Reprinted from:
FRONTIERS OF PATTERN RECOGNITION

A HIERARCHY OF PROBABILITY DENSITY FUNCTION ESTIMATES

Thomas M. Cover

DEPARTMENT OF ELECTRICAL ENGINEERING
& STATISTICS
STANFORD UNIVERSITY

1. Abstract and Summary

The purpose of this paper is to consider a hierarchy of probability density function estimation procedures. The discussion will culminate in a speculation about a universal procedure. First the histogram approach will be investigated. We shall then consider the orthogonal function expansion approach, of which the histogram approach is a special case. Next in complexity is the slightly more complicated window function approach suggested by Parzen and Rosenblatt. One degree of freedom more is found in the approach of Loftsgaarden and Quesenberry in which the window size itself is allowed to be a function of the data.

Finally we ask ourselves what is meant by large, small and medium size samples. How should the number of degrees of freedom of the fitting densities grow with the size of the sample? What is the notion of intrinsic complexity of the underlying true distribution?

2. Introduction

Let $x_1, x_2, x_3, \ldots$ be a sequence of independent identically distributed random variables drawn according to some unknown underlying density function $f(x)$. It is desired to form a sequence of probability density function estimates $\hat{f}_n(x)$ depends only on $x_1, x_2, \ldots, x_n$. An excellent survey of various probability density function estimation procedures together with the modes of convergence, and in some cases the rates of convergence, may be found in Wegman [77] and Rosenblatt [57]. A fairly extensive list of references is also provided at the end of this paper [1-87].

3. The Histogram Approach

To fix ideas we shall consider only univariate random variables. The literature contains the multivariate extensions. In the histogram approach we have a partition of the real line into sets $S_1, S_2, \ldots$. A good estimate $\hat{f}_n(x)$ is given by letting $\hat{f}_n(x)$ be a constant over region S_i, where the constant is given by the proportion of the x_j's among $x_1, x_2, \ldots, x_n$ which fall in set S_i. Thus if g_i is the indicator function for the set S_i, we may write

$$\hat{f}_n(x) = \frac{1}{n} \sum_{j=1}^{n} g_i(x_j)\, g_i(x)$$

We may then readily verify that

$$E_{x_1, x_2, \ldots, x_n} \hat{f}_n(x) = \int_{S_i} f(x')dx', \quad \text{for } x \in S_i, \quad i = 1, 2, \ldots, n.$$

Moreover, the variance of $\hat{f}_n(x)$ tends to zero as $1/n$.

The problems with this procedure are many. First, the partitioning of the space gives an undesirable quantization of the probability distribution--thus yielding estimates which are piecewise constant over the partitioning sets. Unless by some chance the true distribution were piecewise constant over these sets, there would be no hope of asymptotic convergence. Next, the partitioning has to be designed before the data is seen. This allows the possibility that almost all of the probability mass of the underlying distribution may lie in just one cell of the histogram partitioning and therefore that no useful density estimate will be obtained. Attempts to refine the partition as the number of observations tends to infinity are possible but are very clumsy in this framework. We shall see that the subsequent schemes due to Parzen, Rosenblatt, Loftsgaarden and Quesenberry are superior for this purpose.

4. The Orthogonal Function Approach

Let $\psi_1(x), \psi_2(x), \psi_3(x), \ldots$ be a sequence of orthonormal functions defined on the real line. Based on the sample $x_1, x_2, \ldots$ we wish to find a probability density function estimate $f_n(x)$ of the form

$$\hat{f}_n(x) = \sum_{i=1}^{k} c_i^{(n)}\, \psi_i(x)$$

In practice, the orthonormal functions are sometimes chosen to be the

Hermite polynomials, an apt choice when the underlying density is basically Gaussian with perhaps some correction terms. See for example the work by Schwartz [62-64] and also the review by Wegman [77].

Now let

$$J_n = \int \left(f(x) - \hat{f}_n(x)\right)^2 \, dx$$

Suppose first, for the sake of argument, that the underlying density is known, and it is desired to find the set of coefficients $c_1, c_2, \ldots c_k$ minimizing J_n. Setting to zero the partial derivatives of J_n with respect to c_i we find that

$$c_i^* = \int f \, \psi_i = E\, \psi_i(X).$$

Now, since in fact f is not really known, a wise procedure would be to estimate the optimal coefficients from the data $X_1, X_2, \ldots, X_n$. Since the optimal coefficient c_i^* is equal to $E\, \psi_i(X)$, we estimate the expected value by

$$\hat{c}_i^* = \frac{1}{n} \sum_{j=1}^{n} \psi_i(x_j).$$

Note in particular that

$$E\, \hat{c}_i^* = \frac{1}{n} \sum_{j=1}^{n} E\, \psi_i(x_j) = c_i^*$$

Thus the estimate of the coefficients is unbiased. Moreover, since the $\psi_i(x_j)$ are independent random variables, the variance of the sum is the sum of the variances. Thus

$$\text{Var } \hat{c}_i^* = \frac{1}{n} \; n\left(\text{Var } \psi_i(x)\right) = \sigma_i^{\,2}/n,$$

which tends to zero in the limit as n tends to infinity. Thus $\hat{c}_i^*$ is a consistent estimate of the optimal coefficient c_i^*.

Writing out the estimate $\hat{f}_n(x)$ using the estimates of the optimal coefficients, we have the nice formula

$$\hat{f}_n(x) = \frac{1}{n} \sum_{i=1}^{k} \sum_{j=1}^{n} \psi_i(x_j)\, \psi_i(x).$$

This expression if very reminiscent of the potential function formulation of

Aizerman, Braverman and Rozonoer.

Continuing, for known f we find that the mean squared error is now given by

$$J_n = \int f^2 - \int \hat{f}^2,$$

where $\hat{f}$ represents the projection of f onto the linear space spanned by $\psi_1, \ldots, \psi_k$. In case of unknown f with $x_1, x_2, \ldots, x_n$ known, we find that

$$E\, J_n = E\int (f-\hat{f}_n)^2 = E\int (f-\hat{f})^2 + E\int (\hat{f}-f_n)^2 = \int f^2 - \int \hat{f}^2 + \frac{1}{n}\sum_{i=1}^{n} \sigma_i^2$$

Thus we find that the expected value of J_n is equal to the projection error plus a statistical error. Luckily the statistical error tends to zero in the limit as $n \to \infty$. But a basic deficiency of this procedure, at least for a finite number of orthonormal functions is that there is a steady state projection error.

We wish to make the following comments on pdf estimation by orthogonal functions: 1) This approach is simply a generalized histogram approach. We see that the histogram approach can be expressed as an orthogonal function expansion by letting the ψ's be indicator functions for the cells S_i. 2) There is often the possibility that the orthogonal function $\hat{f}_n(x)$ expansion will be negative for some values of x and therefore not a possible probability density function. 3) There will be, in general, a steady-state projection error. 4) The scale of the ψ_i must be selected before the data is observed. Thus it may happen that all of the true density f lies in one cell of the histogram, so to speak; or, to put it another way, that the major part of f lies outside of the linear space spanned by the ψ_i. In any case, it would be nice to "peek" at the data before selecting the orthogonal functions for the expansion. One may actually observe a few observations in order to set the scale and translation parameters before one applies the procedure. However, in this case, the analysis which we have undertaken is invalid because all of the critical parameters such as ψ_i now depend on the data in a way not taken into account. For this reason, we need more sophisticated procedures where the data plays a larger role in the selection of the estimator function. Although extensions of the orthogonal functions approach have been made, we shall drop this approach for the time being.

5. Rosenblatt Estimator

Now let us consider an improvement on the histogram approach, sometimes called the naive pdf estimator. The Rosenblatt [53] estimate is of the form

$$\hat{f}_n(x) = \left(F_n(x+h) - F_n(x-h)\right) /2h,$$

where F_n is the empirical cdf of $x_1, x_2, \ldots, x_n$. In other words, the estimate of $f(x)$ at the point x, based on $x_1, x_2, \ldots, x_n$, is given by the proportion of hits of the x_i's in the window of width $2h$ centered at x. This proportion is then divided by $2h$ to form the estimate. Certainly the estimate appears to be good, since the proportion of hits in the cell is an estimate of the probability content of the cell and $2h$ is the total content of the cell. The ratio as $h \to 0$ is the probability density.

Another way of expressing $\hat{f}_n(x)$ is to define the indicator function

$$\psi_x(x') = \begin{cases} 1, & x' \in [x-h, x+h] \\ 0, & \text{otherwise} \end{cases}$$

Then $\hat{f}_n(x)$ is equal to

$$\frac{1}{2nh} \sum_{i=1}^{n} \psi_x(x_i).$$

Following Rosenblatt, let us calculate the expected value and variance of $\hat{f}_n$. First,

$$E\left(\hat{f}_n\right) = \frac{1}{2nh} \sum_{i=1}^{n} E\ \psi_x(x_i).$$

But the $\psi_x(x_i)$'s are Bernoulli random variables taking on the values 1 and 0 with the probabilities

$$F(x+h) - F(x-h) \text{ and } 1 - F(x+h) + F(x-h)$$

respectively. Thus

$$E\ \hat{f}_n(x) = \frac{1}{2h}\left(F(x+h) - F(x-h)\right).$$

Expanding this in a Taylor series expansion, and assuming the existence of the necessary derivatives, we obtain

$$E\ f_n(x) = f(x) + \frac{h^2}{6} f''(x) + O(h^4).$$

Thus, for a window size h tending to zero, the bias of this estimate tends to zero.

In fact, these calculations hold for any sample size $n = 1,2,3, \ldots$. Thus we see that we can make the bias as small as we wish, even in the single sample case. Of course, the single sample case would result in a probability density function estimate which was extremely large a very small proportion of

the time. Unbiased, yes, but the variance would be tremendous.

Now we calculate the variance of $\hat{f}_n(x)$.

$$\begin{aligned}\text{Var } \hat{f}_n(x) &= \text{Var } \frac{1}{2nh}\sum_{i=1}^{n} \psi_x(X_i) \\ &= \frac{1}{4n^2h^2}\left(n \text{ Var } \psi_x(X_i)\right) \\ &= \frac{p(1-p)}{4nh^2} ,\end{aligned}$$

because, for independent random variables, the variance of the sum is the sum of the variances, and

$$p = \Pr[\psi_x(X_i) = 1] = 2hf + \frac{2h^3 f''(x)}{6} + 0(h^5) .$$

Collecting iterms, we have

$$\begin{aligned}E\left(\hat{f}_n(x) - f(x)\right)^2 &= E\left(E\,\hat{f}_n(x) - \hat{f}_n(x)\right)^2 + \text{Var } \hat{f}_n(x) \\ &= \frac{f(x)}{2hn} + \frac{h^4}{36}\,|\,f''(x)\,|^2 + o\left(\frac{1}{hn} + h^4\right).\end{aligned}$$

Thus the squared error of the approximation can be made to go to zero as long as h tends to zero and hn tends to infinity. Letting $h = kn^{\alpha}$, we see that the dominant terms can be minimized by setting the exponents equal, thus resulting in $h = kn^{-1/5}$ as the optimal rate of decrease of the window size. More careful analysis shows that the constant k should be $k = \left(9f(x)/2|f''(x)|^2\right)^{1/5}$.

6. Parzen Estimators

Parzen, at about the same time, investigated a general class of density estimation procedures based on his work on the estimation of spectral density functions from finite data. Consider

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

For analysis of the behavior of this family of estimates, we will need the following lemma due to Bochner.

Lemma: Let K be bounded, absolutely integrable, and let

$|xK(x)| \to 0$ as $|x| \to \infty$. Let g be absolutely integrable and define

$$g_n(x) = \frac{1}{h(n)} \int K\left(\frac{x-y}{h(n)}\right) g(y)\, dy$$

Then

$$g_n(x) \to g(x) \int K(y)\, dy, \text{ as } h(n) \to 0,$$

at every point of continuity of g.

The proof of Bochner's theorem is repeated in Parzen's 1965 paper [43]. Examples of kernal functions K which satisfy the hypotheses of Bochner's lemma are: a) the window function of Rosenblatt, b) $e^{-|x|}$, c) the normal density, d) the Cauchy density, and e) $(\sin^2 x)/x^2$.

We obtain the consistency of $\hat{f}_n(x)$ by applying Bochner's lemma twice. First, using the identical distribution of the X_i's,

$$\begin{aligned} E\,\hat{f}_n(x) &= \frac{1}{n}\sum_{i=1}^{n} E\,\frac{1}{h(n)} K\left(\frac{x-X_i}{h(n)}\right) \\ &= E\,\frac{1}{h(n)} K\left(\frac{x-X_i}{h(n)}\right) \\ &= \int \frac{1}{h(n)} K\left(\frac{x-y}{h(n)}\right) f(y)\, dy \\ &\to f(x) \int K(y)\, dy \\ &= f(x) \end{aligned}$$

if $\int K(y)\, dy = 1$.

Thus $\hat{f}_n(x)$ is asymptotically unbiased as $h(n) \to 0$. This result holds even when the X_i's are dependent random variables with marginal pdf $f(x)$.

Secondly, since the X_i's are independent,

$$\begin{aligned} \text{Var}\,\hat{f}_n(x) &= \text{Var}\,\frac{1}{n}\sum \frac{1}{h(n)} K\left(\frac{x-X_i}{h(n)}\right) \\ &= \frac{1}{n^2}\; n\; \text{Var}\;\frac{1}{h(n)} K\left(\frac{x-X_i}{h(n)}\right) \\ &\le \frac{1}{n} E\left(\frac{1}{h(n)} K\left(\frac{x-X_i}{h(n)}\right)\right)^2 \end{aligned}$$

$$= \frac{1}{nh(n)} \int \frac{1}{h(n)} \; K^2\left(\frac{x-y}{h(n)}\right) f(y)\, dy$$

Now,

$$\int \frac{1}{h(n)} \; K^2\left(\frac{x-y}{h(n)}\right) f(y)\, dy \to f(x) \int K^2(y)\, dy < \infty.$$

Thus

$$\text{Var } \hat{f}_n(x) \to 0 \quad \text{as} \quad nh(n) \to \infty.$$

Hence

$$E\left(f_n(x) - f(x)\right)^2 \to 0$$

if $h(n) \to 0$ (thus sending the squared bias term to zero) and $nh(n) \to \infty$ (thus sending the variance term to zero). Thus $\hat{f}_n(x)$ is a consistent estimate of $f(x)$ at every point of continuity of x if $h(n) \to 0$ and $nh(n) \to \infty$.

We note in particular that the obvious choice of scale $h(n) = \frac{1}{n}$ leaves a residual variance in the limit as $n \to \infty$. This is due to the fact that we are not smoothing the data enough. Similarly if $h(n)$ is equal to a constant, we are left with a steady-state bias term due to the over-smoothing of the density function.

We know by the Cramer-Rao lower bound that the expected squared error must be greater than or equal to $\frac{1}{nI}$ where I is an appropriate information measure. Parzen estimates can be chosen to converge at rates arbitrarily close to the rate $\frac{1}{n}$ if sufficient differentiability of f is assumed.

7. The Method of Loftsgaarden and Quesenberry

One of the problems with the two previous methods is that the scale parameters cannot be easily set. For example, in the Parzen procedure, the smoothing at time n corresponds to the scale parameter $h(n)$. Now if it just so happens that the range of the observations x_1 through x_n is less than or equal to $h(n)$, the smoothing will be much too great and the answer will be ridiculous. Parzen's theorem guarantees only that as n tends to infinity, $h(n)$ and n will bear a nice relationship guaranteeing a consistent estimate.

In practice we should cheat when applying Parzen's estimate. At time n, we should look at the data, estimate the amount of smoothing that it will tolerate, adjust $h(n)$ accordingly, and then calculate the Parzen estimate of the density. The trouble with all of this is that the analysis has been made under the assumption that $h(n)$ is a deterministic function independent of the data. Thus, the entire analysis would be invalidated by such a procedure

and we would necessarily be forever suspicious of its convergence properties.

The method of Loftsgaarden and Quesenberry [36] effectively allows $h(n)$ to depend on the data. Their method is as follows. Suppose that $f(x)$ is to be estimated at the point x on the real line and that $x_1, x_2, \ldots, x_n$ are i.i.d. random variables drawn according to $f(x)$. Let k_n be some integer less than or equal to n. And let $h(n)$ be the distance to the k_nth closest point to x among the samples $x_1, x_2, \ldots, x_n$. Thus $h(n)$ is a random variable depending on the data. Now the length of the interval $[x-h(n), x+h(n)]$ is $2h(n)$. An estimate of the probability content of this interval is just the proportion of samples which fall in it, which we already know to be k_n/n. Thus it is reasonable to use

$$\hat{f}_n(x) = (k_n/n)/2h(n) = \frac{k_n}{2nh(n)}$$

as an estimate of the probability density function $f(x)$. Since we wish a local estimate of the pdf, choose k_n so that the proportion k_n/n tends to zero. If $k_n \to \infty$, $k_n/n \to 0$, it comes as no surprise that $\hat{f}_n(x)$ can be shown to be a consistent estimate of $f(x)$ at every point of continuity of x. This is shown on the paper by Loftsgaarden and Quesenberry [36] and will not be shown here.

This procedure offers several refinements over the usual histogram approach. First, we are "peeking" at the data in order to center the histogram cell about the desired point x; secondly, we are scaling the size of the histogram cell in such a manner that it contains neither too many nor too few data points. Too many would give over-smoothing and low variance; too few would give under-smoothing and high variance. Presumably, k_n is chosen so that the number of samples is just right in proportion to n. $k_n = \sqrt{n}$ is often used in practice.

8. Further Refinements

The existence of a succession of probability density function estimation procedures which takes into account ever more refined properties of the underlying density functions suggests that we may continue the refining process.

One problem with a more sophisticated approach is that it usually requires a larger sample size to yield good estimates. Ultimately, however, we can expect a more sophisticated procedure to converge faster and over a wider class of underlying densities.

The three factors at work in determining the appropriate degree of sophistication of the pdf estimator are 1) the number of degrees of freedom of the estimator, 2) the expected number of degrees of freedom of the

underlying density, and 3) the number of degrees of freedom of the data, i.e., the sample size n.

It is useful to consider the parametric density function estimation problem for a moment. Suppose that $x_1, x_2, \ldots, x_n$ are independent, identically distributed according to a normal distribution $N(\mu,1)$ with unknown mean μ. The estimation which is usually performed in this case is to estimate μ by some good estimator such as $\bar{x}_n$ and then estimate the density by $N(\bar{x}_n,1)$. This works well because we are dealing with a parametric family of distributions, and a good estimate of the parameter yields a good estimate for the density (obviously if we had parametrized the distributions in a terrible and discontinuous way, this statement would not hold).

It is probably possible to classify the family of all pdf's by their complexity. The low complexity density functions would include, for example, the uniform distributions over fixed intervals, finite unions of uniform distributions, and normal distributions, and other distributions which have very simple descriptions. Next in complexity would come the distributions which are highly multimodal but relatively smooth. Obviously, these distributions will require a large number of samples before we are in any position to estimate the distribution precisely. For example, if the underlying density has n modes, it would certainly be foolish to expect that we could identify the position of these modes before we had observed roughly n samples or more. So we see quickly that for any sample size n it is very easy to come up with a probability density function that will be poorly estimated by n samples. For example, we might have a uniform distribution except for a spike of width ϵ^2 and height $\frac{1}{\epsilon}$. Thus the sample size necessary to detect the position of this spike would be at least $\frac{1}{\epsilon^2}$, and until $n \geq 1/\epsilon^2$, the estimate of the pdf would differ in sup-norm form from the true pdf by at least $\frac{1}{\epsilon}$.

We wish to have a hierarchy of pdf estimators of increasing complexity. For example, if the true underlying density function is a normal distribution with unknown mean, but this is not known a priori, then after observing a sufficiently large number of samples, we would begin to suspect that this is indeed the case, that in fact the underlying density function is a normal distribution--which is what often happens in practice. In this situation we would then simply estimate the mean and variance and correction terms. This procedure would look for small correction terms to the normal distribution and would be extremely effective. Certainly additional samples might throw some doubt on our tentative conclusion that the underlying density function is essentially normal. Then we would be free to jump to some other conclusion and adapt our procedure about it.

The question is whether this disorderly jumping around from procedure

to procedure on the basis of some peeking at the data will still allow convergence. Practically speaking, we know that convergence will continue to hold, although the theoretical justification will probably be quite difficult. However, what is more to the point is whether we can actually converge to the true pdf as fast as a person who has a great deal more a priori knowledge and who realizes that the unknown pdf is a member of some parametric family. All that would be necessary for our convergence rate to equal that of the parametric pdf estimator would be that we would eventually be able to guess that we belonged in the appropriate parametric family.

Therefore I propose the following approach. First, let us enumerate all finite algorithms generating probability density functions. Next, at time n calculate the likelihoods of $x_1, x_2, \ldots, x_n$ under each of this infinite collection of pdfs. Among the first $\tau(n)$ pdfs choose that one for which $x_1, x_2, \ldots, x_n$ is most likely. This in some sense is the likeliest, simplest pdf. We shall let $\tau(n)$ be a function growing slowly with n. Certainly all density functions like the normal and the Cauchy and the union of uniforms will lie somewhere on this list and $\tau(n)$ will eventually grow without bound and include any particular density function on the list. And if, in fact, this density function is the true one, the likelihood of $x_1, x_2, \ldots, x_n$ under this density will soon dominate all the other likelihoods.

And just so long as $\tau(n)$ does not grow so fast that spurious candidates for pdfs have extremely high likelihoods on insufficiently small samples of data, we are guaranteed to be in the position of essentially guessing the parametric family and then estimating the underlying parameters. We have a chance of doing extremely well; certainly better than the procedures mentioned in the previous sections.

This program is far from precise and it is by no means clear whether the details and theoretical analysis could be carried out in practice. In any case, the practical applications of such an ultimate procedure would not be very great. However, in defense of these comments it should be said that all of empirical physics, chemistry and science has been accomplished more in the spirit of this last section than in the spirit of the previous sections. Thus we "know" that noise distributions on measurements are Gaussian and we "know" that the laws of physics read $F = ma$ and not $F = ma^{1.000001}$. A kernal function estimator would treat 1.0000 . . . . and 1.00001 much the same and would miss forever the thrill of leaping from conviction to conviction.

REFERENCES

1. Aizerman, M. A., E. M. Braverman, and L. I. Rozonoer, "The Probability Problem of Pattern Recognition Learning and the Method of Potential Functions", Automation & Remote Control, Vol. 26, 1964.
2. Anderson, Gary (1969), "A comparison of probability density estimates," Presented at the Institute of Mathematical Statistics Annual Meeting, August 19-22, New York.
3. Bartlett, M. S. "Periodogram analysis and continuous spectra" Biometrika 37, 1950, 1-16.
4. Bartlett, M. S. (1963), "Statistical estimation of density functions," *Sankhyā (A), 25*, pp. 245-254.
5. Bhattacharya, P. K. (1967), "Estimation of a probability density function and its derivatives", *Sankhyā (A), 29*, part 4, pp. 373-382.
6. Brillinger, D. R. and Rosenblatt, M., "Asymptotic theory of estimates of k^{th} order spectra" *Advanced Seminar on Spectral Analysis of Time Series* (ed. B. Harris) 1967, 153-188.
7. Brillinger, D. R. and Rosenblatt, M., "Computation and interpretation of k^{th} order spectra" *Advanced Seminar on Spectral Analysis of Time Series* (ed. B. Harris) 1967, 189-232.
8. Brillinger, D. R. "The spectral analysis of stationary interval functions" Proc. 6th Berkeley Symposium on Probability and Math. Stat.
9. Cacoullos, Theophiles (1966), "Estimation of a multivariate density," *Annals of the Institute of Statistical Mathematics, 18*, pp. 178-189.
10. Čencov, N. N. (1962), "Evaluation of an unknown distribution density from observations," *Soviet Math., 3*, pp. 1559-1562.
11. Cramér, H. and Leadbetter, M. R. *Stationary and Related Processes* 1967, John Wiley.
12. Craswell, W. J. (1965), "Density estimation in a topological group," *Ann. Math. Statist., 36*, pp. 1047-1048.
13. Daniels, H. "Saddlepoint approximations in statistics" Ann. Math. Statist. 25, 1964, 631-650.
14. Edgeworth, F. Y. (1904), "The law of error," *Trans. Camb. Phil. Soc., 20*, pp. 36 and 113.
15. Elderton, W. P. and Johnson, N. L. (1969), *Systems of Frequency Curves*, Cambridge University Press.
16. Elkins, T. A. (1968), "Cubical and Spherical estimation of a multivariate probability density," *JASA, 63*, pp. 1495-1513.
17. Epanechnikov, V. A. "Nonparametric estimates of a multivariate probability density" Theor. Prob. Appl. 14, 1969, 153-158.
18. Farrell, R. H. (1967), "On the lack of a uniformly consistent sequence of estimators of a density function in certain cases," *Ann. Math. Statist., 38*, pp. 471-474.
19. Fisher, L. and J. W. Van Ness, "Distinguishability of Probability Measures", Ann. Math. Stat., Vol. 40, pp. 381-392, 1969.

20. Fisher, R. A. (1912), "On an absolute criterion for fitting frequency curves," *Mess. of Math.*, *41*, pp. 155-160.
21. Fix, Evelyn and Hodges, J. L., Jr. (1951), "Nonparametric discrimination: consistency properties." Report Number 4, Project Number 21-49-004, USAF School of Aviation Medicine, Randolph Field, Texas, February, 1951.
22. Frenkiel, F. N. and Klebanoff, P. S. "Two-dimensional probability distribution in a turbulent field" Phys. Fluids 8, 1965, 2291-2293.
23. Gessaman, M. P. (1970), "A consistent nonparametric multivariate density estimator based on statistically equivalent blocks," to appear *Ann. Math. Statist.*, August, 1970.
24. Grenander, Ulf (1956), "On the theory of mortality measurement, Part II," *Skan. Aktuarieltidskr.*, *39*, pp. 125-153.
25. Grenander, U. and Rosenblatt, M. *Statistical Analysis of Stationary Time Series* 1957, John Wiley.
26. Hasselman, K., Munk, W. and MacDonald, G. "Bispectra of ocean waves" *Time Series Analysis* (ed. M. Rosenblatt) 1963.
27. Hodges, J. L. Jr. and Lehmann, E. L. "The efficiency of some nonparametric competitors of the t-test" Ann. Math. Statist. 27, 1956, 324-335.
28. Huber, P. J., Kleiner, B., Gasser, Th. and Dumermuth, G. "Statistical methods for investigating phase relations in stationary stochastic processes" to appear in the IEEE Transactions on Audio and Electroacoustics.
29. Kashyap, R. L., and C. C. Blaydon, "Estimation of Probability Density and Distribution Functions", IEEE Trans. on Info. Theory, Vol. IT-14, pp. 459-556, 1968.
30. Kendall, M. G. (1948), *The Advanced Theory of Statistics*, Vol. 1, Charles Griffin and Co., Ltd., London.
31. Kendall, M. G. and Stewart, A. (1958), *The Advanced Theory of Statistics*, Vol. 1, Charles Griffin and Co., Ltd., London.
32. Kolmogorov, A. "On the approximation of distributions of sums of independent summands by infinitely divisible distributions" Sankhya Ser. A 25, 1963, 159-174.
33. Kowalski, C. and Tarter, M. (1968), "On the simultaneous estimation of density and distribution functions with application to C-type density estimation," an unpublished manuscript.
34. Kronmal, R. and Tarter, M. (1968), "The estimation of probability densities and cumulatives by Fourier Series methods," *JASA*, *38*, pp. 482-493.
35. Leadbetter, M. R. (1963), "On the non-parametric estimation of probability densities," Technical Report No. 11, Research Triangle Institute. (Doctoral dissertation at the University of North Carolina at Chapel Hill.)
36. Loftsgaarden, D. O. and Quesenberry, C. P. (1965), "A Nonparametric Estimate of a Multivariate Density Function", Ann. Math. Stat., Vol. 38, pp. 1261-1265.
37. Loftsgaarden, D. O. and Quesenberry, C. P. (1965), "A non-parametric estimate of a multivariate density function," *Ann. Math. Statist.*, *38*, pp. 1261-1265.
38. Moore, P. S. and Henrichon, E. G. (1969), "Uniform consistency of some estimates of a density function," *Ann. Math. Statist.*, *40*, pp. 1499-1502.

39. Murthy, V. K. (1965), "Estimation of probability density," *Ann. Math. Statist., 36*, pp. 1027-1031.
40. Nadaraya, É. A. (1963), "On estimation of density functions of random variables" *Soobach. Akad. Nauk. Grugin SSR, XXXII, 2*, pp. 277-280. (In Russian.)
41. Nadaraya, É. A., (1965), "On non-parametric estimates of density functions and regression curves," *Theory Prob. Appl., 10*, pp. 186-190.
42. Parzen, E. "Mathematical considerations in the estimation of spectra" Technometrics 3, 1961, 167-190.
43. Parzen, E. (1962), "On the estimation of a probability density function and the mode," *Ann. Math. Statist., 33*, pp. 1065-1076.
44. Pearson, E. S. (1969), "Some historical reflections traced through the development of the use of frequency curves," Technical Report 38, Department of Statistics, Southern Methodist University.
45. Pearson, K. (1902), "On the systematic fitting of curves to observations and measurements, I," *Biometrika, 1*, pp. 265-303.
46. Pearson, K. (1902), "On the systematic fitting of curves to observations and measurements, II," *Biometrika, 2*, pp. 1-23.
47. Pickands, J. (1969), "Efficient estimation of a probability density function," *Ann. Math. Statist., 40*, pp. 854-864.
48. Quesenberry, C. P. and Scheult, A. H. (1969), "On unbiased estimation of density function," Abstracted *Ann. Math. Statist., 40*, pp. 2224.
49. Rao, B. L. S. P. (1969), "Estimation of a unimodal density," *Sankhyā (A)*, 31, pp. 26-36.
50. Révész, Pal (1968), *The Laws of Large Numbers*, Academic Press.
51. Robertson, T. (1967), "On estimating a density which is measurable with respect to a σ lattice," *Ann. Math. Statist., 38*, pp. 482-493.
52. Robertson, T., Cryer, J. D., and Rogg, R. V. (1968), "On non-parametric estimation of distributions and their modes," an unpublished manuscript.
53. Rosenblatt, M. (1956), "Remarks on some nonparametric estimates of a density function," *Ann. Math. Statist., 27*, pp. 832-837.
54. Rosenblatt, M. "Statistical analysis of stochastic processes with stationary residuals" *H. Cramér Volume* (ed. U. Grenander) 1959, 246-275.
55. Rosenblatt, M. "Conditional probability density and regression estimates" *Multivariate Analysis Vol. 2* (ed. Krishnaiah) 1969, 25-31.
56. Rosenblatt, M. "Density estimates and Markov sequences" *Nonparametric Techniques in Statistical Inference* (ed. M. Puri) 1970, 199-210.
57. Rosenblatt, M. "Curve Estimates", presented at the Institute of Mathematical Statistics Meeting, August 25-27, 1970, Laramie, Wyoming.
58. Roussas, G. "Nonparametric estimation in Markov processes" Ann. Inst. Statist. Math. 1969, 21, 73-87.
59. Sazanov, V. V. "On the multidimensional central limit theorem" Sankhya Ser. A. 30, 1968, 181-204.
60. Schuster, E. F. (1969), "Estimation of a probability density function and its derivatives," *Ann. Math. Statist., 40*, pp. 1187–1195.

61. Schuster, E. F. (1970), "Note on the uniform convergence of density estimates," *Ann. Math. Statist.*, Vol. *41*, August, 1970, 1347-48.
62. Schwartz, S. C. (1967), "Estimation of a probability density by an orthogonal series," *Ann. Math. Statist., 38*, pp. 1261-1265.
63. Schwartz, Stewart (1969), "Estimation of density functions by orthogonal series and an application to hypothesis testing." Presented at the Institute of Mathematical Statistics Annual Meeting, August 19-22, New York.
64. Schwartz, S., "An Example of Nonsupervised Adaptive Pattern Classification", IEEE Transactions on Automatic Control, Vol. AC-13, pp. 107-108, 1968.
65. Sclove, S. L. and Van Ryzin, J., "Estimating the Parameters of a Convolution", Journal of the Royal Statistical Society, Series B, Vol. 31, No. 1, pp. 181-191, 1969.
66. Tarter, M. E., Holcomb, R. L., and Kronmal R. A. (1967), "A description of new computer methods for estimating the population density," *Proc. A. C. M.*, Thompson Book Company, 22, pp. 511-519.
67. Tarter, M. and Kronmal, R. (1970), "On multivariate density estimates based on orthogonal expansions," *Ann. Math. Statist.*, 41, pp. 718-722.
68. Tsypkin, Y. Z., "Use of the Stochastic Approximation Method in Estimating Unknown Distribution Densities from Observations", Automatika i Telemekhanika, Vol. 27, No. 3, pp. 44-96, 1966.
69. Tukey, J. W. "An introduction to the measurement of spectra" *H. Cramér Volume* (ed. U. Grenander) 1959, 300-330.
70. Van Ness, J. "Asymptotic normality of bispectral estimates" Ann. Math. Statist. 37, 1966, 1257-1272.
71. Van Ryzin, J. (1969), "On strong consistency of density estimates," *Ann. Math. Statist., 40*, pp. 1765-1772.
72. Van Ryzin, J. (1970) "On a histogram method of density estimation." Presented at the Institute of Mathematical Statistics Meeting, April 8-10, Dallas, Tech. Report #226, Dept. of Stat., Univ. Wisconsin.
73. Venter, J. H. (1967), "On estimation of the mode," *Ann. Math. Statist., 38*, pp. 1446-1455.
74. Wahba, Grace (1970), "A polynomial algorithm for density estimation," submitted to *Ann. Math. Statist.*
75. Watson, G. S. and Leadbetter, M. R. (1963), "On estimating a probability density, I," *Ann. Math. Statist., 34*, pp. 480-491.
76. Watson, G. S. (1969), "Density estimation by orthogonal series," *Ann. Math. Statist., 40*, pp. 1496-1498.
77. Wegman, E. J. (1969), "Nonparametric probability density estimation," invited paper at the Institute of Mathematical Statistics meeting, April 8-10, 1970, Dallas; submitted to *JASA*.
78. Wegman, E. J. (1969), "A note on estimating unimodal density," *Ann. Math. Statist., 40*, pp. 1661-1667.
79. Wegman, E. J. (1970), "Maximum likelihood estimation of a unimodal density function," *Ann. Math. Statist.*, 41, pp. 457-471.

80. Wegman, E. J. (1969), "Maximum likelihood histograms," Institute of Statistics Mimeo Series #629, University of North Carolina at Chapel Hill.
81. Wegman, E. J. (1970), "Maximum likelihood estimation of a unimodal density, II" to appear *Ann. Math. Statist.*, December, 1970.
82. Weiss, L. and Wolfowitz, J. (1967), "Estimation of a density at a point," *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete, 7*, pp. 327-335.
83. Whittle, P. (1958), "On smoothing of probability density functions," *JRSS (B), 20*, pp. 334-343.
84. Woodroofe, M. (1967), "On the maximum deviation of the sample density," *Ann. Math. Statist., 38*, pp. 475-481.
85. Woodroofe, M. (1968), "On choosing a delta sequence," Technical Report No. 10, Department of Statistics, Carnegie-Mellon University, Pittsburgh, Pennsylvania. (Abstracted, *Ann. Math. Statist., 39*, pp. 700.)
86. Yaglom, A. M. "The influence of fluctuations in energy dissipationoon the shape of turbulence characteristics in the inertial interval" Soviet Physics–Doklady 11, 1966, 26-29.
87. Yakowitz, S., "A Consistent Estimator for the Identification of Finite Mixtures", Ann. Math. Stat., Vol. 40, p. 1728, 1969.