Toward an Optimal Theory and Computer-Based Implementation of Pattern Recognition Feature Selection

Lokesh Datta

A Thesis

in

The Department

of

Electrical Engineering

Presented in Partial Fulfilment of the Requirements
for the Degree of Doctor of Philosophy at
Concordia University
Montréal, Québec, Canada

December 1984

ISBN 0-315-35563-8

ABSTRACT

Toward an Optimal Theory and Computer-Based Implementation of Pattern Recognition Feature Selection

Lokesh Datta, Ph.D.
Concordia University, 1984

This work deals with a difficult problem of discriminating a weakly stationary complex Gaussian stochastic process against another when the mean values are similar and the covariance matrices (patterns) are different. The high level of mathematical or computational difficulty encountered in minimizing the probability of classification error has many times led workers in the areas of pattern recognition, information theory, communications, and control theory to utilize suboptimal statistical distance measures for feature selection. A new feature selection scheme is presented which deals quite directly with the Bayesian error expression. The scheme is developed using a combination of classical results, information-theoretic techniques, and concepts of distribution function theory. As opposed to asymptotic results, the scheme is found to be accurate for finite sample size. The approach to feature selection is shown, by use of numerous examples, to be superior to the conventional Kadota-Shepp strategy which employs distance measures and asymptotics in its formulation. A detailed analysis of the computational complexity of a pattern classifier incorporating the new feature selection strategy is included with an eye toward a computer-based implementation. The proposed configuration of the classifier consists of three modes of operation,

viz., the training mode, the processing mode, and the decision-directed mode. The system parameters are established in the training mode, the classification task is performed during the processing mode, and the system parameters are updated using the decision-directed mode to account for realistic quasi-stationarity of patterns. A combination of some well-known and a variety of new computationally efficient results are proposed in order to realize an efficient pattern classifier. In the process, we present characteristic equation reducibility results on centrosymmetric and centrohermitian matrices which provide a significant reduction in the arithmetic complexity encountered in the principal component (eigenvalue/eigenvector) extraction. In addition, an approximation of Toeplitz covariances by circulants is proposed which replaces the principal component extraction by the discrete Fourier transformation (DFT). The DFT can then be performed quite efficiently by the fast Fourier transformation (FFT) algorithm or by the Winograd fast transformation algorithm (WFTA). Results on the feature selection method are further substantiated by a variety of important numerical results on the effects of a free parameter found in the theory, a priori probabilities, and the number of features selected on the probability of classification error. This study on the theory and computer-based implementation of feature selection, including numerical examples and comparisons, may prove useful in stochastic signal classification applications such as image analysis/object recognition, speech analysis/speaker recognition, and robotics. In

addition, the theory presented here is a useful tool for evaluating the performance of competing feature selection schemes in situations when the error probability is extremely low, and thus, simulation is impractical.

TO MY PARENTS

ACKNOWLEDGEMENTS

The author wishes to express sincere gratitude to Dr. Salvatore D. Morgera for his supervision, guidance, encouragement, timely suggestions, and cooperation during the course of this work. His friendship and help certainly deserve my compliments.

Ms. Marie Berryman is to be thanked for an excellent job done with a difficult manuscript in a short period of time.

Mr. P. Misra, Mr. M.S.O. Sharma, and Mr. E. Wingrowicz of Concordia University deserve a special mention, as do the other friends at Concordia, for being there when needed. A certain friend who has been an inspiration for the completion of this work merits special thanks.

Last but not least, it is with pleasure and pride that the motivation provided by my parents and sisters is acknowledged.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF SYMBOLS AND ABBREVIATIONS

$\mathbb{C}^n$	n-dimensional complex space
$(\bullet)^H$	Matrix complex conjugate transpose
$(\bullet)^T$	Matrix transpose
$(\bullet)^c$	Matrix approximation by circulant
E_N	(N×N)-dimensional contra-identity matrix
I_N	(N×N)-dimensional identity matrix
$\mathbb{K}^{N\times N}$	Class of (N×N)-dimensional centrosymmetric matrices
$\mathbb{P}^{N\times N}$	Class of (N×N)-dimensional centrohermitian matrices
MVN	Multivariate normal
K-L	Karhunen-Loève
K-S	Kadota-Shepp
M-D	Morgera-Datta
CS	Centrosymmetric
CH	Centrohermitian
DFT	Discrete Fourier transform
FFT	Fast Fourier transform
WFTA	Winograd Fourier transform algorithm

CHAPTER 1

INTRODUCTION

1.1. GENERAL

Can a machine be devised that accurately recognizes a pattern? A concerted effort to answer this question by researchers in diverse areas of interest led to the conception of pattern recognition nearly three decades ago. The challenge presented by the idea of developing intelligent machines has translated into a considerable progress on both the theoretical and practical fronts of pattern recognition. Pattern recognition now finds applications in a wide variety of areas such as biomedical diagnostics, texture analysis for industrial inspection, earth resource satellite multispectral classification, radar remote sensing, speech analysis and speaker recognition, image analysis and object recognition, and industrial robotics.

A wide applicability of pattern recognition is a direct consequence of the inherent generality of the adopted concept of a pattern recognition system. Figure 1.1 depicts a pattern recognition system consisting of a sequence of three stages, viz., pattern representation, feature selection, and pattern classification. The pattern representation stage involves gathering data measurements and converting them into a suitable form for machine processing. The feature selection stage of the pattern recognition system is, perhaps, the most important in that it is chiefly responsible for the performance of the system. The main purpose of the feature selection stage is to reduce the

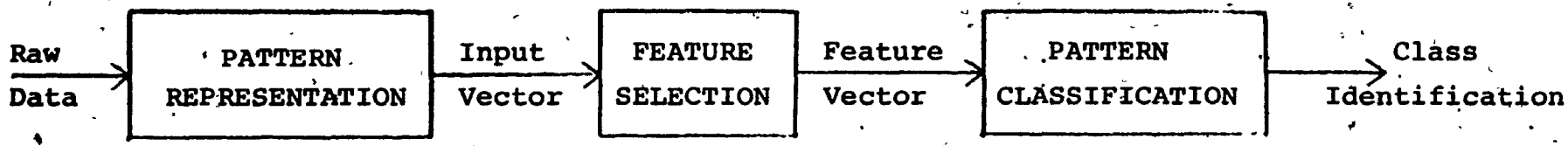

Figure 1.1 Basic Pattern Recognition System

dimensionality of the problem which may be necessitated by constraints of either a technical or economical nature. An important part of the feature selection process is to ensure the reliability of the system by extracting such information from the data vectors that is the most relevant to classification. The reduced dimensionality feature vectors thus obtained are recognized by the pattern classification stage. From a theoretical point of view, it is difficult to draw a boundary between the feature selection and classification stages. An ideal feature selector would make the pattern classification stage trivial, an almighty classifier would eliminate the need for the feature selector. Unfortunately, a pattern recognition system without the feature selection component, though theoretically feasible and plausible, may not be realizable given the practical constraints of high dimensionality of realistic problems and computing power available. Consequently, it generally is mandatory to incorporate the feature selection stage as a process of reducing the dimensionality of the problem preceding the classification stage.

The performance of a pattern recognition system is measured in terms of two important criteria: the probability of classification error and computational complexity. The problem of classification is primarily one of partitioning the feature space in such a manner that the decisions are never wrong. In the case when this cannot be achieved, an attempt is made to minimize the probability of classification error and, if some errors are more expensive than others, the average cost of errors. The second criterion of computational complexity is viewed in terms of the cost and speed of a practical implementation of the system.

1.2. SCOPE OF THE THESIS

This work is primarily concerned with the feature selection stage of the pattern recognition system. The problem considered here is that of discriminating a weakly stationary Gaussian stochastic process against another. The Gaussian stochastic processes are assumed to have similar mean vectors and different covariance matrices (patterns).

A number of feature selection methods are discussed in Chapter 2. These schemes approach a minimization of the probability of classification error in an "indirect" manner due to a general feeling that the probability of classification error expression is either mathematically or numerically intractable. Most of these methods of feature selection asymptotically approach the minimum error probability and are not workable and/or accurate for finite sample size. This motivates the development of a feature selection strategy which deals directly with the probability of error expression and provides accurate classification results for finite sample size.

Chapter 3 presents a new and accurate finite dimensionality information-theoretic strategy for feature selection. The scheme deals in more direct fashion with the error probability expression. The technique is shown, by means of numerous examples, to be superior to the well-known Kadota-Shepp (K-S) method [1.1]. The K-S method is a typical example of conventional feature selection schemes in that it employs asymptotics and statistical distance measures in its formulation. A more direct use of statistical distance measures, for example, the Bhattacharyya distance [1.2] is shown to provide "loose" bounds on the error probability, thereby, failing to provide useful information

for feature selection. The proposed scheme provides a tool not only for feature selection itself, but for a general performance evaluation of competing pattern classification schemes in the case when an extremely low error probability renders a computer simulation impractical.

The new feature selection strategy is developed by using a combination of classical results due to Laplace with further refinements by Polya et al [1.3], distribution function theory, and information theory. An appealing performance of the scheme in terms of the error probability motivates an investigation into the computational complexity of the technique in order to determine its feasibility for a computer-based implementation.

We begin by developing in Chapter 4 a variety of new results on matrix theory with an eye toward an efficient computer-based implementation of the feature selection scheme of Chapter 3. Chapter 4 presents the reducibility results on two classes of matrices of interest, namely, centrosymmetric (CS) and centrohermitian (CH) matrices. The results on CS matrices are a specialization of the results in [1.4] and a generalization of the results in [1.5]. The reducibility results on CH matrices are the first of this kind to appear in the literature. The results are useful for efficient principal component extraction required by the feature selection process. In passing, we mention that real symmetric and Hermitian Toeplitz matrices are special cases of the class of CS and CH matrices, respectively. In order to significantly enhance the computational efficiency of principal component extraction, Chapter 4 also includes a method of approximating real Toeplitz covariances by circulants. Principal component extraction can then be

replaced by the discrete Fourier transform (DFT). The DFT can be performed quite efficiently by using the well-known fast Fourier transform (FFT) or Winograd fast transform algorithm (WFTA).

The results of Chapter 4 offer a significant reduction in the computational complexity of a computer-based implementation of a pattern classifier employing the feature selection scheme of Chapter 3. A complexity analysis of the classifier is presented in Chapter 5. The classifier operates in three modes, viz., the training mode, the processing mode, and the decision-directed mode. The system parameters are established in the training mode, the classification of patterns is performed in the processing mode, and the decision-directed mode, in conjunction with the training mode, updates the system parameters taking into account a realistic quasi-stationarity of patterns. The study on implementation proposes efficient algorithms and corresponding computational complexity analysis for every step in the classification process for realizing an efficient pattern classifier. A number of the proposed algorithms are new.

Computer simulation results of the pattern classifier are presented in Chapter 6. The probability of classification error results for the feature selection scheme of Chapter 3 are compared with that of the conventional K-S method. This comparison substantiates the claim made in Chapter 3 regarding a superior performance of the new scheme in terms of the error probability. The error probability results for the new scheme are also analyzed in view of error bounds on the probability of error using the Bhattacharyya distance. For the examples examined, the bounds are repeatedly found to be quite loose and devoid of any useful information for feature selection.

Effect of *a priori* probabilities of the patterns on the overall error probability for finite sample size is also considered in Chapter 6. Although, asymptotically the error probability is independent of *a priori* probabilities [1.6], it is observed, as expected, that the error in classification for finite sample size decreases as the *a priori* probability of one pattern is increased relative to the other. Behavior of the classification error versus the number of features selected is examined. We find, consistent with classical thought, that the error decreases sharply as the number of features is increased to a certain value and then the decrease in error begins to taper off gradually.

1.3. MAJOR CONTRIBUTIONS OF THE WORK

This work offers a variety of new, interesting, and useful results on the theory and computer-based implementation of pattern recognition feature selection for finite sample size. The information-theoretic approach to feature selection, developed by utilizing classical methods and distribution function theory, deals directly with the Bayes error expression. The work demonstrates the suboptimality of conventional feature selection schemes employing statistical distance measures and asymptotic formulation. A number of examples considered here and comparisons therof with a typical conventional scheme such as that of Kadota and Shepp (K-S) [1.1] substantiate that the new scheme is always *at least* as good as, and sometimes better, by *an order of magnitude*, than the K-S method for finite sample size. The study finds the use of statistical distance measures, by means of examples, often inadequate.

All numerical examples considered discriminate between Toeplitz covariances. Toeplitz covariances are of extreme practical importance in that they are often used for information representation and modelling, e.g., speech representation for speaker recognition. An appreciable data compression ratio of 0.25 (or 75% compression) is used for all examples. These examples not only permit us to demonstrate a better performance of the new scheme in comparison to the conventional methods for feature selection, but also provide better understanding and insight into the problem. Moreover, the comparisons are the first of this type to be found in the literature.

A computer-based implementation of the pattern classifier employing the new feature selection scheme is proposed. As with many feature selection schemes, a principal component extraction is initially required. Computationally efficient algorithms for this task are presented which utilize certain a priori known structure of the covariances or covariance products involved, i.e., Toeplitz, or centrosymmetric and centrohermitian, respectively. The reducibility results on centrohermitian matrices are the first to appear in the literature. In order to increase speed significantly, approximation of the Toeplitz covariances by circulants is proposed. This approach leads to satisfactory error rates when there is sufficient statistical independence within each data vector, and allows the principal component extraction to be replaced by the discrete Fourier transform (DFT). The implementation study continues with the presentation of most suitable algorithms and corresponding complexity analysis for each step of the feature selection process in order to realize an efficient pattern classifier. Several aspects of the implementation study may also be used to enhance

the computational efficiency of many other feature selection schemes. A detailed complexity analysis of the pattern classifier employing the new feature selection scheme shall prove to be an important contribution for practical applications of stochastic signal classification such as encountered in passive sonar, as well as in the areas of image processing, speech recognition, and robotics.

REFERENCES

[1.1] T. Kadota and L.A. Shepp, "On the Best Set of Linear Observables for Discriminating Two Gaussian Signals," IEEE Trans. Inform. Theory, vol. IT-13, pp. 278-285, Apr. 1967.

[1.2] T. Kailath, "The Divergence and Bhattacharyya Distance Measures in Signal Selection," IEEE Trans. Comm. Tech., vol. COM-15, pp. 52-60, Feb. 1966.

[1.3] G. Polya and G. Szego, Problems and Theorems in Analysis, vol. 1. New York: Springer-Verlag, 1976.

[1.4] A.L. Andrew, "Eignevectors of Certain Matrices," Linear Algebra App., vol. 7, pp. 151-162, 1973.

[1.5] A. Cantoni and P. Butler, "Eigenvalues and Eigenvectors of Symmetric Centrosymmetric Matrices," Linear Algebra App., vol. 13, pp. 275-288, 1976.

[1.6] U. Grenander, Abstract Inference. New York: Wiley, 1981.

CHAPTER 2

METHODS IN PATTERN RECOGNITION FEATURE SELECTION

2.1. INTRODUCTION

Let $\underline{x}$ be an $(N\times 1)$-dimensional complex stochastic data vector with multivariate normal (MVN) distribution, $N(\underline{0}, R_i)$, under hypothesis H_i $i=1,2$ for the binary or two-class hypothesis testing problem. The hypothesis H_i is assumed to have a priori probability π_i $i\neq 1,2$ with $\pi_1+\pi_2 = 1$ and $\pi_1 \neq 0,1$. The general problem of extracting n features $(n<N)$ may be viewed as that of either selecting n suitable measurements from the N elements of $\underline{x}$, or of selecting n appropriate linear functionals constituting an n-dimensional linear space $\mathcal{J}^n$. In either case, it is desired to have a probability measure corresponding to each pattern class in feature space. The effectiveness of feature selection relates to the performance of the pattern classifier, usually in terms of probability of error or misclassification. Thus, the solution to the feature extraction problem lies in choosing a subset of the N elements such that the probability of classification error is minimized.

The mathematical techniques of feature selection in pattern recognition may be broadly classified into two categories, namely, feature selection in the measurement space, and feature selection in the transformed space. The development of feature selection schemes belonging to the first category has been based on the implicit assumption that the acquisition of data or measurements representing the input

patterns is costly. The main objective in this case then is to minimize the cost associated with measurement extraction and achieve a reduction in the dimensionality of the problem by reducing the number of measurements required to recognize the patterns. This can be accomplished by eliminating the measurements which provide redundant, irrelevant, or insignificant information. Reducing the number of initial measurements can, for example, lead to savings in sensor hardware and computing power for data processing. A number of feature selection schemes in the measurement space are discussed in Section 2.2.

Feature selection schemes belonging to the second category utilize the entire representation vector to obtain a feature vector of lower dimension. The elimination of redundant, irrelevant, or insignificant information is achieved by applying a transformation which maps the patterns from the representation space to a lower dimension feature space. The key to resolving the feature selection problem here is to construct an optimal transformation which minimizes the probability of misclassification. Some interesting feature selection schemes belonging to this category are discussed in Section 2.3. In passing, we mention that the new and accurate technique of feature selection presented in Chapter 3 also belongs to this particular category.

It is interesting to note that the feature selection methods in the transformed space can be applied not only to the representation vector for dimensionality reduction but also, to achieve further data compression, in the feature space determined by feature selection

schemes in the measurement space. Although the feature selection schemes of these two categories are not mutually exclusive in their applicability, such a simplistic approach to the classification of methods permits us to discuss various aspects of these techniques; for example, the performance reliability in terms of error probability.

2.2. FEATURE SELECTION IN THE MEASUREMENT SPACE

The key to resolving the feature selection problem in the measurement space is to obtain a subset of n features $\underline{y}$ from the set of N measurements of the data vector $\underline{x}$ such that the probability of misrecognition is minimized with respect to any other combination of n features selected from $\underline{x}$. Unfortunately, no simple expressions for classification error are available for establishing the best set of features and, in practice, one has to be satisfied with a compromise of selecting a feature set $\underline{y}^*$ which optimizes some criterion $J(\underline{y})$, i.e.,

$$J(\underline{y}^*) = \max_{\{\underline{y}_i\}} \{J(\underline{y}_i)\} \tag{2.1a}$$

or,

$$J(\underline{y}^*) = \min_{\{\underline{y}_i\}} \{J(\underline{y}_i)\} \tag{2.1b}$$

with an assumption that $J(\underline{y})$ can be related to error probability [2.1]. The members of the family $\{\underline{y}_i\}$ in (2.1) are all the possible combinations of n features that may be selected from N measurements of the pattern $\underline{x}$. The following subsections discuss a number of ways in which the task of optimizing $J(\underline{y})$ may be accomplished.

2.2.1. STATISTICAL DISTANCE MEASURES

The notion of "distance" between two hypothesis or patterns has been defined in many different ways in mathematical statistics and all distance measures are qualitatively related to the probability of misclassification in a similar manner. The underlying concept for the use of distance measures is that the larger the distance established between two patterns by the features selected, the better the performance of the classifier (or lower error probability) [2.1-2.5].

Let $p_i(\underline{y})$ be the probability density function (pdf) of the feature vector under hypothesis H_i $i=1,2$. The classification of features may be based, for example, on the log-likelihood ratio [2.1, 2.2, 2.4, 2.6], given by,

$$\ln L(\underline{y}) = \ln \left[\frac{p_1(\underline{y})}{p_2(\underline{y})}\right] \underset{H_2}{\overset{H_1}{\gtrless}} T \qquad (2.2)$$

where the quantity T is a certain threshold value. It is interesting to observe that the probability of error would be small if the average value of $\ln L(\underline{y})$ is large for the patterns belonging to H_1 and small for those belonging to H_2. Define,

$$E_i\,[\ln L(\underline{y})] \triangleq \int_{\mathcal{C}^n} [\ln L(\underline{y})]\, p_i(\underline{y})d\underline{y} \quad i=1,2 \qquad (2.3)$$

where the quantity $E_i[\bullet]$ may be interpreted as the average information for discrimination against H_j $j=1,2$ with $i \neq j$. The <u>J-divergence</u>, defined as,

$$J_n \triangleq E_1\,[\ln L(\underline{y})] - E_2\,[\ln L(\underline{y})] \qquad (2.4)$$

is therefore a useful distance measure for the discrimination of two

classes [2.1-2.9]. Note that when the classes are separable, i.e., $p_1(\underline{y}) = 0$ if $p_2(\underline{y}) > 0$ and vice-versa, the patterns are classified without any error and $J=\infty$. In contrast, when the patterns are indistinguishable, i.e., $p_1(\underline{y}) = p_2(\underline{y})$, we have $J=0$.

Another commonly used distance measure for feature selection, known as the <u>Bhattacharyya distance</u>, is defined as [2.1-2.9];

$$B_n = -\ln \rho_n \tag{2.5a}$$

where ρ_n is the Bhattacharyya coefficient, given by,

$$\rho_n = \int_{\mathcal{C}^n} [p_1(\underline{y})p_2(\underline{y})]^{\frac{1}{2}}\, d\underline{y} \tag{2.5b}$$

The integral in (2.5b) is also known as Hellinger's integral [2.10]. Hellinger's integral may be interpreted as the inner product of two vectors of unit norm, namely, $\sqrt{p_1(\underline{y})}$ and $\sqrt{p_2(\underline{y})}$, with ρ_n being the cosine of the angle between the two vectors. Also, we have [2.4, 2.6],

$$0 < \rho_n < 1 \tag{2.6a}$$

and, therefore,

$$0 < B_n < \infty. \tag{2.6b}$$

It is noted that when the patterns are separable and may be classified without error, the coefficient $\rho_n=0$ and, on the other hand, when the patterns are indistinguishable and classification is not possible, the coefficient $\rho_n=1$. Thus, in order to obtain an optimal classifier, the selection of features must be performed such that the Bhattacharyya coefficient is minimized or, equivalently, the Bhattacharyya distance is maximized.

In general, all distance measures share some common properties,

e.g., the distance measures are non-negative and they attain minimum or maximum values when the classes are indistinguishable or separable, respectively. The suitability of distance measures for feature selection can be justified by the arguments presented above but their potential could be assessed only if their relationships to error probability were known. Unfortunately, the probability of misclassification cannot be expressed precisely in terms of these measures, but it is comforting to note that various bounds are known which relate the error probability to some of these measures [2.1-2.7], e.g.,

$$\tfrac{1}{2} - \tfrac{1}{2}\left[1-4\pi_1\pi_2\,\rho_n^2\right]^{\frac{1}{2}} < P_e(n) < \left[\pi_1\pi_2\right]^{\frac{1}{2}}\rho_n \qquad (2.7)$$

and,

$$\tfrac{1}{2}\min(\pi_1,\pi_2)\exp(-J_n/8) < P_e(n) < \left[\pi_1\pi_2\right]^{\frac{1}{2}}(J_n/4)^{-\frac{1}{2}} \qquad (2.8)$$

for the Bhattacharyya distance and J-divergence, respectively, where $P_e(n)$ is the probability of misclassification based on n features.

Feature selection schemes based on probabilistic distance measures generally involve some form of optimization of these measures or some criterion utilizing the measures. Although only two most commonly used distance measures have been mentioned, several other distance measures and the bounds which relate them to the probability of error are available in the literature [2.1,2.5]. The lack of exact expressions for error probabilty in terms of distance measures suggests that, ideally, one would use a distance measure which provides tighter bounds with lower error probability than the others, but, in practice, other aspects such as computational complexity may be taken into account with the final choice being a suitable problem-dependent compromise.

2.2.2 DEPENDENCE MEASURES

Section 2.2.1 has presented some criteria for feature selection which are based on the "distance" between two pattern classes. This section deals with the criteria that are based on the statistical dependence. The most commonly used measure of probablistic dependence for feature selection is the mutual information [2.1,2.4,2.11,2.12]. Let

$$p(\underline{y}) = \sum_{i=1}^{2} \pi_i p_i(\underline{y}) \tag{2.9}$$

be the mixture density where π_i is the a priori probability and $p_i(\underline{y})$ is the pdf under H_i $i=1,2$. Define,

$$I(\underline{y}) = \sum_{i=1}^{2} \pi_i E_i \left[\ln \frac{p_i(\underline{y})}{p(\underline{y})} \right] \tag{2.10}$$

where $E_i[\bullet]$ is defined as in (2.3). Each term in (2.10) represents the information for discrimination in favour of H_i against the overall mixture. We may rewrite (2.10) as,

$$I(\underline{y}) = \sum_{i=1}^{2} \int_{\mathcal{D}^n} \pi_i p_i(\underline{y}) \ln \left[\frac{p_i(\underline{y})}{p(\underline{y})} \right] d\underline{y} \tag{2.11a}$$

or, equivalently,

$$I(\underline{y}) = G(\underline{y}) - \sum_{i=1}^{2} \pi_i \, G(\underline{y}|H_i) \tag{2.11b}$$

where $G(\underline{y})$, the entropy or average uncertainty of $\underline{y}$, is given by

$$G(\underline{y}) = - \int_{\mathcal{D}^n} p(\underline{y}) \ln p(\underline{y}) d\underline{y} \tag{2.11c}$$

and the conditional entropy $G(\underline{y}|H_i)$ for a given H_i is given by,

$$G(\underline{y}|H_i) = - \int_{\mathcal{D}^n} p_i(\underline{y}) \ln p_i(\underline{y}) d\underline{y} \quad i=1,2 \tag{2.11d}$$

The quantity $I(\underline{y})$ in (2.11) is known as the mutual information between $\underline{y}$ and the set $\{H_i \quad i=1,2\}$, and may be interpreted as the information about H_i obtained by observing the random vectors $\underline{y}$.

The use of mutual information is justifiable for feature selection due to the following reason. When a feature set has high information about a pattern class, one is fairly certain about which pattern class is present but, on the other hand, when the information provided by the feature set is low, one is uncertain about which class the measurement was taken from. Low or high quantity of information, as one would expect, results in low or high probability of error, respectively [2.4]. A feature set that does not provide any discrimination between the classes gives a minimum of information, whereas a feature set that provides perfect discrimination attains a maximum of mutual information. In practice, however, a feature vector would provide mutual information which lies somewhere in between these two extremes.

Similar to the situation for distance measures, no exact relationship is available which relates error probability to mutual information. A variety of dependence measures have been developed utilizing the concept of mutual information and various bounds are available which relate some of the dependence measures to error probability [2.1, 2.11]. For example, the Bhattacharyya dependence [2.11], given by,

$$B_n^D = \ln \rho_n^D \tag{2.12a}$$

where

$$\rho_n^D = \sum_{i=1}^{2} \pi_i \int_{\mathcal{C}^n} [p_i(\underline{y})\, p(\underline{y})]^{\frac{1}{2}}\, d\underline{y} \tag{2.12b}$$

provides error probability somewhere within the region specified by,

$$0 < P_e(n) < \rho_n^D - \tfrac{1}{2} \qquad (2.12b)$$

for equal a priori probabilities.

Dependence measures, like probabilistic distance measures, are a useful tool for feature selection. In fact, they share the same common properties of non-negativity, and attain maximum or minimum when the classes are indistinguishable or separable, respectively.

2.2.3. EUCLIDEAN DISTANCE MEASURES

Euclidean distance measures for feature selection in the measurement space have originated from an intuitive argument that the greater the Euclidean distance between the elements of different classes, the better the classification performance. Euclidean measures are not commonly used techniques for feature selection even though they are significantly less complex. The low complexity of these measures is attractive but is compromised at the expense of system performance in terms of classification error [2.1]. Some well-known Euclidean measures are discussed in [2.12-2.14].

2.3. FEATURE SELECTION IN THE TRANSFORMED SPACE

The methods of feature selection in the transformed space are conceptually different from the methods in the measurement space in that they utilize all measurements in representing the pattern. As mentioned earlier, the feature selection schemes of Section 2.2 are based on the assumption that the aquisition of data for measurement is

expensive. On the contrary, it is on rare occasions in practice that the aquisition of data is associated with very high costs. In most cases the data may be aquired without significantly increasing the cost and it becomes advantageous to utilize all available information to design a pattern classifier. Although the methods of feature selection in the measurement space still remain a subject of theoretical discussion, feature selection techniques in the transformed space have emerged as the methods of practical importance.

Linear feature selection in the transformed space involves mapping an $(N\times 1)$-dimensional data vector $\underline{x}$ into an $(n\times 1)$-dimensional feature vector $\underline{y}$ $(n<N)$ by applying an $(n\times N)$-dimensional transformation A, i.e.,

$$\underline{y} = A\ \underline{x} \tag{2.13}$$

The data reducing transformation A may be any linear vector function of $\underline{x}$. Several possible ways of obtaining a suitable transformation for linear feature selection are discussed below.

2.3.1. KARHUNEN LOEVE TRANSFORM

The Karhunen-Loève (K-L) transform is a well-known technique for representing a sample function of a square-integrable stochastic process [2.15]. It has been shown that the K-L transform is an optimal transform in a statistical sense under a variety of criteria [2.16-2.18]. The K-L transform completely decorrelates any sequence in the transformed space permitting us to process one transform coefficient without affecting the others. It provides most energy (or variance)

in the fewest number of coefficients of the transform. The property which makes this transform quite appealing for feature selection, albeit principally for data compression, is that it provides minimum mean square error (mse) between the reconstructed and original data. In addition, the K-L transform minimizes the total entropy of the reconstructed sequence.

A brief description of the K-L transform is as follows. For a given sequence $\underline{x}$, the basis function of the transform are the eigenvectors of its covariance matrix R. The K-L transform is a unitary matrix, K, whose columns are the normalized eigenvectors of R, such that,

$$K^H R K = \Lambda \tag{2.14a}$$

where H denotes the matrix complex conjugate transpose, and the diagonal matrix Λ, given by,

$$\Lambda = \mathrm{diag}\{\lambda_1, \lambda_2, \ldots, \lambda_N\} \tag{2.14b}$$

contains the eigenvalues of R. The columns of K in (2.14) are arranged such that $\lambda_1 > \lambda_2 > \ldots > \lambda_N$. In view of the relationship expressed in (2.14), the K-L transform is sometimes called the principal component transform. If only the n $(n<N)$ eigenvectors corresponding to the first n eigenvalues (features) are selected for data compression, the mse in data reconstruction is given by,

$$\text{mse} = \sum_{i=n+1}^{N} \lambda_i \tag{2.15}$$

The mse given by (2.15) is minimum for the K-L transform compared to

that of any other discrete unitary transform [2.16,2.18].

The application of the K-L transform for respresenting a pattern class is discussed by Watanabe [2.17]. A necessary condition under which more than one random process may by represented by a single expansion is presented by Chien et al [2.19]. Fukunaga et al [2.17] emphasized the extraction of features that enhance the class separability rather than the information preserving aspects of the K-L is representation. This was done by obtaining a diagonalizing transformation for the mixture covariance in the manner of (2.14). The mixture covariance is simply the sum of two covariance matrices representing the two pattern classes. It turns out that under hypothes H_1, we have $1>\lambda_1^{(1)}>\lambda_2^{(2)}> \ldots >\lambda_N^{(1)}> 0$ and under hypothesis H_2, $0<\lambda_i^{(2)} = 1-\lambda_i^{(1)}< 1$. The recommendation of Fukunaga et al is that the λ_i be selected such that $|\lambda_i-0.5|$ is largest, i.e., the eigenvalues closest to 0 or 1. This criterion for feature selection has been disputed by Foley [2.20]. There are several other feature selection schemes available that are based on the K-L transform, e.g., see [2.1-2.5, 2.14, 2.21].

2.3.2. SEPARABILITY MEASURES

The separability measures for feature selection are based on the assumption that the pattern classes under consideration may be represented adequately by the second order statistics, i.e., the mean vectors and covariance matrices. It is then desired to determine an optimal transformtion matrix which maximizes some separability measure, for example, the statistical distance measures of Section 2.2.1. Such a

a treatment of the pattern recognition feature selection has been proposed, for example, by Kadota and Shepp [2.22]. The Kadota-Shepp (K-S) strategy stems from the maximization of the J-divergence resulting in a transformation matrix whose columns are the first n eigenvectors such that,

$$\lambda_1 + \frac{1}{\lambda_1} > \lambda_2 + \frac{1}{\lambda_2} > \ldots > \lambda_N + \frac{1}{\lambda_N} \qquad (2.16)$$

Where λ_i's are the eigenvalues of the product matrix $R_1^{-\frac{1}{2}} R_2 R_1^{-\frac{1}{2}}$ for binary classification. This feature selection scheme has been disputed by Chesler et al [2.23], where a counterexample to the optimality of the scheme is presented. Morgera et al [2.24] have recently put forth the reasons for the suboptimality of the K-S technique. The separability measures, such as the dependence measures of Section 2.2.2, may also be optimized to construct suitable data reducing transformations [2.25].

2.3.3. NON-ORTHOGONAL MAPPING

Some non-orthogonal schemes have been proposed for the feature selection task [2.26]. It has been shown that it is possible to find a feature space of lower dimensionality by the use of non-orthogonal projections. The reliability of such schemes is questionable as the criteria used for feature selection depend on Euclidean interclass distances rather than on error probability. Moreover, no analytic expressions are available which relate the performance of these techniques to the classification error; thus, the reliability of the schemes can be assessed only by means of experiments.

2.4. DISCUSSION

Several feature selection schemes for pattern classification have been briefly discussed in this chapter. Feature selection techniques have been classified into two categories depending on their application in the measurement space or in the transformed domain.

Feature selection methods in the measurement space, where one is mainly concerned with saving sensor hardware, suffer from one major disadvantage of excessive computational requirements. To determine the best subset of n features from N measurement, it is required that some separability criterion be computed $\binom{N}{n}$ times. It is obvious that this number attains an astronomical value for large N rendering these schemes infeasible for practical utility. For example, when $N=40$ and $n=10$, it is required that some separability measure be evaluated nearly 848 million times.

The separability measures are a useful tool for feature selection regardless of their application in the measurement space or the transformed space. The choice of a feature selection method is, in general, a problem-dependent compromise between the computational complexity and the reliability of the scheme. The methods based on probablistic distance or dependence measures may be considered better since they optimize criteria which many times can be related to error probability.

The challenge presented by this extremely important problem of feature selection has motivated many researchers in pattern recognition over the last two decades. A multitude of suboptimal and sometimes ad hoc solutions to the problem is available in the literature.

Unfortunately, these schemes do not provide explicit and exact expressions for classification error. The reliability of the schemes may be based on some bounds on error probability but, as we shall see in the sequel, these bounds, at times, are quite "loose" and fail to provide any useful information.

It has not been the intention in this chapter to undertake the impossible task of enumerating all feature selection schemes that are available, but simply to demonstrate the necessity of developing a feature selection strategy which deals more directly with error probability expression. Such a scheme is proposed in Chapter 3. The scheme has been found to be quite accurate for finite data dimension.

REFERENCES

[2.1] J. Kittler, "Mathematical Methods of Feature Selection in Pattern Recognition," Int. J. Man-Machine Studies, vol. 7, pp. 609-637, 1975.

[2.2] P.A. Devijver and J. Kittler, Pattern Recognition. New York: Prentice-Hall, 1982.

[2.3] C. Chen, Statistical Pattern Recognition. New Jersey: Hayden, 1978.

[2.4] T.Y. Young and T.W. Calvert, Classification, Estimation and Pattern Recognition. New York: Elsevier, 1974.

[2.5] L. Kanal, "Patterns in Pattern Recognition: 1968-1974," IEEE Trans. Inform. Theory, vol. IT-20, pp. 697-722, Nov. 1974.

[2.6] T. Kailath, "The Divergence and Bhattacharyya Distance Measures in Signal Selection," IEEE Trans. Comm. Tech., vol. COM-15, pp. 52-60, Feb. 1967.

[2.7] G.T. Toussaint, "Comments on the Divergence and Bhattacharyya Distance Measures in Signal Selection," IEEE Trans. Comm. Tech., vol. COM-20, p. 485, 1972.

[2.8] D. Kazakos and P. Papantoni-Kazakos, "Spectral Distance Measures Between Gaussian Processes," IEEE Trans. Automat. Contr., vol. AC-25, pp. 950-959, Oct. 1980.

[2.9] A. Caprihan and R.J.P. De Figueiredo, "On the Extraction of Pattern Features from Continuous Measurements," IEEE Trans. Syst., Sci., Cybern., vol. SSC-16, pp. 110-115, Apr. 1970.

[2.10] C.S. Rao and V.S. Vardarajan, "Discrimination of Gaussian Processes," Sankhyā, ser. A, vol. 25, pp. 303-330, 1963.

[2.11] T.R. Vilmansen, "Feature Selection with Measures of Probablistic Dependence," IEEE Trans. Comput., vol. C-22, pp. 381-388, Apr. 1973.

[2.12] M. Michael and W. Lin, "Experimental Study of Information Measures and Inter-Intra Class Distance Ratios on Feature Selection and Orderings," IEEE Trans. Syst., Sci., Cybern., vol. SMC-3, pp. 172-181, Mar. 1973.

[2.13] S. Wilks, Mathematical Statistics. New York: Wiley, 1962.

[2.14] R.O. Duda and P.E. Hart, Pattern Classification and Scene Analysis. New York: Wiley, 1973.

[2.15] A. Papoulis, Probability, Random Variables and Stochastic Processes. New York: McGraw-Hill, 1965, p. 457.

[2.16] S. Watanabe, "Karhunen-Loève Expansion and Factor Analysis: Theoretical Remarks and Applications," Proc. 4th Prague Conf. Inform. Theory, Prague, 1965.

[2.17] K. Fukunaga and W.L.G. Koontz, "Application of the Karhunen-Loève Expansion to Feature Selection and Ordering," IEEE Trans. Comput., vol. C-19, pp. 311-318, Apr. 1970.

[2.18] D.F. Elliott and K.R. Rao, Fast Transforms: Algorithms, Analyses, Applications. New York: Academic, 1982, p. 382.

[2.19] Y.T. Chien and K.S. Fu, "On the Generalized Karhunen-Loeve Expansion," IEEE Trans. Inform. Theory, vol. IT-13, pp. 518-520, July 1967.

[2.20] D.H. Foley, "Orthonormal Expansion Study for Waveform Processing," Rome Air Develop. Center, AF Systems Command, Grifiss AFB, New York, Tech. Rep. RADC-TR-73-168, July 1973.

[2.21] J. Kittler and P.C. Young, "A New Approach to Feature Selection Based on the Karhunen-Loève Expansion," Pattern Recognition, vol. 5, p. 335, 1973.

[2.22] T.T. Kadota and L.A. Shepp, "On the Best Set of Linear Observables for Discriminating Two Gaussian Signals," IEEE Trans. Inform. Theory, vol. IT-13, pp. 278-285, Apr. 1967.

[2.23] D.A. Chesler and R.L. Greenspan, " Comments on Choosing Observable for Discriminating Two Gaussian Signals," IEEE Trans. Inform. Theory, vol. IT-14, pp. 820-822, Nov. 1968.

[2.24] S.D. Morgera and L. Datta, "Toward a Fundamental Theory of Optimal Feature Selection: Part I," IEEE Trans. Pattern Anal. Machine Intell., vol. PAMI-6, pp. 601-616, Sept. 1984.

[2.25] M. Ichino and K. Hiramatsu, "Suboptimal Linear Feature Selection in Multiclass Problem," IEEE Trans. Syst., Man., Cybern., vol. SMC-4, p. 28, 1974.

[2.26] T.W. Calvert, "Nonorthogonal Projections for Feature Extraction in Pattern Recognition," IEEE Trans. Comput., vol. C-19, pp. 447-452, May 1970.

CHAPTER 3

TOWARD AN OPTIMAL THEORY OF FEATURE SELECTION

3.1. INTRODUCTION

This chapter presents a feature selection scheme for discriminating a weakly-stationary Gaussian stochastic process against another when the mean vectors are similar and the pattern classes (covariances) differ. The scheme deals more directly with the Bayesian probability of error expression than the existing methods discussed in Chapter 2. The approach to feature selection described here has been motivated by the desire to achieve more accurate and precise results by minimizing the error probability expression, as opposed to optimizing some probablistic measure which may provide "loose" bounds on the classification error. A variety of important examples are considered to demonstrate the performance of the scheme. Although the examples deal with only one important class of Toeplitz covariances, the scheme is applicable to covariances of a general nature. In passing, we mention that some of the results presented in this chapter have recently been reported in [3.1, 3.2].

Let $\underline{x}$ be an $(N\times 1)$-dimensional complex stochastic data vector with multivariate normal (MVN) distribution, specified by mean vector zero and covariance matrix R_i under hypothesis H_i $i=1,2$. We wish to discriminate H_1 against H_2 in the Bayesian manner assuming that H_i has an *a priori* probability π_i $i=1,2$ with $\pi_1+\pi_2 = 1$ and

$\pi_1 \neq 0,1$. Linear feature selection may be accomplished by transforming the (N×1)-dimensional data vector $\underline{x}$ into an (n×1)-dimensional feature vector $\underline{y}$ (n<N) by applying an (n×N)-dimensional data reducing transformation A, i.e.,

$$\underline{y} = A\,\underline{x} \tag{3.1}$$

The rows of A are assumed to be linearly independent (ℓ.i.); thus, A has rank n. The (n×1)-dimensional feature vector $\underline{y}$ has a MVN distribution, $N(\underline{0},S_i)$, under hypothesis H_i, where the transformed covariance, S_i is defined as, $S_i \triangleq A^H R_i A$ i=1,2, where H denotes the matrix complex conjugate transpose.

The pattern classification may now be based on the feature vector $\underline{y}$ with an average probability of error, $P_e(n)$, given by,

$$P_e(n) = \pi_1 I_1(n) + \pi_2 I_2(n) \tag{3.2a}$$

where,

$$I_1(n) = \int_W p_1(\underline{y})d\underline{y}\ ;\quad I_2(n) = \int_{W^c} p_2(\underline{y})d\underline{y} \tag{3.2b}$$

and,

$$W = \{\underline{y}|\underline{y}^H[S_1^{-1}-S_2^{-1}]\underline{y} > \ell n|S_2|-\ell n|S_1| + 2\ell n(\frac{\pi_1}{\pi_2})\} \tag{3.2c}$$

where $p_i(\underline{y})$ is the probability density function (pdf) of $\underline{y}$ under hypothesis H_i i=1,2; and the region $W \varepsilon \mathcal{C}^n$ is the critical region for rejecting H_1.

Ideally, one would like to select the transformation A for optimal feature selection such that $P_e(n)$ of (3.2) is minimized. Due a general consensus expressed by the workers in pattern recognition,

information theory, communications, and control systems that the minimization of $P_e(n)$ is often difficult to carry out, several researchers have utilized suboptimal schemes such as probablistic distance measures to develop some bounds on the probability of classification error [3.3-3.10]. The application of these schemes has been successful in practice, e.g., discriminating between seismic records [3.11], dynamical model approximation [3.12], speech recognition [3.13], and signal selection in communication and radar systems [3.14]. Although it is not possible to dispute the utility of these feature selection schemes, the results presented here indicate that some of these suboptimal approaches to feature selection can indeed be quite inferior to a more direct use of the classification error.

3.2. BAYESIAN DISCRIMINATION - FINITE SAMPLE SIZE

The problem of feature selection may be cast into a simpler form as follows. Assume that the covariance matrices R_i $i=1,2$ of two weakly-stationary Gaussian stochastic processes are positive definite. There exists a non-singular $(N\times N)$-dimensional transformation matrix L and a diagonal matrix Λ with elements arranged in descending order of magnitude such that [3.15],

$$L^H R_1 L = I_N \ , \ L^H R_2 L = \Lambda \tag{3.3}$$

where I_N is the $(N\times N)$-dimensional identity matrix. The diagonal elements $\lambda_1 > \lambda_2 > \ldots > \lambda_N > 0$ of Λ are the roots of the following determinantal equation,

$$|R_2 - \lambda R_1| = 0 \tag{3.4}$$

i.e., the λ_i $1 \le i \le N$ are the eigenvalues with respect to the matrix pair (R_1, R_2). The ith column $\underline{\ell}_i$ of the transformation matrix L in the simultaneous reduction of (3.3) is an eigenvector of the pair (R_1, R_2) associated with the eigenvalue λ_i due to the fact that,

$$R_2 L = R_1 L \Lambda \tag{3.5a}$$

or, equivalently,

$$R_2 \underline{\ell}_i = \lambda_i R_1 \underline{\ell}_i \qquad 1 \le i \le N \tag{3.5b}$$

Moreover, all λ_i are positive and, if all λ_i are <u>distinct</u>, then the transformation L may be determined uniquely except for the sign of every column.

Consider a situation where the transformation L is applied to the data vector $\underline{x}$ prior to feature selection; then (3.1) can be reformulated as,

$$\underline{y} = A\,\underline{x}^* = AL^H\underline{x} \tag{3.6}$$

where the feature vector $\underline{y}$ is MVN distributed as $N(\underline{0}, AA^H)$ or $N(\underline{0}, A\Lambda A^H)$ under hypothesis H_1 or H_2, respectively. Let L^* be the non-singular $(n \times n)$-dimensional transformation matrix and Λ^* be the diagonal matrix with elements λ_i^* arranged as $\lambda_1^* > \lambda_2^* > \ldots > \lambda_n^*$ for the simultaneous reduction of the matrix pair $(AA^H, A\Lambda A^H)$ in the manner of (3.3). Noting that the feature vector,

$$\underline{y}^* = L^{*H}\underline{y} \tag{3.7}$$

leads to the same hypothesis test as the feature vector of (3.6), the probability of error expression of (3.2) may be reformulated as,

$$P_e(n;\underline{\lambda}^*) = \pi_1 I_1(n;\underline{\lambda}^*) + \pi_2 I_2(n;\underline{\lambda}^*) \tag{3.8a}$$

where,

$$I_1(n;\underline{\lambda}^*) = \text{Prob}\ \{\sum_{i=1}^{n}(1-\frac{1}{\lambda_i^*})z_i^2 > \sum_{i=1}^{n} \ln\lambda_i^* + 2\ln(\frac{\pi_1}{\pi_2})\} \tag{3.8b}$$

$$I_2(n;\underline{\lambda}^*) = \text{Prof}\ \{\sum_{i=1}^{n}(\lambda_i^*-1)z_i^2 < \sum_{i=1}^{n} \ln\lambda_i^* + 2\ln(\frac{\pi_1}{\pi_2})\} \tag{3.8c}$$

where z_i $1<i<n$ are statistically independent (s.i.) $N(0,1)$ variates. The critical region for rejecting H_1, based on the feature vector $\underline{y}^*$ of (3.7) for $p_e(n;\underline{\lambda}^*)$ of (3.8), is given by,

$$W(\underline{\lambda}^*) = \{\underline{y}^* | \underline{y}^{*H}[I_n - \Lambda^{*-1}]\underline{y}^* > \sum_{i=1}^{n} \ln\lambda_i^* + 2\ln(\frac{\pi_1}{\pi_2})\} \tag{3.9}$$

The error probabilities of (3.8) are dependent <u>only</u> on the λ_i^* ; thus, an $(n\times 1)$-dimensional vector $\underline{\lambda}^*$ formed from λ_i^* is called a <u>canonical parameter</u> in the feature space $\mathcal{C}^n$.

It shall be important to investigate the discrete empirical distribution function of the eigenspectrum $\{\lambda_i^*\ \ 1<i<n\}$ in view of the concept presented by Okamoto [3.16] that the values of λ_i^* distant from unity, larger or smaller, lead to low probability of error. We state his result as follows:

<u>Theorem 3.1.</u> (Okamota, 1961). For fixed <u>a priori</u> probabil ties π_1 and π_2, the probability of classification error, $P_e(n;\underline{\lambda}^*)$, is strictly monotonically:

i) increasing in λ_i^* if $0\ \lambda_i^*<1$, and

ii) decreasing in λ_i^* if $\lambda_i^*>1$

for each value of the index i $i=1,2, \ldots, n$.

<u>Proof</u>. ii) Let the critical region for rejecting H_2 be,

$$W^c(\underline{\lambda}^*) = \{\underline{y}^* | Q(\underline{y}^*) < k(\underline{\lambda}^*)\} \tag{3.10a}$$

where the quadratic form $Q(\underline{y}^*)$ is given by,

$$Q(\underline{y}^*) = \underline{y}^{*H}[I_n - \Lambda^{*-1}]\underline{y}^* \tag{3.10b}$$

and the decision threshold,

$$k(\underline{\lambda}^*) = \sum_{i=1}^{n} \ln\lambda_i^* + 2\ln(\frac{\pi_1}{\pi_2}) \tag{3.10c}$$

Assume $\lambda_i^*>1$ for all i $1<i<n$, and let $\underline{\lambda}^{*'}$ and $\underline{\lambda}^{*''}$ be two canonical parameters which differ by one component such that $\lambda_i^{*''}>\lambda_i^{*'}$ for any $i\in[1,n]$. Define the generalized errors, $I_2(n;\underline{\lambda}^*,\underline{\lambda}^{*'})$ and $I_2(n;\underline{\lambda}^*,\underline{\lambda}^{*''})$, under hypothesis H_2 using (3.8) and (3.10) as,

$$I_2(n;\underline{\lambda}^*,\underline{\lambda}^{*'}) = \text{Prob}\ \{\sum_{j=1}^{n}(1-\frac{1}{\lambda_j^*})\ \lambda_j^{*'}z_i^2 < k(\underline{\lambda}^*)\} \tag{3.11a}$$

$$I_2(n;\underline{\lambda}^*,\ \underline{\lambda}^{*''}) = \text{Prob}\ \{\sum_{j=1}^{n}(1-\frac{1}{\lambda_j^*})\ \lambda_j^{*''}z_i^2 < k(\underline{\lambda}^*)\} \tag{3.11b}$$

since $(1-1/\lambda_j^*)\lambda_j^{*''} > (1-1/\lambda_j^*)\lambda^{*'}$ for each j $1<j<n$, we have from (3.11),

$$I_2(n;\underline{\lambda}^*,\lambda^{*''}) < I_2(n;\underline{\lambda}^*,\lambda^{*'}) \tag{3.12}$$

Therefore, the value of $I_2(n;\underline{\lambda}^*,\underline{\lambda}^{*'})$ is monotonically <u>non-increasing</u> in $\lambda_i^{*'}$ for each i if $\lambda_i^*>1$. It can be readily shown in a similar manner that $I_2(n;\underline{\lambda}^*,\underline{\lambda}^{*'})$ is monotonically <u>non-decreasing</u> in λ_i^* for each i if $0<\lambda_i^*<1$

We wish to establish a similar result for the probability of classification error, $P_e(n;\underline{\lambda}^*)$. Assume $\lambda_i^*>1$ for all i $1<i<n$ and let $\underline{\lambda}^*$ and $\underline{\lambda}^{*'}$ be two canonical variates which differ by one component such that $\lambda_i^{*'}>\lambda_i^*$ for any $i\in[1,n]$. Using (3.8) and (3.12),

we have,

$$P_e(n;\underline{\lambda}^*) = \pi_1 I_1(n;\underline{\lambda}^*) + \pi_2 I_2(n;\underline{\lambda}^*,\underline{\lambda}^*)$$
$$> \pi_1 I_1(n;\underline{\lambda}^*) + \pi_2 I_2(n;\underline{\lambda}^*,\underline{\lambda}^{*\prime})$$

With a change of critical regions $W(\underline{\lambda}^*)$ and $W^c(\underline{\lambda}^*)$ to $W(\underline{\lambda}^{*\prime})$ and $W^c(\underline{\lambda}^{*\prime})$, respectively, and noting that the Bayes discrimination threshold $k(\underline{\lambda}^{*\prime})$ does not coincide with $k(\underline{\lambda}^*)$, we obtain

$$P_e(n;\underline{\lambda}^*) > \pi_1 I_1(n;\underline{\lambda}^{*\prime}) + \pi_2 I_2(n;\underline{\lambda}^{*\prime},\underline{\lambda}^{*\prime})$$
$$= P(n;\underline{\lambda}^{*\prime})$$

This completes the proof for ii). The proof for i) may be developed in a similar manner. ∎

We see from Theorem 3.1 that when the eigenvalues λ_i are more distant from unity, larger or smaller, the smaller is the probability of classification error; thus, the distribution of the eigenspectrum $\{\lambda_i\}$ about unity determines the performance of the pattern classifier.

In view of Theorem 3.1 and deductions thereof, we define the <u>discrete empirical distribution function</u> (df) of the eigenspectrum $\{\lambda_i^*\}$ as,

$$F_n^*(x) \triangleq \frac{1}{n} \bullet \# \{\lambda_i^* | \lambda_i^* < x \quad 1 < i < n\} \tag{3.13}$$

where the symbol # reads as "the number of". The df of (3.13) has an increment 1/n and may also be written as,

$$F_n^*(x) = \frac{1}{n} \bullet \sum_i m_i u(x-\lambda^*_{n+1-i})$$

where the index i ranges over distinct λ_i^* and m_i denotes the respective multiplicities. The step function $u(x)=0$ for $x<0$ and unity otherwise. The df $F_n^*(x)$ is <u>non-decreasing</u> with a finite number of discontinuities of the first kind, i.e., it is possible to determine the limits $F_n^*(x^+)$ and $F_n^*(x^-)$ everywhere. The number of discontinuities in $F_n^*(x)$ is given by the number of <u>distinct</u> λ_i^*. The quantities $F_n^*(1)$ and $[1-F_n^*(1)]$ represent the cumulative distribution of eigenvalues below and above unity, respectively. We wish to select an optimal data reducing transformation A for (3.1) such that a "suitable" df $F_n^*(x)$ is obtained in the sense of Theorem 3.1. Clearly, the best choice for $F_n^*(x)$ is dependent on the eigenspectrum $\{\lambda_i\}$ of the covariance matrix pair (R_1, R_2). Therefore, we define the df of the eigenspectrum $\{\lambda_i\}$ in the manner of (3.13),

$$F_N(x) \triangleq \frac{1}{N} \bullet \# \{\lambda_i \mid \lambda_i < x \quad 1 \le i \le N\} \tag{3.14}$$

Note that $F_N(x)$ has an increment $1/N$, and the quantities $F_N(1)$ and $[1-F_N(1)]$ determine the cumulative eigenvalue distribution below and above unity, respectively.

As we shall see in the sequel, the detailed dependence of the df $F_n^*(x)$ on the transformation A is extremely important, and is vital to understanding the feature selection problem. The application of various distance measures does not show any regard for this dependence and may lead to a "forced" selection of certain features which could be quite suboptimal in terms of the Bayesian error probability. Chesler <u>et al</u> [3.9] noticed, but did not formalize, this in a communication theory context.

3.3. FEATURE SELECTION - FINITE SAMPLE SIZE

In this section we discuss the optimal selection of linear functionals (features), i.e., the best choice from the subset of possible transformations A, any one of which projects the N-dimensional transformed observation vector $L^H \underline{x} \in \mathcal{C}^N$ of (3.6) onto an n-dimensional feature space $(n<N)$, $\mathcal{C}^n \subseteq \mathcal{C}^N$. We begin by assuming that the transformation L for the simultaneous reduction of the covariance pair (R_1, R_2) is determined uniquely; thus, all λ_i $1<i<N$ are distinct. In this case, the vector $\underline{y}$ of (3.6) is MVN distributed as $\mathcal{N}(0, \Lambda\Lambda^H)$ and $\mathcal{N}(0, \Lambda\Lambda\Lambda^H)$ under hypotheses H_1 and H_2, respectively, and the transformation L^* used in (3.7) is simply an unitary transformation. It is noted that either or both of these above restrictions may be relaxed; however, it is unnecessary at this stage to do so.

Consider the <u>pencils of quadratic forms</u> $\underline{x}^{*H}[\Lambda-\lambda I_N]\underline{x}^*$ and $\underline{y}^H[\Lambda^H-\lambda^* I_n]\underline{y}$ associated with the simultaneously diagonalized pair (R_1, R_2) and the associated pair (S_1, S_2), respectively. An examination of the transformation of (3.6) shows that we have implicitly imposed (N-n) l.i. constraints on $\underline{x}^*$ of the form $L_k(\underline{x}^*) = 0$ $k=1,2,\ldots,(N-n)$, where $L_k(\underline{x}^*)$ are the linear forms of the variables $\underline{x}^*_1, \underline{x}^*_2, \ldots, \underline{x}^*_N$. Therefore, the roots λ^* are bounded from above and below by certain roots λ in the manner of Poincaré.

<u>Theorem 3.2</u>. (Poincaré Separation Theorem [3.17]). Let $\lambda_1>\lambda_2>\ldots>\lambda_N$ be the eigenvalues of the pencil of quadratic forms $\underline{x}^{*H}[\Lambda-\lambda I_N]\underline{x}^*$ and let $\lambda_1^*>\lambda_2^*>\ldots>\lambda_n^*$ be the eigenvalues of the same pencil subject

to (N-n) ℓ.i. constraints, then we have,

$$\lambda_i > \lambda_i^* > \lambda_{i+(N-n)} \qquad 1<i<n \tag{3.15}$$

Theorem 3.2 permits us to directly determine a bound on the df $F_n(x)$ of (3.13) in terms of the df $F_N(x)$ of (3.14), as done in [3.18]. This result is stated as a lemma.

<u>Lemma 3.1</u>. Let $F_N^s(x)$ be the df of the n smallest eigenvalues $\{\lambda_{i+(N-n)} \quad i=1,2,\ldots,n\}$ of the eigenspectrum $\{\lambda_i\}$ and $F_N^\ell(x)$ be the df of the n largest eigenvalues $\{\lambda_i \quad i=1,2,\ldots,n\}$, given by,

$$F_N^s(x) = \min\{1, \frac{N}{n} F_N(x)\} \tag{3.16a}$$

$$F_N^\ell(x) = \max\{0, 1-\frac{N}{n} F_N(x)\} \tag{3.16b}$$

for all x, respectively. Then, the df $F_n^*(x)$ is bounded from above by $F_N^s(x)$, and from below by $F_N^\ell(x)$.

<u>Proof</u>. It is obvious from the definition (3.14) of the df $F_N(x)$ and the required properties of any df [3.19] that (3.16a) and (3.16b) describe the df's associated with the n smallest and the n largest eigenvalues of the eigenspectrum $\{\lambda_i\}$. The equivalent bounds on the df $F_n^*(x)$, given by,

$$F_N^s(x) > F_n^*(x) > F_N^\ell(x) \quad \text{all } x \tag{3.17}$$

are directly obtained from Theorem 3.2 as desired. The quantity N/n, the inverse of the data compression ratio appearing in (3.16), simply adjusts the increment of $F_N(x)$ to that of the df $F_n^*(x)$. ∎

The upper and lower bounds of (3.17) can be achieved by selecting the rows of A, $\underline{a}_i^T$ $1<i<n$, as certain columns of the transformation matrix L which simultaneously diagonalizes the pair (R_1, R_2), i.e., by choosing certain eigenvectors associated with the pair (R_1, R_2) as the rows of A. More specifically,

$$\text{if} \quad \underline{a}_i^T = \begin{cases} \underline{\ell}_{i+(N-n)} \\ \\ \underline{\ell}_i \end{cases} \quad 1<i<n, \quad \text{then} \quad F_n^*(x) = \begin{cases} F_N^s(x) \\ \\ F_N^\ell(x) \end{cases} \tag{3.18}$$

for all x, where $\underline{\ell}_i$ is the ith eigenvector corresponding to the eigenvalue λ_i of the pair (R_1, R_2). It is shown in the sequel that, in general, the <u>best</u> choice of features in the sense of minimizing the Bayesian classification error is <u>neither</u> of the cases of (3.18), but depends on the <u>underlying structure</u> of the df $F_N(x)$.

We now consider five examples of covariance matrix pairs (patterns) in this work. Figures 3.1-3.5 display the df $F_N(x)$, and the upper and lower bounds, $F_N^s(x)$ and $F_N^\ell(x)$, respectively, for the five examples of Table 3.1. In all examples, the data and feature dimensions are selected as $N=40$ and $n=10$, respectively; thus, the data <u>compression ratio</u> $n/N=0.25$ (or 75% compression). For purposes of illustration, consider a case when the rows of A are selected arbitrarily as $\underline{a}_i^T = \underline{\delta}_i^T$ $1<i<n$, where $\underline{\delta}_i$ is a vector with unity in the ith position and zeroes elsewhere. The resulting df $F_n^*(x)$ for Example I is displayed in Figure 3.6.

TABLE 3.1

Toeplitz Covariance Matrix Pairs (R_1,R_2)
Selected as Examples

EXAMPLE	DESCRIPTION (R_1,R_2)	PARAMETERS
I	(R_1,R_2) first order Markov	
	$\rho_{k,\lvert i-j\rvert}=e^{-\alpha_k\lvert i-j\rvert}$ $k=1,2$	$\alpha=1$, $\alpha_2=0.5$
II	(R_1,R_2) first order Markov	$\alpha_1=1$, $\alpha_2=0.25$
III	(R_1,R_2) second order Markov	$\beta_1=1$, $\gamma_1=1.429$
	$\rho_{k,\lvert i-j\rvert}=e^{-\beta_k}\rho_{k,\lvert i-j\rvert-1}+e^{-\gamma_k}\rho_{k,\lvert i-j\rvert-2}$ $k=2$	$\beta_2=0.5$, $\gamma_2=2$
IV	R_1 first order Markov R_2 second order Markov	$\alpha_1=2$ $\beta_1=0.5$, $\gamma_2=2$
V	R_1 "triangular"	
	$\rho_{1,\lvert i-j\rvert}=1-\varepsilon\lvert i-j\rvert$	$\varepsilon_1=0.025$
	R_2 first order Markov	$\alpha_2=1$

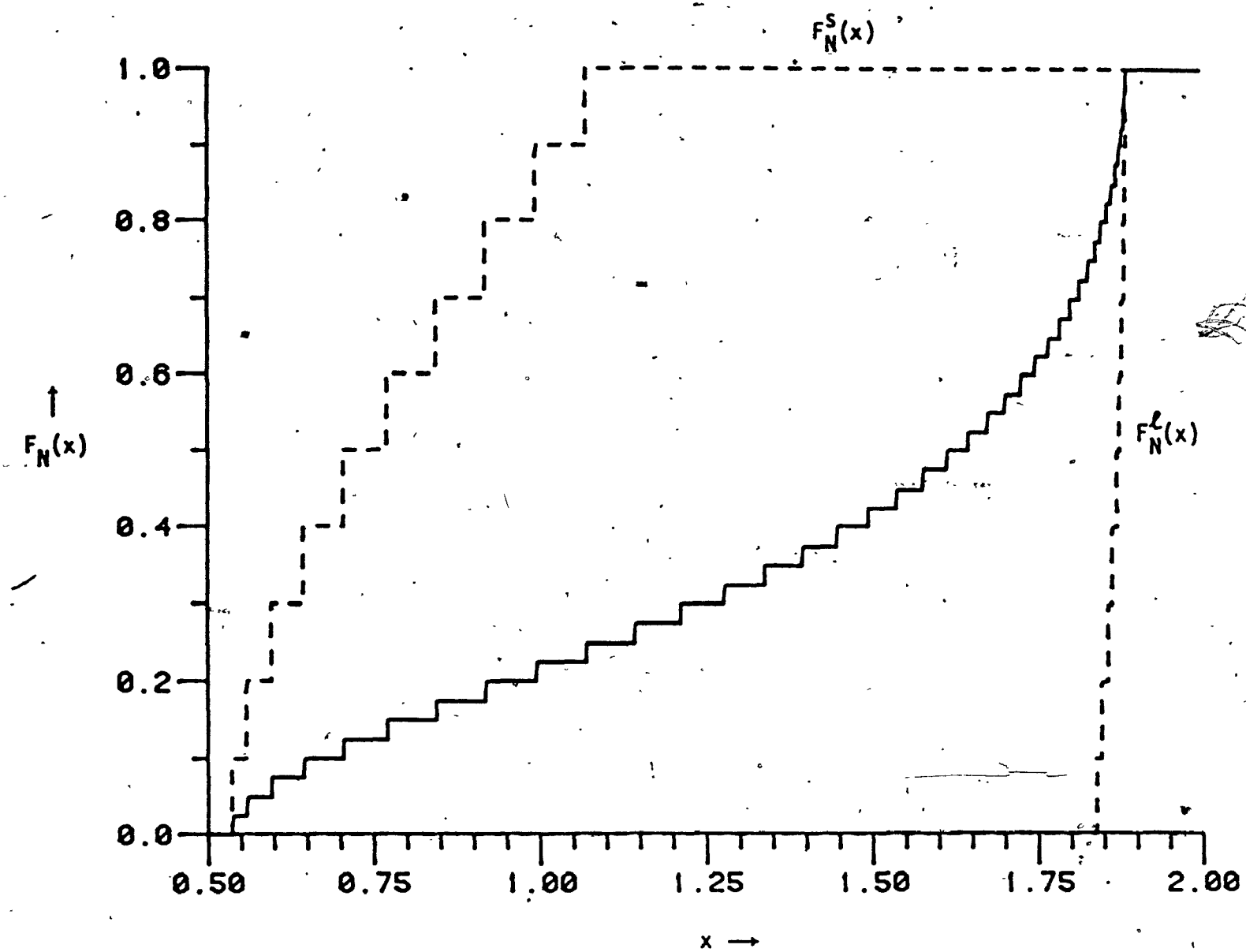

Figure 3.1 Distribution Function $F_N(x)$ and Bounds, $F_N^s(x)$ and $F_N^\ell(x)$. Covariance Example I.

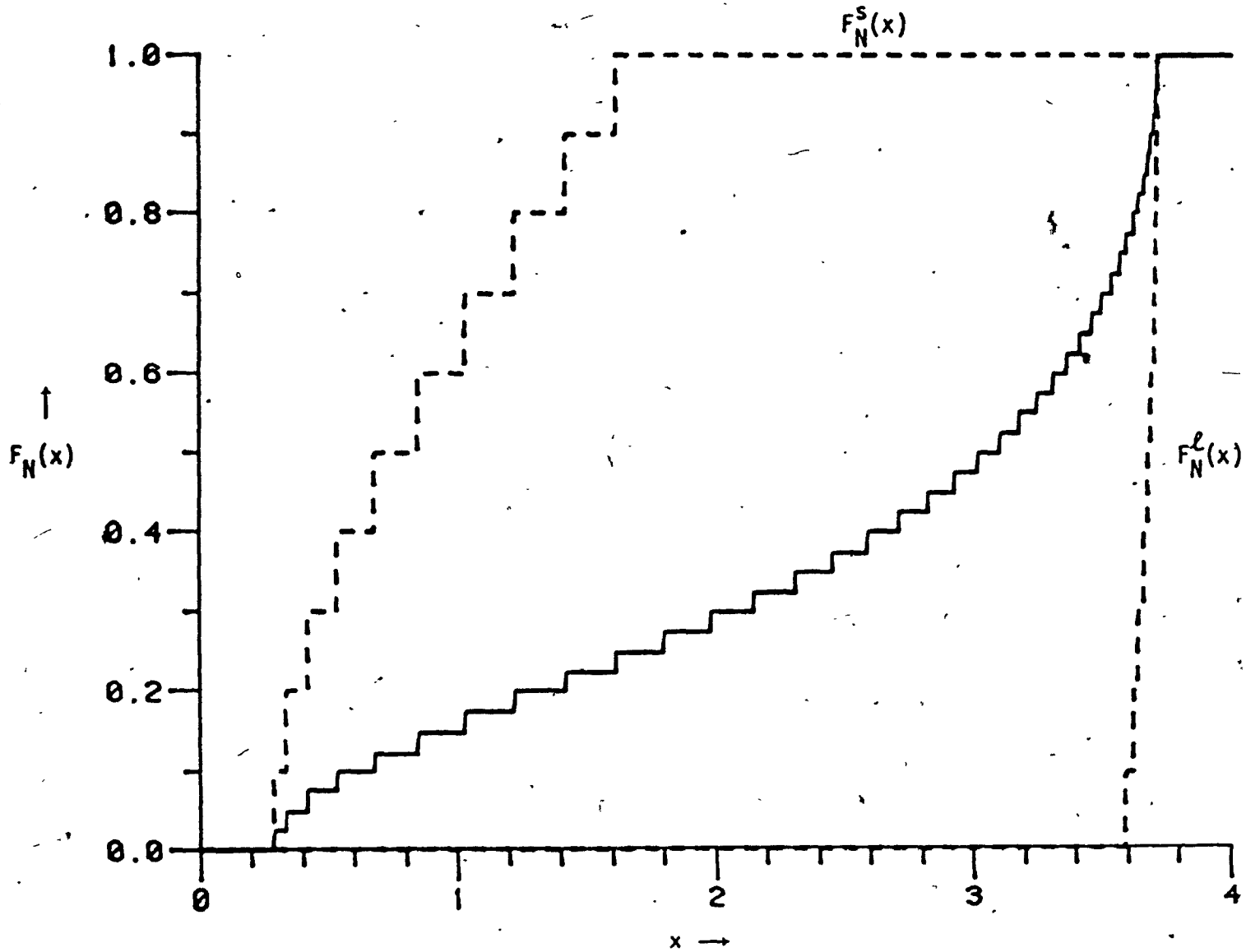

Figure 3.2 Distribution Function $F_N(x)$ and Bounds, $F_N^s(x)$ and $F_N^\ell(x)$. Covariance Example II.

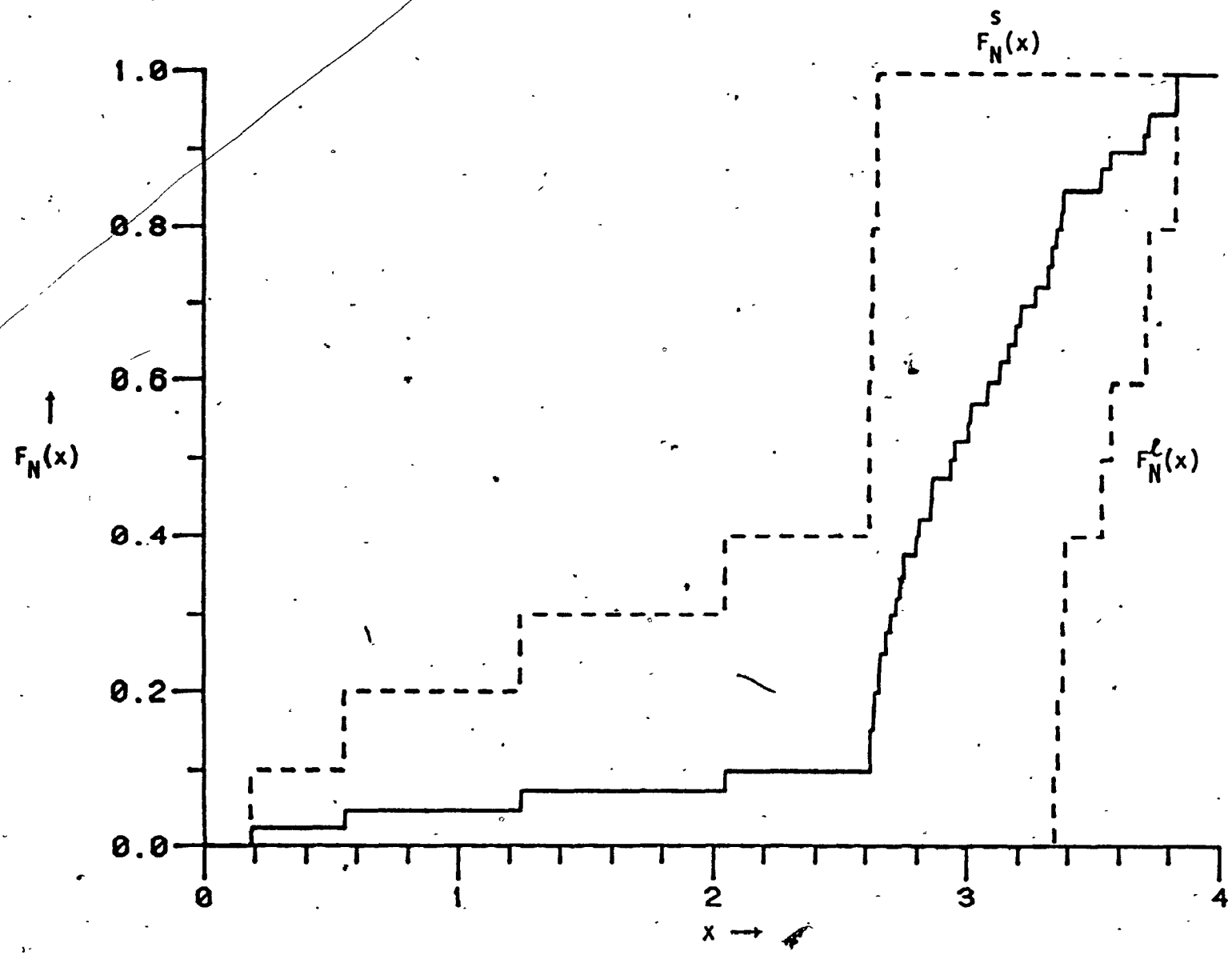

Figure 3.3 Distribution Function $F_N(x)$ and Bounds, $F_N^s(x)$ and $F_N^\ell(x)$. Covariance Example III.

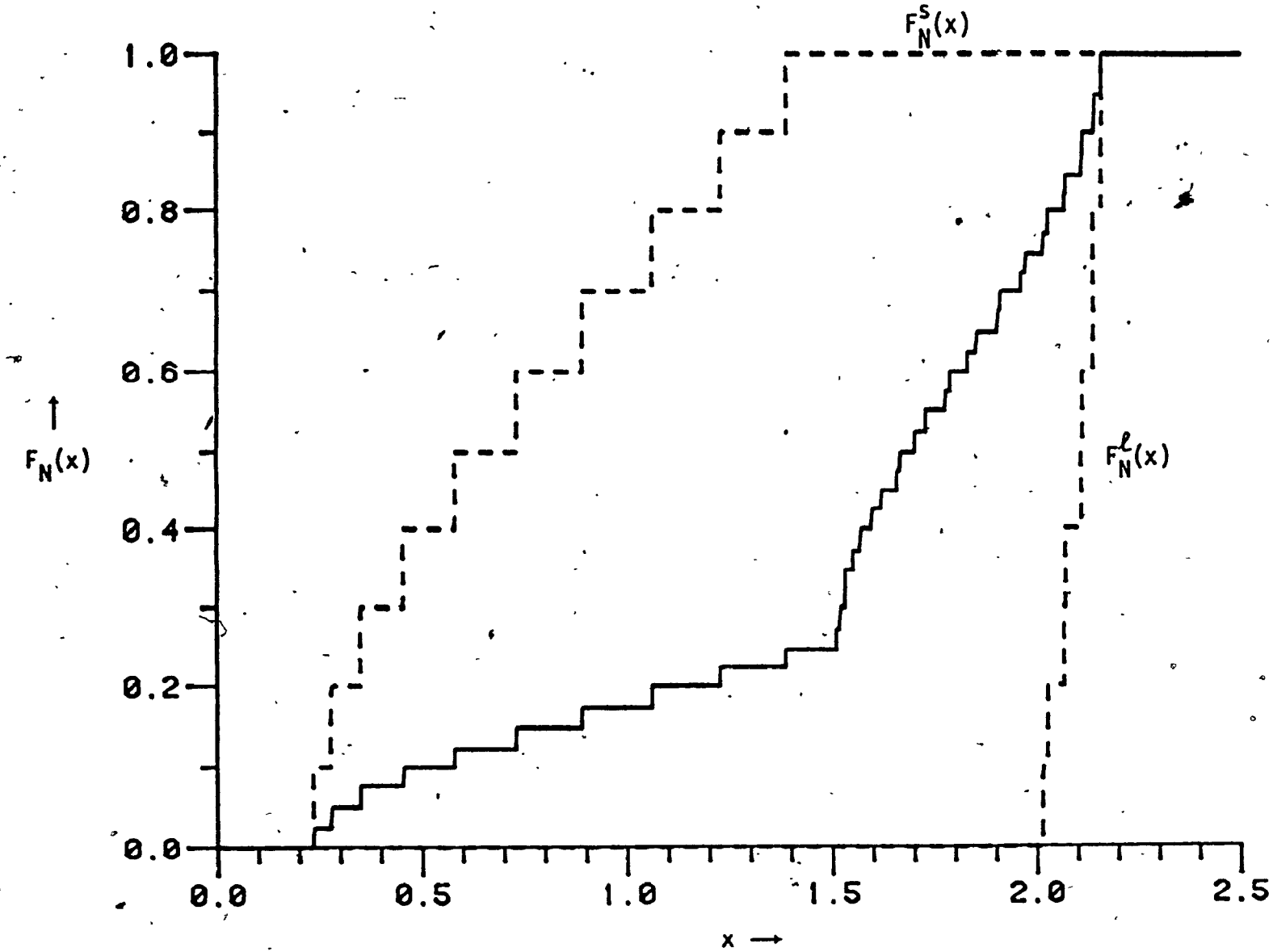

Figure 3.4 Distribution Function $F_N(x)$ and Bounds, $F_N^s(x)$ and $F_N^\ell(x)$. Covariance Example IV.

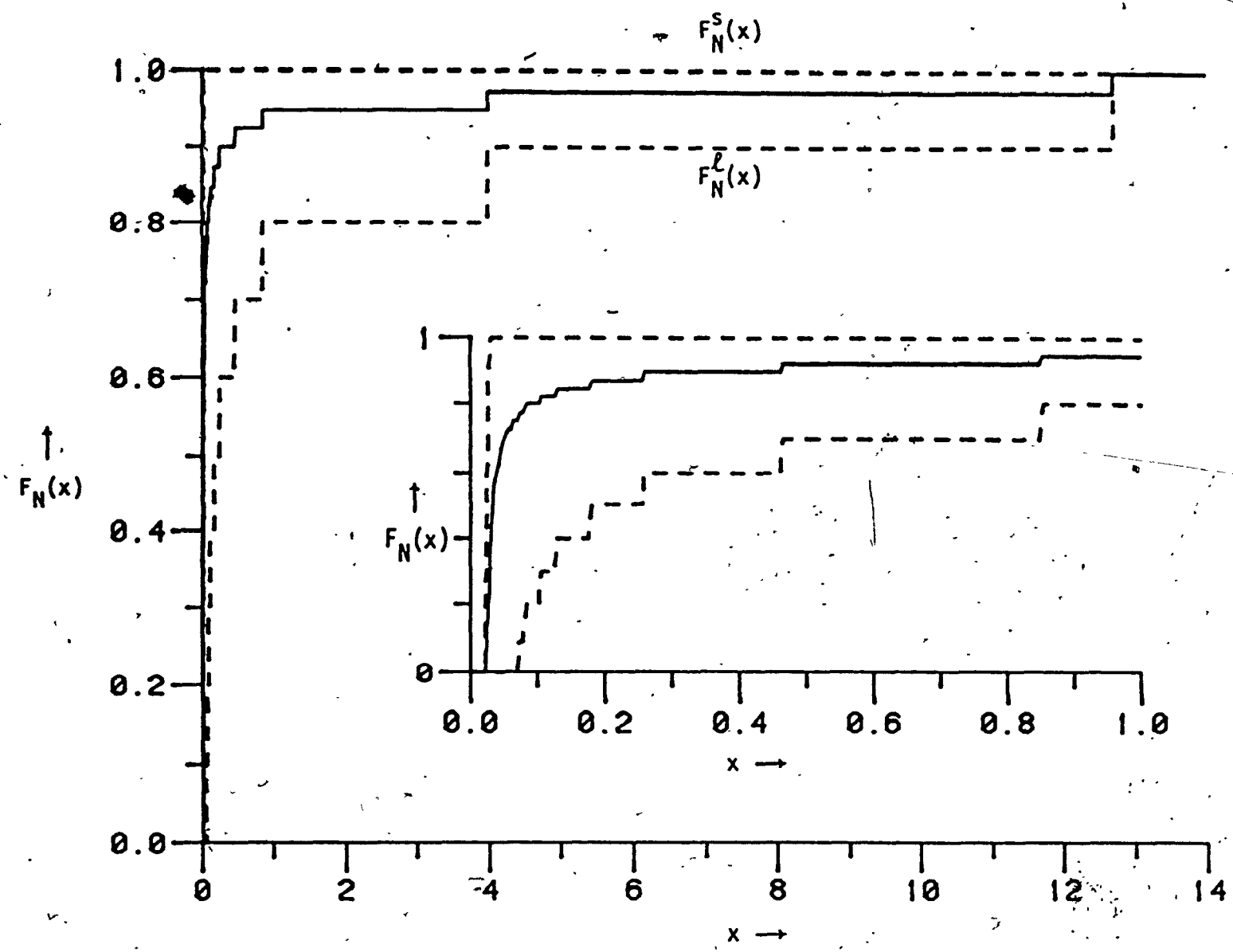

Figure 3.5 Distribution Function $F_N(x)$ and Bounds, $F_N^s(x)$ and $F_N^\ell(x)$. Covariance Example V.

Note: Region for $x \in [0,1]$ expanded to show finer detail.

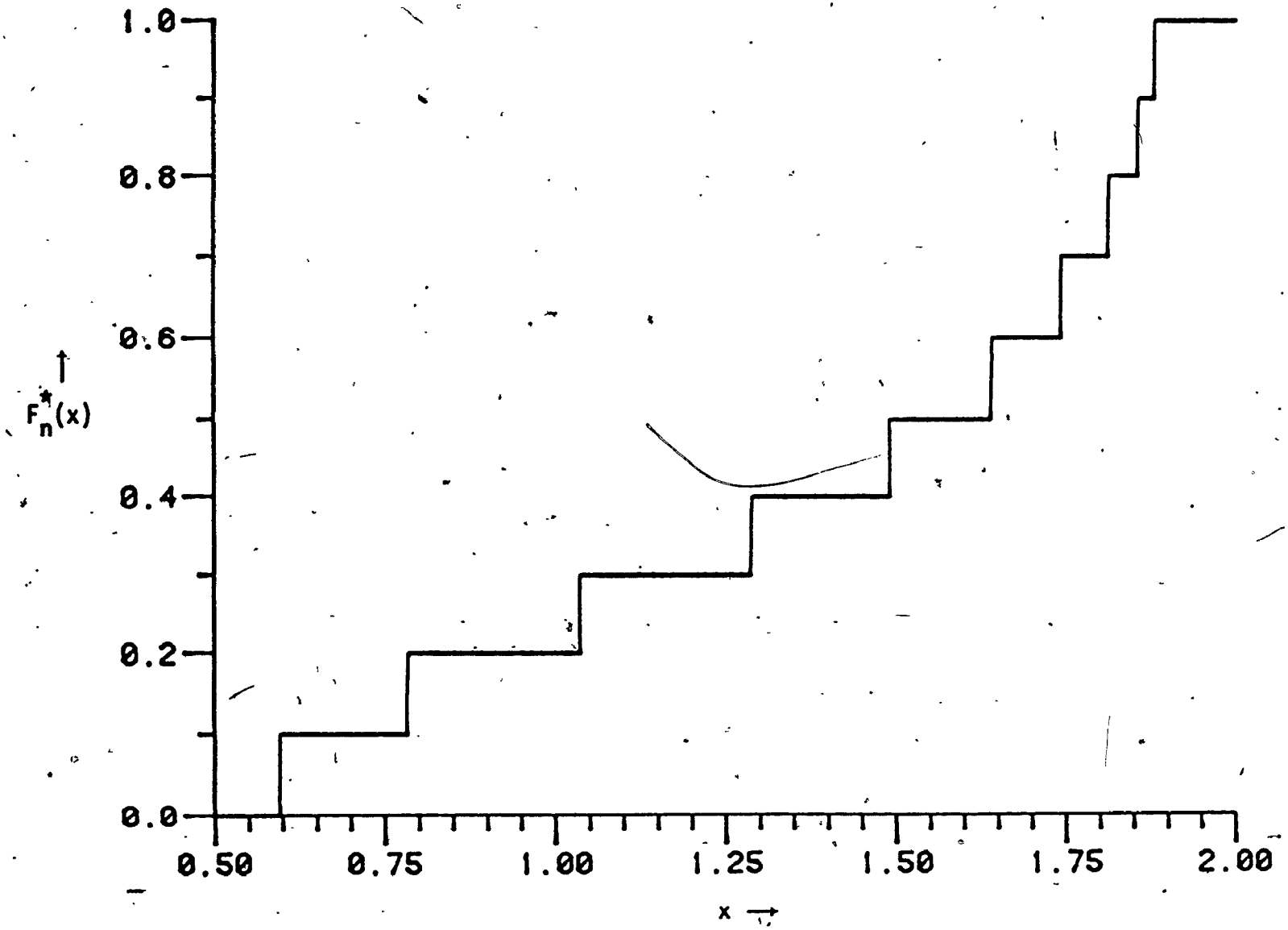

Figure 3.6 Distribution Function $F_n^*(x)$. Rows of A Selected as $\underline{a}_i = \underline{\delta}_i$ $i = 1,2,\ldots, n$. Covariance Example I.

We are now in as position to make some observations which may prove useful for a better understanding of the feature selection problem. It is obvious from the df of Figure 3.6 for Example I and Theorem 3.1, that this arbitrary choice of A presents more problems than solutions. The reason for this is that some elements of the eigenspectrum $\{\lambda_i\}$ are close to unity in contrast to the results of Theorem 3.1 and deductions thereof for low classification error. Furthermore, it is apparent from Figure 3.1 for the same example, that a reasonable selection of features can be made by restricting the choice to the eigenvectors associated with eigenvalues which are quite distant from unity, due to the considerable clustering of large eigenvalues. Examples of such a selection are the eigenvectors which achieve the bounds of (3.18) as shown in Figure 3.1. In this case, the eigenvectors which provide the upper bound $F_N^{\ell}(x)$ appear to be a better choice with associated eigenvalues clustered about 1.85, as opposed to the eigenvectors providing the lower bound $F_N^{s}(x)$ for which the clustering of eigenvalues occurs at about 0.54. It is shown in the sequel that the optimal choice of features, in general, leads to a mix of the eigenvectors associated with the two extremal sets of eigenvalues.

The problem of finding the best subset of possible transformation A has been treated by Okamoto [3.16]. A discussion of his results is presented below in view of the above distribution functions and their bounds.

Theorem 3.3. (Okamoto, 1961). For the n-dimensional feature space, a basis $\underline{y}^*$ which minimizes the probability of classification error

$P_e(n,\underline{\lambda}^*)$ is given by <u>one</u> of the $(n+1)$ variates $(x_1^*, x_2^*, \ldots, x^*_p, x^*_{p+(N-n)+1}, \ldots, x_N^*)$ where $0<p<n$. The subsequence x_i^* where $1<i<p$ are the coordinates corresponding to the p largest eigenvalues of the quadratic pencil of the form $\underline{x}^{*H}[\Lambda-\lambda I_N]\underline{x}^*$, and, if $p=0$, no coordinates from this subsequence are chosen; the subsequence $x^*_{i+(N-n)}$ where $(p+1)<i<n$ are the coordinates corresponding to the $(n-p)$ smallest eigenvalues of the same pencil, and, if $p=n$, no coordinates from the subsequence are chosen.

<u>Proof</u>. The underlying assumptions in this context are that the stochastic data vector $\underline{x} \sim N(\underline{0}_i, R_i)$ under H_i $i=1,2$, and that $AA^H=I_n$; thus, $\underline{y} \sim N(\underline{0},I_n)$ under H_1 and $N(\underline{0},A\Lambda A^H)$ under H_2. Forming the product,

$$\underline{y}^* = L^{*H}\underline{y}$$

where L^* is now a unitary transformation, we then obtain the stochastic MVN feature vector $\underline{y}^* \sim N(\underline{0},I_n)$ under H_1 and $N(\underline{0},\Lambda^*)$ under H_2. We recall that the roots of the quadratic forms $\underline{x}^{*H}[\Lambda-\lambda I_N]\underline{x}^*$ and $\underline{y}^{*H}[\Lambda^*-\lambda I_n]\underline{y}^*$ are $\lambda_1>\lambda_2>\ldots>\lambda_N$ and $\lambda_1^*>\lambda_2^*>\ldots>\lambda_n^*$, respectively. It has been established by Theorem 3.1 that $P_e(n;\underline{\lambda}^*)$ grows smaller as the λ_i^* are more distant from unity. From Theorem 3.2, we note that $P_e(n;\underline{\lambda}^*)$ is minimized at $\lambda^* = (\lambda_1,\lambda_2,\ldots,\lambda_n)^T$ if all $\lambda_i>1$, or at $\underline{\lambda}^* = (\lambda_{(N-n)+1},\lambda_{(N-n)+2}, \ldots,\lambda_N)^T$ if all $\lambda_i<1$. In the general case, when some λ_i are below unity and some above, the minimum point $\underline{\lambda}^*$ selects a mix of the largest and the smallest λ_i; some largest $\lambda_i>1$ and some smallest $\lambda_i<1$ such that a total of n are selected. The solution is determined uniquely if the number of largest (or smallest) λ_i selected is p (or $(n-p)$). We have $p=n$ when all $\lambda_i>1$, or

$p=0$ when all $\lambda_i<1$. ∎

Theorem 3.3 indicates that the optimal choice for the rows of A is one of the following possible $(n+1)$ subsets of eigenvectors associated with the matrix pair (R_1,R_2),

$$\underline{a}_i^T = \begin{cases} \underline{\ell}_i & i\in i^{\ell}(p) \\ \underline{\ell}_i & i\in i^{s}(p) \end{cases} \tag{3.19a}$$

such that,

$$\sigma[i^{\ell}(p) \quad i^{s}(p)] = n \tag{3.19b}$$

where $i^{\ell}(p) = \{1,2,\ldots,p\}$, $i^{s}(p) = \{p+(N-n)+1,\ p+(N-n)+2,\ldots,N\}$ with $p\ [0,n]$ and $\sigma[\bullet]$ denotes the order of the indicated set. Using the choice (3.19) for A in (3.6), we note that $L^* = I_n$, i.e., $\underline{y}^*$ of (3.7) is exactly the same as $\underline{y}$ of (3.6); thus, (3.8) becomes,

$$P_e(n;\underline{\lambda}^*) = \pi_1 I_1(n;\underline{\lambda}^*) + \pi_2 I_2(n;\underline{\lambda}^*) \tag{3.20a}$$

where,

$$I_1(n;\underline{\lambda}^*) = \text{Prob}\ \{\sum_{i\ J} (1-\frac{1}{\lambda_i})z_i^2 > k(\underline{\lambda})\} \tag{3.20b}$$

$$I_2(n;\underline{\lambda}^*) = \text{Prob}\ \{\sum_{i\ J} (\lambda_i-1)z_i^2 < k(\underline{\lambda})\} \tag{3.20c}$$

with,

$$k(\underline{\lambda}^*) \triangleq \sum_{i\ J} \ln \lambda_i + 2\ln(\frac{\pi_1}{\pi_2}) \tag{3.20d}$$

and,

$$J \equiv i^{\ell}(p) \cup i^{s}(p) \tag{3.20e}$$

The canonical parameter $\underline{\lambda}^*$ is explicitly maintained on the left-hand-side of (3.20), since the expressions are applicable only to the feature space.

An important question that must be addressed is which integer

$p \in [0,n]$ delimits the integer sets $i^{\ell}(p)$, $i^{s}(p)$ such that the probability of classification error $P_e(n,\underline{\lambda}^*)$ is minimized. For an answer to this, we turn to a <u>parameterization</u> of the df $F_n^*(x)$ in terms of the bounds $F_N^s(x)$ and $F_N^{\ell}(x)$. Using the concepts of distribution theory [3.20], consider the one parameter family F^* of <u>extremal df's</u>* given by,

$$F^* = \{F^*_{n;\alpha}(x) \mid F_N^s(1) > \alpha > F_N^{\ell}(1)\} \tag{3.21}$$

Since the df $F_n^*(x)$ has increment α, the parameter α is quantized in steps of $1/n$ and $\alpha = (n-p)/n$ where $p \in [0,n]$ with the family F^* defined for each $\alpha \in [F_N^{\ell}(1), F_N^s(1)]$. The parameter α may assume a <u>maximum</u> of $(n+1)$ values, but for any particular case, the bounds of (3.16) evaluated at unity will generally restrict the number that must be considered as <u>possible</u> solutions to the feature selection problem. Each member $F^*_{n;\alpha}$ of the family F^* is defined assuming $F_N^s(1^-) > F_N^{\ell}(1)$, the typical situation, as follows,

$$F^*_{n;\alpha}(x) = \begin{cases} F_N^s(x) & x < x^s(\alpha) \\ \beta & x^s(\alpha) < x < x^{\ell}(\alpha) \\ F_N^{\ell}(x) & x > x^{\ell}(\alpha) \end{cases} \tag{3.22a}$$

where,

$$\begin{aligned} \beta &= \min\{\alpha, F_N^s(1^-)\} \\ F_N^s(\alpha) &= \min\{x \mid F_N^s(x) > \beta\} \\ F_N^{\ell}(\alpha) &= \max\{x \mid F_N^{\ell}(x) < \beta\} \end{aligned} \tag{3.22b}$$

Figure 3.7 displays the family F^* of extremal df's for Example IV, bearing in mind that each family member $F^*_{n;\alpha} \in F^*$ is a <u>potential</u>

---

* An extremal df is a df corresponding to the p(or (n-p)) largest (or smallest) λ_i, i.e., a df resulting from the choice (3.19)

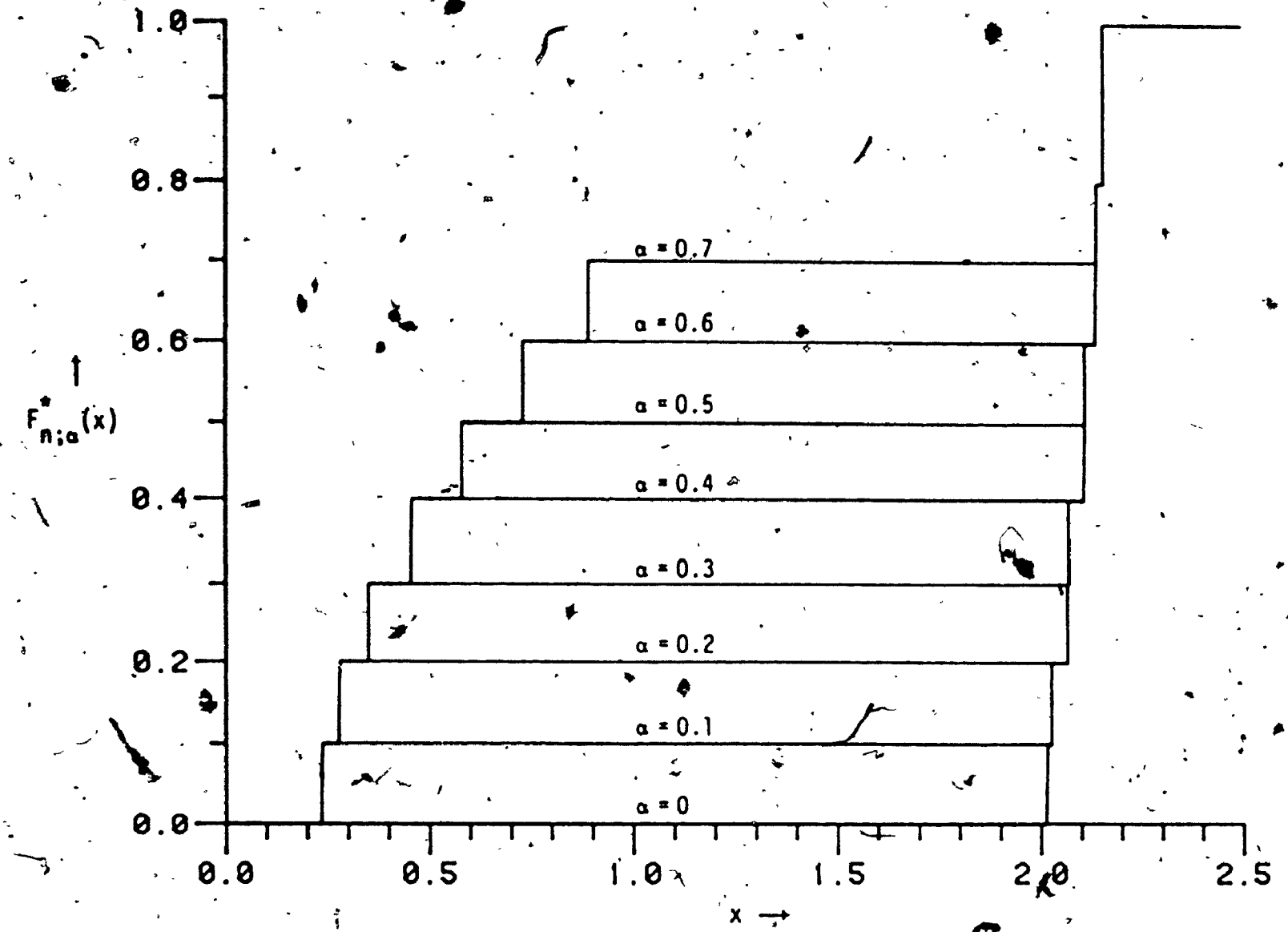

Figure 3.7 Extremal Distribution Function Family F^*.
Covariance Example IV.
Note: Values of $\alpha = 0.8, 0.9, 1.0$ are inadmissible.

solution to the feature selection problem. In the case that $F_N^s(1^-) < F_N^\ell(x)$, the family F^* is obtained from (3.22) with $x^s(\alpha) = x^\ell(\alpha) = 1$. We now present the df counterpart to Theorem 3.3.

Theorem 3.4. For the n-dimensional feature space, a basis for the subspace which minimizes the probability of classification error $P_e(n;\underline{\lambda}^*)$ is given by the eigenvectors associated with the eigenvalues of one member $F^*_{n;\alpha}(x)$ of the family of extremal df's F^* for each df $F^*_n(x)$ which satisfies the eigenvalue bounds of (3.17); if $F_N^s(1^-) = 0$, or if $F_N^\ell(1) = 1$, then the solution is determined uniquely, i.e., F^* contains exactly one df.

Proof. It readily follows form the definition of the family F^* given in (3.21) that for each df $F^*_{n;\alpha_0}(x)$ satisfies,

$$\left.\begin{array}{ll} F^*_n(x) < F^*_{n;\alpha_0}(x) & x<1 \\ F^*_n(x) > F^*_{n;\alpha_0}(x) & x>1 \end{array}\right\} \text{ all } x \qquad (3.23)$$

where $\alpha_0 = F^*_n(1)$. Also, from definition (3.21), it is clear that the extremal df $F^*_{n;\alpha_0}(x)$ associated with $F^*_n(x)$ relates to one of the variates $(x_1^*, x_2^*, \ldots, x^*_{p_0}, x^*_{p_0+(N-n)+1}, \ldots, x_N^*)$ referred to in Theorem 3.3. Thus, the extremal df $F^*_{n;\alpha_0}$ results in the selection of the rows $\underline{a}_i$ of A as in (3.22) as the eigenvectors of the pair (R_1, R_2) corresponding to the $n\alpha_0$ (or $n(1-\alpha_0)$) smallest (or largest) eigenvalues. It is clear from Theorem 3.1 that this choice results in a lower $P_e(n;\underline{\lambda}^*)$. The parameter α is quantized in steps of $1/n$, in accordance with the increment size of $F^*_n(x)$, and all admissible values of p are considered; therefore, the probability of classification error $P_e(n;\underline{\lambda}^*)$ is minimized by one member of the family F^*.

The uniqueness follows directly from the definition (3.21) for the case when $F_N^s(1^-) = 0$ or $F_N^\ell(1) = 1$. ∎

We have, thus far, established an optimality in the sense of Theorems 3.3 and 3.4 relative to any df $F_n^*(x)$ which satisfies the eigenvalue bounds of (3.17), but not an absolute optimality of $P_e(n;\underline{\lambda}^*)$ over the eigenspectrum of the pair (R_1, R_2). We may now write the two types of probability of classification error (3.20), in terms of the family F^* of extremal df's, as follows,

$$P_e(n; F^*) = \pi_1 I_1(n; F^*) + \pi_2 I_2(n; F^*) \tag{3.24a}$$

where,

$$I_1(n;F^*) = \text{Prob}\left\{\sum_{i(\alpha)} (1-\frac{1}{\lambda_i})z_i^2 > k(\underline{\lambda})\right\} \tag{3.24b}$$

$$I_2(n;F^*) = \text{Prob}\left\{\sum_{i(\alpha)} (\lambda_i-1)z_i^2 < k(\underline{\lambda})\right\} \tag{3.24c}$$

and,

$$k(\underline{\lambda}) = \sum_{i(\alpha)} \ln \lambda_i + 2\ln(\frac{\pi_1}{\pi_2}) \tag{3.24d}$$

with $\alpha \in [F_N^\ell(1), F_N^s(1)]$, where the index set $i(\alpha)$ is the set of integers which corresponds precisely to that set of eigenvalues associated with the extremal df $F_{n;\alpha}^*$. We are now in a position to develop an explicit expression for the two forms of classification error; however, we shall return to a further optimization of $P_e(n;F^*)$ over the family F^*, or equivalently, over the parameter α.

3.4. PROBABILITY OF CLASSIFICATION ERROR - EXPLICIT FORM

This section deals with the development of an explicit form for the classification error $P_e(n;F^*)$ which is accurate for n finite. We follow the asymptotic approach of Grenander [3.18, 3.21] but bring

to bear more powerful results to accomplish the task. We shall then show in the sequel the manner in which the integer set $i(\alpha)$ may be determined such that the probability of classification error $P_e(n; F^*)$ is absolutely minimized.

Let us first consider the one type of classification error $I_1(n;F^*)$ of (3.24b) with the assumption that all λ_i are distinct. Noting that the variates z_i^2 are χ^2-distributed with one degree of freedom, the total number of degrees of freedom is the sum $n_1 = \underset{i(\alpha)}{tr}[I_n - \Lambda^{-1}]$, where $\underset{i(\alpha)}{tr}[\bullet]$ is a "partial trace," choosing only the diagonal elements consistent with the set $i(\alpha)$. Using the functional form of the χ^2-distribution, with a change of variables, we may write $I_1(n;F^*)$ as,

$$I_1(n;F^*) = c_n \int_{\mathcal{U}} \prod_{i(\alpha)} e^{-nu_i/2} u_i^{-\frac{1}{2}} d\underline{u} \tag{3.25a}$$

where $\mathcal{U}$ is the n-dimensional region of integration, given by,

$$\mathcal{U} = \{u_i | u_i > 0, \ i \in i(\alpha); \ \sum_{i(\alpha)} q_i u_i \geq k/n\} \tag{3.25b}$$

and

$$c_n = (n/2\pi)^{n/2} \tag{3.25c}$$

We have defined $q_i \triangleq 1 - 1/\lambda_i$ and $k \triangleq k(\underline{\lambda})$ in (3.25b). Let,

$$\phi(\underline{u}) = \prod_{i(\alpha)} e^{-u_i/2} u_i^{1/2n} \tag{3.26a}$$

$$\phi(\underline{u}) = \prod_{i(\alpha)} e^{-\gamma u_i/2} u_i^{(\gamma/2n)-1} \tag{3.26b}$$

we have,

$$I_1^{1/n}(n;F^*) = c_n^{1/n} \{ \int_{\mathcal{U}} \phi^{n-\gamma}(\underline{u}) \, \phi(\underline{u}) d\underline{u} \}^{1/n} \tag{3.27}$$

Clearly, $\phi(\underline{u})$ is integrable for $\gamma>0$. The quantity γ introduced in (3.27) is a free parameter, but shall be restricted in the sequel.

We now examine the behavior of $\phi(\underline{u})$ in the n-dimensional region of integration U. The partial derivatives of $\ln[\phi(\underline{u})]$ w.r.t. the u_i show that the <u>unconstrained</u> maximum of $\phi(\underline{u})$ occurs when $u_i=1/n$ $i\in i(\alpha)$. These coordinates, however, do not belong to the region U, since,

$$\sum_{i(\alpha)} q_i u_i = \frac{1}{n}\sum_{i(\alpha)}(1-1/\lambda_i) < \frac{1}{n}\sum_{i(\alpha)}\ln\lambda_i + \frac{2}{n}\ln\left(\frac{\pi_1}{\pi_2}\right) = \frac{k}{n} \qquad (3.28)$$

assuming $\pi_1>\pi_2$. The <u>constrained</u> maximum cannot be attained as $\phi(\underline{u})$ of (3.26a) vanishes when any $u_i=0$; thus, the constrained maximum may be attained only on the <u>simplex</u>,

$$U = \{u_i \mid u_i>0,\ \ i\in i(\alpha);\ \ \sum_{i(\alpha)} q_i u_i = k/n\} \qquad (3.29)$$

A maximization of $\ln[\phi(\underline{u})]$ by the method of Lagrangian multipliers yields the coordinates,

$$u_i^* = \frac{1}{n(1+\mu q_i)} \qquad i\in i(\alpha) \qquad (3.30a)$$

where a factor of 2 has been absorbed in the parameter μ which must satisfy the condition,

$$\sum_{i(\alpha)} q_i u_i^* = \frac{1}{n}\sum \frac{q_i}{1+\mu q_i} = k/n \qquad (3.30b)$$

The solution for μ in (3.30) is facilitated by the introduction of a

functional $u(t)$ defined as,

$$u(t) = \frac{1}{n} \sum \frac{q_i}{1+tq_i} \tag{3.31}$$

The quantity μ may now be determined by $u(t=\mu)=k/n$. The coordinates u_i^* are strictly positive and, thus, the independent variable t in (3.31) must be restricted in a region (t_{min}, t_{max}) with,

$$t_{min} = -[\max_{i(\alpha)} \{q_i^+\}]^{-1} \tag{3.32a}$$

and

$$t_{max} = -[\min_{i(\alpha)} \{q_i^-\}]^{-1} \tag{3.32b}$$

where $\{q_i^+\}$ and $\{q_i^-\}$ denote the subsets of positive and negative q_i, respectively.

An examination of the functional $u(t)$ is important to the development of the theory. The functional $u(t)$ exhibits large discontinuities over the permissible range (t_{min}, t_{max}) corresponding to the points $t=\lambda_i/(1-\lambda_i)$, where $i \in i(\alpha)$; however, the root μ of (3.30) must lie in the interval $(-1,0)$. This observation is attributed to the fact that $\mu_i^*>0$ <u>iff</u> we have $1+\mu q_i>0$; thus, <u>iff</u> $0<\lambda_i<1$, then $\mu<0$, and <u>iff</u> $\lambda_i>1$, then $\mu>-1$. A compound requirement may be stated as $\mu \in (-1;0)$. It is interesting to note that in this interval, $u(t)$ is a continuous and monotonically decreasing function of t. Moreover, there is always a root for each integer set $i(\alpha)$ in this interval such that $u(\mu)=k/n$, noting from (3.24d), that k is also a function of $i(\alpha)$. The question we shall address in the sequel is which root μ, or equivalently, which integer set $i(\alpha)$, i.e., value of α, provides the lowest $P_e(n; F^*)$.

Figures 3.8 and 3.9 for Example I demonstrate the typical nature of the function $u(t)$ in the restricted domain of (t_{min}, t_{max}). Figures 3.8 and 3.9 correspond to the choices $i(\alpha)\equiv i(o)\equiv i^{\ell}(p=n)$ and $i(\alpha)\equiv i(\alpha=0.9)\equiv i^{s}(p=1)\cup i^{\ell}(p=1)$ in (3.31), respectively. Henceforth, we shall use a more convenient and obvious notation that since $n=10$, Figure 8 represents the eigenvalue selection $10^{\ell}/0^{s}$ and Figure 9, the selection $1^{\ell}/9^{s}$. The <u>a priori</u> probabilities are $\pi_1=\pi_2=\frac{1}{2}$ in both figures. The root μ such that $u(t=\mu)=k/n=0.6251$ is $\mu=-.5482$ in Figure 3.8; whereas, in Figure 3.9 the root μ such that $u(t=\mu)=-.2386$ is $\mu=-.4782$. The roots for eigenvalues selections inbetween the extremes represented by these figures fall somewhere inbetween the two values above. We shall see that the eigenvalue selection of Figure 3.8 is the <u>optimum</u> choice for Example I in terms of minimum $P_e(n;F^*)$ and, furthermore, it is <u>not</u> the selection obtained by the Kadota-Shepp method. In addition, asymptotically ($N\to\infty$, n/N fixed) the Kadota-Shepp method requires that the root $\mu=-.5$ for the optimal df. In passing, we also mention that the root μ in <u>dependent</u> on the <u>a priori</u> probabilities (π_1,π_2), e.g., the root μ for the selection of Figure 3.8 such that $u(t=\mu)=k/n=0.8448$ is now $\mu=-.9155$. We summarize the discussion in the form of a lemma.

<u>Lemma 3.2.</u> Let $\phi(\underline{u}^*)$ denote the maximum of the function $\phi(\underline{u})$ in the region $\mathcal{U}$ for the <u>a priori</u> probabilities $\pi_1>\pi_2$. We have,

$$\phi(\underline{u}^*) = n^{-\frac{1}{2}} e^{\frac{1}{2n}\sum_{i(\alpha)} \ln(u_i^*-u_i^*)} \tag{3.33a}$$

where,

$$u_i^* = \frac{1}{1+\mu q_i}, \quad q_i = (1-1/\lambda_i) \quad i\in i(\alpha) \tag{3.33b}$$

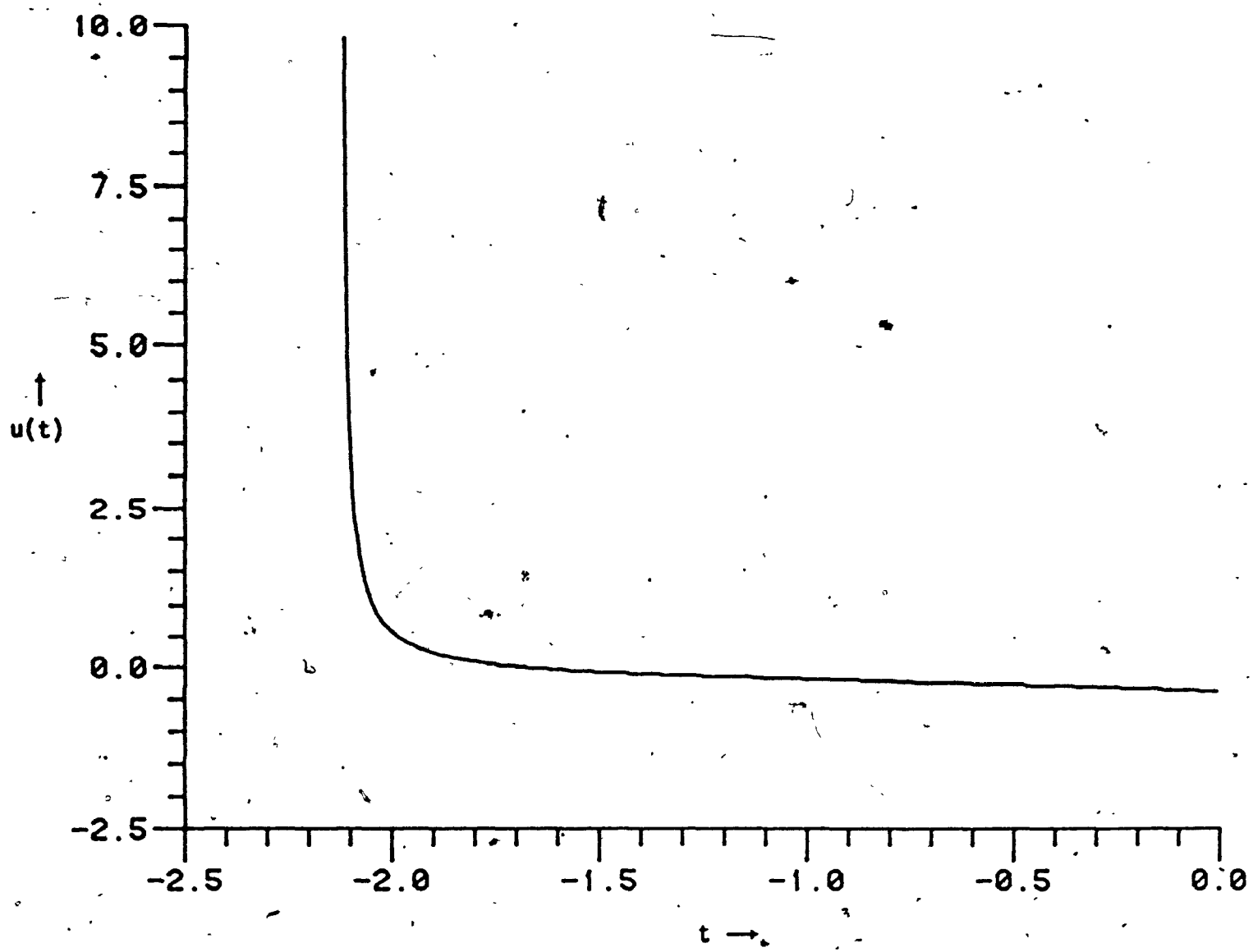

Figure 3.8 Optimum Coordinate Functional u(t). Eigenvalue Selection is the One Largest and Nine Smallest λ_i ($1^{\ell}/9^{s}$). Covariance Example I.

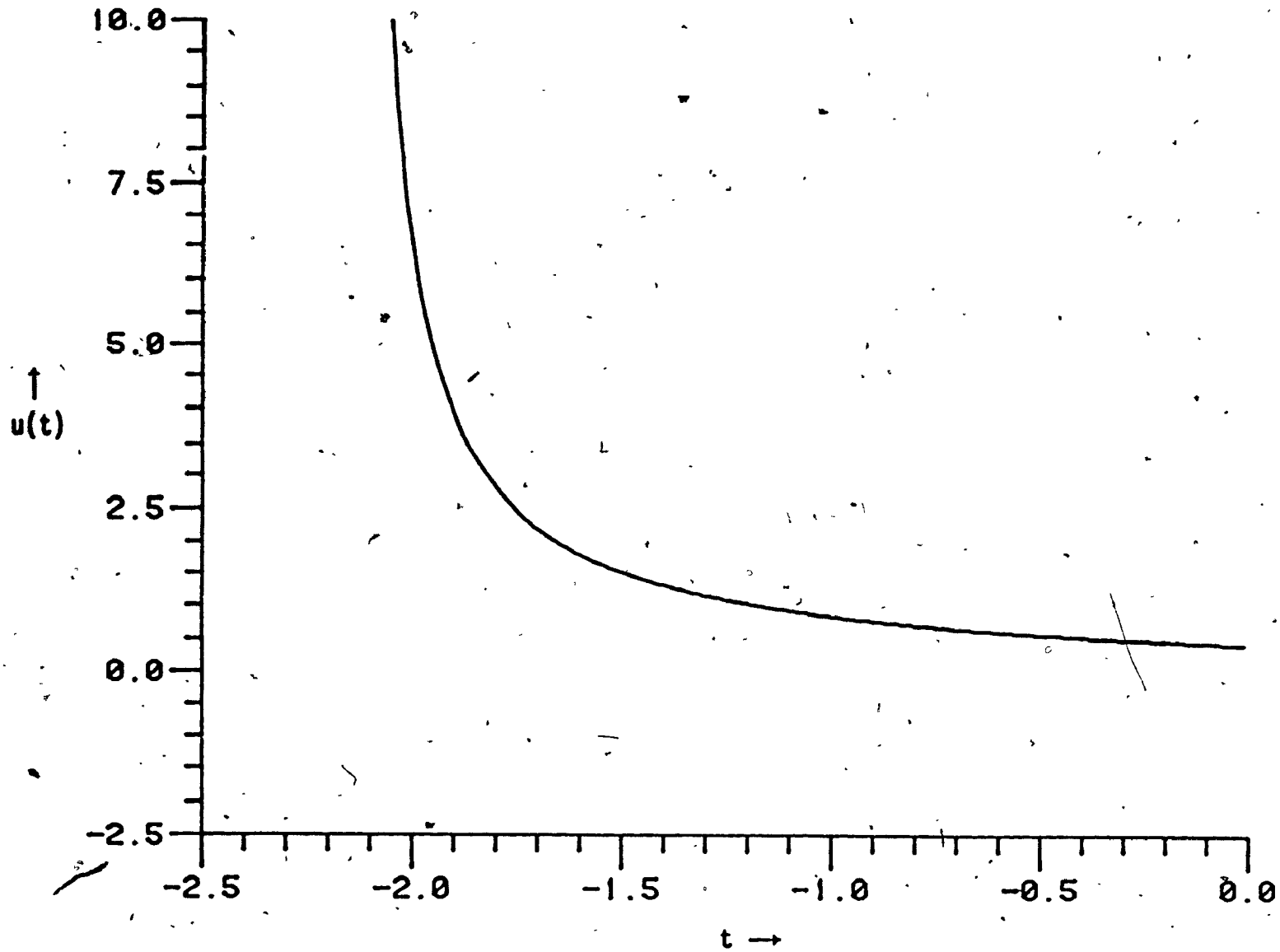

Figure 3.9 Optimum Coordinate Functional u(t). Eigenvalue Selection is the Ten largest λ_i ($10^{\ell}/0^{s}$). Covariance Example I.

with $\mu\in(-1,0)$ is the root of the equation,

$$\frac{1}{n}\sum_{i(\alpha)}\frac{q_i}{1+\mu q_i} = \frac{1}{n}\sum_{i(\alpha)}\ln\lambda_i + \frac{2}{n}\ln\left(\frac{\pi_1}{\pi_2}\right) \qquad (3.33c)$$

<u>Proof</u> From (3.26a), we obtain,

$$\ln[\phi(\underline{u})] = \sum_{i(\alpha)}\left[-\frac{u_i}{2} + \frac{1}{2n}\ln u_i\right]$$

Note that the coordinates u_i^* in (3.33b) have been redefined forconvenience. Using (3.30a) and (3.33b), we have,

$$\ln[\phi(\underline{u}^*)] = \sum_{i(\alpha)}\left[-\tfrac{1}{2}\left(\frac{u_i^*}{n}\right) + \frac{1}{2n}\ln\left(\frac{u_i^*}{n}\right)\right]$$

$$= \frac{1}{2n}\left[\sum_{i(\alpha)}(\ln u_i^* - u_i^*) - \sum_{i(\alpha)}\ln n\right]$$

The last term in the above equality is a constant independent of $i(\alpha)$, i.e.,

$$\sum_{i(\alpha)}\ln n = n\ln n$$

Thus,

$$\ln[\phi(\underline{u}^*)] = \ln n^{-\frac{1}{2}} + \frac{1}{2n}\sum_{i(\alpha)}(\ln u_i^* - u_i^*)$$

and the desired result readily follows. ∎

The new definition of u_i^* introduced in the preceding lemma is employed in order to simply notation. We now utilize the absolute integrability of $\phi^{n-\gamma}(\underline{u})\phi(\underline{u})$ in the region U as we consider only a neighbourhood of $\underline{u}^*$ in expanding the exponent of $\phi(\underline{u})$ in powers of $(\underline{u}-\underline{u}^*)$ up to second order terms. We obtain the following asymptotic

formulae for the classification error using the classical method of Laplace, as described in [3.22] and the refined method of Laplace as described in [3.23].

<u>Theorem 3.5.</u> Fix the parameter $\gamma > 0$ of (3.27) and let γ be sufficiently small, say $\gamma/n < \mathcal{O}(10^{-5})$. Then the following asymptotic formulae represent the two types of classification error as the number of features n, and the data dimension N increase continuously to infinity such that the compression ratio n/N remains constant,

$$I_1^{1/n}(n;F^*) \sim A(\underline{u}^*) \bullet B^{1/n}(\underline{u}^*) \tag{3.34a}$$

$$I_2^{1/n}(n;F^*) \sim A(\underline{v}^*) \bullet C^{1/n}(\underline{v}^*) \tag{3.34b}$$

where,

$$A(\underline{u}^*) = (2\pi)^{-\frac{1}{2}} e^{\frac{1}{2n}\sum_{i(\alpha)}(\ln u_i^* - u_i^*)} \tag{3.34c}$$

$$A(\underline{v}^*) = A(\underline{u}^*) \bullet \left(\frac{\pi_1}{\pi_2}\right)^{1/n} \tag{3.34d}$$

$$B(\underline{u}^*) = \phi(\underline{u}^*/n) \bullet \left[\frac{2\pi}{\frac{1}{2}\sum_{i(\alpha)}(u_i^*/n)^{-2}}\right]^{\frac{1}{2}} \tag{3.34e}$$

$$c(\underline{u}^*) = \phi(\underline{v}^*/n) \bullet \left[\frac{2\pi}{\frac{1}{2}\sum_{i(\alpha)}(v_i^*/n)^{-2}}\right]^{\frac{1}{2}} \tag{3.34f}$$

$$u_i^* = \frac{1}{1+\mu q_i}, \quad v_i^* = u_i^*/\lambda_i \quad i \in I(\alpha) \tag{3.34g}$$

with $q_i = (1-1/\lambda_i)$ and $\mu \in (-1,0)$ is the root of the equation,

$$\frac{1}{n}\sum_{i(\alpha)} \frac{q_i}{1+\mu q_i} = \frac{1}{n}\sum_{i(\alpha)} \ln \lambda_i + 2\ln\left(\frac{\pi_1}{\pi_2}\right) \tag{3.34h}$$

for <u>a priori</u> probabilities $\pi_1 > \pi_2$. Finally, the overall probability of classification error is given by the following asymptotic formula,

$$P_e^{1/n}(n;F^*) \sim A(\underline{u}^*) \bullet \{\pi_1[B(\underline{u}^*) + C(\underline{v}^*)]\}^{1/n} \qquad (3.34i)$$

Proof. The asymptotic relationship of (3.34a) follows directly from Lemma 3.2 and the refined method of Laplace [3.23]. The accuracy of the two types of errors in (3.34a) and (3.34b), and consequently that of the probability of classification error in (3.34h) for $\gamma/n<\mathcal{O}(10^{-5})$ follows from extensive computer simulations. Care must be exercised in fixing the value of γ. This point shall be discussed in Chapter 6. The relationship of (3.34b) follows in an identical manner where the change of variables $v_i^* = u_i^*/\lambda_i$ is incorporated so as to obtain the region of integration corresponding to $I_2(n;F^*)$ in (3.24c). Noting that the total sum of degrees of freedom in the sum of (3.24c) is $n_2 = \underset{i(\alpha)}{tr}[\Lambda - I_N]$, the coordinates of the two extrema are then immediately seen to relate as in (3.34g). The root μ as defined in (3.34h) is common to both sets of coordinates, since both maxima are attained on the same simplex (3.29). This establishes a similar constant term $A(\underline{u}^*)$ in both (3.34a) and (3.34b). The equality of the constant terms is critical to the form (3.34i) and is revealed by an examination of the following difference,

$$\begin{aligned}
&\frac{1}{2n}\sum_{i(\alpha)}(\ln u_i^* - u_i^*) - \frac{1}{2n}\sum_{i(\alpha)}(\ln v_i^* - v_i^*) \\
&= \frac{1}{2n}\sum_{i(\alpha)}\left[\ln\left(\frac{u_i^*}{v_i^*}\right) - (u_i^* - v_i^*)\right] \\
&= \frac{1}{n}\sum_{i(\alpha)}\left[\ln \lambda_i - \frac{q_i}{1+\mu q_i}\right] \\
&= -\frac{1}{n}\ln\left(\frac{\pi_1}{\pi_2}\right) \qquad (3.35)
\end{aligned}$$

Thus, from (3.35), we have (3.34d), and, consequently, the result of (3.34b). This illustrates why the non-constant term of (3.34i) only contains the a priori probability π_1. Of course, it is already known that the value of the root μ of (3.34h) is influenced by both a priori probabilities, π_1 and π_2. ∎

It is noted that Theorem 3.5 provides an asymptotic formula for $P_e(n;F^*)$, not an asymptotic result. The terms approaching unity as n tends to infinity are retained, and the formulae are found to be quite accurate for $n \geq 10$. Some calculations shall be provided in the next section.

We note that the important factor in $P_e(n;F^*)$ of (3.34i), in particular as n grows, is the constant term $A(\underline{u}^*)$, the value of which is dictated by the set $i(\alpha)$, or the df family F^*. There is an extremal df $F^*_{n;\alpha}(x) \in F^*$ and an associated root μ which minimizes $A(\underline{u}^*)$. It is noted from (3.34) that $P_e(n;F^*)$, in the limit, exhibits a geometric decrease as observed by Grenander [3.18, 3.21] and Kazakos [3.24]. This approach to feature selection, using a rather direct application of the Bayesian error probability and its explicit dependence on the extremal df F^* sets these results apart from other commonly used methods, notably that of Kadota et al [3.8].

Explicit dependence of the results on the family F^* may be viewed by rewriting (3.34h) as,

$$\frac{1}{n}\sum_{\lambda_{min}}^{\lambda_{max}} \left[\frac{x-1}{x+\mu(x-1)}\right] f^*_{n;\alpha}(x) = \frac{1}{n}\sum_{\lambda_{min}}^{\lambda_{max}} (\ln x) f^*_{n;\alpha}(x) + \frac{2}{n}\ln\left(\frac{\pi_1}{\pi_2}\right) \quad (3.36)$$

where $\lambda_{min} = \min_{i(\alpha)} \{\lambda_i\} > 0$, $\lambda_{max} = \max_{i(\alpha)} \{\lambda_i\} < tr(R_1^{-1}R_2)$ and $f^*_{n;\alpha}(x)$ is

the empirical pdf, consisting of n Kronecker delta functions in the case of distinct eigenvales. Other components of (3.34), in particular $A(\underline{u}^*)$, which also depends on $i(\alpha)$, may be rewritten in a similar manner. We now briefly discuss the limit point of our results (3.34) in view of the asymptotic result of Grenander [3.18, 3.21].

In the limit, the non-constant terms of (3.34) tend to unity, and to reconcile the seeming difference, we must take into account the possible eigenvalue multiplicity in the constant term $A(\underline{u}^*)$. This may be done easily be replacing $1/n$ in the arguments of $A(\underline{u}^*)$ by a quantity ρ_i such that $\sum_{i=1}^{k} \rho_i = 1$; thus, there are $k<n$ distinct eigenvalues. This affects the integrand of (3.25a) since the χ^2-variates z_i^2 $=1,2,\ldots,k$ of (3.24) now each have $\rho_i n$ degrees of freedom. Also, the coefficient c_n of (3.25c) is now more complicated, representing χ^2-variates of differing degrees of freedom. Application of Stirling's asymptotic expression leads to a coefficient factor of the form $(\prod_{i=1}^{k} / \rho_i^{\rho_i})^{\frac{1}{2}}$, which introduces the additional $e^{\frac{1}{2}}$ in Grenander's expression for $A(\underline{u}^*)$. Also, his coefficient should also contain a factor of $(2\pi)^{k/2n}$, which can account for the quantity $(2\pi)^{-\frac{1}{2}}$ which multiplies $A(\underline{u}^*)$ in (3.34a) and (3.34b). Assuming that the df $F_n^*(x)$ converges, Reimann sums of the type shown in (3.36) may then, in the limit, be expressed as Lebesque integrals. Numerically, the eigenvalues in the neighbourhood $\lambda_i(\lambda_{max}-\lambda_{min})/(2n\lambda_{max})$ of λ_i are considered to be multiple eigenvalues. The distinct eigenvalues are then determined as the average of distinct sets of multiple eigenvalues.

3.5. OPTIMIZATION OF CLASSIFICATION ERROR

In this section we study the behavior of the term $A(\underline{u}^*)$ in the

expression (3.34i) for $P_e^{1/n}(n;F^*)$. The value of $A(\underline{u}^*)$ is determined by the set $i(\alpha)$ and the normalized sum appearing in the exponential argument. Under the stated assumptions, there is an extremal df $F^*_{n;\alpha}(x) = F^*$ and associated root μ which minimizes $A(\underline{u}^*)$. We present the following result. Lemma 3.3. The constant term $A(\underline{u}^*)$ of (3.34) is minimized by selecting the extremal df $F^*_{n;\alpha}(x)$ such that the forward difference $\left|\frac{\Delta J(\underline{u}^*)}{\Delta\alpha}\right|$ is a minimum, where,

$$J(\underline{u}^*) = -\frac{1}{2n}\sum_{i(\alpha)}(\ln u_i^* - u_i^*) \tag{3.37}$$

The forward difference $\Delta J(\underline{u}^*;\alpha) = J(\underline{u}^*;\alpha+\Delta\alpha) - J(\underline{u}^*;\alpha)$, where the increment $\Delta\alpha = 1/n$, and the dependence on α is explicitly included. Proof. The quantity $J(\underline{u}^*)$ appearing in (3.37) is the negative of the exponent of the term $A(\underline{u}^*)$. Taking the forward difference of $J(\underline{u}^*)$ with respect to α using the notations of (3.36) and bearing in mind that the root μ is also a function of α as shown in (3.34h), we obtain,

$$\frac{\Delta J(\underline{u}^*)}{\Delta\alpha} = \frac{\mu(1+\mu)}{2n}\sum_{i^c(\alpha)}\frac{q_i/\lambda_i}{1+\mu q_i} \tag{3.38}$$

where $i^c(\alpha)$ denotes the complement of the set $i(\alpha)$, i.e., the set containing the (N-n) integers not contained in $i(\alpha)$. The relationship of (3.38) presents the discrete counterpart of the variational result in [3.18, 3.21]. A continuous fit to the discrete function of (3.38) exhibits a well-behaved concave function of α. The discrete quantities in (3.38) are seen to range over both positive and negative values; thus, the extremal point, or the point closest to zero is

established by examining $\left|\frac{\Delta J(\underline{u}^*)}{\Delta\alpha}\right|$. ∎

The function (3.38) is evaluated at a maximum of $(n+1)$ points α, where $F_N^s(1) > \alpha > F_N^\ell(1)$. Within the constraints of the values n and N, we shall obtain an extremal df, bearing in mind that this point may not be precisely zero due to the quantization introduced by the finite sample size. The nature of (3.38) is that of information, or negative entropy content of the eigenvalues (eigenvectors) which are not selected, i.e., the eigenvalues indexed by the set $i^c(\alpha)$ [3.25]. The complementary information content is seen to relate to the left-hand-side of the equation for the root μ (3.34h), indexed by $i(\alpha)$. Due to the well-behaved nature of (3.38) and the constraints imposed on α by (3.21), these calculations need not be carried out for a maximum of $(n+1)$ values of α. Figure 3.10 outlines the basic strategy for feature selection and Chapter 5 shall provide a detailed computer-based classifier utilizing this strategy. Figures 3.11 -3.15 display the function (3.38) versus a convenient domain $\alpha p = (n-p)$ for each of the examples of Table 3.1; thus, $\alpha n=0$ and $\alpha n=10$ correspond to the eigenvalue selection $10^f/0^s$ and $0^f/10^s$, respectively, and other choices lie inbetween these two extremes. Due to the restricted range of α, the inadmissible selections arise and are circled.

The performance of the above strategy, called the M-D method, in terms of classification error may be appreciated by comparing the probability of error actually obtained by using this method with the probability of error obtained employing a conventional method, such as the Kadota-Shepp (K-S) method. As mentioned in Chapter 2, the K-S method simply chooses the eigenvectors (rows of A) corresponding to the

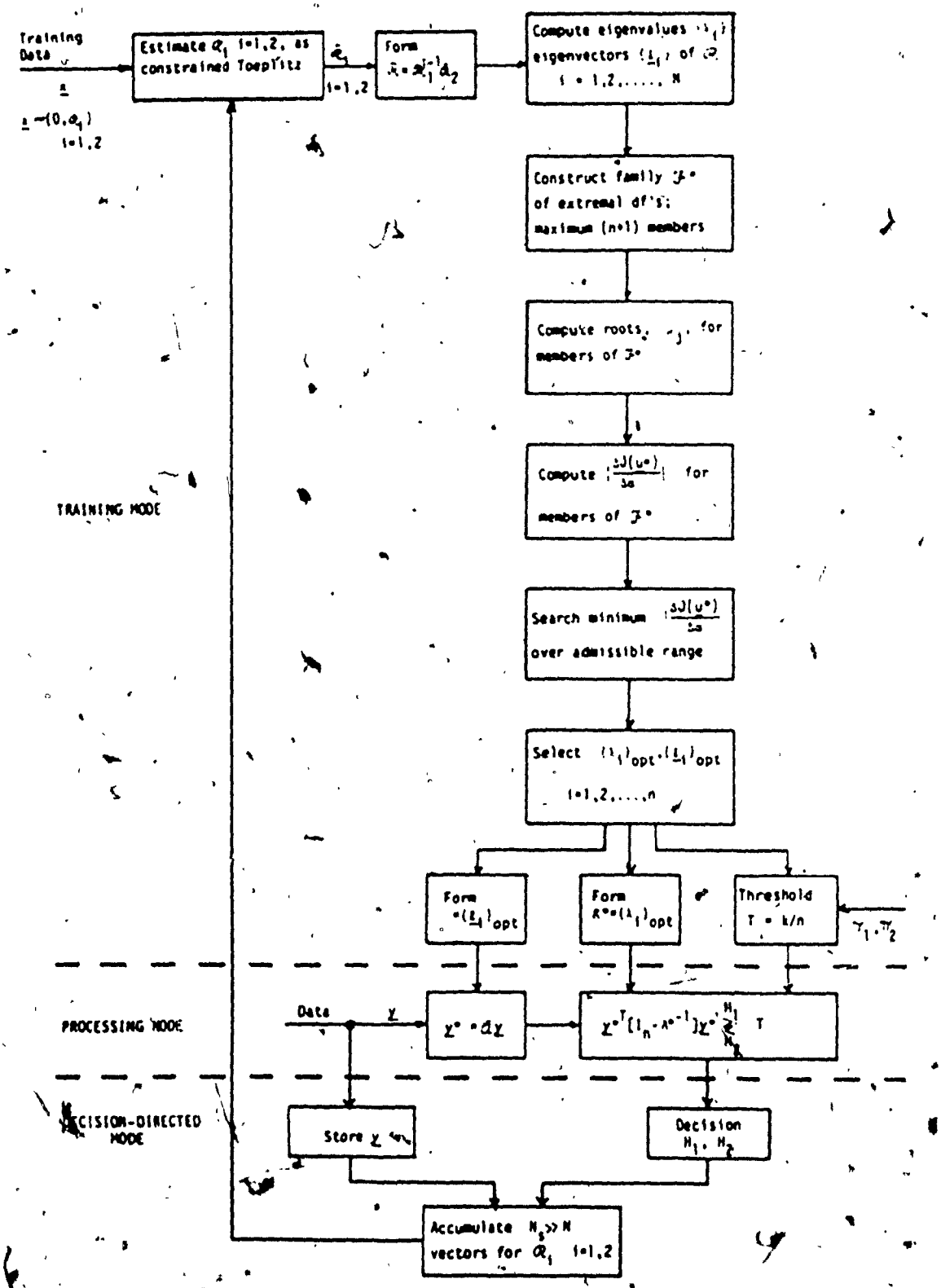

Figure 3.10 The strategy for Optimal Feature Selection

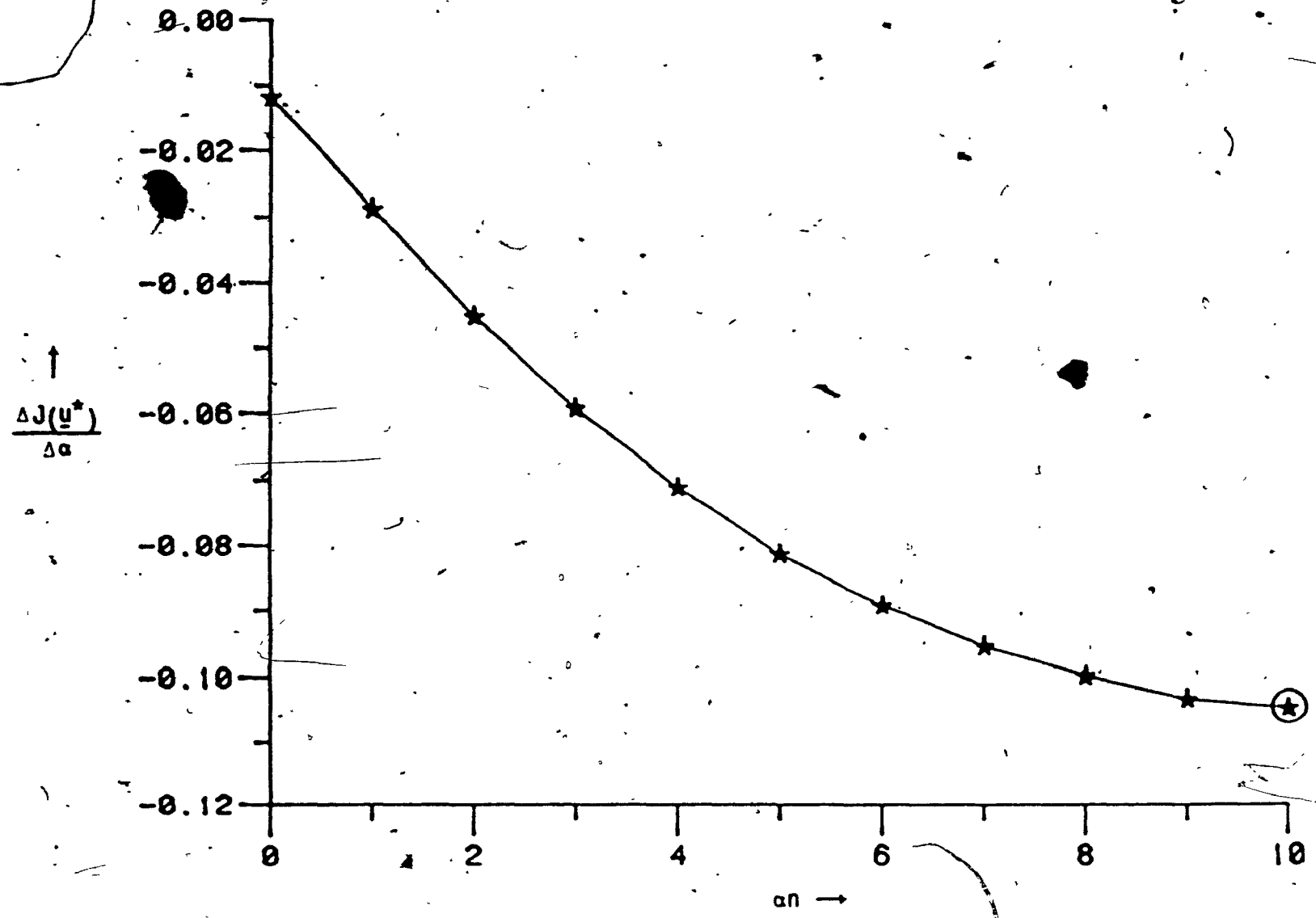

Figure 3.11 Information Functional $\frac{\Delta J(\underline{u}^*)}{\Delta \alpha}$.
Covariance Example I.
Note: Circled points are inadmissible.

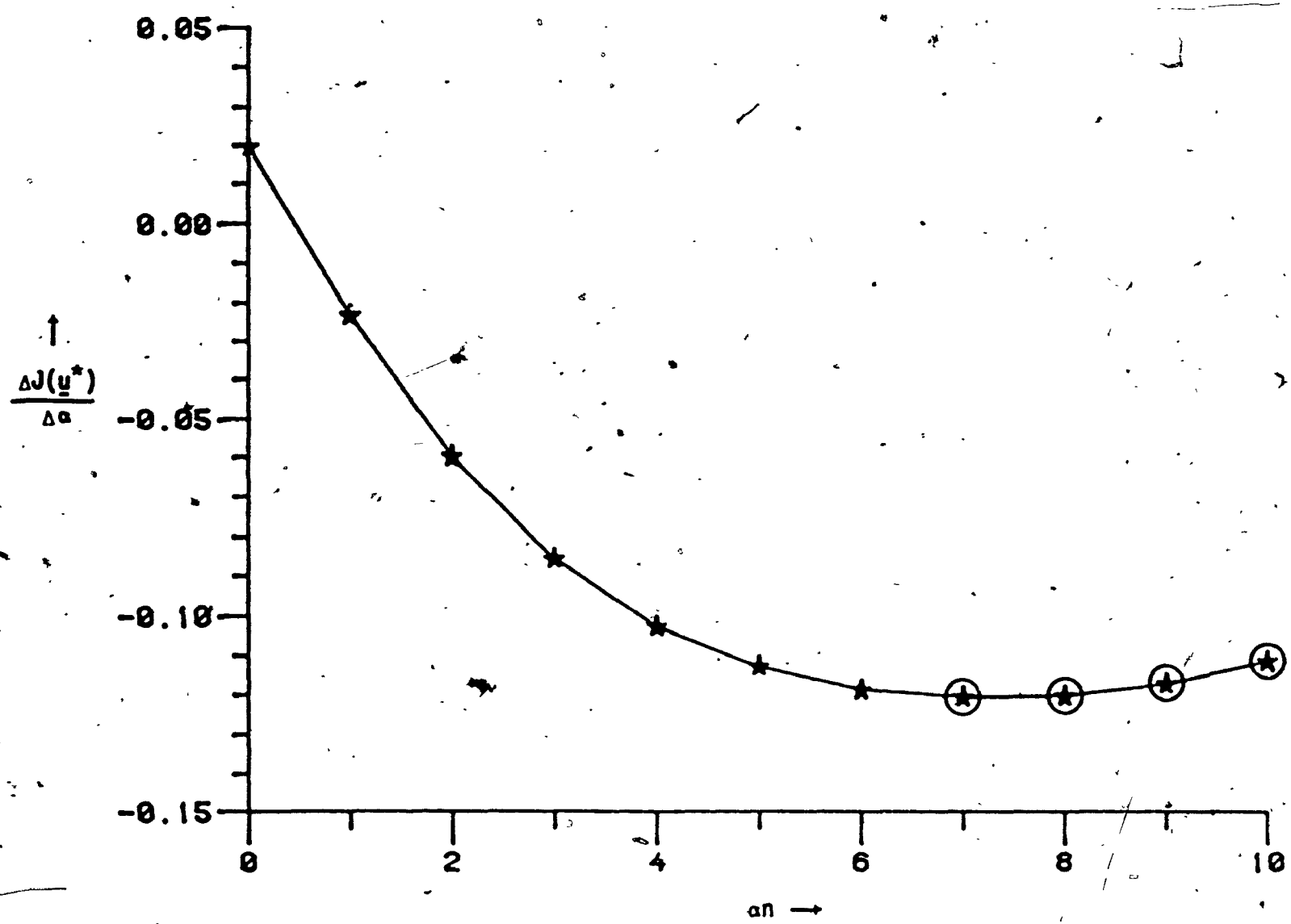

Figure 3.12 Information Functional $\frac{\Delta J(\underline{u}^*)}{\Delta \alpha}$.
Covariance Example II.
Note: Circled points are inadmissible.

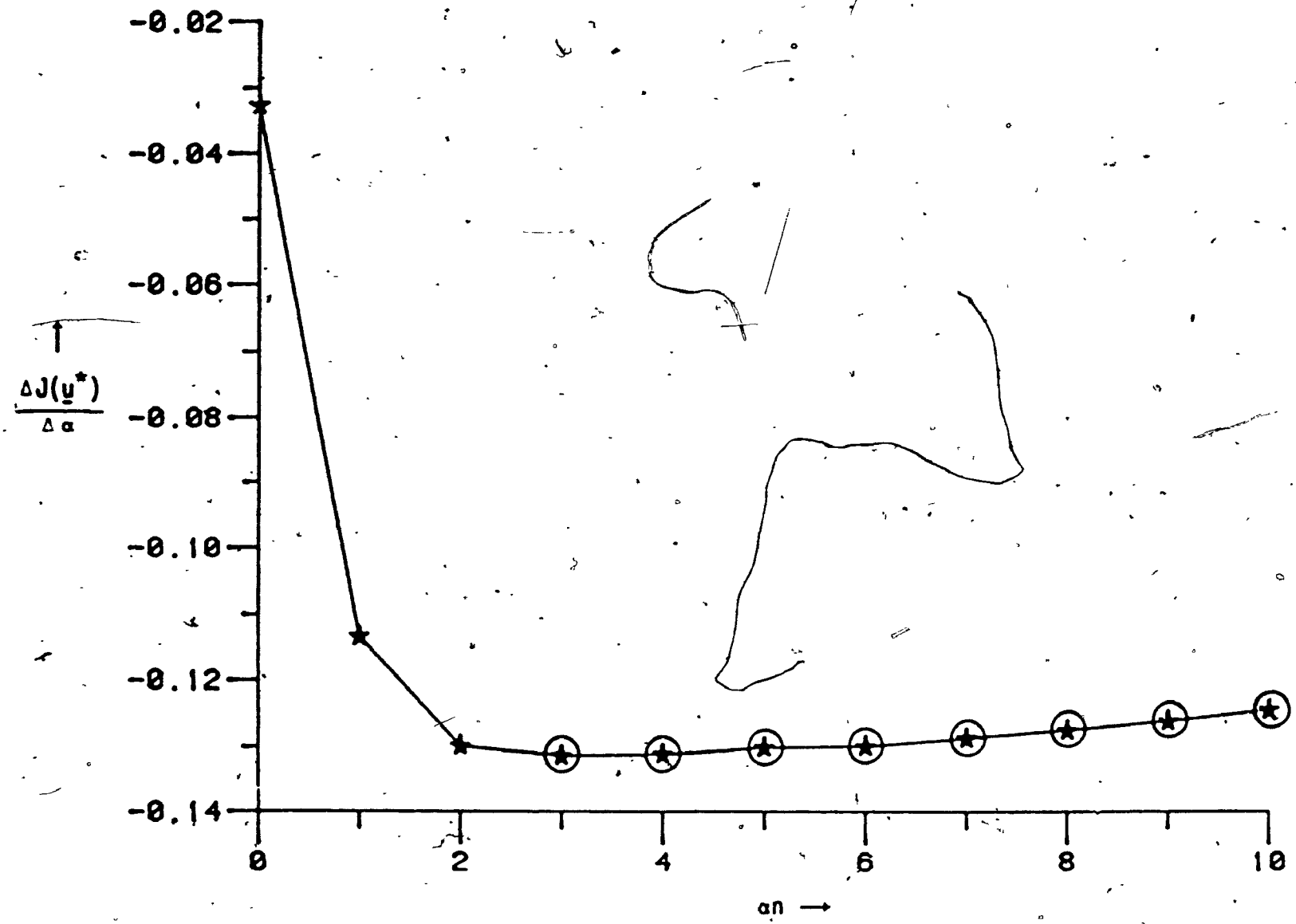

Figure 3.13 Information Functional $\frac{\Delta J(\underline{u}^*)}{\Delta \alpha}$.
Covariance Example III.
Note: Circled points are inadmissible.

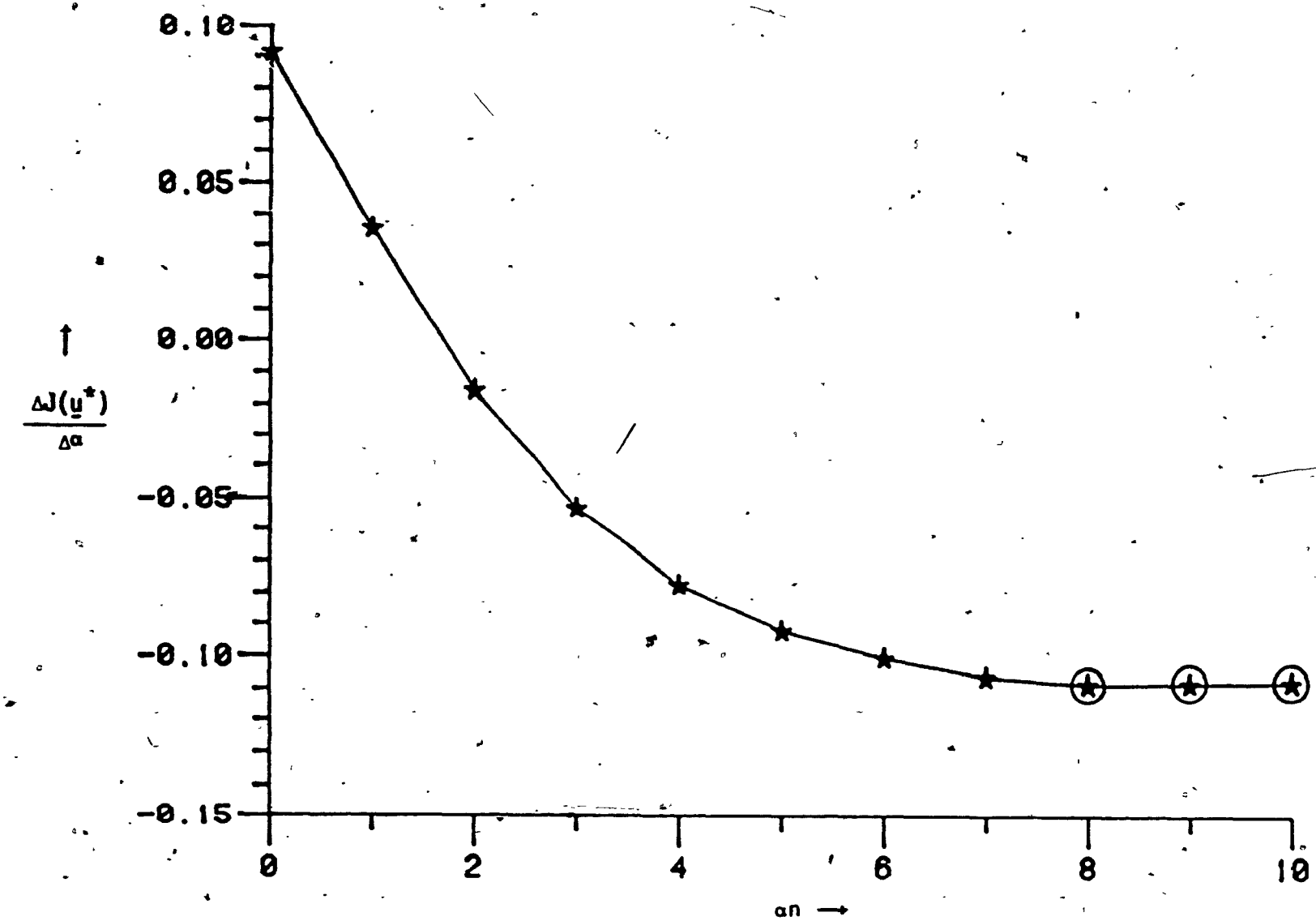

Figure 3.14 Information Functional $\frac{\Delta J(\underline{u}^*)}{\Delta\alpha}$.
Covariance Example IV.
Note: Circled points are inadmissible.

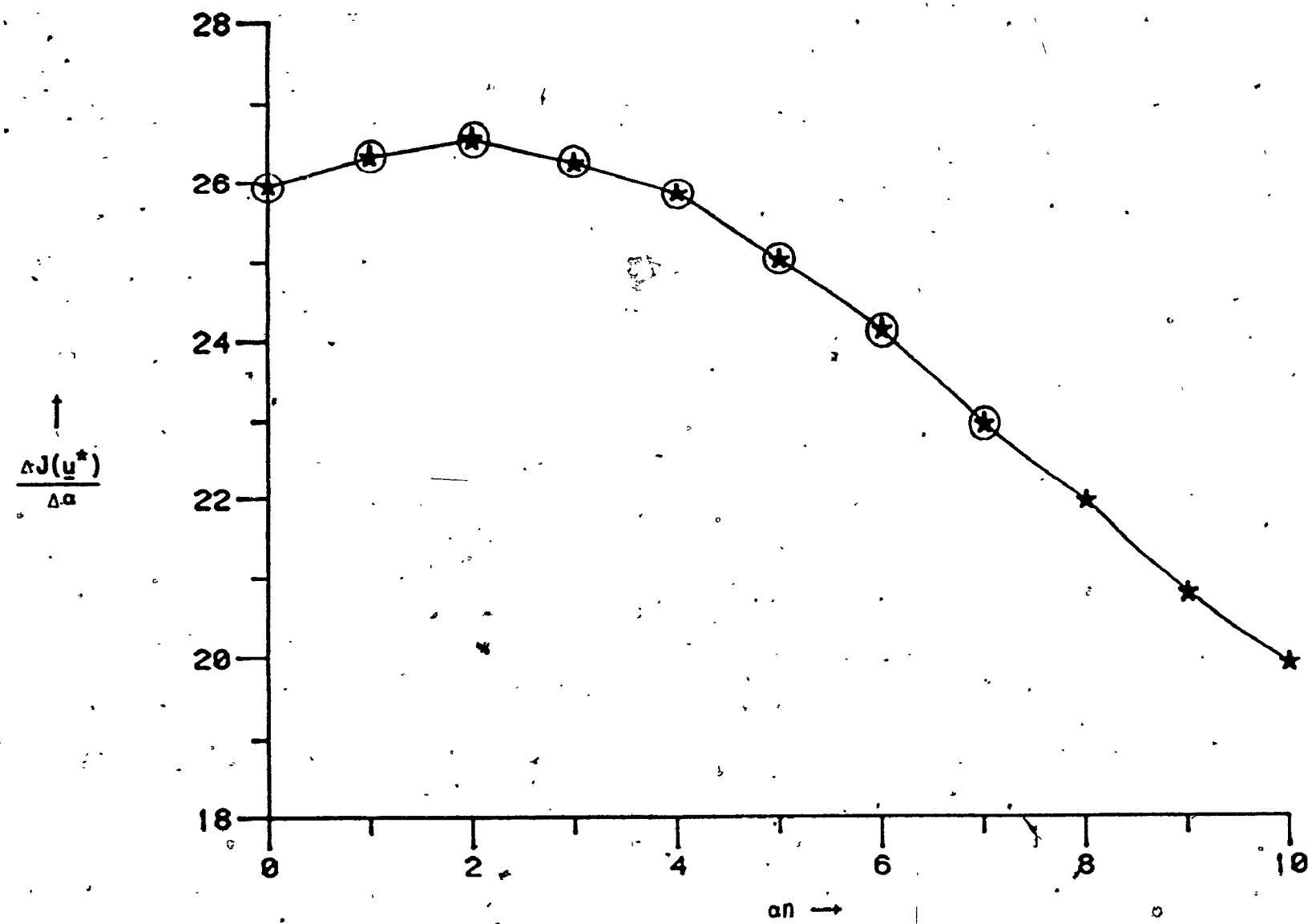

Figure 3.15 Information Functional $\frac{\Delta J(\underline{u}^*)}{\Delta \alpha}$.
Covariance Example V.
Note:

eigenvalues of the matrix pair (R_1, R_2) such that $\lambda_i + 1/\lambda_i$ is maximum. Table 3.2 presents a comparison of the probability of classification error $P_e(n)$ for N=40 and n=10 with $\pi_1=\pi_2=0.5$ for the M-D and K-S methods for each of the covariance pair examples of Table 3.1. The root μ obtained for each method from (3.34h) is also included in Table 3.2.

TABLE 3.2

$P_e(n)$ For New M-D Method and Conventional K-S Method With Respective Eigenvalue Selections; N=40, n=10, $\pi_1=\pi_2$

COVARIANCE EXAMPLE	EIGENVALUE SELECTION M-D	EIGENVALUE SELECTION K-S	ROOT, μ M-D	ROOT, μ K-S	$P_e(n)$ M-D	$P_e(n)$ K-S
I	$10^\ell/0^s$	$9^\ell/1^s$	-.5525	-.5425	6.07×10^{-8}	6.16×10^{-7}
II	$10^\ell/0^s$	$10^\ell/0^s$	-.605	-.605	2.80×10^{-8}	2.80×10^{-8}
III	$10^\ell/0^s$	$9^\ell/1^s$	-.6025	-.5725	2.91×10^{-8}	3.18×10^{-7}
IV	$8^\ell/2^s$	$6^\ell/4^s$	-.4875	-.4625	2.35×10^{-7}	1.44×10^{-6}
V	$0^\ell/10^s$	$0^\ell/10^s$	-.2500	-.2500	8.33×10^{-11}	8.33×10^{-11}

The features (eigenvalues) selected by the M-D method are clearly seen, from Table 3.2, to be consistent with Figures 3.11-3.15 by choosing the points at which the magnitude of each curve is closest to zero. It is interesting to note from Table 3.2 that the M-D method is always <u>at least</u> as good as the K-S method, and sometimes better by an <u>order of magnitude</u>, in terms of error probability, $P_e(n)$.

It is noted here that although the selections made by the M-D method are based on a minimization of the term $A(\underline{u}^*)$ of (3.34i) only in (3.37), computer simulation results indicate that the remaining factor $\{\pi\ [B(\underline{u}^*) + C(\underline{v}^*)]\}^{1/n}$ does not significantly influence the shape of the function $J(\underline{u}^*)$. A typical effect of this factor, say, for Example I is that of simply scaling the function $J(\underline{u}^*)$ by 0.92 for $n=10$. Moreover, this factor for the same example is seen to range from 0.92 for n=10 to 0.98 for n=30, and the factor approaches unity, as expected, as the number of features is increased.

The appreciable decrease in error probability achieved by the M-D method implies that the optimum selection of features cannot be obtained simply in terms of the inverse-symmetric function $\lambda_i + 1/\lambda_i$ as the in K-S method. The proper selection of features must be made, as in the M-D method, by examining $P_e(n)$, or $\frac{\Delta J(\underline{u}^*)}{\Delta\alpha}$, for all members of the extremal df family F^*. This is similar to the approach presented in [3.26] to determine the solution to the optimal diversity communications channel problem, where the scheme provides, as an optimal choice, a mix of eigenvalues such that λ_i or $1/\lambda_i$ is as large as possible. We note from Table 3.2 that a different value for the root $\mu\in(-1,0)$ is involved in each optimal selection using the M-D method, and, even asymptotically, μ need not have the same value for different extremal df's. This is a significant and crucial difference between the M-D and K-S methods. The following result helps to amplify this point in the limit.

Lemma 3.4. Let the number of features n and the data dimension N increase continuously to infinity with the compression ratio n/N

fixed. Assume that in the limit the df $F_N(x)$, denoted by $F(x)$, is absolutely continuous with respect to Lebesque measure on $[\lambda_{min},\lambda_{max}]$ and that the limiting pdf $f(x) = \frac{dF(x)}{dx}$ is continuous and positive. Under the stated assumptions, the K-S method <u>requires</u> $\mu=-\frac{1}{2}$ for the optimal extremal df in the family F^*.

<u>Proof.</u> Using the definition (3.22) for the members $F^*_{n;\alpha}(x)$ of the family F^* and the above assumptions, we have $F^s_N(x)\to F^s(x)$ and $F^\ell_N(x)\to F^\ell(x)$, with the limiting df's absolutely continuous and invertible; $\beta = \min\{\alpha, F^s(1^-)\} = \alpha$; and $x^s(\alpha) = F^{s^{-1}}(\alpha)$, $x^\ell(\alpha) = F^{\ell^{-1}}(\alpha)$. The functional (3.38) then becomes,

$$\frac{\partial J(\underline{u}^*)}{\partial\alpha} = \frac{\mu(1+\mu)}{2}\int_{F^{s^{-1}}(\alpha)}^{F^{\ell^{-1}}} \frac{x-1}{x[x+\mu(x-1)]}\,dx \qquad (3.39)$$

After integrating (3.39) by parts and considerable algebraic manipulation, we obtain,

$$\frac{\partial J(\underline{u}^*)}{\partial\alpha} = -\frac{1}{2}\log[-\mu+(1+\mu)x] + \frac{(1+\mu)}{2}\log x - \frac{\mu}{2}\Bigg|_{F^{s^{-1}}(\alpha)}^{F^{\ell^{-1}}(\alpha)} \qquad (3.40)$$

The K-S selection criteria of choosing eigenvectors corresponding to the eigenvalues λ_i such that λ_i+1/λ_i is maximum, implies that $F^{s^{-1}}(\alpha) = 1/F^{\ell^{-1}}(\alpha)$ or, equivalently, we have, in the limit, $x^s(\alpha) = 1/x^\ell(\alpha)$. Now, if the lower integration limit of (3.40) is set in this manner, we find that a <u>necessary</u> condition for $\frac{\partial J(\underline{u}^*)}{\partial\alpha}$ to equal zero is that the root μ <u>must</u> equal $-\frac{1}{2}$. Thus, we see that in the limit, if the solution to (3.34h) is a root $\mu\neq-\frac{1}{2}$, the K-S method will <u>not</u> be optimal.

One final remark on the preceding lemma pertains to the asymptotic behavior of $P_e(n)$. From (3.34i), we note that in the asymptotic case $P_e(n)$ is independent of the a priori probabilities π_1 and π_2. Thus, a statement of the nature of Lemma 3.4 is not possible for finite sample size, as can be surmised from Table 3.2

We now consider an interesting problem to examplify the superior performance in terms of error probability of the M-D method as compared to the K-S method. Let $N=4$, and the eigenvalues of the pair (R_1, R_2) be $\lambda_1=\lambda_2=\rho>1$ and $\lambda_3=\lambda_4=1/\rho<1$. The empirical pdf $f_N(x)$, and the extremal pdf's $f^*_{n;\alpha}(x)$ are given by,

$$f_N(x) = \tfrac{1}{2}\,\delta(x-1/\rho) + \tfrac{1}{2}\,\delta(x-\rho) \tag{3.41a}$$

$$f^*_{n;\alpha}(x) = \alpha\,\delta(1-1/\rho) + (1-\alpha)\,\delta(x-\rho) \quad 0<\alpha<1 \tag{3.41b}$$

Let us choose a data compression of 50%, i.e., $n=2$. There are three possible values of α, viz., $\alpha=0,1,\tfrac{1}{2}$. We obtain, using (3.34h), the following value for the corresponding root, denoted by μ_α, for $\pi_1=\pi_2=\tfrac{1}{2}$,

$$\mu_\alpha = \begin{cases} \mu_0 = \dfrac{(1-\rho)+\rho\ln\rho}{(1-\rho)\ln\rho} & \alpha=0 \\[2ex] \mu_1 = \dfrac{(1-\rho)+\ln\rho}{(1-\rho)\ln\rho} & \alpha=1 \\[2ex] \mu_{\frac{1}{2}} = -\tfrac{1}{2} & \alpha=\tfrac{1}{2} \end{cases} \tag{3.42}$$

Note, from (3.42), that $\mu_{\frac{1}{2}}$ is independent of the value of ρ, and that $\mu_1=-\mu_0-1$, $\mu_{\frac{1}{2}}=(\mu_0+\mu_1)/2$.

The optimal coordinates u^*_i of (3.34g) are calculated for the roots (3.42), and then only the constant term of $P_e(n;\alpha)$, as in (3.34i), is computed to obtain,

$$P_e(n;\alpha) = \begin{cases} P_e(2;0) \cong \left(\frac{1}{2\pi}\right) \frac{\ln\rho}{(\rho-1)} e^{-\left(\frac{\ln\rho}{\rho-1}\right)} & \alpha=0 \\ P_e(2;1) = P(2;0) & \alpha=1 \\ P_e(2;\tfrac{1}{2}) = \left(\frac{1}{\pi}\right) \frac{\sqrt{\rho}}{(\rho+1)} e^{-1} & \alpha=\tfrac{1}{2} \end{cases} \quad (3.43)$$

The K-S method selects two eigenvectors corresponding to the eigenvalues with λ_i+1/λ_i being largest. Assume that λ_1 is increased infintesimally simultaneously with an infinitestimal decrease in λ_4. Then, the K-S method selects λ_1 and λ_4 with error,

$$P_e(2;\tfrac{1}{2}) = \left(\frac{1}{\pi}\right) \frac{\sqrt{\rho}}{(\rho+1)} e^{-1} \quad (3.44)$$

whereas, the M-D method chooses <u>either</u> λ_1 and λ_2 <u>or</u> λ_3 and λ_4 since,

$$P_e(2;0) = P_e(2;1) = \left(\frac{1}{2\pi}\right) \frac{\ln\rho}{(\rho-1)} e^{-\left(\frac{\ln\rho}{\rho-1}\right)} < P_e(2;\tfrac{1}{2}) \quad \rho>1 \quad (3.45)$$

The inequality in (3.45) is strict for $\rho > 1$.

A <u>considerable</u> improvement in classification can be seen if we consider $\rho \gg 1$, e.g., if $\rho=100$, the errors for the K-S and M-D method are approximately 2.32% and 0.73%, respectively. The above result indicates that, for the example considered, selection of <u>either</u> set of features $2^\ell/0^s$ or $0^\ell/2^s$ ($\alpha=0$ or $\alpha=1$) is <u>equivalent</u>, and is <u>superior</u> to a mixed set of features $1^\ell/1^s$ ($\alpha=\tfrac{1}{2}$). In the context of communication theory, $P_e(2;0)$ or $P_e(2;1)$ relates to binary OOK and $P_e(2;\tfrac{1}{2})$ to binary FSK in flat fading. One final remark in regard to the preceding example is that the problem may be generalized to any even value of N to obtain similar error probability results.

3.6. DISCUSSION

In this chapter we have dealt with the problem of testing one weakly stationary Gaussian stochastic process against another. The stochastic processes are assumed to have similar means and different covariances (patterns). We have demonstrated the suboptimality of employing certain statistical distance measures for feature selection as opposed to a more direct application of the Bayesian error probability.

A feature selection scheme, developed by combining classical results on pattern recognition with modern concepts of distribution function theory, minimizes the Bayes error more directly than the conventional schemes. In the development of the theory, we defined the empirical distribution functions (df's) of the eigenvalues of covariance pairs for both the N-dimensional data space, $F_N(x)$, and the n-dimensional feature space, $F_n^*(x)$. This permits a detailed examination of the eigenvalue distribution of one covariance relative to the other covariance. Such an eigenvalue examination, necessary for optimal feature selection, is not permitted by the use of statistical distance measures.

Consistent with the classical thought, there is a family of extremal df's F^*, defined in the n-dimensional feature space. The members of F^* are $F_{n;\alpha}^*(x)$, which are the df's $F_n^*(x)$ parameterized by the quantity α. The family F^* contains a maximum of $(n+1)$ such members, where this number may be restricted to a somewhat lower value according to the eigenspectrum of the covariance pair $(\mathcal{R}_1, \mathcal{R}_2)$. Among the admissible $F_{n;\alpha}^*(x)$ of F^*, there is one which minimizes the

probability of classification error. The key to resolving the feature selection problem then is to determine this member (or the value of α) and select the n data eigenvectors corresponding to the $n\alpha$ smallest eigenvalues and the $n(1-\alpha)$ largest eigenvalues.

The optimal feature selection strategy is developed by following the asymptotic approach of Grenander and using the results of Laplace to obtain an explicit form for the Bayesian error which appears quite accurate for finite sample size. This method leads to an error expression containing variables (coordinates) dependent on the root μ of an equation involving the threshold for hypothesis testing. It is shown that the root μ, for each member of F^*, lies somewhere in a restricted range $(-1,0)$. The Kadota-Shepp (K-S) method for feature selection, developed by examining the extremal points of statistical distance measure, requires that, in the limit, $\mu=-\frac{1}{2}$, i.e., the root μ must lie at the mid-point of the allowable range with no regard for the family F^*. Therein lies the suboptimality of certain statistical distance measures.

An optimization of the classification error expression with respect to the parameter α leads a discrete function which takes the form of an information content (negative entropy) of discarded eigenvalues. The information content of the eigenvalues retained is maximized by minimizing this discrete function. Probability of error results are obtained using the new strategy and the K-S method, for five pairs of Toeplitz covariance matrices. The new scheme always performs at least as good as, and sometimes better, by an order of magnitude than the conventional K-S method.

The complexity of the feature selection scheme presented here questions the feasiblility of the scheme for real-time practical applications, although the most computationally complex part of the scheme, the extraction of eigenvalues and eigenvectors of the pair (R_1, R_2), is also required by the K-S and other methods. For a more detailed answer to this, we present, in Chapter 5, a computational complexity analysis of the scheme in view of its implementation on a mini/micro-computer-based pattern discriminator. To achieve our objectives of Chapter 5, we present some recent results on matrix theory in Chapter 4. It is our belief that the results of Chapter 4 and 5 will lead to an appealing pattern classification system which may prove useful for real-time computer-based image processing and robotics.

REFERENCES

[3.1] S.D. Morgera and L. Datta, "Toward an Fundamental Theory of Optimal Feature Selection: Part I," IEEE Trans. Pattern Anal. and Mach. Intell., vol. PAMI - 6, pp. 601-616, Sept. 1984.

[3.2] S.D. Morgera, and L. Datta, "Optimal Feature Selection: Part I - Theory," Proc. Seventh Intern. Conf. on Pattern Recognition, Montreal, Canada, July 30-Aug. [illegible], 1984.

[3.3] J. Kittler, "Mathemactical Methods of Feature Selection in Pattern Recognition," Int. J. Man-Machine Studies, vol. 7, pp. 609-637, 1975.

[3.4] L. Kanal, "Patterns in Pattern Recognition: 1968-1974," IEEE Trans. Inform. Theory, vol. IT-20, pp. 697-722, Nov. 1974.

[3.5] T. Kailath, "The Divergence and Bhattacharyya Distance Measures in Signal Selection," IEEE Trans. Comm. Tech., vol. COM-15, pp. 52-60, Feb. 1967.

[3.6] G.T. Toussaint, "Comments on the Divergence and Bhattacharyya Distance Measures in Signal Selection," IEEE Trans. Comm. Tech., vol. COM-20, p. 485, 1972.

[3.7] D. Kazakos and P. Papantoni-Kazakos, "Spectral Distance Measures Between Gaussian Processes," IEEE Trans. Automat. Contr., vol. AC-25, pp. 950-959, Oct. 1980.

[3.8] T.T. Kadota and L.A. Shepp, "On the Best Set of Linear Observables for Discriminating Two Gaussian Signals," IEEE Trans. Inform. Theory, vol. IT-13, pp. 278-285, Apr. 1967.

[3.9] D.A. Chesler and R.L. Greenspan, "Comments on Choosing Observables for Discriminating Two Gaussian Signals," IEEE Trans. Inform. Theory, vol. IT-14, pp. 820-822, Nov. 1968.

[3.10] J.T. Tou and R.P. Hedorn, "Some Approaches to Optimal Feature Extraction", in Computer and Information Sciences. vol. 2, J.T. Tou, New York: Academic, 1966.

[3.11] R.H. Shumway and A.N. Unger, "Linear Discriminant Functions for Stationary Time Series," J. Amer. Stat. Ass., vol. 69, pp. 948-956, Dec. 1974.

[3.12] B.D.O. Anderson, J.B. Moore, and R.M. Hawkes, "Model Approximation via Prediction Error Identification," Automatica, vol. 14, pp. 615-622, 1978.

[3.13] A.H. Gray and J.D. Markel, "Distance Measures for Speech Processing," IEEE Trans. Acoust., Speech, and Signal Processing, vol. ASSP-24, pp. 380-391, Oct. 1976.

[3.14] T.L. Grettenberg, "Signal Selection in Communication and Radar Systems," IEEE Trans. Inform. Theory, pp. 265-275, Oct. 1965.

[3.15] F.E. Hohn, Elementary Matrix Algebra. New York: MacMillan, 1964, pp. 347-348.

[3.16] M. Okamoto, "Discrimination for Variance Matrices," Osaka Math. J., vol. 13, pp. 1-39, June 1961.

[3.17] F.R. Gantmacher, The Theory of Matrices, Volume 1. New York: Chelsea, 1959, p. 324.

[3.18] U. Grenander, Abstract Inference. New York: Wiley, 1981, p. 290.

[3.19] A.H. Zemanian, Distribution Theory and Transform Analysis. New York: McGraw-Hill, 1965.

[3.20] P. Antosik, J. Mikusinski and R. Sikorski, The Theory of Distributions, The Sequential Approach. Amsterdam: Elsevier, 1973.

[3.21] U. Grenander, "Large Sample Discrimination Between Two Gaussian Processes with Different Spectra," Annals of Statistics, vol. 2, pp. 347-352, 1974.

[3.22] G.H. Hardy, J.E. Littlewood, and G. Polya, Inequalities. Great Britain: Cambridge University Press, 1952.

[3.23] G. Polya and G. Szego, Problems and Theorems in Analysis, Volume 1. New York: Springer-Verlag, 1976.

[3.24] D. Kazakos, "On Resolution and Asymptotic Discrimination Between Gaussian Stationary Vector Processes and Dynamic Models," IEEE Trans. Automat. Contr., vol. AC-25, pp. 294-296, Apr. 1980.

[3.25] D.A. Bell, Information Theory and its Engineering Applications. London: Pitman, 1953.

[3.26] G.L. Turin, "On Optimal Diversity Reception, II," IRE Trans. Comm. Syst., vol. CS-10, pp. 22-31, Mar. 1962.

APPENDIX 3.A

CLASSICAL METHOD OF LAPLACE [3.23]

Frequently in probability theory, situations arise where one has to deal with expressions which contain many factors and terms and the number of events considered is large. Often such expressions are analytically intractable. The classical method of Laplace provides a mechanism of evaluating definite integrals of complex integrands consisting of many factors and terms one of which may be raised to a large power. Let $p(x)$, $h(x)$, and $f(x) = e^{h(x)}$ be defined over a finite or infinite interval $[a,b]$ satisfying the conditions: a) $p(x)[f(x)]^n = p(x)e^{nh(x)}$ is absolutely integrable over $[a,b]$ for $n>0$, b) $h(x)$ uniquely attains maximum at ζ and the second derivative $h''(x)$ of $h(x)$ exists and is continuous in a neighbourhood of ζ with $h''(x)<0$, c) $p(x)$ is non-zero and continous at $x=\zeta$. Under the stated assumptions, the following asymptotic formula holds as $n\to\infty$,

$$\int_a^b p(x)\ [f(x)]^n dx \sim p(\zeta)e^{nh(\zeta)} \left[-\frac{2\pi}{nh''(\zeta)}\right]^{\frac{1}{2}} \qquad (3.A.1)$$

It must be noted here that in applying the Laplace's method to the present problem, the variables are assumed continuous, since $n\to\infty$. A simple example of the Laplace's method is as follows. Consider the definite integral,

$$\int_a^b e^{-kn(x-\zeta)^2}\, dx \qquad (3.A.2)$$

where the constant $k>0$ and $a<\zeta<b$. Let $y = (kn)^{\frac{1}{2}}(x-\zeta)$, we may reformulate (3.A.2) as,

$$\frac{1}{\sqrt{kn}} \int_{-(kn)^{\frac{1}{2}}(\zeta-a)}^{-(kn)^{\frac{1}{2}}(\zeta-b)} e^{-y^2} dy \qquad (3.A.3)$$

For a,b,ζ,k fixed as $n\to\infty$, the integral converges to $\sqrt{\pi}$. Thus,

$$\int_a^b e^{-kn(x-\zeta)^2} dx \sim \left[\frac{\pi}{kn}\right]^{\frac{1}{2}} \qquad (3.A.4)$$

CHAPTER 4

EFFICIENT PRINCIPAL COMPONENT EXTRACTION FOR PATTERN RECOGNITION FEATURE SELECTION

4.1. INTRODUCTION

This chapter presents a variety of new results leading to efficient principal component (eigenvalue/eigenvector) extraction for two classes of matrices. These results shall assist us in developing a strategy for a computer-based implementation of the feature selection scheme introduced in Chapter 3. It is appropriate at this juncture to briefly discuss the necessity and importance of the results to be presented in relation to feature selection.

Let $\underline{x}$ be an $(N\times 1)$-dimensional data vector with multivariate normal (MVN) distribution, $N(\underline{0}, R_i)$, under hypothesis H_i $i=1,2$ for the binary or two-class hypothesis testing problem. Dimensionality reduction is achieved by applying an $(n\times N)$-dimensional data reducing transformation $A\,(n<N)$ to the data vector $\underline{x}$, resulting in the formation of an $(n\times 1)$-dimensional feature vector $\underline{y}$ given by,

$$\underline{y} = A\,\underline{x} \tag{3.1}$$

The matrix A is assumed to be of rank n. The rows of A are appropriately selected eigenvectors of the covariance matrix pair (R_1, R_2).

Let the $(N\times N)$-dimensional covariance matrices, R_1 and R_2, representing two weakly stationary Gaussian stochastic processes, be

of Topelitz form and assume that R_1 is positive definite. The feature selection process requires the solution of the following determinantal equation,

$$|R_2 - \lambda R_1| = 0 \tag{3.3}$$

or, equivalently,

$$|R_1^{-1}R_2 - \lambda I| = 0$$

Extraction of the principal components of the product matrix $R_1^{-1}R_2$ is the most complex step of the feature selection process, as can be surmised from Figure 3.10. Thus, we study in detail the structure of such a product in order to achieve computational efficiency.

There are two cases of interest, namely, when the elements of the data vector $\underline{x}$ are in the field of real numbers, or in the field of complex numbers. In the case of a real data space, we obtain covariance matrices, R_1 and R_2 which are symmetric Toeplitz. The product matrix $R_1^{-1}R_2$, belongs to the class of centrosymmetric (CS) matrices. CS matrices possess an interesting structure which is investigated in Section 4.2 for efficient principal component extraction. The covariance matrices obtained in the case of a complex data space are of Hermitian Toeplitz form. The product matrix $R_1^{-1}R_2$, in this case, belongs to the class of centrohermitian (CH) matrices. Section 4.3 deals with the structure of the CH matrices with a view toward reducing the computational complexity of principal component extraction.

In addition, we discuss in Section 4.4 a technique for approximating symmetric Toeplitz matrices by circulant matrices for the case when the input data vectors are real. This approximation has been found useful in some situations. Numerical examples are presented to demonstrate the utility of the method.

4.2. ON THE REDUCIBILITY OF CENTROSYMMETRIC MATRICES

This section presents the reducibility results on the class of centrosymmetric (CS) matrices. The discussion here on CS matrices is restricted to results pertaining to the feature selection problem. We begin by presenting some definitions and results in order to develop the terminology and notations for the sequel.

Recall that the covariance matrices, R_1 and R_2, are symmetric Toeplitz for a real data space. It has been shown that the symmetric Toeplitz matrices belong to the class of symmetric centrosymmetric matrices and that the symmetric centrosymmetric matrices have reducible characteristic equations [4.1-4.4]. Let R be an (N×N)-dimensional symmetric centrosymmetric matrix, then R satisfies the following equality [4.1-4.4],

$$R = E_N R E_N \tag{4.1}$$

where E_N is the (N×N)-dimensional contra-identity‡ matrix containing

‡ Also known as reflection, exchange, or permutation matrix.

ones along the cross-diagonal and zeroes elsewhere. The premultiplication of a matrix by E_N results in a permutation of the rows in the reverse order, while the postmultiplication by E_N permutes the columns of the matrix in the reverse order. Note that the condition (4.1) is a necessary, but not sufficient condition for R to be symmetric centrosymmetric [4.4].

Cantoni et al [4.2] and Makhoul [4.3] have demonstrated that the inverse of a non-singular symmetric centrosymmetric matrix is also symmetric centrosymmetric, i.e.,

$$R^{-1} = E_N R^{-1} E_N \tag{4.2}$$

Let us now examine the structure of the product matrix $R_1^{-1}R_2$. The $(N\times N)$-dimensional matrices R_1^{-1} and R_2 both satisfy the necessary condition for symmetric centrosymmetry; thus, we have,

$$R_1^{-1} = E_N R_1^{-1} E_N \tag{4.3}$$

$$R_2 = E_N R_2 E_N \tag{4.4}$$

Taking the matrix product and using (4.3) and (4.4), we obtain,

$$R^{-1}{}_1 R_2 = E_N R_1^{-1} R_2 E_N \tag{4.5}$$

since $E_N E_N = I_N$, where I_N is the $(N\times N)$-dimensional identity matrix. Clearly, the matrix $R_1^{-1}R_2$ exhibits the relationship (4.1) and, since

the sufficiency of (4.1) does not hold, we examine the structure of a matrix R which satisfies (4.1), bearing in mind that R may <u>not</u> necessarily be symmetric centrosymmetric.

<u>Definition 4.1.</u> [4.5-4.7] Let $\mathbb{K}^{N\times N}$ be the class of $(N\times N)$-dimensional matrices such that

$$R \in \mathbb{K}^{N\times N} \ \underline{\text{iff}} \quad = E_N R E_N \tag{4.6a}$$

then $\mathbb{K}^{N\times N}$ is said to be the class of centrosymmetric (CS) matrices.∎

Let the elements of R be denoted by $\{r_{ij} \quad 1\leq i,j\leq N\}$, then, using Definition 4.1, we obtain,

$$r_{ij} = r_{N+1-i,N+1-j} \tag{4.6b}$$

From the above, we have the following result.

<u>Lemma 4.1.</u> The product of symmetric centrosymmetric matrices belongs to the class of centrosymmetric matrices.

The proof is straightforward. ∎

The members of the class of CS matrices possess an interesting structure which we exploit in order to achieve a reduction in the computational complexity for the extraction of principal components.

<u>Lemma 4.2</u> Let $R\in\mathbb{K}^{N\times N}$, then R can be partitioned as follows,

i) $N=2M$ (even order)

$$R = \left[\begin{array}{c|c} A & B \\ \hline E_M B E_M & E_M A E_M \end{array}\right] \tag{4.7a}$$

ii) $N=2M+1$ (odd order)

$$R = \left[\begin{array}{c|c|c} A & \underline{s} & B \\ \hline \underline{t}^T & \rho & \underline{t}^T E_M \\ \hline E_M B E_M & E_M \underline{s} & E_M A E_M \end{array}\right] \qquad (4.7b)$$

where, A and B are $(M\times M)$-dimensional matrices with no particular structure, $\underline{s}$ and $\underline{t}$ are $(M\times 1)$-dimensional vectors, and ρ is a scalar. The matrix E_M is a contra-identity matrix of order M. The proof is straightforward. ∎

<u>Lemma 4.3.</u> Let $R \in \mathbb{K}^{N\times N}$, and if,

i) N is even $(N=2M)$, then the matrices,

$$R = \left[\begin{array}{c|c} A & B \\ \hline E_M B E_M & E_M A E_M \end{array}\right] \quad \text{and} \quad \hat{R} = \left[\begin{array}{c|c} A-BE_M & 0 \\ \hline 0 & A+BE_M \end{array}\right] \qquad (4.8a)$$

are similar.

ii) N is odd $(N=2M+1)$, then the matrices,

$$R = \left[\begin{array}{c|c|c} A & \underline{s} & B \\ \hline \underline{t}^T & \rho & \underline{t}^T E_M \\ \hline E_M B E_M & E_M \underline{s} & E_M A E_M \end{array}\right] \quad \text{and} \quad \hat{R} = \left[\begin{array}{c|c|c} A-EB_M & \underline{0} & 0 \\ \hline \underline{0}^T & \rho & \sqrt{2}\,\underline{t}^T \\ \hline 0 & \sqrt{2}\,\underline{s} & A+BE_M \end{array}\right] \qquad (4.8b)$$

are similar.

<u>Proof</u> i) Let the $(N\times N)$-dimensional matrix L be given by,

$$L = \frac{1}{\sqrt{2}} \left[\begin{array}{c|c} I_M & -E_M \\ \hline I_M & E_M \end{array}\right] \qquad (4.9)$$

where I_M is the (M×M)-dimensional identity matrix. Clearly, L is an orthogonal matrix. Forming the matrix product,

$$L R L^{-1} = \hat{R} = \left[\begin{array}{c|c} A-BE_M & 0 \\ \hline 0 & A+BE_M \end{array}\right] \tag{4.10}$$

establishes the proof for part (i). The corresponding (N×N)-dimensional matrix L for part (ii) is,

$$L = \frac{1}{\sqrt{2}} \cdot \left[\begin{array}{c|c|c} I_M & \underline{0} & -E_M \\ \hline 0^T & \sqrt{2} & \underline{0}^T \\ \hline I_M & \underline{0} & E_M \end{array}\right] \tag{4.11}$$

from which the desired result follows. ∎

We now present some definitions before stating the theorem on efficient principal component extraction of CS matrices.

<u>Definition 4.2</u>. An (N×1)-dimensional vector $\underline{x}$ is said to be symmetric <u>iff</u>,

$$\underline{x} = E_N \underline{x}$$

<u>Definition 4.3</u>. An (N×1)-dimensional vector $\underline{x}$ is said to be skew-symmetric <u>iff</u>,

$$\underline{x} = -E_N \underline{x}$$

<u>Theorem 4.1</u> Let $R \in \mathbb{K}^{N\times N}$ and assume that R, has distinct eigenvalues.

i) Let N be even $(N=2M)$ and R be partitioned as,

$$R = \left[\begin{array}{c|c} A & B \\ \hline E_M B E_M & E_M A E_M \end{array}\right]$$

then the principal components of R may be extracted from the solutions of the following characteristic equations,

$$(A-BE_M)\underline{v}_i = \lambda_i \underline{v}_i \quad 1\le i\le M \tag{4.12a}$$

$$(A+BE_M)\underline{u}_j = \gamma_j \underline{u}_j \quad 1\le j\le M \tag{4.12b}$$

The M linearly independent ($\ell.i.$) skew-symmetric eigenvectors of R corresponding to the eigenvalues λ_i are given by $\underline{x}_i = (1/\sqrt{2})[\underline{v}_i, -E_M\underline{v}_i]^T$, $1\le i\le M$. The other M symmetric eigenvectors of R corresponding to the eigenvalues γ_j are given by $\underline{y}_j = (1/\sqrt{2})[\underline{u}_i, E_M\underline{u}_i]^T$, $1\le j\le M$. Moreover, the set $\{\underline{x}_i, \underline{y}_j \mid 1\le i,j\le M\}$ is a $\ell.i.$ set of eigenvectors of R.

ii) Let N be odd $(N=2M+1)$ and R be partitioned as,

$$R = \left[\begin{array}{c|c|c} A & \underline{s} & B \\ \hline \underline{t}^T & \rho & \underline{t}^T E_M \\ \hline E_M B E_M & E_M \underline{s} & E_M A E_M \end{array}\right]$$

then the principal components of R may be extracted from the solutions of the following characteristic equations,

$$(A-BE_M)\underline{v}_i = \lambda_i \underline{v}_i \qquad 1 \le i \le M \tag{4.13a}$$

$$\left[\begin{array}{c|c} \rho & \sqrt{2}\,\underline{t}^T \\ \hline \sqrt{2}\underline{s} & A+BE_M \end{array}\right] \left[\begin{array}{c} \alpha_j \\ \hline \underline{u}_j \end{array}\right] = \gamma_j \left[\begin{array}{c} \alpha_j \\ \hline \underline{u}_j \end{array}\right] \qquad 1 \le j \le (M+1) \tag{4.13b}$$

The M $\ell.i.$ skew-symmetric eigenvectors of R corresponding to the eigenvalues λ_i are given by $\underline{x}_i = (1/\sqrt{2})[\underline{v}_i, 0, -E_M\underline{v}_i]^T$, $1 \le i \le M$. The other $(M+1)$ symmetric eigenvectors of R corresponding to the eigenvalues γ_j are given by $\underline{y}_j = (1/\sqrt{2})\,[\underline{u}_j, 2\alpha, E_M\underline{u}_j]^T$, $1 \le j \le (M+1)$. Moreover, the set $\{\underline{x}_i, \underline{y}_j \mid 1 \le i \le M,\ 1 \le j \le (M+1)\}$ is a $\ell.i.$ set of eigenvectors of R.

<u>Proof:</u> i) The matrix product of (4.10) is,

$$LRL^{-1} = \left[\begin{array}{c|c} A-BE_M & 0 \\ \hline 0 & A+BE_M \end{array}\right].$$

There exist two non-singular $(M \times M)$-dimensional matrices X and Y such that,

$$X^{-1}(A-BE_M)X = \mathrm{diag}(\lambda_i) \qquad 1 \le i \le M \tag{4.14a}$$

and,

$$Y^{-1}(A+BE_M)Y = \mathrm{diag}(\gamma_j) \qquad 1 \le j \le M \tag{4.14b}$$

Forming an $(N\times N)$-dimensional partitioned matrix Z as,

$$Z = \left[\begin{array}{c|c} X & 0 \\ \hline 0 & Y \end{array}\right] \tag{4.15}$$

then, we have that,

$$Z^{-1}L\,R\,L^{-1}Z = \left[\begin{array}{c|c} \mathrm{diag}(\lambda_i) & 0 \\ \hline 0 & \mathrm{diag}(\gamma_j) \end{array}\right] \quad 1\le i,j\le M \tag{4.16}$$

clearly, the matrix $L^{-1}Z$ diagonalizes R and, therefore, contains the eigenvectors of R as columns. Forming the matrix product,

$$L^{-1}Z = \left[\begin{array}{c|c} X & Y \\ \hline -E_M X & E_M Y \end{array}\right] \tag{4.17}$$

we see that the M eigenvectors of R corresponding to the eigenvectors of the sub-matrix $(A-BE_M)$ are skew-symmetric, and the remaining M eigenvectors of R corresponding to the eigenvectors of the sub-matrix $(A+BE_M)$ are symmetric.

The proof for part (ii) can be readily established in a similar manner. ∎

The restriction imposed in Theorem 4.2 that R must have distinct eigenvalues may be relaxed with certain consequences which are as follows, noting that Lemma 4.3 implies that the scheme proposed in Theorem 4.1 determines the eigenvalues of R with exact multiplicities. Let N be even $(N=2M)$, $\underline{x}$ be an eigenvector of corresponding to the eigenvalue λ, and the vector $\underline{x}$ be partitioned as,

$$\underline{x} = [\underline{x}_1, \underline{x}_2]^T \tag{4.18}$$

where $\underline{x}_1$ and $\underline{x}_2$ are $(M\times 1)$-dimensional vectors. Consider the eigenvalue problem,

$$R\underline{x} = \lambda\underline{x} \tag{4.19a}$$

or, equivalently, using (4.16), we obtain,

$$L\,R\,L^{-1}L\underline{x} = \lambda L\underline{x} \tag{4.19b}$$

where the matrix L is given by (4.9). From (4.9), (4.18) and (4.19b), we have the following uncoupled eigenvalue problems,

$$[(A-BE_M) - \lambda I_M]\underline{y} = 0 \tag{4.20a}$$

$$[(A+BE_M) - \lambda I_M]\underline{z} = 0 \tag{4.20b}$$

where, the $(M\times 1)$-dimensional vectors $\underline{y}$ and $\underline{z}$ are given by,

$$\underline{y} = [\underline{x}_1 - E_M\underline{x}_2] \tag{4.20c}$$

$$\underline{z} = [\underline{x}_1 + E_M\underline{x}_2] \tag{4.20d}$$

The following three cases immediately follow from the above formulation:

a) If λ is an eigenvalue of $(A-BE_M)$ and not of $(A+BE_M)$, we must have $\underline{z}=0$ implying that $\underline{x}$ is skew-symmetric.

b) If λ is an eigenvalue of $(A+BE_M)$ and not of $(A-BE_M)$, we must have $\underline{y}=0$ implying that $\underline{x}$ is symmetric.

c) If λ is an eigenvalue of both $(A-BE_M)$ and $(A+BE_M)$, we may obtain non-trivial solutions for $\underline{y}$ and $\underline{z}$. We may choose $\underline{y}=0$ to obtain a symmetric $\underline{x}$, or choose $\underline{z}=0$ to obtain a skew-symmetric $\underline{x}$; but, if non-trivial solutions for $\underline{y}$ and $\underline{z}$ are selected, the resultant eigenvector $\underline{x}$ of R may neither be symmetric nor skew-symmetric.

Similar conclusions can be arrived at for the case when the order N of the matrix is odd.

Theorem 4.1 demonstrates that the problem of principal component extraction of a matrix $R \in K^{N\times N}$ has been reduced to the problem of principal component extraction of two $(N/2\times N/2)$ matrices, (or one $(N-1)/2 \times (N-1)/2)$ and one $((N+1)/2 \times (N+1)/2)$ matrix) for even (odd) order R. These results lead to nearly a 75% reduction in the multiplicative complexity involved in solving the characteristic equation of a CS matrix. The factorization results presented here provide a reduction in the complexity similar to the savings achieved for symmetric Toeplitz matrices [4.1].

In this section, we have presented computationally efficient results for the principal component extraction of CS matrices. The results are a generalization of the results of [4.2] and a

specialization of the results of [4.6]. These results shall prove useful in the implementation of the feature selection scheme of Chapter 3, in the situation when the data vectors are real. The CS matrices are also encountered in a number of other areas such as antenna theory, mechanical and electrical systems, and quantum physics. A detailed discussion on many other interesting properties along with the examples of the applications in the above areas of CS matrices is presented in [4.5].

4.3. ON THE REDUCIBILITY OF CENTROHERMITIAN MATRICES

This section relates to the efficient principal component extraction for the feature selection scheme of Chapter 3. We study the structure of the product matrix $R_1^{-1}R_2$ for the case when the data vectors contain complex elements and the resulting covariances, R_1 and R_2, are of Hermitian Toeplitz form. Some properties of Hermitian Toeplitz matrices are discussed below to establish the structure of $R_1^{-1}R_2$.

Let R be an (N×N)-dimensional Hermitian Toeplitz matrix. Since R is Hermitian, we have,

$$R = R^H \tag{4.21}$$

where H denotes the complex conjugate matrix transpose. It has been shown that Toeplitz matrices belong to a broader class of persymmetric matrices and satisfy the following equality [4.8],

$$R = E_N R^T E_N \tag{4.22}$$

Thus, if R is a Hermitian Toeplitz matrix, the following condition,

obtained by using (4.21) and (4.22), is satisfied,

$$R = E_N R^* E_N \tag{4.23a}$$

where $*$ denotes the complex conjugate. Let the element of R be denoted by $\{r_{ij} \quad 1 \leq i,j \leq N\}$, then we have,

$$r_{ij} = r^*_{N+1-i,N+1-j} \tag{4.23b}$$

Note that the condition (4.3) is a necessary but <u>not</u> sufficient condition for R to be Hermitian persymmetric [4.9]. The reason for the lack of sufficiency is that the conditions (4.21) and (4.22) must <u>both</u> be satisfied for R to be Hermitian persymmetric, a deduction that cannot be made from (4.23) alone.

In view of the above, we study the structure of the product matrix $R_1^{-1}R_2$. We note that the inverse of a non-singular Hermitian persymmetric matrix is also Hermitian persymmetric; the following correspondences for the Hermitian Toeplitz matrices R_1 and R_2, using (4.23), may be established,

$$R_1^{-1} = E_N R_1^{-1*} E_N \tag{4.24a}$$

and,

$$R_2 = E_N R_2^* E_N \tag{4.24b}$$

Forming the matrix product, using (4.24), we obtain

$$R_1^{-1}R_2 = E_N(R_1^{-1}R_2)^*E_N \tag{4.25}$$

Comparing (4.23) and (4.25), and observing the lack of sufficiency of (4.23) for Hermitian persymmetry, we study the structure of an obviously <u>more general</u> class of matrices which satisfy (4.23) but are <u>not</u> necessarily Hermitian persymmetric.

<u>Definition 4.4.</u> [4.10] Let $\mathbb{P}^{N\times N}$ be the class of $(N\times N)$-dimensional matrices such that,

$$R \in \mathbb{P}^{N\times N} \quad \underline{\text{iff}} \quad R = E_N R^* E_N$$

then $\mathbb{P}^{N\times N}$ is the class of centrohermitian (CH) matrices. ∎

From the above, we present the following result.

<u>Lemma 4.4</u> The product of Hermitian persymmetric matrices is a member of the class of centrohermitian matrices. ∎

The structure of CH matrices is exploited for efficient principal component extraction by using a <u>particular representation</u> for complex matrices. This representation permits us to represent an $(N\times N)$-dimensional complex matrix by a $(2N\times 2N)$-dimensional real matrix for numerical purposes [4.11]. The consequences of such a representation on the eigenvalue problem are briefly discussed in the sequel.

Let R be an (N×N)-dimensional matrix with elements in the field of complex numbers. The matrix R may be written as,

$$R = A + jB \tag{4.26a}$$

where the (N×N)-dimensional real matrices, A and B, are given by,

$$A = \tfrac{1}{2}(R + R^*) \tag{4.26b}$$

$$B = \frac{1}{2j}(R - R^*) \tag{4.26c}$$

The (N×N)-dimensional complex matrix R may be represented by a (2N×2N)-dimensional partitioned real matrix $\hat{R}$,

$$\hat{R} = \left[\begin{array}{c|c} A & -B \\ \hline B & A \end{array}\right] \tag{4.27}$$

By imposing the restriction that the complex matrix $R \in |P^{N\times N}$, we obtain, by using Definition 4.4, the following relationships,

$$A = E_N A E_N \tag{4.28a}$$

$$B = -E_N \, B \, E_N \tag{4.28b}$$

Using (4.28), the matrix $\hat{R}$ may be cast in the form,

$$\hat{R} = \left[\begin{array}{c|c} A & E_N B E_N \\ \hline B & E_N A E_N \end{array}\right] \tag{4.29}$$

The (2N×2N)-dimensional matrix $\hat{R}$ is clearly a member of the class of CS matrices, discussed in Section 4.2.

The centrosymmetric matrices have been shown to possess a reducible characteristic equation; thus, the principal components of $\hat{R}$ may be obtained with a significant reduction of nearly 75% in the multiplicative complexity by using Theorem 4.1. The eigenvectors of may be either symmetric or skew-symmetric, but the question remains as to how to relate the eigenvectors of $\hat{R}$ to those of the original complex matrix R. For an answer to this, we present a brief discussion [4.11].

Let $\underline{x}_i$ be an eigenvector of the (N×N)-dimensional complex matrix R corresponding to the complex eigenvalue λ_i, then the (2N×2N)-dimensional partitioned real matrix $\hat{R}$ has eigenvalues λ_i and λ_i^*, with the corresponding eigenvectors, $[\underline{x}_i, -j\underline{x}_i]^T$ and $[\underline{x}_i^*, -j\underline{x}_i^*]^T$. For each real eigenvalue λ_i and the corresponding eigenvector $\underline{x}_i = \underline{u}_i + j\underline{v}_i$ of R, the matrix $\hat{R}$ has an eigenvalue λ_i of multiplicity two, and the corresponding eigenvectors, $[\underline{u}_i, \underline{v}_i]^T$ and $[-\underline{v}_i, \underline{u}_i]^T$. The problem

of determining which one of the complex conjugate pair of eigenvalues of $\hat{R}$ is an eigenvalue of R is resolved by computing the corresponding eigenvectors of $\hat{R}$. Partitioning a typical eigenvector of $\hat{R}$ as $[\underline{u},\underline{v}]^T$, we see that corresponding to one of the complex conjugate pair of eigenvalues, the vector $\underline{u}+j\underline{v}$ is null; the corresponding eigenvalue is to be discarded. In case of a real eigenvalue of multiplicity two of $\hat{R}$, either one of the corresponding eigenvectors of $\hat{R}$ is an eigenvector of R.

We have shown here the manner in which the problem of extracting the principal components of an (N×N)-dimensional CH matrix may be resolved by solving the characteristic equations of two (N×N)-dimensional real matrices. In the context of an implementation of the feature selection scheme of Chapter 3, the results relate to the situation when the weakly stationary Gaussian stochastic input data vectors to the classifier are complex.

4.4. APPROXIMATION OF TOEPLITZ MATRICES BY CIRCULANTS: A WAY OF IMPROVING COMPUTATIONAL COMPLEXITY

This section presents matrix approximations which may be used to improve algorithm efficiency when the input data vectors are real. Recall that the covariance matrices R_1 and R_2 representing the two weakly stationary stochastic processes are symmetric Toeplitz. A suitable situation for this approximation is when the magnitude of the covariance diminishes sufficiently rapidly with respect to its size N. In this case, covariance matrices are either banded Toeplitz or may be truncated to form banded Toeplitz matrices. However, care must be

exercised while truncating the covariance matrices to ensure that the corresponding processes are being represented reasonably accurately. Numerical examples are included in the sequel to demonstrate the manner in which the matrix truncation may be performed in practice.

Let the (N×N)-dimensional banded Toeplitz matrix R with elements $\{r_{ij} | 1 \le i,j \le N\}$ be of the form,

$$R = \begin{bmatrix} r_{11} & \cdots & r_{1k} & & 0 \\ \vdots & & & \ddots & r_{1k} \\ r_{k1} & & & & \vdots \\ & \ddots & & & \\ 0 & & r_{k1} & \cdots & r_{11} \end{bmatrix} \tag{4.30}$$

i.e., $r_{1\ell} = r_{\ell 1} \equiv 0$ for $\ell > k$. The banded Toeplitz matrix may be approximated by a circulant matrix R^c of the form,

$$R^c = \begin{bmatrix} r_{11} & \cdots & r_{1k} & & r_{k1} & \cdots & r_{21} \\ \vdots & & & \ddots & & & \vdots \\ r_{1k} & & & & 0 & & r_{k1} \\ & \ddots & & & & \ddots & \\ r_{k1} & & 0 & & & & r_{1k} \\ \vdots & & & \ddots & & & \vdots \\ r_{12} & \cdots & r_{k1} & & r_{1k} & \cdots & r_{11} \end{bmatrix} \tag{4.31}$$

We note that cyclic matrices are a special case of CS matrices [4.5]. It can be shown the matrices R and R^c are asymptotically equivalent [4.12]. The motivation for the above approximation is an attractive property of circulant matrices. A circulant matrix can be easily

diagonalized by means of the discrete Fourier transform (DFT) [4.1, 4.13, 4.14]. The eigenvalues of a cyclic matrix may be obtained by computing the DFT of the first row of elements. The orthogonal diagonalization, $\Lambda^c = F^{-1} R^c F$, is performed by selecting the element f_{mn} of the $(N \times N)$-dimensional Fourier matrix F as $f_{mn} = W_N^{mn}$, where $W_N = e^{j2\pi/N}$, the nth primitive root of unity; and F^{-1}, the inverse Fourier matrix with elements $f_{mn}^{-1} = W_N^{-mn}$, $0<m,n<(N-1)$. The columns of F are the eigenvectors of R^c corresponding to the respective eigenvalues in the diagonal matrix Λ^c. The eigenvalues of R^c may be obtained quite efficiently by padding the N elements of first row with zeros to a length $N'=2^k$. Then, for example, the fast Fourier transform (FFT) of the padded sequence may be computed to provide the eigenvalues of R^c. The computational complexity of the procedure for extracting the principal components of a cyclic matrix in this manner is $N'\log_2 N'$.

We study the above results in the context of the feature selection problem. The banded Toeplitz covariance matrices R_1 and R_2 may be both approximated by circulant matrices R_1^c and R_2^c, respectively. Noting that the inverse of a non-singular circulant matrix is a circulant matrix and the product of circulant matrices is also circulant [4.14], we have that the product matrix $R_1^{c^{-1}} R_2^c$ is a circulant matrix. In view of these properties, the approximation of banded Toeplitz matrices by circulant matrices is quite attractive since the principal components of the matrix $R_1^{c^{-1}} R_2^c$ may be extracted with $N'\log_2 N'+N'$ multiplications, where $N'>N$.

The above approximation is applied to the covariance pair examples of Table 3.1. The covariance matrices R_1 and R_2 for each example are truncated to contain only 14 diagonals above and below the principal diagonal. The banded symmetric Toeplitz structures thus obtained are approximated by circulant matrices. It is determined that the approximation is successful for Examples I, II and IV only. The error probability results obtained using such a circulant matrix approximation are shown in Table 4.1. As can be surmised from

TABLE 4.1

$P_e(n)$ with Circulant Matrix Approximation for New M-D Method and Conventional K-S Method; N=40, n=10. Covariance Examples Refer to Table 3.1.

COVARIANCE EXAMPLE	EIGENVALUE SELECTION M-D	K-S	ROOT, μ M-D	K-S	$P_e(n)$ M-D	K-S
I	$10^{\ell}/0^{s}$	$9^{\ell}/1^{s}$	-.5525	-.5425	1.7441×10^{-7}	1.9566×10^{-6}
II	$10^{\ell}/0^{s}$	$9^{\ell}/1^{s}$	-.6125	-.6	7.321×10^{-8}	8.9967×10^{-7}
IV	$8^{\ell}/2^{s}$	$7^{\ell}/3^{s}$	-.49	-.47	3.9995×10^{-7}	1.9115×10^{-6}

Table 4.1, a deterioration in the error performance of the both M-D and K-S feature selection schemes is encountered. However, it is interesting to note that in some instances, for example, in Example I,

the error probability for the M-D method after circulant matrix approximation in still lower than the error probability for the K-S method without any approximation. The root μ for each method is nearly the same with or without the approximation.

The approximation of symmetric Toeplitz covariance matrices by circulant matrices may not always be possible, but whenever a suitable situation is found for the approximation, a significant reduction in the computational complexity is obtained for principal component extraction by using the results discussed above. Some details on the implementation of such a scheme are included in Chapter 5.

4.5. DISCUSSION

This work offers factorization results on centrosymmetric and centrohermitian matrices. The reducibility of the matrices of both classes is discussed in view of an implementation of the feature selection scheme of Chapter 3. The other areas that may benefit from the results presented here are antenna theory, electrical and mechanical systems, estimation and detection, and speech processing.

The principal component extraction step is computationally the most intensive step of the feature selection strategy. A good deal of effort is devoted to this task in order to develop computationally efficient algorithms. The weakly stationary Gaussian stochastic processes to be discriminated are assumed to have Toeplitz covariances R_1 and R_2. The feature selection scheme requires the principal components of the product matrix $R_1^{-1}R_2$. In general, nothing is a priori assumed about the

structure of $R_1^{-1}R_2$ and the general algorithms used for solving the characteristic equation of $R_1^{-1}R_2$ lead to a highly complex computational process; thus, the structure of $R_1^{-1}R_2$ is investigated.

There are two cases of interest, namely, when the stochastic input vectors are real, or complex. We find, in the case of real input vectors, that the covariances R_1 and R_2 are symmetric Toeplitz, and the product matrix $R_1^{-1}R_2$ is a member of the class of centrosymmetric matrices. The centrosymmetric matrices are shown to possess a reducible characteristic equation. This a priori knowledge about the structure of $R_1^{-1}R_2$ introduces nearly a 75% reduction in extracting the principal components of $R_1^{-1}R_2$. In the case of complex stochastic input vectors, we find that the covariance matrices R_1 and R_2 are Hermitian Toeplitz, and the product matrix $R_1^{-1}R_2$ is a member of the class of centrohermitian (CH) matrices. We show the manner in which the characteristic equation of an (N×N)-dimensional CH matrix can be related to the characteristic equations of two (N×N)-dimensional real matrices. This introduces, as in the case of real input vectors, nearly a 75% reduction in the algorithm computational complexity for determining the principal components of $R_1^{-1}R_2$.

We also consider an approximation of Toeplitz matrices by cyclic matrices for the case of real input data vectors. The results presented are quite appealing in view of the fact that the principal component extraction in certain situations may be performed in $N'\log_2 N'+N'$ multiplications, where $N'>N$ and $N'=2^k$.

The results presented in this chapter relate to an efficient implementation of the feature selection scheme. Such an implementation is presented in Chapter 5. It is felt that the implementation of a pattern classifier using this feature selection scheme shall prove useful in computer-based real-time image processing systems and robotics.

REFERENCES

[4.1] S.D. Morgera, "On the Reducibility of Finite Toeplitz Matrices - Applications in Speech Analysis and Pattern Recognition," Signal Processing, vol. 4, pp. 425-443, Oct. 1982.

[4.2] A Cantoni and P. Butler, "Eigenvalues and Eigenvectors of Symmetric Centrosymmetric Matrices," Linear Algebra App., vol. 13, pp. 275-288, 1976.

[4.3] J. Makhoul, "On the Eigenvectors of Symmetric Toeplitz Matrices," IEEE Trans. Acoust., Speech, and Signal Processing, vol. ASSP-29, pp. 868-872, Aug. 1981.

[4.4] L. Datta and S.D. Morgera, "Comment and Corrections on "On the Eigenvectors of Symmetric Toeplitz Matrices"," IEEE Trans. Accoust., Speech, and Signal Processing, vol. ASSP-32, pp. 440-441, Apr. 1984.

[4.5] L. Datta and S.D. Morgera, "On the Reducibility of Centrosymmetric Matrices - Applications in Engineering problems," submitted for review to J. Franklin Institute, Sept. 1984.

[4.6] A.L. Andrew, "Eigenvectors of Certain Matrices," Linear Algebra App., vol. 7, pp. 151-162, 1973.

[4.7] A.L. Andrew, "Solutions of Equations Involving Centrosymmetric Matrices," Technometrics, vol. 15, pp. 405-407, May 1973.

[4.8] S. Zohar, "Toeplitz Matrix Inversion - The W.F. Trench Algorithm," J. Ass. Comput. Mach., vol. 16, pp. 592-602, Oct. 1969.

[4.9] L. Datta and S.D. Morgera, "Further Additions to "Comment and Corrections on 'On the Eigenvectors of Symmetric Toeplitz Matrices'," submitted for review to IEEE Trans. Acoust., Speech, and Signal Processing, Sept. 1984.

[4.10] A. Lee, "Centrohermitian and Skew-Centrohermitian Matrices," Linear Algebra App., vol. 29, pp. 205-210, 1980.

[4.11] J.H. Wilkinson and C. Reinsch, Linear Algebra. New York: Springer-Verlag, 1971.

[4.12] R.M. Gray, "On the Asymptotic Eigenvalue Distribution of Toeplitz Matrices," IEEE Trans. Inform. Theory, vol. IT-18, pp. 725-730, 1972.

[4.13] J. Biemond, J. Rieske and J.J. Gerbrands, "A Fast Kalman Filter for Images Degraded by Both Blur and Noise," IEEE Trans. Acoust., Speech, and Signal Processing, vol. ASSP-31, pp. 1248-1256, Oct. 1983.

[4.14] P.J. Davis, Circulant Matrices. New York: Wiley, 1979.

CHAPTER 5

TOWARD AN IMPLEMENTATION OF OPTIMAL FEATURE SELECTION

5.1. INTRODUCTION

This work deals with the complexity analysis of a computer-based pattern classifier employing the feature selection scheme developed in Chapter 3. The classifier discriminates between two weakly stationary Gaussian stochastic processes. The stochastic processes are assumed to have similar means and different covariances (patterns). The feature selection strategy, outlined in Figure 3.10, shall be referred to frequently and, therefore, for convenience, the same figure is included here as Figure 5.1.

The operation of the pattern classifier, employing the feature selection scheme of Figure 5.1, may be divided into three modes of operation, viz., the training mode, the processing mode, and the decision-directed mode. The training mode of the classifier, used for initializing the system parameters, is the most computationally intensive one. The actual classification of data vectors takes place during the processing mode which is, computationally, the least complex. The system parameters are updated in the decision-directed mode in order to take into account a realistic quasi-stationarity of the patterns. In this mode many of the same functions of the training mode are operative.

A detailed complexity analysis is presented for each of the three

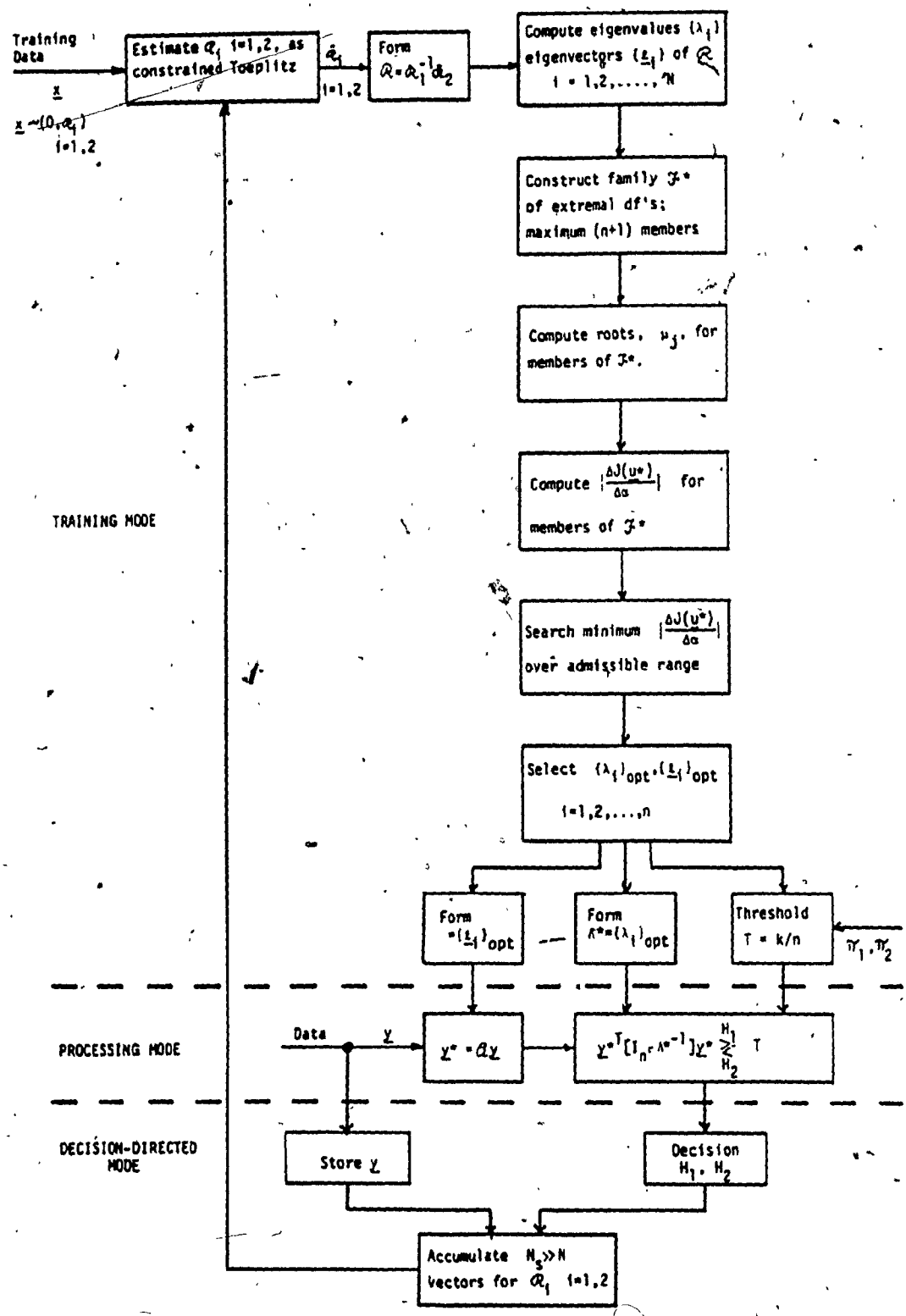

Figure 5.1 The Strategy for Optimal Feature Selection Consisting of Three Modes of Operation.

modes of operation. We bring to bear some well-known and some new computationally efficient algorithms in order to realize an efficient pattern classifier. The pattern classifier thus obtained shall prove useful for computer-based real-time image processing and robotic systems.

5.2. THE TRAINING MODE

The main objective of the training mode is to establish the system parameters of the pattern classifier. This consists of the construction of an $(n\times n)$-dimensional diagonal matrix Λ^*, the formation of an $(n\times N)$-dimensional data reducing transformation A, and the evaluation of a decision-threshold T. The diagonal elements of Λ^*, $\lambda_1^*>\lambda_2^*> \ldots > \lambda_n^*$, are the appropriately selected eigenvalues of the covariance matrix pair (R_1, R_2), i.e., the selection of n among N eigenvalues of (R_1, R_2) suggested by the feature selection scheme of Chapter 3. The rows of the transformation matrix A are the n eigenvectors of (R_1, R_2) corresponding to the eigenvalues λ_i^* $1<i<n$.

We begin by assuming that initial estimates of the Toeplitz covariance matrices R_1 and R_2 are known. We return in the sequel to the question of how to carry out the estimation of the covariances R_1 and R_2. The training mode performs the following steps sequentially to achieve the objective of establishing the system parameters.

Step 1. Matrix Inversion.

The $(N\times N)$-dimensional covariance matrix R_1 is assumed to be a

positive definite Toeplitz matrix. It has been noted in Chapter 4 that the matrix R_1 is symmetric (or Hermitian) Toeplitz depending on whether the data space is real (or complex). This a priori knowledge about R_1 permits us to obtain its inverse efficiently by using some well-known algorithms. The Trench algorithm [5.1] and the refined algorithm of Zohar [5.2] for the Hermitian Toeplitz case, and the specialized algorithm of Preis [5.3] for the symmetric Toeplitz case, can be ulitized to efficiently invert the matrix R_1. A Fortran routine to accomplish symmetric Toeplitz matrix inversion, using the scheme proposed by Preis, has appeared in [5.4]. The bordering scheme for matrix inversion used in the above-mentioned algorithms is appealing since the multiplicative complexity involved is $O(N^2)$. The matrices R_1 and R_1^{-1} are both centrosymmetric matrices and, thus, the storage requirement for R_1^{-1} is $N(N+1)/4$ $(\text{or}(N+1)^2/4)$ for even (or odd) N.

Step 2. Matrix Product, $R \triangleq R_1^{-1}R_2$.

The eigenvalues and eigenvectors of the covariance matrix pair (R_1, R_2), required for feature selection, are the principal components of the matrix product $R_1^{-1}R_2$; thus, we need to compute $R \triangleq R_1^{-1}R_2$. The matrix R, as shown in Chapter 4, is a centrosymmetric or a centrohermitian matrix depending on whether the data vectors are real or complex, respectively. Moreover, the matrix R is completely defined by $N^2/2$ or $(N^2+1)/2$ real (complex) elements for even or odd N, respectively, for the case when R is a centrosymmetric (centrohermitian) matrix. The elements of the centrosymmetric (centrohermitian)

R may be obtained by performing <u>at most</u> $5N^3/24+O(N^2)$ real (complex) multiplications as follows.

Let the elements of R be denoted by $\{\rho_{ij}|1\le i,j\le N\}$, the elements of R_1^{-1} be denoted by $\{\sigma_{ij}|1\le i,j\le N\}$, and the elements of R_2 be denoted by $\{\tau_{ij}|1\le i,j\le N\}$. We first consider the case when the data vectors are real and the covariance matrices are symmetric Toeplitz; also, the matrices R_1^{-1} and R_2 are both centrosymmetric. From the centrosymmetry of R_1^{-1} and R_2, we obtain the following correspondences,

$$\sigma_{ij} = \sigma_{N+1-i,N+1-j} = \sigma_{ji} \qquad 1\le i,j\le N \tag{5.1a}$$

$$\tau_{ij} = \tau_{N+1-i,N+1-j} = \tau_{ji} \qquad 1\le i,j\le N \tag{5.1b}$$

Noting that the elements of the symmetric Toeplitz matrix R_2 are the functions of $|i-j|$ rather than of i and j independently as in the case of a general matrix; thus, let

$$\tau_{ij} = \tau_{|i-j|+1} \qquad 1\le i,j\le N \tag{5.1c}$$

The elements of the matrix R may be obtained as,

$$\rho_{ij} = \sum_{k=1}^{N} \sigma_{ik}\tau_{kj} \qquad 1\le i,j\le N \tag{5.2}$$

Using the relations of (5.1), (5.2) becomes,

$$\rho_{ij} = \sum_{k=1}^{i-1}\sigma_{ki} \bullet \tau_{|k-j|+1} + \sum_{k=i}^{N+1-i} \sigma_{ik} \bullet \tau_{|k-j|+1} + \sum_{k=N+2-i}^{N}\sigma_{N+1-k,N+1-i} \bullet \tau_{|k-j|+1} \qquad 1\le i,j\le N \tag{5.3}$$

The product of the matrices R_1^{-1} and R_2 and, thus, the elements of R

may be quite efficiently obtained by using (5.3). It may not be obvious from (5.3), but <u>at most</u> the multiplications of σ_{ij} with τ_k for $1 \le i \le N$, $i \le j \le N-i+1$ and $1 \le k \le N+1-i$ need be performed to completely define the product matrix R, thus, the multiplicative complexity is $5N^3/24+O(N^2)$.

We now address the problem of obtaining the elements of R when the data vectors are complex and the resulting covariances R_1 and R_2 are Hermitian Toeplitz matrices; also, the matrices R_1^{-1} and R_2 are both Hermitian persymmetric. The elements of R_1^{-1} and R_2 satisfy the following equalities,

$$\sigma_{ij} = \sigma^*_{N+1-i,N+1-j} = \sigma^*_{ji} \qquad 1 \le i,j \le N \tag{5.4a}$$

$$\tau_{ij} = \tau^*_{N+1-i,N+1-j} = \tau^*_{ji} \qquad 1 \le i,j \le N \tag{5.4b}$$

Since the matrix R_2 is Hermitian Toeplitz, the elements of R_2 are functions of the difference $(i-j)$ rather than i and j independently; thus, let

$$\tau_{ij} = \begin{cases} \tau^*_{(i-j)+1} & i > j \\ \tau_{(j-i)+1} & i < j \end{cases} \qquad 1 \le i,j \le N \tag{5.4c}$$

Using (5.4), the matrix product (5.2) may now be cast in the following form,

$$\rho_{ij} = \sum_{k=1}^{i-1} \sigma^*_{ki} \bullet \tau_{(j-k)+1} + \sum_{k=i}^{N+1-i} \sigma_{ik} \bullet \tau_{(j-k)+1} + \sum_{k=N+2-i}^{N} \sigma_{N+1-k,N+1-i} \bullet \tau_{(j-k)+1} \qquad 1 \le i,j \le N \tag{5.5}$$

The relation (5.5) may be used to efficiently obtain the elements of

the product matrix R. Care must be taken for the values of $\tau_{(\bullet)+1}$ whenever $(\bullet)$ is negative, in which case, the quantity $\tau^*_{-(\bullet)+1}$ must be used instead. The computation of ρ_{ij} $1<i,j<N$, the elements of R, requires <u>at most</u> the complex multiplications of σ_{ij} with τ_k $1<i<N$, $i<j<N-i+1$ and $1<k<N+1-i$; thus, the multiplicative complexity‡, in terms of complex multiplies, is $5N^3/24+O(N^2)$.

<u>Step 3.</u> Principal Component Extraction of $R \triangleq R_1^{-1}R_2$.

This step of the feature selection process is computationally the most intensive. A good deal of effort has been devoted to develop the results of Chapter 4 for efficient principal component extraction of the product matrix R. The matrix R is either a centrosymmetric or a centrohermitian matrix depending on whether the input data vectors are real or complex, respectively. In either case, the matrix R has been shown to possess a reducible characteristic equation. The principal components of the $(N\times N)$-dimensional centrosymmetric matrix R may be obtained from the solutions of two characteristic equations of order $N/2$ each (one $(N+1)/2$ and one $(N-1)/2$) for even (odd) N. For the other case, when R is an $(N\times N)$-dimensional centrohermitian matrix, the principal component of R may be related to the characteristic equations of two $(N\times N)$-dimensional real matrices. The resulting

---

‡ Note that once the product of two complex numbers z_1 and z_2 is performed, other product combinations such as $z_1^*z_2^*$, $z_1^*z_2$ and $z_1z_2^*$ may be obtained without any additional multiplications. Let $z_1=a+jb$, $z_2=c+jd$, then

$$z_1z_2 = (ac-bd) + j(ad+bc)$$

Now, for example, the product,

$$z_1^*z_2 = (ac+bd) + j(ad-bc)$$

may be obtained from z_1z_2 <u>without</u> any additional multiplications.

sub-problems for the both cases can be solved by the general techniques available in the literature [5.5,5.6]. This realization introduces a significant savings of nearly 75% in the multiplicative complexity of Step 3.

We now study the situation when the real symmetric Toeplitz covariance matrices R_1 and R_2 may be approximated by the circulant matrices, R_1^c and R_2^c, respectively. Figure 5.2 outlines the feature selection strategy for this case. As illustrated in Figure 5.2, Steps 1 and 2 may be bypassed in this case and the strategy begins directly from step 3. We now discuss the manner in which Step 3 can provide the principal components of the product matrix $R_1^{c^{-1}} R_2^c$.

Recall from Chapter 4 that the matrices R_1^c, R_2^c and $R_1^{c^{-1}} R_2^c$ are <u>all</u> circulant matrices. In addition, the eigenvalues of a circulant matrix may be obtained by computing the discrete Fourier transform (DFT) of the elements of the first row. Then the procedure for determining the eigenvalues of $R_1^{c^{-1}} R_2^c$ is as follows. The reciprocals of the DFT of the first row of R_1^c provides the eigenvalues of $R_1^{c^{-1}}$. The eigenvalues of R_2^c may be obtained by computing the DFT of the first row of the matrix R_2^c. The eigenvalues of the product matrix $R_1^{c^{-1}} R_2$ are simply the point-by-point multiplications of the eigenvalues of the matrices $R_1^{c^{-1}} R_2^c$ since the factors of the product commute. The eigenvectors of the product matrix $R_1^{c^{-1}} R_2$ are equal to the columns of the N-dimensional Fourier matrix [5.7-5.9]. The

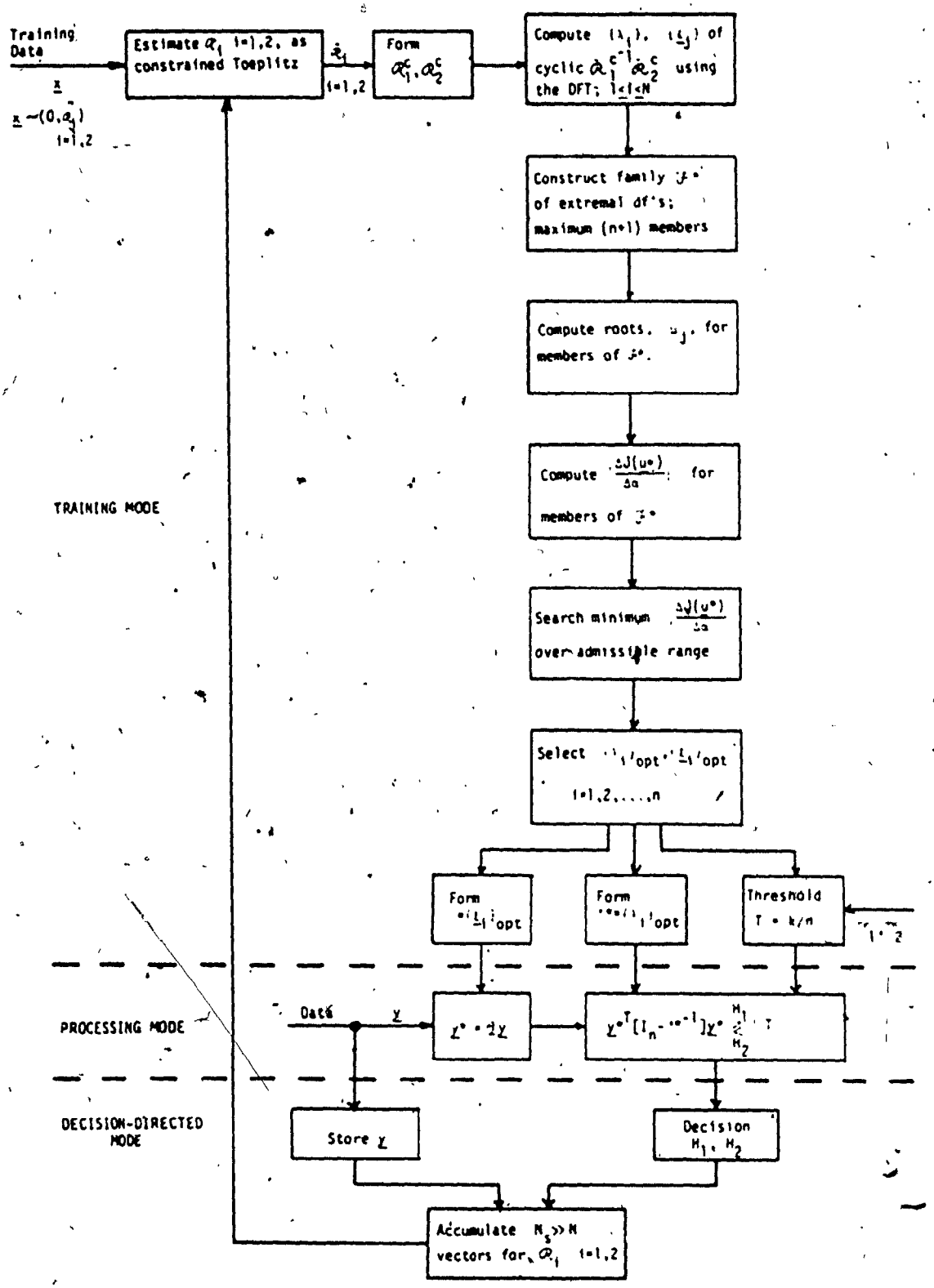

Figure 5.2 The Strategy for Optimal Feature Selection in the Case of Circulant Approximation of Toeplitz Covariances.

efficiency of performing the DFT may be enhanced by padding the N-dimensional sequence of the elements of the first row of a circulant matrix by zeroes to form an N'-dimensional sequence such that $N'=2^m$ $(N'>N)$, where m is any positive integer. The fast Fourier transform (FFT) on the N'-dimensional zero-padded sequence may be performed in only $N'\log_2 N'$ multiplications [5.10,5.11]. Additional computational savings may be achieved by using the Winograd fast transform algorithm (WFTA) [5.11,5.12]. In fact, a denser set of lengths is available with the WFTA as opposed to the FFT.

It is interesting to note that circulant matrices are a special case of CS matrices and, thus, have either symmetric or skew-symmetric eigenvectors [5.13]. The remaining discussion on the implementation of the pattern classifier is applicable to this approximation as well.

Step 4. Construction of the Family F^* of Extremal df's.

The construction of the family F^* of extremal df's is merely a selection, based on the criterion (3.22), of certain df's among $F_n^*(x)$ parameterized by α; this step does not require any computation. The number of smallest (or largest) eigenvalues, a maximum of n, below (or above) unity determines the family F^*. The family F^* may contain a maximum of $(n+1)$ members, where n is the dimension of the feature space. Each member of F^* is a potential solution to the feature selection problem. Step 4 may invariably restrict the number of possible solutions over which the error probability expression need be optimized.

Step 5. Computation of the Root, μ.

This step in the feature selection process appears to be quite complex but, in fact, this is not the case. Here, the solution of the following equality is desired,

$$\frac{1}{n}\sum_{i(\alpha)}\frac{q_i}{1+\mu q_i} = \frac{1}{n}\sum_{i(\alpha)} \ln \lambda_i + \ln\left(\frac{\pi_1}{\pi_2}\right) = k/n \qquad (3.30b)$$

for the root μ $(-1,0)$, using the functional $u(t)$, given by

$$u(t) = \frac{1}{n}\sum_{i(\alpha)}\frac{q_i}{1+tq_i} \qquad (3.31)$$

where $q_i \triangleq (1-1/\lambda_i)$ and $i(\alpha)$ determines the eigenvalue selection λ_i $1<i<n$ for each df member of the family F^*. The condition for determining μ is that $u(t=\mu)=k/n$. We know that $\mu\in(-1,0)$ and the functional $u(t)$ is a well-behaved monotonically decreasing function of t in the restricted region as typified by Figures 3.8 and 3.9. The manner in which the root μ may be determined efficiently is as follows.

Let the function $u(t)$ be discretized at m points within the restricted region of $t\in(-1,0)$ with a quantization step of $1/m$. A search for μ such that $u(t=\mu)=k/n$ can be performed quite efficiently. Due to the regular behavior of the functional $u(t)$, the well-known binary search [5.14] may be utilized to accomplish the task in <u>at most</u> $\log_2 m+1$ steps, provided an appropriate tolerance is incorporated to account for the quantization of $u(t)$ in the comparison of $u(t)$ with k/n. Note that the discrete value of $u(t)$ need be computed at a <u>maximum</u> of $\log_2 m+1$ discrete values of t, as suggested by the

binary search procedure. This process must be repeated for each member, a maximum of $(n+1)$, of F^*, to obtain the corresponding root μ.

Step 6. Determining the Optimal df.

Step 6 is the final step in obtaining the solution for feature selection. This step requires the computation of $\left|\frac{\Delta J(\underline{u}^*)}{\Delta\alpha}\right|$, given by,

$$\frac{\Delta J(\underline{u}^*)}{\Delta\alpha} = \frac{\mu(1+\mu)}{2n} \sum_{i^c(\alpha)} \frac{q_i/\lambda_i}{1+\mu q_i} \tag{3.38}$$

where $i^c(\alpha)$ is the complement of the set of eigenvalues determined by $i(\alpha)$, i.e., the $i^c(\alpha)$ determines the set of $(N-n)$ eigenvalues not contained in the set determined by $i(\alpha)$. The expression (3.38) should be evaluated once for each of a maximum of $(n+1)$ members of the family F^*. The df that provides $\min_{i(\alpha)} \left|\frac{\Delta J(\underline{u}^*)}{\Delta\alpha}\right|$ is the optimal df.

In the training mode, Steps 1 through 6 are performed in order to establish the system parameters. The eigenvalues λ_i^* $1<i<n$ of the covariance matrix pair (R_1,R_2), appropriately selected by the optimal df, are arranged in descending order to form the diagonal matrix Λ^*, and the corresponding eigenvectors are arranged as the rows of the $(n\times N)$-dimensional data reducing transformation matrix A. The decision threshold $T=k/n$ is computed using (3.30b). Having established the system parameters, we are ready to initiate processing input vectors for classification.

5.3. THE PROCESSING MODE

The classification of patterns is performed during the processing mode. The dimensionality reduction, or data compression, is achieved by applying the $(n \times N)$-dimensional $(n<N)$ transformation A to the $(N \times 1)$-dimensional data vector $\underline{x}$ to obtain the $(n \times 1)$-dimensional feature vector $\underline{y}$, i.e.,

$$\underline{y} = A \, \underline{x} \tag{3.1}$$

The structure of the transformation A must be examined in order to perform (3.1) efficiently. Note that the rows of A are the n appropriately selected eigenvectors of the matrix pair (R_1, R_2), or equivalently, that of the product matrix $R \triangleq R_1^{-1} R_2$. Two cases of interest immediately follow, viz., R is a centrosymmetric matrix for real input vectors, or, R is a centrohermitian matrix for complex input vectors.

We first examine the case when R is a centrosymmetric matrix. It has been established, by Theorem 4.1, that the eigenvectors of R are either symmetric or skew-symmetric; thus, the rows of transformation A are either symmetric or skew-symmetric. In view of this, we state the following result on the inner product of two vectors, given that one of them is either symmetric or skew-symmetric.

Lemma 5.1. Let $\underline{f}$ and $\underline{g}$ be $(N \times 1)$-dimensional vectors.

i) Let N be even $(N=2M)$, and the vector $\underline{g}$ be partitioned as,

$$\underline{g} = [\underline{g}_1, \underline{g}_2]^T$$

where $\underline{g}_1$ and $\underline{g}_2$ are $(M \times 1)$-dimensional vectors.

a) If $\underline{f}$ is a symmetric vector partitioned as,

$$\underline{f} = [\underline{f}_1, E_M \underline{f}_1]^T$$

then the inner product of the vectors $\underline{f}$ and $\underline{g}$ is given by,

$$\underline{f}^T \underline{g} = \underline{f}_1^T [\underline{g}_1 + E_M \underline{g}_2]$$

where $\underline{f}_1$ is an (M×1)-dimensional vector, and E_M is the (M×M)-dimensional contra-identity matrix.

b) If $\underline{f}$ is a skew-symmetric vector partitioned as,

$$\underline{f} = [\underline{f}_1, -E_M \underline{f}_1]^T$$

then the inner product of the vectors $\underline{f}$ and $\underline{g}$ is given by,

$$\underline{f}^T \underline{g} = \underline{f}_1^T [\underline{g}_1 - E_M \underline{g}_2]$$

ii) Let N be odd (N=2M+1), and the vector $\underline{g}$ be partitioned as,

$$\underline{g} = [\underline{g}_1, \alpha, \underline{g}_2]^T$$

where $\underline{g}_1$ and $\underline{g}_2$ are (M×1)-dimensional vector, and α is scalar.

a) If $\underline{f}$ is a symmetric vector partitioned as,

$$\underline{f} = [\underline{f}_1, \beta, E_M \underline{f}_1]^T$$

then the inner product of the vectors $\underline{f}$ and $\underline{g}$ is given by,

$$\underline{f}^T \underline{g} = \underline{f}_1^T [\underline{g}_1 + E_M \underline{g}_2] + \alpha \beta$$

where $\underline{f}_1$ is an (M×1)-dimensional vector, E_M is the (M×M)-dimensional contra-identity matrix, and β is a scalar.

b) If $\underline{f}$ is a skew-symmetric vector partitioned as,

$$\underline{f}_1 = [\underline{f}_1, 0, E_M\underline{f}_1]^T$$

then the inner product of the vectors $\underline{f}$ and $\underline{g}$ is given by,

$$\underline{f}^T\underline{g} = \underline{f}_1^T[\underline{g}_1 - E_M\underline{g}_2]$$

The proof is straightforward. ∎

The operation (3.1) may now be performed, using Lemma 5.1, with a nearly 50% reduction in the multiplicative complexity.

The classification of patterns, based on the feature vector $\underline{y}$, is performed in accordance with binary hypothesis testing rule,

$$\underline{y}^T[I_n - \Lambda^{*-1}]\underline{y} \underset{H_2}{\overset{H_1}{\gtrless}} T \qquad (5.6)$$

where I_n is the (n×n)-dimensional identity matrix. The quantites Λ^* and T have already been computed in the training mode. The left-hand-side of the inequality (5.6), $\underline{y}^T[I_n-\Lambda^{*-1}]\underline{y}$, may be directly computed in 2n multiplications. The overall complexity of the processing mode is then approximately (nN/2)+2n, in the case when is a centrosymmetric matrix.

We now examine the complexity of the processing mode for the case when the input vectors are complex and the covariance R_1 and R_2 are Hermitian Toeplitz matrices. In this case, the product matrix $R \triangleq R_1^{-1}R_2$ is a member of the class of centrohermitian matrices. We have

discussed in Section 4.3 that an $(N\times N)$-dimensional centrohermitian matrix R may be represented by a $(2N\times 2N)$-dimensional real centrosymmetric matrix $\hat{R}$ for numerical purposes. The manner in which the characteristic equation of R may be related to $\hat{R}$ has also been discussed in Section 4.3. Due to the a priori knowledge that R has strictly real positive eigenvalues, the discussion here is restricted to this particular case. For each real eigenvalue λ_k and and the corresponding eigenvector $\underline{\ell}_k=\underline{u}_k+j\underline{v}_k$ of the complex matrix R, the real matrix $\hat{R}$ has a double eigenvalue λ_k with corresponding $\ell.i.$ eigenvectors given by $[\underline{u}_k,\underline{v}_k]^T$ and $[-\underline{v}_k,\underline{u}_k]^T$, where $\underline{\ell}_k$, $\underline{u}_k$ and $\underline{v}_k$ are $(N\times 1)$-dimensional vectors. Moreover, corresponding to a double real eigenvalue of $\hat{R}$, the eigenvector $\underline{u}_k+j\underline{v}_k$ derived from either of the eigenvectors $[\underline{u}_k,\underline{v}_k]^T$ and $[-\underline{v}_k,\underline{u}_k]^T$ of $\hat{R}$ is an eigenvector of the matrix R. However, the centrosymmetric matrix $\hat{R}$ has eigenvectors which are either symmetric or skew-symmetric. Thus, the matrix $\hat{R}$ has $(2N\times 1)$-dimensional partitioned eigenvectors of the form $[\underline{u}_k,E_N\underline{u}_k]^T$ or $[\underline{u}_k,-E_N\underline{u}_k]^T$ with vector $\underline{\ell}_k=\underline{u}_k+jE_N\underline{u}_k$ or $\underline{\ell}_k=\underline{u}_k-jE_N\underline{u}_k$, respectively, corresponding to the eigenvectors of the matrix R. The rows of the transformation A, the n appropriately selected eigenvectors of R, are of the form $\underline{\ell}_k=\underline{u}_k\pm jE_N\underline{u}_k$. In view of the above, we examine the inner product of two vectors of the form $\underline{f}=f_1\pm jE_N\underline{f}_1$ and $\underline{g}=\underline{g}_1+j\underline{g}_2$, where $\underline{f},\underline{g}$ are $(N\times 1)$-dimensional complex vectors, and $\underline{f}_1,\underline{g}_1,\underline{g}_2$ are $(N\times 1)$-dimensional real vectors.

Lemma 5.2. Let the (N×1)-dimensional complex vectors $\underline{f}$ and $\underline{g}$ be of the form,

$$\underline{f} = \underline{f}_1 \pm jE_N\underline{f}_1$$

$$\underline{g} = \underline{g}_1 + j\underline{g}_2$$

where $\underline{f}_1$, $\underline{g}_1$ and $\underline{g}_2$ are (N×1)-dimensional real vectors, and E_N is the (N×N)-dimensional contra-identity matrix. The inner product of the vectors $\underline{f}$ and $\underline{g}$ is given by

$$\underline{f}^T\underline{g} = \underline{f}_1^T\,[(\underline{g}_1 \mp E_N\,\underline{g}_2) + j(\underline{g}_2 \pm E_N\,\underline{g}_1)]$$

The proof is straightforward. ∎

Lemma 5.2 offers, similar to Lemma 5.1, a 50% reduction in the multiplicative complexity involved in performing the operation (3.1). The computations required to perform the decision threshold testing of (5.6) in this case are similar to those of the previous case of real input vectors. The decision rule in this case is,

$$\underline{y}^H[I_n - \Lambda^{*-1}]\underline{y} \underset{H_2}{\overset{H_1}{\gtrless}} T \qquad (5.7)$$

where H denotes the matrix complex conjugate transpose.

Sections 5.2 and 5.3 have demonstrated that a pattern classifier employing the feature selection scheme of Chapter 3 may be implemented quite efficiently. The pattern classifier incorporating only the training mode and the processing mode may be extended to include a decision-directed mode. The decision-directed mode is used to update the system parameters using the training mode functions, thereby taking into account a realistic quasi-stationarity of the environment, in

which the features evolve slowly with time.

5.4. THE DECISION-DIRECTED MODE

The initial estimates of the covariance matrices R_1 and R_2 may be obtained from training data as,

$$r^i_{\ell k} = E\{x^i_\ell x^{i*}_k\} \qquad 1 \leq \ell, k \leq N \tag{5.8}$$

where $r^i_{\ell k}$ denotes the ℓkth entry in the R_i $i=1,2$ matrix. The quantities x^i_j $1 \leq j \leq N$ are the entries of the input data vector $\underline{x}^i$ under hypothesis H_i $i=1,2$. The covariance matrices thus obtained are used to establish the system parameters Λ^*, A, and T as discussed in Section 5.2 before the pattern classification process of Section 5.3 is initiated. The complexity of the processing mode is nominal as can be surmised from Section 5.3 and the processor can easily operate in real-time. This mechanism of pattern classification is satisfactory provided the input to the processor is of stationary nature, but the problem becomes somewhat more complex if quasi-stationarity of patterns is assumed. This is attributed to the fact that the system parameters Λ^*, A, and T must be updated as we move from one interval of stationarity to the next. Therefore, a means of estimating the covariances R_1 and R_2, and establishing the system parameters in parallel with the processing mode is required.

The underlying assumption of the following discussion on an iterative estimation of covariances is that the classification error of

the pattern classifier is low, say $O(10^{-6})$. The classification decisions made by the classifier can then be utilized to estimate iteratively the matrices R_1^{-1} and R_2 in the following manner. Let G be an (N×N)-dimensional covariance matrix related to the (N×1)-dimensional sample vectors $\underline{g}$. Then the estimates of G_m and G_m^{-1} at the mth iteration may be obtained from $\hat{G}_{m-1}$ and $\hat{G}_{m-1}^{-1}$, respectively, by using the relations [5.15],

$$\hat{G}_m = (1-\frac{1}{m})\,\hat{G}_{m-1} + \frac{1}{m}\,\underline{g}_m\underline{g}_m^H \tag{5.9a}$$

and,

$$\hat{G}_m^{-1} = \frac{m}{m-1}\,\hat{G}_{m-1}^{-1} - \frac{\hat{G}_{m-1}^{-1}\underline{g}_m\underline{g}_m^H\hat{G}_{m-1}^{-1}}{(m-1)+\underline{g}_m^H\hat{G}_{m-1}^{-1}\underline{g}_m} \tag{5.9b}$$

here ^ denotes an estimate of the quantity enclosed. Some adaptive mechanisms such as least-mean-squares (LMS) method [5.16], or modified Kalman filtering technique [5.17], may also be incorporated to estimate the covariance matrices iteratively. The system parameters may be computed periodically via the training mode, in parallel with processing mode, after good estimates for R_1^{-1} and R_2 are obtained.

An alternate approach of estimating the covariances R_1 and R_2 as constrained Toeplitz matrices is due to Morgera et al [5.18, 5.19]. The method is simple and appealing, as the estimation is quite accurate when the number of input vectors is small. However, it does not seem possible to extend the approach to a recursive form similar to that of (5.9) due to the fact then an instantaneous covariance estimate is not rank one.

5.5. DISCUSSION

A detailed complexity analysis of the pattern classifier is presented in view of a computer-based real-time implementation. The pattern classifier incorporates the feature selection scheme proposed in Chapter 3. The training mode, the processing mode, and the decision-directed mode are three modes of operation of the classifier. The training mode, used to establish the system parameters, is quite complex and the results of Chapter 4 are used to reduce the complexity of this mode significantly. The processing mode during which the classification of patterns is performed, is shown to have nominal complexity. The decision-directed mode is used to update the system parameters via the training mode to account for quasi-stationarity of patterns.

A number of new results along with some well-known results are incorporated to realize an efficient pattern classifier. The computational complexity involved in every step of the classifier is discussed. This study shall prove useful in the practical implementation of mini/micro-computer-based pattern classifiers for image processing system and robotics.

REFERENCES

[5.1] W.F. Trench, "An Algorithm for the Inversion of Finite Toeplitz Matrices," J. Soc. Indust. Appl. Math., vol. 12, pp. 512-522, Sept. 1954.

[5.2] S. Zohar, "Toeplitz Matrix Inversion: The W.F. Trench Algorithm," J. Ass. Comput. Mach., vol. 16, pp. 592-601, Oct. 1969.

[5.3] D.H. Preis, "The Toeplitz Matrix: Its Occurence in Antenna Problems and a Rapid Inversion Algorithm," IEEE Trans. Antennas Propagat., pp. 204-206, Mar. 1972.

[5.4] D.C. Farden and L.L. Sharf, Author's reply to Butler's comments on, "Statistical Design of Nonrecursive Digital Filters," IEEE Trans. Acoust., Speech, and Signal Processing, vol. ASSP-23, pp. 495-497, Oct. 1975.

[5.5] J.H. Wilkinson, The Algebraic Eigenvalue Problem. Oxford: Claredon, 1965.

[5.6] J.H. Wilkinson and C. Reinsch, Linear Algebra. New York: Springer-Verlag, 1971.

[5.7] P.J. Davis, Circulant Matrices. New York: Wiley, 1979.

[5.8] S.D. Morgera, "On the Reducibility of Finite Toeplitz Matrices - Applications in Speech Analysis and Pattern Recognition," Signal Processing, vol. 4, pp. 425-443, Oct. 1982.

[5.9] J. Biemond, J. Rieske and J.J. Gerbrands, "A Fast Kalman Filter for Images Degraded by Both Blur and Noise," IEEE Trans. Acoust., Speech, and Signal Processing, vol. ASSP-31, pp. 1248-1256, Oct. 1983.

[5.10] J.W. Cooley and J.W. Tukey, "An Algorithm for Machine Computation of Complex Fourier Series," Math Comput., pp. 297-302, Apr. 1965.

[5.11] J.S. Ward, P. Barton, J.B.G. Roberts, and B.J. Stanier, "Figures of Merit for VLSI Implementations of Digital Signal Processing Algorithms," Proc. IEE, Part F, vol. 131, pp. 64-70, Feb. 1984.

[5.12] S. Winograd, "On Computing the Discrete Fourier Transform," Math. Comput., vol. 32, pp. 175-199, Jan. 1978.

[5.13] L. Datta and S.D. Morgera, "On the Reducibility of Centrosymmetric Matrices - Applications in Engineering Problems," submitted for review to J. Franklin Institute, Sept. 1984.

[5.14] L.I. Kronsjo, Algorithms: Their Complexity and Efficiency. New York: Wiley, 1979, pp. 293-297.

[5.15] J. Karhunen, "Adaptive Algorithms for Estimating Eigenvectors of Correlation Type Matrices," Proc. IEEE Inter. Conf. Acoust. Speech, and Signal Processing, San Diego, California, Mar. 19-21, 1984.

[5.16] B. Widrow, P.E. Mantey, L.J. Griffiths and B.B. Goode, "Adaptive Antenna Systems," Proc. IEEE, vol. 55, pp. 2143-2159, Dec. 1967.

[5.17] S.D. Morgera and L. Datta, "On Reducing the Computational Complexity of Kalman Filtering for Narrowband Interference Rejection," IEEE Intern. Symp. Inform. Theory, Les Arcs, France, June 21-16, 1982.

[5.18] S.D. Morgera and D.B. Cooper, "Structured Estimation: Sample Size Reduction for Adaptive Pattern Classification", IEEE Trans. Inform Theory, vol. IT-23, pp. 728-741, Nov. 1977.

[5.19] S.D. Morgera, "Structured Estimation, Part II - Multivariate Probability Density Estimation", IEEE Trans. Inform. Theory, vol. IT-27, pp. 607-622, Sept. 1981.

CHAPTER 6

PERFORMANCE EVALUATION OF THE FEATURE SELECTION SCHEME AND EFFECTS OF CERTAIN PARAMETERS ON ERROR PROBABILITY

6.1. INTRODUCTION

The performance of a pattern recognition feature selector is, in general, evaluated on the basis of two important criteria, viz., reliability and economics. The reliability of the classification system is measured in terms of the probability of misclassification or error, whereas the economics of the system is related to the cost or computational complexity of the system. A detailed complexity analysis of a classification system employing the feature selection scheme of Chapter 3 has been presented in Chapter 5. This chapter deals with assessing the performance of the classifier in terms of error probability in comparison with conventional feature selection schemes, notably that of Kadota et al [6.1].

This work presents numerical results regarding various aspects of the pattern classifier. Numerical simulation results are included on the error probability achieved by using the Morgera-Datta (M-D) and Kadota-Shepp (K-S) methods. These results are then analyzed in view of the error bounds available using statistical distance measures, for example, the error bounds for feature selection utilizing the Bhattacharyya distance as discusssed in [6.2].

We examine the behavior of the error probability expression of (3.34) in relation to variation in the values of certain system

parameters. Effects of a priori probabilities π_1, π_2 are studied on the root μ of (3.34h) and, thus, on the classification error. We discuss the reasons for restricting the value of the free parameter γ of (3.26b) in Theorem 3.5. In addition, we present the changes observed in the probability of classification error as the feature dimension n is varied while maintaining the data dimension N fixed.

6.2. COMPUTER SIMULATION OF PATTERN CLASSIFIER-PERFORMANCE EVALUATION

Simulation of the pattern classifier was carried out on CDC Cyber 172 computer. Three examples of covariance matrix pairs (patterns) of Table 6.1 were considered. Due to limited computer resource allocation to a user, some difficulties were encountered for generating a sufficient number of data vectors of size $N=40$ with specified second order statistics. This problem was overcome by reducing the data and feature dimensions to $N=12$ and $n=3$, respectively, thereby maintaining the data compression ratio $n/N=.25$ (or 75% compression) for all the examples of Table 6.1. This value of the compression ratio is in agreement with the examples of Table 3.1. We note that the usage of the examples of Table 6.1 shall be restricted to this section only.

Feature selection for the pattern classifier simulation is performed by employing both the M-D and K-S methods. The (12×1)-dimensional data vectors with zero mean and multivariate normal (MVN) distribution with covariance matrix specified in Table 6.1 are generated using the International Mathematical and Statistical Library

(IMSL) subroutine GGNSM. The algorithm used in GGNSM for generating MVN distributed random vectors with specified second order statistics is similar to that discussed in [6.3]. For any particular covariance matrix pair example, 2000 data vectors are generated for each pattern. The simulated pattern classifier processes 4000 input vectors of two weakly stationary Gaussian stochastic processes for each example of Table 6.1.

The number of data vectors used in the simulation is deemed to be statistically adequate. Assuming that the number of misclassified samples is distributed according to a binomial probability distribution, a 95% confidence interval for the error probability estimates in the range of 0.1 to 0.25 is extremely tight for sample size in excess of 1000 vectors [6.7]. Using this type of model we calculate, for example, that $\text{Prob}\{ P_e(n)-P_e(n) < 1.56\times10^{-2}\}=0.95$, where $P_e(n)=.15$ is the <u>true</u> error rate of the classifier and $P_e(n)$ is the error rate estimated by simulation using 2000 vectors. We shall present the error probability results of the simulated classifier in the sequel, but now we study the feature selection for each example by using M-D and K-S methods.

Table 6.2 presents the eigenvalue selections and a comparison of the probability of classification error $P_e(n)$ for the M-D and K-s methods for each covariance pair example of Table 6.1. The results of Table 6.2 are obtained for equal <u>a priori</u> probabilities, i.e., $\pi_1=\pi_2=0.5$. The value of $P_e(n)$ listed in Table 6.2 is the theoretical value obtained by using the error probability expression (3.34). The root μ of (3.34) for each method is also included in Table 6.2. It is interesting to note from Table 6.2 that the M-D method is always <u>at</u>

TABLE 6.1

Toeplitz Covariance Matrix Pairs (R_1, R_2)

Selected as Covariance Examples

Example	DESCRIPTION (R_1, R_2)	PARAMETERS
I	(R_1, R_2) first order Markov $\rho_{k,\lvert i-j\rvert} = e^{-\alpha_k \lvert i-j\rvert}$	$\alpha=1,\ \alpha_2=0.5$
II	(R_1, R_2) first order Markov	$\alpha_1=1,\ \alpha_2=2$
III	R_1 first order Markov R_2 second order Markov	$\alpha_1=1$
	$\rho_{\lvert i-j\rvert} = e^{-\beta}\rho_{\lvert i-j\rvert-1} + e^{-\gamma}\rho_{\lvert i-j\rvert-2}$	$\beta=0.5,\ \gamma=2$

TABLE 6.2

$P_e(n)$ for New M-D Method and Conventional K-S Method with Respective Eigenvalue Selection;, N=12, n=3, $\pi_1=\pi_2$

COVARIANCE EXAMPLE	EIGENVALUE SELECTION M-D	K-S	ROOT, μ M-D	K-S	$P_e(n)$ M-D	K-S
I	$3^\ell/0^s$	$3^\ell/0^s$	-.5500	-.5500	.1716	.1716
II	$0^\ell/3^s$	$1^\ell/2^s$	-.4625	-.4850	.1439	.1557
III	$3^\ell/0^s$	$2^\ell/1^s$	-.555	-.5175	.1287	.1378

least as good as the K-S method and sometimes better in terms of error probability, $P_e(n)$. The decrease in the error probability obtained by the M-D method implies that the optimum selection cannot be achieved simply in terms of the inverse symmetric function λ_i+1/λ_i as in the K-S method. The proper selection of features must be made, as in the M-D method, by examining $\frac{\Delta J(\underline{u}^*)}{\Delta\alpha}$ of (3.38) for all members of the extremal df family F^* of (3.21). We shall now substantiate this claim as we return to presenting the classification error results of the computer simulated classifier.

Tables 6.3-6.5 present the simulation results for the probability of classification error $P_e(n)$ for each of the three examples of Table 6.1. Each table presents $P_e(n)$ achieved for all possible eigenvalue selections for a particular example. In addition, the error bounds on $P_e(n)$ obtained by using the Bhattacharyya distance for each eigenvalue selection are included. The error probability for the eigenvalue selection obtained by the M-D method, as can be surmised from Tables 6.3-6.5 by examining the respective eigenvalue selections, is always at least as good as the K-S method. It is also obvious from the same tables that the error bounds on $P_e(n)$ obtained by using the Bhattacharrya distance measure are quite "loose" and fail to provide any useful information for feature selection, e.g., the error bounds on the eigenvalue selection $3^\ell/0^s$ and $2^\ell/1^s$ in Table 6.4 are identical for Example II. For the same examples, these eigenvalue selections seem to be equally good and best for feature selection, whereas it can be seen from the simulation results of Table 6.4 and the theoretical results of Table 6.2 that the optimum choice of features is given by

TABLE 6.3

Error Bounds using Bhattacharyya Distance and Simulation Results for $P_e(n)$, for all Possible Eigenvalue Selections for Example I of Table 6.1

EIGENVALUE SELECTION	$P_e(n)$ CLASSIFIER SIMULATION	ERROR BOUNDS
$3^\ell/0^s$	.185	$.058<P_e(n)<.233$
$2^\ell/1^s$	.207	$.058<P_e(n)<.234$
$1^\ell/2^s$	.227	$.063<P_e(n)<.238$
$0^\ell/3^s$	.235	$.064<P_e(n)<.244$

TABLE 6.4

Error Bounds using Bhattacharyya Distance and Simulation Results for $P_e(n)$, for All Possible Eigenvalue Selections for Example II of Table 6.1

EIGENVALUE SELECTION	$P_e(n)$ CLASSIFIER SIMULATION	ERROR BOUNDS
$3^\ell/0^s$	.221	$.061<P_e(n)<.240$
$2^\ell/1^s$	.223	$.061<P_e(n)<.240$
$1^\ell/2^s$	.215	$.062<P_e(n)<.241$
$0^\ell/3^s$	.203	$.063<P_e(n)<.244$

TABLE 6.5

Error Bounds using Bhattacharyya Distance and Simulation Results for $P_e(n)$, for All Possible Eigenvalue Selections for Example III of Table 6.1

EIGENVALUE SELECTION	$P_e(n)$ CLASSIFIER SIMULATION	ERROR BOUNDS
$3^\ell/0^s$	.186	$.058<P_e(n)<.234$
$2^\ell/1^s$	.205	$.056<P_e(n)<.230$
$1^\ell/2^s$	.223	$.059<P_e(n)<.236$
$0^\ell/3^s$	.231	$.062<P_e<.2417$

the eigenvalue selection $0^{\ell}/3^{s}$. A similar situation is encountered for Example III as shown in Table 6.5. In this case, the eigenvalue selection $2^{\ell}/1^{s}$ made by the K-S method appears to be the best in terms of the respective error bounds, but it is obvious from the simulation results for $P_e(n)$ that this is not the case and and the selection of the M-D method is optimum.

To reconcile the seeming differences between the theoretical and simulation values of $P_e(n)$, we present the following discussion. This discrepency may be attributed to two reasons, namely, the size of the problem and the numerical inaccuracies. The feature selection approach of Kadota et al uses asymptotics in the formulation, and the M-D method optimizes the asymptotic formulation for finite sample size. The data and feature dimensions of $N=12$ and $n=3$, respectively, are too small to handle by either method. However, it is interesting to note that regardless of the discrepencies in the theoretical and simulation values of $P_e(n)$, the M-D method always makes the optimum selection. The second argument relates to the numerical inaccurracies embedded in the data generation. On computing the second order statistics of the generated data for simulation, the mean vectors were seen to be non-zero valued and some discrepencies in the covariances were apparent. As a direct consequence, the eigenvalues of the matrix pair (R_1, R_2) of the generated data are in a maximum of 8% error relative to that obtained from the desired means and covariances. Therefore, the small size of the problem compounded with the numerical inaccuracies share part of the blame for differences between the theoretical and simulation values of $P_e(n)$.

6.3. EFFECT OF A PRIORI PROBABILITIES ON ROOT μ

The parameter μ is the root of (3.34h) and lies in the restricted range of (-1,0) as shown in Chapter 3. In the limit, as $N \to \infty$ with n/N fixed, the K-S method requires that the root μ be $-\frac{1}{2}$ for the optimal df. A critical difference between the K-S method and the M-D method is that for the M-D method, a different value for the root $\mu \in (-1,0)$ is involved in each optimal selection and, even asymptotically, μ need not have the same value for different extremal df's. Although it is true that in the asymptotic case $P_e(n)$ is independent of the a priori probabilities [6.4-6.6], we study the effects of a priori probabilities on the root μ and, thus, on $P_e(n)$ for finite sample size. This section deals with the variations in the value of the root μ as the a priori probabilities π_1, π_2 change. Section 6.5 shall present the effects of a priori probabilities on the probability of classification error $P_e(n)$. We consider five covariance pair (R_1, R_2) examples of Table 3.1 in this section and for the remainder of this chapter. For each of the examples, the data and feature dimensions are taken as N=40 and n=10, respectively, resulting in a data compression ratio of 0.25 (or 75% compression).

Tables 6.6-6.10 illustrate the change in the value of the root μ for the M-D and K-S methods as the a priori probabilities π_1, π_2 vary, for each of the five examples of Table 3.1. Since the results of Chapter 3 are valid for $\pi_1 > \pi_2$, the value of π_1 is varied form 0.5 to 0.95 in steps of 0.5. We observe, from Tables 6.6-6.10, that an increase in the value of π_1, or, equivalently, a decrease in the value of π_2 causes the root μ to be more negative in the restricted range

TABLE 6.6

Effect of A Priori Probabilities π_1, π_2 on Root μ for Example I of Table 3.1.

A PRIORI PROBABILITIES π_1	π_2	ROOT μ M-D	K-S
0.5	0.5	-.5525	-.5425
0.55	0.45	-.6475	-.6425
0.6	0.4	-.735	-.7325
0.65	0.35	-.815	-.8175
0.7	0.3	-.8925	-.8975
0.75	0.25	-.9675	-.9675
0.8	0.2	-.9975	-.9975
0.85	0.15	-.9975	-.9975
0.9	0.1	-.9975	-.9975
0.95	0.05	-.9975	-.9975

TABLE 6.7

Effect of A Priori Probabilities π_1, π_2 on Root μ for Example II of Table 3.1

A PRIORI PROBABILITIES π_1	π_2	ROOT μ M-D and K-S
0.5	0.5	-.605
0.55	0.45	-.6275
0.6	0.4	-.65
0.65	0.35	-.6725
0.7	0.3	-.695
0.75	0.25	-.7175
0.8	0.2	-.74
0.85	0.15	-.7675
0.9	0.1	-.8
0.95	0.05	-.845

TABLE 6.8

Effect of A Priori Probabilities π_1, π_2 on Root μ for Example III of Table 3.1.

A PRIORI PROBABILITIES π_1	π_2	ROOT μ M-D	K-S
0.5	0.5	-.6025	-.5725
0.55	0.45	-.6275	-.6
0.6	0.4	-.65	-.625
0.65	0.35	-.6725	-.6425
0.7	0.3	-.695	-.6725
0.75	0.25	-.72	-.7
0.8	0.2	-.745	-.725
0.85	0.15	-.7725	-.7575
0.9	0.1	-.805	-.7925
0.95	0.05	-.8525	-.845

TABLE 6.9

Effect of A Priori Probabilities π_1, π_2 on Root μ for Example IV of Table 3.1.

A PRIORI PROBABILITIES π_1	π_2	ROOT μ M-D	K-S
0.5	0.5	-.4875	-.4625
0.55	0.45	-.545	-.5175
0.6	0.4	-.605	-.575
0.65	0.35	-.665	-.635
0.7	0.3	-.725	-.6975
0.75	0.25	-.79	-.7675
0.8	0.2	-.8575	-.8425
0.85	0.15	-.9325	-.9275
0.9	0.1	-.9975	-.9975
0.95	0.05	-.9975	-.9975

TABLE 6.10

Effect of A Priori Probabilities π_1, π_2 on Root μ for Example V of Table 3.1.

A PRIORI PROBABILITIES π_1	π_2	ROOT μ M-D and K-S
0.5	0.5	-.25
0.55	0.45	-.2525
0.6	0.4	-.2575
0.65	0.35	-.26
0.7	0.3	-.2625
0.75	0.25	-.2675
0.8	0.2	-.2725
0.85	0.15	-.28
0.9	0.1	-.2875
0.95	0.05	-.305

of (-1,0). The probability of error expression (3.34) is directly influenced by π_1 and π_2, and the root μ which itself is a function of π_1, π_2. The effect of the a priori probabilities on the error probability is discussed in Section 6.5.

6.4 EFFECT OF THE PARAMETER γ ON THE PROBABILITY OF CLASSIFICATION ERROR

The parameter γ, introduced in (3.26b) as a free parameter, was restricted in Theorem 3.5. The reasons for restricting the value of γ in relation to the feature dimension n are discussed in this section. But first, we present the effect of the value of γ on the error probability $P_e(n)$ of (3.34). Tables 6.11-6.15 show the quantity $P_e(n)$ as a function of the parameter γ for each of the five examples of Table 3.1. For each case, the root μ is appropriately selected with $\pi_1=\pi_2=0.5$. It is clear from Tables 6.11-6.15 that the parameter γ influences the probability of classification error, $P_e(n)$, of (3.34). Since the parameter γ was artificially introduced to derive the expression for $P_e(n)$, it is desirable that the effect of γ on $P_e(n)$ be slight. It may be observed from Tables 6.11-6.15 that a large value of γ implies a lower probability of error, with the error probability increasing as the value of γ is reduced until a certain point, for example, $\gamma=10^{-4}$, when the error probability appears to be numerically independent of further reduction in the value of γ. This observation, consistent for all the examples of Table 3.1, has established the constraint of Theorem 3.5 on the value of γ that γ

TABLE 6.11

Effect of Parameter γ on Error Probability, $P_e(n)$, for Example I of Table 3.1. The Root μ is Appropriately Selected for $\pi_1=\pi_2$.

PARAMETER	ERROR PROBABILITY, $P_e(n)$	
γ	M-D	K-S
1	3.5807×10^{-8}	3.5686×10^{-7}
0.2	5.4543×10^{-8}	5.5184×10^{-7}
0.04	5.9443×10^{-8}	6.0278×10^{-7}
8×10^{-3}	6.0479×10^{-8}	6.1354×10^{-7}
1.6×10^{-3}	6.0688×10^{-8}	6.1571×10^{-7}
3.2×10^{-4}	6.0731×10^{-8}	6.1616×10^{-7}
6.4×10^{-5}	6.0739×10^{-8}	6.1624×10^{-7}
1.28×10^{-5}	6.074×10^{-8}	6.1626×10^{-7}
2.56×10^{-6}	6.0741×10^{-8}	6.1626×10^{-7}
5.12×10^{-7}	6.0741×10^{-8}	6.1627×10^{-7}
1.024×10^{-7}	6.0741×10^{-8}	6.1627×10^{-7}
2.048×10^{-8}	6.0741×10^{-8}	6.1627×10^{-7}
4.096×10^{-9}	6.0741×10^{-8}	6.1627×10^{-7}
8.192×10^{-10}	6.0741×10^{-8}	6.1627×10^{-7}
1.6384×10^{-10}	6.0741×10^{-8}	6.1627×10^{-7}

TABLE 6.12

Effect of Parameter γ on Error Probability, $P_e(n)$, for Example II of Table 3.1. The Root μ is Appropriately Selected for $\pi_1=\pi_2$.

PARAMETER γ	ERROR PROBABILITY, $P_e(n)$ M-D and K-S
1	1.5033×10^{-8}
0.2	2.4501×10^{-8}
0.04	2.7226×10^{-8}
8×10^{-3}	2.7815×10^{-8}
1.6×10^{-3}	2.7935×10^{-8}
3.2×10^{-4}	2.7959×10^{-8}
6.4×10^{-5}	2.7963×10^{-8}
1.28×10^{-5}	2.7964×10^{-8}
2.56×10^{-6}	2.7965×10^{-8}
5.12×10^{-7}	2.7965×10^{-8}
1.024×10^{-7}	2.7965×10^{-8}
2.048×10^{-8}	2.7965×10^{-8}
4.096×10^{-9}	2.7965×10^{-8}
8.192×10^{-10}	2.7965×10^{-8}
1.6384×10^{-10}	2.7965×10^{-8}

TABLE 6.13

Effect of Parameter γ on Error Probability, $P_e(n)$, for Example III of Table 3.1. The Root μ is Appropriately Selected for $\pi_1=\pi_2$.

PARAMETER	ERROR PROBABILITY, $P_e(n)$	
γ	M-D	K-S
1	1.5726×10^{-8}	1.6421×10^{-7}
0.2	2.555×10^{-8}	2.7785×10^{-7}
0.04	2.8365×10^{-8}	3.097×10^{-7}
8×10^{-3}	2.8973×10^{-8}	3.1654×10^{-7}
1.6×10^{-3}	2.9096×10^{-8}	3.1792×10^{-7}
3.2×10^{-4}	2.9121×10^{-8}	3.182×10^{-7}
6.4×10^{-5}	2.9126×10^{-8}	3.1826×10^{-7}
1.28×10^{-5}	2.912710^{-8}	3.1827×10^{-7}
2.56×10^{-6}	2.9127×10^{-8}	3.1827×10^{-7}
5.12×10^{-7}	2.9127×10^{-8}	3.1827×10^{-7}
1.024×10^{-7}	2.9127×10^{-8}	3.1827×10^{-7}
2.048×10^{-8}	2.9127×10^{-8}	3.1827×10^{-7}
4.096×10^{-9}	2.9127×10^{-8}	3.1827×10^{-7}
8.192×10^{-10}	2.9127×10^{-8}	3.1827×10^{-7}
1.6384×10^{-10}	2.9127×10^{-8}	3.1827×10^{-7}

TABLE 6.14

Effect of Parameter γ on Error Probability, $P_e(n)$, for Example IV of Table 3.1. The Root μ is Appropriately Selected for $\pi_1=\pi_2$.

PARAMETER	ERROR PROBABILITY, $P_e(n)$	
γ	M-D	K-S
1	1.3818×10^{-7}	8.7247×10^{-7}
0.2	2.1097×10^{-7}	1.3017×10^{-6}
0.04	2.2973×10^{-7}	1.410×10^{-6}
8×10^{-3}	2.3364×10^{-7}	1.4329×10^{-6}
1.6×10^{-3}	2.3444×10^{-7}	1.4375×10^{-6}
3.2×10^{-4}	2.346×10^{-7}	1.4384×10^{-6}
6.4×10^{-5}	2.3463×10^{-7}	1.4386×10^{-6}
1.28×10^{-5}	2.3464×10^{-7}	1.4386×10^{-6}
2.56×10^{-6}	2.3464×10^{-7}	1.4386×10^{-6}
5.12×10^{-7}	2.3464×10^{-7}	1.4386×10^{-6}
1.024×10^{-7}	2.3464×10^{-7}	1.4386×10^{-6}
2.048×10^{-8}	2.3464×10^{-7}	1.4386×10^{-6}
4.096×10^{-9}	2.3464×10^{-7}	1.4386×10^{-6}
8.192×10^{-10}	3.3464×10^{-7}	1.4386×10^{-6}
1.6384×10^{-10}	2.3464×10^{-7}	1.4386×10^{-6}

TABLE 6.15

Effect of Parameter γ on Error Probability, $P_e(n)$, for Example V of Table 3.1. The Root μ is Appropriately Selected for $\pi_1=\pi_2$.

PARAMETER	ERROR PROBABILITY, $P_e(n)$	
γ	M-D	K-S
1	2.2436×10^{-11}	
0.2	6.1066×10^{-11}	
0.04	7.817×10^{-11}	
8×10^{-3}	8.2237×10^{-11}	
1.6×10^{-3}	8.3079×10^{-11}	
3.2×10^{-4}	8.3249×10^{-11}	
6.4×10^{-5}	8.3283×10^{-11}	
1.28×10^{-5}	8.329×10^{-11}	
2.56×10^{-6}	8.3291×10^{-11}	
5.12×10^{-7}	8.3292×10^{-11}	
1.024×10^{-7}	8.3292×10^{-11}	
2.048×10^{-8}	8.3292×10^{-11}	
4.096×10^{-9}	8.3292×10^{-11}	
8.192×10^{-10}	8.3292×10^{-11}	
1.6384×10^{-10}	8.3292×10^{-11}	

be sufficiently small in relation to the feature dimension n, say $\gamma/n \sim O(10^{-5})$.

6.5. INFLUENCE OF A PRIORI PROBABILITIES ON THE PROBABILITY OF CLASSIFICATION ERROR FOR FINITE SAMPLE SIZE

The probability of classification error, in the limit, has been shown to be independent of the *a priori* probabilities [6.4-6.6]. The discussion here deals with the manner in which the *a priori* probabilities influence the error probability for finite sample size. It has been noted in Section 6.3 that the root μ of (3.34h) is affected by a priori probabilities. π_1, π_2. The root μ is appropriately selected from Tables 6.5-6.10 for Tables 6.16-6.20, where the variations in the value of $P_e(n)$ are presented as a function of *a priori* probabilities for the five examples of Table 3.1. The parameter γ is chosen to be 10^{-4} for all the examples.

We find, from Tables 6.16-6.20, that consistent with classical thought, the probability of classification error is the largest for *a priori* probabilities $\pi_1=\pi_2=0.5$. The error probability is seen to decrease, as expected, as the *a priori* probability for one pattern class in increased in relation to the other. This behavior of the error probability is consistent for all the five examples considered.

6.6. PERFORMANCE ENHANCEMENT BY INCREASING THE NUMBER OF FEATURES

This section deals with observing the improvement achieved in the probability of classification error, $P_e(n)$ by increasing the number

TABLE 6.16

Effect of A Priori Probabilities π_1, π_2 on Error Probability, $P_e(n)$, for Example I of Table 3.1. The Root μ is Appropriately Selected for π_1, π_2 and $\gamma = 10^{-4}$.

A PRIORI PROBABILITIES		ERROR PROBABILITY, $P_e(n)$	
π_1	π_2	M-D	K-S
0.5	0.5	6.0738×10^{-8}	6.1623×10^{-7}
0.55	0.45	6.067×10^{-8}	6.1368×10^{-7}
0.6	0.4	5.9592×10^{-8}	6.0713×10^{-7}
0.65	0.35	5.6358×10^{-8}	6.0021×10^{-7}
0.7	0.3	5.1404×10^{-8}	5.4888×10^{-7}
0.75	0.25	4.939×10^{-8}	4.846×10^{-7}
0.8	0.2	4.0487×10^{-8}	4.4209×10^{-7}
0.85	0.15	3.8308×10^{-8}	4.2427×10^{-7}
0.9	0.1	3.603×10^{-8}	4.0646×10^{-7}
0.95	0.05	3.3751×10^{-8}	3.8865×10^{-7}

TABLE 6.17

Effect of A Priori Probabilities π_1, π_2 on Error Probability, $P_e(n)$, for Example III of Table 3.1. The Root μ is Appropriately Selected for π_1, π_2 and $\gamma = 10^{-4}$.

A PRIORI PROBABILITIES		ERROR PROBABILITY, $P_e(n)$	
π_1	π_2	M-D	K-S
0.5	0.5	2.7963×10^{-8}	
0.55	0.45	2.7625×10^{-8}	
0.6	0.4	2.6823×10^{-8}	
0.65	0.35	2.5436×10^{-8}	
0.7	0.3	2.3883×10^{-8}	
0.75	0.25	2.1786×10^{-8}	
0.8	0.2	1.9306×10^{-8}	
0.85	0.15	1.6072×10^{-8}	
0.9	0.1	1.2153×10^{-8}	
0.95	0.05	7.2513×10^{-8}	

TABLE 6.18

Effect of A Priori Probabilities π_1, π_2 on Error Probability, $P_e(n)$, for Example III of Table 3.1. The Root μ is Appropriately Selected for π_1, π_2 and $\gamma = 10^{-4}$.

A PRIORI PROBABILITIES π_1	π_2	ERROR PROBABILITY, $P_e(n)$ M-D	K-S
0.5	0.5	2.9125×10^{-8}	3.1825×10^{-7}
0.55	0.45	2.8715×10^{-8}	3.1065×10^{-7}
0.6	0.4	2.7938×10^{-8}	2.9943×10^{-7}
0.65	0.35	2.6697×10^{-8}	2.9099×10^{-7}
0.7	0.3	2.5006×10^{-8}	2.645×10^{-7}
0.75	0.25	2.2693×10^{-8}	2.3649×10^{-7}
0.8	0.2	1.9961×10^{-8}	2.0769×10^{-7}
0.85	0.15	1.6651×10^{-8}	1.6834×10^{-7}
0.9	0.1	1.2633×10^{-8}	1.255×10^{-7}
0.95	0.05	7.4401×10^{-9}	7.0829×10^{-8}

TABLE 6.19

Effect of A Priori Probabilities π_1, π_2 on Error Probability, $P_e(n)$, for Example IV of Table 3.1. The Root μ is Appropriated Selected for π_1, π_2 and $\gamma = 10^{-4}$.

A PRIORI PROBABILITIES π_1	π_2	ERROR PROBABILITY, $P_e(n)$ M-D	K-S
0.5	0.5	2.3463×10^{-7}	1.4386×10^{-6}
0.55	0.45	2.3401×10^{-7}	1.4133×10^{-6}
0.6	0.4	2.3319×10^{-7}	1.3887×10^{-6}
0.65	0.35	2.2324×10^{-7}	1.3691×10^{-6}
0.7	0.3	2.0749×10^{-7}	1.3523×10^{-6}
0.75	0.25	1.8492×10^{-7}	1.3185×10^{-6}
0.8	0.2	1.5663×10^{-7}	1.1476×10^{-6}
0.85	0.15	1.2213×10^{-7}	9.1423×10^{-7}
0.9	0.1	8.8289×10^{-8}	6.1748×10^{-7}
0.95	0.05	7.2389×10^{-8}	6.8993×10^{-7}

TABLE 6.20

Effect of A Priori Probabilities π_1, π_2 on Error Probability, $P_e(n)$, for Example V of Table 3.1. The Root is Appropriately Selected for π_1, π_2 and $\gamma = 10^{-4}$.

A PRIORI PROBABILITIES		ERROR PROBABILITY, $P_e(n)$	
π_1	π_2	M-D	K-S
0.5	0.5	8.3278×10^{-11}	
0.55	0.45	8.3188×10^{-11}	
0.6	0.4	8.2928×10^{-11}	
0.65	0.35	7.9876×10^{-11}	
0.7	0.3	7.7326×10^{-11}	
0.75	0.25	7.6024×10^{-11}	
0.8	0.2	7.3841×10^{-11}	
0.85	0.15	7.0001×10^{-11}	
0.9	0.1	6.8327×10^{-11}	
0.95	0.05	6.5016×10^{-11}	

of features n, with the data dimension N fixed. The error probability for each of the five examples of Table 3.1 is shown in Tables 6.21-6.25 as a function of the feature dimension n. The quantity n is varied from 5 to 30 in steps of 5 to provide data compressions of 80% to 25%, respectively. Tables 6.21-6.25 also include the eigenvalue selections and the root, μ for the M-D and K-S method for each value of n. Figures 6.1-6.5 display $\ln P_e(n)$ vs n for the M-D method and a continuous exponential fit to the discrete values of $\ln P_e(n)$. It is noted, however, that this continuous fit is a crude approximation of the behavior of $P_e(n)$ in view of the results in [6.4-6.6] that $P_e(n)$ exhibits a geometric decrease as n is increased.

We observe from Tables 6.21-6.25 an improvement in the error probability is obtained, as expected, by increasing the number of features. However, the improvement in the performance tapers as the number of features becomes large in relation to the data dimension. addition, we note that $P_e(n)$ for both the M-D and K-S methods approach the value for large feature dimensions. It is interesting that, for $n>10$, the M-D and K-S schemes may begin with different eigenvalue selections, but, more often than not, approach a similar selection for $n=30$. The problem with the small feature dimension $(n=5)$ observed he is consistent with the observations of Section 6.2 in that the error probability is large. The <u>most important</u> observation of this section

TABLE 6.21

Probability of Classification Error as a Function of Feature Dimension n for Example I of Table 3.1; $N=40$, $\pi_1=\pi_2$, $\gamma=10^{-5}$.

n	EIGENVALUE SELECTION		ROOT, μ		ERROR PROBABILITY, $P_e(n)$	
	M-D	K-S	M-D	K-S	M-D	K-S
5	$5^\ell/0^s$	$5^\ell/0^s$	-.5525	-.5525	2.3325×10^{-4}	2.3325×10^{-4}
10	$10^\ell/0^s$	$9^\ell/1^s$	-.5525	-.5425	6.0738×10^{-8}	6.1623×10^{-7}
15	$15^\ell/0^s$	$13^\ell/2^s$	-.55	-.54	2.463×10^{-11}	1.8257×10^{-10}
20	$18^\ell/2^s$	$18^\ell/2^s$	-.54	-.54	2.585×10^{-12}	2.585×10^{-12}
25	$22^\ell/3^s$	$22^\ell/3^s$	-.5375	-.5375	2.0552×10^{-13}	2.0551×10^{-13}
30	$26^\ell/4^s$	$25^\ell/5^s$	-.535	-.535	3.8259×10^{-14}	6.1381×10^{-14}

TABLE 6.22

Probability of Classification Error as a Function of Feature Dimension n for Example II of Table 3.2; $N=40$, $\pi_1=\pi_2$, $\gamma=10^{-5}$.

n	EIGENVALUE SELECTION		ROOT, μ		ERROR PROBABILITY, $P_e(n)$	
	M-D	K-S	M-D	K-S	M-D	K-S
5	$5^{\ell}/0^{s}$	$5^{\ell}/0^{s}$	-.605	-.605	1.5866×10^{-4}	1.5866×10^{-4}
10	$10^{\ell}/0^{s}$	$10^{\ell}/0^{s}$	-.605	-.605	2.7963×10^{-8}	2.7963×10^{-8}
15	$14^{\ell}/1^{s}$	$14^{\ell}/1^{s}$	-.5925	-.5925	1.4645×10^{-10}	1.4645×10^{-10}
20	$18^{\ell}/2^{s}$	$19^{\ell}/1^{s}$	-.5875	-.59	7.6887×10^{-13}	1.3403×10^{-12}
25	$23^{\ell}/2^{s}$	$23^{\ell}/2^{s}$	-.5875	-.5875	1.2407×10^{-13}	1.2407×10^{-13}
30	$27^{\ell}/3^{s}$	$27^{\ell}/3^{s}$	-.585	-.585	5.8262×10^{-14}	5.8262×10^{-14}

TABLE 6.21

Probability of Classification Error as a Function of Feature Dimension n for Example III of Table 3.1; $N=40$, $\pi_1=\pi_2$, $\gamma=10^{-5}$.

n	EIGENVALUE SELECTION		ROOT, μ		ERROR PROBABILITY, $P_e(n)$	
	M-D	K-S	M-D	K-S	M-D	K-S
5	$5^{\ell}/0^{s}$	$4^{\ell}/1^{s}$	-.6075	-.5475	1.0756×10^{-4}	4.8363×10^{-4}
10	$10^{\ell}/0^{s}$	$9^{\ell}/1^{s}$	-.6025	-.5725	2.9125×10^{-8}	3.1065×10^{-7}
15	$15^{\ell}/0^{s}$	$14^{\ell}/1^{s}$	-.6	-.58	1.084×10^{-11}	2.1954×10^{-10}
20	$20^{\ell}/0^{s}$	$19^{\ell}/1^{s}$	-.6	-.5825	7.9939×10^{-14}	1.5992×10^{-13}
25	$25^{\ell}/0^{s}$	$24^{\ell}/1^{s}$	-.5975	-.5825	8.9209×10^{-15}	1.8309×10^{-14}
30	$29^{\ell}/1^{s}$	$29^{\ell}/1^{s}$	-.5825	-.5825	2.5425×10^{-15}	2.5425×10^{-15}

TABLE 6.24

Probability of Classification Error as a Function of Feature Dimension n for Example V of Table 3.1; N=40, $\pi_1=\pi_2$, $\gamma=10^{-5}$.

n	EIGENVALUE SELECTION		ROOT, μ		ERROR PROBABILITY, $P_e(n)$	
	M-D	K-S	M-D	K-S	M-D	K-S
5	$4^{\ell}/1^{s}$	$1^{\ell}/4^{s}$	-.4825	-.415	1.009×10^{-4}	5.8279×10^{-3}
10	$8^{\ell}/2^{s}$	$6^{\ell}/4^{s}$	-.4875	-.4625	2.3463×10^{-7}	1.4386×10^{-6}
15	$13^{\ell}/2^{s}$	$11^{\ell}/4^{s}$	-.505	-.4875	4.8537×10^{-10}	5.3816×10^{-9}
20	$17^{\ell}/3^{s}$	$16^{\ell}/4^{s}$	-.5025	-.4975	5.0426×10^{-11}	2.8976×10^{-10}
25	$22^{\ell}/3^{s}$	$20^{\ell}/5^{s}$	-.5075	-.5025	5.1384×10^{-12}	9.6727×10^{-11}
30	$26^{\ell}/4^{s}$	$25^{\ell}/5^{s}$	-.505	-.505	1.0933×10^{-12}	4.2809×10^{-11}

TABLE 6.25

Probability of Classification Error as a Function of Feature Dimension n for Example V of Table 3.15; N=40, $\pi_1=\pi_2$, $\gamma=10^{-5}$.

n	EIGENVALUE SELECTION		ROOT, μ	ERROR PROBABILITY, $P_e(n)$
	M-D	K-S	M-D and K-S	M-D and K-S
5	$0^{\ell}/5^{s}$	$0^{\ell}/5^{s}$	-.25	9.0785×10^{-6}
10	$0^{\ell}/10^{s}$	$0^{\ell}/10^{s}$	-.25	8.3278×10^{-11}
15	$0^{\ell}/15^{s}$	$0^{\ell}/15^{s}$	-.25	5.912×10^{-15}
20	$0^{\ell}/20^{s}$	$0^{\ell}/20^{s}$	-.2525	4.2024×10^{-16}
25	$0^{\ell}/25^{s}$	$0^{\ell}/25^{s}$	-.255	1.4151×10^{-16}
30	$0^{\ell}/30^{s}$	$0^{\ell}/25^{s}$	-.2575	1.1766×10^{-16}

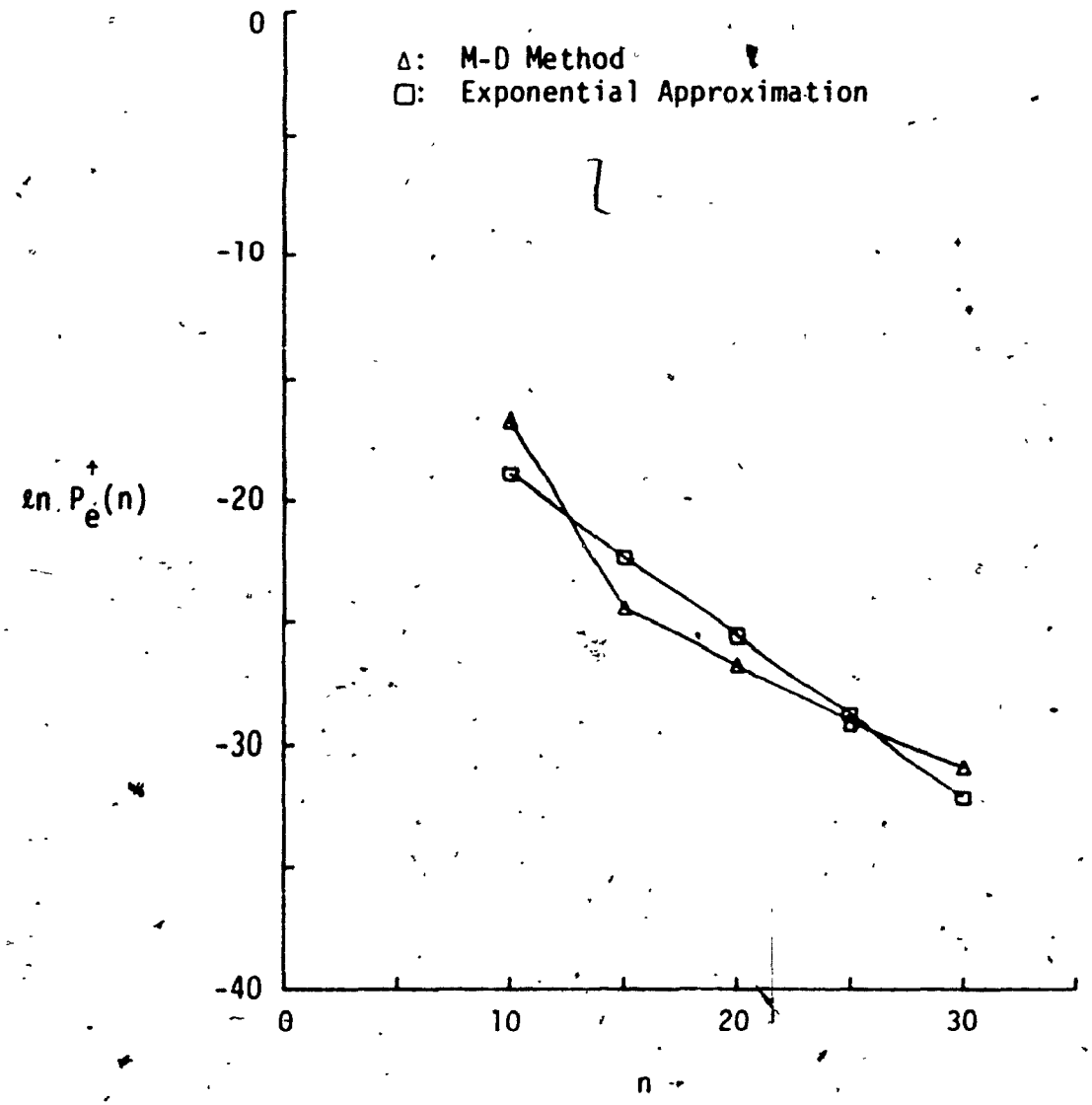

Figure 6.1 Ln $P_e(n)$ vs. n. Covariance Example I.

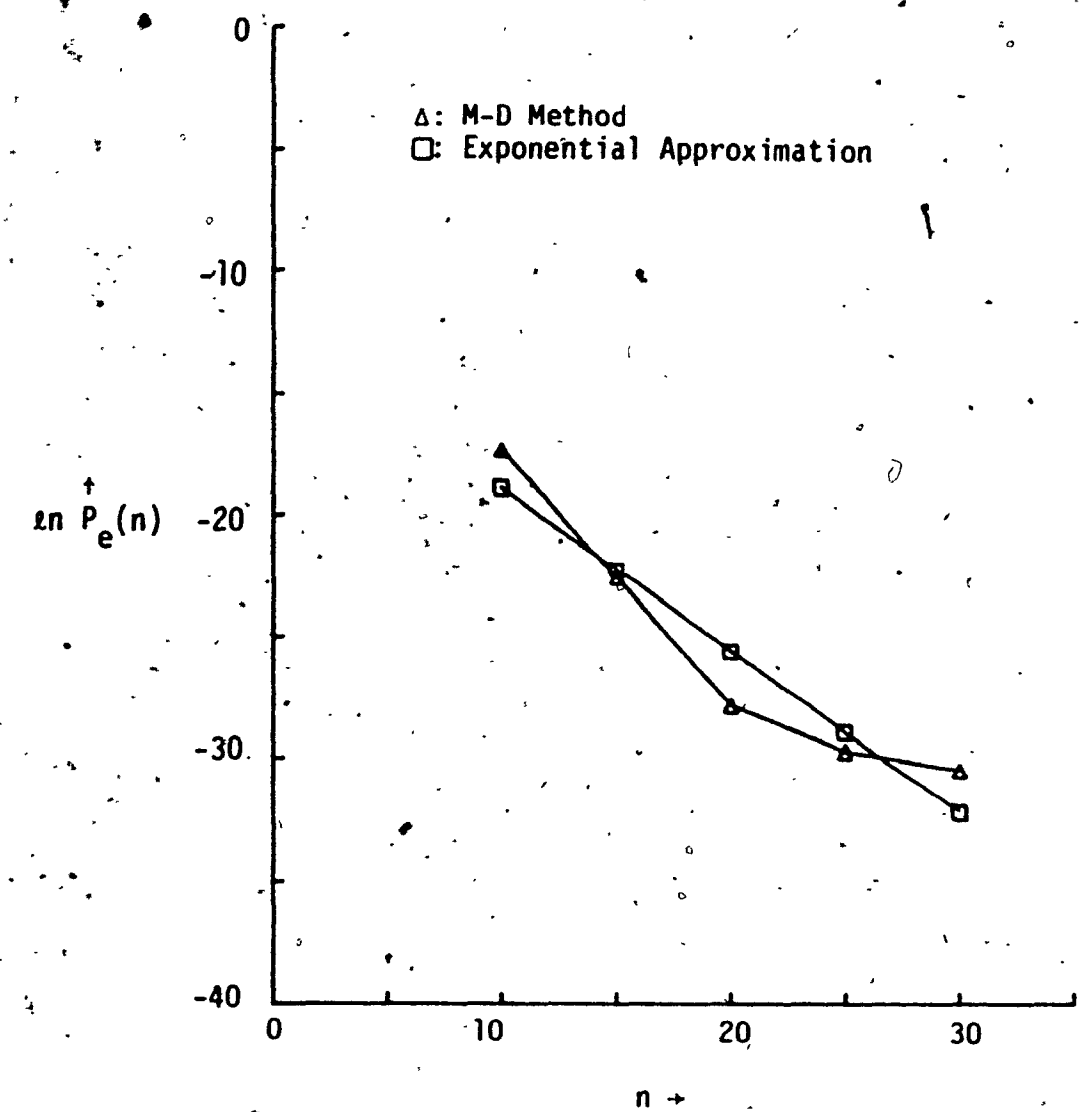

Figure 6.2 Ln $P_e(n)$ vs. n. Covariance Example II.

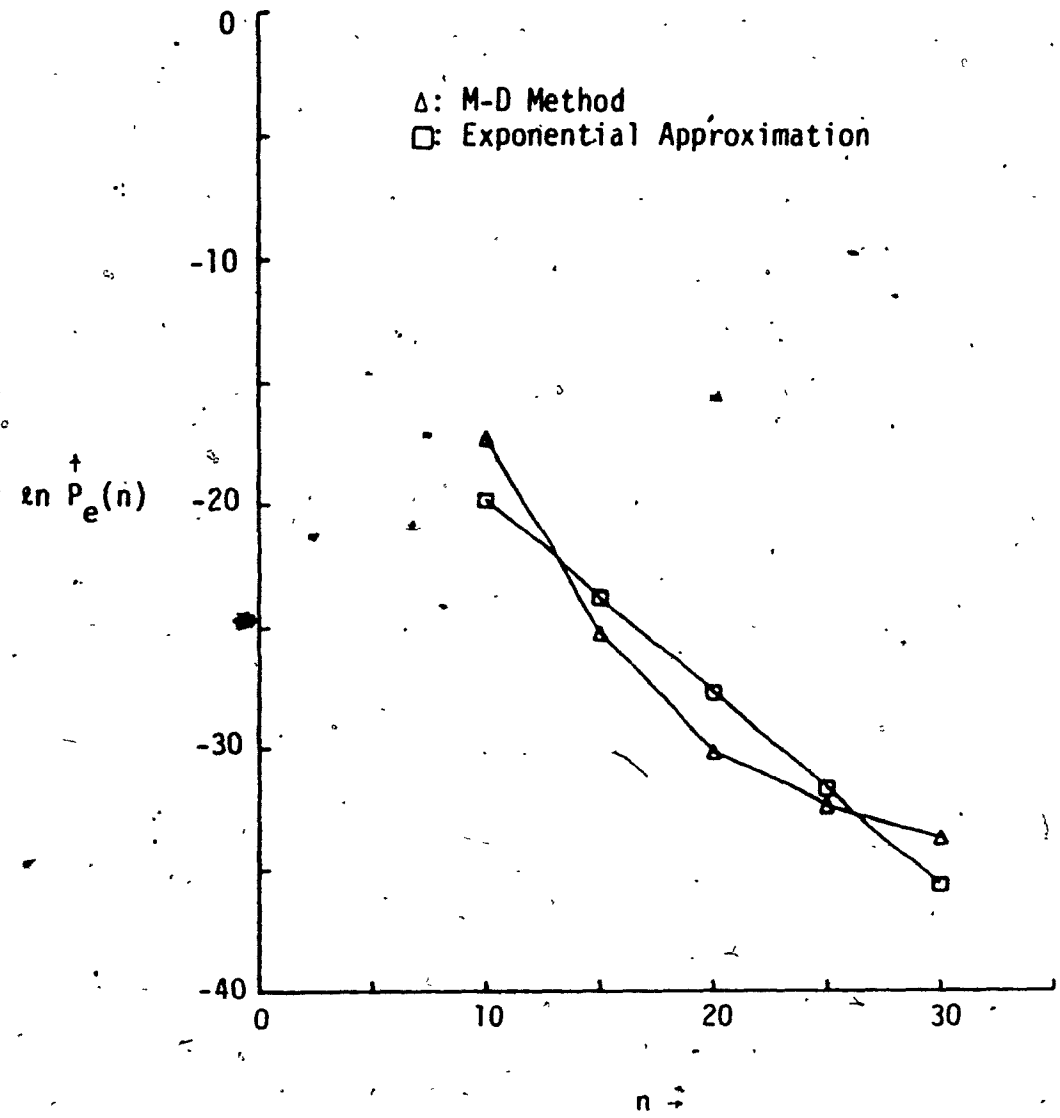

Figure 6.3 Ln $P_e(n)$ vs. n. Covariance Example III

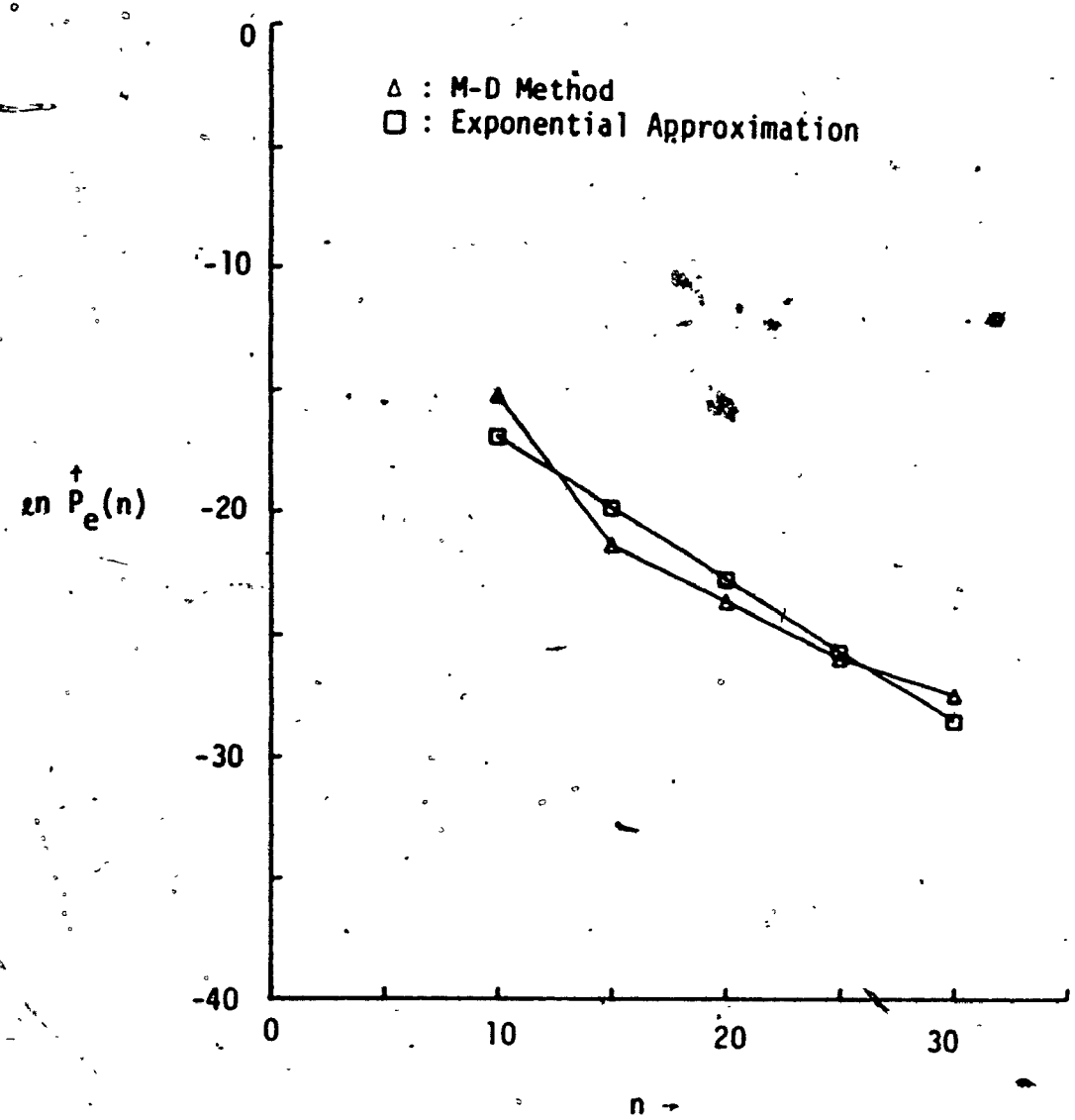

Figure 6.4 Log $P_e(n)$ vs n. Covariance Example IV.

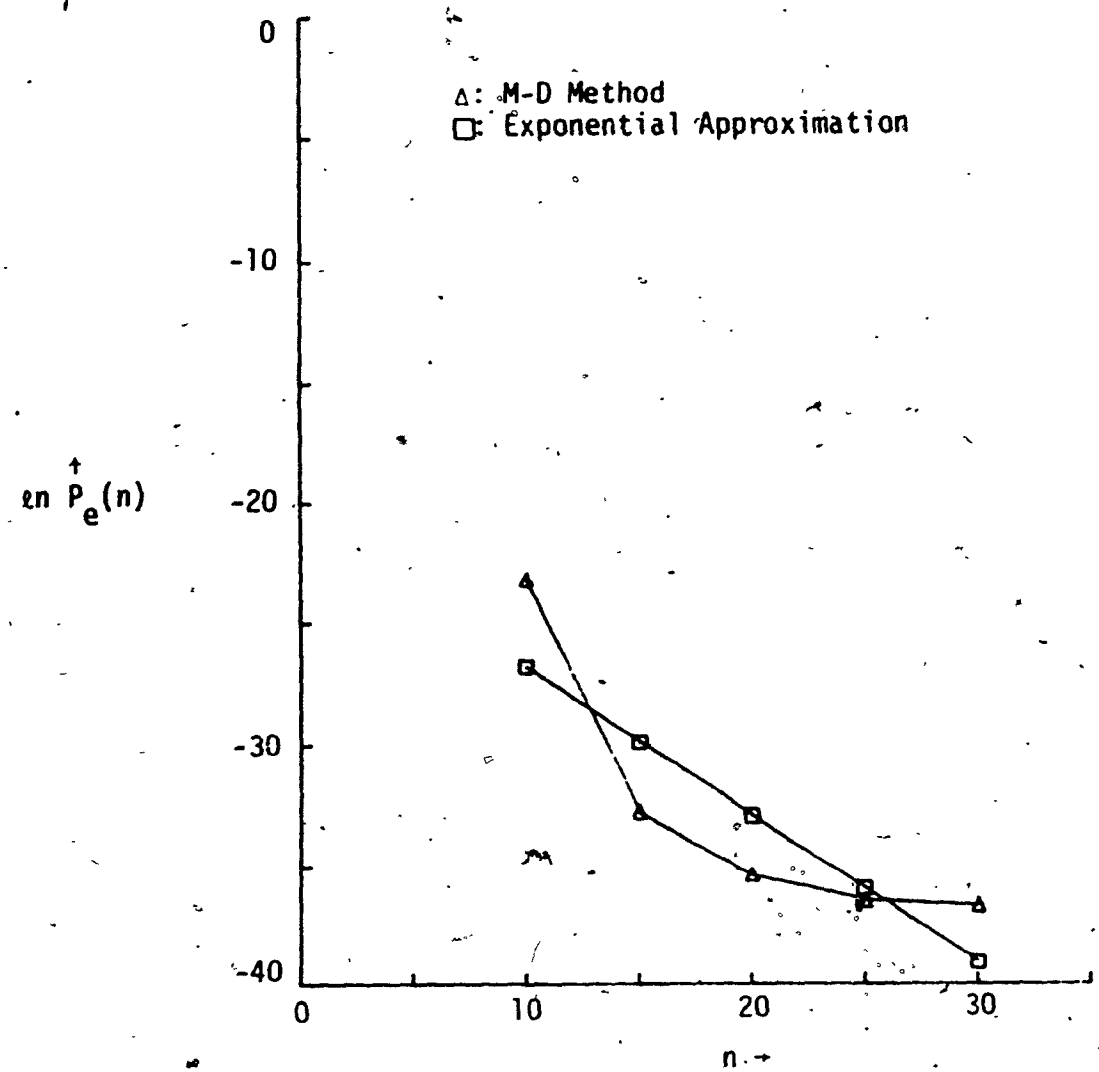

Figure 6.5 Log $P_e(n)$ vs n. Covariance Example V.

is that regardless of the feature dimension the eigenvalue selection made by the M-D method is always at least as good as and sometimes better, by an order of magnitude, than the K-S method in terms of the probability of classification error.

6.7. DISCUSSION

This chapter takes a detailed look, by means of numerical examples, at the probability of error expression derived in Chapter 3 for feature selection in order to provide a better understanding and insight to the problem. Computer simulations of a pattern classifier indicate that the eigenvalue selection provided by the Morgera-Datta method is always at least as good as and sometimes better than the conventional Kadota-Shepp method. The results also demonstrate that the error bounds provided by the Bhattacharyya distance measure are quite "loose" and, at times, fail to provide any useful information for feature selection.

We also study in detail the effect of certain parameters on the error probability. The manner in which the *a priori* probabilities of the pattern classes affect the error probability is consistent with classic thought, i.e., an enhanced error performance is observed as the *a priori* probability of one class increases in relation to the other. The effect of the *a priori* probabilities on error probability is compounded with their influence on the root, μ.

The reasons of restricting the value of the parameter γ are

discussed. It is observed that for sufficiently small values of γ relative to the feature dimension n, the probability of error expression is independent of γ. Since the parameter γ is artificial introduced in the error expression, it is recommended that γ must be chosen such that $\gamma/n < (10^{-5})$ to eliminate its influence on the error probability. In passing, we mention that the parameter γ does not play any role in the actual implementation of the feature selection strategy.

We examine the behavior of the error probability $P_e(n)$ as a function of the feature dimension n. It is observed, as expected, that $P_e(n)$ decreases as n increases. The results also indicate that the M-D and K-S methods may converge to the same error probability for the large feature dimensions.

REFERENCES

[6.1] T.T. Kadota and L.A. Shepp, "On the Best Finite Set of Linear Observables for Discriminatring Two Gaussian Signals," IEEE Trans. Inform. Theory, vol. IT-13, pp. 278-284, Apr. 1967.

[6.2] T. Kailath, "The Divergence and Bhattacharyya Distance Measures in Signal Selection," IEEE Trans. Comm., vol. COM-15, pp. 52-60, Feb. 1966.

[6.3] L.J. Griffiths, "Signal Extraction Using Real-Time Adaptation of Linear Multichannel Filter," Tech. Report 6788-1, Syst. Theory Lab., Stanford, Feb. 1968.

[6.4] U. Grenander, Abstract Inference. New York: Wiley, 1981.

[6.5] U. Grenander, "Large Sample Discrimination Between Two Gaussian Processes with Different Spectra," Annals of Statistics, vol. 2, pp. 347-352, 1974.

[6.6] D. Kazakos, "On Resolution and Asymptotic Discrimination Between Gaussian stationary Vector Processes and Dynamic Models," IEEE Trans. Automat. Contr., vol. AC-25, pp. 294-296, Apr. 1980.

[6.7] R.O. Duda and P.E. Hart, Pattern Classification and Scene Analysis. New York: Wiley, 1973, pp. 73-76.

CHAPTER 7

CONCLUSIONS

7.1. INTRODUCTION

This work has dealt with the difficult problem of discriminating two weakly stationary complex Gaussian stochastic processes with similar mean values and different covariances (patterns). High Dimensionality of data so often encountered in practical pattern recognition applications necessitates that the classification task be preceded by feature selection. Dimensionality reduction of data vectors is performed by feature selection, a mechanism for selecting only those features which are most relevant to the classification objective, i.e., in such a manner that the probability of classification error is minimized. A minimization of the error probability is often either mathematically or computationally difficult to carry out. This general feeling among many researchers in the areas of pattern recognition, communication, control systems, and information theory has led to the use of suboptimal and sometimes ad hoc feature selection strategies. The schemes are devised to optimize a certain criterion instead of dealing with the classification error, for example, optimizing a statistical distance measure like the Bhattacharyya distance between patterns which provides bounds on the error probability. In addition, most schemes provide only asymptotic results. A brief discussion on the available feature selection schemes is presented in Chapter 2 to emphasize the need for a feature selection technique which deals more directly with

the error probability expression and exhibits reasonably accurate results for finite sample size.

7.2. CONCLUDING REMARKS ON THE THESIS

In this work, a new feature selection scheme, referred to as the M-D method, has been proposed in Chapter 3. This scheme deals directly with the Bayes error probability expression. The results are obtained by utilizing classical results on discrimination, information theory, and concepts of distribution function theory. The work demonstrates the suboptimality of certain statistical distance measures employed in the feature selection or data compression mechanisms as opposed to a direct application of the error probability expression. It is noted that for extremely small error probabilities, simulation is impractical. In this regime of importance, the proposed technique provides a tool not only for feature selection itself, but for a general performance evaluation of competing pattern classification schemes. Some interesting aspects of the theory merit a brief mention.

Using concepts of distribution function theory, the empirical distribution functions (df's) of the eigenvalues of the covariance matrix pair associated with the two hypotheses are defined for both the N-dimensional data space, $F_N(x)$, and the n-dimensional feature space, $F_n^*(x)$. The quantity $n/N<1$ is defined as the compression ratio. A detailed characterization of the eigenvalues of one covariance relative to the other in both the feature and data spaces, not possible using statistical distance measures, is permitted by the empirical df's.

A family of extremal df's F^* is defined in the feature space.

The family F^* contains a maximum of $(n+1)$ members with a typical member $F^*_{n;\alpha}(x)$ being parameterized by the quantity α. However, only one member of F^* yields minimum error probability. The problem of feature selection is reformulated to that of determining the parameter α or, equivalently, the df $F^*_{n;\alpha}(x)$ such that error probability is minimized. The optimal feature selection strategy is to choose n data eigenvectors corresponding to the $n\alpha$ smallest and $n(1-\alpha)$ largest eigenvalues as the n rows of an $(n\times N)$-dimensional transformation matrix which premultiplies the input data vectors.

The theory is developed by following the excellent asymptotic approach of Grenander and the results of Laplace to obtain an error probability expression accurate for finite sample size. The error expression contains coordinates dependent on the root μ of an equation involving the threshold for hypothesis testing. The value of the root μ depends on the family F^* and lies in a restricted range of $(-1,0)$. This fact sets the results presented here apart from, say, the conventional Kadota-Shepp (K-S) method for feature selection developed by examining the extremal points of a statistical distance measure known as J-divergence. The suboptimality of the K-S method also become apparent by noting that, in the limit as $N\to\infty$ with n/N fixed, the method requires that the root μ be $-\frac{1}{2}$ with no regard to the family F^*.

The final step in the development of the theory is the determination of the extremal point of the error expression with respect to the

parameter α. This leads to a functional taking the form of an information content (negative entropy) of discarded eigenvalues. This functional is minimized in order to maximize the information content of the eigenvalues (eigenvectors) to be retained. The performance of the scheme is shown, by extensive use of examples, to be at least as good as, and sometimes better, by an order of magnitude, than conventional techniques. The ideas presented here are also discussed in the context of previous inferences drawn from statistical communications theory.

Performance of the new feature selection strategy in terms of error probability encouraged an investigation into the computational complexity of the scheme to assess its feasibility for practical applications. Some well-known and a variety of new algorithms are proposed which are suitable for realizing an efficient computer-based feature selection and pattern classification system. In the process, we present the characteristic equation reducibility results on two class of matrices, viz., centrosymmetric (CS) and centrohermitian (CH) matrices. These matrices are encountered in pattern recognition feature selection among many other areas, such as antenna theory, electrical and mechanical systems, and quantum physics. The reducibility results on CS and CH matrices provide a significant savings in the process of principal component (eigenvalue/eigenvector) extraction required by the feature selection strategy. In addition, a method of approximating Toeplitz covariances by circulant matrices is discussed for a further reduction in the complexity. This approach leads to satisfactory classification error when there is sufficient statistical independence within each data vector, and allows the

principal component extraction to be replaced by the discrete Fourier transformation (DFT). Several efficient algorithms such as the fast Fourier transformation (FFT) and Winograd fast transformation algorithm (WFTA) are available for computing the DFT.

A detailed complexity analysis of each step in the pattern classification process is presented. The proposed configuration utilizing well-known and new algorithms render the computer-based pattern classification system quite appealing for practical applications.

The proposed computer-based pattern classifier consists of three modes of operation, viz., the training mode, the processing mode, and the decision-directed mode. The training mode is used to perform feature selection and establish the system parameters. The actual classification of patterns takes place in the processing mode. The pattern classifier configuration also provides a means of taking into account a realistic quasi-stationarity of patterns, in which case, the features evolve slowly with time. The decision-directed mode updates the system parameters from one interval of stationarity to the next via the training mode.

A variety of numerical results based on a computer simulation of the pattern classifier are included. The new feature selection scheme is consistently found to be at least as good as, and frequently better, than the conventional K-S method. The error bounds obtained, for example, by using the Bhattacharyya distance are repeatedly found to be quite loose and, thereby, failing to provide any meaningful deductions for feature selection.

Some extremely important numerical results on the theory are presented in view of classical thoughts. For example, an improved error performance may be obtained by increasing the number of features. The usual reduction in the error probability is also observed when the a priori probability of one pattern is increased in relation to the other. The numerical exexamples are felt to be an important contribution because ideas are not readily accepted in pattern recognition, particularly, if the theory is highly complex. Although the examples considered are confined to those associated with weakly stationary stochastic processes, it is important to note that the discrimination theory presented is generally applicable to other underlying probability structues; however, the implementation complexity has to be reexamined accordingly.

This work on the development of a feature selection strategy and its computer based implementation shall prove useful in stochastic signal classification problems such as encountered in image processing and robotics. It is felt that the new approach to feature selection may contribute to a better understanding of the problem.

7.3. IDEAS FOR FUTURE WORK

Performance of the feature selection and pattern classification system proposed here is quite attractive in terms of the error proprobability and computational complexity. It would be useful and interesting to develop a VLSI implementation of the proposed system. The system shall be useful for many applications, e.g., image analysis/-

object recognition, speech analysis/speaker recognition and robotics.

The development of a VLSI architecture requires a considerable effort, A number of matrix arithmetic networks must be developed for large-scale matrix computations such as multiplication, inversion, and principal component extraction. The efficiency of networks may be significantly enhanced by utilizing special covariance structures, e.g., Toeplitz and circulant approximation of Toeplitz matrices.

A complete study on VLSI implementation of the new scheme would consist of the following tasks. A study of various relevant VLSI arc-architectures avaliable is needed. This may be followed by the development of new architectures and a comparison of various alternatives in terms of the speed and space requirements of the system. The goal of the proposed study is, of course, to develop a VLSI layout and specifications for custom chip(s).

The proposed computer-based implementation of the pattern classifier is somewhat restrictive in that it assumes certain a priori structures of covariances, i.e., Toeplitz and circulant approximations. Although these structures are frequently encountered in practice and are of extreme practical interest, it would be worthwhile to expand the scope of this work to, say, near Toeplitz or other useful covariance structures. However, computational complexity of such an implementation must be of prime concern in view of its practical utility.

An interesting class of covariances to which the proposed implementation may be extended is that of separable covariances. Separable covariance matrices are used to represent certain two

dimensional random fields where an appropriate choice of coordinates for horizontal and vertical components renders the row and column covariances uncoupled. Such covariances arise in image analysis for texture and seismic signature analysis. A separable covariance matrix S has the following form,

$$S = C \otimes R$$

where C is the (M×M)-dimensional within-column covariance matrix, R is the (N×N)-dimensional within-row covariance matrix, and the operator $\otimes$ denotes the Kronecker product.

Let S_1 and S_2 be the separable covariance matrices representing the two weakly stationary Gaussian stochastic processes under hypothesis H_1 and H_2, respectively. The proposed feature selection scheme requires the principal components of the product matrix $S_1^{-1}S_2$ to obtain the transformation matrix which premultiplies the input data vectors for dimensionality reduction. The product matrix $S_1^{-1}S_2$ exhibits the separability property similar to that of the individual covariances S_1 and S_2, i.e.,

$$S_1^{-1}S_2 = C_1^{-1}C_2 \otimes R_1^{-1}R_2$$

where C_i and R_i relate to the separable covariance S_i i=1,1 as within-column and within-row covariances, respectively. Now assume that the matrices C_i and R_i are Toeplitz as found in a variety of practical applications. The results of Chapter 4 on efficient principal component extraction, and Chapter 5 on the computer based implementation for pattern classification are applicable for the realization of an efficient clasifier.