Technical Note

1974-59

C. W. Therrien

A Generalized Approach to Linear Methods of Feature Extraction

30 December 1974

Prepared for the Ballistic Missile Defense Program Office,
Department of the Army,
under Electronic Systems Division Contract F19628-73-C-0002 by

Lincoln Laboratory

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

LEXINGTON, MASSACHUSETTS

The work reported in this document was performed at Lincoln Laboratory, a center for research operated by Massachusetts Institute of Technology. This program is sponsored by the Ballistic Missile Defense Program Office, Department of the Army; it is supported by the Ballistic Missile Defense Advanced Technology Center under Air Force Contract F19628-73-C-0002.

This technical report has been reviewed and is approved for publication.

FOR THE COMMANDER

Eugene C. Raabe, Lt. Col., USAF
Chief, ESD Lincoln Laboratory Project Office

MASSACHUSETTS INSTITUTE OF TECHNOLOGY
LINCOLN LABORATORY

A GENERALIZED APPROACH TO LINEAR METHODS OF FEATURE EXTRACTION

C. W. THERRIEN
Group 92

TECHNICAL NOTE 1974-59

30 DECEMBER 1974

LEXINGTON MASSACHUSETTS

ABSTRACT

An approach to feature extraction based on functions of the class correlation matrices is described. If linear functions of the correlation matrices are chosen, the present method extends the methods of feature extraction proposed by Fukunaga and Koontz. If certain types of non-linear functions are employed, the method reduces to the orthogonal subspace method of Watanabe and Pakvasa.

Optimization of selected features through selection of appropriate functions is discussed briefly. Preliminary results of classification of radar signatures using the feature extraction methods described here are presented.

TABLE OF CONTENTS

I. INTRODUCTION

The goal of feature extraction in pattern recognition is to reduce the dimensionality of the space in which classes of data are represented without greatly reducing the separability of the classes. An approach to linear methods of feature extraction is described which is based on applying certain functions to the correlation matrices of the classes to be separated. This approach to feature extraction was motivated by experience with two other methods–that of Fukunaga and Koontz[1] and that of Watanabe and Pakvasa.[2] The present report shows a relation between these two methods, and provides a natural extension of the Fukunaga-Koontz method to the multiclass case. In addition, the present formulation provides enough flexibility to in principle optimize class separability in a very general way. This point is discussed in the report.

II. FORMULATION OF LINEAR FEATURE EXTRACTION

Consider the problem of generating features to classify patterns into one of K distinct classes. The patterns are originally represented by vectors $\underline{x}$ in an n-dimensional linear vector space (the "observation space"). The correlation matrix for each class is defined by

$$R_k = E_k\left[\underline{x}\underline{x}^T\right] = \int \underline{x}\underline{x}^T p_k(\underline{x})\, d\underline{x} \qquad k = 1, 2, \ldots, K \qquad (1)$$

where E_k denotes expectation carried out using the probability density p_k of class k. It is assumed that the correlation matrices satisfy the condition

$$\| R_k \| \equiv \lambda_{max} \leq 1; \qquad k = 1, 2, \ldots, K \qquad (2)$$

where λ_{max} is the largest eigenvalue of R_k. (This results in no loss of generality since (2) can always be achieved by a linear scaling of the observation space.) Thus, since the correlation matrix is positive definite, all of the eigenvalues of R_k lie between 0 and 1.

In order to motivate the general approach, let $\left\{ \underline{u}_j,\ j = 1,\ 2,\ \ldots,\ n \right\}$ be an orthonormal basis for the observation space. Further, let $\underline{x}$ be any random vector and let $\hat{\underline{x}}$ be a truncated expansion of $\underline{x}$ using $m < n$ of the $\underline{u}_j$.

$$\hat{\underline{x}} = \sum_{j=1}^{m} b_j \underline{u}_j \qquad ; \quad b_j = \underline{u}_j^T \underline{x} \tag{3}$$

A suitable set of features for the i^{th} class would result if one could choose the $\underline{u}_j$ such that the mean-square error

$$E_i\left[\left| \underline{x} - \hat{\underline{x}} \right|^2 \right] = \int \left| \underline{x} - \hat{\underline{x}} \right|^2 p_i(\underline{x})\, d\underline{x} \tag{4}$$

is minimum and simultaneously the mean-square error

$$E_k\left[\left| \underline{x} - \hat{\underline{x}} \right|^2 \right] = \int \left| \underline{x} - \hat{\underline{x}} \right|^2 p_k(\underline{x})\, d\underline{x} \qquad \begin{matrix} k = 1,\ 2,\ \ldots,\ K \\ k \neq i \end{matrix} \tag{5}$$

is maximum. Minimizing (4) without conditions (5) leads to the well-known Karhunen-Loève expansion[1] an optimum representation of a vector of class i with m terms. The additional conditions (5) however, if satisfied, would insure that the basis chosen to optimally represent a vector as a member of class i would simultaneously be non-optimal for representing it as a member of the other classes.

Since it is usually not possible to minimize (4) and maximize (5) simultaneously, a related criterion will be derived. This leads to a generalization of the Karhunen-Loève expansion that applies to problems where class separability must be preserved.

Note first that the mean-square error in representation can be expressed as*

$$E_k\left[\left| \underline{x} - \hat{\underline{x}} \right|^2 \right] = \sum_{j=m+1}^{n} \underline{u}_j^T R_k \underline{u}_j \qquad k = 1,\ 2,\ \ldots,\ K \tag{6}$$

---

* Although this result is well known, a proof is given in the Appendix for convenience.

Then by virtue of (2) and the positive definite property of R_k, (6) is bounded by

$$n - m \geq E_k\left[\left| \underline{x} - \hat{\underline{x}} \right|^2 \right] \geq 0 \qquad k = 1, 2, \ldots, K \quad (7)$$

As a result, maximizing (5) for $k \neq i$ is equivalent to minimizing

$$(n - m) - E_k\left[\left| \underline{x} - \hat{\underline{x}} \right|^2 \right] = \sum_{j=m+1}^{n} (\underline{u}_j^T \underline{u}_j - \underline{u}_j^T R_k \underline{u}_j) = \sum_{j=m+1}^{n} \underline{u}_j^T (I - R_k) \underline{u}_j \quad (8)$$

A single combined criterion is taken therefore as the sum of the criteria (4) and (8) normalized by K, the number of classes, that is

$$C_i = \frac{1}{K} \left\{ F_i\left[\left| \underline{x} - \hat{\underline{x}} \right|^2 \right] + \sum_{\substack{k=1 \\ k \neq i}}^{K} \left(n - m - E_k\left[\left| \underline{x} - \hat{\underline{x}} \right|^2 \right] \right) \right\} \quad (9)$$

where C_i is to be minimized. If (6) and (8) are substituted into (9) then C_i can be expressed as

$$C_i = \sum_{j=m+1}^{n} \underline{u}_i^T \hat{G}_i \underline{u}_i \quad (10)$$

where

$$\hat{G}_i = \frac{1}{K} \left[R_i + \sum_{\substack{k=i \\ k \neq i}}^{K} (I - R_k) \right] \quad (11)$$

The vectors $\underline{u}_j$ that minimize (10) are the eigenvectors of $\hat{G}_i$ corresponding to the n - m smallest eigenvalues.* Since (10) is to be minimized for any $m < n$, the optimal basis $\left\{ \underline{u}_j \right\}$ is the set of eigenvectors of $\hat{G}_i$, and the eigenvectors chosen to express $\hat{\underline{x}}$ should be those corresponding to the m largest eigenvalues of $\hat{G}_i$.

---

* Op. cit., p. 2.

Note that if $\underline{e}$ is a normalized eigenvector of $\hat{G}_i$, then the corresponding eigenvalue μ can be expressed as

$$\mu = \underline{e}^T \hat{G}_i \underline{e} = \frac{1}{K}\left[\underline{e}^T R_i \underline{e} + \sum_{\substack{k=1 \\ k\neq i}}^{K} (1 - \underline{e}^T R_k \underline{e}) \right] \tag{12}$$

Equation (2) and the positive definite property imply that each of the quadratic products in (12) has a value between 0 and 1. Thus μ lies between 0 and 1 and is close to 1 only if $\underline{e}^T R_i \underline{e}$ is close to 1 and all of the $\underline{e}^T R_k \underline{e}$ ($k \neq i$) are simultaneously close to 0. Thus the eigenvectors of $\hat{G}_i$ corresponding to eigenvalues near 1 relate to important distinguishing features of class i.

This approach can be generalized as follows. If A is a real symmetric matrix, then the matrix function f (A) for any scalar function f can be defined as

$$f(A) = V \begin{bmatrix} f(\lambda_1) & & & \\ & f(\lambda_2) & & \\ & & \ddots & \\ & & & f(\lambda_n) \end{bmatrix} V^T \tag{13}$$

where λ_j are the eigenvalues of A, and V is the orthonormal transformation that diagonalizes A.* Since the columns of V are the eigenvectors of A, the function f serves to "weight" the eigenvalues of A without changing its eigenvectors.

The foregoing concept can be applied to feature extraction. Define the matrices G_i and H_i for i = 1, 2, . . . , K by

---

* For purposes of this report (13) is taken to be the definition of a function of a symmetric matrix. This definition does not make any assumptions of analyticity on the function f which are required for the extension of the matrix function concept to more general matrices.

$$G_i = \frac{1}{K}\left[f_i(R_i) + \sum_{\substack{k=1 \\ k \neq i}}^{K} (I - f_k(R_k)) \right] \tag{14a}$$

$$H_i = h(G_i) \tag{14b}$$

where the functions $\{f_k\}$ and h are any functions mapping the interval [0, 1] into [0, 1]. We refer to the functions $\{f_k\}$ and h as the "preweighting" functions and the "postweighting" function, respectively. Features are defined in terms of the post-weighted matrices H_i by one of two methods:

Method 1 - Features are chosen as the projection of the data along selected eigenvectors of the matrices G_i. The post-weighting function h can serve to select the appropriate eigen-vectors, that is $h(\lambda)$ is 1 for a selected eigendirection and 0 for an eigendirection that is not selected. *

Method 2 - Features are defined by the relation

$$z_i = \underline{x}^T H_i \underline{x}$$

Each of the features z_i, $i = 1, 2, \ldots, K$ can be thought of as a weighted projection of the observation vector $\underline{x}$ into a subspace of the observation space.

III. THE TWO-CLASS CASE

For the special case of K = 2 the matrices G_1 and G_2 defined by (14a) satisfy the relation

$$G_2 = I - G_1 \tag{15}$$

* Since there is no guarantee that the selected eigenvectors from different G_i will be independent, it may be necessary to eliminate those eigenvectors that can be represented as linear combinations of the others.

Therefore G_2 and G_1 have identical eigenvectors and their eigenvalues are related by

$$\lambda_j^{(2)} = 1 - \lambda_j^{(1)} \qquad j = 1, 2, \ldots, n \quad (16)$$

Since the $\lambda_j^{(k)}$ all lie in the interval [0, 1], (16) shows that the eigenvectors of G_1 that provide the "most important" features for class 1 provide the "least important" features for class 2 and vice-versa. This is the principle upon which the method of Fukunaga and Koontz is based (see Section IV).

IV. LINEAR WEIGHTING

A simple form of pre-weighting function is a linear function

$$f_k(R_k) = a_k R_k \qquad (0 < a_k < 1) \qquad k = 1, 2, \ldots, K \quad (17)$$

When this form of preweighting is used in the two-class case, the results can be related to the Fukunaga-Koontz method of feature extraction.

For K = 2 (14a) becomes

$$\begin{aligned} G_1 &= 1/2\,(a_1 R_1 - a_2 R_2 + I) \\ G_2 &= 1/2\,(a_2 R_2 - a_1 R_1 + I) \end{aligned} \quad (18)$$

Fukunaga and Koontz first perform a linear transformation of the observation space which forces the correlation matrices R_1' and R_2' in the transformed space to satisfy

$$a_1 R_1' + a_2 R_2' = I \quad (19)$$

The transformed correlation matrices automatically satisfy (2). Under this condition (18) reduces to

$$\begin{aligned} G_1 &= a_1 R_1' \\ G_2 &= a_2 R_2' \end{aligned} \quad (20)$$

and the two methods become identical.

When the two classes have different mean vectors $\underline{m}_1$ and $\underline{m}_2$ but equal covariance matrices, and are weighted equally ($a_1 = a_2 = a$), (18) becomes

$$G_1 = 1/2 \left(a (\underline{m}_1 \underline{m}_1^T - \underline{m}_2 \underline{m}_2^T) + 1 \right)$$
$$G_2 = 1/2 \left(a (\underline{m}_2 \underline{m}_2^T - \underline{m}_1 \underline{m}_1^T) + 1 \right) \qquad (21)$$

Such matrices have only two eigenvalues that are not equal to 1/2, and only the corresponding two eigendirections contribute to the separation of the classes. [1]

Linear weighting is, of course, applicable to the multiclass case. Further, when the correlation matrices are transformed to satisfy the condition

$$\sum_{k=1}^{K} a_k R_k' = I \qquad (22)$$

then linear weighting becomes an extension of the Fukunaga-Koontz method. The matrices in (14a) assume the form

$$G_i = \frac{1}{K} \left[2 a_i R_i' + (K - 2) I \right] \qquad (23)$$

Although (16) has no direct analogy, the eigenvectors of G_i corresponding to eigenvalues that are close to 1 are "most important" for representing class i and simultaneously "least important" for representing the other classes. Therefore these eigenvectors can be expected to produce the best features if Method 1 is employed.

V. NONLINEAR WEIGHTING

When the nonlinear functions shown in Fig. 1 (unit step functions) are used for pre-weighting the correlation matrices, the result can be interpreted in terms of the subspace method of feature extraction developed by Watanabe and Pakvasa.

$$f_k(x) = u(x - a_k) \qquad 0 \le a_k \le 1$$

TN74-59 (1)

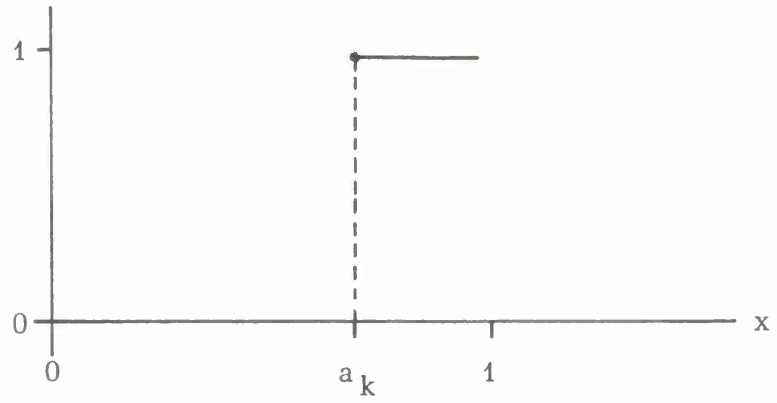

Fig. 1. Unit step function.

The functions $f_k(x) = u(x - a_k)$ map the eigenvalues of the correlation matrices into 0 and 1 and thereby transform the correlation matrices into so-called orthogonal projection operators. The projection operator $P_k = f_k(R_k)$ corresponds to a subspace $S(P_k)$ of the observation space spanned by the eigenvectors of R_k whose eigenvalues are greater than or equal to a_k. The projection operator P_k transforms any vector $\underline{x}$ into another vector $\underline{x}_k$ called the projection of $\underline{x}$ into $S(P_k)$. A geometrical interpretation is given in Fig. 2. The matrices G_i of (14a) are expressed in terms of the projection operators by

$$G_i = \frac{1}{K} \left[P_i + \sum_{\substack{k=1 \\ k \neq i}}^{K} (I - P_k) \right] \qquad i = 1, 2, \ldots, K \qquad (24)$$

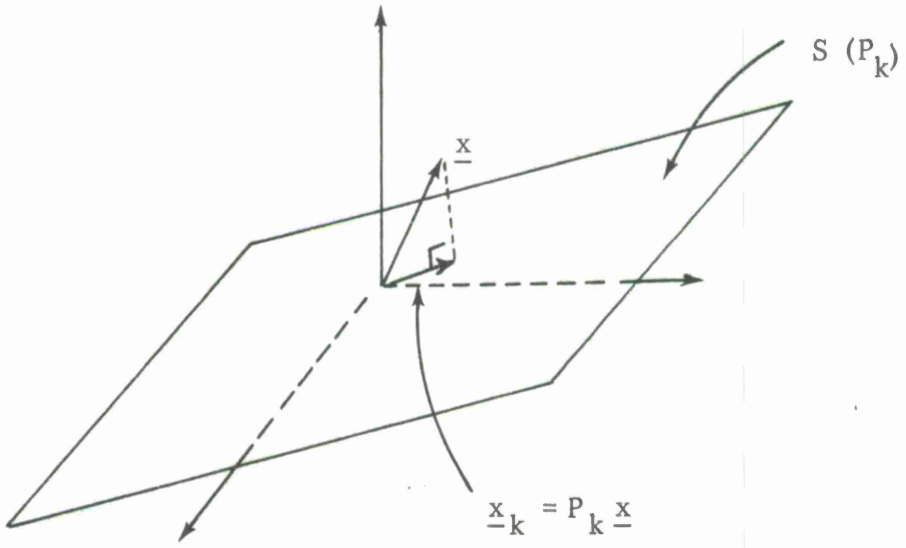

Fig. 2. Geometric interpretation of projection operators.

Watanabe calls the subspaces S (P_k) "representation subspaces" since they are spanned by the eigenvectors of each class that provide the optimal mean-square representation of that class. Feature subspaces are formed by removing the intersection of the representation classes. If S (P_i') is the i^{th} feature subspace then

$$S(P_i') = S(P_i) \cap \left[\bigcap_{\substack{k=1 \\ k \neq i}}^{K} \overline{S(P_k)} \right] \tag{25}$$

where $\overline{S(P_k)}$ denotes the complement of S (P_k).

It can be shown [3] that the eigenvectors of (24) corresponding to an eigenvalues of 1 span $S(P_i')$. In particular, if the postweighting function $h(x)$ is defined by

$$h(x) = \begin{cases} 1 & \text{if } x = 1 \\ 0 & \text{otherwise} \end{cases} \tag{26}$$

then the projection operators for the feature subspaces are given by

$$P_i' = h(G_i) \qquad i = 1, 2, \ldots, K \tag{27}$$

Features for the subspace technique are defined by Method 2, resulting in one feature for each class.*

VI. REMARKS ABOUT OPTIMAL WEIGHTING

The Fukunaga-Koontz results show that equal linear weighting optimizes the Divergence measure of separability in the two-class case with equal covariances. [1] It is probably very difficult to analytically determine weighting functions that would optimize any measure of class separability in more general cases. It does seem feasible to numerically optimize almost any criterion within certain classes of parameterized weighting functions. Both the linear functions and step functions described here are suitable choices for this type of optimization. Polynomial or piecewise-linear functions could also be used.

When using Method 1 for defining features, one would choose the postweighting function to select eigendirections corresponding to the largest m_k eigenvalues of each G_k and optimize the parameters of the preweighting functions. When using Method 2 for defining features both the preweighting and the postweighting functions must be optimized simultaneously and a dynamic programming approach may be appropriate.

---

* When the number of classes is small, the portion of the observation space not common to any of the feature subspaces may also be used to generate a feature.

VII. APPLICATION TO RADAR SIGNATURE CLASSIFICATION

Preliminary results from the application of the weighting function methods to the classification of radar signatures are reported in this section. The data to be classified consisted of 300 simulated radar signatures from each of two distinct objects (a reentry vehicle and a decoy) in ballistic trajectories. Each signature was represented in the observation space by a 30-dimensional vector corresponding to a set of sequential returns received by the radar. The data were then mapped into a 3-dimensional feature space using the Fukunaga-Koontz technique, the linear weighting technique, and the subspace technique.* Features were chosen according to Method 1 for the Fukunaga-Koontz and linear weighting techniques and according to Method 2 for the subspace technique. Fig. 3 shows results of classification in the observation space and in each of the three feature spaces. A quadratic classifier using the "leave-one-out" method was employed to produce the operating characteristics. For a three-dimensional feature space, the Fukunaga-Koontz, linear weighting, and the subspace methods show comparable performance. When the feature spaces for the Fukunaga-Koontz and linear weighting methods are increased to twelve dimensions, performance appoaches that of the classifier in the 30-dimensional observation space. These examples show that it is possible to considerably reduce the dimensionality of data through suitable linear transformations without greatly reducing the separability.

---

* Equal weighting $a_2 = a_1 = 1$ was used for the former two techniques.

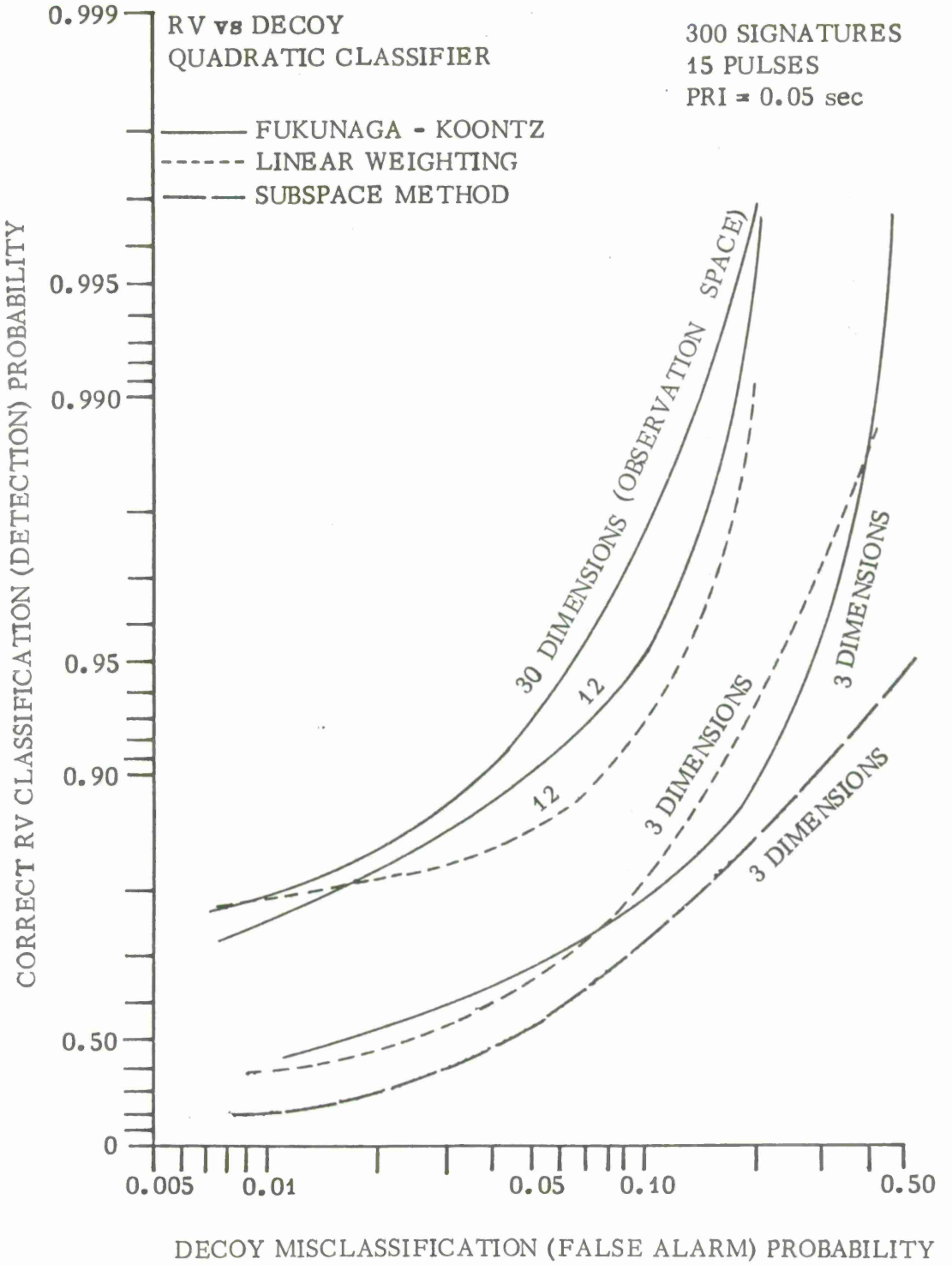

Fig. 3. Operating characteristic for quadratic classifier in feature spaces.

REFERENCES

1. K. Fukunaga and W. L. G. Koontz, "Application of the Karhunen-Loeve Expansion to Feature Selection and Ordereing," IEEE Trans, Computers C-19, 311-318 (1970).

2. S. Watanabe and N. Pakvasa, "Subspace Methods in Pattern Recognition," Proc. 1st Intl. Joint Conf. on Pattern Recognition, Washington, D.C., 30 October - 1 November 1973.

3. C. W. Therrien, "Eigenvalue Properties of Projection Operators and Their Application to the Subspace Method of Feature Extraction," Technical Note 1974-41, Lincoln Laboratory, M.I.T. (15 August 1974), DDC AD-784863/3.

4. C. W. Therrien, W. H. Schoendorf, G. L. Carayannopoulos and B. J. Burdick, "Application of Feature Extraction to Radar Signature Classification," Proc. 2nd Intl. Joint Conf. on Pattern Recognition, Lyngby, Denmark, 13-15 August 1974.

APPENDIX

Proof of Results Relating to the Optimal Basis

1. Proof of Equation (6)

Given any orthonormal basis $\left\{ \underline{u}_j,\ j = 1,\ 2,\ \ldots,\ n \right\}$, a vector $\underline{x}$ in the observation space can be represented by

$$\underline{x} = \sum_{j=1}^{n} b_j \underline{u}_j \tag{A.1}$$

where

$$b_j = \underline{x}^T \underline{u}_j \qquad j = 1,\ 2,\ \ldots,\ n \tag{A.2}$$

Let $\hat{\underline{x}}$ be a truncated representation of $\underline{x}$ given by (3). Then for any class k one can write

$$E_k\left[\left| \underline{x} - \hat{\underline{x}} \right|^2 \right] = E_k\left[\left(\sum_{j=m+1}^{n} b_j \underline{u}_j^T \right) \left(\sum_{j=m+1}^{n} b_j \underline{u}_j \right) \right] = \sum_{j=m+1}^{n} E_k\left[b_j^2 \right] \tag{A.3}$$

where the last equality derives from the orthonormal property of the basis. If (A. 2) and (1) are used in (A. 3), the latter equation becomes

$$E_k\left[\left| \underline{x} - \hat{\underline{x}} \right|^2 \right] = \sum_{j=m+1}^{n} E_k\left[\underline{u}_j^T \underline{x} \cdot \underline{x}^T \underline{u}_j \right] = \sum_{j=m+1}^{n} \underline{u}_j^T R_k \underline{u}_j \tag{A.4}$$

where R_k is the correlation matrix for class k.

2. Proof of Optimal Properties of the Eigenvectors of $\hat{G}_i$

It is desired to find the set of vectors $\underline{u}_{m+1}, \underline{u}_{m+2}, \ldots, \underline{u}_n$ that minimizes (10) subject to the normality constraint

$$\underline{u}_j^T \underline{u}_j = 1 \qquad j = m+1,\ m+2,\ \ldots,\ n \tag{A.5}$$

Let $\mu_{m+1}, \mu_{m+2}, \ldots, \mu_n$ be Lagrange multipliers. A necessary condition for the minimum is

$$\frac{\partial}{\partial \underline{u}_k}\left[\sum_{j=m+1}^{n} \underline{u}_j^T \hat{G}_i \underline{u}_j + \sum_{j=m+1}^{n} \mu_j (1 - \underline{u}_j^T \underline{u}_j)\right] = \underline{0} \qquad k = m+1, m+2, \ldots, n \qquad (A.6)$$

which reduces to the eigenvalue equation

$$\hat{G}_i \underline{u}_k - \mu_k \underline{u}_k = \underline{0} \qquad k = m+1, m+2, \ldots, n \qquad (A.7)$$

where $\underline{u}_k$ are eigenvectors and μ_k are the eigenvalues. If (A.7) is used in (10) then the criterion C_i becomes

$$C_i = \sum_{j=m+1}^{n} \mu_j \qquad (A.8)$$

Therefore to minimize C_i, one must choose the eigenvectors of $\hat{G}_i$ corresponding to the n - m smallest eigenvalues.

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER: ESD-TR-74-348	2. GOVT ACCESSION NO.
3. RECIPIENT'S CATALOG NUMBER	
4. TITLE *(and Subtitle)*: A Generalized Approach to Linear Methods of Feature Extraction	5. TYPE OF REPORT & PERIOD COVERED: Technical Note
6. PERFORMING ORG. REPORT NUMBER: Technical Note 1974-59	
7. AUTHOR(S): Therrien, Charles W.	8. CONTRACT OR GRANT NUMBER(S): F19628-73-C-0002
9. PERFORMING ORGANIZATION NAME AND ADDRESS: Lincoln Laboratory, M.I.T. P.O. Box 73 Lexington, MA 02173	10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS: Project No. 8X363304D215
11. CONTROLLING OFFICE NAME AND ADDRESS: Ballistic Missile Defense Program Office Department of the Army 1320 Wilson Boulevard Arlington, VA 22209	12. REPORT DATE: 30 December 1974
	13. NUMBER OF PAGES: 22
14. MONITORING AGENCY NAME & ADDRESS *(if different from Controlling Office)*: Electronic Systems Division Hanscom AFB Bedford, MA 01731	15. SECURITY CLASS. *(of this report)*: Unclassified
	15a. DECLASSIFICATION DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT *(of this Report)*

Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT *(of the abstract entered in Block 20, if different from Report)*

18. SUPPLEMENTARY NOTES

None

19. KEY WORDS *(Continue on reverse side if necessary and identify by block number)*

radar signatures
feature extraction
ballistic missile defense

20. ABSTRACT *(Continue on reverse side if necessary and identify by block number)*

An approach to feature extraction based on functions of the class correlation matrices is described. If linear functions of the correlation matrices are chosen, the present method extends the methods of feature extraction proposed by Fukunaga and Koontz. If certain types of non-linear functions are employed, the method reduces to the orthogonal subspace method of Watanabe and Pakvasa. Optimization of selected features through selection of appropriate functions is discussed briefly. Preliminary results of classification of radar signatures using the feature extraction methods described here are presented.