NASA CR-
140297

[redacted]

(NASA-CR-140297) CLASSIFICATION BY MEANS OF B-SPLINE POTENTIAL FUNCTIONS WITH APPLICATIONS TO REMOTE SENSING (Rice Univ.) 9 p HC $4.00 CSCL 12A

N74-34985

Unclas
G3/19 51189

ICSA

INSTITUTE FOR COMPUTER SERVICES AND APPLICATIONS

RICE UNIVERSITY

Classification by Means of B-Spline Potential Functions with Applications to Remote Sensing

by
J.O. Bennett*, R.J.P. de Figueiredo
and J.R. Thompson
Dept. of Mathematical Sciences
Rice University

ABSTRACT

This paper presents the method of potential functions using B-splines as potential functions in the estimation of likelihood functions (probability density functions conditioned on pattern classes) or of the resulting discriminant functions. Integrated means square consistency of this technique is discussed. Experimental results of using the likelihood functions thus obtained in the classification of remotely sensed data are given.

Institute for Computer Services Applications
Rice University
Houston, Texas

May, 1974

Research supported in part by NASA contract NAS 9-12776
*Now with Esso Production Research, Houston, Texas

Reprint of an Invited Paper from the PROCEEDINGS OF THE SIXTH SOUTHEASTERN SYMPOSIUM ON SYSTEM THEORY, Baton Rouge, La. (Sponsored by IEEE).

CLASSIFICATION BY MEANS OF B-SPLINE POTENTIAL FUNCTIONS WITH APPLICATION TO REMOTE SENSING

John O. Bennett*
Esso Production Research Co.
Houston, Texas

Rui J. P. de Figueiredo+
Rice University
Houston, Texas

James R. Thompson‡
Rice University
Houston, Texas

Abstract

This paper presents the method of potential functions using B-splines as potential functions in the estimation of likelihood functions (probability density functions conditioned on pattern classes) or of the resulting discriminant functions. Integrated mean square consistency of this technique is discussed. Experimental results of using the likelihood functions thus obtained in the classification of remotely sensed data are given.

***

The method of "potential functions" (also called "kernel functions") for the direct construction of likelihood functions and discriminant functions has been widely discussed in the literature on statistics and pattern classification (see for example [4] and the references therein). In what follows, first we review very briefly this method. Second, we present its integrated-mean-square (IMS) consistency and give a formula for the value of the mesh parameter h(N) (to be defined in section 2) which is optimal with respect to IMS convergence. Next, we discuss the use of multivariate B-splines as potential functions, bringing into the discussion the IMS consistency criteria mentioned above. Finally, we present some of the experimental results obtained when likelihood functions constructed by means of B-spline potential functions were used to classify remotely sensed data pertaining to the Purdue LARS flight line C1.

1. Likelihood Functions and Discriminant Functions in Pattern Classification

As a preamble to our results, let us briefly recall the Bayesian solution to the pattern classification problem.

Suppose that observations made on patterns, which are to be classified as pertaining to one of the pattern classes $H^1, \ldots, H^M$, appear as n-vectors belonging to the real Euclidian space R^n. Then any given observation $x = \text{col}\,(x_1, \ldots, x_n)$ may be viewed as a realization of a random vector $X = \text{col}\,(X_1, \ldots, X_n)$. Associated with each pattern class H^j, $j = 1, \ldots, M$, there is the conditional probability density function** $f_X(x/H^j)$, called the likelihood function for the class H^j, and the prior probability H^j for that class. The Bayes decision rule, which minimizes the probability of misclassification, consists of classifying any observed x as arising from H^j if

$$P_j f_X(x/H^j) - P_i f_X(x/H^i) > 0, \quad i \neq j, \ i=1,\ldots,n \tag{1}$$

The left side of (1)

$$g_{ji}(x) \equiv P_j f_X(x/H^j) - P_i f_X(x/H^i) \tag{2}$$

is called a discriminant function. Since (1) is equivalent to

$$\tilde{g}_{ji}(x) \equiv \log(f_X(x/H^j)/f_X(x/H^i)) + \log(P_j/P_i) > 0, \tag{1a}$$

$\tilde{g}_{ji}(x)$ is also sometimes called a discriminant function.

For $j = 1, \ldots, M$, let there be given the n-vectors $y_1^{(j)} = \text{col}(y_{11}^{(j)}, \ldots, y_{1n}^{(j)})$, $y_2^{(j)}, \ldots, y_{N_j}^{(j)} = \text{col}(y_{N_j 1}^{(j)}, \ldots, y_{N_j n}^{(j)})$ constituting the training set $T_j(N_j)$ belonging to the pattern class H^j. The problems to which we will be addressing are:

(a) Given $T_j(N_j)$ construct an estimate $\hat{f}_X(x/H^j, T_j(N_j))$ of $f_X(x/H^j)$;

(b) Given $T_j(N_j)$ and $T_i(N_i)$ construct an estimate

$$\hat{g}_{ji}(x; T_j(N_j), T_i(N_i)) \text{ of } g_{ji}(x).$$

For simplicity in notation, from now on we will drop the superscript and subscript j whenever it is clear that we are referring to the estimation of a likelihood function pertaining to a given class H^j, and rewrite $f_X(x/H^j)$ and $\hat{f}_X(x/H^j, T_j(N_j))$ simply as $f_X(x)$ and $\hat{f}_X(x/T(N))$, respectively.

2. The Method of Potential Functions

We will now indicate how the method of potential functions is used in the solution of problems (a) and (b) above.

According to this method, in the solution of problem (a), the estimate of $f_X(x/T(N))$ is constructed in the form

$$\hat{f}_X(x/T(N)) = N^{-1} \sum_{k=1}^{N} \varphi(x, y_k), \tag{3}$$

where $\varphi(x,z)$ called a "potential function" or "kernel function" is a real-valued function of the n-vectors x and z, satisfying appropriate conditions.

* Supported by NASA contract NAS-9-12776 while at Rice University.
\+ Supported by the NSF Grant GK-36375.
‡ Supported by ONR Grant NR-042-283.

** Even though $f(x)$ is the correct notation for the "value" at x of a function f or $f(\cdot)$, we often use the same notation $f(x)$ for the "function" and "function values" when the meaning is clear from the context.

For the one-dimensional case, i.e. for x and y_k in R^1, Parzen [7] was one of the first investigators to suggest the construction of a probability density function using (3) and for this reason (3) is often called a Parzen estimator of the probability density function $f_X(x)$.

Parzen suggested specifically potential functions of the form

$$\varphi(x,z) = h^{-1}(N)K(h^{-1}(N)(x-z)), \qquad (4)$$

where h(N), is a "mesh parameter", dependent on N, sufficiently small so as to validate the assumption of $f_X(x)$ being nearly constant on any interval (z-h(N), z+h(N)). Parzen [7] gave conditions on K and h(N), which guarantee mean square consistency of (3) for a wide class of densities. He also gave a formula for optimal h(N), i.e. the values of h(N), n = 1, 2, ..., which maximize the rate of convergence of an approximation to the mean square error to zero.

Parzen's results were extended to the n-dimensional case by Murthy [6] and Cacoullos [1]. Note that in this case, we have in general n scalar mesh parameters $h_1(N)$, ..., $h_n(N)$ and (4) is replaced by

$$\varphi(x,z)=(h_1(N)h_2(N)\ldots h_n(N))^{-1}K(H^{-1}(N)(x-z)), \qquad (5)$$

where

$$H(N) = \mathrm{Diag}(h_1(N), \ldots, h_n(N)) \qquad (6)$$

However, by suitable normalization, we may make all $h_i(N)$, i = 1, ..., n the same, i. e.

$$h_1(N) = h_2(N) = \ldots = h_n(N) = h(N), \qquad (7)$$

which when substituted in (5) leads to

$$\varphi(x,z)=h^{-n}(N)K(h^{-1}(N)(x-z)). \qquad (8)$$

So from now on, without loss in generality we will assume equation (7) and equation (8) hold.

Referring next to Problem (b), it is clear that the above technique can be used to estimate $g_{ji}(x)$ directly from the training sets $T_j(N_j)$ and $T_i(N_i)$. In fact, let the elements of $T_j(N_j)\cup T_i(N_i)$, ordered in any arbitrary way, be labeled z_k, k = 1, ..., N_j+N_i, and define the function

$$\psi_{ji}(x,z_k)\equiv N_i P_j \varphi(x,z_k)u_j(z_k) - N_j P_i \varphi(x,z_k)u_i(z_k), \qquad (9)$$

where, for ℓ = j,i

$$u_\ell(z_k) = \begin{cases} 1, & \text{if } z_k \in T_\ell(N_\ell) \\ 0, & \text{otherwise.} \end{cases} \qquad (10)$$

Then substituting (3) in (2) and using (9), we obtain

$$\hat{g}_{ji}(x;T_j(N_j),T_i(N_i))=(N_jN_i)^{-1}\sum_{k=1}^{N_j+N_i}\psi_{ji}(x,z_k) \qquad (11)$$

It clearly follows from this definition that our consistency results developed for (3) apply also to (11).

3. Integrated Mean Square Consistency and Optimality

The above-mentioned consistency and optimality results are with respect to the mean square error

$$E\{(\hat{f}_X(x/T(N)) - f_X(x))^2\} \qquad (12)$$

for a given x. However, in approximating f_X one is most likely to have some a priori knowledge of the "global behavior", like "global smoothness", of the function to be approximated rather than its "local behavior". In such circumstances, the integrated mean square (IMS) error criterion in the coice of optimal h(N) becomes more meaningful as we shall explain in section 4. The IMS error is defined by

$$V(T(N)) = \int E\{(\hat{f}_X(x/T(N)) - f(x))^2\}dx, \qquad (13)$$

where the integration is performed over R^n.

The following result is proved in the Appendix.

Theorem 1. Suppose

(I) The random samples y_1, ..., y_N are independently and identically distributed as X whose density is f_X;

(II) $f_X \in L^2(R^n)$;

(III) $K: R^n \to R^+$ (where R^+ = set of nonnegative real numbers) is such that

(III-1) $K \in L^2(R^n)$,

(III-2) $\int_{R^n} K(x)dx = 1$,

(III-3) $\operatorname{ess\,sup}_{x\in R^n} K(x) < \infty$,

(III-4) $\lim_{\|x\|\to\infty} \|x\|K(x) = 0$,

where $\|x\| = (x_1^2 + \ldots + x_n^2)^{1/2}$, and

(IV) h(N) is such that

(IV-1) $\lim_{N\to\infty} h(N) = 0$

and

(IV-2) $\lim_{N\to\infty} Nh^n(N) = \infty$.

Then $\hat{f}_X(x/T(N))$ is a consistent estimator of $f_X(x)$ in the IMS error sense, that is

$$\lim_{N\to\infty} V(T(N)) = 0 .$$

We next seek a formula for the value $h^*(N)$ of h(N), which optimizes the IMS rate of convergence, of $\hat{f}_X$ to f_X. This is obtained by modifying Cacoullos' [1] result for the mean square convergence case as follows:

Theorem 2. Let the hypotheses of Theorem 1 hold and assume, in addition, that K is symmetric (i.e. $K(-x) = K(x)$) and f_X is thrice differentiable and such that the second partial derivative of f_X are in $L^2(R^n)$. Then within $o(h^4(N))$ the IMS error (defined by (13)) is minimized by choosing

$$h^*(N) = N^{-1/(n+4)} \left[\frac{n\|K\|_2^2}{\left\| \sum_{i=1}^{n} \sum_{j=1}^{n} \mu_{ij} \frac{\partial^2 f_X}{\partial x_i \partial x_j} \right\|_2^2} \right]^{\frac{1}{n+4}}, \quad (14)$$

where, for a function g,

$$\|g\|_k = \left(\int_{R^n} |g(x)|^k dx \right)^{1/k} \quad (15)$$

and

$$\mu_{ij} = \int_{R^n} x_i x_j K(x) dx . \quad (16)$$

The optimal rate of convergence, corresponding to the choice (14), is

$$V^*(T(N)) = (4^{-1}n+1)n^{-\frac{n}{n+4}} N^{-\frac{4}{n+4}} \left\| \sum_{i=1}^{n} \sum_{j=1}^{n} \mu_{ij} \frac{\partial^2 f_X}{\partial x_i \partial x_j} \right\|_2^{\frac{2n}{n+4}} \left\| K \right\|_2^{\frac{8}{n+4}} + o(h^{*4}(N)). \quad (17)$$

If, as we shall assume in the following section, the kernel K has the product form

$$K(x) = \prod_{i=1}^{n} K_0(x_i) , \quad (18)$$

where K_0 is an even one-dimensional kernel, then Theorem 2 further simplifies to:

Theorem 3: Under the condition on K just stated, the results (14) and (17) of the preceding theorem assume the forms (19) and (20) below:

$$h^*(N) = N^{-\frac{1}{n+4}} \left[\frac{n\|K_0\|_2^{2n}}{\sigma_0^4 \left\| \sum_{i=1}^{n} \frac{\partial^2 f_X}{\partial x_i^2} \right\|_2^2} \right]^{\frac{1}{n+4}}, \quad (19)$$

$$V^*(T(N)) = (4^{-1}n+1)n^{-\frac{n}{n+4}} N^{-\frac{4}{n+4}} \Lambda_1(f_X)\Lambda_2(K_0), \quad (20)$$

where

$$\sigma_0^2 = \int_{-\infty}^{\infty} x_i^2 K_0(x_i) dx_i , \quad (21)$$

$$\Lambda_1(f_X) = \left\| \sum_{i=1}^{n} \frac{\partial^2 f_X}{\partial x_i^2} \right\|_2^{\frac{2n}{n+4}} \quad (22)$$

and

$$\Lambda_2(K_0) = \sigma_0^{\frac{4n}{n+4}} \left\| K_0 \right\|_2^{\frac{8n}{n+4}} \quad (23)$$

We will call $\Lambda_2(K_0)$ the optimal kernel-dependent rate of convergence since it represents the part of the right side of (20) which depends on K_0. We omit proofs of theorems 2 and 3 above since they represent straightforward extensions of those in [1].

4. B-Splines as Potential Functions

It is clear from the formulas in (19) and (20) that the choice of the optimal $h^*(N)$, and hence $V^*(T(N))$, depends on the properties of f_X and on the structure of the product kernel K.

The L^2 norm of the second derivative of a function represents a measure of the "global smoothness" of the function. In this sense, according to (22), $\Lambda_1(f_X)$ is a monotonically increasing function of the "global smoothness" of f_X. Thus an a priori knowledge of the smoothness can be incorporated in the formulas (19), (22), and (20) by assigning a value to

$$\left\| \sum_{i=1}^{n} \frac{\partial^2 f_X}{\partial x_i^2} \right\|_2^2$$

in those formulas.

In picking K one would like to choose its structure so that the optimal kernel-dependent rate of convergence $\Lambda_2(K_0)$ is minimized. Then the choice of the support of K_0 represents a compromise between the minimization of its second moment and its L^2 norm. K_0 must also be at least as smooth as f_X, particularly if only a few training samples are available. Based on these considerations, we suggest for the structure of K_0 a univariate B-spline, and hence for K a product of n such splines. Such a choice for K_0 constitutes a compromise between a Gaussian kernel and a square kernel

$$K_0(x_i) = \begin{cases} \frac{1}{2} , & |x_i| \leq 1 \\ 0 , & \text{otherwise,} \end{cases} \quad (24)$$

We note that a multivariate polynomial, as suggested by Specht [10], while certainly adequate for approximating a large class of discriminant functions given in the form (1a), is certainly unsuitable for the representation of K over the entire R^n since it violates the conditions (III) of Theorem 1.

Let any given component variable x_i of x be denoted by ξ. Then a univariate B-spline of degree m-1 with support on (ξ_0, ξ_m) and knots

$$\xi_0 < \xi_1 < \ldots < \xi_m , \quad (25)$$

is defined by [2,3,9]:

$$M_m(\xi) = \sum_{i=0}^{m} (m(\xi_i - \xi)_+^{m-1} / \omega'(\xi_i)), \quad (26)$$

where

$$\omega(\xi) = (\xi-\xi_0)(\xi-\xi_1)\ldots(\xi-\xi_m) \quad (27)$$

and

$$g_+(\xi) = \begin{cases} g(\xi), & \text{if } g(\xi) \geq 0, \\ 0 , & \text{if } g(\xi) < 0 . \end{cases} \quad (28)$$

If we: (1) assume that the degree of the B-spline is odd, i.e., $r \equiv m-1 \equiv 2k-1$, k a positive integer; (2) center the spline about the origin as required by Theorems 2 and 3; and (3) let the knots of the spline occur at integer values; then we may obtain the B-spline representation for K_0 (see [9]).

$$K_0(\xi) = M_{r+1}(\xi) = \frac{1}{r!} \sum_{\ell=-k}^{k} (-1)^{\ell+k} \binom{2k}{\ell+k} (\ell-\xi)_+^r, \quad (29)$$

where, as indicated above, $m = r+1 = 2k$.

Substituting (29) in (21), (23) and then (19) and (20), we obtain the formula for optimal $h^*(N)$ and the optimal kernel-dependent rate of convergence $\Lambda_2(K_0)$:

$$h^*(N) = N^{-\frac{1}{n+4}} \left[\frac{144n(M_m(0))^n}{m^2 \left\| \sum_{i=1}^{n} \frac{\partial^2 f_X}{\partial x_i^2} \right\|_2^2} \right]^{\frac{1}{n+4}} \qquad (30)$$

and

$$\Lambda_2(K_0) = \left(\frac{m}{12}\right)^{2n/(n+4)} \gamma_m^{4n/(n+4)} , \qquad (31)$$

where

$$\gamma_m = \frac{1}{(2r+1)!} \sum_{j=1}^{r+1} (-1)^{j+r+1} \binom{2(r+1)}{(j+r+1)} j^{2r+1} . \qquad (32)$$

Numerical values for $\Lambda_2(K_0)$ given by (31), for $r = 1,3$, and 5, are listed in Table I.

TABLE I

r	$\Lambda_2(K_0)$
1	.353075
3	.357836
5	.359683

Let $\eta(\mu,\Sigma;x)$ denote the value at x of the normal density with mean μ and covariance matrix Σ . For $K_0(x) = \eta(0,\sigma^2;x)$ we have

$$\|\eta^{(2)}(0,\sigma^2; \cdot)\|_2^2 = (3/8\pi^{-.5}\sigma^{-5}). \qquad (33)$$

Using (33), with $\sigma = 1$, in (30), we get the formulas for $h^*(N)$, for any dimension n and $r = 1,3$, presented in Table II.

TABLE II

r	$h^*(N)$
1	$\left[\frac{36n(.66666)^n}{.2115}\right]^{1/(n+4)} N^{-1/(n+4)}$
3	$\left[\frac{9n(.49365)^n}{.2115}\right]^{1/(n+4)} N^{-1/(n+4)}$

5. Computer Simulation Results

In this section we present some of the computer simulation results performed on the Rice University IBM 370/155 digital computer for the purpose of testing how well the B-spline potential function algorithm performs in the construction of likelihood functions.

Given a set of samples $T(N) = \{y_1,\ldots,y_n\}$, where each y_i is an independent realization of a random variable X with density f_X, let the sample mean $\tilde{\mu}$ and sample covariance matrix $\tilde{\Sigma}$ be defined in the usual way, i.e.

$$\tilde{\mu} = \frac{1}{N} \sum_{i=1}^{N} \mu_i \qquad (34)$$

and

$$\tilde{\Sigma} = \frac{1}{N} \sum_{i=1}^{N} (y_i - \tilde{\mu})(y_i - \tilde{\mu})^T , \qquad (35)$$

where the superscript T denotes the transpose. Then $\eta(\tilde{\mu},\tilde{\Sigma};x)$ will be called the "sample normal density".

The simulation results mentioned above are shown in Figs. 1 through 6.

Fig. 1 is a graphic display of the bimodal density

$$f_X(x) = p\eta(\mu_1,\Sigma;x) + (1-p)\eta(\mu_2,\Sigma;x) \qquad (36)$$

with the mixing parameter $p = .5$, and $\mu_1 = \text{col}(-2,0)$, $\mu_2 = \text{col}(6,0)$ and

$$\Sigma = \begin{pmatrix} 5.75 & 4.34 \\ 4.34 & 6.64 \end{pmatrix} . \qquad (37)$$

Figs. 2 and 3 are displays of $\hat{f}_X(x/T(N))$ obtained by the B-spline potential function algorithm corresponding respectively to $N=50$ and $N = 300$ samples from the above density. Fig. 4 shows the sample normal density approximation of the same density on the basis of 50 samples.

To show the effect of the increase in dimensionality on the performance of our algorithm, we present in Figs. 5 and 6 the cross-sections through the x_1-axis of the density estimates, obtained by the B-spline potential function algorithm, of the bimodal density (36) on four- and six-dimensional spaces under the conditions given in those figures.

In all this work we used cubic B-splines with the mesh parameter $h^*(N)$ equal to the second entry in Table II.

From the above few results we conclude that the B-spline potential function algorithm appears to fare well in the construction of likelihood functions from only a modest number of samples and with densities that are not necessarily unimodal and on spaces that are not necessarily of too low a dimension.

6. Application to Remote Sensing

To test the effectiveness of the B-spline potential function algorithm for classification, discriminant functions were obtained from the likelihood functions generated by the algorithm, for Bayesian classification of agricultural crops. The algorithm was based on cubic B-splines chosen as in section 4 of this paper, and was implemented in the LARSYSAA VERSION 2 which was developed at the Laboratory for Applications of Remote Sensing, Purdue University, Lafayette, Indiana. We also performed classification using sample normal densities as likelihood functions. (See the beginning of section 5 for the definition of "sample normal density.")

The data used in our experiments pertained to the Purdue LARS flight line C1, which has been widely employed for testing algorithms on remote sensing. This data consists of the output of a twelve channel spectrometer which analyzes the reflected radiance from the object being

sensed. Let a given crop field belonging, say, to the pattern class H^j, be discretized (partitioned) into N_j points, called "resolution elements". The spectrometer maps each resolution element into a 12-tuple of real numbers, say $y_k^{(j)} = \text{col}(y_{k1}^{(j)},\ldots,y_{k,12}^{(j)})$, and the whole field provides N_j such 12-dimensional vector samples $y_1^{(j)},\ldots,y_{N_j}^{(j)}$, which can be used to train a classifier in the acquisition of the likelihood function corresponding to H^j.

Our first example is designed to show the poor results obtained when normality is assumed on bimodal data. In this example, we used data from only one channel, namely Channel 1 (.40μ to .44μ), to classify data corresponding to the two bimodal pattern classes: H^1: RED CLOVER HAY and CORN1; and H^2: BARE SOIL1 and ALFALFA1. Figs. 7 and 8 show the histograms of the classes. The percentages of the number of correct classifications, for a typical set of observations corresponding to H^1 and H^2, are indicated in Table III, both for the algorithm presented here and for the sample normal classification algorithm.

TABLE III

	Potential Function Algorithm	Sample Normal Algorithm
H^1	86%	26%
H^2	98%	.3%

It is clear that the much superior performance of the potential function algorithm in relation to the sample normal algorithm may be attributed to the bi-modality of the data.

Our second example is for the purpose of testing the effectiveness of the potential function algorithm for normal data. In this example, we used 3 channels, namely Channels 1 (.40μ - .44μ), 10 (.66μ - .72μ), and 12 (.60μ - 1.00μ), to classify H^1: SOYBEANS, H^2: CORN, H^3: OATS, and H^4: WHEAT. The percentages of correct classifications are displayed in Table IV, for H^1 and H^2.

TABLE IV

	Potential Function Algorithm	Sample Normal Algorithm
H^1	97%	99%
H^2	94%	99%

Even though the efficiency of classification by the potential function algorithm is lower than by the sample normal, we note that the ability of the potential function algorithm to classify effectively data that is normal is comparable with the sample normal in quality of classification.

Acknowledgement

We are much indebted to Dr. D. Van Rooy for implementing the classification algorithm into the LARSYSAA system and to Mr. Ken Baker of the Lyndon B. Johnson Space Center for his advice on the remote sensing problem investigated. It is also a pleasure to acknowledge a helpful comment from Professor Polking on the proof of Theorem 1 and the encouragement received in carrying out this research from Dr. M. S. Lynn, Director of the Rice University Institute for Computer Services and Applications.

Appendix

Proof of Theorem 1

For notational convenience let

$$K_h(x) = h^{-n}K(h^{-1}x). \tag{A-1}$$

Then, clearly* (see for example [5, p. 172]),

$$E\{(\hat{f}_X(x/T(N)) - f_X(x))^2\}$$

$$= \text{Var}(\hat{f}_X(x/T(N))) + B^2(\hat{f}_X(x/T(N))), \tag{A-2}$$

where

$$\text{Var}(\hat{f}_X(x/T(N)) = E\{\hat{f}_X^2(x/T(N))\} - E^2\{\hat{f}_X(x/T(N))\} \tag{A-3}$$

and

$$B(\hat{f}_X(x/T(N))) = E\{\hat{f}_X(x/T(N))\} - f_X(x). \tag{A-4}$$

Hence it follows that

$$\lim_{N\to\infty}\|E\{(\hat{f}_X(x/T(N)) - f_X(x))^2\}\|_1$$

$$= \lim_{N\to\infty}\|\text{Var}(\hat{f}_X(x/T(N)))\|_1$$

$$+ \lim_{N\to\infty}\|B(\hat{f}_X(x/T(N)))\|_2^2 . \tag{A-5}$$

The proof will consist of showing that each term on the right side of (A-5) tends to zero.

Since the random variables Y_i, $i = 1,\ldots,N$ are each independently distributed as X, we have, in accordance with (3), that

$$E\{\hat{f}_X(x/T(N))\} = E\{K_h(x - X)\}, \tag{A-6}$$

and

$$\text{Var}\{\hat{f}^2(x/T(N))\} = N^{-1}\text{Var}\{K_h(x-X)\}. \tag{A-7}$$

Now

$$K_h(x)*f_X(X) \equiv \int_{R^n} K_h(x-z)f_X(z)dz$$

$$= E\{K_h(x-X)\}$$

$$= E\{\hat{f}(x/T(N))\}, \tag{A-8}$$

where, in going from the third to the last member, we have used (A-6).

Similarly,

*In this Appendix, we use capital letters for symbols denoting random variables and corresponding small letters for realizations of these random variables. In (A-2), $\hat{f}_X(x/T(N))$ is to be regarded as a function of the random variables $Y_1,\ldots,Y_N$ the realizations of which are the training samples $y_1,\ldots,y_N$.

$$K_h^2(x) * f_X(x) = E\{K_h^2(x-X)\} \quad . \tag{A-9}$$

Hence, from (A-7), using (A-8) and (A-9) we get

$$\|\mathrm{Var}(f_X(x/T(N)))\|_1 = \int_{R^n} \mathrm{Var}(f_X(x/T(N)))dx$$

$$= \int_{R^n} [E\{K_h^2(x-X)\} - E^2\{K_h(x-X)\}]dx$$

$$= \|E\{K_h^2(x-X)\}\|_1 - \|E\{K_h(x-X)\}\|_2^2$$

$$= (Nh^n)^{-1}\|(K^2(x))_h * f_X(x)\|_1$$

$$-N^{-1}\|K_h(x) * f_X(x)\|_2^2 \quad , \tag{A-10}$$

where

$$(K^2(x))_h = h^{-n}K^2(h^{-1}x). \tag{A-11}$$

By Young's inequality (see [8], p. 148), we have

$$\|K_h(x) * f_X(x)\|_2^2 \leq \|f_X(x)\|_2^2 \tag{A-12}$$

since

$$\|K_h(x)\|_1^2 \quad , \tag{A-13}$$

and again by Young's inequality

$$\|(K^2(x))_h * f_X(x)\|_1 \leq \|(K^2(x))_h\|_1 \tag{A-14}$$

since

$$\|f_X(x)\|_1 = 1 \quad . \tag{A-15}$$

By a change of variables we obtain

$$\|(K^2(x))_h\|_1 = \|K(x)\|_2^2 \quad , \tag{A-16}$$

and so (A-14) becomes

$$\|(K^2(x))_h * f_X(x)\|_1 \leq \|(x)\|_2^2 \quad . \tag{A-17}$$

Using the triangle inequality on the right side of (A-10), and then substituting into it (-17) and (A-12), we obtain

$$\|\mathrm{Var}(\hat{f}_X(x/T(N))\|_1 \leq (Nh^n)^{-1}\|K(x)\|_2^2$$

$$+ N^{-1}\|f_X(x)\|_2^2 \quad . \tag{A-18}$$

Finally, resorting to the hypotheses II, III-1, and IV-2 of the theorem, we have

$$\lim_{N\to\infty}\|\mathrm{Var}(\hat{f}(x/T(N)))\|_1 = 0 \quad . \tag{A-19}$$

Now consider the bias term (A-4) and use (A-8) to write it in the form

$$B(f_X(x|T(N))) = K_h(x) * f_X(x) - f_X(x). \tag{A-20}$$

Then by Theorem 2, Part (c) in Stein [11, p.62] we get

$$\lim_{N\to\infty}\|B(\hat{f}_X(x/T(N)))\|_2 = 0. \tag{A-21}$$

Equations (A-5), (A-19), and (A-21) show that $\hat{f}_X(x/T(N))$ is a consistent estimator of $f_X(x)$ in the IMS sense.

BIBLIOGRAPHY

[1] Cacoullos, T., Estimation of a Multivariate Density, Annals of the Institute of Statistical Mathematics (Tokyo), Volume 18, pp. 179-189, 1966.

[2] Curry, H.B., and Schoenberg, I.J., On Polya Frequency Functions IV. The Fundamental Spline Functions and their Limits, Journal d'Analyse Mathematique (Jerusalem), Volume 17, pp. 71-107, 1966.

[3] de Boor, C., Package for Calculating with B-splines, Mathematics Research Center (Madison, Wisconsin), Technical Summary Report #1333 (April), 1973.

[4] Fu, K.S., Sequential Methods in Pattern Recognition and Machine Learning, Academic Press (New York) 1968.

[5] Mood, A.M., and Graybill, F.A., Introduction to the Theory of Statistics, McGraw-Hill (New York) 1963.

[6] Murthy, V.K., Non-Parametric Estimation of Multivariate Densities with Applications, Multivariate Analysis II, International Symposium on Multivariate Analysis at Wright State University (Dayton, Ohio), Academic Press (New York), 1966.

[7] Parzen, E., On Estimation of Probability Density Function and Mode, The Annals of Mathematical Statistics, (Baltimore, Maryland) Volume 33, # 3(September), pp. 1065-1076,1962.

[8] Rudin, W., Real and Complex Analysis, McGraw-Hill Book Company (New York), 1966.

[9] Schoenberg, I.J., Cardinal Interpolation and Spline Functions, Journal of Approximation Theory, (New York), Volume 2, pp. 167-206, 1969.

[10] Specht, D.F., Generation of Polynomial Discriminant Functions for Pattern Recognition, Institute of Electrical and Electronic Engineers Transactions on Electronic Computers (New York), Volume EC-16, # 3,(June) 1967.

[11] Stein, E.M., Singular Integrals and Differentiability Properties of Functions, Princeton University Press (Princeton, N.J.), 1970.

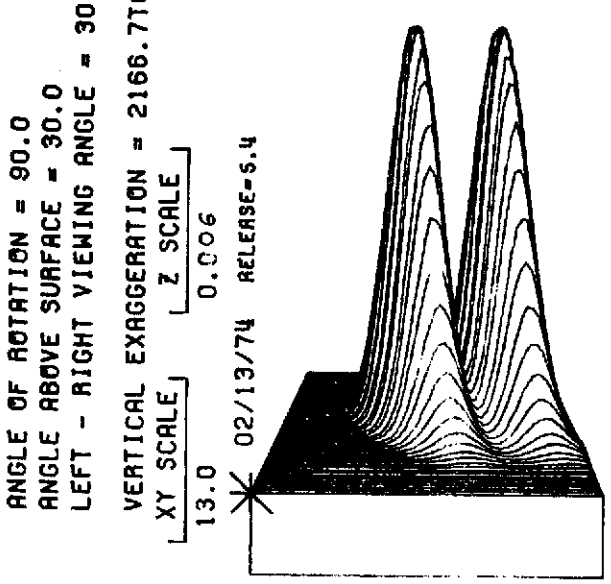


FIG.1 PERSPECTIVE VIEW OF

A BINORMAL DENSITY

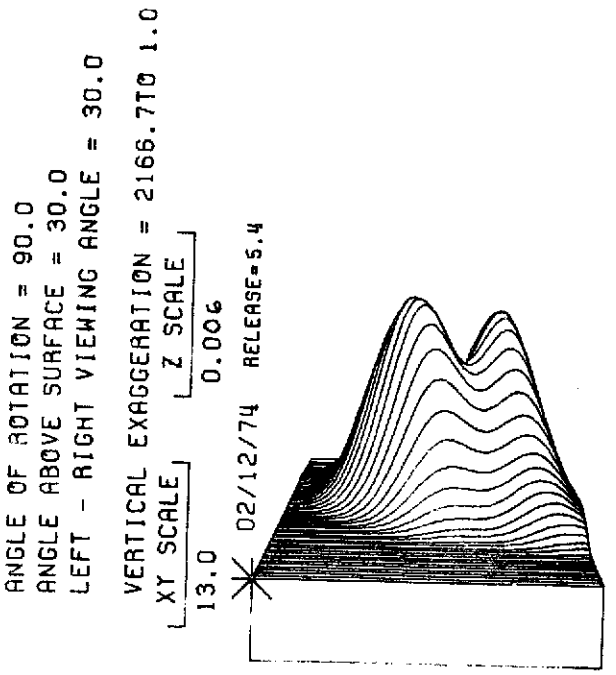


FIG 2. PERSPECTIVE VIEW OF

DENSITY KERNEL ESTIMATOR,

SAMPLE SIZE=50, CUBIC B-SPLINE

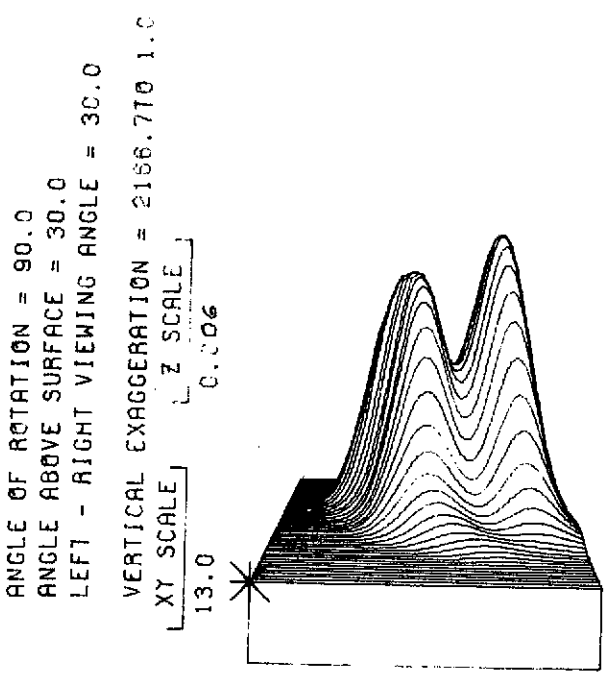


FIG.3 DENSITY KERNEL ESTIMATOR

SAMPLE SIZE =300, CUBIC B-SPLINE BASE

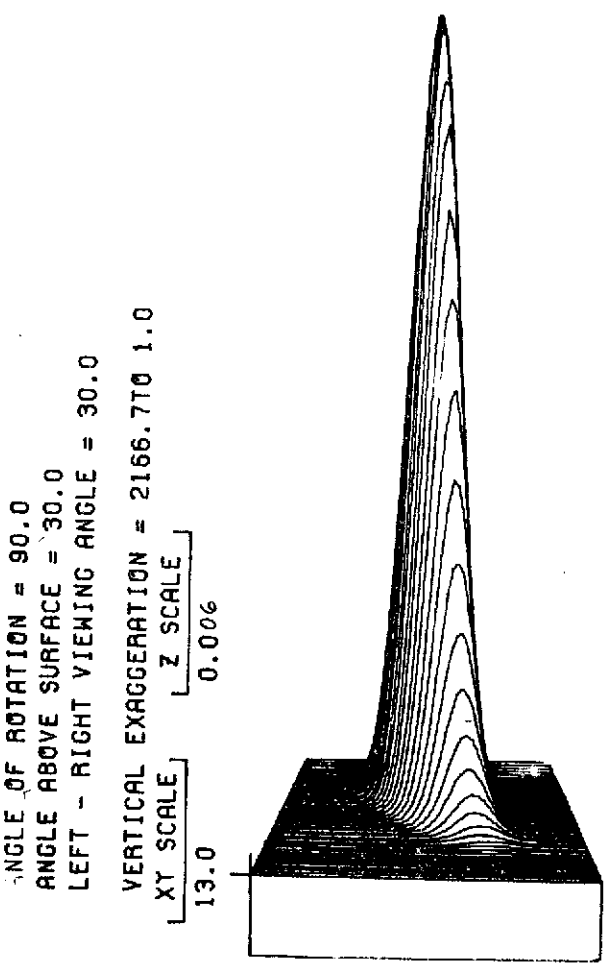


FIG4.PERSPECTIVE VIEW OF SAMPLE NORMAL

WITH SAMPLE SIZE = 50

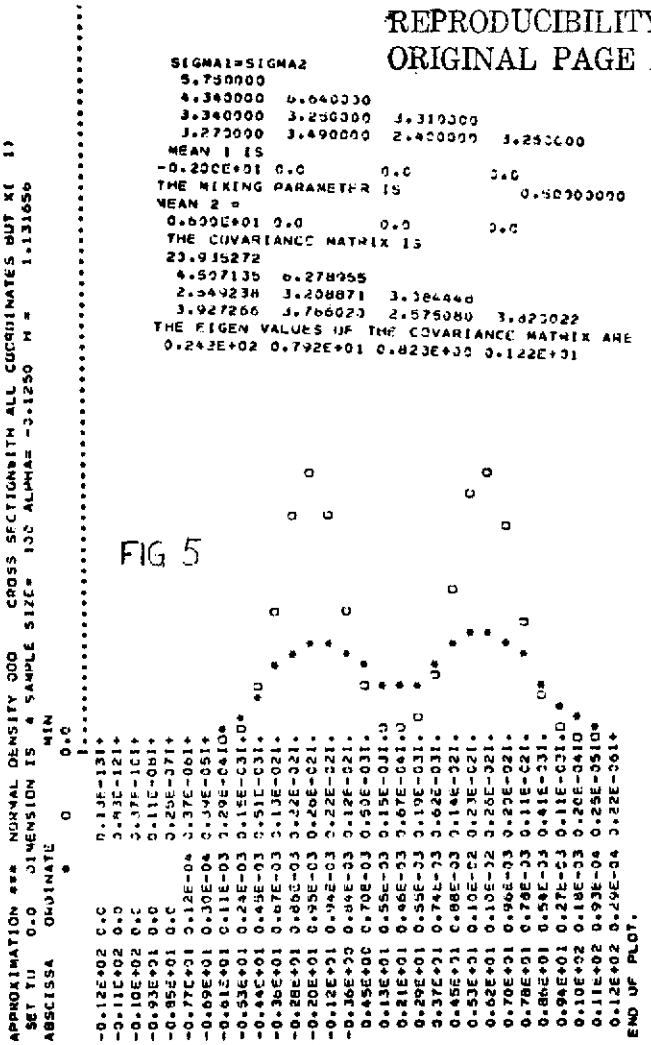


FIG 5

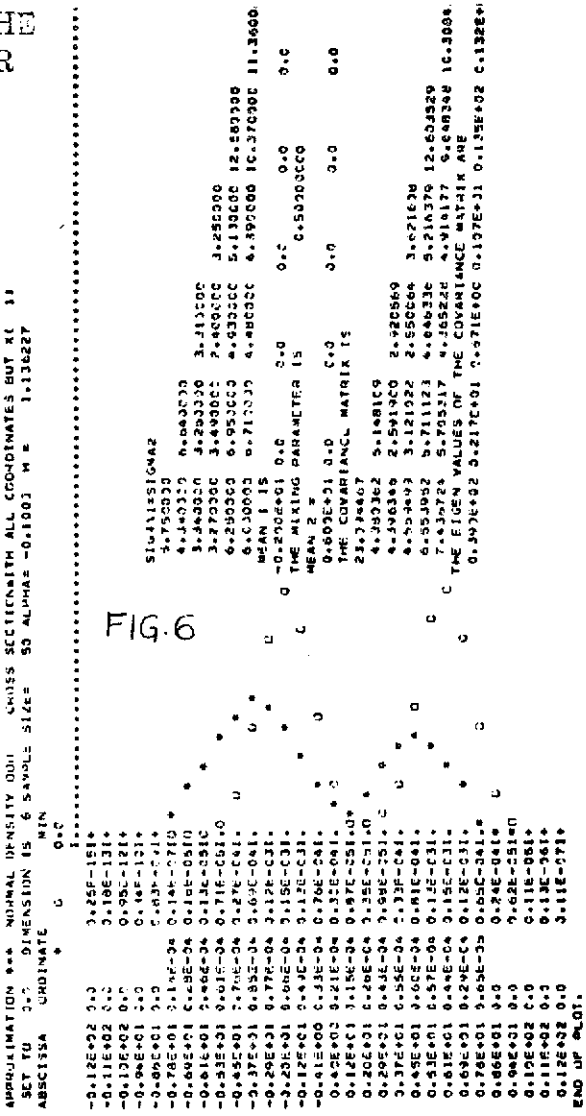


FIG.6

Pattern 1 Histogram

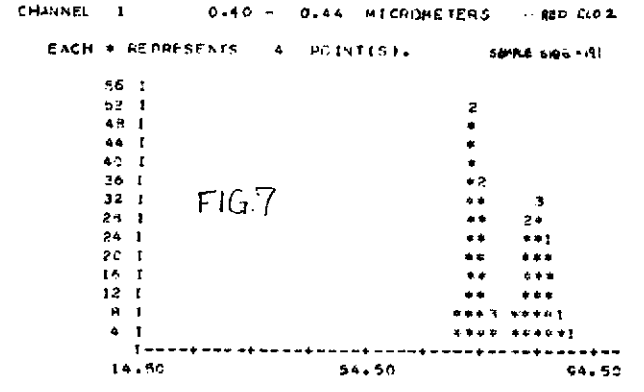


FIG.7

Table 6.2

Pattern 2 Histogram

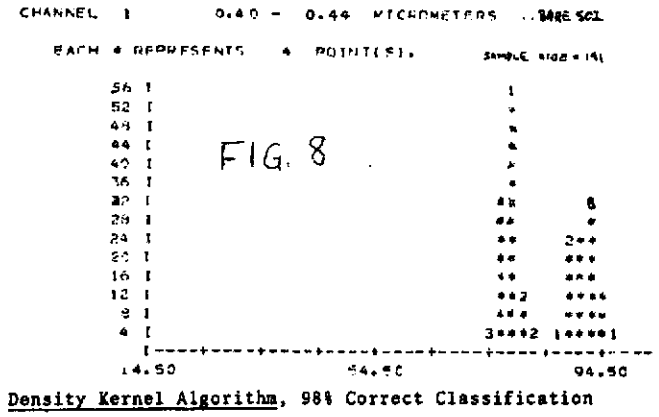
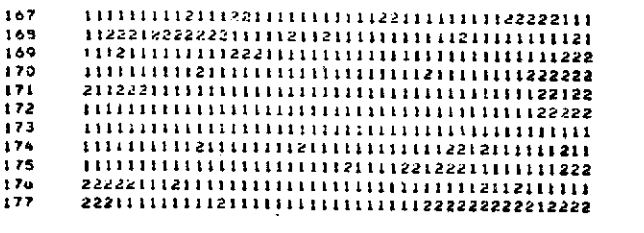
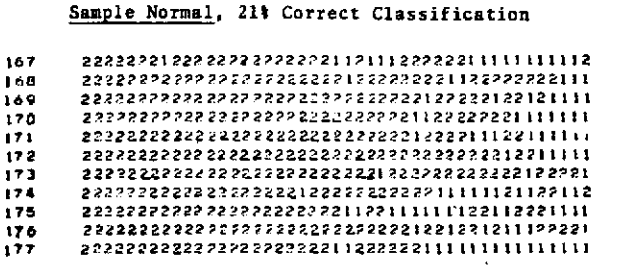


FIG. 8

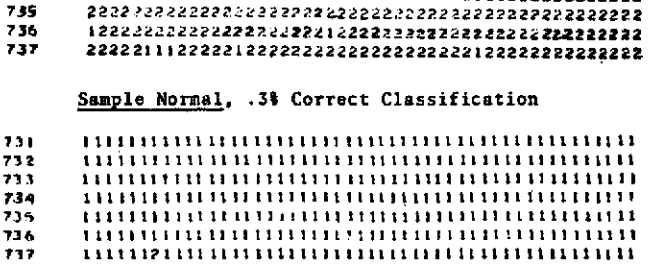
Density Kernel Algorithm, 82% Correct Classification



Sample Normal, 21% Correct Classification



Density Kernel Algorithm, 98% Correct Classification



Sample Normal, .3% Correct Classification