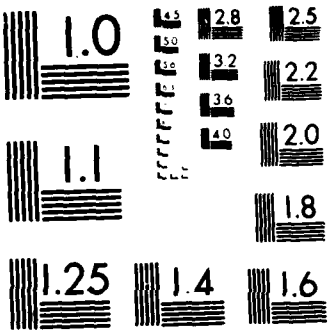
AD-A178 455 ON NON-PARAMETRIC DISCRIMINATION USING SERIES EXPANSIONS AND PROJECTION PURSUIT(U) CATHOLIC UNIV OF AMERICA WASHINGTON DC L K JONES 26 FEB 87 1/1

UNCLASSIFIED N00014-84-C-2245 F/G 17/8 NL



MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS 1963-A

N.R.L. FINAL REPORT

N00014-84-C-2245

2 Feb 87

Abstract

A method for non-parametric discrimination using series expansions is presented, and a "uniform" consistency property is proven. A completely automatic projection pursuit method for constructing a suitable series expansion is described and an application to optical detection is given.

Key Words

Series expansions. Discrimination. Nearest neighbor error. Projection pursuit.

ON NON-PARAMETRIC DISCRIMINATION USING SERIES EXPANSIONS AND PROJECTION PURSUIT

by

Lee K. Jones*

The Catholic University of America

I. Introduction and Summary

In this article we ~~first~~ develop a rigorous non-parametric theory of optimal binary discrimination using series expansions. In ~~II~~ a relationship between minimum scatter and limiting nearest neighbor error rate and its pertinence to optimal discrimination is presented. A general consistency result for series expansions is then given ~~in III~~. This motivates a data driven consistent projection pursuit alogorithm ~~in IV~~ for constructing an orthonormal basis for discriminant expansions. It is seen for projection pursuit discrimination that limiting nearest neighbor error rate plays an ~~important~~ role as mean squared residual error and relative entropy do in projection pursuit regression and density estimation, respectively. Finally, ~~in V~~, an application to motion detection of optical point sources is given and a numerical experiment is carried out.

Research performed with the facilities of the Naval Research Laboratory and supported, in part, by ONR contract#N00014-84-C2245

II Background: Scatter Criteria, Optimal Discriminants and and Limiting Nearest Neighbor Error

Let p_1, p_2 be two distinct bounded measureable probability densities on a fixed open subset Q, of R^d. We will require a regularity assumption: $p_1(X) > \kappa q(X)$ for some known small positive constand κ and a known positive probability density, q, on Q. This ensures the continuity of various functionals of the likelihood ratio p_2/p_1, provides some robustness to our model, and servees as a regularization for the inverse problem solved in III. Finally let $L_2(Q)$ be the real Hilbert space of measurable functions on Q with inner product - $(f,g) = \int_Q fgq$.

We consider the standard binary classification experiment with a class 1 occurence having prior probability P_1 and class 2 proior probability $P_2 (=1-P_1)$. (The P_i's may or may not be known). Class i is then manifested by a d-dimensional observation of a random variable with density p_i.

A binary discriminant $(f,t,\tau)$is a rule for deciding the class of an observation $X \in Q$:

$f(X) > t$	decide class 2
$= t$	decide class 2 with probability τ, class 1 with probability $1 - \tau$
$< t$	decide class 1

Without loss of generality we will consider only two types of optimality for discriminants: minimum expected total error (assuming P_1 and P_2 are known) and minimum class 2 error given class 1 error = $\bar{\alpha}$ (Neyman-Pearson problem at level $\bar{\alpha}$). The measurable map f: $Q \to R$ will be called the discrminant function.

A discriminant function f will be called optimal if (f,t,τ) is optimal for some t and τ. Since finding t and τ (given f) is only a univariate estimation problem, f will be our major concern.

For a given discriminant function f, the <u>between class scatter</u> of f, B(f), is given by

$$(1)\quad B(f) = (E_2 f - E_1 f)^2 = \left(\int_Q f p_2 - \int_Q f p_1\right)^2$$

and the <u>within class scatter of weight α</u> $(0<\alpha<1)$ of f, $W_\alpha(f)$, is given by

$$(2)\quad W_\alpha(f) = \alpha \, VAR_1 \, f + (1-\alpha) VAR_2 f$$

Various combinations of (1) and (2) have been uesed to measure the effectiveness of f. For instance $\frac{1}{2} B(W_{1/2})^{-1}$, known as the <u>Fisher criterion</u>, can be used to choose a near optimal linear f (see [1], [2], [3]). Since we are interested in general non-linear f, we restrict ourselves henceforth to $f \in \mathcal{F} = \{g: g: Q \to R, \int_Q g p_2 = 1, \int_Q g p_1 = 0\}$.

For this case (2) reduces to

$$(2')\quad W_\alpha(f) = \int_Q f^2 (\alpha p_1 + (1-\alpha) p_2) - (1-\alpha)$$

The following result is similar to those in [4], [5]; which characterize optimal discriminants as functions maximizing certain scatter criteria.

<u>Theorem 1</u> The global minimum $\bar{f}$ of $W_\alpha(f)$ for $f \in \mathcal{F}$ satisfies:

$$\text{(a)}\quad \bar{f} = \frac{\left[(1-\alpha)-\lambda\right]\left(\frac{p_2}{p_1}\right)+\lambda}{(1-\alpha)\left(\frac{p_2}{p_1}\right)+\alpha} = \frac{[1-\alpha-\lambda]p_2 + \lambda p_1}{\alpha p_1 + (1-\alpha) p_2}$$

where (b) $\lambda = \dfrac{(1-\alpha)\int_Q \dfrac{p_1 p_2}{\alpha p_1+(1-\alpha)p_2}}{\int_Q \dfrac{(p_2-p_1)p_1}{\alpha p_1+(1-\alpha)p_2}} = -W_\alpha(\bar{f})$

(c) $\bar{f}$ is an optimal discriminant fuction

(d) For $g_i \in \mathcal{F}$, $W_\alpha(g_i) \to W_\alpha(\bar{f})$ implies $g_i \to \bar{f}$ in $L_2(Q)$

Proof That the forms in (a), (b) are necessary conditions for a minimum follows by a tedious calculation using elementary variational techniques. To avoid repeating the calculation we will prove the theorem directly from the given formulas (a), (b):

$$\text{First} \quad \int_Q \bar{f} p_1 = \int_Q \frac{(1-\alpha)p_2 p_1 - \lambda(p_2-p_1)p_1}{\alpha p_1 + (1-\alpha)p_2}$$

which is 0 iff λ is given by the first half of (b).

$$\text{Also} \quad \int_Q \bar{f} p_2 = \int_Q \bar{f}\left(\frac{\alpha p_1 + (1-\alpha)p_2}{1-\alpha}\right) = \int_Q \frac{(1-\alpha-\lambda)p_2+p_1\lambda}{(1-\alpha)}$$

which clearly is unity. By (a) and (2')

$$W_\alpha(\bar{f}) = \int_Q \bar{f}\left[(1-\alpha-\lambda)p_2+\lambda p_1\right] - (1-\alpha) = -\lambda$$

which verifies the second half of (b).

Next $\bar{f}$ is rational and increasing in the quantity (p_2/p_1). Hence it is an optimal discriminant function (for either the Bayes of Neyman-Pearson problems considered).

Finally the rest of the theorem holds by the following argument: If $g \in \mathcal{F}$ then

$$\int_Q (g-\bar{f})^2((1-\alpha)p_2+\alpha p_1) = \int_Q (g^2+\bar{f}^2)((1-\alpha)p_2+\alpha p_1)$$

$$-2\int_Q g\bar{f}((1-\alpha)p_2+\alpha p_1) = W_\alpha(g)+W_\alpha(\bar{f})$$

$$-2\int_Q g((1-\alpha-\lambda)p_2+p_1) + 2(1-\alpha) =$$

$$W_\alpha(g) + W_\alpha(\bar{f}) - 2W_\alpha(\bar{f}) = W_\alpha(g) - W_\alpha(\bar{f})$$

By the regularity assumption this yields

$$(3) \quad W_\alpha(g) - W_\alpha(\bar{f}) > \alpha\kappa \int_Q (g-\bar{f})^2 q$$

Clearly the above reduction of the search for an optimal discriminant function to that of minimizing $W_\alpha(f)$ has many advantages. One not so obvious advantage is that $W_\alpha(\bar{f})$ has the very beautiful non-parametric consistent estimator described below: Set the quantity

$$\epsilon_\infty = 2\alpha(1-\alpha) \int_Q \frac{p_1\, p_2}{\alpha p_1 + (1-\alpha) p_2} .$$

Note that the denominator of the middle term in (b) may be written as

$$\int_Q \frac{\left(p_2 - \left(\frac{\alpha p_1 + (1-\alpha) p_2}{\alpha} - \frac{1-\alpha}{\alpha} p_2\right)\right) p_1}{\alpha p_1 + (1-\alpha) p_2}$$

which reduces to

$$\left(\frac{1}{2\alpha(1-\alpha)} + \frac{1}{2\alpha^2}\right) \epsilon_\infty(\alpha) - 1/\alpha$$

Combining with the rest of (b) yields

$$(4) \quad W_\alpha(\bar{f}) = \frac{\epsilon_\infty(\alpha)}{2 - \left(\frac{1}{1-\alpha} + \frac{1}{\alpha}\right) \epsilon_\infty(\alpha)}$$

Now $\epsilon_\infty(\alpha)$, known as the limiting nearest neighbor error rate, has the well known consistent non-parametric estimator $\hat{\epsilon}_N(\alpha)$ described in Theorem 2 (see [1], [6]), due to Cover and Hart in the case of continuous densities p_1, p_2. For completeness we prove the case where the densities are only assumed to be measurable and bounded.

Theorem 2 For the classification problem described at the outset suppose $P_1 = \alpha$. According to the rules of the classification experiment let N sample points be generated independently. Call the sample $X_1, X_2, \ldots X_N$ and suppose their correct classes are known. Then classify an N + 1st independent sample point X, as the class of its nearest neighbor (WRT Euclidean distance or another suitable distance generating the standard topology of R^d) in $\{X_1, X_2, \ldots X_N\}$. Call the expected error of such a procedure $\hat{\epsilon}_N(\alpha)$. Then $\hat{\epsilon}_N(\alpha) \xrightarrow[N\to\infty]{} \epsilon_\infty(\alpha)$ in probability. The consistent estimator of $W_\alpha(\bar{f})$ is then given by

$$(4') \quad \hat{W}_\alpha = \frac{\hat{\epsilon}_N(\alpha)}{2 - \left(\frac{1}{1-\alpha} + \frac{1}{\alpha}\right) \hat{\epsilon}_N(\alpha)}$$

Proof: For the moment let X be fixed and let r_N be the distance from X to the second nearest neighbor in $\{X_1, X_2, \ldots X_N\}$. Clearly $r_N \to 0$ in probability. Now we write the error of classifying X by the nearest neighbor rule given r_N as

$$\epsilon(X \mid r_N) = \int_{d(X,Y) < r_N} \left[\Pr(1 \mid X)\Pr(2 \mid Y) + \Pr(2 \mid X)\Pr(1 \mid Y)\right] d\mu$$

where μ is the conditional distribution of the first nearest neighbor Y, given r_N = distance to second meighbor. This is clearly given by

$$d\mu = \frac{\alpha p_1 + (1-\alpha) p_2}{\int_{d(X,\tilde{y}) < r_N} (\alpha p_1 + (1-\alpha) p_2)\, d\tilde{y}} \, dy$$

Now by the Lebesque Density Theorem a.e. X is a point of density one for the functions

$$Pr(i|x) = \frac{P_i p_i(x)}{P_1 p_1(x) + P_2 p_2(x)}$$

i.e. $Pr(i|y)$ is arbitrarily close to $Pr(i|x)$ for an arbitrarily high percentage of the set $\{y : d(x,y) < r_N\}$ as r_N gets arbitrarily small.

It is now straightforward that

$$\hat{\epsilon}_N(\alpha) = E_X(E_{r_N}(E(X|r_N))) \xrightarrow[N\to\infty]{}$$

$$\int_Q \frac{2P_1P_2p_1p_2}{P_1p_1 + P_2p_2} = \int_Q \frac{2\alpha(1-\alpha)p_1p_2}{\alpha p_1 + (1-\alpha)p_2} = \epsilon_\infty(\alpha)$$

where E_X and E_{r_N} are expectations over X and r_N (considered as random variables determined by our classification experiment).

Actually, in a particular application, only an estimate of $\hat{\epsilon}_N(\alpha)$, based on the data $X_1, X_2, \ldots X_N$, can be given. (We shall still denote such an estimate by $\hat{\epsilon}_N(\alpha)$ for notational convenience.) Usually the L-method of estimation is employed - classify each X_i according to the class of its nearest neighbor in $\{X_j\}_{j\neq i}$. Use the error percentage for the N samples as $\hat{\epsilon}_N(\alpha)$. By straightforwardly but tediously amending the preceding proof, this can be proven to be a consistent extimate of $\epsilon_\infty(\alpha)$. Furthermore the error in estimating $\epsilon_\infty(\alpha)$ by the above techniques can be reduced by choosing a proper distance measure for the data set. This has been successfully demonstrated by several authors. (See for example [11].)

treat this problem here but will use the above naive "L" estimate for our numerical experiment in V. The performance sensitivity to this estimate will then be examined by classifying an independent data set.

III Series Expansions for Minimum Within-Class Scatter and a General Consistency Property

Suppose we generate N samples according to the rules of our classification experiment with the correct class known for each sample. Let $\varphi_1, \varphi_2, \ldots$ be a complete set of limearly independent functions spanning $L_2(Q)$. Since we will be solving a linear problem on spans of the form $\langle \varphi_1, \varphi_2, \ldots \varphi_M \rangle$, we will assume W.L.O.G. that $\varphi_1 \equiv 1, \varphi_2, \varphi_3, \ldots$ is an orthonormal basis for $L_2(Q)$. (This can be accomplished by adding the unity function and applying the usual Gramm-Schmidt procedure with weighting function $q(x)$.) Let μ_N, ν_N be the empirical densities* determined by the class 1 and class 2 samples respectively. Since we will let N get arbitrarily large assume there are samples present from each class. Now consider the variational problem

(5) minimize $J_M(f) = \alpha \operatorname{Var}_{\mu_N} f + (1-\alpha) \operatorname{Var}_{\nu_N} f$

subject to the conditions $f = \sum_1^M a_i \varphi_i$

$$E_{\mu_N} f = 0$$

$$E_{\nu_N} f = 1$$

The optimal coefficients a_i can be found straightforwardly by the method of Lagrange multipliers. The solution is:

$$(a_1, a_2, \ldots a_M)^t = \frac{(V_1^t K^{-1} V_2) K^{-1} V_1 - (V_1^t K^{-1} V_1) K^{-1} V_2}{(V_1^t K^{-1} V_2)(V_2^t K^{-1} V_1) - (V_2^t K^{-1} V_2)(V_1^t K^{-1} V_1)}$$

where V_1, V_2 are the class 1, 2 sample mean vectors of $\Phi = (\varphi_1(x), \varphi_2(x), \ldots \varphi_M(x))^t$ and K is the weighted sum of the class 1,2 sample correlation matrices for Φ with weights $\alpha, (1-\alpha)$ respectively.

Our estimate, f_N, of $\bar{f}$ is then obtained by specifying M: First

*These are just averages of Dirac delta functions for the sample points.

notice that the minimum scatter in (5) is decreasing as a function of M. In fact it decreases to zero with probability one. (given N fixed and at least one sample point from each class). This can be shown by approximating the indicator function of a set of intervals, containing the class 2 sample points but not the class 1 sample points, using a finite linear combination, h, of $\varphi_1, \varphi_2, \ldots$ and considering a suitable $Ah + B$ as f.

Second we may restrict M by the regularity condition. Recall $p_1 > \kappa q$. This provides the motivation for the regularization constraint on the domain of J_M:

$$\text{Var}_{\mu_N} g > \kappa \int_Q g^2 q$$

$$\text{for } g \in \langle \varphi_1, \varphi_2, \ldots \varphi_M \rangle$$

$$\text{S.T.} \quad E_{\mu_N} g = 0$$

$$E_{\nu_N} g = 1$$

With probability one, this constraint restricts our choice of M to $1 \leq M \leq \bar{M}$. This is demonstrated by a simple Fourier analytic construction similar to that in the next to last paragraph.

Finally M is chosen to minimize

$$(6) \qquad \left| \min J_M - \frac{\hat{\epsilon}_N(\alpha)}{2 - \left(\frac{1}{1-\alpha} + \frac{1}{\alpha}\right)\hat{\epsilon}_N(\alpha)} \right| \qquad M = 1, 2, \ldots \bar{M}$$

Before proving the consistency of this procedure we give a simple algorithm for finding f_N:

Using (5′) compute successively $\min J_M$, checking that the regularity constraint is satisfied before going to M+1. The procedure terminates when either $\min J_M < \dfrac{\hat{\epsilon}_N(\alpha)}{2 - \left(\frac{1}{1-\alpha} + \frac{1}{\alpha}\right)\hat{\epsilon}_N(\alpha)}$ or the regulari-

*The regularization parameter κ is usually determined by some non-statistical reasoning. This choice may be critical in small sample problems.

zation constraint is not satisfied. If termination coincides with a regularity violation, use $\bar{f}$ estimate for M-1. Otherwise at termination use M or M-1, whichever gives scatter closer to nearest neighbor scatter estimate. Note that the regularization may be checked by minimizing

$$\mathrm{VAR}_{\mu_N} \frac{\left(\sum_1^M a_i \varphi_i\right)}{\sum_1^M a_i^2}$$

subject to $E_{\mu_N} g = 0$, $E_{\nu_N} g = 1$. This can be solved by minimizing first $\mathrm{VAR}_{\mu_N}\left(\sum_1^M a_i \varphi_i\right)$ subject to $E_{\mu_N} g = 0$, $E_{\nu_N} g = 1$, $\sum_1^M a_i^2 = \eta$ using linear algebra and then searching over η. The regularization constraint serves to prevent spurious over-fitting of the data to the given basis. We prove our main result now.

Theorem 3 For any basis $\varphi_1 \equiv 1, \varphi_2, \ldots$ of $L_2(Q)$, $f_N \xrightarrow[L_2(Q)]{} \bar{f}$ in probability.

Proof: Consider the subspace of $L_2(Q)$ given by $L_3 = \left\{f: \int f p_1 = 0\right\} \cap L_2\langle \alpha p_1 + (1-\alpha)p_2\rangle$, where $L_2\langle p\rangle$ denotes the space of square integrable functions WRT. a measure whose density is p. L_3 is normed by $\|(\ )\|_3 = \sqrt{\int (\ )^2 (\alpha p_1 + (1-\alpha)p_2)}$. Use $\|\ \|_2$ to denote the assumed norm for $L_2(Q)$. We may construct a sequence $\varphi_2', \varphi_3', \ldots$, which is linearly independent and dense in L_3, by simply adding appropriate const a nts to $\varphi_2, \varphi_3, \ldots$

Now form a complete orthonormal basis $\xi_2, \xi_3, \ldots$ of L_3 where each ξ_i is a linear combination of $\varphi_2', \varphi_3', \ldots$ and hence a linear combination of $\varphi_1, \varphi_2, \ldots$. Let $c_i = \int \xi_i p_2$. Then $\bar{f} = \sum_2^\infty b_i \xi_i$ where b_i is the solution of $\min \sum_2^\infty b_i^2$ s.t. $\sum_2^\infty c_i b_i = 1$. This is just given by $b_i = c_i / (\sum_2^\infty c_i^2)^2$. Notice $\bar{w} = (\sum_2^\infty c_i^2)^{-1}$ in the minimum scatter in (2').

By mimicking the same sequence of steps for $L_4 = \{f: \int f d\mu_N = 0\} \cap L_2\langle \alpha\mu_N + (1-\alpha)\nu_N\rangle$ with norm $\|(\ )\|_4 = \sqrt{\int (\ )^2 (\alpha\mu_N + (1-\alpha)\nu_N)}$, we construct an othonormal basis $\eta_2^N, \eta_3^N, \ldots \eta_R^N$ for $L_4 \cap \langle \varphi_1, \ldots \varphi_M\rangle$. (Note R may be less than M since some of the φ_i's may be linearly dependent in L_4.) Let $d_i^N = \int \eta_i^N \nu_N$. The solution of (5) is $f_N = \sum_2^R d_i^N \eta_i^N / \sum_2^R (d_i^N)^2$ with $W_N = (\sum_2^R (d_i^N)^2)^{-1}$ the approximately optimal scatter. Also by the construction (if we let $N \to \infty$ with M and R varying appropriately)

$\sum_2^R (d_i^N)^2 \to \sum_2^\infty c_i^2$ in probability and $d_i^N \to c_i$, $\eta_i^N \xrightarrow{L_4(\alpha)} \xi_i$ in probability for each i.

Now consider the simple inequality

$$\|f_N - \bar{f}\|_2 \le \left\| \sum_2^{\ell} \left(d_i^N \eta_i^N W_N^2 \quad c_i \xi_i \bar{W}^2 \right) \right\|_2 + \bar{W}^2 \left\| \sum_{\ell}^{\infty} c_i \xi_i \right\|_2 + W_N^2 \left\| \sum_{\ell}^{R} d_i^N \eta_i^N \right\|_2 .$$

Since in general $\| \ \|_2 < (\alpha K)^{-1/2} \| \ \|_3$ by the regularity assumption, the second term of the right hand side can be made arbitrarily small by choosing ℓ sufficiently large. Having fixed ℓ the first term will be small with probability close to one by taking N sufficiently large. Finally by the regularization constraint

$\left\| \sum_{\ell}^{R} d_i^N \eta_i^N \right\|_2 \le (\alpha K)^{-1/2} \left\| \sum_{\ell}^{R} d_i^N \eta_i^N \right\|_4 = (\alpha K)^{-1/2} \sum_{\ell}^{R} (d_i^N)^2$ which is small with probability close to one since $\sum_{\ell}^{R} (d_i^N)^2 \to \sum_{\ell}^{\infty} c_i^2$ in probability.

Hence $\|f_N - \bar{f}\|_2$ approaches zero in probability. This completes the proof of the theorem.

The same proofyields the following.

Corollary: The consistency remains valid if we amend our algorithm as follows: First remove sequentially any φ_i from the basis sequence whic

causes the regularization constraint to be violated for $\langle \varphi_1'', \varphi_2'', \ldots \varphi_a'', \varphi_i \rangle$ where $\varphi_1'', \varphi_2'', \ldots \varphi_a''$ are the previously unremoved φ_i's in $\{ \varphi_1, \varphi_2, \ldots \varphi_{i-1} \}$. Then apply the minimum scatter procedure to initi spans of the remaining orthonormal sequence, stopping when neares neighbor scatter exceeds the current estimate or when $M = N$ (which is only theoretically necessary to avoid some degenerate situations).

If one is using a fixed basis it is recommended that the procedure of the corollary be implemented. Of course one would simultaneously improve estimates of $\bar{f}$ while removing basis functions in a computer program. Although for practical small sample problems some regularization seems necessary to prevent coincidental fits of highly oscillatory basis functions, it is not clear that is necessary for consistency. The author believes there exist counterexamples but has none at this writing.

IV Projection Pursuit Method for Constructing Series Expansions

Standard multidimensional orthonormal bases involve products of univariate basis functions. Since the number of such products grows exponentially with the number of univariate possibilities, solving the minimum scatter problem (5) using these bases is infeasible with today's computers. Since our consistency result holds for any basis it seems natural to construct the basis functions directly from the data. Using the principle of projection pursuit (for background and applications to regression and density estimation see [7], [8], [9], [10]), we give an algorithm which simultaneously generates the basis functions and solves the associated optimal discriminant problem. We treat only the case of Q the unit cube in R^d centered at the origin, and $q(X)$ the uniform density on Q. Other cases may be treated similarly. The algorithm is described as follows:

Let $\phi_0 \equiv 1, \phi_1, \phi_2, \phi_3, \ldots$ be linearly independent and dense in $L_2(-\sqrt{d}/2, +\sqrt{d}/2)$. These are the univariate approximating functions and should be chosen appropriately according to the particular application. Let $T_2, T_3, \ldots$ be a non-decreasing sequence of integers converging to ∞. T_M is the number of ϕ functions used at the Mth stage and again should be judiciously chosen for a given application. Now apply the indicated steps.

[a] Initialize -- $\varphi_1 \equiv 1$, $M = 2$

[b] Minimize over a_i, and C of norm 1 in R^d

$$J_M(f) = \alpha \mathrm{VAR}_{\mu_N} f(C^t X) + (1-\alpha)\mathrm{VAR}_{\nu_N} f(C^t X)$$

subject to conditions

$$f = \sum_1^{M-1} a_i \varphi_i(X) + \sum_1^{T_M} a_{i+M-1} \phi_i(C^t X)$$

$$E_{\mu_N} f = 0$$

$$E_{\nu_N} f = 1$$

[c] For optimal a_i, C apply Gram-Schmidt procedure to $\sum_1^{T_M} a_{i+M-1} \Psi_i(C^t X)$ and $\varphi_1(X), \varphi_2(X), \ldots \varphi_{M-1}(X)$ in $L_2(Q)$ to generate φ_M.

[d] Set M = M + 1.

[e] Stop if regularization is not staisfied or nearest neighbor scatter estimate exceeds current estimate. Use $\bar{f}$ estimate for M or M - 1, accordingly.

[f] Return to [b].

Proof of Consistency (Convergence in Probability)

We establish that this procedure is consistent by first showing convergence of the non-sampling form of the above (that is, at each stage M, we know the first two moments of $1, \varphi_2, \ldots \varphi_M, \Psi_1(C^t X), \ldots, \Psi_{T_M}(C^t X)$ for any C and can therefore solve [b] with the actual densities): Using the φ_M's generated construct the sequence as in the proof of theorem 3 - $\xi_2, \xi_3, \ldots$. If this spans L_3 then f_M converges to $\bar{f}$ in $L_2(Q)$ trivially so we assume this is not the case. Now construct a sequence of the form $h_i = A_i \Psi_i(C_i^t X) + B_i - \sum_2^\infty s_\ell^i \xi_\ell$ which spans $\langle \xi_2, \xi_3, \ldots \rangle^\perp \cap \{f : \int f p_1 = 0\}$. If these all have zero integral wrt. p_2 then again $f_M \to \bar{f}$ in $L_2(Q)$ (by orthonormalizing these and solving the minimum scatter problem directly as in the proof of theorem 3). Otherwise we find h_i with $\int h_i \rho_1 > 0$. But then we get smaller scatter than with $\lim f_N$ by solving the problem on

$\langle h_i, \xi_2, \xi_3, \ldots \rangle$ which means for some (very large) M we get smaller scatter on $\langle \xi_2, \xi_3, \ldots \xi_M, A_i \psi_i(C_i^t X) + B_i - \sum_{2}^{M} S_\ell^i \xi_\ell \rangle$ than for lim f_M. This is a contradiction. Hence $f_M \to \bar{f}$ in $L_2(Q)$.

Now for the convergence in probability: If not, we can find a sequence of samples of increasing sizes N_i and associated increasing iterates M_i with the properties:

a. $\|f_{M_i} - \bar{f}\|_2 > \epsilon > 0$

b. sample (i) cov $(1, \psi_{T_2}(C_2^t X), \ldots, \psi_{T_M}(C_M^t X))$

$\to$ cov $(1, \psi_{T_2}(C_2^t X), \ldots, \psi_{T_M}(C_M^t X))$

for any M; $C_2, C_3, \ldots C_M$ of norm 1.

(Note we needed lots of subsequence taking for this. Also this ensures no early violation of the regularization constraint (by a compactness argument) so that M_i is increasing.) By taking further subsequences we can get an orthonormal sequence 1, $\bar{\varphi}_2, \bar{\varphi}_3, \ldots$ such that the Mth basis functions in the ith sample converge to $\bar{\varphi}_M$. Standard arguments imply that 1, $\bar{\varphi}_2, \bar{\varphi}_3, \ldots$ result from a non-sampling application of our algorithm. But then for some (large $\bar{M}$) we have $\|f_{\bar{M}} - \bar{f}\|_2 < \epsilon/3$ for an infinity of samples and also $\|f_{M_i} - f_{\bar{M}}\|_2 < \epsilon/3$ for these samples by an easy application of the regularity constraint as in the proof of theorem 3. This is a contradiction.

V. An Application to Optical Detection of a Randomly Moving Point Source

A. Description of Experiment

Intensity measurements from a photodetector were modeled: Pixel radiances were simulated for two 4X4 square Pixel arrays assuming a small time gap between arrays. See figure 1. A difference frame was formed with entries consisting of differences of corresponding Pixel radiances. This was X. For a set of background scences of interest, which drifted .4 Pixel in ΔT secs., X was $\approx$ uncorrelated white noise with mean 0 and standard deviation 32 (in grey levels). Hence we used this distribution to generate class 1 (background) sample points.

Class 2 (target plus background) was simulated as follows: At T_0 a point source (+) of intensity 256 was generated with a uniform distribution in the shaded region of size one Pixel. See figure 2. The corresponding radiances were then calculated using Gaussian blurring with a blur circle of radius that of a Pixel width. At $T_0 + \Delta T$ the target was moved one Pixel width in a random direction with a uniform distribution in angle. Gaussian blurring as above was used to generate the radiances and then the difference frame was generated as in the background case. Class 2 samples were generated by adding an independently generated background sample point. For the data set used in the numerical prodecure the spatial standard deviation was 31 for targets before the background addition. The corresponding mean was on the order of -1/2 grey level.

A total of two thousand sample points were generated according to the

rules of section II with $P_1 = P_2 = 1/2$. Although the target class had a simple and intuitive stochastic construction it seems extremely hard to obtain a numerically feasible form for its density. Hence we applied our non-parametric analysis.

B. Numerical Procedure and Results

The data was normalized via an affine transformation so that the combined class samples had zero mean vector and whitened covariance with 1/2 correspoinding to 2.33 standard deviations for any projection. With q uniform and Q the unit cube centered at the origin, we set $\kappa = .01$. Although the background was normally generated we proceeded without this knowledge and "$\kappa = .01$" corresponded roughly to the statement "all background possibilities (in Q) are equally likely more than 1% of the time." We set $\alpha = 1/2$ and, because most of the pooled data was inside a ball of radius $\frac{1}{2}$, we used

$$\psi_i(y) = \begin{cases} \cos\left[\pi i\left(y+\frac{1}{2}\right)\right] & -\frac{1}{2} \leq y \leq \frac{1}{2} \\ 1 & y < -\frac{1}{2} \\ (-1)^i & y > \frac{1}{2} \end{cases}$$

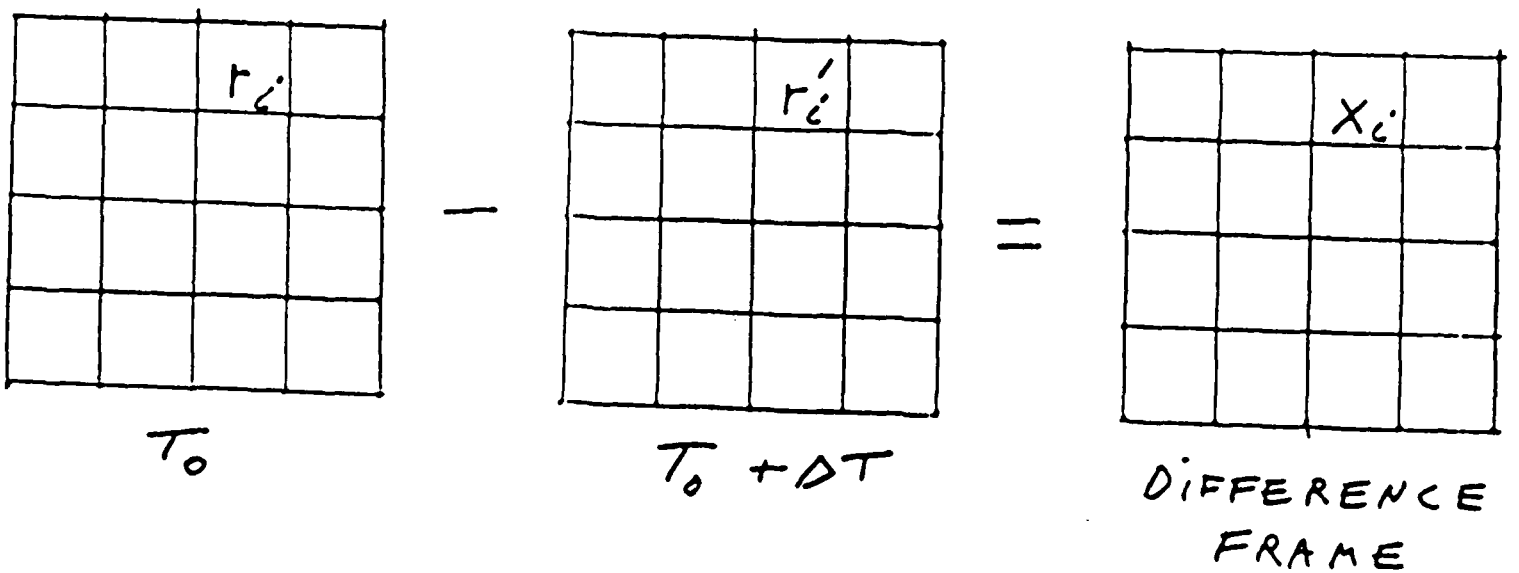
for $i = 1, 2, 3, \ldots 15$. Because the sample size was large (for univariate estimation) we set $T_2 = T_3 = \ldots = T_9 = 15$.

Using the "L" method we found $\hat{\epsilon}_{2000}(1/2) = 17.1\%$ corresponding to a minimum scatter estimate of .13. The algorithm was then run on a VAX 11780 at the Naval Research Laboratory. Stages M = 2, 3, ... 9 were first performed without orthonormalization. The regularization constraint was then checked for $\varphi_1, \varphi_2, \ldots \varphi_9$ and found not violated. Minimization was done using ZX Min of IMSL. Stage M = 5 yielded the estimate f_N of section III. The error rates for each M were estimated by classifying (using threshold t = .5) independent data consisting of 2000 sample points. For this experiment there seems to be relatively little sensitivity to $\hat{\epsilon}_N(\alpha)$ (provided it lies in [13%, 20%]). The results are summarized in Figure 3. Hopefully further

research may yields feasible resampling techniques for better estimating both $\epsilon_{\infty}(\alpha)$ and M.

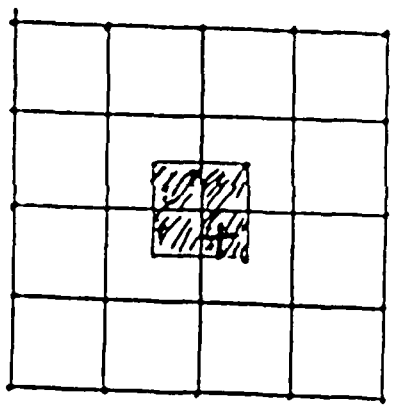
The author is indebted to Dr. Thomas Flick of the Naval Research Laboratory for adapting his program "Projection Pursuit Regression Using Fourier Series" to this experiment.

FIGURE 1

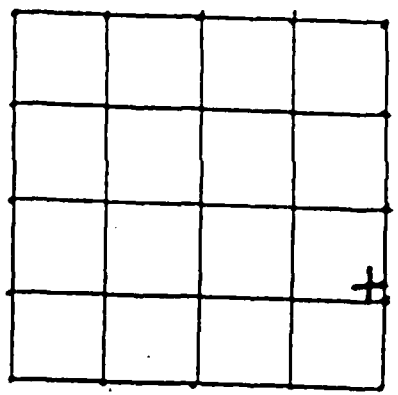


$$r_i - r_i' = X_i$$

FIGURE 2

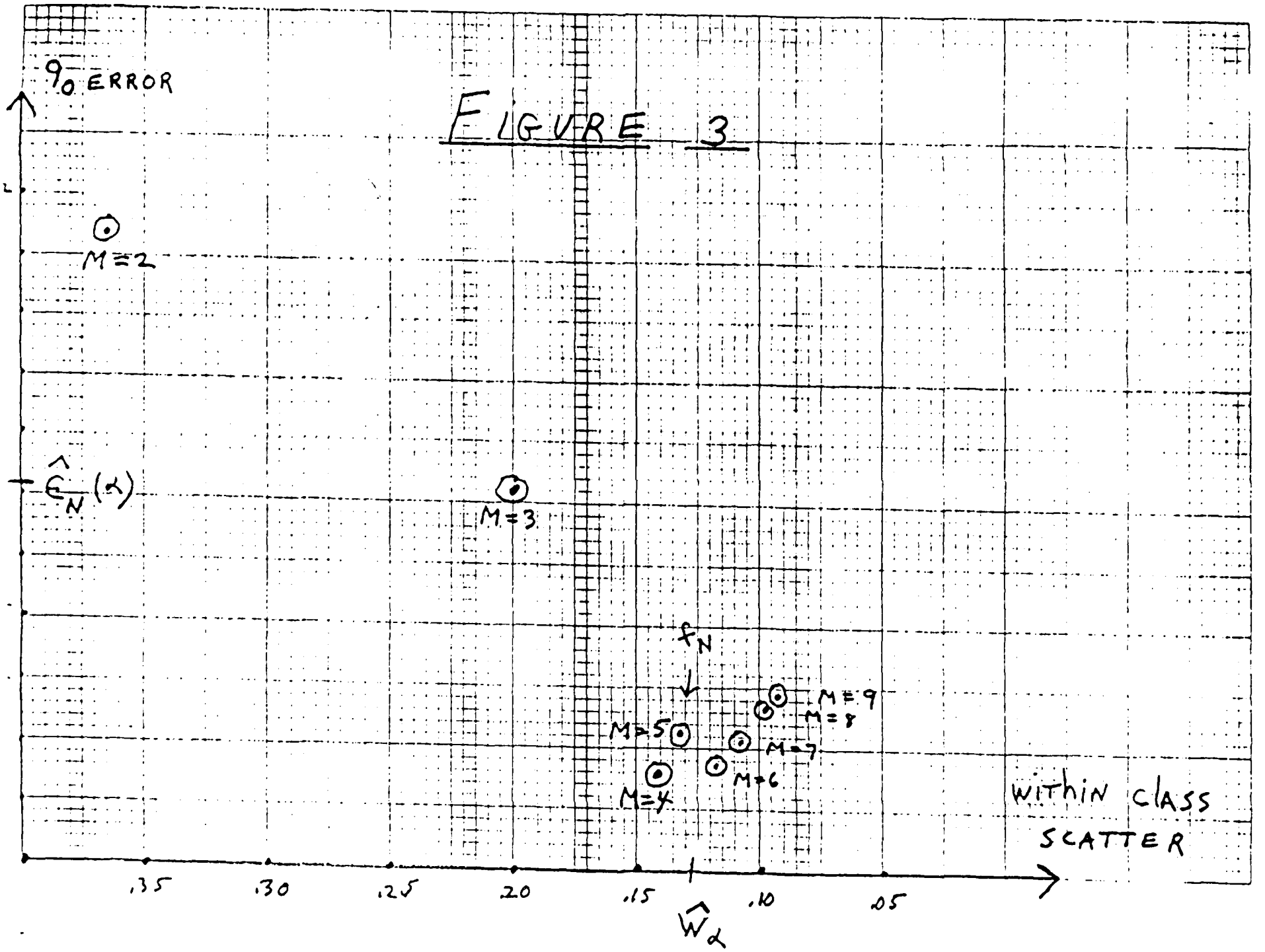


T_0



$T_0 + \Delta T$

FIGURE 3


% ERROR
M=2
$\hat{\epsilon}_N(\alpha)$
M=3
f_N
M=9
M=8
M=5
M=7
M=6
M=4
WITHIN CLASS SCATTER
.35
.30
.25
.20
.15
.10
.05
$\hat{W}_\alpha$

REFERENCES

1. Fukunaga, K. Introduction to Statistical Pattern Recognition, Academic Press, New York, 1972.

2. Friedman, H.P. and Rubin, J., "On Some Invariant Criteria for Grouping Data," American Statistical Association Journal, Vol. 62, 1962, pp. 1159-1178.

3. Peterson, D.W. and Mattson R.L., "A Method of Finding Linear Discriminants for a Class of Performance Criteria," I.E.E.E. Trans. on Inf. Th., IT-12, 1966, pp. 380-387.

4. Fukunaga, K. and Ando, S. "The Optimum Nonlinear Features for a Scatter Criterion in Discriminant Analysis," I.E.E.E. Trans. on Inf. Th., July 1977, Vol. IT-23, No. 4, pp. 453-459.

5. Otso, N. "An Optimal Nonlinear Transformation Based on a Variance Criterion for Pattern Recognition"-I, Bull. of the Electrotechnical Lab., Japan, Vol. 36, 1972, pp. 815-830.

6. Cover, T.M. and Hart, P.E. "Nearest Neighbor Pattern Classification," I.E.E.E. Trans. on Inf. Th., Jan. 1967, Vol. IT-13, pp. 27-31.

7. Huber, P.J. "Projection Pursuit," Harvard Univ. Research Report, Dept. of Statistics, August 1981.

8. Huber, P.J. "Density Estimation and Projection Pursuit Methods", Harvard Univ. Research Report, Dept. of Statistics, Sept. 1981.

9. Friedman, J.H. and Stuetzle, W. "Projection Pursuit Regression", J. of Amer. Stat. Assoc., Dec. 1981, Vol. 76, No. 376, pp. 817-823.

10. Friedman, J.H., Stuetzle W. and Schroeder, A. "Projection Pursuit Density Estimation", J. of Amer. Stat. Assoc., Sept. 1984, Vol. 79, No. 387, pp. 599-608.

11. Fukunaga, K. and Flick, T.E. "An Optimal Global Nearest Neighbor Metric", I.E.E.E. Trans. on P.A.M.I., Vol. PAMI-6, No. 3, May 1984, pp. 314-318.

END

4-87

DTIC