Monte Carlo Study of Three Data-Based Nonparametric Probability Density Estimators

DAVID W. SCOTT and LYNETTE E. FACTOR*

Although the theoretical properties of modern nonparametric probability density estimators have been studied for 25 years, there remains the practical problem of how to specify the amount of bias or smoothing in a density estimate based on a random sample. In this paper we review and evaluate three recently developed data-based algorithms that completely specify a density estimate from a random sample. Using Monte Carlo techniques, we compare the statistical accuracy of these algorithms as measured by the integrated mean squared error. In addition, we examine the sensitivity of these algorithms to outliers and estimate computer time requirements. One conclusion we draw is that the statistical accuracy of these data-based algorithms seems comparable to levels predicted by theoretical models.

KEY WORDS: Nonparametric probability density estimation; Data-based smoothing; Kernel estimator; Orthogonal series estimator; Monte Carlo; Data analysis.

1. INTRODUCTION

During the past 25 years the theoretical development of nonparametric probability density estimation has been rapid. Several different approaches have received extensive treatment: kernel methods (Rosenblatt 1956, Parzen 1962, Watson and Leadbetter 1963, Cacoullos 1966, and Epanechnikov 1969), orthogonal series methods (Kronmal and Tarter 1968, Watson 1969, Wahba 1977, and Brunk 1978), penalized-likelihood methods (Good and Gaskins 1971, de Montricher, Tapia, and Thompson 1975, Scott, Tapia, and Thompson 1980), k-nearest neighbor methods (Loftsgaarden and Quesenberry 1965), Bayesian-spline methods (Whittle 1958, Boneva, Kendall, and Stefanov 1971, Wahba 1976, and Stewart 1979) and maximum-likelihood or histogram-like methods (Robertson 1967, Wegman 1970, and Van Ryzin 1973). See Tapia and Thompson (1978) for a recent and more complete bibliography. An important observation is that each of these methods has a single parameter that controls the amount of bias or smoothness in the corresponding density estimate. For the classical histogram, the bin width plays the role of this smoothing parameter—too great a bin width implies large bias and an oversmoothed estimate, while too small a bin width implies large variance and too rough an estimate.

The great potential of these nonparametric density estimators in data analysis is not being fully realized, primarily because of the practical difficulty associated with choosing the smoothing parameter given only data $x_1, x_2, \ldots, x_n$. From a theoretical viewpoint, the choice of an optimal smoothing parameter is known; however, knowledge of the true underlying sampling density f is required. As a result, various ad hoc methods for choosing the smoothing parameter have been suggested. For example, for the histogram, various authors (Guttman and Wilks 1965) suggest dividing the sample range into 10 to 25 bins. Although the ad hoc procedures work reasonably well, their successful application has required the skills of statisticians, who understand the many ways in which the ad hoc approaches may fail.

Recently, several data-based algorithms for choosing a smoothing parameter have been proposed. By a data-based algorithm we mean an algorithm that can be embodied in a computer subroutine whose input is the data and whose output is a value of the smoothing parameter that is approximately equal to the theoretically optimal, but unknown, value of the smoothing parameter. Moreover, we require that the algorithm have this approximation property for a large class of true densities. Most ad hoc procedures can be computer based but, in general, do not have good large-sample or small-sample properties.

In this paper we present a Monte Carlo simulation designed to compare three data-based algorithms and evaluate their small-sample properties. The first algorithm was proposed by Duin (1976) and also by Hermans and Habbema (1976) and is based on a kernel density estimator. The second algorithm is also based on a kernel density estimator and is attributable to Scott, Tapia, and Thompson (1977). The third algorithm is based on an orthogonal series estimator and is attributable to Wahba (1977). A parametric maximum likelihood Gaussian estimate is also included as a benchmark. For convenience, we shall refer to the first two algorithms by the initials DHH and STT, respectively. We begin by presenting a

* David W. Scott is Associate Professor, Department of Mathematical Sciences, Rice University, Houston, TX 77001, and Lynette E. Factor is a graduate student at Rice University. Some of the numerical results in this paper were part of the second author's Master's thesis. This research was supported in part by the National Heart, Lung, and Blood Institute, NIH, through the National Heart and Blood Vessel Research and Demonstration Center, Grant 17269, while the authors were at Baylor College of Medicine, Houston, TX 77030. The authors wish to thank the referees and editors for their helpful suggestions.

© Journal of the American Statistical Association
March 1981, Volume 76, Number 373
Applications Section

brief description of the estimators and data-based algorithms.

2. THE DATA-BASED ALGORITHMS

2.1 The Kernel Estimator

Given a kernel function K, which is a probability density function symmetric about zero, a positive smoothing parameter h, and a sample $x_1, x_2, \ldots, x_n$, the kernel estimate of f at each fixed point x is

$$\hat{f}(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right). \tag{2.1}$$

The kernel estimate is nonnegative and integrates to one. Since the shape of the true density is of most interest, a relevant criterion is the integrated mean squared error (IMSE) (Rosenblatt 1971, p. 1817) defined as

$$\begin{aligned} \text{IMSE} &= \int E[\hat{f}(x) - f(x)]^2 \, dx \\ &= E \int [\hat{f}(x) - f(x)]^2 \, dx, \end{aligned} \tag{2.2}$$

which, for a kernel estimate, may be expressed as

$$\begin{aligned} &\frac{1}{nh} \int K(y)^2 \, dy \\ &\quad + \tfrac{1}{4}h^4 \left[\int K(y) y^2 \, dy\right]^2 \cdot \int f''(x)^2 \, dx + 0\left(\frac{1}{n}\right) \end{aligned} \tag{2.3}$$

(Parzen 1962). Minimizing over the first two terms in expression (2.3), we obtain the well-known asymptotically optimal choice for the smoothing parameter h, which depends on the true density f, the choice of a kernel, and the sample size

$$h = \alpha(K) \cdot \beta(f) \cdot n^{-1/5} \tag{2.4}$$

where

$$\alpha(K) = \left[\int K(y)^2 \, dy\right]^{1/5} \cdot \left[\int K(y) y^2 \, dy\right]^{-2/5} \tag{2.5}$$

and

$$\beta(f) = \left[\int f''(x)^2 \, dx\right]^{-1/5}. \tag{2.6}$$

With the choice (2.4), the IMSE decreases in proportion to $n^{-4/5}$. Extensive numerical work has demonstrated that this choice of h is, on the average, nearly optimal for samples as small as $n = 25$ (Scott, Tapia, and Thompson 1977). Early work focused on the choice of an optimal kernel function K; however, it is now known that many symmetric unimodal kernel functions are nearly optimal. The Gaussian kernel is popular, although finite support kernels have less computational overhead. We shall use the Gaussian kernel exclusively in this paper.

2.1.1 The Data-Based Algorithm of Duin and Hermans and Habbema (DHH). Consider the following problem with a maximum likelihood criterion for choosing h:

$$\underset{h \geq 0}{\text{maximize}} \; L(h) = \prod_{k=1}^{n} \hat{f}(x_k). \tag{2.7}$$

From (2.1) it may be seen that $h = 0$ maximizes $L(h)$, corresponding to an estimate with a Dirac function at each sample point. Thus Duin (1976) and Hermans and Habbema (1976) were led to consider problem (2.7) with a slightly modified maximum-likelihood criterion:

$$\underset{h \geq 0}{\text{maximize}} \; L'(h) = \prod_{k=1}^{n} \hat{f}_k(x_k), \tag{2.8}$$

where

$$\hat{f}_k(x_k) = \frac{1}{nh} \sum_{\substack{i=1 \\ i \neq k}}^{n} K\left(\frac{x_k - x_i}{h}\right). \tag{2.9}$$

The modified maximum likelihood approach has been shown to be related to a certain jackknife procedure (Wong 1979). We shall denote a solution to problem (2.8) by h_D.

2.1.2 The Data-Based Algorithm of Scott, Tapia, and Thompson (STT). Scott, Tapia, and Thompson (STT) (1977) developed an iterative algorithm based on the theoretical choice (2.4) for h. For a particular choice of h, an estimate of $\beta(f)$ is available, namely $\beta(\hat{f}_h)$, where the subscript emphasizes the value of the smoothing parameter. Given an initial choice $h(0)$, they form the sequence

$$h(i+1) = \alpha(K) \cdot \beta(\hat{f}_{h(i)}) n^{-1/5}, \tag{2.10}$$

where

$$\begin{aligned} \beta(\hat{f}_h)^{-5} = \frac{3}{8\sqrt{\pi} n^2 h^9} \sum_{j=1}^{n} \sum_{k=1}^{n} &[h^4 - (x_j - x_k)^2 h^2 \\ &+ \tfrac{1}{12}(x_j - x_k)^4] \cdot \exp[-(x_j - x_k)^2 / 4h^2] \end{aligned} \tag{2.11}$$

for a Gaussian kernel estimate. The problem of the convergence of the sequence (2.10) is well defined since the solution $h = 0$ always exists. They choose the largest nonnegative solution of problem (2.10). It is possible to show that choosing $h(0)$ to be the sample range guarantees convergence to the largest solution, which we shall denote by $h_S = h(\infty)$. If the largest solution is $h_S = 0$, we shall call the corresponding Dirac spike estimate *degenerate*.

2.2 The Series Estimator and Wahba's Data-Based Algorithm

Wahba considers an estimator based on the Fourier series expansion of a true density f with support on the interval [0, 1]

$$f(t) \sim 1 + \sum_{\substack{\nu = -\infty \\ \nu \neq 0}}^{\infty} f_\nu \varphi_\nu(t), \tag{2.12}$$

where $\varphi_\nu(t) = \exp(2\pi i \nu t)$ and the coefficients f_ν are estimated by

$$\hat{f}_\nu = \frac{1}{n} \sum_{j=1}^{n} \varphi_\nu^*(x_j). \tag{2.13}$$

Here φ^* denotes the complex conjugate of φ.

There are two distinct approaches to smoothing the series estimator. The first approach includes a relatively small number of terms in expansion (2.12). Cencov (1962) considers truncating high frequency terms, while Tarter and Kronmal (1976) recommend an additional inclusion rule. The second approach, used by Watson (1969), Wahba (1977), and Brunk (1978), is to apply a low-pass window to the series expansion. Wahba's estimate has the form, letting $\sum_* = \sum_{\nu=-n/2,\nu\neq 0}^{n/2}$,

$$\hat{f}(t) = 1 + \sum_* \frac{\hat{f}_\nu}{1 + \lambda(2\pi\nu)^4}\varphi_\nu(t), \tag{2.14}$$

where λ is the smoothing or window-shaping parameter. It is convenient to assume that n is even for Wahba's algorithm. This estimate integrates to one but is not nonnegative in general. In practice, the finite support requirement is not restrictive, since nonparametric estimates far out in the tails are poor. (Using other orthogonal functions, Wahba (1978) has relaxed the finite support requirement.) We have chosen the exponent of four in the denominator of (2.14) (Wahba allows for any positive even integer) because some recent work of Byrd, Tapia, and Thompson (1978) has shown that the resultant denominator is the optimal smoothing window for a related spectral density estimation problem. The IMSE also decreases at the rate $n^{-4/5}$.

Wahba (1978) observed that the theoretical integrated mean squared error of her estimate can be written as the sum of two terms

$$\frac{1}{n-1}E\varphi(\lambda \mid x_1, \ldots, x_n) + 0(n^{-4}), \tag{2.15}$$

where

$$\varphi(\lambda \mid x_1, \ldots, x_n) = \sum_* \left(\frac{\lambda}{\lambda_\nu + \lambda}\right)^2 (n \mid \hat{f}_\nu \mid^2 - 1) + \left(\frac{\lambda_\nu}{\lambda_\nu + \lambda}\right)^2 (1 - \mid \hat{f}_\nu \mid^2), \tag{2.16}$$

where $\lambda_\nu = (2\pi\nu)^{-4}$. The data-based choice for λ minimizes $\varphi(\lambda \mid x_1, \ldots, x_n)$ and will be denoted by λ_1.

Wahba (1977) originally presented a data-based choice λ_2 that minimizes

$$\varphi'(\lambda) = n\left[\sum_* \frac{\mid \hat{f}_\nu \mid^2}{(\lambda_\nu + \lambda)^2}\right] \cdot \left[\sum_* \frac{1}{\lambda_\nu + \lambda}\right]^{-2} \tag{2.17}$$

and is obtained by a curve-fitting argument. We shall use λ_2 for the Monte Carlo work and comment on the apparent relationship of λ_1 and λ_2.

3. MONTE CARLO SIMULATION RESULTS

To evaluate the performance of these data-based algorithms, we designed two Monte Carlo experiments. For the first, 25 pseudorandom standard Gaussian samples were generated by using subroutine GGNMP (International Mathematical and Statistical Libraries, Inc. 1979) for sample sizes of 50 and 100. The data-based algorithms were applied to each of the generated samples. To estimate the error, the integrated squared error (ISE) defined by

$$\text{ISE} = \int [\hat{f}(x) - f(x)]^2\,dx \tag{3.1}$$

was evaluated numerically by Simpson's rule, using a mesh over the interval $(-5, 5)$ with increments of a tenth. Averaging the ISE obtained for the 25 repetitions gives an estimate of the IMSE (see Eq. (2.2)). We also computed the standard deviation of the 25 ISE estimates.

In Table 1 we have summarized the results of this Monte Carlo experiment. As may be expected, the parametric maximum likelihood approach has a lower estimated IMSE than the more general nonparametric estimates. For the nonparametric algorithms, there are no significant differences among the IMSE estimates. However, the ordering DHH-Wahba-STT is observed in both. Since the true sampling density is known in this experiment, we may compute the theoretically optimal smoothing parameter for a kernel estimate by using (2.4) and the resulting integrated mean squared error by using (2.3). For $n = 50$, the optimal choice is $h = .48$, and for $n = 100$, $h = .42$, with resulting theoretical IMSE's given by .0146 and .0084, respectively. The averages of the data-based choices for h are very close to these optimal choices. The data-based estimates seem to perform at least as well as the predicted optimal integrated mean squared error. We are unaware of any corresponding closed-form formula for Wahba's method, but her method has the same asymptotic rate of convergence as the kernel method.

The second experiment (see Table 2) paralleled the first except that a mixture of Gaussians $\frac{1}{2}N(-\frac{3}{2}, 1) + \frac{1}{2}N(\frac{3}{2}, 1)$ was used. We chose to focus on a bimodal density rather than a heavy-tailed density because nonparametric techniques are well suited for multimodal data but, as noted before, not particularly efficient in the tails. To demonstrate the danger of incorrectly specifying a parametric density form, we also fit a single Gaussian to the mixture data. Not surprisingly, the use of the wrong parametric form leads to substantially larger IMSE's than does the nonparametric approach. For $n = 100$ DHH performed statistically significantly better than the other methods.

Table 1. Results of Data-Based Algorithms for Gaussian Data: Estimated IMSE for 25 Repetitions

Sample Size	Method	Degenerate Cases	Smoothing Parameter Mean (Std. Dev.)	ISE Mean (Std. Dev.)
50	$N(\bar{x}, s^2)$	0	– (–)	.0042 (.0036)
50	DHH	0	.47 (.15)	.0130 (.0103)
50	STT	2	.43 (.13)	.0158 (.0110)
50	Wahba	1	.000070 (.000054)	.0132 (.0115)
100	$N(\bar{x}, s^2)$	0	– (–)	.0028 (.0031)
100	DHH	0	.40 (.11)	.0070 (.0056)
100	STT	0	.38 (.09)	.0074 (.0054)
100	Wahba	0	.000044 (.000029)	.0071 (.0067)

Table 2. Results of Data-Based Algorithms for Gaussian Mixture: Estimated IMSE for 25 Repetitions

Sample Size	Method	Degenerate Cases	Smoothing Parameter Mean (Std. Dev.)	ISE Mean (Std. Dev.)
50	$N(\bar{x}, s^2)$	0	– (–)	.0189 (.0072)
50	DHH	0	.64 (.17)	.0104 (.0045)
50	STT	1	.64 (.22)	.0113 (.0042)
50	Wahba	1	.000041 (.000042)	.0116 (.0047)
100	$N(\bar{x}, s^2)$	0	– (–)	.0169 (.0023)
100	DHH	0	.58 (.09)	.0056 (.0030)
100	STT	1	.50 (.16)	.0073 (.0040)
100	Wahba	0	.000028 (.000031)	.0069 (.0035)

Again, for the kernel method we have that the optimal smoothing parameters are .66 and .58 for $n = 50$ and 100, respectively, with corresponding IMSE's given by .0106 and .0061. The data-based algorithms have error rates similar to the optimal rate predicted by repeated use of a fixed smoothing parameter.

4. MONTE CARLO IMPLEMENTATION REMARKS

We now comment on the degenerate cases listed in Tables 1 and 2. Using STT resulted in $h_S = 0$ for four of the 200 samples, corresponding to a Dirac solution. Another simulation reported using this technique (Scott, Tapia, and Thompson 1977) found that the degenerate cases occurred at the same rate. The estimates of the IMSE are based on the nondegenerate cases only. The nondegenerate estimates of h_S were computed to four significant digits, using Newton's method.

For application of Wahba's algorithm, the data summarized in Tables 1 and 2 were linearly transformed from the intervals $[-3, 3]$ and $[-5, 5]$ to the interval $[0, 1]$, respectively. Data values outside these intervals were trimmed. The average of the data-based smoothing parameters λ_2 for these scaled data are given in the tables; however, we rescaled each $[0, 1]$-density estimate back to the original intervals to compute the IMSE. For Gaussian data, use of the intervals $[-3, 3]$, $[-6, 6]$ (i.e., too wide) and $[-9, 3]$ (i.e., not symmetric about zero) for rescaling Wahba's estimate gave similar IMSE results. Thus the precise support interval is apparently not important.

We used a discrete line search to find the approximate minimizer of (2.17) suggested by Wahba (1977, p. 446), namely, $\lambda_2 = 10^{(k-21)/3}$ for integers $0 \le k \le 18$. The final estimate was obtained by perturbing λ_2 using the factor $10^{-1/12}$. We also called an estimate degenerate if the best λ_2 corresponded to the extreme values $k = 0$ or $k = 18$. This happened in only two of the 200 samples generated and not with the same samples that led to degenerate cases for STT.

We also remark that we found only small differences in the data-based estimates of λ using criteria (2.16) and (2.17). Both criteria occasionally had a local minimum in addition to a global minimum for values of λ in the range given previously. These samples had 25 points, and we have not seen any multiple local minima for larger sample sizes; see, however, Section 7.

For DHH a similar but linear line search was used to find h_D to an accuracy of .01. No multiple maxima were observed with the criterion function in problem (2.8).

5. SENSITIVITY ANALYSIS

On the average, the three data-based methods resulted in similar integrated mean squared error levels. These levels were comparable to the theoretically optimal IMSE levels of kernel estimates. We believe that no one data-based algorithm should be accepted over all others. Rather, we suggest constructing several data-based estimates for the same data so that the occasional poor estimates may be eliminated. This level of subjectivity is similar to that encountered in linear regression, where the researcher reserves the right to examine the result for the possibility of outlier effects. Thus we designed a sensitivity analysis not to eliminate an algorithm but to provide better insight into the pecularities of each data-based algorithm.

Specifically, we focused on the effect of a single outlier point x in a fixed data set of 25 standard Gaussian points. For values of x ranging from -7 to 7, we recomputed the smoothing parameter of each algorithm for the 26 samples. We defined an insensitive data-based algorithm as one that produced values of the smoothing parameter that were nearly independent of the value of x. In Figure 1 we have plotted a trace of the values of the smoothing parameter for each of the three algorithms as x ranges from

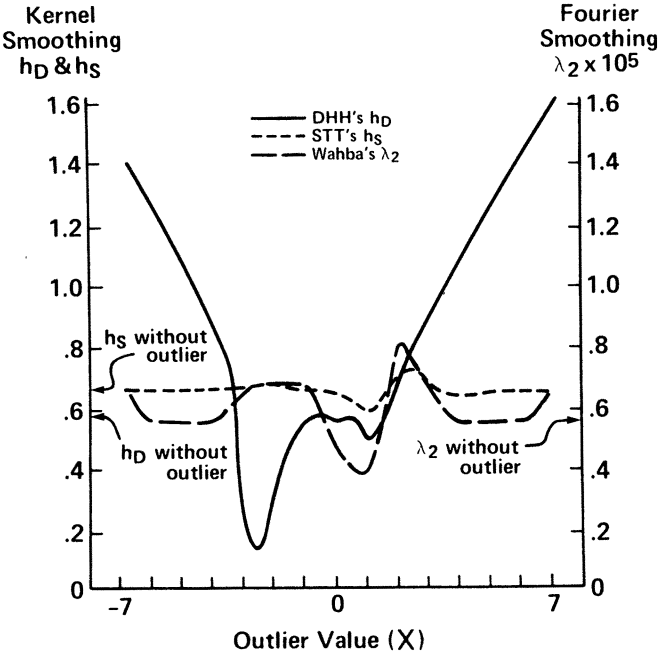

Figure 1. Sensitivity Analysis for the Smoothing Parameters of the Three Algorithms in the Presence of an Outlier

-7 to 7. For STT the trace was nearly constant, with only some insignificant variation, and differed only slightly from the value $h_S = .67$ obtained for the original 25 points. On the other hand, it was clear that DHH was quite sensitive and that the addition of a 26th point substantially changed the smoothing parameter from $h_D = .58$ for the original data. In particular, h_D increased linearly for values of the outlier outside the interval $(-4, 2)$, corresponding to increasingly oversmoothed estimates. Surprisingly, for $x = -2.5$, $h_D = .14$, corresponding to a very rough estimate. Examination of the original 25 data points revealed the presence of a single sample point at -2.5. Thus duplicate outliers presented still another problem for DHH. However, in spite of the sensitivity, DHH performed well on the average in the Monte Carlo simulation. Attempts to correct the problem by trimming one or two data points in each tail resulted in rough estimates. Wahba's algorithm, for which the data were linearly rescaled from the interval $[-7, 7]$ to $[0, 1]$, was also quite insensitive to outliers. The trace shown varies insignificantly from $\lambda_2 = .000056$ for the original data. Notice that the scale for Wahba's method is not directly comparable with the kernel parameter scale.

6. ESTIMATING COMPUTER RUNTIME REQUIREMENTS

We measured the CPU time required for each algorithm on a DEC 10 with a KI processor and 512K core memory. For standard Gaussian samples with $n = 10, 25, 50, 100,$ and 200, we recorded the amount of CPU time in seconds required, first, to find the smoothing parameter and second, to evaluate the corresponding estimate 101 times over the mesh required to estimate the integrated squared error in Section 3 (see Figure 2).

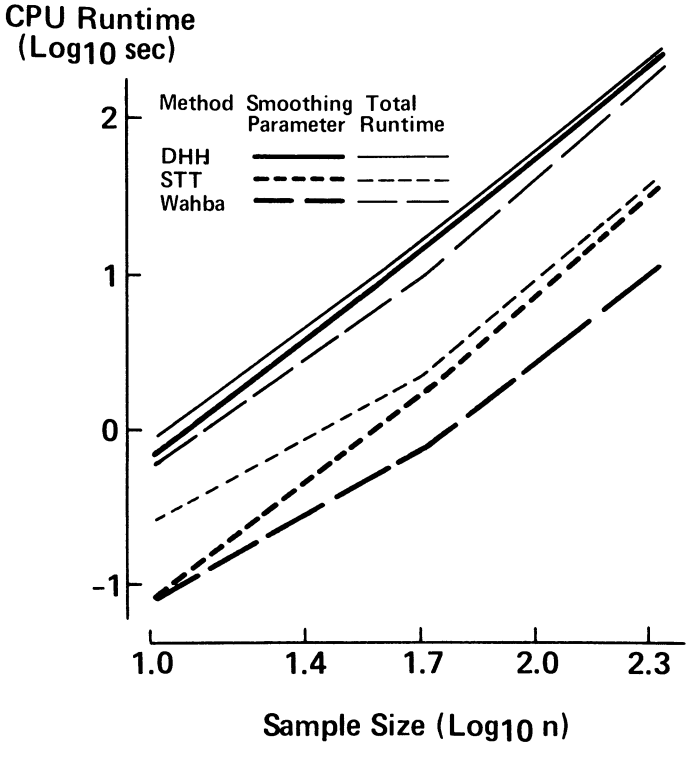

Figure 2. Computer CPU Time Required for Each Algorithm to Estimate Its Smoothing Parameter and Evaluate the Estimate 101 Times

Table 3. Bin Counts for the Brightness Data (read from left to right)

1	0	1	4	6	13	20	20	32	24	31	33	30	35	28
34	29	20	27	15	19	11	10	9	15	9	8	15	9	8
11	12	10	6	10	10	12	6	12	4	7	6	7	5	1
3	5	0	5	3	3	2	2	5	3	3	0	2	4	0
0	4	0	1	1	1	2	1	0	1					

Wahba's algorithm required the least CPU time to find a smoothing parameter; however, the evaluation of the corresponding density was more expensive than for the kernel estimate. To find the kernel-smoothing parameter, STT took significantly less time than DHH. Overall, STT was the least expensive of the three. The asymptotic rates of increase in cost were similar for all three algorithms, increasing in proportion to the square of the sample size.

7. EXAMPLES

Some examples of the application of these data-based algorithms may be found in Scott et al. (1978), Scott (1980), and Scott et al. (1980). Here we shall examine some continuous data that have been discretized because of hardware constraints. These LANDSAT data, kindly made available by John Potter and George Terrell at Lockheed in Houston, represent a brightness component of the reflectance from corn fields. The histogram bin counts for the 686 points in 70 bins, which we shall associate without loss of generality with the values 1, 2, . . . , 70, are shown in Table 3. With discrete data, the data-based algorithms are drawn towards the (discrete) Dirac solution ($h = 0$), as can be seen in Figures 3, 4, and 5, but secondary optima give appropriately smoothed

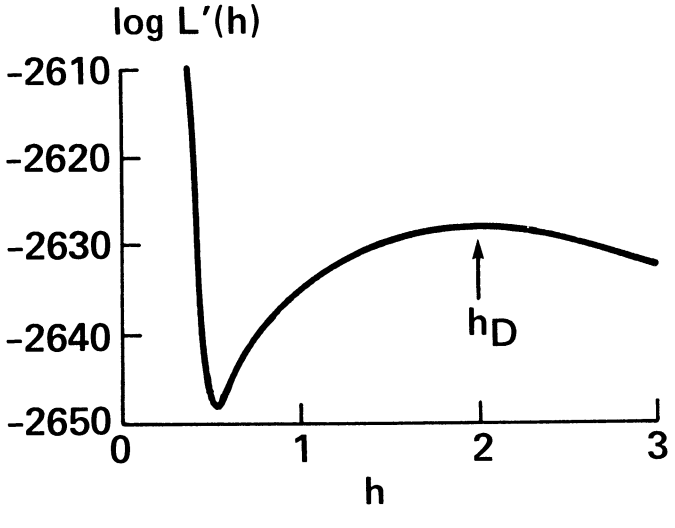

Figure 3. Modified Likelihood Function (Eq. 2.8) for the Brightness Data

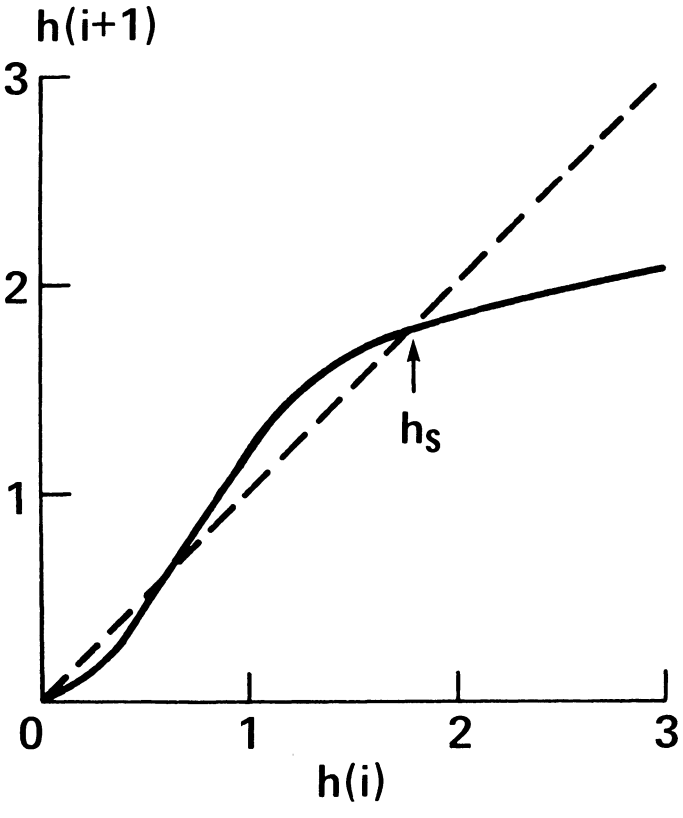

Figure 4. Equation (2.10) As a Function of h = h(i) for the Brightness Data. The Dotted Line Is h(i + 1) = h(i)

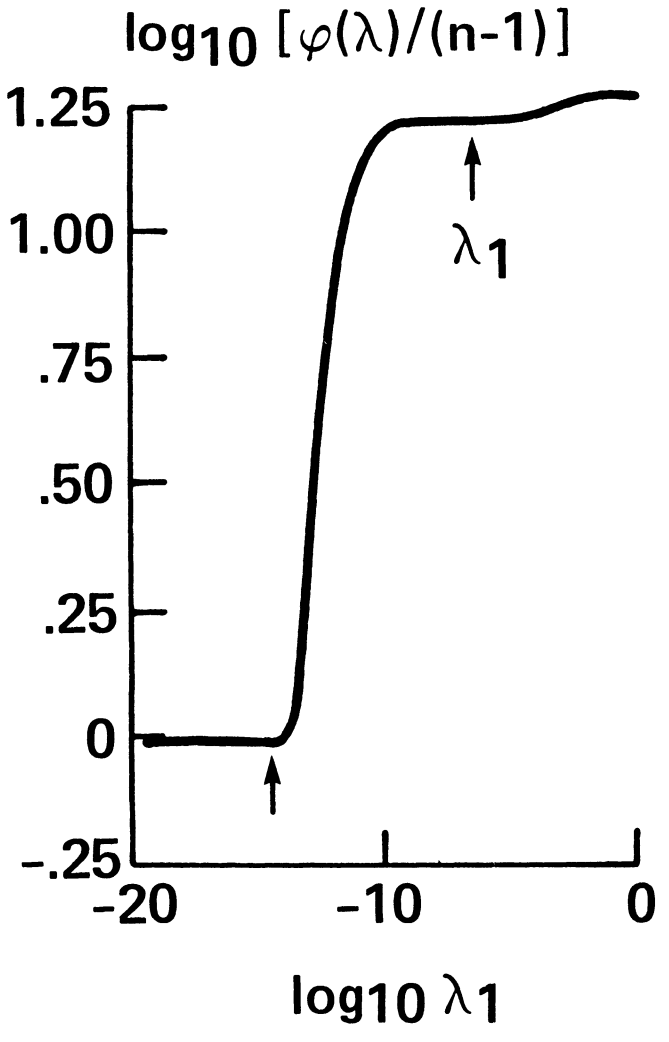

Figure 5. Wahba's Cross-Validation Function (2.16) for the Brightness Data

density estimates. For example, in Figure 3, problem (2.8) has a local maximum at $h_D = 2.0$. The global maximum occurs for h slightly greater than zero. In Figure 4, there are three fixed points, $h = 0, .50$, and 1.79; therefore, $h_S = 1.79$, the largest. For the series estimator, the data were transformed to the values $x = .16, .17, \ldots, .85$. In Figure 5, the global minimizer of (2.16) occurs for $\lambda_1 = 10^{-14.4}$, a very rough estimate comparable to a Gaussian kernel estimate with $h = .03$ on the original data scale. A secondary minimum occurs for $\lambda_1 = 10^{-6.68}$, comparable again to $h = 1.90$. Thus all three data-based estimates are virtually identical, being right skewed with an interesting bump around $x = 32$. We remark that when the data discretization is very coarse (only a few bins), the data-based algorithms may not have any secondary optima, with the Dirac-like solution resulting. Data discretization leads to reduced variance but increased bias of density estimates.

8. DISCUSSION AND SUMMARY

We compared three data-based algorithms for nonparametric density estimation, two very different algorithms for a kernel estimate, and a third algorithm for an orthogonal series estimate. Using an integrated mean squared error criterion for evaluating quality, we found that the three algorithms perform similarly on the average and quite close to the optimal level predicted for kernel

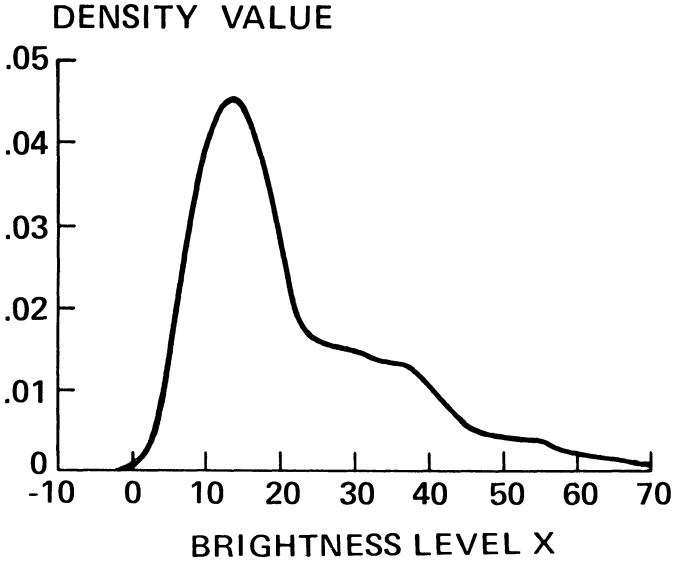

Figure 6. Gaussian Kernel Estimate of the Brightness Data With h = 2.0

density theory. Although DHH performs best, it is quite sensitive to outliers in practice. If many or large data sets are involved, then the differences in computer time required by these algorithms may be important. However, the possibility of occasionally poor estimates with DHH resulting from outliers and from degeneracy in the other algorithms suggests that a prudent and not too expensive course is to construct all three estimates and then choose any reasonable estimate consistent with prior belief about the degree of smoothness in the true sampling density.

The timing results suggest that hybrid algorithms and clever implementations may be useful since the kernel estimate is cheaper to evaluate than the orthogonal series estimate, but the kernel smoothing parameter is more expensive to obtain. In particular, it is easy to evaluate (2.6) for a Fourier series estimate. Thus a quick search gives λ and then a ready estimate for h by way of (2.6) and (2.4).

For multivariate data, the extension of DHH is straightforward and has been used in two dimensions for some practical data analysis (Scott et al. 1978). The other univariate algorithms have multivariate extensions that are under development at this time (Nezames 1980).

[*Received May 1979. Revised July 1980.*]

REFERENCES

BONEVA, L., KENDALL, D.G., and STEFANOV, I. (1971), "Spline Transformations: Three New Diagnostic Aids For The Statistical Data-analyst," *Journal of the Royal Statistical Society*, Ser. B, 20, 1–70.

BRUNK, H.D. (1978), "Univariate Density Estimation by Orthogonal Series," *Biometrika*, 65, 521–528.

BYRD, R.H., TAPIA, R.A., and THOMPSON, J.R. (1978), "Optimal Smoothing of Direct Estimates of the Power Spectrum," *Communications in Statistics—Simulation and Computation*, B7, 335–344.

CACOULLOS, T. (1966), "Estimation of a Multivariate Density," *Annals of the Institute of Statistical Mathematics*, 18, 179–189.

CENCOV, N.N. (1962), "Evaluation of an Unknown Distribution Density From Observations," *Soviet Mathematics*, 3, 1559–1562.

DE MONTRICHER, G.M., TAPIA, R.A., and THOMPSON, J.R. (1975), "Nonparametric Maximum Likelihood Estimation of Probability Densities by Penalty Function Methods," *Annals of Statistics*, 6, 1329–1348.

DUIN, R.P.W. (1976), "On the Choice of Smoothing Parameters for Parzen Estimators of Probability Density Functions," *IEEE Transactions on Computers*, C–25, 1175–1179.

EPANECHNIKOV, V.A. (1969), "Nonparametric Estimates of a Multivariate Probability Density," *Theory of Probability and Its Applications*, 14, 153–158.

GOOD, I.J., and GASKINS, R.A. (1971), "Nonparametric Roughness Penalties for Probability Densities," *Biometrika*, 58, 255–278.

GUTTMAN, I., and WILKS, S.S. (1965), *Introductory Engineering Statistics*, New York: John Wiley & Sons.

HERMANS, J., and HABBEMA, J.D.F. (1976), *Manual for the ALLOC Discriminant Analysis Programs*, University of Leiden, Dept. of Medical Statistics.

INTERNATIONAL MATHEMATICAL AND STATISTICAL LIBRARIES, INC. (1979), Houston, Tex.

KRONMAL, R.A., and TARTER, M.E. (1968), "The Estimation of Probability Densities and Cumulatives by Fourier Series Methods," *Journal of the American Statistical Association*, 63, 925–952.

LOFTSGAARDEN, D.O., and QUESENBERRY, C.P. (1965), "A Nonparametric Estimate of a Multivariate Density Function," *Annals of Mathematical Statistics*, 36, 1049–1051.

NEZAMES, D. (1980), "Some Results for Estimating Bivariate Densities Using Kernel, Orthogonal Series and Penalized-Likelihood Procedures," unpublished PhD dissertation, Rice University.

PARZEN, E. (1962), "On Estimation of a Probability Density Function and Mode," *Annals of Mathematical Statistics*, 33, 1065–1076.

ROBERTSON, T. (1967), "On Estimating a Density Which Is Measurable With Respect to a Sigma Lattice," *Annals of Mathematical Statistics*, 33, 482–493.

ROSENBLATT, M. (1956), "Remarks on Some Nonparametric Estimates of a Density Function," *Annals of Mathematical Statistics*, 27, 832–837.

——— (1971), "Curve Estimates," *Annals of Mathematical Statistics*, 42, 1815–1842.

SCOTT, D.W. (1980), "Comment" on "Density Estimation and Bump-Hunting by the Penalized Likelihood Method Exemplified by Scattering and Meteorite Data" by I.J. Good and R.A. Gaskins, *Journal of the American Statistical Association*, 75, 61–62.

SCOTT, D.W., GORRY, G.A., HOFFMAN, R.G., BARBORIAK, J.J., and GOTTO, A.M. (1980), "A New Approach for Evaluating Risk Factors in Coronary Artery Disease: A Study of Lipid Concentrations and Severity of Disease in 1847 Males," *Circulation*, 62, 477–484.

SCOTT, D.W., GOTTO, A.M., COLE, J.S., and GORRY, G.A. (1978), "Plasma Lipids As Collateral Risk Factors in Coronary Artery Disease—A Study of 371 Males With Chest Pain," *Journal of Chronic Diseases*, 31, 337–345.

SCOTT, D.W., TAPIA, R.A., and THOMPSON, J.R. (1977), "Kernel Density Estimation Revisited," *Journal of Nonlinear Analysis, Theory, Methods and Applications*, 1, 339–372.

——— (1980), "Nonparametric Probability Density Estimation by Discrete Maximum Penalized-Likelihood Criteria," *Annals of Statistics*, 8, 820–832.

STEWART, L. (1979), "Multiparameter Univariate Bayesian Analysis," *Journal of the American Statistical Association*, 74, 684–693.

TAPIA, R.A., and THOMPSON, J.R. (1978), *Nonparametric Probability Density Estimation*, Baltimore: Johns Hopkins University Press.

TARTER, M.E., and KRONMAL, R.A. (1976), "An Introduction to the Implementation and Theory of Nonparametric Density Estimation," *The American Statistician*, 30, 105–112.

VAN RYZIN, J. (1973), "A Histogram Method of Density Estimation," *Communications in Statistics*, 2, 493–506.

WAHBA, G. (1976), "Histosplines With Knots Which Are Order Statistics," *Journal of the Royal Statistical Society*, Ser. B, 38, 140–151.

——— (1977), "Optimal Smoothing of Density Estimates," in *Classification and Clustering*, ed J. Van Ryzin, New York: Academic Press, 423–458.

——— (1978), "Data-Based Optimal Smoothing of Orthogonal Series Estimates," *Annals of Statistics*, in press.

WATSON, G.S. (1969), "Density Estimation by Orthogonal Series," *Annals of Mathematical Statistics*, 40, 1496–1498.

WATSON, G.S., and LEADBETTER, M.R. (1963), "On the Estimation of the Probability Density, I," *Annals of Mathematical Statistics*, 34, 480–491.

WEGMAN, E.J. (1970), "Maximum Likelihood Estimation of a Unimodal Density, II," *Annals of Mathematical Statistics*, 41, 2169–2174.

WHITTLE, P. (1958), "On Smoothing of Probability Density Function," *Journal of the Royal Statistical Society*, Ser. B, 20, 334–343.

WONG, W.H. (1979), "Expected Information Criterion for the Smoothing Parameter of Density Estimates," Technical Report No. 589, University of Wisconsin—Madison, Dept. of Statistics.