BNL-23305
AMD-778

COMPARISON OF DENSITY ESTIMATORS*

Samuel Kao
Applied Mathematics Department
Brookhaven National Laboratory
Upton, NY

and

John F. Monahan
Applied Mathematics Department
Brookhaven National Laboratory
Upton, NY

September 1977

*Work performed under the auspices of the ERDA.

DISCLAIMER

1. Introduction

Considerable new work has been done in the field of probability density estimation since the survey paper and simulation study done by Wegman (1972a,b). Notable in this work is the introduction of some new methods, such as the polynomial and spline methods and the nearest neighbor method. In addition, the asymptotic properties have been studied in depth and with unified results in the work of Farrell (1972) and Wahba (1975b).

Part of the comparisons done in this paper will be a summary of this work. In addition, the computational complexity of the various algorithms will be analyzed, along with an analysis of some simulations. Here the objective is to compare the performance of the various methods in small samples, their sensitivity to change in their parameters, and also to attempt to discover at what point a smaple is so small that density estimation can no longer be worthwhile.

2. Common Methods of Estimation

The problem considered here is the estimation of an unknown probability density function $f(x)$ (assumed to exist) from a sample X_i, $i = 1,\ldots,N$ arising from a distribution with such a density. In the histogram method, the support of the distribution is broken into intervals. For x in interval j, the estimate of the density is $\hat{f}(x) = Y_j/N\ell_j$ where Y_j observations fell in the interval length ℓ_j.

For orthogonal series estimation (Tarter and Kronmal (1976) and references cited there), f is assumed to be square integrable with respect to a weight function r. Let $\{\varphi\}$ be an orthonormal basis for $L_2(r)$. then f can be expanded in its Fourier Series:

$$f(x) = \sum_j a_j \varphi_j(x). \quad (5)$$

The Fourier coefficients are then estimated by

$$\hat{a}_{jN} = N^{-1} \sum_i \varphi_j(X_i) r(X_i)$$

for $j = 1,\ldots,M$, $\hat{a}_{jN} = 0$ for $j > M$; hence the density is estimated by

$$\hat{f}(x) = \sum_{j=1}^{M} \hat{a}_{jN} \varphi_j(x).$$

In the spline method (see Wahba (1975a, 1976) and references

cited there), the empirical distribution function is interpolated through M regularly spaced points (or regularly chosen order statitstics) using a cubic spline. The boundary conditions are made on f at the endpoints either a priori or estimated, say, using the histogram method. The density estimate is the spline's derivative.

To use the polynomial method (Wahba, 1971), M regularly spaced order statistics are chosen. The estimate of $f(x)$ is obtained by interpolating the empirical distribution function with an r^{th} degree polynomial through $r + 1$ order statistics near x and differentiating.

The kernel method (Parzen, 1962) can be viewed as a smoothing of the lumps of probability in the empirical d.f. The estimate is given by

$$\hat{f}(x) = (HN)^{-1} \sum_i k((x - X_i)/H(N))$$

and $H \to 0$ (less smoothing) as $N \to \infty$.

The nearest neighbor method (see Loftsgaarden and Quesenberry (1965) and references in Moore and Yackel (1977)) is a modification of the kernel method where the convergence of H depends on the data. Here, given r, $H(x,N)$ is the distance from x to the r^{th} closest observation.

One criterion for comparing estimates may be on aesthetic grounds. An estimator which is smooth, nonnegative on its support and zero elsewhere and normalized, i.e. integrates to 1, is highly preferred. However, the histogram is not "smooth" and is limited

to distributions of finite support, though is nonnegative and integrates to 1. Orthogonal series estimates can be negative, but with the appropriate choice of $\{\varphi_i\}$ such as trigometric series or Legendre polynomials on a finite interval or Hermite polynomials for R^1 can handle any type of support. They are smooth, but need not integrate to 1 although this may be enforced with an appropriate choice of φ_1 and r. Both spline and polynomial methods are smooth and integrate to 1 but they might be negative. For the kernel method, if the kernel is chosen to be a density function, the density estimate is a density. However, a minor drawback is that some smoothing may stray beyond the boundaries of a finite support. The nearest neighbor method is not normalized even if the kernel is a density function. The kernel, though, may not be chosen this way, and while gaining some asymptotic properties, it loses others. As a result, the histogram is the only "clean" estimate, but its use is limited and it is not as smooth as could be preferred.

III. Computational Complexity

The criteria here for comparison are the time and space computing requirements for large problems. Large problems vary in size depending on the capabilities of machines. What are important are the orders of magnitude because these decide whether a certain task is feasible; scalar factors are important in dollars and cents.

We are concerned here with the computation required to obtain the estimate of a density at L points with a sample of N observations. The orders of magnitude for the tasks required by the various methods are given in Table 1. Although many of the entries in the table are self-evident, there are two advances in computer science which are useful here. First is that the work required (e.g. comparisons) to sort a list of N numbers is $O(N\ell gN)$ (ℓg = base 2 logarithm). Second is that to place an element in a sorted list or to check for membership in such a list requires an additional (ℓgN). A useful tool for statisticians, fast order statistics algorithms, taking $O(N)$ are not applicable here. The reader is referred to Knuth (1973) or Aho, Upcroft and Ullman (1974) for the derivation and explanation of these algorithms.

Evident from the table is that some methods can be costly to use. First to be noted is the storage requirements. The kernel method, hence, nearest neighbor, requires that all of the data

be stored. For the other methods, the storage needed is considerably smaller. If L is small, the set-up and sorting cost preponderate and the orthogonal series becomes costly and the kernel method cheap. If L is large, say for doing detailed plotting, the last line of the table tells the tale. The kernel method is most costly, N^2, with the use of a finite kernel and sorting the data drop the cost to $N^{3/2}$, the same as orthogonal series. The polynomial method is next, marginally more costly than the histogram and spline methods. The price paid for smoothing the histogram by using the spline method is only a scalar factor. The nearest neighbor method, not included in the table, requires sorting of the data in addition to the same costs as the kernel method.

Finally, the human cost of coding, checking and using these algorithms is usually avoided by computer scientists, but is very important in the field of statistical computing. We are happy to report that, in the opinion of one of the authors, most of the methods are very easy to code, debug and use. The only non-trivial problems are the $O(N\ell gN)$ sort, the spline method and the evaluation in the polynomial method. Code for the first is ubiquitous; the last one takes only a few minutes following perusal of a good numerical analysis text. The spline method does require some research but less than one man-day and appropriate code is available in many program libraries.

Table 1

Computational Complexity

Order of magnitude of calculations for evaluating density estimate at L points

Method	Hist.	Orthog. Series	Spline	Kernel	*	Poly-nomial
Parameter	M = # bins	M = # terms	M = # knots	H = Scale Factor		M = # Intervals
Sort ?	No	No	No	No	Yes	Yes
Storage	M	M	M	N	N	M
Set-up	N	MN	N+M	None	None	M
Density	L	LM	L	LN	HNL+ LlgN	L+ LlgM
Parameter rate m=1, p=2	$N^{\frac{1}{2}}$	$N^{\frac{1}{2}}$	$N^{\frac{1}{2}}$	$N^{-\frac{1}{2}}$	$N^{-\frac{1}{2}}$	$N^{\frac{1}{2}}$
Total	L+N	$LN^{\frac{1}{2}} + N^{3/2}$	L+N	LN	$LN^{\frac{1}{2}} +$ NlgN	LlgN + NlgN
L=N	N	$N^{3/2}$	N	N^2	$N^{3/2}$	NlgN

*Kernel with finite support

IV. Asymptotic Behavior of the Density Estimators

It is indicated in a recent paper by Wahba (1975) that the mean squared errors of density estimators obtained using polynomial, kernel, and orthogonal series methods have the same optimal order of magnitude when some appropriate assumptions on the true density function and the selected kernel etc. are satisfied. To be more explicit, we define

$$W_p^{(m)} = \{f : f^{(\nu)} \text{ abs. conti., } \nu=0,1,\ldots,m-1, f^{(m)} \in L_p\},$$

where m is a positive integer, p is a real number and is no less than 1, and let $W_p^{(m)}(\Lambda) = \{f : f \in W_p^{(m)} \text{ and } \|f^{(m)}\|_p \leq \Lambda\}$, where $\|\ \|_p$ means the L_p-norm. Then if the true density function f belongs to $W_p^{(m)}(\Lambda)$ for some Λ, the density estimators by the orthogonal series method and the kernel method have their mean squared errors optimally in the order of magnitude of $0(N^{-\phi(m,p)})$, where N is the sample size and $\phi(m,p) = (2m-2/p)/(2m+1-2/p)$. In addition, if it is also assumed that, at the point of interest x, there exists λ such that $0 < \lambda \leq f(u) \leq \Lambda$ for u in a neighborhood of x, $|uF(u)|$ bounded for $u \leq x$, and $|u(1-F(u))|$ bounded for $u \geq x$, then the density estimator by the polynomial method has its mean squared error optimally in the order of magnitude of $0(N^{-\phi(m,p)})$.

The generalized nearest-neighbor method is similar to the kernel method, except that the sample-independent scale

parameter H_N is replaced by a sample-dependent scale. Therefore, though not yet proved in literature, it is conjectured that the mean squared error of density estimator by this method has an optimal order of magnitude of $0(N^{-\phi(m,p)})$ also.

The order of magnitude of the mean squared error strongly depends on the smoothness of the true density function. The histogram method utilizes little of the smoothness of the true density function. For this reason, the density estimator by the histogram method has the optimal order of magnitude $0(N^{-\phi(1,p)})$ for its mean squared error.

The density estimator obtained by using cubic spline interpolation has been analyzed by Wahba (1975a, b). Its mean squared error has the optimal order of magnitude $0(N^{-\phi(m,p)})$ for (i) $m=1$, $p=2$ or ∞, (ii) $m=2$, $1\leq p\leq 2$ or $p=\infty$, and (iii) $m=3$, $1\leq p\leq 2$ or $p=\infty$. (These are the only cases having been proved so far.) It is conjectured that this optimal order of magnitude may hold for cases when $m>3$ if corresponding m-degree spline polynomial is used.

Let $\hat{f}$ be the density estimator for the true density f, then the mean squared error $E(\hat{f}-f)^2$ is the sum of two terms, which are the variance $V(\hat{f})$ and the square of bias $b_N^2 = (E\hat{f}-f)^2$. Since reducing $V(\hat{f})$ increases b_N^2, and vice versa, then optimality will be reached when $V(\hat{f})$

and b_N^2 have the same order of magnitude. In doing so, asymptotic normality for the obtained density estimator is not attained. In fact, it is attained only when the square of bias is negligible as compared against the variance as the sample size tends to infinity (i.e. $\lim_{N\to\infty} b_N^2/V(\hat{f}_N) = 0$). Nevertheless, to attain asymptotic normality for the density estimators derived using these available methods is practically unnecessary. The reason is in what follows. One of the major advantages to attain asymptotic normality is to enable us to obtain confidence interval for the true density. The confidence interval obtained through asymptotic normality and the standard-normal table has a width in the order of magnitude of $(V(\hat{f}))^{\frac{1}{2}}$ for which (i) $\lim_{N\to\infty} b_N^2/V(\hat{f}) = 0$ has to be satisfied, and we assume it is of $0(v_N)$. On the other hand, the confidence interval obtained using an argument on the optimal asymptotic mean squared error and the Tchebyshev's inequality has a width in the order of magnitude of $(V(\hat{f}_N))^{\frac{1}{2}}$ for which (ii) $\lim_{N\to\infty} b_N^2/V(\hat{f}_N) = c > 0$ is satisfied, and we assume it is of $0(u_N)$. Then, since from the conditions of (i) and (ii) we can see that $u_N/v_N \to 0$ as $N \to \infty$, we thus find that asymptotically (as $N \to \infty$) a better confidence interval (namely, with narrower width but with same confidence coefficient) may be obtained by using the optimal mean squared error and the Tchebyshev's inequality.

Accordingly, there seems to be no need for dealing with the asymptotic normality for these density estimators.

Results on pointwise consistency properties of density estimators under various assumptions are available in literature noticeably for kernel, nearest-neighbor, and histogram type methods. For the methods considered in this paper, the pointwise weak consistency is attained by the density estimators each of which has the optimal asymptotic mean squared error for its corresponding method. This is easily seen using Tchebyshev's inequality that gives

$$P(|\hat{f}-f|>\epsilon) \leq \frac{E(\hat{f}-f)^2}{\epsilon^2} = O(\epsilon^{-2} N^{-\phi(m,p)}) \to 0 \text{ as } N \to \infty$$

for any $\epsilon>0$. In fact, the estimator $\hat{f}$ may have the pointwise weak convergence property as long as both the variance and the bias of $\hat{f}$ tend to 0 as $N \to \infty$. From the fact that $\phi(m,p) \leq 1$ for all m and p as defined the pointwise strong consistency property does not necessarily hold for the density estimator with optimal asymptotic mean squared error, since for the harmonic series $\Sigma_N 1/N$ diverges. Van Ryzin (1969) proved for the pointwise strong consistency of the kernel type density estimator, where a form of Lipschitz condition on the kernel was required. Strong uniform consistency was proved to hold under the additional assumptions

that the continuity of f is uniform and the kernel K has absolutely integrable characteristic function. Accordingly, such a proof does not apply to uniform kernel. Most recently Moore and Yackel (1977) constructed the proof for uniform kernel while investigating consistency properties of generalized nearest-neighbor density estimators. The nearest-neighbor density estimator proposed by Loftsgaarden and Quesenberry (1965) is the special case when uniform kernel is taken for the generalized nearest-neighbor density estimator. They showed that the nearest-neighbor density estimator is pointwise weak consistent at continuity point of f, while Moore and Yackel (1977) came up with the extensive conclusion that, roughly stated, any consistency theorem true for the kernel type density estimator (including the case of uniform kernel) remains true for the generalized nearest-neighbor density estimator. In this manner the properties of weak and strong consistency, pointwise and uniform, for the generalized nearest-neighbor density estimator may easily be observed by referring to that for the kernel type density estimators.

For the histogram-type density estimators, Van Ryzin (1973) proposed a method and proved for the pointwise weak and strong consistency properties of the estimators that are obtained from the method. Strong uniform consistency for the same histogram-type density estimators was later proved by Van Ryzin and Kim (1975).

To investigate the behavior of density estimates when the sample size is small, a simulation experiment was performed. The objectives were to compare the performance of the various methods, their sensitivity to changes in their parameters and to discover at what point a sample is so small that density estimation is not worthwhile.

The computations were done on a CDC 6600 computer in Fortran under Scope 3.4.4 and COPE. The source of pseudorandom numbers was the uniform generator GFSR due to Lewis and Payne (1973). Five methods were used: histogram, spline with equispaced knots, orthogonal series using cosines, and the kernel method using triangular and uniform kernels. Samples numbering 25, 50, and 200 were obtained from three distributions on (-1, +1),

$$f_1(x) = 1-|x|$$

$$f_2(x) = \frac{3}{4}(1-x^2)$$

$$f_3(x) = \frac{1}{2} + \frac{x}{4} .$$

A finite support was chosen to place all of the methods on an equal footing. One thousand replications were done for each combination of method, parameter values and sampling distribution, (except for the experiments with Kernel method with N = 200, where 250 replications were performed). Four types of error were recorded.

1) $N^{-1} \sum_i \left| \frac{\hat{f}_n(X_i)}{f(X_i)} - 1 \right| \approx \int |\hat{f} - f| dx$ L_1

2) $N^{-1} \sum_i \left| \frac{\hat{f}_n(X_i)}{f(X_i)} - 1 \right|^2 f(X_i) \approx \int |\hat{f}_n - f|^2 dx$ L_2

3) $N^{-1} \sum |\hat{f}(X_i) - f(X_i)|^2 \approx \int |\hat{f} - f|^2 f dx$ Average Square error

4) $\sup_i \left| \int_{\infty}^{X_i} \hat{f}(x) dx - F_n(X_i) \right|$ Approximate Kolmogorov-Smirnov

The parameters for the different methods were chosen according to an optimal search policy, either a golden section search for the scale parameter in the kernel method or a lattice search for the number of terms or number of intervals. While in reality four criteria are used, they typically moved in unison with changes in the parameters.

In examination of the results given in table 2, a few general comments can be made. First, note that a trivial estimate, such as, $\hat{f}(x) = 0.5$, yields L_2 errors of $\frac{1}{6}$, $\frac{1}{10}$, and $\frac{1}{24}$ for f_1, f_2, and f_3, respectively. Note that for f_3, 25 observations is just too few to do density estimation, since one could do nearly as well with $\hat{f}(x) = 0.5$. However, the performance of kernel estimates for f_1 and f_2 is quite commendable. The spline and orthogonal series do

well for f_1 while the histogram does well on some of the distributions when the sample sizes is quite small.

For small (50) and moderate (200) sample sizes the picture appears cloudy. But note that the ordering of the best for each distribution is nearly the same for the two sample sizes. In general, note that the histogram never does very well. The orthogonal series method place high consistently, but it can be very sensitive to parameter changes, i.e., the number of terms. The spline method never does poorly or extremely well, and is stable to parameter changes. The kernel method seems handicapped when the true density is very smooth, like f_3, but does well for f less smooth and for very small (25) sample sizes.

Legend for Table 2

1. Entries are means for 1000 replications (250 for kernel, N = 200) of

$$\frac{1}{N}\sum_{i=1}^{N}\left|1-\frac{\hat{f}(X_i)}{f(X_i)}\right| f(X_i).$$

2. All standard errors for these entries are less than 7.7%.

3. Parameters

Histogram	M = No. of Intervals
Orthogonal Series	M = No. of Terms
Spline	M = No. of Intervals
Kernel	H = Scale Factor

4. The means and standard errors from all 4 error criteria are available from the authors

Mean Approximate L_2 Error

$$f_1(x) = 1 - |x|$$

N=25		N=50		N=200	
Histogram					
M=2	.18196	M=4	.07037	M=5	.03222
3	.08730	5	.06101	6	.03074
4	.09881	6	.07107	7	.02751
		7	.07544	8	.02831
Orthogonal Series					
M=1	.18902	M=1	.17769	M=1	.16603
2	.02632	2	.01474	2	.00567
3	.05196	3	.02749	3	.00847
		4	.03906	5	.01427
Spline					
M=4	.06057	M=4	.04061	M=4	.02509
5	.07982	5	.04510	5	.01811
		6	.05209	6	.01811
		8	.07644	8	.02071
Uniform Kernel					
H=.23	.06404	H=.23	.03196	H=.14	.01443
.38	.03332	.38	.03004	.23	.00877
.47	.03072	.47	.02189	.38	.00986
.61	.03914	.61	.03423	.61	.03014
Triangular Kernel					
H=.38	.05551	H=.38	.02638	H=.23	.01173
.52	.03372	.52	.01872	.38	.00743
.61	.02884	.61	.01837	.47	.00756
.76	.02890	.76	.02280	.61	.01094

Mean Approximate L_2 Error

$$f_2(x) = \frac{3}{4}(1-x^2)$$

N=25		N=50		N=200	
Histogram					
M=2	.16065	M=3	.12043	M=3	.09039
3	.16569	4	.10687	4	.06548
4	.17509	5	.12011	5	.06672
		7	.15348	7	.06672
Orthogonal Series					
M=1	.18361	M=1	.15310	M=1	.13471
2	.11923	2	.06869	2	.04036
3	.15918	3	.09014	3	.04956
		5	.12974	5	.05530
Spline					
M=4	.15927	M=4	.10633	M=4	.07141
5	.18751	5	.10895	5	.06050
		6	.12407	6	.06304
		8	.16402	8	.06622
Uniform Kernel					
H=.38	.07198	H=.38	.04338	H=.38	.02420
.52	.04168	.52	.02586	.52	.01634
.61	.03522	.61	.02343	.61	.01704
.76	.03791	.76	.03104	.76	.02771
Triangular Kernel					
H=.38	.13560	H=.38	.07477	H=.38	.03621
.61	.06390	.61	.03719	.61	.02059
.76	.04437	.76	.02668	.76	.01637
.85	.03786	.85	.02389	.85	.01616

Mean Approximate L_2 Error

$$f_3(x) = \frac{1}{2} + \frac{x}{4}$$

N=25		N=50		N=200	
Histogram					
M=2	.02954	M=3	.02553	M=3	.01009
3	.04568	4	.03322	4	.01087
4	.06437	5	.04495	5	.01287
		7	.06789	7	.01709
Orthogonal Series					
M=1	.02070	M=1	.01091	M=1	.00341
2	.04464	2	.02210	2	.06623
3	.06703	3	.03256	3	.00835
		5	.05468	5	.01380
Spline					
M=4	.04854	M=4	.02562	M=4	.00799
5	.07154	5	.03589	5	.01011
		6	.04979	6	.01321
		8	.07658	8	.01882
Uniform Kernel					
H=.38	.04439	H=.23	.04063	H=.14	.02184
.47	.04141	.38	.03319	.23	.01950
.52	.04158	.47	.03441	.38	.02422
.61	.04355	.61	.03996	.61	.03735
Triangular Kernel					
H=.38	.06049	H=.38	.03395	H=.23	.01842
.61	.03659	.52	.02865	.38	.01727
.76	.03445	.61	.02839	.47	.01875
.85	.03498	.76	.03032	.61	.02259

REFERENCES

1. Ahlberg, J. H., Nilson, E.N., Walsh, J. L. (1967) The Theory of Splines and Their Applications. Academic Press.

2. Aho, Alfred, V., Hopcroft, John E., Ullman, Jeffrey D. (1974) The Design and Analysis of Computer Algorithms. Addison-Wesley.

3. Farrell, R. H. (1972) On Best Obtainable Asymptotic Rates of Convergence in Estimation of a Density Function at a Point. Ann. Math. Stat. Vol. 43, pp. 179-180.

4. Knuth, Donald E. (1973) The Art of Computer Programming. Vol. 3: Sorting and Searching. Addison-Wesley.

5. Lewis, T. G., Payne, W. H. (1973) Generalized Feedback Shift Register Pseudorandom Number Algorithm. JACM, Vol. 20, No. 3, pp. 456-468.

6. Loftsgaarden, D. O. and Quesenberry, C. P. (1965) A Nonparametric Estimate of a Multivariate Density Function. Ann. Math. Stat. Vol. 36, pp. 1049-1051.

7. Moore, David, S., Yackel, James, W. (1977) Consistency Properties of Nearest Neighbor Density Function Estimators. Ann. Stat., Vol. 5, No. 1, pp. 143-154.

8. Parzen, E. (1962) On the Estimation of a Probability Density Function and the Mode. Ann. Math. Stat. Vol. 33, pp. 1065-1076.

9. Tarter, M. E., Kronmal, R. A. (1976) An Introduction to the Implementation and Theory of Nonparametric Density Estimation. The American Statistician, Vol. 20, No. 3, pp. 105-112.

10. Van Ryzin, J. (1969) On Strong Consistency of Density Estimates. Ann. Math. Stat., Vol. 40, pp. 1765-1772.

11. Van Ryzin, J. and Kim, B. K. (1975) Uniform Consistency of a Histogram Density Estimator and Modal Estimation. Comm. Stat., Vol. 4, pp. 303-316.

12. Wahba, Grace. (1971) A polynomial Algorithm for Density Estimation. Ann. Math. Stat. Vol. 42, pp. 1870-1886.

13. Wahba, Grace. (1975a). Interpolating Spline Methods for Density Estimation. I.Equispaced knots. Ann. Statist. 3, pp. 30-48.

14. Wahba, Grace. (1975b) Optimal Convergence Properties of Variable Knot, Kernel and Orthogonal Series Methods for Density Estimation. Ann. Statistics Vol. 3, pp. 15-29.

15. Wahba, Grace. (1976) Histoplines with Knots which are Order Statistics. JRSSB, Vol. 38, No. 2, pp. 140-151.

16. Wegman, Edward, J. (1972a) Nonparametric Probability Density Estimation: I.A Summary of Available Methods. Technometrics, Vol. 14, No. 3, pp. 533-546.

17. Wegman, Edward J. (1972b) Nonparametric Probability Density Estimation: II.A Comparison of Density Estimation Methods. J. Statist. Comput. Simul., Vol. 1, pp. 225-245.