J. Inst. Maths Applics (1977) **20**, 335–354

A Review of Some Non-parametric Methods of Density Estimation

M. J. FRYER
Mathematics Department, University of Essex

[Received 13 October 1975 and in revised form 13 July 1976]

This paper lists and reviews most of the many papers published on the subject of density estimation. The four main categories of estimators (kernel, spline, orthogonal series and histogram) are compared not only for their theoretical properties but also for their applicability to real life problems.

1. Introduction

DENSITY estimation is possibly the most important topic in applied statistics, for unless we know the density $f(x)$ (in which case we are in the realms of probability) we must infer its characteristics from a sample $X_1, \ldots, X_n$ before we can make predictions. It is for this reason that in elementary texts the histogram plays such a central rôle. Not only are we told that in the limit areas tend to probability but also shown how best to form a histogram from (usually univariate) data, balancing the numbers of observations against the class widths. We are taught to check on the range, the shape, the skewness and to note any tendency to multimodality, the last being often very difficult to ascertain because of the subjectivity involved in drawing a histogram. This initial screening of the data often leads us to hypothesize that the data come from one of a particular *parametric* family of density curves and then the process of estimating and hypothesis testing tells us whether or not this hypothesis is tenable.

It is somewhat surprising to find it was not until 1951 that some real improvement was suggested on the time honoured method of producing a histogram, reducing the subjectivity to some extent. This soon led to estimates of $f(x)$ which are continuous and so to some extent could short circuit the usual inference chain of parametric density estimation.

Whilst for univariate data it is relatively easy to produce histograms and modify them until one "looks right", for bivariate data to produce a "good" histogram is very tedious and time consuming. It is in this situation that the continuous non-parametric bivariate density estimate comes into its own—for with the aid of computer graphics we can plot contours or perspective views of $f(x, y)$ and visually choose that picture which looks "best"—contours being much easier to digest visually than a bivariate histogram.

In the later literature more and more applications have been made in which the whole process has been computerized, ending in a single "best" $f(x)$.

The types of estimators suggested fall roughly into four categories, kernel (or window), spline, orthogonal series and histogram-type estimators, although there is inevitably some overlap. In the following sections, the papers have been grouped by subject content and are not necessarily in chronological order.

2. The Kernel Estimator

In 1951 Fix & Hodges (1951) wrote a paper on non-parametric discrimination, in which they required an estimate of a univariate density. Rather than assume an underlying normal distribution or choose the usual histogram as an estimate (being very subjective), they managed to eliminate the "starting position" problem by using a "running histogram". That is, subjectively they chose an interval width h and then estimated the density at any given point as being proportional to the number of observations falling within an interval of width h, centred at the point under consideration. They then went on to consider several alternative estimators, but it was this running histogram or naïve estimator which led Rosenblatt (1956) to define a class of univariate estimators, known as kernel or window estimators, which can be written as

$$\hat{f}_n(x) = \frac{1}{nh}\sum_{i=1}^{n} k_n\left(\frac{x-X_i}{h}\right)$$

where $X_1, \ldots, X_n$ are assumed to be i.i.d. with $f(\cdot)$ the unknown density and $k_n(\cdot)$ the kernel. For the naïve estimator $k_n(x) = \frac{1}{2}$ for $|x| < 1$.

The larger the value of h, the coarser the grouping, so that as n becomes larger, h can (and should) become smaller. If $k_n(\cdot)$ is symmetric and satisfies $\int k_n(u)\,du = 1$, $\int k_n(u)^2\,du < \infty$, $\int k_n(u)|u|^3\,du < \infty$ Rosenblatt showed the class of estimators to be pointwise and integratedly consistent in quadratic mean (MSE, MISE) provided $h = h_n$ is chosen suitably. In fact he showed that if we write $h = Pn^{-\alpha}$, the optimum choice for α is $\frac{1}{5}$, but P_{opt} should be a function of $f(x)$ and its derivates!—an impossible situation, but it does show that the optimum rate of convergence of the MSE or MISE is $O(n^{-\frac{4}{5}})$. Note that some properties of $k_n(\cdot)$ are transferred to $\hat{f}_n(\cdot)$, e.g. if $k_n(\cdot)$ is smooth so is $\hat{f}_n(\cdot)$, and similarly if $k_n(\cdot)$ is a density so is $\hat{f}_n(\cdot)$. The fact that there is no uniformly unbiased estimator for all continuous densities is not surprising though rather disappointing.

Parzen (1962) imposing further constraints on $k_n(\cdot)$, showed asymptotic unbiasedness, and then listed seven forms for the kernel which satisfy these constraints, amongst which are the rectangular, triangular, normal and Cauchy density functions. If $nh_n \to \infty$, Parzen showed MSE $\to 0$ and under further conditions showed the asymptotic normality of $\{\hat{f}_n(x)\}$ (for fixed x).

Demonstrations of the effect of varying h and $f(\cdot)$ for a normal kernel are to be found in Fryer (1971).

Many authors have followed in Parzen's footsteps, changing the assumptions about $f(x)$ and the conditions imposed on $k(\cdot)$ and $\{h_n\}$ and proving consistency properties for $\hat{f}_n(\cdot)$. Amongst them are Nadaraye (1963, 1965), Murthy (1965*a*), Woodroofe (1967), Bhattacharrya (1967), Schuster (1969, 1970) and Silverman (1977*a*). Craswell (1965) generalizes Parzen's results to estimation on a topological group, whereas Borwanker (1971) considers strictly stationary processes.

Other authors concern themselves with finding (asymptotically) optimal forms for $k(\cdot)$. Bartlett (1963), for example, proposes

$$k(u) = \frac{9}{8h}(1 - 5u^2/3h^2) \text{ for } |u| \leqslant h$$
$$= 0 \text{ otherwise,}$$

which optimizes a larger group of terms in the asymptotic expansion of the MSE. This kernel is of course negative for $|u| > h\sqrt{0{\cdot}6}$ and hence the corresponding $\hat{f}_n(\cdot)$ is not a density.

Watson & Leadbetter (1963) use the MISE as a criterion and arrive at

$$\Phi_k(t) = |\Phi_f(t)|^2 \Big/ \left(\frac{1}{n} + [(n-1)/n]/|\Phi_f(t)|^2\right),$$

where $\Phi_g(t)$ is the Fourier transform of $g(\cdot)$, assuming Φ_f to be square integrable. Hence the form of k again depends on f. They demonstrate the optimal form for k corresponding to various f and show that the MISE cannot be better than $O(n^{-1})$. If the rate of decrease of Φ_f is known, they give estimators which are asymptotically optimal.

Woodroofe (1968) presents a 2-stage procedure to estimate $f(\cdot)$ when the kernel has been specified. After two initial guesses for h_n which are used to obtain rough estimates for f and $f^{(r)}$ respectively (where r is the first non-vanishing moment of k) a new value, h_n, for h is computed. This h_n is used to estimate $f(\cdot)$ in the usual way. Woodroofe shows that this method converges asymptotically in MSE. Involved in this process, however, is yet another sequence which is defined just as vaguely as $\{h_n\}$: fortunately $\hat{f}_n(\cdot)$ is not very sensitive to its value. Although this method can be shown to work reasonably well, it involves much computing and no multivariate extension has as yet been proposed. Nadaraye (1974) provides a similar 2-stage procedure but based on a different optimality criterion.

Pickands (1969) presents another "self contained" estimator for a specific (large) class of densities, but admits that it is considerably more difficult to compute even than that of Woodroofe.

Whittle (1958) suggests a linear estimator of $f(x)$ of the form $\hat{f}(x) = \Sigma w_x(x_j)/N$, where w is a weight function to be optimized and then considers a Poissonization of the problem $N \sim$ Poi (mean M) in which he estimates $\phi(x) = Mf(x)$ by

$$\hat{\phi}(x) = \Sigma w_x(x_j)M/N,$$

using as optimization criterion $\min E_p E_s|\hat{\phi}(x) - \phi(x)|^2$ where the suffices refer to the prior distribution of ordinates and sampling fluctuations respectively. He obtains a rather complicated integral equation for w, which has a δ-function as asymptotic solution. One important property enjoyed by this estimator is that smoothings applied to different scalings of the data (non-singular transformations of the variate scale) are equivalent. The asymptotic behaviour of the MSE is also discussed. Several criticisms of this paper, including that the estimates are not constrained to be non-negative, are considered in Dickey (1968*b*), and some modifications to the estimator are suggested.

Anderson (1969*a*,*b*), after a fairly extensive study, concludes that the actual kernel function $k(\cdot)$ used makes little difference to the optimum value of the MISE, but that the optimal value of h_n differs for different kernels. The normal kernel is shown to perform satisfactorily when estimating normal and relatively non-skew densities, but not when estimating the negative exponential. This facet has been taken up by Ojo (1974), Fryer (1976) and Copas & Fryer (1977), in which it is recommended that skew data should be transformed nearer to symmetry before the estimation procedure, and

then the resulting estimate transformed back. Fryer also considers the robustness of the normal kernel and provides plots of the sensitivity of the MSE and MISE to variations in h and $f(\cdot)$, one of the main conclusions being that it is usually preferable to over- rather than under-estimate h.

Nadaraye (1964*b*) and Watson (1964) are concerned with estimating the regression curve of Y on X,

$$E(Y/X) = \frac{\int yf(x, y)\,dy}{\int f(x, y)\,dy} \equiv m(x).$$

As estimator they both choose

$$\hat{m}(x) = \frac{\sum^{n} Y_i k[(x-X_i)/h]}{\sum^{n} k[(x-X_i)/h]}$$

with a symmetric $k(\cdot)$ and prove some asymptotic results. Watson goes on to give the results of some Monte Carlo experiments. For example he compares these estimates with the true value when $m(x) = 0{\cdot}8x$. He also computes $\hat{f}_n(\cdot)$, uses the best estimate to fix h_n, and concludes (as Anderson) that it is h_n and not $k(\cdot)$ which is important in the estimate. He applies the results to a practical problem in biology.

The convergence properties of the obvious estimates of $\int f^2\,dx$ are considered in Bhattacharrya & Roussas (1969) and Schuster (1974).

In a later paper, Nadaraye (1965) considers the regression problem of Y on X where $X = Y+Z$ and $Z \sim N(0, \sigma^2)$ but Y is of unknown density. Since $E(Y/x) = \sigma^2 f'(x)/f(x)+x$, he proposes $\hat{m}(x) = \sigma^2\psi_n(x)/\hat{f}_n(x)+x$, where $\psi_n(x) = (\hat{f}_n(x+h)-\hat{f}_n(x-h))/2h$ and proves some consistency properties with $h_n = n^{-\theta}$, $0 < \theta < \frac{1}{2}$.

Another group of papers is concerned with estimating the hazard function $Z(x) = f(x)/(1-F(x))$. Watson & Leadbetter (1964*a*,*b*) use as estimators

$$\hat{f}_n(x)\Big/\left(1-\int_0^x f_n(t)\,dt\right) \quad \text{and} \quad \hat{f}_n(x)/[1-F_n(x)],$$

where $F_n(x)$ is the proportion of observations $\leqslant x$, and show them both to be asymptotically unbiased under suitable conditions. In a numerical example a triangular window is shown to perform nearly as well as the optimal window even for such a skew density as the exponential.

Copas & Fryer (1977) use the second of the above estimators for the age specific suicide rate in a study to test the significance of an apparently increased rate near the commencement of the time on test. Since the histogram of data is very skewed (and despite the above observation) a log transformation is made, the density estimated and then transformed back. This eliminates the obvious inaccuracy near the origin when estimating an exponential-type distribution. Some further theoretical results are given in Murthy (1965*b*). Significance tests and confidence intervals are dealt with at length in Nadaraye (1970), Bickel & Rosenblatt (1973) and Rosenblatt (1975).

3. The Multivariate Form

$$\hat{f}_n(x_1 \ldots x_p) = \frac{1}{nh_1 \ldots h_p} \sum_{i=1}^{n} K_n[(x_1-X_{1i})/h_1, (x_2-X_{2i})/h_2, \ldots, (x_p-X_{pi})/h_p].$$

Rosenblatt's results for the naïve estimator are extended to the bivariate case by Maniya (1961). He shows the optimal rate of convergence in MSE and MISE is now only of order $n^{-\frac{2}{3}}$, corresponding to $\alpha_1 = \alpha_2 = \frac{1}{6}$. Cacoullos (1964, 1966) obtains the (p-dimensional) multivariate equivalents of much of Parzen's work. In particular $\mathrm{MSE}_{\mathrm{opt}} = O(n^{-4/p+4})$ when $h_1 = h_2 = \ldots = h_p = O(n^{-1/p+4})$. He points out that the product kernel, i.e.

$$K_n(y) = \prod_{i=1}^{p} k_n(y_i),$$

where $k_n(\cdot)$ is a univariate kernel, has stronger invariance properties than the general vector kernel since it is invariant under different scale transformations in each dimension: the property being essential when incommensurable characteristics are being considered (e.g. height, weight). Van Ryzin (1969) strengthens the consistency properties obtained by Cacoullos. He notes that the kernels tabulated by Parzen satisfy all the conditions imposed on $k_n(\cdot)$, and that suitable $\{h_n\}$ are given by $h_n = Cp^{-q}$, $0 < q < \beta/p$, $C > 0$ and $\beta = \frac{1}{2}$ (or, except for the rectangular kernel, $\beta = 1$).

Wertz (1969) and Silverman (1976) are similarly concerned with the consistency properties of the product kernels.

Epanechnikov (1967) also considers the use of (non-negative) product kernels but assumes them, amongst other properties, to have Taylor expansions in all their arguments about each point. Besides proving several types of consistency, he shows that the optimum kernel (in Relative MISE) is:

$$k_n(y) = \frac{3}{4\sqrt{5}}(1 - y^2/5) \text{ for } |y| \leqslant \sqrt{5}, \quad = 0 \text{ elsewhere.}$$

He tabulates the relative efficiencies of various "standard" kernels (e.g. Normal: 0·95; Rectangular: 0·93).

Murthy (1966) extends his previous univariate results to the multivariate case using Cacoullos' vector estimator.

Martz & Krutchkoff (1969) use $k_n(y) = (2 \sin (y/2)/y)^2/2\pi$ (see Parzen, 1962) as the kernel of a product estimator for obtaining the solution to simple multiple regression by Empirical Bayes techniques. They take $h_n = n^{-\frac{1}{3}}$ after standardizing all the data. Rosenblatt (1969) is concerned with estimating both the conditional density and regression associated with a joint density. For the marginal density he takes the integral of the usual bivariate estimator. He then takes the quotient of the joint and marginal estimators to estimate the conditional density, and its mean to estimate the regression function. He proves two theorems concerning the (pointwise) asymptotic behaviour of the estimators.

Silverman (1977*b*) also considers the problem of choosing the h_n appropriate to a given sample in both the univariate and the multivariate cases. His method involves the consideration of "test graphs" of the second derivative of the density estimate for various values of h_n. A theoretical result shows that, under certain conditions, if the sequence h_n is chosen to give the most rapid uniform convergence of the estimates to the true density, then the Laplacian of the estimate will exhibit fluctuations of size

$$\pm K \sup_{(-\infty, \infty)} |\nabla^2 f|$$

about its trend. The constant K depends explicitly on the kernel and the factor $\sup |\nabla^2 f|$ can be estimated from the test graphs. Silverman's method is to choose the estimate corresponding to the test graph which has fluctuations of the right size. Although the method has some drawbacks (notably the occasional difficulty of distinguishing random fluctuation in the test graph from true variation in $\nabla^2 f_n$ and also the problem of the presentation of the test graph in the multivariate case) the examples considered show that the method can work well in practice.

Specht (1967*a*,*b*) brings a refreshingly practical approach to the whole subject when applying the techniques to the problem of classifying vectorcardiographic outputs into one of two categories. Since he is concerned with $p = 46$ dimensions, computer time and space are somewhat at a premium. He has first to estimate the underlying densities using "training samples" and then uses a Bayes' strategy to classify new data. He chooses h_n to give maximum discrimination of the "training sample". He assumes a product normal kernel, but a direct application leads to an inordinate amount of computing. He therefore approximates part of the kernel by a quadratic polynomial function of the observations $P(\mathbf{x})$, whence

$$\hat{f}_n(\mathbf{x}) = \frac{1}{\sigma^4 2\pi} \exp\left(\frac{-\mathbf{x}'\mathbf{x}}{2\sigma^2}\right) . P(\mathbf{x}) \text{ for each of the classes.}$$

He gives an algorithm for calculating the coefficients of $P(\mathbf{x})$, and for reducing their number to just those which have a significant effect on $\hat{f}_n(\mathbf{x})$. He chooses for a given sample that classification for which $p_r l_r P(\mathbf{x})$ is a maximum, where p_r is the *a priori* probability of pattern type r and l_r is the associated loss. The fitting of a polynomial to $k_n(\cdot)$ is shown not to be subject to the overfitting problem that can arise when fitting a polynomial to $f(\cdot)$. Specht (1971) considers the normal kernel, leading to a series expansion for $f(\cdot)$ in terms of Tchebychev–Hermite polynomials, although in practical applications he recommends using a theoretically equivalent power series-type expansion which he claims yields higher accuracies when truncated to the same number of terms, and is computationally simpler.

$$\hat{f}_n(x) = N(0, \sigma^2) \sum_{r=0}^{\infty} C_r x^r$$

where

$$C_r = \frac{1}{r!} \frac{1}{\sigma^{2r}} \frac{1}{n} \sum X_i^r \exp\left(\frac{-x_i^2}{2\sigma^2}\right),$$

and $N(0, 1)$ is the standardized normal density function.

This estimator is essentially non-negative, and the round-off error problem of Tchebyshev–Hermite polynomials is eliminated since the series consists of positive terms only rather than having alternating signs.

Konakov (1973) suggests the kernel $(\pi x)^{-1} \sin x$ (the Fourier integral estimate) and shows that under stated regularity conditions on f and convergence properties for $\{h_n\}$, the estimator is asymptotically unbiased, consistent and asymptotically normal.

Davis (1974, 1975) shows that the rate of decrease of the bias for this estimator of a "smooth" f is much faster than for kernels belonging to $L^1(-\infty, \infty)$ (e.g. when $f = N(\mu, \sigma^2)$).

Fukunaga & Kessel (1971) compare the classification errors associated with the parametric and non-parametric estimation of the underlying multivariate densities when using a Bayes' classification rule. After some experimentation with h_n (not very sensitive in their $p = 8$ dimensional problem(?)) they find that an estimator with a multivariate normal kernel performs only slightly less well than the parametric estimator when it is known that the density does in fact come from the parametric family.

Glick (1972) also considers the classification problem and its associated non-error rate.

Habbema *et al.* (1974*a*,*b*) use the multivariate product kernel estimators in discriminant analysis. However, in Habbema *et al.* (1974*a*) they propose a novel approach to estimating h_n. After the data have been standardized to have sample variance 1, a modification of the likelihood function is maximized (in each series of the product, they leave out that term which makes the likelihood infinite with h_n zero) to give an estimate to h_n. This criterion, although not based on any theoretical reasoning, seems to give reasonable results. The allocation rule used is the Bayes' minimum expected loss criterion. As they point out, this technique can only be used in quite small scale problems due to the large amount of computing involved especially in the teaching stage (Hermans & Habbema, 1976). This technique is used and compared in Hermans & Habbema (1975).

TABLE 1

A check on the optimality of Habbema's estimator

	Habbema		max MSE		Theoretical optimum	
n	mean h_n	mean ASE	mean h_{no}	mean ASE	h_{nT}	mean ASE
10	0·706	0·0180	0·735	0·0038	0·759	0·0049
20	0·601	0·0074	0·604	0·0027	0·642	0·0033

In order to test the optimality of Habbema's estimator, I took 1000 samples each of $n = 10, 20$ drawn from a univariate normal population, and using a normal kernel obtained an estimate $\hat{h}_n$ for h_n. At the same time, using the sample average square error (ASE: estimating the MISE) as criterion, I found the optimal value h_{no} for h_n, the corresponding ASE, and, for interest, the ASE corresponding to h_{nT}, the theoretical optimal value for h_n. The results are summarized in Table 1.

Habbema's method is seen to produce significantly worse results than those obtained by using a fixed h_{nT}, presumably because of the consistent underestimation involved. Overestimation by the same amount would have had a less severe effect (see Fryer, 1976).

4. Splines

Boneva *et al.* (1971) present an extremely lucid and interesting paper (with discussion) introducing a transformation of the data to be used for diagnostic rather than estimation purposes. They start with the data in the form of a frequency function (H)

(with the cell width, ε, specified) and find a 1–1, linear invariant, and bi-continuous mapping onto a Hilbert space of smooth functions, the resulting form being called a histospline ($s \in S$), so named as S turns out to consist of those continuous and continuously differentiable functions such that s is a quadratic in each fixed interval, and is square integrable i.e. a quadratic spline. They are also constrained to integrate to the same as H over every cell but are not necessarily non-negative. The δ-spline ($k(\,:\varepsilon)$) is the unit from which the histospline is built and can be quite easily stored on a computer requiring a *maximum* of 39 constants (Boneva *et al.*, 1975). Both theoretical and practical suggestions are made as to how the histospline should be modified to allow for densities defined on finite intervals or on the circle. They then modify their transformation to apply to raw data in much the same way as the naïve estimator is a modification of the histogram—the spline transformation

$$\sigma(x) = \sum_{i=1}^{n} k(x - x_i : \varepsilon)$$

which is of kernel-type and so enjoys the attributes of that group of estimators. A bivariate analogue is also considered.

Wahba (1975*b*) considers the statistical properties of a slight generalization of the histospline for densities of finite support. Instead of the second derivative being assumed zero at the ends of the interval their values can be estimated from the data. An upper limit for the MSE is obtained under stated conditions (U.S.C.) and shown, not surprisingly, to have the same asymptotic rate of consistency as the usual kernel estimators.

Wahba (1971) considers a local estimate of $f(x)$ (at x) based on the derivative of an mth degree polynomial estimate of $F(x)$ in the neighbourhood of x obtained by Lagrangean interpolation formulae. The estimator depends on the number of data points in the neighbourhood and m, which is related to the continuity assumed for f and its derivatives. The estimator is shown to be pointwise consistent in MSE at a slightly faster rate than Parzen's (U.S.C.), but under strong conditions to have the same (higher) rate. It is noted that a special case of this estimator is given in Van Ryzin (1970).

5. Series Estimators

In 1962, the following estimator was proposed by Čencov (1962*a*,*b*), the same formulation being used by later authors. Suppose $x_1, \ldots, x_n$ are i.i.d. $\sim f(x)$, $x \in \mathbb{R}$, and $r(\cdot)$ is a (fixed) weight function so that the inner product

$$(\phi, \psi) = \int_{\mathbb{R}} \phi(x)\psi(x)r(x)\,dx$$

defines a Hilbert space $L_2(r)$, and suppose an orthonormal basis $\{\phi_{KN}\}$ exists for the N-dimensional subspace E_N, then

$$f_N(x) = \sum_{K=1}^{N} a_{KN}\phi_{KN} \equiv \sum_{k=1}^{N} (\phi_{KN}, f)\phi_{KN}$$

is the mean square approximation to $f(x)$.

Now

$$a^*_{KN} = \frac{1}{n}\sum_{K=1}^{n} \phi_{KN}(x_i)r(x_i)$$

is a strongly consistent estimator for a_{KN} and so $\hat{f}_n(x) = \Sigma a^*_{KN}\phi_{KN}$ is proposed as an estimator for $f(x)$. The choice of E_N and n both contribute to the closeness of $\hat{f}_n(x)$ to $f(x)$.

Čencov proves several theorems relating to the degree of approximation and proposes a "stopping rule" for N, the number of terms of the series. This estimator is not necessarily always non-negative.

Schwartz (1967) considers the case $r(x) = 1$, $N = N(n)$ and $\phi_k(\cdot)$ the kth Hermite function (over the real line). He proves several consistency properties, requiring conditions such as $N/n \to 0$ as $N \to \infty$ (c.f. $1/h_n$).

Blaydon (1967) in a generalization considers estimating both $F(\cdot)$ and $f(\cdot)$ by a linear combination of functions using the criterion of minimum least squares. Kashyap & Blaydon (1968) evaluate a^*_{KN} by a gradient-type technique, and give an example using the first three Laguerre polynomials over [0, 4] to estimate the distribution function corresponding to an exponential density and compare the speed of convergence of the various algorithms.

Watson (1969) introduces a general weight function λ into the estimator:

$$\hat{f}_n = \sum_{K=1}^{\infty} \lambda_{K(N)} a_{KN}\phi_{KN},$$

with the intention of improving upon Čencov's results, but is forced to conclude that $\lambda_{K(N)} = 1$ for $K = 0, \ldots, N$, $= 0$ elsewhere is perhaps the best experimental form!

Kronmal & Tarter (1968) and Tarter & Kronmal (1967) extend the Čencov model to cover both $F(\cdot)$ and $f(\cdot)$ but only with finite support. They choose the trigonometric functions $\{\cos K\pi x\}$, $\{\sin K\pi x\}$ and $\{\cos K\pi x, \sin K\pi x\}$ for their ϕ's since they require the orthogonal series used for $F(\cdot)$ still to be orthogonal when differentiated to give $f(\cdot)$. They defend negative estimates by saying that they are a warning that insufficient data are available to provide estimates in those regions. They propose the following "stopping rule" for the $\{\cos K\pi x, \sin K\pi x\}$ series: the mth term should be included iff

$$\sum_{i}^{n}\sum_{j}^{n} \cos m\pi\left(\frac{x_i - x_j}{b-a}\right)\Big/ n^2 \geqslant 2/(n+1)$$

(as opposed to $1/(n+1)$ proposed by Čencov).

Fellner (1974) in a synthesis of the papers by Whittle (1958), Tarter & Kronmal (1967) and Kronmal & Tarter (1968) produces a multistep estimating procedure which by-passes the usual "stopping rule" problem by using an hypothesis testing technique. Crain (1974) proposes a maximum likelihood approach to estimating the coefficients of the orthogonal series (finite support), demonstrating his results with examples using Legendre polynomials.

Several authors (including Watson, 1969; Fellner & Tarter, 1971; and Tarter & Raman, 1971) note the theoretical equivalence (U.S.C.) of the Fourier and Kernel estimators, but it is all too evident in practice that the resulting estimates are (locally) very different, especially when n is relatively small.

In a brief comparison with the results of Parzen (1962) and Rosenblatt (1956), Schwartz (1967) finds that stronger conditions are required on $f(\cdot)$ for the same asymptotic rate of convergence in the univariate case, but as the rate of convergence of his estimator does not depend on ρ, he concludes that it is (theoretically) preferable in higher dimensions. Kashyap & Blaydon (1968) note two operational advantages of the series estimator:

(i) it does not require all samples to be stored during computation; and

(ii) the final result is easy to store, since it is not in the form of a complicated analytic function.

Anderson (1969*a*,*b*) considers trigonometric, Hermite and Laguerre ϕ's in a Monte Carlo study using his own truncation rule. His results indicate that the cosine trigonometric series estimator is probably the best (when estimating a function with finite support), and because of its smaller computing requirements, superior in practice (but not in MISE) to the kernel estimates.

Wegman (1972*b*) also in a Monte Carlo study, finds the performances of the kernel estimates (represented by the triangular kernel) and the series estimators (represented by the cosine series) to be about the same in terms of either the average square error or the likelihood.

Schwartz (1969) considers an estimator for the convolution density $f(x)$ of $X = N+Z$ where N and Z are independent with $N \sim N(0, \sigma^2)$ and Z has unknown density. The Hermite polynomials arise naturally in this context and lead to an MISE of order $\ln(n)/n$.

Tarter & Silvers (1975) introduce a new application of density estimation—a cluster-type analysis in which the contours of the (estimated) bivariate population density (or bivariate marginal density) (of finite support) are interpreted subjectively and modified interactively with the aid of an on-line computer CRT terminal. The estimator used is as in Tarter & Kronmal (1970). If one assumes the underlying distributions to be a mix of bivariate normals, then the multiplication of the formula for the estimated density by a factor which changes the covariance structure will change the shape of the estimated contours. The idea behind the algorithm presented is to modify the structure so that the contours of one of the mix (usually the "strongest") tend to straight lines and hence separate from the rest of the mix. The parameters of the corresponding normal distribution are then estimated graphically. The process is repeated until all the elements of the mix are estimated. This technique is obviously more robust to changes in the underlying mix of densities than methods based solely on parameter estimation in mixture decomposition.

Aizerman *et al.* (1964*b*) and Cooper (1964) apply series estimators to estimating the reliability of classification procedures.

6. Maximum Likelihood Estimators

Grenander (1956) was the first to derive the MLE for a non-increasing density $f(\cdot)$, corresponding to an absolutely continuous $F(\cdot)$. It arose from studies of the force of mortality (age specific death rate) determined from mortality tables. He showed $\hat{f}_n(\cdot)$ to be a step function, the derivative of the greatest convex minorant of the empirical distribution function $F_n(x)$.

Marshall & Proschan (1965) consider MLE's in the context of a non-decreasing hazard rate and show functionally uniform consistency with probability 1 (U.S.C.).

Robertson (1967) is concerned with finding the MLE for a unimodal density measurable on a σ-lattice L of subsets of $\mathbb{R}$ and subject to various other conditions. He assumes that the position of the mode is known ($x = a$), and finds the MLE to be

$$\hat{f}_n(x) = E_\mu\left[\sum_{i=1}^{n} \frac{n_i}{n} \frac{I_{Ai}}{\mu(A_i)}\middle| \mathscr{L}n\right]$$

where n_i/n is the relative frequency of observations occurring in the interval $[x_{(i)}, x_{(i+1)})$ if $x_{(i+1)} < a$, and $(x_{(i)}, x_{(i+1)}]$ if $x_{(i)} > a$ of length A_i, and I_{Ai} is the indicator function. He shows pointwise consistency with Pr 1 and uniform consistency with Pr 1 over any (finite) closed interval in which $f(\cdot)$ is continuous.

Weiss & Wolfowitz (1967) consider three classes of densities to which $f(\cdot)$ may belong (the classes being defined mainly by the existence of derivatives in the intervals) and obtain as MLE

$$\hat{f}_n(x) = n_i/\{n[2\varepsilon_n + k(\varepsilon_n)]\} \text{ for } x \text{ in } I_i = A \pm \varepsilon_n$$

where $\varepsilon_n = n^{-\alpha}$, $\alpha \leqslant \frac{1}{3}$ and k is a function of ε_n, and $f(\cdot)$ assumed known. Since k will not be known in general, they suggest replacing it by $\hat{k} = 2\hat{k}_2\varepsilon_n^3/3$ where $\hat{k}_2 = 12n^2\beta(n\beta Q_n - \frac{1}{2})$,

$$Q_n = \sum_{x_i \in J_i} (x_i - A)/n_i,$$

$0 < \beta < \frac{1}{3}$, $\beta < \alpha$ and J_i is the interval $(A \pm n^{-\beta})$. They show that these estimators are more efficient than the usual histogram estimates.

Rao (1969) also considers estimating a unimodal density in the case when the mode is assumed known. Following Grenander (1956) he shows the MLE to be the slope of the greatest convex minorant of $F_n(x)$ to the left of the mode and of the least concave majorant to the right. Pointwise consistency in probability is established and the asymptotic distribution derived.

Wegman (1969*b*,*c*,1970*a*,*b*) takes Robertson's results one stage further by not assuming the modal position known, but estimated by one of the strongly consistent estimators of Venter (1967) or Nadaraye (1965). This estimator has the tendency to peak too severely around the mode, so in a follow-up Wegman modifies the estimator by fixing the width of the modal interval in advance. He goes on to prove some consistency properties and in the final paper derives the asymptotic distribution. An alternative method of tackling this problem is proposed by McGilchrist (1975). Good (1971) and Good & Gaskins (1971, 1972) consider maximizing a score $\omega = L - \Phi(f)$, where L is the sample log likelihood and Φ a non-negative roughness penalty functional (flamboyancy functional) of $f(\cdot)$. (An optimization of L alone just results in δ-functions at each of the data points.) They prove pointwise consistency in probability (U.S.C.). They take

$$\Phi(f) = 4\alpha\int\gamma'^2\,dx + \beta\int\gamma''^2\,dx,$$

where $f = \gamma^2$, $\alpha \geqslant 0$, $\beta \geqslant 0$, $\alpha + \beta > 0$ and assume

$$\gamma(x) = \sum_{m=0} \gamma_m\phi_m(x),$$

where γ_m are real coefficients and $\phi_m(x)$ are Hermite polynomials. This leads to a set of simultaneous non-linear equations to be solved iteratively for $\gamma_1, \gamma_2, \ldots, \gamma_R$, where $\gamma_{R+1} \ldots$ are assumed zero. An iterative method for estimating β is proposed and from graphical considerations a value of zero for α is often assumed. A natural extension of Φ to multivariate densities is given and an invariant formulation discussed. A formulation suitable for data in the form of histograms is also given. A comparison with other methods and a condensed literature review is included with a bibliography. The main points in the comparison are:

(i) this method estimates the true density and irons out non-significant bumps (as opposed to splines);

(ii) is more widely applicable than Fourier series methods;

(iii) is non-negative (as opposed to Fourier series methods); and

(iv) is perhaps more efficient in the use of observations than kernel methods.

7. Histogram-type Estimators

Just as the MLE estimators were histograms with their intervals chosen in an optimal fashion, so many other histogram-type estimators have been proposed, (usually) requiring the sample to be ordered.

The first of these was by Fix & Hodges (1951, 1952), the interval widths being defined before the relative frequencies calculated (as with the usual histogram). Discussion follows as to the rate of decrease of these widths to obtain consistency.

Loftsgaarden & Quesenberry (1965) estimate the (multivariate) density in the reverse order—fix the number of observations per interval and then draw the smallest (spherical) intervals to include this number of points around each point of interest. They prove pointwise consistency and find the asymptotic distribution. This estimator is simple to calculate when only point estimates are required (as in discriminant analysis) but gives computationally complicated results when the density is required over a region in more than one dimension. It is somewhat akin to the naïve estimator in this respect.

Elkins (1968) considers a "cubic" estimator, which counts the number of observations within a cube of side $2h$ (a multivariate naïve estimator), and the corresponding "spherical" estimator (radius r). He finds h and r to optimize the MSE, evaluates the MISE and concludes that the spherical estimator is to be preferred as the number of dimensions increases. The paper includes a brief review of the subject.

Moore & Henrichon (1969) consider a univariate modification of that given in Loftsgaarden & Quesenberry (1965) in which the estimator is a step function with the discontinuities at the observations only, hence being much easier to compute. They prove uniform convergence in probability provided f is positive and uniformly continuous on $\mathbb{R}$.

Pelto (1969) uses the spherical results of Elkins (1968) in the problem of assigning a p-dimensional random variable to one of two populations determined only by observations. He uses the "leaving-one-out" method to estimate the probabilities of misclassification (as a function of r). The expected loss is then plotted and the value of r corresponding to the minimum used. Sampling experiments on multivariate

normal populations indicate that the efficiency of the method is about the same as of the standard parametric procedure, in which the form of the density is assumed known.

Gessaman (1970) provides an estimator for which $\hat{f}_n(\cdot)$ is constant over rectangular blocks, and is given by

$$\hat{f}_n(x) = k_n/[(n+1)Ax_n]$$

where $k_n = [n^{\frac{1}{2}}]$ and Ax_n is the area of the block containing x. The algorithm requires the ordering of the sample to split up the plane into rectangular blocks. Pointwise consistency in probability is proved (U.S.C.).

Mucciardi & Gose (1970) describe a fully automated algorithm based on Sebestyen & Edie (1966) for non-parametric cluster analysis using hyperellipsoidal cells. A new observation defines a new cluster cell provided it neither falls within an existing cell nor in a "guard zone" surrounding an existing cell, the centres and shapes of the cells being updated as more data are collected. A second pass over the data "refines" the clusters. Provision is made for combining nearly empty cells with their neighbours. Applications to medical data in 80 and 36 dimensions are given, together with corresponding results using different models for the density. An analysis of classification errors demonstrates the usefulness of the algorithm.

Van Ryzin (1970) uses a modification of the Loftsgaarden & Quesenberry (1965) estimator and besides strong consistency (U.S.C.) he gives asymptotic confidence intervals for $\hat{f}_n(x)$. In the final sections of the paper, he asserts that the estimates presented compare quite favourably with the kernel estimators, and under certain circumstances (e.g. where estimating the tails of $f(x) = N(0, 1)$ with a rectangular kernel) can be much more efficient.

8. General

Farrell (1967, 1972) obtains several theoretical results concerning the general estimation problem. He considers a sequence of estimators $\delta N(x_1, \ldots, x_N)$ of $f(0)$, and using a square error loss function shows that no uniformly consistent sequence of estimators exists relative to the class $C_\alpha = \{F(x) : F(x) \in \mathbb{R}, F' = f$ and $F'' = f'$ are defined and continuous on $\mathbb{R}$, $\sup f = \alpha\}$. He then finds the optimum asymptotic rate of convergence of $\hat{f}_n$ to $f(0)$ (U.S.C.) as attained in Epanechnikov (1967). Wahba (1975*a*) modifies Farrell's theorem and then evaluates the "rate constant" of several types of estimates achieving the optimal rate of convergence (in MSE): the kernel estimate, the trigonometric series of Kronmal & Tarter (1968) and Tarter & Kronmal (1967), the polynomial method of Wahba (1971) and the usual histogram method.

9. Comparisons and Reviews

Văduva (1968) comes nearest to a text book on the state of the art (as in 1967), but unfortunately it is in Rumanian!

Anderson (1969*a*,*b*) reviews the literature (to 1969) on both kernel and orthogonal series estimates. He considers three kernels (normal, double exponential and uniform) and the estimation of four densities (gamma, exponential, uniform, mix of normals). He concludes that for estimating a fixed $f(\cdot)$, the actual optimum values of the MISE

are relatively independent of the kernels, but that on the other hand the corresponding values of h are very different. For the orthogonal series estimates, he considers the trigonometric, Hermite and Laguerre types and finds their (theoretical) MISE when estimating the normal and gamma densities. The final chapter gives the results of some Monte Carlo comparisons using an estimate of the MISE as criterion. The normal kernel is used to represent the family of kernel estimates using the theoretically optimum h_n. Results using Woodroofe's two-stage procedure are also included but not compared. All the estimators considered seem to perform reasonably satisfactorily when estimating the normal and not too skew members of the gamma family, but not when estimating the exponential. In no case did the kernel estimates display significantly smaller MISE than the orthogonal ones when his own stopping rule was used. He concludes that the cosine trigonometric estimator is probably the best of the series type and because of its smaller computing requirements, superior in practice to the kernel estimators.

The poor performance relative to the exponential density led Ojo (1974) to consider transforming the data prior to estimating (with a kernel estimate) and then transforming back. He concludes that this procedure gives superior estimates (in MISE) and almost eliminates the obvious bias near the origin.

Van Ryzin (1966), in considering the theoretical problem of classifying outcomes into one of two categories by a Bayes' risk criterion, estimates the underlying densities with training samples by three methods: (i) finite orthonormal expansion, (ii) the Čencov series, and (iii) the kernel method. He derives results concerning asymptotic rates of convergence which might help one to decide on the best method to use in a given classification problem.

Gessaman & Gessaman (1972) give the results of a Monte Carlo study to compare various estimators in both forced and partial discrimination problems. The non-parametric estimators chosen were (i) the bivariate normal kernel, (ii) the Loftsgaarden & Quesenberry (1965) estimator and (iii) the Gessaman (1970) estimator. They conclude that amongst these (iii) does least well in the forced comparison situation.

Rosenblatt (1971) gives a review of some theoretical topics concerned with kernel and orthogonal series estimators and in particular describes the relationships between the results for density estimation and spectral analysis. He agrees with the views expressed by Boneva *et al.* (1971) when he writes "In most cases it is clear that one will not use the techniques . . . in estimating a density function unless there is a good deal of data . . . , little *a priori* information . . . , but a great need to get additional information about the density function, even if it is fairly crude". He adds the caveat to the theoretical results "it is a mistake to take asymptotic results too literally from a finite sample point of view". He derives Epanechnikov's non-negative optimal kernel again and suggests another kernel, this time somewhere negative, which has better asymptotic properties. (Bias $= O(h_n^4)$.)

Wegman (1972*a*) gives a straight résumé of all the methods to 1972 and a large, but by no means complete bibliography. He lists the various types of convergence involved and is quite detailed (in some cases) in the conditions imposed on $f(\cdot)$ and $\hat{f}_n(\cdot)$ in many of the theorems. In his next paper (1972*b*), however, Wegman gives some results of Monte Carlo trials to test the practical effectiveness of the estimators in the light of the conclusions drawn by Anderson. He uses the naïve estimator to

represent the kernel estimators (why not Bartlett's or Epanechnikov's?) and uses the theoretical optimal value of h_n for each of the given densities (when possible). In the cases where this procedure was not possible, an experimental MISE (the ASE) was minimized. (Since h_n is a statistic, I would have preferred it also to be optimized for each particular sample in order to be able to compare it satisfactorily with other sample-based-estimators.) To represent the series type, the cosine series and associated stopping rule was used, since it was shown to behave well.

Two forms of histogram were used, the one in which the number of (equal) class intervals is fixed in advance, and the other in which it is estimated. The densities estimated were uniform, triangular, exponential, Cauchy and normal, the region of support being the sample range and the criteria the average square error (ASE) and the likelihood function. As a crude summary of his results, when using the ASE the naïve and trigonometric estimates proved to be best, followed by the two histogram estimates and, way behind, the MLE estimates. When using the likelihood function the ordering was reversed. He notes the following pros and cons (Table 2).

TABLE 2

Some advantages and disadvantages of the various types of estimator

Orthogonal series	Kernel	Histogram	MLE
For:			
Optimal Stopping Rule exists	Can be density		Few arbitrary choices
Easy to compute	Can have infinite support		Can have infinite support
Good MISE	Good MISE and rate of convergence	Easy to compute Error rate seems higher than $O(n^{-0.8})$	
Against:			
Choice of series critical	Kernel to be chosen although the choice is not critical	Very poor MISE	
Not usually a density	Critical choice of h_n		
Often finite support	Large amount of computer time required	Interval length or number of observations per interval to be chosen	

10. Additional Comments

Any attempt at drawing conclusions or giving recommendations based on a survey such as this seems destined to failure, since one's choice of method must depend to a large extent on the use to which it is going to be put—initial screening of the data at one end of the scale, to a "plug-in" estimator for some taxonomic problem at the other.

Taking my table of Wegman's comments as a basis, there are some further observations I would like to make:

(1) The cosine series estimator seems to perform reasonably in most circumstances.

(2) A quadratic form for the kernel estimator seems to perform well, and there are some theoretical reasons for choosing it. It is basically also easier to compute than the kernels of infinite support. The papers by Specht, however, do provide an efficient algorithm based on the normal kernel.

(3) Several methods exist for estimating h_n in the kernel estimator, although further Monte Carlo studies are required to check on their properties for small n. They all seem to be computationally long, although the Silverman technique (still requiring some subjectivity) is perhaps the shortest. My personal preference (for data screening) is to plot the estimates of $f(\cdot)$ for several values of h_n and choose subjectively—usually taking h just large enough to eliminate bumps at outlying observations. Studies suggest (a) the kernel method performs best for symmetric $f(\cdot)$—i.e. transform to near-symmetry, and (b) over-estimation of h_n is preferable to under-estimation. In some problems such as classification where a teaching sample is used, h_n can be chosen to minimize the classification errors of the teaching sample.

(4) Spline estimators also require the estimation of a parameter corresponding to h_n.

(5) The modified maximum likelihood estimators of Good & Gaskins suffer from the same subjectivity problems although the authors do suggest several "rules of thumb" for the values to be given to the parameters. The iterative nature of their algorithm suggests it to be expensive in computing facilities. However further Monte Carlo studies seem justified.

(6) The histogram-type estimators seem to be non-starters unless a particularly simple step-function form is required.

The author would like to thank a referee for bringing to his attention several references: Silverman (1976, 1977*a*,*b*) and Boneva *et al.* (1975).

References

Aizerman, M. A., Braverman, E. M. & Rozonoer, L. I. 1964*a* *Automn remote Control* **25**, 821–837.

Aizerman, M. A., Braverman, E. M. & Rozonoer, L. I. 1964*b* *Automn remote Control* **26**, 1175–1190.

Aizerman, M. A., Braverman, E. M. & Rozonoer, L. I. 1967 *Automn remote Control* **27**, 7–12.

Anderson, G. D. 1969*a* *A Comparison of Probability Density Estimates.* Presented at IMS Annual meeting, 19–22 August, New York.

Anderson, G. D. 1969*b* *Ph.D. dissertation.* University of Washington.

Bartlett, M. S. 1963 *Sankhyā* (A) **25**, 245–254.

Bartlett, M. S. & Macdonald, P. D. M. 1971 *Nature* **229**, 125–126.

Bhattacharrya, P. K. 1967 *Sankhyā* (A) **29**, 373–382.

Bhattacharrya, G. K. & Roussas, G. G. 1969 *Skand. Aktuarietidskr.* **52**, 201–206.

Bickel, P. J. & Rosenblatt, M. 1973 *Ann. Statist.* **1**, 1071–1095.

Blaydon, C. C. 1967 *Proc. IEEE* **55**, 231–232.

Boneva, L. I., Kendall, D. G. & Lambert, J. A. 1975 *Tech. Report CC*-740 I. Computer Centre, University of Newcastle, NSW, Australia or Stat. Lab., 16 Mill Lane, Cambridge, England.

BONEVA, L. I., KENDALL, D. G. & STEFANOV, I. 1971 *J.R.S.S.* (B) **33**, 1–71.
BORWANKER, J. D. 1971 *Zeit. Wahrscheinlichkeitsth* **20**, 182–188.
CACOULLOS, T. 1964 *Tech. Report No.* 40. Dept. of Statist., University of Minnesota.
CACOULLOS, T. 1966 *Ann. Inst. Statist. Math.* **18**, 179–190.
ČENCOV, N. N. 1962*a Soviet Math.* **3**, 1559–1562.
ČENCOV, N. N. 1962*b Dokl. Akad. Nauk SSSR.* **147**, 45–48.
CHANDA, K. C. 1967 *Bull. Calcutta Statist. Assn.* **16**, 153–163.
CHERNOFF, H. 1964 *Ann. Inst. Statist. Math.* **16**, 31–41.
COOPER, D. B. 1964 *IEEE Trans. Elec. Computers* **13**, 306–307.
COPAS, J. B. & FRYER, M. J. 1977 *On Suicide Rates for Patients under Psychiatric treatment.* Submitted for publication.
CRAIN, B. R. 1974 *Ann. Statist.* **2**, 454–463.
CRASWELL, K. J. 1965 *Ann. Math. Stat.* **36**, 1047–1048.
DAVIS, K. B. 1974 *Doct. dissertation* (unpublished). University of Washington. (Paper submitted for publication.)
DAVIS, K. B. 1975 *Ann. Statist.* **3**, 1025–1030.
DICKEY, J. M. 1968*a Math. Biosc.* **3**, 249–265.
DICKEY, J. M. 1968*b Ann. Math. Stat.* **39**, 561–566.
ELKINS, T. A. 1968 *J. Am. Stat. Assn.* **63**, 1495–1513.
EPANECHNIKOV, V. A. 1967 *Theor. Prob. Appl.* **14**, 153–158. (In translation.)
FARRELL, R. H. 1967 *Ann. Math. Stat.* **38**, 471–474.
FARRELL, R. H. 1972 *Ann. Math. Stat.* **43**, 170–180.
FELLNER, W. H. 1974 *Biometrika* **61**, 485–492.
FELLNER, W. H. & TARTER, M. E. 1971 *Proc. Comp. Sci. and Stat. 5th Annual Symp. on the Interface.* North Hollywood: Western Periodicals.
FIX, E. & HODGES, J. L., Jr. 1951 *Report No.* 4. *Proj. No.* 21-49-004. USAF School of Aviation Medicine, Randolph AFB, Texas.
FIX, E. & HODGES, J. L., Jr. 1952 *Report No.* 11. USAF School of Aviation Medicine, Randolph AFB, Texas.
FRYER, M. J. 1971 *Bull. I.M.A.* **7**, 3–7.
FRYER, M. J. 1976 *J. Inst. Maths Applics* **18**, 371–380.
FUKUNAGA, K. & KESSELL, D. L. 1971 *IEEE Trans. on Computers* C-**20**, 1521–1527.
GESSAMAN, M. P. 1970 *Ann. Math. Stat.* **41** 1344–1346.
GESSAMAN, M. P. & GESSAMAN, P. H. 1972 *J. Am. Stat. Assn.* **67**, 468–472.
GLICK, N. 1972 *J. Am. Stat. Assn.* **67**, 116–122.
GOOD, I. J. 1971 *Nature* **229**, 29–30.
GOOD, I. J. & GASKINS, R. A. 1971 *Biometrika* **58**, 255–277.
GOOD, I. J. & GASKINS, R. A. 1972 *Va J. Sci.* **23**, 171–193.
GRENANDER, U. 1956 *Skan. Aktuarietidskr.* **39**, 125–153.
GUPTA, S. DAS 1964 *Sankhyā* (A) **26**, 25–30.
HABBEMA, J. D. F., HERMANS, J. & VAN DEN BROEK, K. 1974*a Proceedings of Compstat Conference, Vienna.* (Physicaverlag, Wien.)
HABBEMA, J. D. F., HERMANS, J. & VAN DER BURGT, A. T. 1974*b Biometrika* **61**, 313–324.
HERMANS, J. & HABBEMA, J. D. F. 1975 *EDV Med. Biol.* **6**, 14–19.
HERMANS, J. & HABBEMA, J. D. F. 1976 *Manual for the ALLOC discriminant analysis programs.* Obtainable from the authors at Dept. Med. Stats., University of Leiden.
KASHYAP, R. L. & BLAYDON, C. C. 1968 *IEEE Trans. Inf. Theor.* IT-**14**, 549–556.
KIM, B. K. & VAN RYZIN, J. 1975 *Commun. Statist.* **4**, 303–315.
KONAKOV, V. D. 1973 *Theor. Prob. Appl.* **17**, 361–362.
KRONMAL, R. A. 1964 *Doct. dissertation* (unpublished). University of California.
KRONMAL, R. A. & TARTER, M. 1968 *J. Am. Stat. Assn.* **63**, 925–952.
LEADBETTER, M. R. 1963 *Tech. Report* 11. Research Triangle Inst., University of N. Carolina.
LEADBETTER, M. R. & WATSON, C. S. 1962 *Tech. Report S37 No.* 3. Research Triangle Inst., University of N. Carolina.
LIN, PI-ERH, 1968 *Thesis.* Columbia University.

LOFTSGAARDEN, D. O. & QUESENBERRY, C. P. 1965 *Ann. Math. Stat.* **36,** 1049–1051.
LUMEL'SKII, YA. P. & SAPOZHNIKOV, P. N. 1967 *Theor. Prob. Appl.* **14,** 357–364.
MALTZ, C. 1974 *Ann. Statist.* **2,** 359–361.
MANIYA, G. M. 1960 *Trudy Vycial Centra Akad. Nauk Gruzin. SSR* **1,** 75–96.
MANIYA, G. M. 1961 *Soobshch. Akad. Nauk Gruzin. SSR* **27,** 385–390.
MARSHALL, A. W. & PROSCHAN, F. 1965 *Ann. Math. Statist.* **36,** 69–77.
MARTZ, H. F. & KRUTCHKOFF, R. G. 1969 *Biometrika* **56,** 367–374.
MCGILCHRIST, C. A. 1975 *Sankhyā* (A) **37,** 139–149.
MEISEL, W. S. 1971 *Maths in Science and Engineering.* Vol. 83. Academic Press.
MOORE, P. S. & HENRICHON, E. G. 1969 *Ann. Math. Stat.* **40,** 1499–1502.
MUCCIARDI, A. N. & GOSE, E. E. 1970 *Proc. of* 1970 *IEEE Sys. Sci. & Cybernetics Conf.* Pittsburgh.
MURTHY, V. K. 1965*a* *Ann. Math. Stat.* **36,** 1027–1031.
MURTHY, V. K. 1965*b* *Ann. Math. Stat.* **36,** 1032–1040.
MURTHY, V. K. 1966 In *Multivariate Analysis: Proceedings of an International Symposium Held in Dayton, Ohio, 14–19 June 1965* (Ed. Paruchuri R. Krishnaiah). New York: Academic Press.
NADARAYE, E. A. 1963 *Soob. Akad. N. Gruzin SSR* **32,** 277–280. (In Russian.)
NADARAYE, E. A. 1964*a* *Soob. Akad. Nauk Gruzin SSR* **34,** 19–24. (In Russian.)
NADARAYE, E. A. 1964*b* *Primenon* **9,** 157–159. In translation *SIAM Theor. Prob. Appl.* **9,** 141–142.
NADARAYE, E. A. 1964*c* *Soobshch. Akad. Nauk. Gruzin SSR* **36,** 267–268. (In Russian.)
NADARAYE, E. A. 1964*d* *SIAM Theor. Prob. Appl.* **9,** 497–500.
NADARAYE, E. A. 1965 *Teoriya Verojatnost* **10,** 199–203. In translation 1966 *Theor. Prob. Appl.* **10,** 186–190.
NADARAYE, E. A. 1966 *Sakharth. SSR Mecn. Akad. Gamothol. Centr. Srom.* **7,** 35–42.
NADARAYE, E. A. 1968 *Teoriya Verojatnost* **15,** 139–142. In translation 1970 *SIAM Theor. Prob. Appl.* **15,** 134–137.
NADARAYE, E. A. 1970 *Soobsch. Akad. Nauk Gruz. SSR.* **59,** 33–36.
NADARAYE, E. A. 1974 *SIAM Theor. Prob. Appl.* **19,** 133–141.
OJO, M. O. 1974 *M. Phil. dissertation* (unpublished). University of Essex.
PARZEN, E. 1961 *Technometrics* **3,** 167–190.
PARZEN, E. 1962 *Ann. Math. Stat.* **33,** 1065–1076.
PELTO, C. R. 1969 *Technometrics* **11,** 775–792.
PICKANDS, J. 1969 *Ann. Math. Stat.* **40,** 854–864.
QUESENBERRY, C. P. & GESSAMAN, M. P. 1968 *Ann. Math. Stat.* **39,** 664–673.
QUESENBERRY, C. P. & LOFTSGAARDEN, D. O. 1965 *Tech. Note. D2699* NASA.
RAMAN, S. 1971 *Ph.D. thesis* (unpublished). Dept. of Biostatistics, University of California.
RAO, B. L. S. P. 1969 *Sankhyā* (A) **31,** 23–36.
RÉVÉSZ, P. 1972 *Periodic Math. Hungar.* **2,** 85–110.
ROBERTSON, T. 1967 *Ann. Math. Stat.* **38,** 482–493.
ROBERTSON, T., CRYER, J. D. & HOGG, R. V. *On Non-parametric Estimation of Distributions and Their Modes.* Unpublished.
ROSENBLATT, M. 1956 *Ann. Math. Stat.* **27,** 832–837.
ROSENBLATT, M. 1969 *Multiv. Anal.* **2,** 25–31.
ROSENBLATT, M. 1971 *Ann. Math. Stat.* **42,** 1815–1842.
ROSENBLATT, M. 1975 *Ann. Math. Stat.* **3,** 1–14.
SAMANTA, M. 1974 *Zeit. Wahrscheinlichkeitsth* **28,** 85–88.
SCHUSTER, E. F. 1969 *Ann. Math. Stat.* **40,** 1187–1195.
SCHUSTER, E. F. 1970 *Ann. Math. Stat.* **41,** 1347–1348.
SCHUSTER, E. F. 1974 *Scand. Actuarial J.* **1,** 103–107.
SCHWARTZ, S. C. 1967 *Ann. Math. Stat.* **38,** 1261–1265.
SCHWARTZ, S. C. 1969 *SIAM J. Appl. Maths.* **17,** 447–453.
SCHWARTZ, S. C. & STEWART. 1969 *IMS Annual Meeting 19–22 August.* New York.
SEBESTYEN, G. S. 1962 *I.R.E. Trans. Info. Theory.* IT-**8,** 582–591.

SEBESTYEN, G. S. & EDIE, J. 1966 *IEEE Trans. on Elect. Comp.* EC-**15**, 908–915.
SEHEULT, A. H. & QUESENBERRY, C. P. 1971 *Ann. Math. Stat.* **42**, 1434–1438.
SHANNON, C. E. 1948 *Bell System Tech. J.* **27**, 379–423, 623–656.
SHAPIRO, J. S. 1969 *Smoothing and Approximation of Functions.* New York: Van Nostrand Reinhold.
SILVERMAN, B. W. 1976 *Math. Proc. Camb. Phil. Soc.* **80**, 135–144.
SILVERMAN, B. W. 1977*a Weak and strong uniform consistency of the kernel estimate of a density and its derivatives.* (Submitted for publication.)
SILVERMAN, B. W. 1977*b Choosing a window width when estimating a density.* (Submitted for publication.)
SINGH, R. S. 1976 *J. Multiv. Anal.* **6**, 111–122.
SPECHT, D. F. 1967*a IEEE Trans. Bio. Med. Eng.* BME **14**, 90–95.
SPECHT, D. F. 1967*b IEEE Trans. Elect. Comp.* EC **16**, 308–319.
SPECHT, D. F. 1971 *Technometrics* **13**, 409–423.
STANAT, D. F. 1966 *Tech. Report.* Sensory Intelligence Lab., University of Michigan.
STOLLER, D. S. 1954 *J. Am. Stat. Assn.* **49**, 770–775.
TARTER, M. 1977 *Variance and covariance formulas for evaluations of estimated orthogonal expansions.* (To appear.)
TARTER, M. & FELLNER, W. 1972 *Proc. 5th Conf. Interface Statist. Comput.* North Hollywood: Western Periodicals.
TARTER, M. E. & KRONMAL, R. A. 1967 *Proc. A.C.M.* **22**, 511–519.
TARTER, M. E. & KRONMAL, R. A. 1968 *Proc. A.C.M.* **23**, 491–497.
TARTER, M. E. & KRONMAL, R. A. 1970 *Ann. Math. Stat.* **41**, 718–722.
TARTER, M. E. & RAMAN, S. 1971 *Proc. 6th Berkeley Symp. Math. Statist. Prob.* **4**, 199–222.
TARTER, M. E. & SILVERS, A. 1975 *J. Am. Stat. Assn.* **70**, 47–55.
VĂDUVA, I. 1963 *Studii Cercetări Mat.* **14**, 653–660. (In Rumanian.)
VĂDUVA, I. 1967 *Studii Cercetări Mat.* **19**, 455–460.
VĂDUVA, I. 1968 *Studii Cercetări Mat.* **20**, 1207–1276. (In Rumanian.)
VAN RYZIN, J. 1966 *Sankhyā* (A) **28**, 261–270.
VAN RYZIN, J. 1969 *Ann. Math. Stat.* **40**, 1765–1772.
VAN RYZIN, J. 1970 *Tech. Report* 226. University of Wisconsin (IMS 8–10 April, Dallas). Also 1973 *Commun. Statist.* **2**, 493–506.
VENTER, J. H. 1967 *Ann. Math. Stat.* **38**, 1446–1455.
WAHBA, G. 1971 *Ann. Math. Stat.* **42**, 1870–1886.
WAHBA, G. 1975*a Ann. Stat.* **3**, 15–29.
WAHBA, G. 1975*b Ann. Stat.* **3**, 30–48.
WAHBA, G. & WOLD, S. 1975 *Commun. Statist.* **4**, 125–141.
WALD, A. & WOLFOWITZ, J. 1939 *Ann. Math. Stat.* **10**, 105–118.
WATSON, G. S. 1964 *Sankhyā* (A) **26**, 359–384.
WATSON, G. S. 1969 *Ann. Math. Stat.* **40**, 1496–1498.
WATSON, G. S. & LEADBETTER, M. R. 1963 *Ann. Math. Stat.* **34**, 480–491.
WATSON, G. S. & LEADBETTER, M. R. 1964*a Biometrika* **51**, 175–184.
WATSON, G. S. & LEADBETTER, M. R. 1964*b Sankhyā* (A) **26**, 101–116.
WEGMAN, E. J. 1968 *Dissertation.* University of Iowa.
WEGMAN, E. J. 1969*a Inst. Stat. Mimeo Ser.* 638. University of N. Carolina.
WEGMAN, E. J. 1969*b Ann. Math. Stat.* **40**, 1661–1667.
WEGMAN, E. J. 1969*c Inst. Stat. Mimeo Ser.* 629. University of N. Carolina.
WEGMAN, E. J. 1970*a Ann. Math. Stat.* **41**, 457–471.
WEGMAN, E. J. 1970*b Inst. Stat. Mimeo Ser.* 647. University of N. Carolina. Also *Ann. Math. Stat.* **41**, 2169–2174.
WEGMAN, E. J. 1972*a Technometrics* **14**, 533–546.
WEGMAN, E. J. 1972*b J. Statist. Comp. Simul.* **1**, 225–245.
WEISS, L. & WOLFOWITZ, J. 1967 *Z. Wahrscheinlichkeitstheor. verw. Geb.* **7**, 327–335.
WEISS, L. & WOLFOWITZ, J. 1969 *Ops Res. Verfahren* **8**, 295–299.
WERTZ, W. 1969 *Ops Res. Verfahren* **8**, 300–304.

WHITTLE, P. 1958 *J.R. Stat. Soc.* (B) **20,** 334–343.
WINTER, B. B. 1975 *Ann. Statist.* **3,** 759–766.
WOODROOFE, M. 1967 *Ann. Math. Stat.* **38,** 475–481.
WOODROOFE, M. 1968 *Ann. Math. Stat.* **41,** 1655–1671.
YAMATO, H. 1972 *Bull. Math. Stat.* **15,** 113–131.
YOUNG, T. Y. & CALVERT, T. W. 1974 *Classification, Estimation and Pattern Recognition.* New York: American Elsevier.