to develop equation 1.2 and that higher order (nonzero) terms involving the fourth derivate of f could have been included in the Taylor expansion. Overall flat tails are the cause of not only poor asymptotic approximation but also of strong variation of the optimal value of h along the support. It can be speculated that this is related to the high variance of the estimator because of the scarcity of points observed in such tails. If the experimenter wishes to use the fixed kernel as opposed to the variable one, a larger value of h larger than the optimal value at the mode may be required to reduce the error over the support. It is evident from the foregoing simulation study that there is a need for a variable h that performs well for any distribution and more specifically that h should be increased in the tails, a result already observed by Breiman et al. (1977). When the number of observations reaches 2 000 in the case of the normal distribution with mean 0 and variance 1, the optimal value of h reaches 0.54 which shows a slow convergence of h as n goes towards infinity. The bias of the optimal h theory and the effects of sample size are also depicted by Scott (1985). It is also shown in this study that $L1$ and $L2$ norms in fact lead to essentially the same optimal values of h.

5. References

5.1. References Cited in the Text

Anderson, G.D. (1969) : A Comparison of Methods for Estimating a Probability Density Function. Ph.D. Dissertations, University of Washington.

Bean, S.J. and Tsokos, C.P. (1980) : Developments in Nonparametric Density Estimation. International Statistical Review, 48, pp. 267–287.

Breiman, L., Meisel, W., and Purcell, E. (1977) : Variable Kernel Estimates of Multivariate Densities. Technometrics, 19, pp. 135–144.

Parzen, E. (1962) : On the Estimation of a Probability Density Function and the Mode. Annals of Mathematical Statistics, 40, pp. 1065–1076.

Pearson, K. (1902a) : On the Systematic Fitting of Curves to Observations and Measurements, I Biometrika 1, pp. 265–303.

Pearson, K. (1902b) : On the Systematic Fitting of curves to Observations and Measurements, II Biometrika 2, pp. 1–23.

Rosenblatt, M. (1956) : Remarks on Some Nonparametric Estimates of a Density Function. Annals of Mathematical Statistics, 27, pp. 832–837.

Scott, D.W. (1985) : Frequency Polygons : Theory and Applications. Journal of the American Statistical Association, 80, pp. 348–354.

Tapia, R.A. and Thompson, J.R. (1978) : Nonparametric Probability Density Estimation. Johns Hopkins University Press, Baltimore, Maryland.

Wegman, E.J. (1972) : Nonparametric Density Estimation : I. A Summary of Available Methods. Technometrics 14, pp. 533–546.

Wertz, W. and Schneider, B. (1979) : Statistical Density Estimation: A Bibliography. International Statistical Review, 47, pp. 155–175.

5.2. References Not Cited in the Text

Manija, G.M. (1974) : Statistical Estimation of Probability Distribution. Russian. Publishing House of Tbilisi University, Tbilisi.

Wertz, W. (1978) : Statistical Density Estimation. Vandenhoeck and Ruprecht in Göttingen Publication.

Received September 1984
Revised January 1986