International Journal of Computer and Information Sciences, Vol. 5, No. 2, 1976

Pattern Classes: A Technique for Recovering Their Distributions

Douglas Dorrough[1]

Received November 1972; revised September 1975

Many statistical pattern-recognition techniques depend for their application on the generation of one or more prototype patterns for each decision class. In turn, the determination of prototypes is dependent on the underlying probability distribution associated with a given class and that distribution's relationship to the distributions associated with the remaining classes. If these distributions are known, the problem of classification is considerably less complex than if they are unknown. The problem of recovering an unknown underlying distribution is one that has received considerable attention. The results thus far, however, are nonpractical. A practical technique that makes use of certain parameters related to sample size is presented and verified.

KEY WORDS: Pattern recognition; distribution recovery; density estimation; clustering algorithms; fault isolation.

1. INTRODUCTION

The majority of reported statistical pattern-recognition techniques involve the generation of one or more prototype patterns for each decision class. In general, the choice of a prototype pattern depends on the set of sample patterns in a given class and its interrelation with the sample sets in each of the remaining classes. More specifically, the determination of prototypes is a function of the underlying probability distribution associated with a

This work was supported in part by the Office of Naval Research under Contract No. N00014-72-C-0459.

[1] Former Member of the Professional Staff, Ultrasystems, Inc. Present address: Senior Scientist, McDonnell Douglas Corporation, Huntington Beach, California.

given class and its interrelation with the distributions associated with the remaining classes. If the distributions are known, the classification problem can be greatly simplified. The problem of accurately recovering an unknown underlying distribution (or density) associated with a collection of samples has received considerable attention.[1] Parzen[2] has treated a general class of consistent estimators for the one-dimensional case. Most of his results have been extended to the multidimensional case by Murthy[3,4] and Cacoullos,[5] with stronger consistency results obtained by Nadaraya.[6,7]

All of these results focus on estimators of the form

$$f_M(x) = \int_{-\infty}^{\infty} K_M(x, y)\, dF_M(y)$$

$$= \frac{1}{M} \sum_{i=1}^{M} K_M(x, X_i)$$

where $X_i, \ldots, X_n$ is a sequence of independent identically distributed random variables with probability density function f, F_M denotes the empirical distribution function based on the first M observations, and K denotes the nonnegative Borel "weighting" function made to satisfy differing conditions for each set of results.

The indicated results are within the class of density estimator densities (ded). All of them are asymptotic and consequently of little practical utility. To make them usable, an adequate estimate of certain parameters must be obtained for the sample size involved in the recovery problem. The distribution recovery techniques described below provide a method for estimating the requisite parameters and for supplying relatively accurate estimates of the underlying distributions associated with sample sets of patterns. Consequently, the techniques have strong application to the problem of designing pattern-recognition systems that will permit, among other things, better fault search and fault isolation policies.

2. PATTERN-RECOGNITION PROBLEM

The significant applications of statistical pattern-recognition methodology usually involve patterns of great complexity, where complexity implies input patterns of high dimensionality. The enormous mass of data associated with complex pattern-recognition tasks has prevented elegant solutions to the problem. Implicitly involved in most of the current pattern-recognition techniques are strong assumptions about the underlying distribution from which the sample set of input patterns are drawn. The generalization capability of designs based on such techniques is usually tested by rating the performance of the recognition system using a relatively small set of patterns

that were not used in the design process. Although this method of measuring generalization capability is currently in wide use, it is usually conceded by investigators in the field of pattern recognition that the true generalization capability of a design may differ considerably from the rating calculated by such a testing procedure. It is also conceded that, if accurate estimates of the underlying distributions associated with each pattern class were available, designs with superior generalization capabilities could be produced. This is particularly, though not exclusively, true of pattern-recognition systems that classify on the basis of the principle of maximum likelihood.

In Section 3, a technique for recovering the distribution density associated with a given class of patterns (represented by a collection of N-dimensional vectors) is discussed.

If it is assumed that a good estimate of the distribution associated with a given class of patterns is obtained, the next task is to represent this distributional information using a finite number of "prototype" vectors. For reasons of design economy, this number should be much less than the actual number of sample patterns in the class being considered. In Section 4, a "clustering" technique is discussed. A "cluster point" (or vector) is defined as a point at which the estimate of the underlying distribution density assumes (locally) a maximum value. These cluster points are identified with the "prototype" vectors mentioned earlier. Finally, application of these techniques to pattern-recognition is presented in Section 5.

3. DISTRIBUTION RECOVERY

Let $(x^k)_{k=1}^0$ be a set of identically distributed N-dimensional random vectors. An empirical distribution function F_M is defined by the expression

$$F_M(x_1, x_2, \ldots, x_N) \equiv \frac{1}{M} \cdot \begin{Bmatrix} \text{number of observations } x^k \\ \text{such that } x_j^k \leqslant x_j \end{Bmatrix} \tag{1}$$

where $j = 1, 2, \ldots, N$.

An estimator f_M for the N-variate density f is defined as

$$f_M(x) \equiv \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \prod_{n=1}^{N} b_n H\{b_1(x_1 - y_1), \ldots, b_N(x_N - y_N)\}\, dF_M(y_1, \ldots, y_N) \tag{2}$$

where, for the applications of interest, f is assumed to be everywhere continuous. In the definiendum, the function H satisfies the following conditions:

$$\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} H(x_1, x_2, \ldots, x_N)\, dx_1 \cdots dx_N = 1$$

$$H(x_1, x_2, \ldots, x_N) = H(\pm x_1, \pm x_2, \ldots, \pm x_N) \geqslant 0$$

$$\lim_{\|x\| \to \infty} \|x\| H(x) = 0$$

where $\| x \|$ is given by the definition

$$\| x \| = \left\{ \sum_{n=1}^{N} x_n^2 \right\}^{1/2}$$

and where $b_n > 0$ for any $n = 1, 2, \ldots, N$.

It is easily demonstrable that Eq. (2) is equivalent to

$$f_M(x) = \frac{1}{M} \left\{ \prod_{n=1}^{N} b_n \right\} \sum_{k=1}^{N} H\{b_1(x_1 - x_1^k), \ldots, b_N(x_N - x_N^k)\} \tag{3}$$

For the one-dimensional case, estimators of the type given by Eqs. (2) or (3) have been considered by Parzen.[2] As indicated, his results were extended to the multidimensional case by Murthy,[4] who demonstrated that, if the b_n are functions of the sample size M such that the conditions

$$\text{(i)} \quad \lim_{M \to \infty} b_n = \lim_{M \to \infty} b_n(M) = \infty \qquad \text{for} \quad n = 1, 2, \ldots, N$$

$$\text{(ii)} \quad \lim_{M \to \infty} \prod_{n=1}^{M/N} b_n = \infty$$

are satisfied, then f_M is a consistent estimate of f at every point x. More precisely, if condtitions (i) and (ii) are satisfied, then at every point x

$$\lim_{M \to \infty} E\{f_M(x)\} = f(x)$$

and

$$\lim_{M \to \infty} var\{f_M(x)\} = 0$$

For purposes of application and where M is finite, asymptotic estimations of the form under consideration are useless unless a technique for determining $b = (b_1, b_2, \ldots, b_N)$ is available for a given M. Such a technique does exist. Its theoretical basis is summarized below.

3.1. Applicable Density Function Approximation

In order to evaluate b as a function of M, as well as those properties intrinsic to the sample, it is here assumed that b_n and a parameter

$$\rho_n = \rho(M, c_1, c_2) > 0$$

are related by the equation

$$\frac{1}{b_n} = \frac{1}{M(M-1)} \sum_{i=1}^{M} \sum_{j=1}^{M} a_{ijn} \exp(-\rho_n a_{ijn}) \tag{4}$$

where c_1, $c_2 \equiv$ parameters associated with the sample set and $a_{ijn} \equiv (x_n{}^i - x_n{}^j)^2$, such that i and j refer to the ith and jth sample vector.

The problem of determining b_n is thus traded for the problem of determining ρ_n. The parameters c_1 and c_2 are specifically associated with the set

$$\{a_{ijn}\}_{i,j=1}^{M} \prod_{n=1}^{N} b_n$$

derived from the original sample set $(x^j)_{j=1}^{M}$ of random vectors.

It remains to state and demonstrate certain relationships between b_n and ρ_n such that the indicated trade becomes valid. The required relationships are given by the following two theorems, whose extensive proofs are given elsewhere.[8]

If it is assumed that, for $n = 1, 2, \ldots, N$, b_n is given by Eq. (4), then

$$\lim_{M\to\infty} \rho_n(M, c_1, c_2) = \infty \Rightarrow \lim_{M\to\infty} b_n(M) = \infty \qquad \text{(T-I)}$$

for each n. If

$$\mu_n(M) \equiv \frac{1}{M(M-1)} \sum_{i=1}^{M} \sum_{j=1}^{M} a_{ijn}$$

and

$$\lim_{M\to\infty} \mu_n(M) = \mu_n < \infty$$

then

$$\lim_{M\to\infty} \prod_{n=1}^{M/N} b_n = \infty \Leftrightarrow \rho_n = o(\log M) \qquad \text{(T-II)}$$

These results as well as those of Eq. (3) indicate that an expression for ρ_n of the form

$$\rho_n = \frac{\delta_n}{\mu_n(M)} \log(M \cdot \mu_n(M)^{1/\delta})$$

should be investigated, with

$$\mu_n(M) = \frac{1}{M(M-1)} \sum_{i=1}^{M} \sum_{j=1}^{M} a_{ijn}$$

where $\mu_n(M)$ and δ_n are parameters that either may be extracted from the sample set or whose near optimum value is easily obtained.

3.2. Experimental Results

A goodness-of-fit computer program,[9] called SIMFIT, that checks the accuracy of an estimate against known distributions was applied to (I) the

one-dimensional continuous case, (II) the two-dimensional case, and (III) the two-dimensional discrete case.

If R is used to denote the set of points on the x plane, the known distributions for case I are of the general form

$$f(x) = \sum_{i=1}^{R} p_i g_i(x) \tag{5}$$

with g_i denoting Gaussian density functions having different means and possibly different variances. The normalizer coefficients p_i denote the ratios of the number of samples taken from the ith Gaussian distribution (or mixture) to the total number of samples.

For case II, the two-dimensional extension is straightforward:

$$f(x, y) = \sum_{i=1}^{R} \sum_{j=1}^{R} p_{ij} g_{ij}(x, y) \tag{6}$$

The known distributions for case III are given respectively by the equation

$$f(x, y) = \frac{\binom{6}{x}\binom{7}{y}\binom{5}{4 - x - y}}{\binom{18}{4}} \tag{7}$$

where (x, y) is a two-dimensional random variable and $0 \leqslant x + y \leqslant 4$, and by the equation

$$f(t, \lambda) = \lambda^t e^{-\lambda}/t! \tag{8}$$

where $t = 0, 1, 2, \ldots$ and $\lambda > 0$.

The actual distance D between an estimate f_M and the density function f being estimated is given by SIMFIT according to the general metric[2]

$$D = E\{\|f - f_M\|^2\}$$

Sample generation was accomplished by equally partitioning the unit interval on the probability axis into K subintervals, where K is always equal to the number of samples desired. The method is illustrated in Fig. 1.

The results of executing the indicated computer programs for the one-dimensional continuous case are shown in Figs. 2–18. Figures 2–9 illustrate the results obtained for a known distribution density f given specifically by

$$f(x) = \frac{1}{3\sqrt{2\pi}}\left\{\frac{1}{2}\exp\left[-\frac{(x-2)^2}{8}\right] + \frac{1}{2}\exp\left[-\frac{(x+1)^2}{8}\right] + \exp\left[-\frac{(x-5)^2}{2}\right]\right\} \tag{9}$$

[2] This metric is actually approximated by an average over Monte Carlo trials.

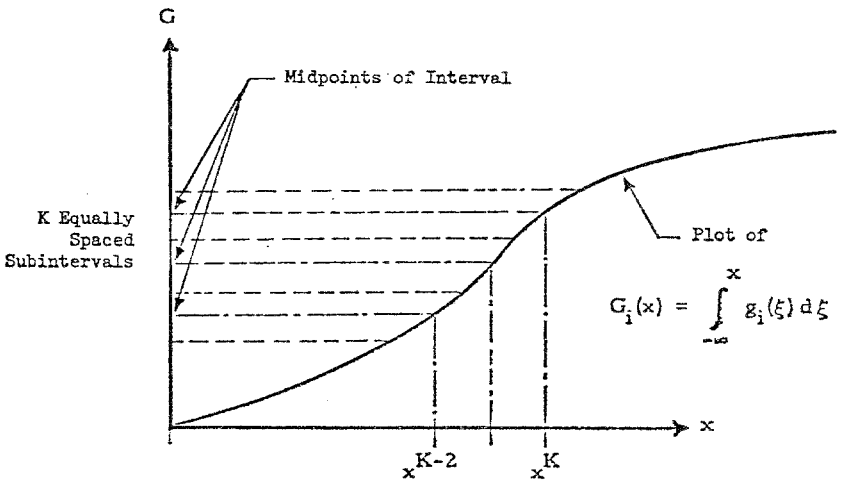

Fig. 1. Sample set generator.

Figures 2–5 give the results for 150 sample points as δ decreases from 0.95 to 0.5 and where $\log M \cdot \mu(M)^{1/\delta}$ is replaced by $\log M$.

Approximation of the distribution of Eq. (9) is given by Figs. 6–9 for 75 sample points. Thus, Fig. 2 represents the best approximation, with accuracy steadily decreasing until the worst case is reached in Fig. 9.

Figures 10–13 represent the results obtained for the case where

$$f(x) = \frac{1}{3\sqrt{2\pi}} \left\{ \exp\left[-\frac{(x+1.873)^2}{2} \right] + \exp\left[-\frac{(x-5.873)^2}{2} \right] \right\} \quad (10)$$

This case was considered to illustrate the ability of the recovery technique to approximate bimodal distributions with sharp peaks that are strongly separated. Once again the figures are arranged in order of decreasing accuracy of approximation. The behavior as δ (or the log function of δ) varies is consistent with that shown in the block of Figs. 2–9. Experiments indicated, of course, that better representation in the neighborhood of the modes would have been obtained if a larger sample had been used.

To illustrate the effect of sample size more strongly, the following case was considered and illustrated in Fig. 14–17.

$$f(x) = \frac{1}{7\sqrt{2\pi}} \left\{ \exp\left[-\frac{(x+2)^2}{8} \right] + 10 \exp\left[-\frac{(x-0.5)^2}{0.02} \right] + \exp\left[-\frac{(x-4)^2}{32} \right] \right\} \quad (11)$$

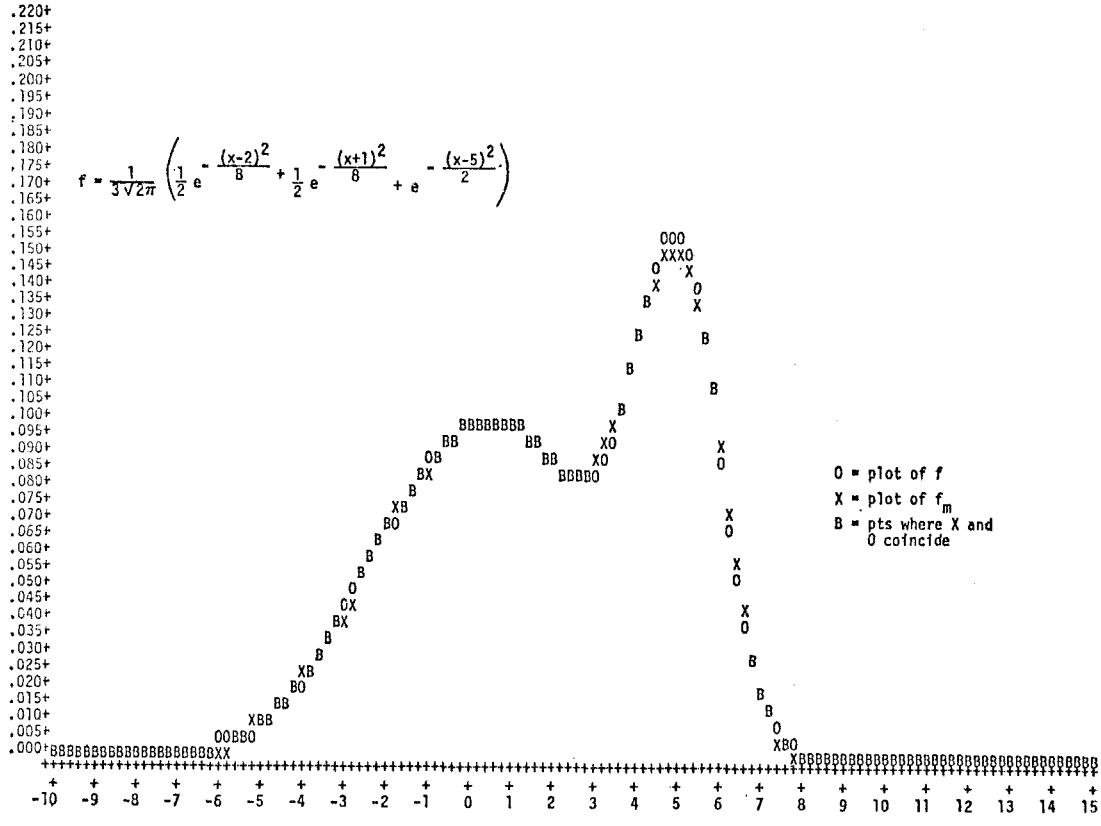

Fig. 2. Density approximation for $\rho = [\delta/\mu(M)] \log [M(M)^{1/\delta}]$, $\delta = 0.95$, 150 sample points.

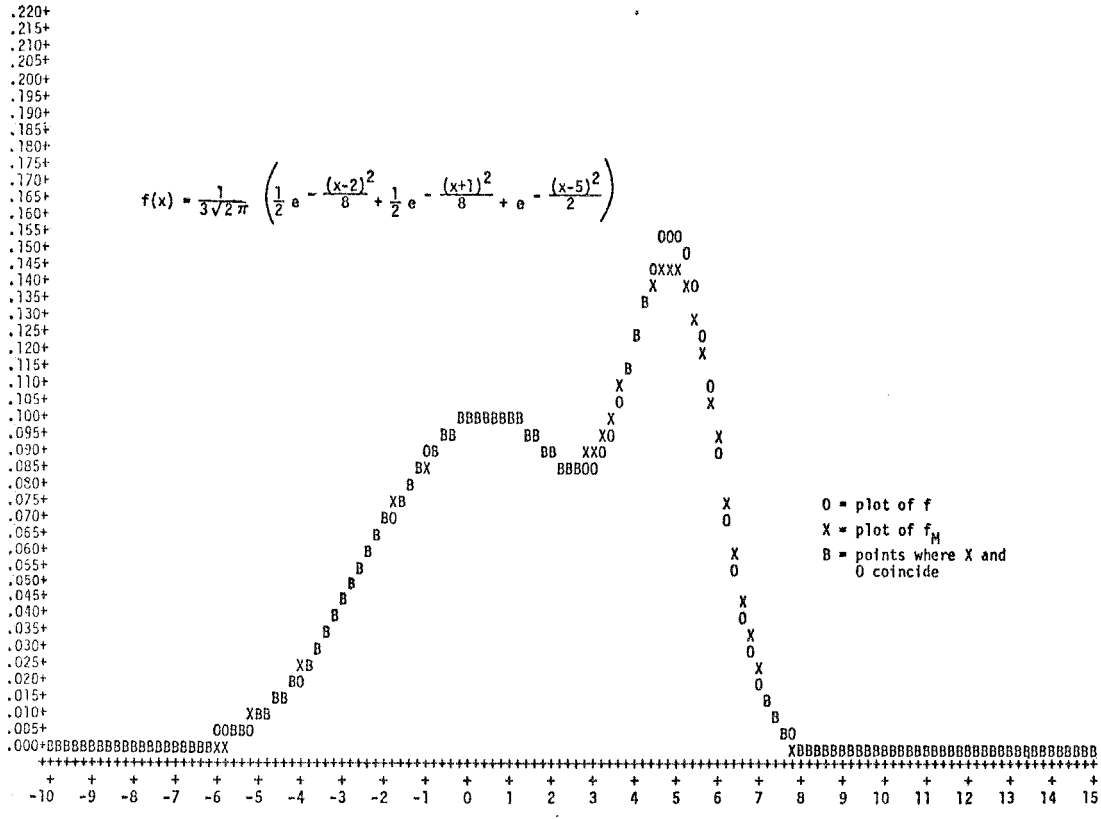

Fig. 3. Density approximation for $\rho = [\delta/\mu(M)] \log [M(M)^{1/\delta}]$, $\delta = 0.5$, 150 sample points.

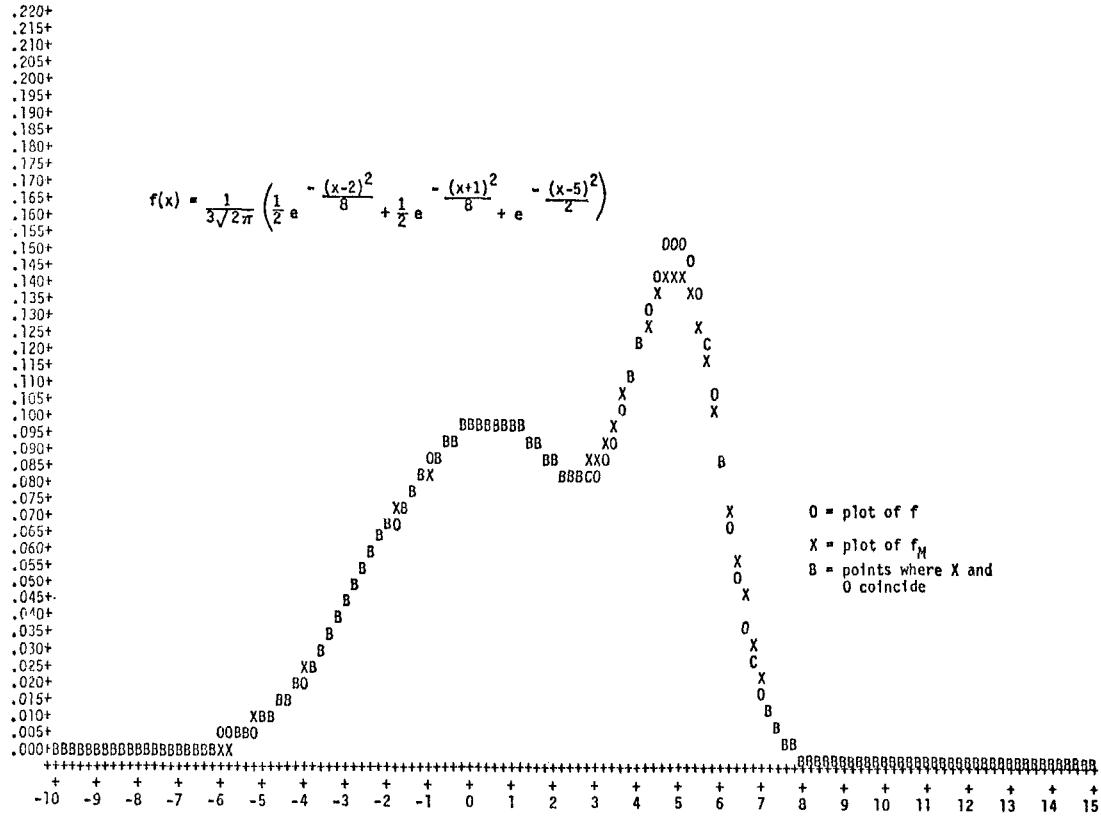

Fig. 4. Density approximation for $\rho = [\delta/\mu(M)]\,\hat{\log} M$, $\delta = 0.95$, 150 sample points.

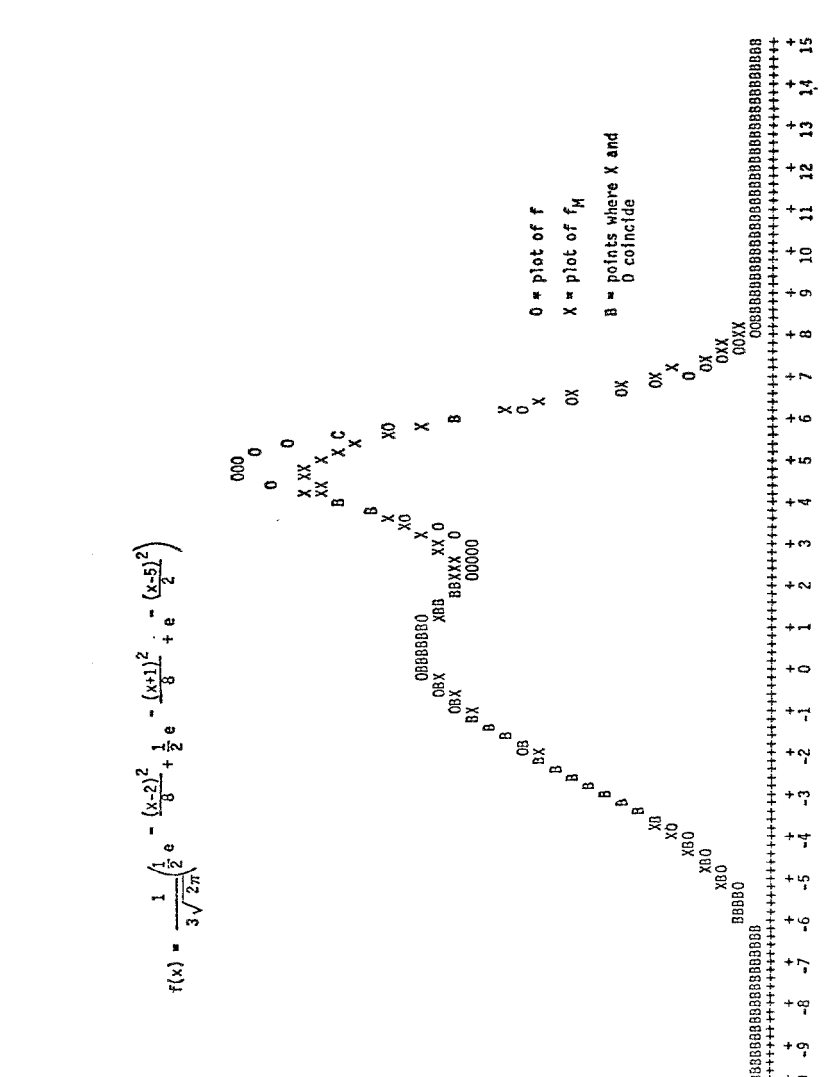

Fig. 5. Density approximation for $\rho = [\delta/\mu(M)]\log M$, $\delta = 0.5$, 150 sample points.

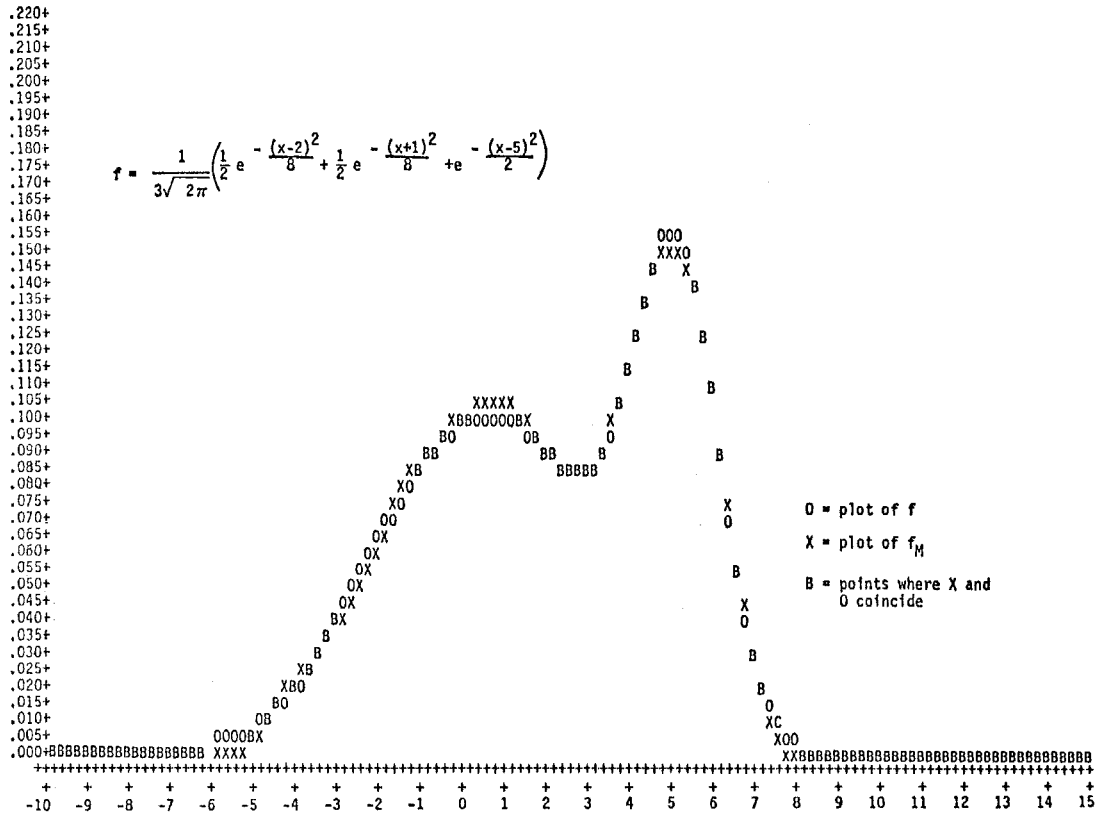

Fig. 6. Density approximation for $\rho = [\delta/\mu(M)] \log [M(M)^{1/\delta}]$, $\delta = 0.95$, 75 sample points.

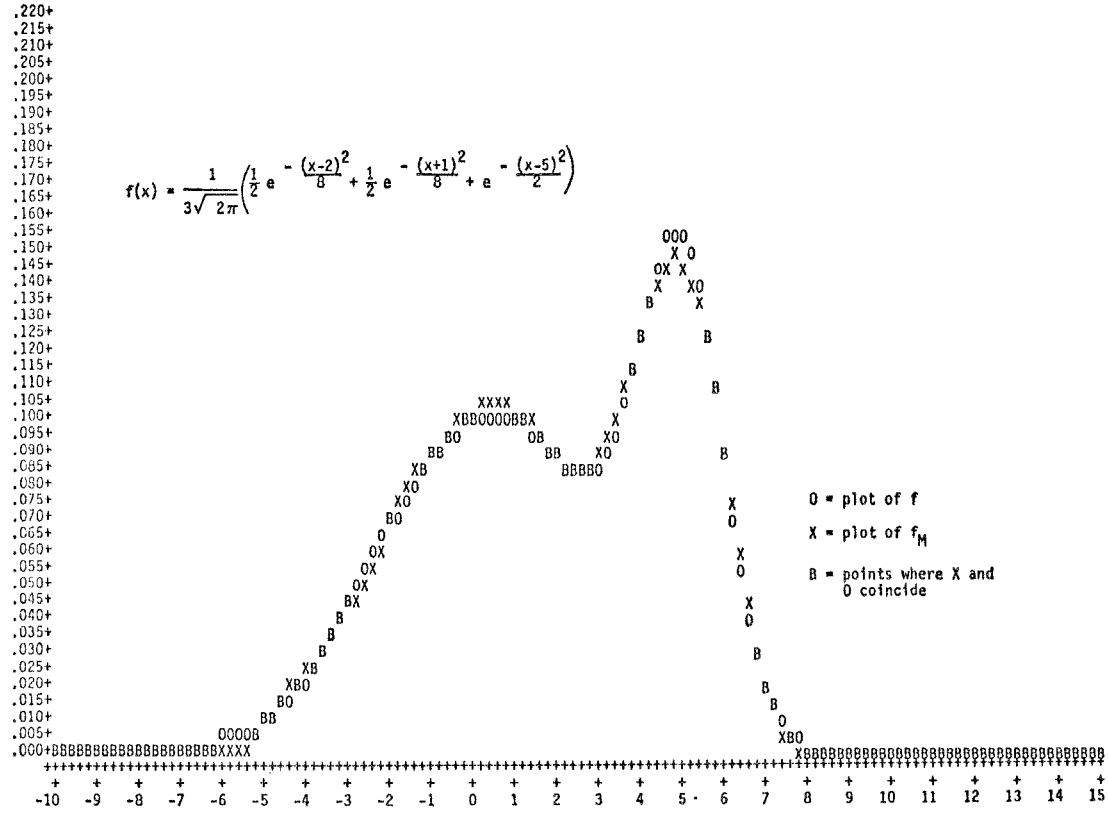

Fig. 7. Density approximation for $\rho = [\delta/\mu(M)]\log[M(M)^{1/\delta}]$, $\delta = 0.5$, 75 sample points.

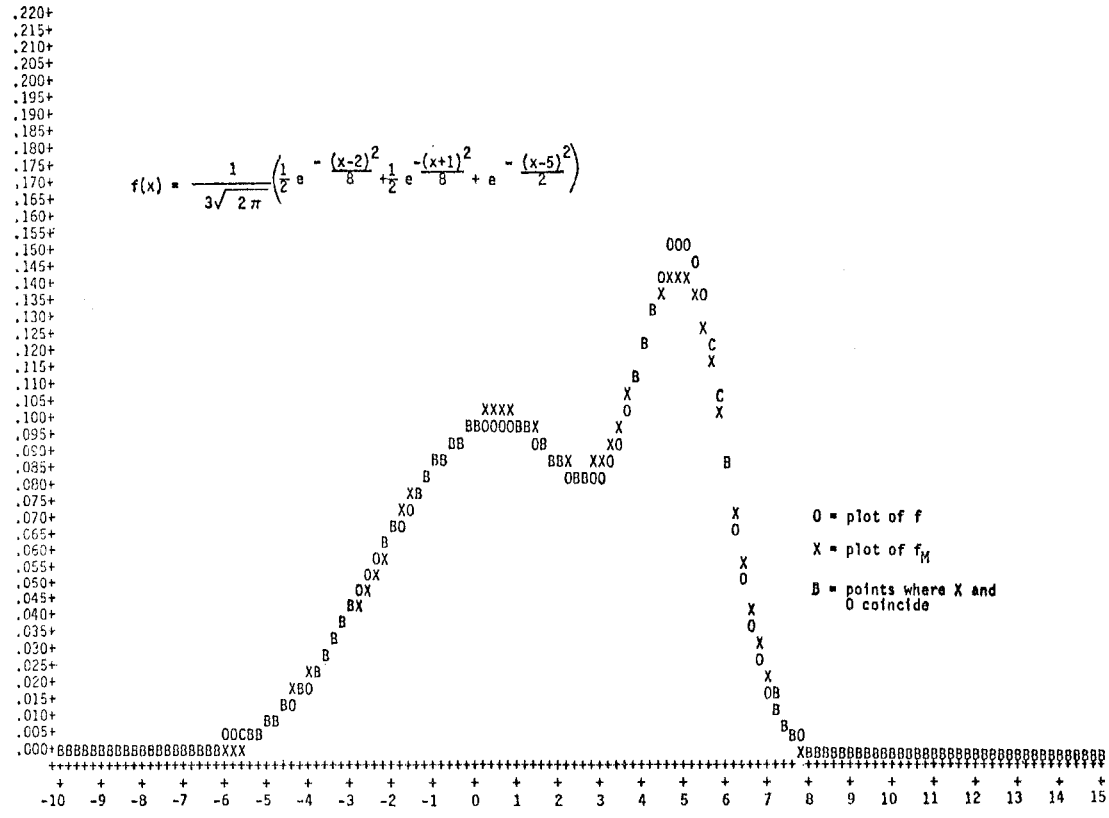

Fig. 8. Density approximation for $\rho = [\delta/\mu(M)] \log M$, $\delta = 0.95$, 75 sample points.

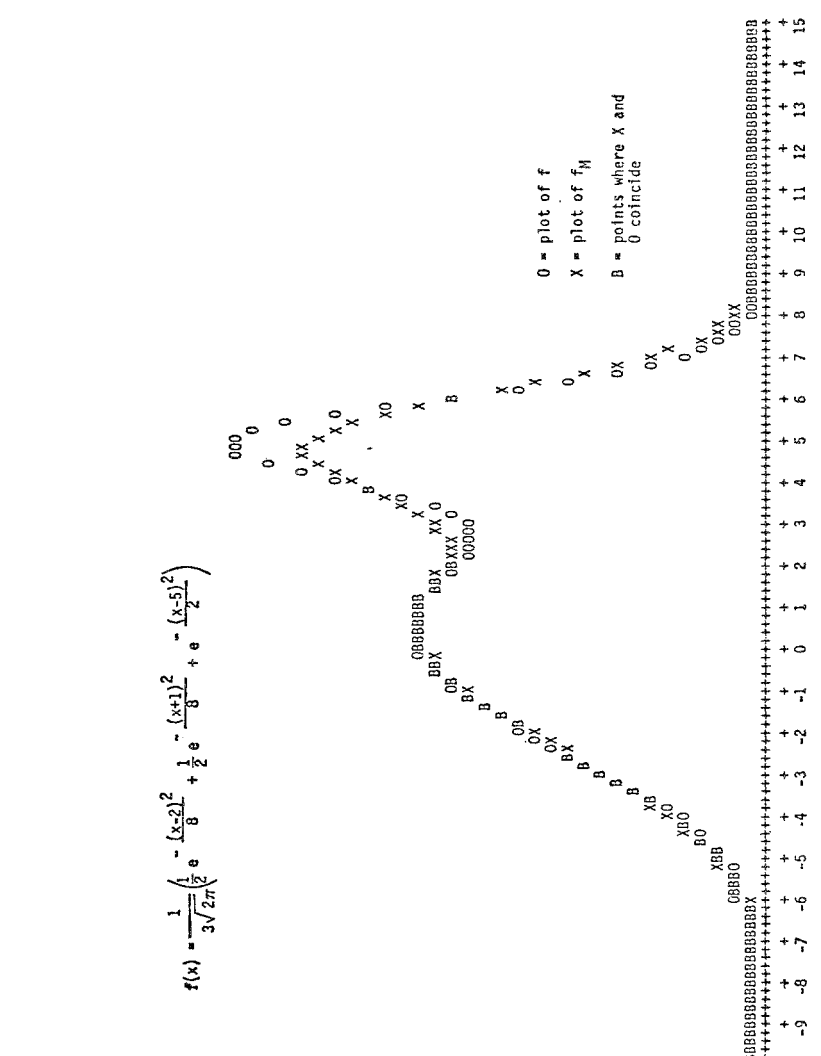

Fig. 9. Density approximation for $\rho = [\delta/\mu(M)]\log M$, $\delta = 0.5$, 75 sample points.

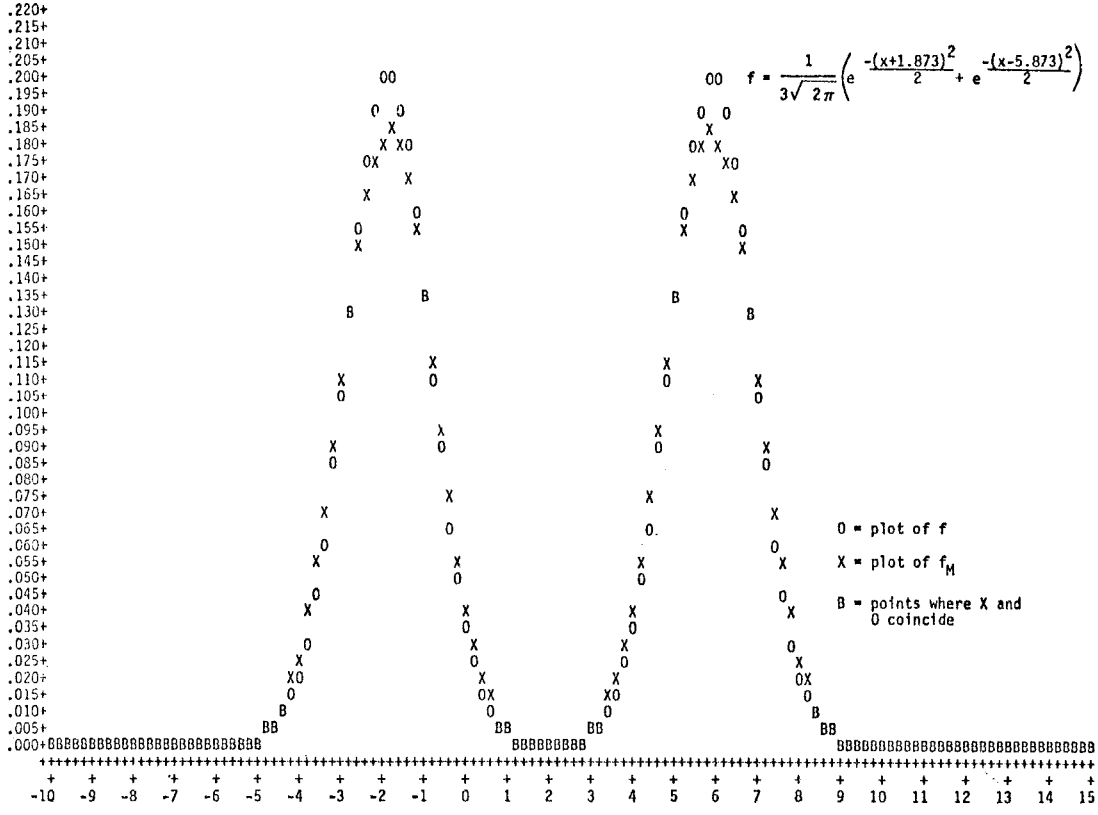

Fig. 10. Density approximation for $\rho = [\delta/\mu(M)] \log [M(M)^{1/\delta}]$, $\delta = 0.95$, 200 sample points.

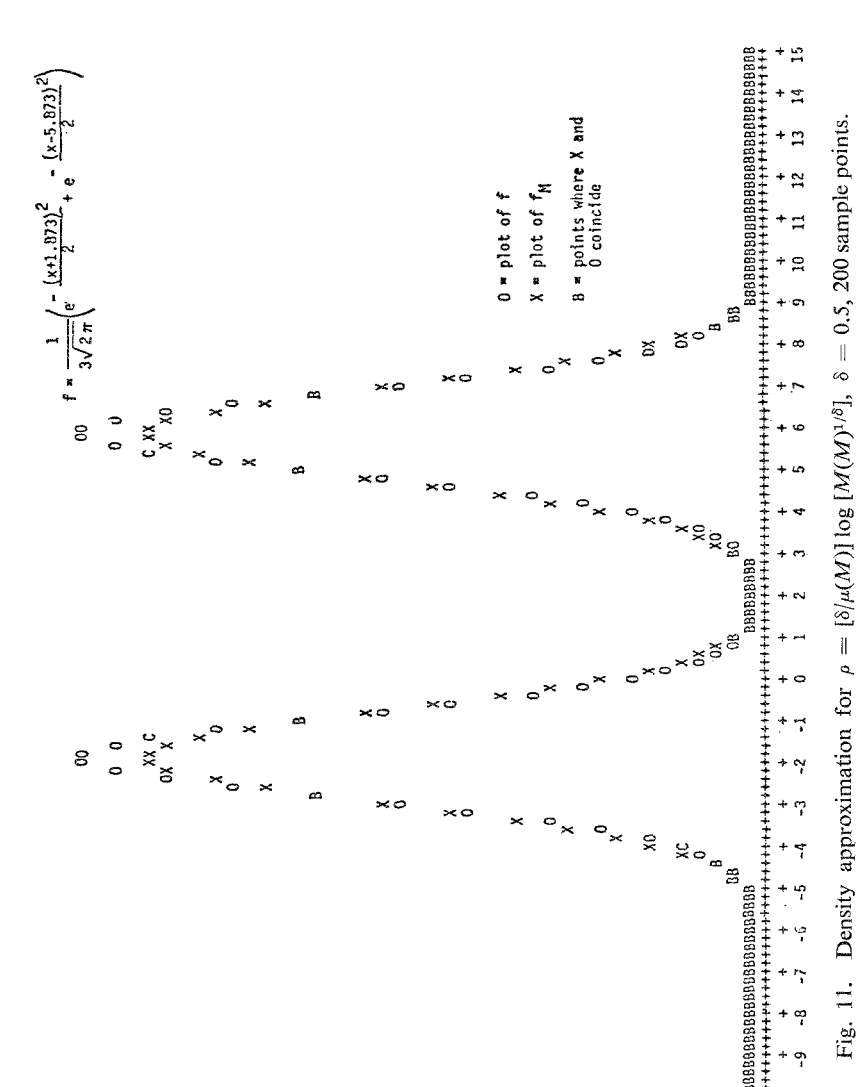

Fig. 11. Density approximation for $\rho = [\delta/\mu(M)] \log [M(M)^{1/\delta}]$, $\delta = 0.5$, 200 sample points.

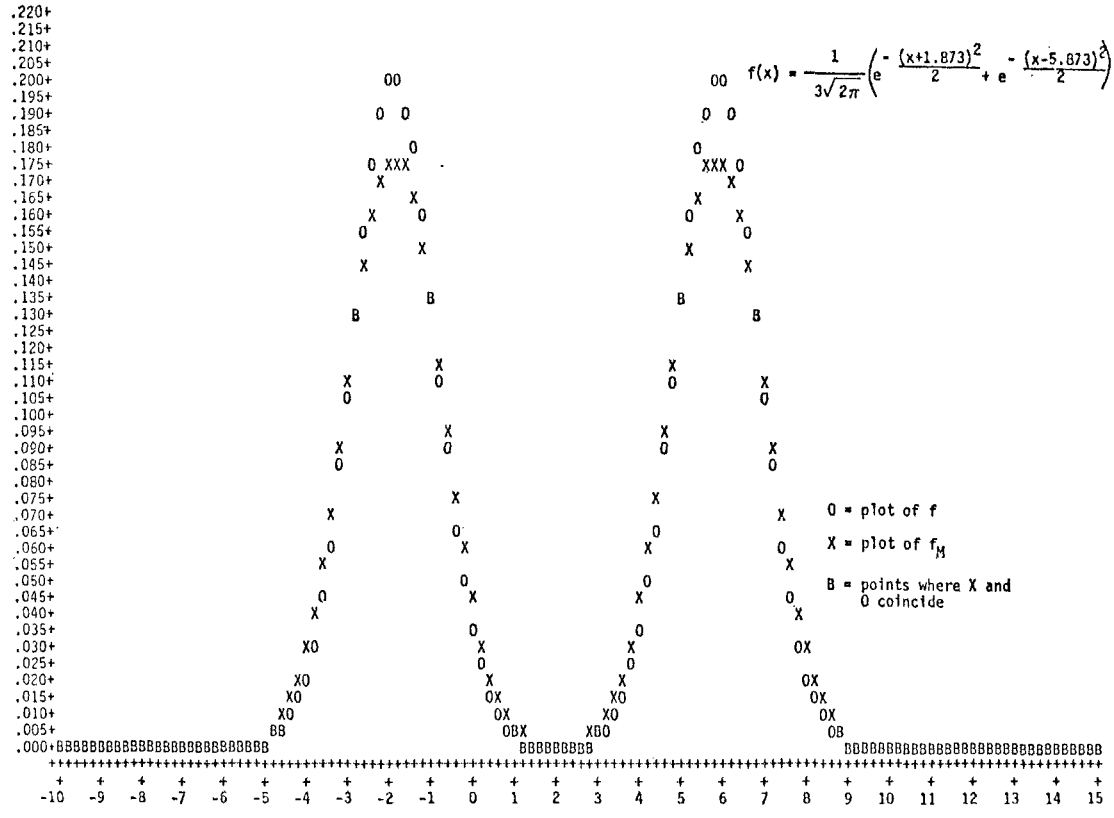

Fig. 12. Density approximation for $\rho = [\delta/\mu(M)] \log M$, $\delta = 0.95$, 200 sample points.

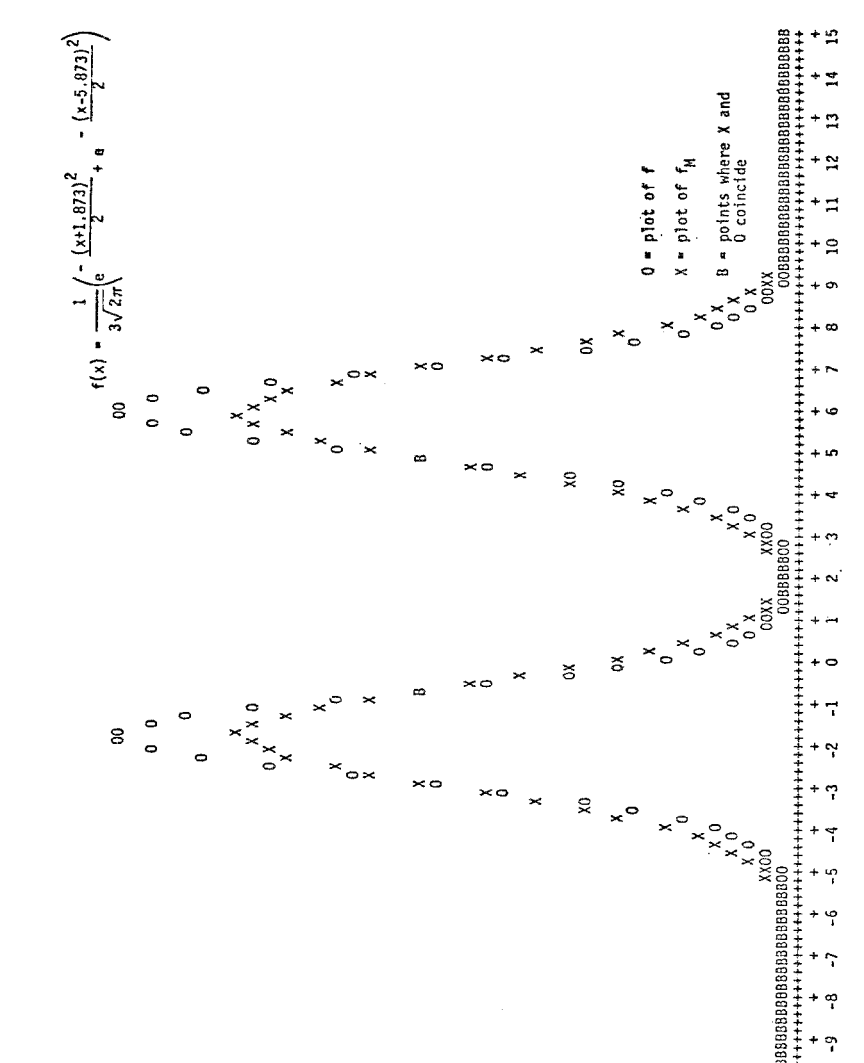

Fig. 13. Density approximation for $\rho = [\delta/\mu(M)] \log M$, $\delta = 0.5$, 200 sample points.

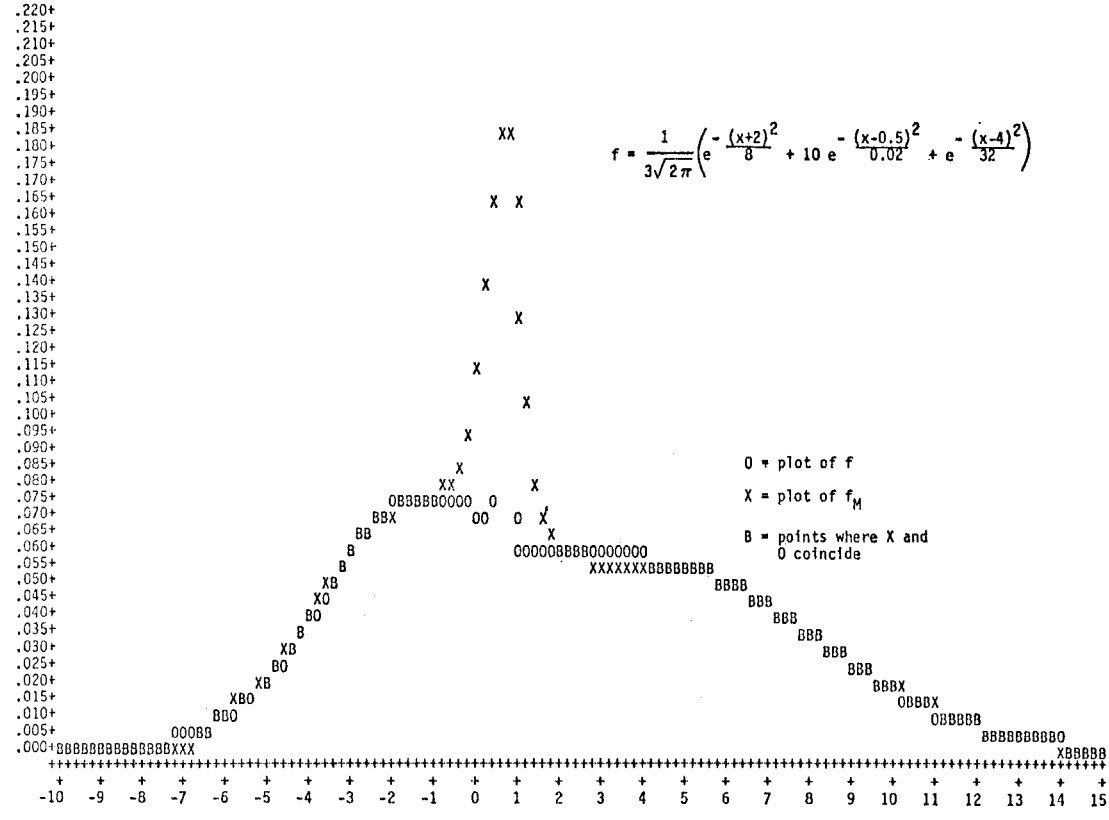

Fig. 14. Density approximation for $\rho = [\delta/\mu(M)] \log [M(M)^{1/\delta}]$, $\delta = 0.95$, 175 sample points.

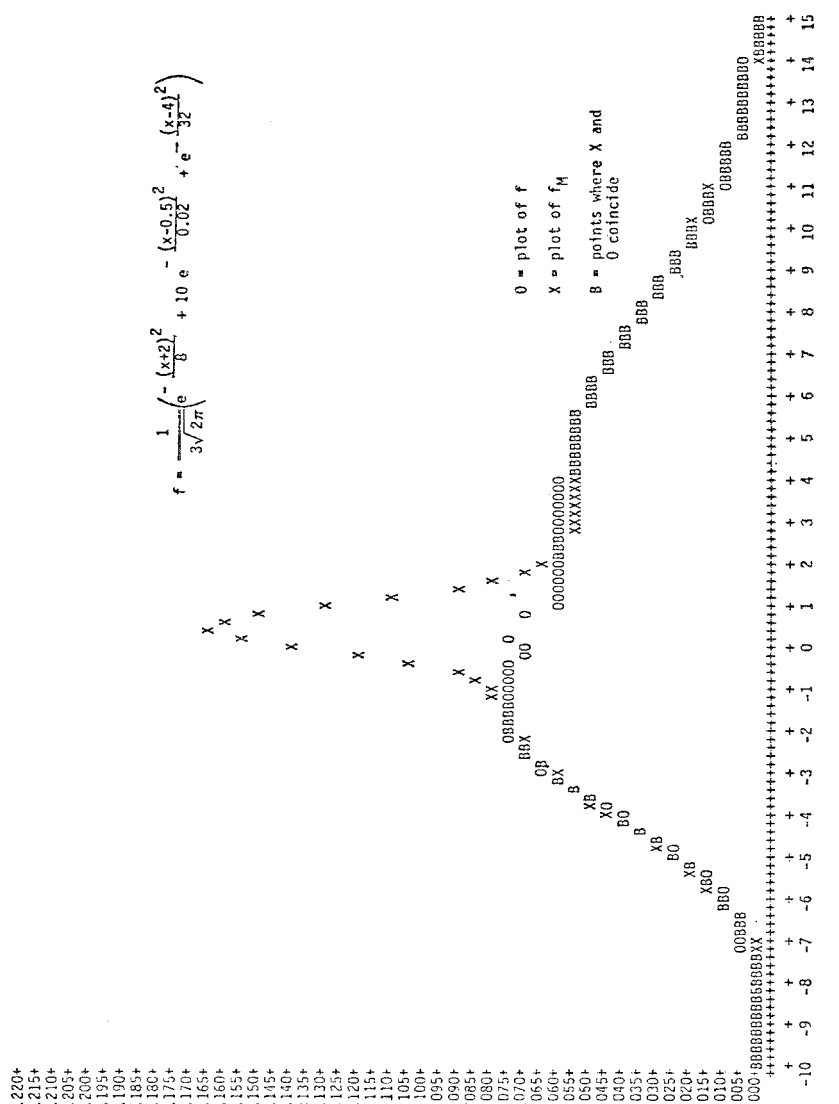

Fig. 15. Density approximation for $\rho = -[\delta/\mu(M)]\log[M(M)^{1/\delta}]$, $\delta = 0.5$, 175 sample points.

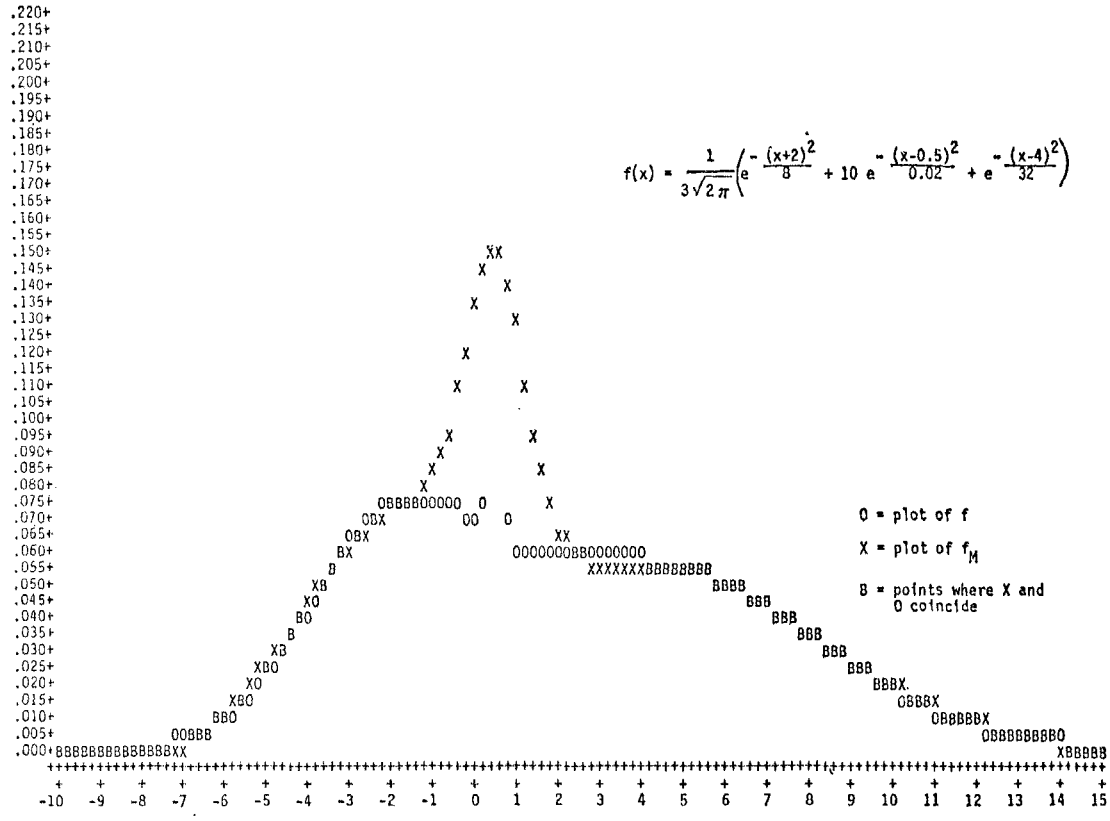

Fig. 16. Density approximation for $\rho = [\delta/\mu(M)] \log (M)$, $\delta = 0.95$, 175 sample points.

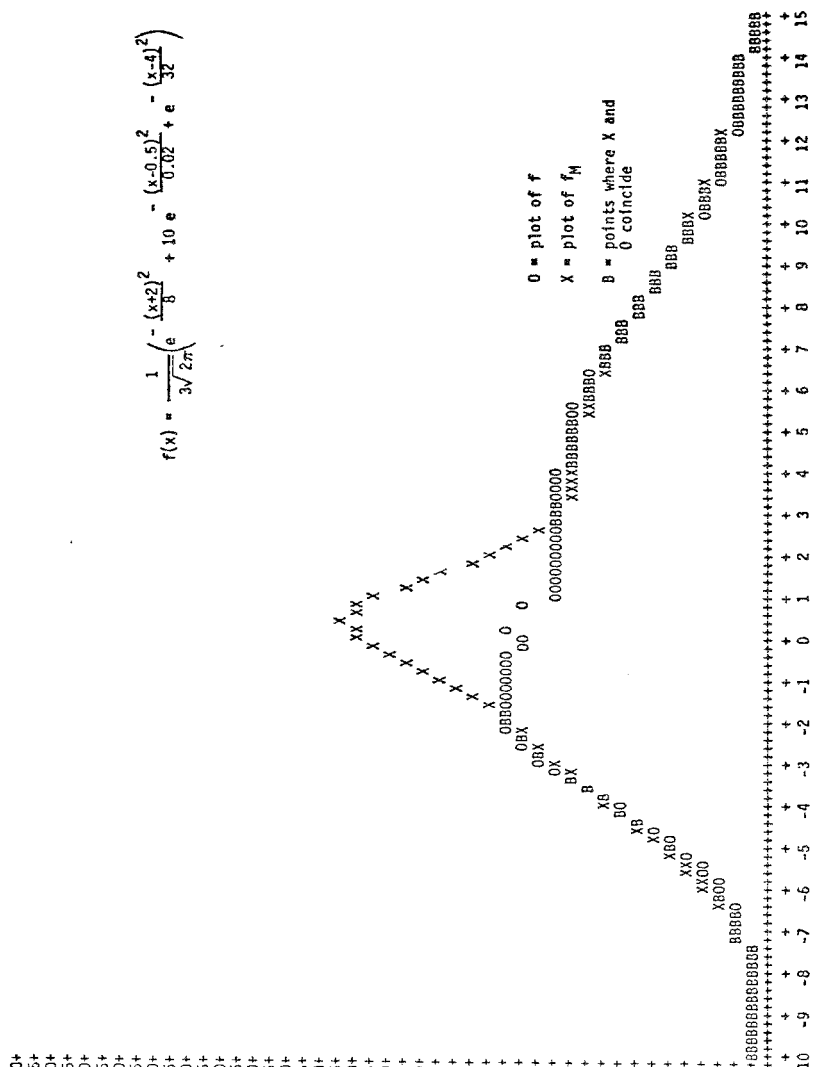

Fig. 17. Density approximation for $\rho = [\delta/\mu(M)]\log(M)$, $\delta = 0.5$, 175 sample points.

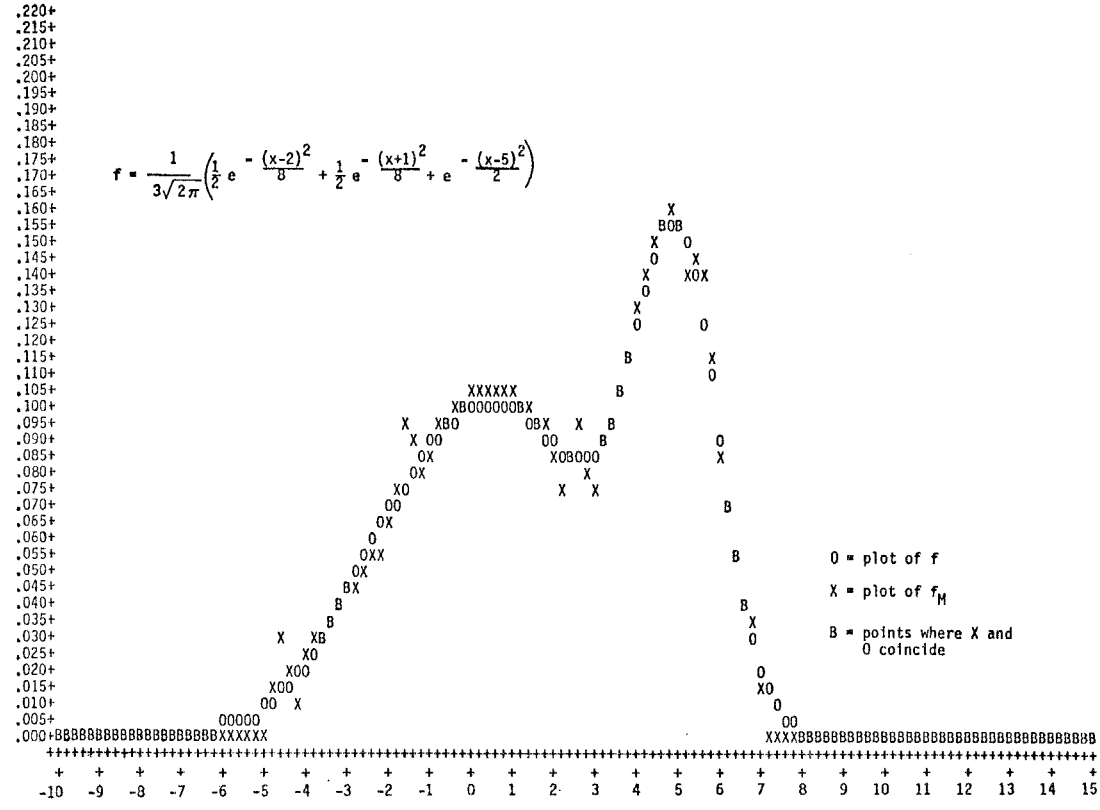

Fig. 18. Density approximation for $\rho = [\delta/\mu(M)] \log [M(M)^{1/\delta}]$, $\delta = 0.95$. The b used is four times the calculated b (compare Fig. 6). Total 75 sample points.

This says that $f(x)$ is a randomized mixture of the form

$$f(x) = \sum_{i=1}^{3} p_i f_i(x)$$

where $f_i(x) \sim N(\mu_i, \sigma_i^2)$ and $p_1 = \frac{2}{7}$, $p_2 = \frac{1}{7}$, $p_3 = \frac{4}{7}$. The small standard deviation in the second term and the fact that the means of the first and third terms are widely separated produce an f with a large narrow peak in the neighborhood of the point $x = 0.5$ even with only 25 points taken from the center distribution.

As the results shown in Figs. 12–15 indicate, the approximation was very accurate in well-represented regions but rather poor in the neighborhood of the point $x = 0.5$. However, the sharp peak was accurately located in each case.

If these results and adequate sample sizes are used, it is clear that extremely close approximations with accurate mode location can be obtained. To stress the importance of the results, one need only note that an arbitrary choice of the vector parameter b could lead to an oversmoothed approximation to the underlying (and unknown) density function or, at the other extreme, could result in an approximation that has as many peaks (modes) as there are sample points. Figure 18 illustrates the degraduation that can occur when too large a value of b is used. The case illustrated in Fig. 18 is identical to that shown in Fig. 6 except that a value of b four times that calculated to plot Fig. 6 was used. Note that already extraneous peaks have begun to appear. The degradation in accuracy would be much more apparent if the sample size were decreased or if b (as used) were increased.

If $[R]$ denotes the set of points on the x, y plane, one of the known distributions for case II is given by a two-dimensional, normal random variable having the joint density

$$f(x, y) = \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}} \times \exp\left\{-\frac{1}{2(1-\rho^2)}\left[\left(\frac{x-\mu x}{\sigma_x}\right) - 2\rho\left(\frac{x-\mu x}{\sigma_x}\right)\left(\frac{y-\mu y}{\sigma_y}\right) + \left(\frac{y-\mu y}{\sigma_y}\right)^2\right]\right\} \tag{12}$$

It can be demonstrated that the density denoted by Eq. (12) is that of the bivariate normal. Thus, the probability that a point (x, y) taken at random will be within the set $[R]$ of points on the x, y plane can be obtained by integrating the density over the region denoted by the set.

Hence,

$$Pr[(x, y) \in R] = \iint_R \delta(x, y)\, dy\, dx \tag{13}$$

Frequently, a proof that

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \delta(x, y)\, dy\, dx = 1 \tag{14}$$

consists of showing that Eq. (12) can be rewritten as the product

$$\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-(w^2/2)}\, dw \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-(v^2/2)}\, dv \tag{15}$$

where

$$v = (y - \mu_y)/\sigma_y)$$

and

$$\mu = (x - \mu_x)/\sigma_x$$

so that

$$w = (\mu - \rho v)/(1 - \rho^2)$$
$$dw = d\mu/(1 - \rho^2)$$

Since Eq. (15) denotes the product of two univariate normal densities, this relation is used by SIMFIT to empirically verify the approximation accuracies of the set of estimators under consideration. The same procedure applies to the bivariate extensions of Eqs. (9), (10), and (11).

The density function defined by Eq. (12) and represented in Fig. 19 was almost perfectly estimated by the subcase of estimators under investigation. The approximation for ρ was respectively given by

$$\rho = \frac{\delta}{\mu M} \log[M\mu(M)^{1/\delta}] \tag{16}$$

and

$$\rho = \frac{\delta}{\mu(M)} \log M \tag{17}$$

under conditions of 150 and 75 sample points, with δ alternatively equal to 0.95 and 0.5. Under such conditions, the estimation technique discussed does very well with an error range of 0.10 % for the density of Eq. (12) to 2 % for the bivariate extensions of Eqs. (9), (10), and (11). These include x distributed according to Eq. (9), y according to (10), x according to (10), y according to (11); x according to (9), y according to (11). A total of 30 cases was examined.

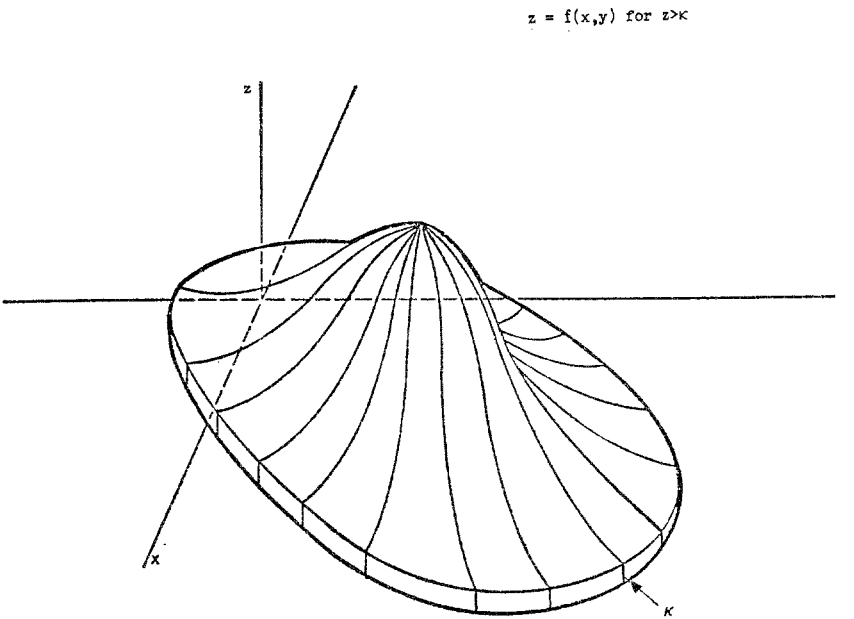

Fig. 19. Bivariate density function.

Thus, experimental evidence indicates that the estimation technique presented approximates very accurately a significant number of uni- and bidimensional distributions of both the Gaussian and mixture types.

During the indicated experimental assessments, it was discovered that the estimator being investigated could be used to recover accurately discrete bivarate densities like the ones given by Eqs. (7) and (8). Graphical representation of the former, together with its recovery (given by the wedges), is supplied in Fig. 20. Recovery error is less than 2%. Recovery representation of the latter is given by Fig. 21, in which a recovery error of 5% obtains. Recovery for both was accomplished by 200 sample points.

4. DESCRIPTION OF AN ALGORITHM FOR CLUSTERING

As indicated earlier in Section 3, each of the M patterns in the sample set of input patterns (associated with a given class of patterns) is assumed to be represented by an N-dimensional vector; e.g., the kth input pattern is written $x^k = (x_1^k, x_2^k, \ldots, x_N^k)$. A value is assigned to each vector y of N-dimensional input space by means of the distribution density function

$$f_M(y) = A \sum_{k=1}^{M} \prod_{n=1}^{N} \exp[-b_n(y_n - x_n^k)^2]$$

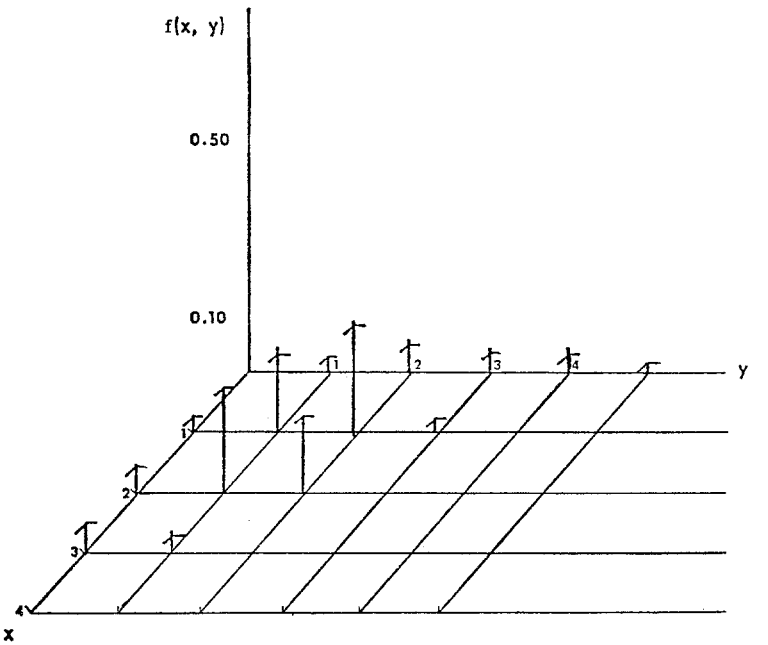

Fig. 20. Graph of density function.

where $b_n > 0$, and $A > 0$ is a normalizing constant. If the proper values of b_n are chosen[3] and the sample set $[x^k]_{k=1}^M$ is large enough, the local maxima of f_M will closely approximate the modes of the underlying distribution from which the sample patterns are drawn. In this investigation, the algorithm for locating the local maxima of f_M involves either an iterative gradient technique or the use of Matyas' random optimization theorems.[10] In the interests of simplicity, where the former technique is used, it is assumed that $b_n = b$ for $n = 1, 2, \ldots, N$, so that the expression for f_M is

$$f_M(y) = A \sum_{k=1}^{M} \exp[-b \| y - x^k \|^2]$$

with

$$\| y - x^k \|^2 = \sum_{n=1}^{N} (y_n - x_n{}^k)^2$$

[3] J. N. Medick, in a personal communication (August 1967), has obtained an approximate lower bound for the choice of values for b that admits of easy numerical approximation:

$$b \geqslant KM^{\delta} \cdot [\int_{R_N} f^2 \, dV_x]^{5/4} / \int_{R_N} (x - \bar{x})^2 f^2 \, dV_x$$

where K is a positive constant that depends for its value on the kernel of the estimator fw expression, and R_N denotes the set of points on the x, y plane.

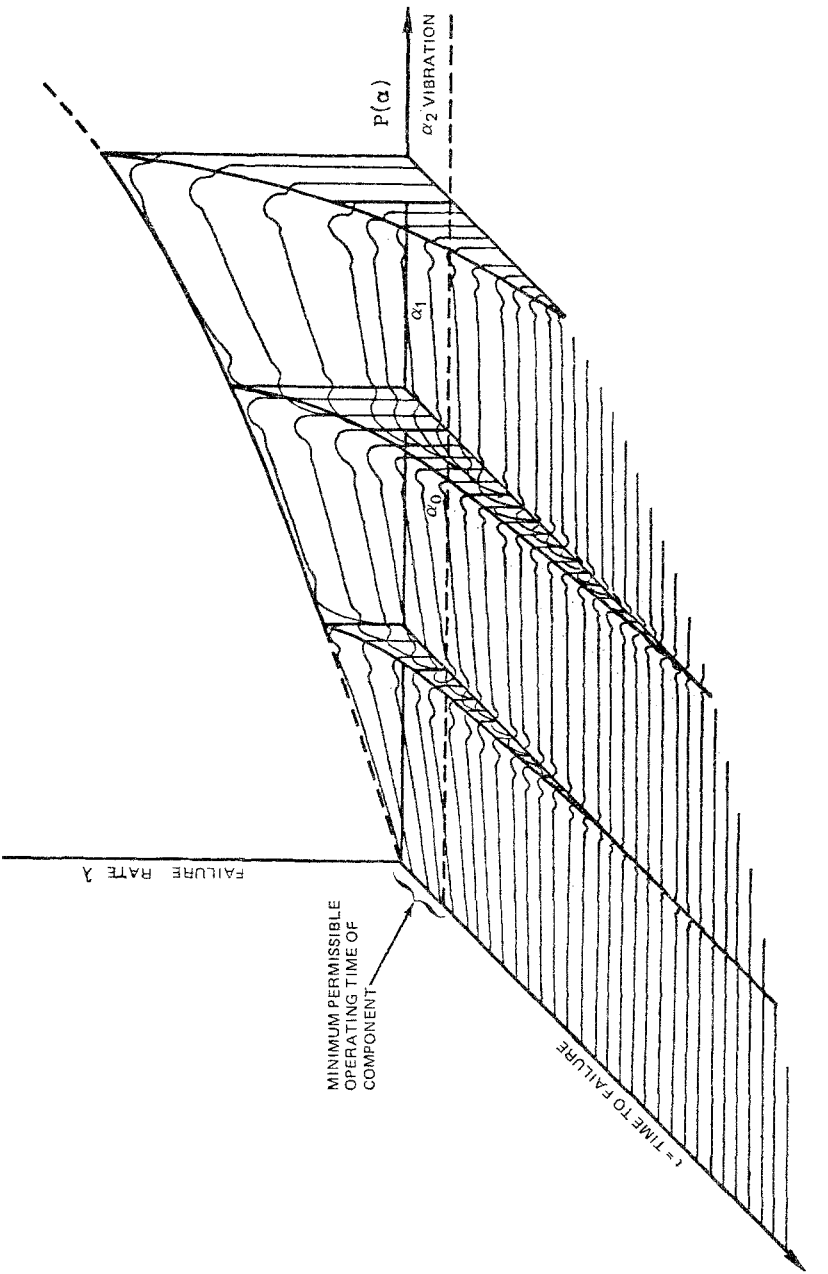

Fig. 21. Vibration frequency response surface.

Let y^0 be an arbitrary starting point (with y^0 not a stationary point of f_M); then the first approximation y^1 is given by

$$y^1 = y^0 + a_0\xi^0$$

where

$$a_0 = \frac{\| \gamma(y^0)\|}{2bf_M(y_0)}$$

$$\gamma(y^0) = \text{gradient vector of } f_M \text{ at } y = y^0$$

$$\xi^0 = \frac{\gamma(y^0)}{\| \gamma(y^0)\|} \text{ [i.e., } \xi^0 \text{ is a unit vector in the direction of } \gamma(y^0)]$$

In general, y^r is given by

$$y^r = y^{r-1} + a_{r-1}\xi^{r-1}$$

where

$$a_{r-1} = \frac{\| \gamma(y^{r-1})\|}{2bf_M(y^{r-1})}$$

$$\xi^{r-1} = \frac{\gamma(y^{r-1})}{\| \gamma(y^{r-1})\|}$$

It can be demonstrated[11] that this choice of a_{r-1} leads to a convergent process that results in a local maximum ("cluster point") of f_M. The technique generalizes to the case of interest, i.e., b_i not necessarily equal to b_j for $i, j = 1,..., N$.

5. APPLICATION TO PATTERN RECOGNITION

Figure 22 shows the conceptual design of a recognition system that decides on the class membership of an input by calculating the probabilities of class membership for each of the individual classes. The actual decision is based on the maximum likelihood principle. The estimate of the underlying distribution density f associated with a given class is approximated by a function $h(x)$ that depends on the cluster points determined by f_M and a set of N parameters that are determined to minimize the "distance" between f_M and h. A brief description of the method follows.

Consider the distribution density estimate f_M (where M is the number of samples in the data base associated with a given class), and the collection of prototypes (local maxima), $z^1, z^2,..., z^{p_1}$, obtained from f_M using the clustering algorithm. The points $[z^k]_{k=1}^{p_1}$ may be used to yield an approximation to the underlying distribution density of the form[4]

$$h_{p_1}(x) = \lambda \sum_{r=1}^{p_1} \xi_r \prod_{n=1}^{N} \exp[-w_n(x_n - z_n{}^r)^2]$$

[4] This form is similar to that used in the distribution recovery technique to obtain f_M.

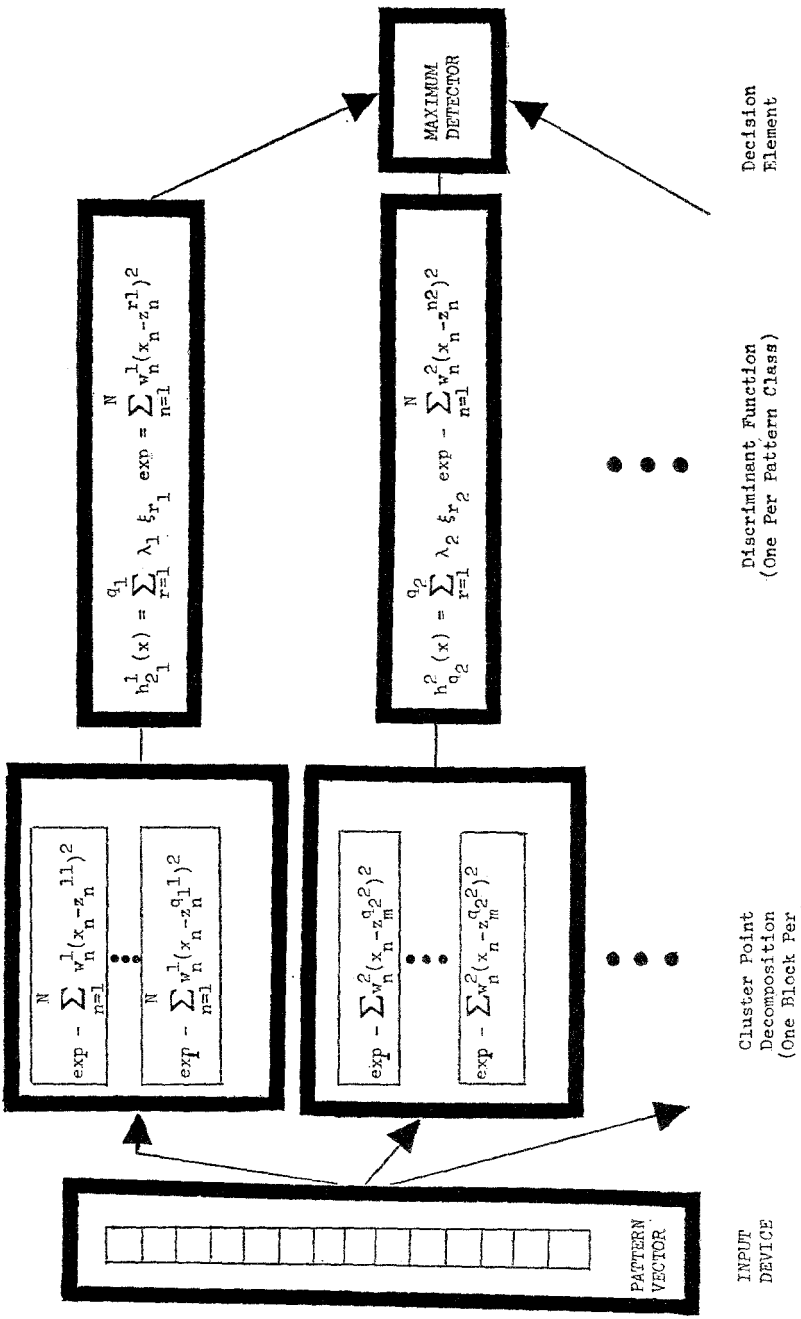

Fig. 22. Pattern decision scheme.

where $\xi_r = f_M(z^r)$, $w_n > 0$, and λ is a normalizing constant. The positive constants w_n are determined in order to minimize the quantity

$$\|f_M(x) - h_{p_1}(x)\|^2$$

If $\|f_M - h_{p_1}\|^2$ is large, the clustering algorithm may be used to find the local maxima of

$$(f_M - h_{p_1}), \qquad \text{i.e.,} \quad z^{p_1+1}, z^{p_1+2}, \ldots, z^{p_1+p_2}$$

The points of set $[z^k]_{k=p_1+1}^{p_2}$ are considered "second-order" cluster points and the augmented set $[z^k]_{k=1}^{p_1+p_2}$ contains more information about the underlying distribution in the sense that $h_{p_1+p_2}$ is, in general, a closer approximation of f than h_{p_1}.

Similarly, third-order, fourth-order, etc., cluster points can be defined and their efficiency in characterizing the distribution assessed using as a criterion the metric $\|f_M - h_{p_1+p_2+\cdots p_L}\|^2$. In this way, a set of prototypes $[z^k]_{k=1}^{p_1+p_2+\cdots+p_L}$ can be generated. Although in general

$$\lim_{L\to\infty} \sum_{j=1}^{L} p_j = \infty$$

the effect of choosing L such that

$$q = \sum_{j=1}^{L} p_j \ll M$$

can be effectively measured.

As indicated, Fig. 22 actually represents a pattern-recognition scheme based on the maximum likelihood principle. For each class, the system develops the function

$$h_{q_j}^j(x) = \lambda_j \sum_{r=1}^{q_j} \xi_{rj} \prod_{n=1}^{N} \exp\{-w_n^j(x_n - z_n^{rj})^2\}$$

where j denotes the class and

$q_j \equiv$ number of cluster points used in jth class

$\lambda_j \equiv$ normalizing constant for jth class

$\xi_{rj} = f_M{}^j(z^{rj})$

$z^{rj} \equiv r$th cluster point of jth class $= (z_1^{rj}, z_2^{rj}, \ldots, z_N^{rj})$

$w^j \equiv (w_1{}^j, w_2{}^j, \ldots, w_n{}^j) =$ vector that minimizes $\|f_M{}^j - h_{q_j}^j\|^2$

A system such as that shown in Fig. 22 can be useful in determining "dominant" regions in pattern space. For class j a "strong" cluster point z^{ij} may be defined as a cluster point that is highly successful in separating the jth class from the remaining classes, where the degree of success can be evaluated by examining the output of the ith subblock of the jth block in Fig. 22 for each of the input patterns. Thus, if z^{ij} is a strong cluster point, one may associate z^{ij} with a "pure signal" representing a dominant feature of class j while neighboring values are z^{ij} degraded by "noise." "Dominant features" (where "features" are normally associated with the individual components of a pattern vector) may be acquired by entering a pattern vector with a subset of its components made equal to zero while measuring the effect on the ability of the ith subblock of the jth block to separate the jth class from the remaining classes.

The FORTRAN program, together with approximate number of statements, for implementing the scheme of Fig. 12 is briefly described below. Computational burden for both the cluster decomposition and discriminant functions appears to be minimal.

Let $F(x) = e^{-x}$

Determine x, where

$$\sum_{i=1}^{N} W_n(x_n - z_n)^2$$

	$x = 0 \cdot 0$	$I = 3,\ W = 2$	
	→DO 1 $i = 1, N$	$I = 2,\ W = 1$	
N	$x = x_n - z_n$	$I = 4,\ W = 2$	~25N
	$x = x + W_n * x_1 * x_1$	$I = 15,\ W = 8$	
	1 continue	$I = 3,\ W = 1$	

Determine e^x

	$y = 1 \cdot 0$	$I = 3,\ W = 2$	
	$F(x) = 1 \cdot 0$	$I = 3,\ W = 2$	
	→DO 2 $i = 1, 10$	$I - 2,\ W = 1$	
10	$F(x) = F(x) + \frac{x}{y}$	$I = 11,\ W = 4$	~280
	$y = y * y + 1$	$I = 9,\ W = 4$	
	$x = x * x$	$I = 6,\ W = 3$	
	2 continue	$I = 2,\ W = 1$	
	$F(x) = \frac{1}{F(s)}$	$I = 6,\ W = 3$	

Let $h(x) = \sum_{r=1}^{8} \lambda_1 \xi_i e - \sum_{j=1}^{N} W_j(x_j - z_j{}^r)^2$

Determine x

	$x = 0 \cdot 0$	$I = 3,\ W = 2$	
	→Do 3 $n = 1, N$	$I = 2,\ W = 1$	
N	$x_3 = x_n - z_n$	$I = 5,\ W = 2$	$\sim 25N$
	$x = x + W_n * x_3 * x_3$	$I = 15,\ W = 8$	
	3 continue	$I = 3,\ W = 1$	

Determine $\sum_{j=1}^{q} \lambda_1 \xi_j e^{-x}$

	$y = 1 \cdot 0$	$I = 3,\ W = 2$	
	$F(x) = 1 \cdot 0$	$I = 3,\ W = 2$	
	→DO 4 $i = 1, 10$	$I = 2,\ W = 1$	
10	$F(x) = F(x) + \frac{x_3}{y}$	$I = 11,\ W = 1$	~ 300
	$x_3 = x_3 * x_3$	$I = 6,\ W = 3$	
	$y = y * y + 1$	$I = 8,\ W = 4$	
	4 continue	$I = 2,\ W = 1$	

Determine $F(x) = \frac{1}{F(s)}$

	$\lambda(x) = 0 \cdot 0$	$I = 3,\ W = 1$	
	→DO 5 $j = 1, q$	$I = 2,\ W = 1$	
q	$\lambda(x) = \lambda(x) + \lambda_1 * \xi_j * F(x)$	$I = 15,\ W = 8$	$20q$
	5 continue	$I = 3,\ W = 1$	

6. CONCLUSIONS

When these techniques are applied (as with any that employ a distribution estimate), it is important to start with a large data base. The degree of generalization that a final design is capable of providing is clearly dependent on how accurately the actual distribution of input patterns is represented by the sample set of input data. A large data base leads to high generalization capability. However, even with a relatively small data base, it appears that the described techniques provide very good generalization for classificatory systems based on maximum likelihood decisions.

Further, the techniques described have the advantage that the design time varies linearly with the number of sample patterns employed. This constrasts markedly with many error-correcting design techniques, in which the sample patterns are introduced cyclically until a satisfactory recognition network is obtained. Under such conditions, design time (i.e., computing time) usually varies as the square or cube of the number of sample patterns.

Finally, there remain several problems to be resolved. Thus, the perfor-

mance of the distribution-recovering technique articulated in Section 3 should be checked against known N-dimensional distributions, where $N > 2$. It is also desirable that some explicit method for determining the values of w_n that minimize $\|f_M - h_g\|^2$ be developed. If this latter cannot be done in closed form, an algorithm for iteratively determining w_n should be generated. But these are problems requiring further investigation.

ACKNOWLEDGMENTS

In the preparation of much of the material for this paper, the author was aided by the insights and suggestions of J. Medick, Drs. R. D. Joseph, V. K. Murthy, and, especially, W. F. Webber, who kindly read the entire revised manuscript. Also valuable were the suggestions of the referees and the programming support furnished by R. J. Brundy.

REFERENCES

1. E. J. Wegman, Nonparametric probability density estimation: a summary of available methods, *Technometrics* **14**(3) (August 1972).
2. E. Parzen, On the estimation of a probability density function and mode, *Ann. Math. Stat.* **33**, 1065–1076 (1962).
3. V. K. Murthy, Estimation of probability density, *Ann. Math. Stat.* **36**, 1027–1031 (1965).
4. V. K. Murthy, "Nonparametric Estimation of Multivariate Densities with Applications," in *Multivariate Analysis* (Academic Press, New York, 1966).
5. T. Cacoullos, Estimation of a multivariate density, *Ann. Inst. Stat. Math.* **10**, 186–190 (1966).
6. E. A. Nadaraya, On the estimation of density functions of random variables, *Trans. Georgian Acad. Sci.* **32**(2) (1963).
7. E. A. Nadaraya, "On Nonparametric Estimates of Density Functions and Regression Curves," in *Theory of Probability with Applications*, Vol. 10, pp. 186–190.
8. D. C. Dorrough, "Distribution Recovery Techniques for Reliability," Interim Report to the Office of Naval Research, Ultrasystems, Incorporated (July 1973); copies available from the author.
9. "SIMFIT: Program Description and Users Manual," System Development Corporation (1965); included in CYBERNET, Control Data Corporation.
10. J. Matyas, Random optimization, *Autom. Remote Control (USSR)* **26**(2), 244–251 (February 1965).
11. D. C. Dorrough, Some applicable results on clustering and distribution recovery, submitted to *Commun. Stat.*