The Laboratory for Applications of Remote Sensing

The Minimum Distance Approach to Classification

A. G. Wacker

D. A. Landgrebe

Purdue University

Information Note 100771

LARS Information Note 100771

THE MINIMUM DISTANCE APPROACH TO CLASSIFICATION[1]

A. G. Wacker
D. A. Landgrebe

TR-EE 71-37
October, 1971

Published by the
Laboratory for Applications of Remote Sensing (LARS)
and
the School of Electrical Engineering
Purdue University
Lafayette, Indiana 47907

[1]This work was supported by the National Aeronautics and Space Administration under Grant No. NGL 15-005-112.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

Wacker, Arthur Gordon, Ph.D., Purdue University, January 1972. Minimum Distance Approach to Classification. Major Professor: D. A. Landgrebe.

In minimum distance classification a group of vectors (sample), known to belong to the same class, is classified into the class whose known or estimated distribution most closely resembles the estimated distribution of the sample to be classified. The measure of resemblance is a distance measure in the space of distribution functions.

The general objective of this work is to advance the state of the art of minimum distance classification. This is accomplished through a combination of some theoretical investigations and a comprehensive experimental investigation based on multispectral scanner data. A thorough survey of the literature for suitable distance measures was conducted and the results of this survey are presented.

Theoretically it is shown that minimum distance classification, using density estimators and Kullback-Leibler numbers as the distance measure, is equivalent to a form of maximum likelihood sample classification. It is also shown that for the parametric case minimum distance classification is equivalent to nearest neighbor classification in the parameter space.

A two class univariate normal problem, in which the set of distributions representing each class is described by a distribution over the parameter space, is analysed for various amounts of overlap of the parameter space densities.

A theoretical investigation of a new separability measure defined in terms of random samples provides insight into some experimentally observed effects of dimensionality.

The experimental investigation of minimum distance classification is based on a supervised parametric (normal) minimum distance classifier PERFIELD and a supervised non-parametric minimum distance classifier (using histogram estimators) LARSYSDC. Each classifier is capable of using any one of three distance measures with only one distance measure common to both classifiers. Classification accuracy of a parametric (normal) maximum likelihood vector classifier is also compared experimentally with minimum distance classification.

In cases where the training set contains a large number of samples, parameter space clustering is experimentally investigated as a technique for combining similar samples.

The principal experimental results pertaining to minimum distance classification of multispectral scanner data are:

1) The Jeffreys-Matusita distance (defined as the square root of the integral squared difference of the square root of two densities) appears to be a good general purpose distance measure.

2) The minimum distance classification accuracy (% samples correct) was typically 5 to 10% greater than the maximum likelihood vector classification accuracy (% vectors correct). Improvements as great as 15% have been observed. The improvement depends on the degree of overlap of the parameter space densities.

3) For the techniques used to define training samples no distance measure was consistently superior for classifying test samples. Neither was the nonparametric classifier LARSYSDC superior to the parametric classifier PERFIELD in these circumstances. For classifying training samples the nonparametric classifier was slightly superior as were certain distance measures.

4) The effect on classifier performance of the number of spectral channels, the number of vectors in a test sample, and the histogram bin size for the nonparametric classifier LARSYSDC are also experimentally investigated. For the data considered classifier accuracy can be improved only slightly by using more than 4 channels and test samples containing more than 60 vectors. The results show that test samples for the nonparametric classifier need not be larger than for the parametric classifier. A bin size of 5 to 10 is indicated.

CHAPTER 1
INTRODUCTION

Making measurements and categorizing objects on the basis of these measurements is an essential aspect of knowledge, and consequently an essential aspect of all sciences. Thus to cite two arbitrary examples from the science of astronomy: A star is classified as a red giant because of its physical size and spectral characteristics; a pulsar is identified primarily by the periodicity in its radiation. Numerous other examples abound in astronomy and all other scientific fields.

A frequent requirement in the categorization process is the ability to manipulate data and carry out computations. Consequently it is not surprising that with the advent of computers man quickly turned to them for assistance in the classification task. Thus evolved the field of pattern recognition which is precisely concerned with the problem of classification or labeling objects on the basis of a set of measurements, usually with the aid of a machine. Many different classification schemes have evolved over the years. Minimum distance classification is one such scheme. In a certain sense minimum distance classification resembles what is probably the simplest approach to pattern recognition,

namely "template matching". In template matching a template is stored for each class of patterns to be recognized (e.g. letters in the alphabet) and an unknown pattern (e.g. an unknown letter) is then classified into the pattern class whose template best fits the unknown pattern on the basis of some previously determined similarity measure. In minimum distance classification the templates and unknown patterns are distribution functions and the measure of similarity used is a distance measure between distribution functions. Thus an unknown distribution function is classified into the class whose distribution function is nearest to the unknown distribution in terms of some predetermined distance measure.

Normally, in practically problems, it is not the distribution function itself that is observed, rather a random set of measurement vectors drawn from the distribution are observed. Consequently, before the distribution function can be classified it must be estimated from a set of observed vectors. It is possible to adopt the view that when a distribution function is classified then in effect all the vectors used to estimate that distribution function are classified. Thus minimum distance classification belongs to a set of classification schemes that we refer to as "sample classification schemes". A basic premise in sample classification schemes is that the vectors to be classified appear in groups or samples, where it is known a priori, or

where it is reasonable to assume, that each vector in the group belongs to the same class. Sample classification schemes contrast with the more conventional pattern recognition schemes where each measurement vector is classified individually.

Our interest in minimum distance classification was prompted by work in the field of Remote Sensing of earth resources. Fu et al.[1] state that "remote sensing technology is primarily concerned with the identification or classification of physical objects through the analysis of these objects made with sensors that are at some distance from the objects". Although not specifically stated it is implied that these measurements are made without coming into physical contact with the objects, and that the information is conveyed from the distant object to the sensor by some force field. Specifically it is the variation of some force field with some parameter such as space, or time, or in the case of electromagnetic radiation wavelength, that conveys the information. Although remote sensing has only recently been identified as a distinct technology, some remote sensing techniques have been in use for many years. Photography is an example of one such technique.

At the present time in the development of remote sensing technology it is possible to identify a duality in the system types utilized. Landgrebe[2] refers to the two types as "image-oriented systems" and "numerically-oriented

systems". The duality exists primarily for historical reasons as a consequence of the independent development of photographically oriented and computer oriented technology. In image-oriented systems a visual image is an essential part of the analysis scheme, while in numerically-oriented systems the visual image plays a secondary role, and may in fact not even be formed. For example an astronomer studying the temporal variation in illumination of a pulsar might conceivably do so by examining a sequence of photographs (an image-oriented system). On the other hand a radio astronomer observing the radio wave-length properties of the same pulsar would probably never generate an image of the star (a numerically oriented system).

In numerically-oriented remote sensing systems it is frequently possible to design the data collection system in such a manner that classification becomes a problem in pattern recognition. This situation prevails if one attempts to study earth resources through the utilization of "multispectral data-images" which is a basic premise on which the research at Purdue's Laboratory for Applications of Remote Sensing (LARS) is based.

The term multispectral data-image requires elaboration. By multispectral image, (i.e. without the modifier "data") we mean two or more spectrally different, superimposed, pictorial images of a scene. The modifier data is added to indicate that the images are stored as

numerical arrays, as opposed to visual images. To obtain a multispectral data-image of a scene, the scene in question is partitioned into small cells and the radiance from each cell, for each wave-length band of interest is measured and stored. We call these cells image resolution elements (IRE's). In other words a multispectral data-image of a scene is an array of measurement vectors, one from each IRE in the scene. The components of the measurement vectors are the radiances observed when viewing the scene through different spectral windows. The spatial coordinates of the IRE are of course also recorded to uniquely identify each measurement vector.

The method of processing multispectral data-images depends on the information being sought. A rather common goal is that of segregating the measurement vectors into a number of classes. For example one may wish to identify crop species in an agricultural scene. In the more conventional pattern recognition schemes each measurement vector would be analysed individually and classified into one of the classes of interest on the basis of some classification rule. In a sample classification scheme, like the minimum distance rule, all vectors to be classified are first segregated into groups, such that all the vectors in a group belong to the same class, and then the group is classified. Note there are two distinct aspects to the problem of minimum distance classification. The first is concerned with

partitioning measurement vectors into homogeneous groups, while the second is concerned with the classification of the groups.

It is clear that for minimum distance classification to be most useful automatic methods must be devised for defining samples (i.e. groups of measurement vectors). While we recognize the importance of this problem, and have done some work on it, we will primarily concern ourselves with only the classification aspect of the problem. We do, however, wish to make a few comments regarding definition of samples.

It frequently occurs for multispectral data-images that many of the adjacent measurement cells belong to the same class. For example in an agricultural scene each physical field typically contains many measurement cells. In fact it is precisely this condition that prompts the investigation of minimum distance classification. In such situations the physical field boundaries serve to define suitable samples for problems like crop species identification, and it is on this basis that minimum distance classification is also referred to as per-field classification. It is apparent that for the situation just described one method of automatically defining samples is to devise a scheme that automatically locates physical field boundaries in the multispectral data-imagery. In this investigation of minimum distance classification physical

field boundaries will actually be used to define the samples, but the field boundaries will be located manually rather than automatically. A second and perhaps more promising approach to the problem of defining samples is via observation space clustering. In this approach vectors from an arbitrary area are clustered in the observation space, and all the vectors assigned to the same cluster constitute a sample irrespective of their location in the arbitrary choosen area. In this case the term "fields" no longer seems appropriate and consequently the term sample classification is preferred over the term per-field classification.

It is apparent that minimum distance classification (or any other sample classification scheme) cannot be used in all situations where a vector by vector approach is possible. A basic requirement is that the data to be classified can either be segregated into homogeneous samples, or occurs naturally in this form. Where the minimum distance scheme can be applied it has several potential advantages over a vector by vector classifier; in particular it is potentially faster and more accurate.

It seems logical that provided the time required to automatically define the samples is not too great, then a minimum distance classifier should be faster than a vector by vector classifier. This is of considerable importance in utilizing a numerically-oriented remote sensing system

to survey earth resources because a characteristic of such surveys is the tremendous volume of data involved. One would also anticipate that the vector classification accuracy of a vector by vector classifier would be lower than the sample classification accuracy for minimum distance classification. The reason for this is that in minimum distance classification all the information conveyed by a group of vectors is used to establish the classification of each vector whereas in a vector by vector classifier each vector is treated separately without reference to any other vector. In a sense minimum distance classification utilizes spatial information because vectors are classified as groups, which naturally have some spatial extent. No spatial information is used in vector by vector classifiers, consequently, minimum distance classification should perform better since spatial information is certainly of some value.

The objectives of this investigating of minimum distance classification can now be stated. The primary objective is to experimentally assess minimum distance classification as a method of classifying multispectral data-images under the basic assumption that all samples are manually defined. An important aspect of the investigation is the comparison of various distance measures as well as a limited parametric vs non parametric assessment of minimum distance classification.

Page Intentionally Left Blank

CHAPTER 2
MINIMUM DISTANCE CLASSIFICATION

In this section we introduce the necessary definitions and notation to formulate the minimum distance classification rule in a decision theoretic framework The diverse literature pertaining to minimum distance classification and distance measures is reviewed and discussed utilizing consistent notation and terminology.

2.1 Basic Concept of the Minimum Distance Classification Procedure

Distance between cdf's is the basic concept upon which the proposed classification scheme is based. In a mathematical sense the terms "distance" and "metric" are sometimes used interchangeably. A metric on a set S is, of course, a real valued function δ defined on S X S (X indicates cartesian product) such that for arbitrary F,G,H in S

$$(a) \quad \delta(F,G) \geq 0 \tag{2.1.1}$$

$$(b)(1) \quad \delta(F,F) = 0 \tag{2.1.2}$$

$$(2) \quad \text{If } \delta(F,G) = 0 \text{ then } F = G \tag{2.1.3}$$

$$(c) \quad \delta(F,G) = \delta(G,F) \tag{2.1.4}$$

$$(d) \quad \delta(F,G) + \delta(G,H) \geq \delta(F,H) \tag{2.1.5}$$

We will not consider the terms "metric" and "distance" to be synonomous, rather we will assume that a

distance has some, though not necessarily all, the properties of a metric. Specifically we will assume a distance on a set S is a real valued function d on S X S such that for arbitrary F,G,H, in S at least metric properties (a), (b) (1), and usually (c) hold. We will specifically point out those instances were (c) is assumed not to hold.

To describe the basic concept of the minimum distance method we consider a particular case. The method is formulated in a more general and rigorous manner in the next section. We assume that the ith class is characterized by a known q-variate cdf $F^{(i)}$, i = 1,2,...,k. Let $\Omega = \{F^{(1)}, F^{(2)},\ldots,F^{(k)}\}$. To classify an unknown sample of N random vectors drawn from a population with cdf F (where $F = F^{(i)}$ for some i) we compute the emperic cdf $\overline{F}_N$ and assign the sample to the ith class in case

$$d(\overline{F}_N, F^{(i)}) = \min_{j=1,\ldots,k} d(\overline{F}_N, F^{(j)}) \qquad 2.1.6$$

It appears that it should be possible, under suitable conditions, to adopt the point of view that this decision rule is a version of the well known nearest neighbor rule[3], except that the items being classified are emperic cdf's representing the class from which the sample (group of vectors) originated rather than vectors representing individual patterns. The validity of this contention is established in Chapter 3 for the parametric case. The nearest neighbor viewpoint seems particularly appealing both theoretically and practically. From a theoretical point of view

it means that theoretical results in connection with nearest neighbor decision rules[3,4,5,6] are directly applicable. From a practical point of view it immediately becomes very logical to view subclasses as different "sample points" (a "sample point" in this context is an emperic cdf) representing the particular class in question. These concepts will subsequently be formulated in a formal manner and their validity and resultant implications investigated.

The decision rule as given above is completely nonparametric. The intention is, however, to investigate the rule in a parametric as well as a nonparametric setting. In the parametric setting the cdf's are assumed to have some parametric form (e.g. q-variate normal) and hence $\Omega = \{F^{(1)}, F^{(2)},\ldots,F^{(k)}\}$ becomes a subset of a parametric family (q-variate normal).

It must also be pointed out that in the particular case considered above we assumed that the true class distributionswere known. The case where they are not known is discussed in the next section. The basic idea in this situation is to replace the unknown class cdf's in 2.1.6 by suitable "estimates" of the cdf's, for example empiric cdf's might be used.

2.2 On Estimating Distribution Functions

As already mentioned, to apply the minimum distance method we must estimate the cumulative distribution function of the sample to be classified, and possibly also the class

distribution functions if these are unknown. Some of the distances we are interested in are expressed in terms of probability density functions (pdf's) rather than cdf's. In such cases we will need to estimate pdf's. Consequently, before we proceed to the formulation of the minimum distance rule we discuss briefly the estimation of pdf's and cdf's and make a number of appropriate definitions.

We will adopt the following conventions regarding the notation for pdf's, cdf's and their estimates. We will distinguish between pdf's and cdf's that refer to the same distribution by means of corresponding lower and upper case letters respectively. A symbol above a quantity designates an estimated quantity. Thus $\dot{F}$ and $\dot{f}$ are the "dot" estimates for F and f respectively. Note that if the "dot" estimator is defined in terms of pdf's, then $\dot{F}$ is computed by first obtaining $\dot{f}$ and then finding the corresponding cdf by integration. Similarly if the "dot" estimator is defined in terms of cdf's then $\dot{f}$ is obtained by differentiating $\dot{F}$.

We will assume in general that the estimated pdf's or cdf's are to be based on a random sample of size N (i.e. $\underline{X}_1, \underline{X}_2, \ldots, \underline{X}_N$) from a q-variate population with distribution function $F(\underline{x})$ and corresponding density $f(\underline{x})$ (if it exists). Thus the $\underline{X}_i$'s are q tuples, $\underline{X}_i = (X_{i1}, X_{i2}, \ldots, X_{iq})$ $i = 1, 2, .., N$ and $\underline{x} = (x_1, x_2, \ldots, x_q)$.

Probably the most natural estimators are the so called empiric estimators.

Definition 2.2.1

The <u>empiric cdf</u> $\bar{F}_N(\underline{x})$ is defined as

$$\bar{F}_N(\underline{x}) = \frac{1}{N} \text{ (Number of } \underline{X}_i\text{'s such that } X_{ij} < x_j,\ j = 1,2,\ldots,q) \qquad 2.2.1$$

Assuming cdf's are continuous on the right the corresponding <u>empiric pdf</u> is

$$\bar{f}_N(\underline{x}) = \frac{1}{N} \sum_{i=1}^{N} \delta(\underline{x}-\underline{x}_i) \qquad 2.2.2$$

where $\delta(\cdot)$ is dirac delta function

There are a number of other estimators of interest whose origins are probably heuristic but which can in general be motivated by the following theoretical result due to Fix and Hodges[7].

Theorem 2.2.1 (Fix and Hodges)

If a density $f(\underline{x})$ is continuous at $\underline{x} = \underline{z}$ and $[\gamma_N]$ is a sequence of sets with nonzero volume $[\Phi_N]$ such that

(1) $$\lim_{N \to \infty} \sup_{\underline{y} \in \gamma_N} \underline{z} - \underline{y} = 0 \qquad 2.2.3$$

(2) $$\lim_{N \to \infty} N\Phi_N = \infty \qquad 2.2.4$$

and if $k(N)$ is the number of independent variables $\underline{X}_1, \underline{X}_2, \ldots, \underline{X}_N$ distributed as $f(\underline{x})$ which are contained in Φ_N then if

$$f^*_N(\underline{x}) = \frac{k(N)}{N\Phi_N} \qquad 2.2.5$$

then $\overset{*}{f}_N(\underline{x})$ approaches $f(\underline{x})$ in probability. The $\overset{*}{f}_N(\underline{x})$ will be referred to as a <u>local density estimate</u> of $f(\underline{x})$ at $\underline{x}$.

Conditions (1) and (2) ensure that as Φ_N decreases with N about $\underline{z}$ it does so in such a manner that the expected number of observations in Φ_N approaches infinity, thus ensuring a consistent density estimate.

Choosing Φ_N to consist of disjoint cells of equal size fixed with respect to the coordinate system leads to "histogram estimates".

Definition 2.2.2

The <u>cumulative histogram</u> $F_N(\underline{x})$ is defined as

$$\overset{v}{F}_N(\underline{x}) = \frac{1}{N} \text{ (Number of } \underline{X}_i\text{'s such that}$$

$$X_{ij} < b\,([x_j]_b + 1),\ j = 1,2,\ldots,q) \qquad 2.2.6$$

Where $[x_j]_b$ is the largest integer less than or equal to x_j/b. The pdf corresponding to $\overset{v}{F}_N(\underline{x})$ is $\overset{v}{f}_N(\underline{x})$ is referred to as the <u>pdf of the cumulative histogram</u>. In 2.2.6 b is the bin edge.

Definition 2.2.3

The <u>density histogram</u> $\dot{f}_N(\underline{x})$ is defined as

$$\dot{f}_N(\underline{x}) = \frac{k(N)}{Nb^q} \qquad 2.2.7$$

where b is the bin edge and k(N) is the number of $\underline{X}_i$'s such that

$$b[x_j]_b \leq X_{ij} < b([x_j]_b + 1)\ j = 1,2,\ldots,q \qquad 2.2.8$$

where $[x_j]_b$ is the largest integer less than or equal to x_j/b. Equation 2.2.8 simply states that $k(N)$ is the number of $\underline{X}_i$'s in the same bin as $\underline{x}$. The cdf corresponding to $\dot{f}_N(\underline{x})$ is $\dot{F}_N(\underline{x})$ and is referred to as the cdf of the density histogram.

Note that $\dot{f}_N$ and $\overset{v}{f}_N$ are quite different estimators in that $\overset{v}{f}_N$ is the summation of N delta functions while $\dot{f}_N$ is the summation of N step functions. If the bins in the estimators $\overset{v}{F}_N$ and $\dot{f}$ are permitted to become smaller and smaller within the framework of Fix and Hodges result then at points of continuity $\dot{f}(\underline{x})$ and $\overset{v}{F}(\underline{x})$ are consistent asymptotically unbiased estimates for $f(\underline{x})$ and $F(\underline{x})$ respectively.

The idea of selecting γ_N to consist of an interval about the estimation point $\underline{x}$ (as opposed to fixed bins) was first investigated by Rosenblatt[8]. This concept can be generalized[6] by replacing γ_N by a suitable weighting function, and considering Φ_N as the volume of the weighting function, and k(N) as the weighted count of the vectors in Φ_N. That is we define

$$\Phi_N = \int_{-\infty}^{\infty} K_N(\underline{y},\underline{x})d\underline{y} \qquad 2.2.9$$

$$k(N) = N \int_{-\infty}^{\infty} K_N(\underline{y},\underline{x})\ \overline{f}_N(\underline{y})d\underline{y} \qquad 2.2.10$$

where $\int_{-\infty}^{\infty}$ indicates an integration over the whole space and $\overline{f}_N(\underline{y})$ is the emperic pdf. k(N) reduces to

$$k(N) = \sum_{j=1}^{N} K_N(\underline{X}_j,\ \underline{x}) \qquad 2.2.11$$

which for K_N an even function of its argument leads to

$$k(N) = \sum_{j=1}^{N} K_N(\underline{x}, \underline{X}_j) \qquad 2.2.12$$

This leads to the following definition:

Definition 2.2.4

The Parzen density estimate $\dot{\overline{f}}_N(\underline{x})$ is

$$\dot{\overline{f}}_N(\underline{x}) = \frac{1}{N\Phi_N} \sum_{i=1}^{N} K_N(\underline{x}, \underline{X}_i) \qquad 2.2.13$$

Parzen density estimates were investigated for the univariate case by Whittle[9] and Parzen[10] and for the multivariate case by Cacoullos[11].

Under relatively weak conditions on $K_N(\cdot, \cdot)$ the Parzen density estimate is consistent and asymptotically unbiased at points of continuity of $f(x)$. The conditions K_N are that it be bounded, absolutely integrable, and that it approach zero sufficiently rapidly for large values of the argument[11].

The estimators of definition 2.2.1 to 2.2.4 can be used to obtain estimates for q-variate populations regardless of whether the distribution function F belongs to a parametric family or not. If the family is parametric we may wish to use pdf's and cdf's based on the estimated parameters.

Definition 2.2.5

If $F(\underline{x})$ is characterized by $\underline{\theta}$ (i.e., $F(\underline{x}) = F(\underline{x}|\underline{\theta})$) then the parametrically estimated cdf $F_N(\underline{x}|\theta)$ is defined as

$$\hat{F}_N(\underline{x}|\hat{\underline{\theta}}) = F(\underline{x}|\hat{\underline{\theta}})$$

where $\underline{\theta} = (\theta_1, \theta_2, \ldots, \theta_s)$ and $\hat{\underline{\theta}}$ is some estimate of $\underline{\theta}$ based on the random sample.

The density corresponding to $\hat{F}_N(\underline{x}|\hat{\underline{\theta}})$ is $\hat{f}_N(\underline{x}|\hat{\underline{\theta}})$ and is referred to as the parametrically estimated pdf. Note that $\hat{f}_N(\underline{x}|\hat{\underline{\theta}}) = f(\underline{x}|\hat{\underline{\theta}})$.

Frequently we will not wish to be specific regarding the estimator to be utilized. For this reason we make the following definition.

Definition 2.2.6

A sample-based estimate of a cdf ($\tilde{F}_N$) or pdf ($\tilde{f}_N$) is any estimate of a cdf or pdf based on a random sample.

In situations where there appears to be no danger of confusion we will drop the adjective sample-based. Thus the term estimate used by itself usually refers to a sample-based estimate.

2.3 Decision Theoretic Formulation of Minimum Distance Classification

In this section we present what essentially amounts to a decision theoretic formulation of minimum distance classification. Two main types of problems will be considered, each with three cases. In Type I problems we assume that distribution functions for all classes and subclasses are known apriori while in Type II problems we assume that estimates of these distributions must be obtained from appropriate random samples. The three cases considered in

each problem Type are a consequence of different apriori assumptions regarding the number of subclasses. Case (a) assumes each class can be represented by an infinite number of distribution functions (i.e., subclasses) while Case (b) assumes the number is finite but larger than unity. Case (c) is concerned with the situation where each class can be represented by a single distribution function. In every case we assume that the number of main classes is finite and greater than unity.

We will be interested not only in determining distances between individual distribution functions but between sets of distribution functions as well. Such distances are defined in Definition 2.3.1.

Definition 2.3.1

Let the distance $d(F,G)$ be defined for all F,G, in A, where A is an arbitrary set of cdf's of interest. If A_1 and A_2 are non-empty subsets of A then we define the distance $d(A_1, A_2)$ between the sets A_1 and A_2 as

$$d(A_1, A_2) = \inf_{\substack{F \in A_1 \\ G \in A_2}} d(F,G) \qquad 2.3.1$$

With regard to the last definition we note that it applies to finite and infinite sets of distribution functions. Of course, if the sets are finite then taking the infimum is equivalent to taking the minimum.

Furthermore, if each set consists only of a single distribution function then the distance between the sets is precisely the distance between the distribution functions. It is also important to note that the above definition includes as a special case the distance between a distribution function and a set of distributions functions.

In order to avoid future misunderstanding it is necessary to make some comments about notation. In particular, the usage of d(F,G) requires clarification. Some of the distance measures we will consider are expressed in terms of pdf's rather than cdf's. The convention we adopt is that we will use the notation d(F,G) and refer to this quantity as the distance between cdf's even though the distance is expressed in terms of the densities of F and G. A comment should perhaps also be made about the class of cdf's that are permitted. This in general depends upon the particular distance measure and the particular estimator used. All that is required is that the particular distance used must exist for all cdf's of interest, including estimated cdf's. This means, for example, that if a distance is expressed in terms of pdf's then the densities must exist, whereas if the distance is expressed in terms of cdf's then the densities need not necessarily exist.

We are now in a position to formulate the problem in a decision theoretic framework. In specifying a statistical problem we must specify

(a) Z - the sample space of the observed random variable.

(b) Ω - the set of states of nature; that is, the set of possible cdf's of the random variable. If the functional form of the cdf is known, then we can identify Ω with the parameter space.

(c) $\overset{*}{A}$ - the action space; that is the set of actions or decisions available to the statistician.

(d) L (a,F) - loss function defined on $\overset{*}{A}X\Omega$ which measures the loss incurred if $F\varepsilon\Omega$ is the true state of nature and action $a\varepsilon\overset{*}{A}$ is the action taken.

The general formulation of the minimum distance problem in this framework follows:

(a) $Z = E^q$ (q-dimensional Euclidean space)

(b) $\Omega = [\Omega^{(1)}, \Omega^{(2)}, \ldots, \Omega^{(k)}]$ where $\Omega^{(i)}$ is the set of possible distribution functions for the ith class, i = 1, 2, ..., k.

(c) $\overset{*}{A} = [a_1, a_2, \ldots, a_k]$ where a_i is the decision to decide the random sample to be classified belongs to the ith class, i = 1, 2, ..., k.

(d) $L(a,F) = 0$ if $F\varepsilon\Omega^{(i)}$ and action a_i was taken
$= 1$ otherwise.

A decision rule is a function defined on Z and

taking values in $\overset{*}{A}$. The minimum distance decision rule is defined below.

Definition 2.3 2

Let $\underline{Y}$ be the vector of all sample observations. The minimum distance decision rule $D_{MD}: Z \rightarrow \overset{*}{A}$ is $D_{MD}(\underline{Y}) = a_i$ (i.e., decide the random sample to be classified belongs to class i) in case

$$d(\tilde{F}_N, \Lambda^{(i)}) = \underset{j = 1,..,k}{\text{Min}} \; d(\tilde{F}_N, \Lambda^{(j)})$$

Where $\Lambda^{(i)}$ is the set of cdf's selected to represent the ith class and $\tilde{F}_N$ is a sample-based estimate of the cdf of the random sample to be classified.

Normally in a parametric problem parametrically estimated cdf's would be used. It is, of course, always possible to treat a given parametric problem in a non-parametric way. That is even if the problem is parametric one could use some nonparametric estimator, but the converse is not true. It is important to note that $\underline{Y}$ includes not only the random sample to be classified, but also any other observations used in the classification procedure. For example, if training samples are used for each class, these are included in $\underline{Y}$. The sets $\Lambda^{(i)}$ also require comment. $\Lambda^{(i)}$ may be the set of all possible distributions for class i(i.e., $\Lambda^{(i)} = \Omega^{(i)}$) or it may be a subset of $\Lambda^{(i)}$ or the sample based estimates of a set of cdf's selected to represent class i.

As already indicated we will consider a number of special cases of the above formulation. The special cases we consider have been selected to assist us in describing work that has been done on this problem. These special cases are basically a consequence of making different assumptions regarding Ω, and $\Lambda = [\Lambda^{(1)}, \Lambda^{(2)}, \ldots, \Lambda^{(k)}]$. We initially deal with Type I problems where the sets of distribution functions representing the classes are known sets. Actually, this problem is not of great interest from a practical point of view, but it is interesting from a theoretical point of view because it is relatively simple.

Type I - The $\Omega^{(i)}$'s are known sets of cdf's

Case (a) The sets $\Omega^{(i)}$ are infinite and $\Lambda^{(i)} = \Omega^{(i)}$

Case (b) The sets $\Omega^{(i)}$ are finite and $\Lambda^{(i)} = \Omega^{(i)}$

Case (c) The set $\Omega^{(i)} = F^{(i)}$ (single cdf/class) and $\Lambda^{(i)} = F^{(i)}$

If the sets $\Omega^{(1)}, \Omega^{(2)}, \ldots, \Omega^{(k)}$ are known to consist of q-variate distributions but are otherwise unknown then we would like to replace each actual cdf by a corresponding sample-based cdf, and base the decision rule on these distributions. In practice we can of course handle only a finite number of distributions. Consequently, if the sets $\Omega^{(i)}$ are infinite, we must somehow replace the infinite sets with representative finite sets. We are also forced to adopt a similar attitude if we know <u>apriori</u> that the sets $\Omega^{(i)}$ are finite, but do not know precisely how many

distribution functions each $\Omega^{(i)}$ contains (i.e., how many subclasses of wheat are there?); or even if we know the precise number, we may not know how to obtain a random sample for each distribution function (i.e., how do we select samples representing different subclasses of wheat?). Finally, in the finite case, even if we can obtain a random sample for each distribution function of interest, their number may be so large that for practical reasons we may wish to use a smaller number of representative distributions. Thus, the need arises for a method to select a representative set of distribution functions from a larger (possibly infinite) set. To do this we will assign a distribution $H^{*(i)}$ to $\Omega^{(i)}$, $i = 1, 2, \ldots, k$. That is the events to which probability mass is assigned by $H^{*(i)}$ are sets of distributions in $\Omega^{(i)}$. To select a random set of cdf's from $\Omega^{(i)}$ (i.e., to select a random set of training samples for the ith class) is now equivalent to selecting a random sample from $H^{*(i)}$.

The above formulation is rather complicated in that we are dealing with a distribution over a space of functions. This complexity can be avoided by restricting consideration to a parametric family characterized by s real parameters. Making the logical assumption that a one to one correspondence exists between cdf's in $\Omega^{(i)}$ and points in the parameter space $\theta^{(i)}(\equiv E^s)$, it is apparent that assigning a distribution $H^{*(i)}$ to $\Omega^{(i)}$ is equivalent to

assigning some other distribution $H^{(i)}$ to the parameter space $\theta^{(i)}$. Consequently, in the parametric case rather than deal with $H^{*(i)}$, which is a cdf on a set of distribution function, we can deal with $H^{(i)}$ which is a cdf in E^s.

Actually as far as the minimum distance classification scheme itself is concerned we do not have any direct interest in $H^{*(i)}$ and $H^{(i)}$. These distributions are introduced to enable us to establish a connection between minimum distance and nearest neighbor decision rules.

It is perhaps worthwhile to restate the above ideas with reference to a specific application, involving multispectral data-imagery from an agricultural scene, before stating them in a more formal manner. In the interest of simplicity and since it is the case of primary interest we will assume Ω is a parametric family characterized in E^s. That is, we assume that the true q-dimensional distribution of the radiance measurements from each field belong to the same parametric family which can be characterized in the parameter space E^s. This family may have a finite or infinite number of members (i.e., subclasses). We assume that all the fields in a class (i.e., wheat) can be described by a suitable distribution $H^{(i)}$ over the parameter space. We select at random a set of training fields for each class. Because of our formulation this is equivalent to selecting a random sample from the parameter space according to the assumed distribution over the parameter

space for that class (i.e., $H^{(i)}$). For each of the randomly selected training fields we use the radiance measurements to get an estimated cdf for that field. In this way we obtain estimated cdf's for a representative set of training fields for each class. An unknown field is then assigned to the class that has a training field whose estimated cdf is nearest to the estimated cdf of the unknown field. Since the problem as stated is parametric, one would normally, though not necessarily, use parametrically estimated cdf's.

We now formally state the Type II problem in which the $\Omega^{(i)}$'s are unknown. While we are primarily interested in the case where Ω is a parametric family we will not restrict ourselves to this case in stating the problem. Also in Type II problems the description of the set $\Lambda^{(i)}$ is rather involved.

Type II - The $\Omega^{(i)}$'s are Unknown Sets of cdf's

Case (a) - The sets $\Omega^{(i)}$ are infinite in number and $\Lambda^{(i)} = \tilde{\Omega}_{M_i}^{(i)}$. We now describe the set $\Omega_{M_i}^{(i)}$. First we select a set of population cdf's corresponding to a representative set of M_i training fields for class i, i = 1,2,..., k. Let $\Omega_{M_i}^{(i)}$ be this set for the ith class. That is $\Omega_{M_i}^{(i)}$ is a random sample of size M_i for $H^{*(i)}$. A sample-based cdf is then obtained for each cdf in $\Omega_{M_i}^{(i)}$ for i = 1,2,..., k. The resultant set of sample-based estimated cdf's is $\tilde{\Omega}_{M_i}^{(i)}$. For the case where

parametrically estimated cdf's are used (i.e., $\sim$ *replaced by* $\hat{}$) $\Omega_{M_i}^{(i)}$ *can also be considered to be* a random sample of size M_i in the parameter space according to a distribution $H^{(i)}$.

Case (b) - The sets $\Omega^{(i)}$ are finite and $\Lambda^{(i)} = \tilde{\Omega}^{(i)}$ or $\Lambda^{(i)} = \tilde{\Omega}_{M_i}^{(i)} \subset \tilde{\Omega}^{(i)}$. Normally if the $\Omega^{(i)}$ are finite (i.e., *finite number of subclasses) we would let* $\Lambda^{(i)} = \tilde{\Omega}^{(i)}$ where $\tilde{\Omega}^{(i)}$ is the set of sample-based estimated cdf's in the ith class. In cases where the number of subclasses is impractically large or only a random sample of training fields is *available, we let* $\Lambda^{(i)} = \tilde{\Omega}_{M_i}^{(i)} \subset \tilde{\Omega}^{(i)}$ *and proceed as* in case (a).

Case (c) - The set $\Omega^{(i)} = F^{(i)}$ (Single cdf per class) and $\Lambda^{(i)} = \tilde{F}_N^{(i)}$.

2.4 Distance Measures

The importance in statistics of distances between cdf's has, of course, long been recognized.[12] According to Samuel and Bachi[13] their use appears to fall into two broad categories.

(a) Used for descriptive purposes. For example, as an indicator to quantitatively specify how near a given distribution is to normal distribution.

(b) Use in hypothesis testing, which is, of course, a special case of decision theory.

There is a tendency for distance functions sufficiently sensitive to detect minor differences in distribution functions (i.e., type (a) use) to be somewhat involved functions of the observations, with the result that their use as test statistics in hypothesis testing has been limited because of the complicated distribution theory. On the other hand, distance functions whose theory is simple enough to be readily used as test statistics often do not distinguish distribution functions sufficiently well. Since we are interested in good discrimination between distribution functions, we must somehow circumvent this problem. We do so by relaxing somewhat our requirements from those usually demanded of test statistics in hypothesis testing. Usually in hypothesis testing it is required that at least the asymptotic distribution of the test statistic under the null hypothesis be known. This is required to enable the experimenter to determine the range of values of the test statistic (critical region) for which the null hypothesis is to be rejected for a specified probability of false rejection of the null hypothesis ($\equiv$ probability of Type I error which is also called the size of the test). Our requirements are somewhat more modest. In particular we attempt only to establish reasonably tight upper bounds on the total probability of error rather than specifying specifically the probability of Type I error. Actually, this approach is more meaningful for the classification problem

than is the classical hypothesis testing approach. In the hypothesis testing approach the size of the test is chosen by the experimenter. Such a procedure controls the probability of false rejection (Type I error) at the desired level, but leaves the power of the test or the probability of false acceptance (Type II error), and consequently the total probability of error to the mercy of the experiment[14]. Such an approach is reasonable if the emphasis is on the null hypothesis as the case in hypothesis testing. In the classification problem interest is more naturally centered on the total probability of error.

It also appears worthwhile mentioning that although distance measures are widely used as test statistics it appears that the distance properties of such test statistics are used rather infrequently, at least directly. This is probably a consequence of the hypothesis testing approach where the emphasis is on the appropriate distribution theory.

We will now turn our attention to specific distance measures. The literature abounds with references to distance measures and no attempt will be made to give a complete bibliography. A representative sample of distance measures is given in Table 2.4.1 along with references.[15-32] We have attempted to include the most widely used distance measures because of their obvious importance, as well as more obscure distance measures whose application to the

Table 2.4.1

Univariate Forms of Distance Measures

Name	Mathematical Form	Reference
Cramer-Von Mises Distance	$W = \{\int_{-\infty}^{\infty} (G(x) - F(x))^2 dx\}^{\frac{1}{2}}$	15,16,17,18
Kolmogorov-Smirnov Distance	$K = \mathrm{Sup}_x \lvert G(x) - F(x) \rvert$	19,20,17,18
Divergence	$J = \int_{-\infty}^{\infty} \mathrm{Ln}(\frac{f(x)}{g(x)})(f(x)-g(x))dx$	21,22,23
Bhattacharyya Distance	$B = -\mathrm{Ln}\int_{-\infty}^{\infty} (f(x)g(x))^{\frac{1}{2}} dx$	23,24
Jeffreys-Matusita Distance	$M = \{\int_{-\infty}^{\infty} (\sqrt{g(x)} - \sqrt{f(x)})^2 dx\}^{\frac{1}{2}}$	21,22,25
Kolmogorov Variational Distance	$K(p) = \int_{-\infty}^{\infty} \lvert p_g g(x) - p_f f(x) \rvert dx$	23,26,27
Kullback-Leibler Numbers	$L_{fg} = \int_{-\infty}^{\infty} \mathrm{Ln}(\frac{f(x)}{g(x)}) f(x) dx$	28,23
Swain-Fu Distance	$T = \frac{[(\mu_f-\mu_g)^2]^{1/2}}{\sqrt{3}\ (\sigma_f+\sigma_g)}$	29
Mahalanobis Distance	$\Delta = \{\frac{(\mu_g-\mu_f)^2}{\sigma^2}\}^{1/2}$	30,31

Table 2.4.1 (Cont'd.)

Name	Mathematical Form	Reference
Samuels-Bachi Distance	$U = \{\int_0^1 [G^{-1}(x) - F^{-1}(x)]^2 dx\}^{\frac{1}{2}}$ Where $F^{-1}(\alpha) = \text{Inf}_x(x:F(x)\geq\alpha)$	13
Kiefer-Wolfowitz Distance	$V = \int_{-\infty}^{\infty} \lvert g(x) - f(x) \rvert e^{-\lvert x \rvert} dx$	32

Notation
(1) F,G are univariate cdf's with pdf's f,g; means μ_f, μ_g; Variances σ_f^2, σ_g^2; and prior probabilities p_f, p_g.

(2) For Mahalanobis distance F and G are normal with means μ_f and μ_g and have common variance σ^2.

(3) | | designates the absolute value or vector norm.

present problem appears reasonable. In addition a few miscellaneous distance measures have been included to give an indication of the variety of distances that have been suggested. Rather than attempt to provide a comprehensive list of references the attempt has been made to reference, in addition to the original source, only those papers containing a number of additional references such as survey papers. The papers by Darling[17], Sahler,[18] and to a certain extent Kailath[23] fall in this latter category.

Table 2.4.1 gives the one dimensional version of the various distance measures because the vast majority of the references cited deal only with this case. The extension to multivariate distributions is in most cases quite natural, except perhaps for the Samuels-Bachi distance. In order to avoid any misunderstanding the multivariate forms of the distances measures in Table 2.4.1 are given in Table 2.4.2 including a possible extension to the multivariate case for the Samuel-Bachi distance.

One of the properties of distance measures with which we shall be concerned is whether or not the distance is a true metric. This property, of course, depends on the set of distribution functions of interest. In Table 2.4.2 the metric properties of the distance measures are shown for three different families of distributions functions. These three families are: C the family of q-variate absolutely continuous distribution functions, MVN the family of q-variate

Table 2.4.2

Multivariate Forms of Distance Measures and Their Metric Properties

Name	Form	Metric in C	Metric in MVN	Metric in MVN_Σ
Cramer-Van Mises	$W = \{\int_{-\infty}^{\infty} (G(\underline{x}) - F(\underline{x}))^2 d\underline{x}\}^{\frac{1}{2}}$	Yes	Yes	Yes
Kolmogorov-Smirnov	$K = \mathrm{Sup}_{\underline{x}} \lvert G(\underline{x}) - F(\underline{x}) \rvert$	Yes	Yes	Yes
Divergence	$J = \int_{-\infty}^{\infty} \mathrm{Ln}(\frac{f(\underline{x})}{g(\underline{x})})(f(\underline{x})-g(\underline{x}))d\underline{x}$	No	No	Yes
Bhattacharyya Distance	$B = -\mathrm{Ln}\int_{-\infty}^{\infty} (f(\underline{x})g(\underline{x}))^{\frac{1}{2}} d\underline{x}$	No	No	Yes
Jeffreys-Matusita Distance	$M = \{\int_{-\infty}^{\infty} (\sqrt{g(\underline{x})} - \sqrt{f(\underline{x})})^2 d\underline{x}\}^{\frac{1}{2}}$	Yes	Yes	Yes
Kolmogorov Variational Distance	$K(p) = \int_{-\infty}^{\infty} \lvert p_g g(\underline{x}) - p_f f(\underline{x}) \rvert d\underline{x}$	Yes	Yes	Yes
Kullback-Liebler Numbers	$L_{fg} = \int_{-\infty}^{\infty} \mathrm{Ln}(\frac{f(\underline{x})}{g(\underline{x})})\ f(\underline{x})d\underline{x}$	No	No	Yes
Swain-Fu Distance	$T = \frac{\lvert \underline{\mu}_f - \underline{\mu}_g \rvert}{D_f + D_g}$	No	No	Yes

Where $D_\cdot = \{\frac{\lvert \underline{\mu}_f - \underline{\mu}_g \rvert^2 (q+2)}{\mathrm{tr}\{\Sigma_\cdot (\underline{\mu}_f - \underline{\mu}_g)(\mu_f - \mu_g)^t\}}\}^{\frac{1}{2}}$

Table 2.4.2 (cont'd.)

Name	Form	Metric in C	Metric in MVN	Metric in MVN_Σ
Mahalanobis Distance	$\Delta = \{(\underline{\mu}_g-\underline{\mu}_f)^t \Sigma^{-1}(\underline{\mu}_g-\underline{\mu}_f)\}^{\frac{1}{2}}$			Yes
Samuels-Bachi Distance	$U = \{\int_0^1 [F^{-1}(\alpha)-G^{-1}(\alpha)]d\alpha\}^{\frac{1}{2}}$ where $F^{-1}(\alpha) = \text{Inf}\{c \mid Q_c \cap Q_\alpha \neq \emptyset\}$ and $Q_c=\{\underline{x} \mid \sum_{i=1}^{q} x_i \leq C\}, Q_\alpha=\{\underline{x} \mid F(\underline{x}) \geq \alpha\}$	No	No	No
Kiefer-Wolfowitz Distance	$V = \int_{-\infty}^{\infty} \lvert F(\underline{x})-G(\underline{x})\rvert e^{-\lvert x\rvert} d\underline{x}$	Yes	Yes	Yes

Notation

(1) F, G are multivariate cdf's with densities f, g; means $\underline{\mu}_f$, $\underline{\mu}_g$; covariances Σ_f, Σ_g; and prior probabilities p_f, p_g.

(2) $\int_{-\infty}^{\infty} (\ )\, d\underline{x}$ designates a multivariate integral.

(3) For Mahalanobis distance F and G are normal with means $\underline{\mu}_f$ and $\underline{\mu}_g$ and have common covariance Σ.

(4) $\lvert\ \rvert$ designates the absolute value or vector norm.

(5) t designates the transpose

normal distribution functions, and MVN_{Σ} the family of q-variate normal distribution functions with equal covariance matrices. Since MVN and MVN_{Σ} are subsets of C it is, of course, true that a metric in C is also a metric in MVN and MVN_{Σ}. A metric in MVN_{Σ} need not, however, be a metric in MVN or C.

Because of the importance of the multivariate normal distribution, expressions for the distance between two such distributions are given in Table 2.4.3 for each of the distances measured in Table 2.4.1 for the cases where the expressions are known.

Probably the best known distance measures in statistics are the Cramer-Von Mises distance (CV distance)[15, 16, 17, 18] and Kolmogorov-Smirnov distance (KS distance).[19, 20, 17, 18] Test statistics based directly on these distance measures, as well as closely related distance measures are in common usage in statistics. The most important characteristic of the test statistics derived from these distance measures is that in the one dimensional case they are distribution-free under the null hypothesis. By distribution-free we mean that the distribution of the test statistic is independent of the underlying distribution. It is this distribution-free property which has lead to widespread use of CV and KS type of test statistics. Sahler[18] provides a comprehensive tabulation of the distribution theory of these and other distribution-free statistics while Darling[17]

Table 2.4.3

Distances Between Two Multivariate Normal cdf's

Name	Distance
Divergence	$J = \frac{1}{2}\,tr[\Sigma_f-\Sigma_g][\Sigma_g^{-1}-\Sigma_f^{-1}]+\frac{1}{2}\,tr[\Sigma_f^{-1}+\Sigma_g^{-1}][\underline{\mu}_f-\underline{\mu}_g][\underline{\mu}_f-\underline{\mu}_g]^t$
Bhattacharyya Distance	$B = \frac{1}{8}(\underline{\mu}_f-\underline{\mu}_g)^t[\frac{\Sigma_f+\Sigma_g}{2}]^{-1}(\underline{\mu}_f-\underline{\mu}_g)+\frac{1}{2}\,Ln\,\frac{det(\frac{1}{2}[\Sigma_f+\Sigma_g])}{\{det(\Sigma_f)det(\Sigma_g)\}^{1/2}}$
Jeffreys-Matusita Distance	$M = [2\{1 - \frac{\{det(\Sigma_f)det(\Sigma_g)\}^{1/4}}{\{det(\frac{1}{2}[\Sigma_f+\Sigma_g])\}^{1/2}}\,exp(-\frac{1}{8}(\underline{\mu}_f-\underline{\mu}_g)t[\frac{\Sigma_f-\Sigma_g}{2}]^{-1}(\underline{\mu}_f-\underline{\mu}_g)\}]^{1/2}$
Kullback-Leibler Numbers	$L_{fg} = \frac{1}{2}\,Ln\,\frac{det\,(\Sigma_f)}{det\,(\Sigma_g)} + \frac{1}{2}\,tr\,\Sigma_f[\Sigma_g^{-1}-\Sigma_f^{-1}]+\frac{1}{2}\,tr\,\Sigma_g^{-1}[\underline{\mu}_f-\underline{\mu}_g][\underline{\mu}_f-\underline{\mu}_g]^t$
Swain-Fu Distance	$T = \frac{\lvert\underline{\mu}_f-\underline{\mu}_g\rvert}{D_f + D_g}$ where $D_. = \{\frac{\lvert\underline{\mu}_f-\underline{\mu}_g\rvert^2(q+2)}{tr\{(\Sigma_.)^{-1}(\underline{\mu}_f-\underline{\mu}_g)(\underline{\mu}_f-\underline{\mu}_g)^t\}}\}^{\frac{1}{2}}$

Table 2.4.3 (cont'd.)

Name	Distance
Mahalanobis Distance	$\Delta = \{(\underline{\mu}_g - \underline{\mu}_g)^t \Sigma^{-1} (\underline{\mu}_g - \underline{\mu}_f)\}^{\frac{1}{2}}, (\Sigma = \Sigma_f = \Sigma_g)$

Notation

(1) t means transpose

(2) det means determinant

(3) tr means trace

(4) The normal distributions involved have means $\underline{\mu}_f$ and $\underline{\mu}_g$ and covariance matrices Σ_f and Σ_g

traces their history and development.

The Divergence (D),[21,22,23] Bhattacharyya distance (B distance),[23,24] Jefferys-Matusita distance (JM distance),[21,22,25] Kolmogorov variational distance (KV distance)[23,26,27] and Kullback-Leibler number (KL numbers)[28, 23] are the next group of distance measures we will discuss. They do not lead to distribution-free statistics even in the one dimensional case and consequently their use has been more restricted than CV and KS type statistics. Some of them, particularly the Divergence and Bhattacharyya distance, have nevertheless gained a certain degree of acceptance.

There are several similarities between these five distance measures. One similarity that is immediately apparent is the fact that each of these distances is defined in terms of pdf's rather than cdf's. This means of course that their use is restricted to a somewhat smaller class of distributions than the CV and KS distances. As already mentioned we shall continue to write d(F,G) to indicate an arbitrary distance between cdf's F and G, with pdf's f and g, even if the distance is expressed in terms of pdf's. A second similarity, which is somewhat more obscure but much more important than the first similarity noted, is that these five distance measures can be written in terms of the likelihood ratio $L(\underline{x})$ where

$$L(\underline{x}) = \frac{f(\underline{x})}{g(\underline{x})} \qquad 2.4.1$$

In the parametric case where $\underline{\theta}^{(f)}$ characterizes f and $\underline{\theta}^{(g)}$ characterizes g we will write

$$L(\underline{x}|\underline{\theta}) = \frac{f(\underline{x}|\underline{\theta}^{(f)})}{g(\underline{x}|\underline{\theta}^{(g)})} . \qquad 2.4.2$$

Not only can these 5 distance measures be written in terms of the likelihood ratio, they can in fact all be written in the following form

$$d'(F,G) = I(E_g[C(L(\underline{x}))]) \qquad 2.4.3$$

where the ' denotes a distance measure of this form.
C is a continuous convex function
E_g is the expectation with respect to $g(\underline{x})$, and
I is any strictly increasing real function of a real variable.
The importance of this property lies in the fact that it enables us to prove the following theorem.

Theorem 2.4.1

Let two q-variate parametric pdf's f and g be characterized by parameters $\underline{\theta}^{(f)}$ and $\underline{\theta}^{(g)}$ and prior probabilities p_f and p_g respectively. Let $\underline{\beta}^{(f)}$ and $\underline{\beta}^{(g)}$ be an alternate set of parameters for f and g. The theorem then states that if

$$d'_{\theta}(F,G) \geq d'_{\beta}(F,G)$$

then there exists a set of prior probabilities $[p_f, p_g]$ such that

$$P_e(\theta,p) < P_e(\beta,p)$$

where $d'_\theta(F,G)$ is a distance measure of form 2.4.3 using the parameter set $[\underline{\theta}^{(f)},\underline{\theta}^{(g)}]$, and $P_e(\theta,p)$ is the probability of error using parameter set $[\underline{\theta}^{(f)},\underline{\theta}^{(g)}]$ and prior probabilities $[p_f,p_g]$. $d'_\beta(F,G)$ and $P_e(\beta,p)$ are similarily defined.

Esentially Theorem 2.4.1 says that if the distance between F and G is greater when using the θ parameter set then when using the β parameter set, then using probability of error as a criterion, there exists a set of prior probabilities for which the θ set is better than the β set. Although the existance of such a set of priors is known, it has not been established how to determine what this set is. Nevertheless, it is primarily this property that has encouraged the use of these distances measures in feature selection.[33]

Karlin and Bradt[34] have proven Theorem 2.4.1 for Divergence, while Kailath[23] has proven it for Bhattacharyya distance. It has not previously been proven in the general form stated; for this reason its proof is given in Section 3.1. The proof essentially parallels Kailath's proof for the Bhattacharyya distance.

Since a number of commonly used distance measures have the form of 2.4.3 it is natural to ask whether or not

2.4.3 could be used to generate other distance measures. Ali and Silvey[35] have in fact shown this to be the case. Starting with four properties that one might reasonably demand of a distance measure, they show that distance measures of the form d'(F,G) possess these properties. In fact, their result is even somewhat more general than suggested by the last statement. They permit $L(\underline{x})$ to be infinite on a set of zero measure. This necessitates that the expectation E is 2.4.3 be replaced by a generalized expectation E*. This generalized expectation reduces to E if L(x) is finite.

Kullback-Leibler numbers[28,23] have been included in the tables of distance measures primarily because they turn out to be important from a theoretical point of view. In general, Kullback-Leibler numbers are not symmetric with respect to the densities involved. Consequently, it is necessary to distinguish between the Kullback-Leibler number of density f with respect to g (L_{fg}), and that of g with respect to f (L_{gf}). A consequence of this lack of symmetry is that Kullback-Leibler numbers are not a metric in either C or MVN. The asymmetry disappears in the sapce of MVN_{Σ} distributions and consequently for this case we drop the subscripts on L. Also in the space of MVN_{Σ} distributions L is a metric. The divergence is a symmetrized form of the KL numbers namely

$$J = L_{fg} + L_{gf} \qquad 2.4.4$$

There are a number of important equalities and inequalities relating the five distance measures under discussion (i.e., Divergence, B distance, JM distance, KV distance and KL numbers) to each other; and to the probability of error in a two class classification problem. It is convenient to define the affinity (or Bhattacharyya coefficient) between two distributions F and G as

$$\rho(F,G) = \int_{-\infty}^{\infty} (f(\underline{x})g(\underline{x}))^{1/2} d\underline{x} \qquad 2.4.5$$

then the Bhattacharyya distance is

$$B = -\ln\rho \qquad 2.4.6$$

The Jeffreys-Matusita distance M and Bhattacharyya distance B are closely related. In fact, from the definition of M and B (Table 2.4.1) it follows directly that

$$M = [2(1-\rho)]^{1/2} = [2(1-e^{-B})]^{1/2} \qquad 2.4.7$$

The reason for considering both of these measures is because M is a metric in the space of all absolutely continuous cdf's but ρ and consequently B are not. Relationships in the form of inequalities also exist between the Divergence J, Kolmogorov variational distance K(p) and the affinity.[23] These are

$$\rho \geq e^{-J/4} \qquad 2.4.8$$

$$[1 - 4p_f p_g \rho^2]^{1/2} \geq 2K(p) \geq [1-2(p_f p_g \rho)^{1/2}] \qquad 2.4.9$$

For the two class problem the probability of error P_e can be bounded above and below in terms of the affinity by

$$1/4\ \rho^2 \leq 1/2(1 - (1-\rho^2)^{1/2}) \leq P_e \leq 1/2\rho \qquad 2.4.10$$

A crude lower bound on the probability of error has also been obtained in terms of Divergence but an upper bound is unknown. Specifically

$$P_e \geq 1/8\ e^{-J/4} \qquad 2.4.11$$

The probability of error is intimately related to K(p) in that

$$P_e = p_f - K(p) \qquad 2.4.12$$

Kailath[23] gives a more complete discussion of these and other inequalities as well as a number of additional references.

The Swain-Fu distance[29] differs from all the other distances in Table 2.4.1 in that is defined in terms of the first and second moments of the distributions, rather than the pdf's or cdf's themselves. Consequently, one would expect it to be a reasonable distance measure only if its use is restricted to distributions that can reasonably be characterized by their first and second moments. The Swain-Fu distance can be interpreted geometrically in the following way. Let the means of distributions F and G be $\underline{\mu}^{(f)}$ and $\underline{\mu}^{(g)}$ respectively. Let D_f be the distance along $(\underline{\mu}^{(g)}-\underline{\mu}^{(f)})$ from

$\mu^{(f)}$ to the surface of the ellipsoid of concentration for the distribution F; and let D_g be defined in an anologous manner for the distribution G. Then the Swain-Fu distance is

$$T = \frac{|\underline{\mu}^{(g)} - \underline{\mu}^{(f)}|}{D_f + D_g} \qquad 2.4.13$$

The ellipsoid of concentration for a distribution F is the ellipsoid over which a uniform distribution has the same first and second moments as the distribution F. Actually the expression given for the Swain-Fu distance for the multivariate and normal cases in Tables 2.4.2 and 2.4.3 differs from the original expression of Swain and Fu[29]. The given expression is much more compact than the original and computationally simpler. In Appendix A we show that the two forms are equivalent.

If $\underline{\mu}^{(f)} = \underline{\mu}^{(g)}$ then T is zero (see Appendix A). Consequently T is not a metric in C or MVN. It is a metric in MVN_Σ.

The next distance in Table 2.4.1 is the Mahalanobis distance Δ[30,31] which has long been used in statistics. The use of this distance measure is restricted to normal distributions with equal covariance matrices (i.e., MVN_Σ). It is worthwhile noting that in MVN_Σ the Bhattacharyya distance, Kullback-Leibler numbers and Divergence are proportional to Δ^2, in fact from Table III we have

$$B = \frac{J}{8} = \frac{L}{4} = \frac{1}{8}\Delta^2 \text{ for distribution in } MVN_\Sigma \qquad 2.4.14$$

The last two distance measures in Table I have been included primarily to demonstrate the variety of distance measures available. We will not make any further comments about the Samuel-Bachi[13] distance but a few remarks about the Kiefer-Wolfowiz[32] distance are in order. Actually this distance is a special case of a more general distance used by Kiefer and Wolfowitz. They were prompted to use this distance as it possessed some theoretical properties they desired. It is readily apparent that the Kiefer-Wolfowitz distance is essentially an exponentially weighted version of the Kolmogorov variational distance with equal priors. The technique of using a weighting function to emphasize certain region of the distribution function, and consequently generate new distance measures has been used in conjunction with other distance measures as well, notably the CV and KS distances.

Recognizing the large variety of distance measures available, the problem naturally arises as to which distance measure to use in a given problem. Unfortunately, no answer is available to this question at present, but some general comments regarding the selection of a distance measure can be made. The distribution-free properties that make the CV and KS distance so popular in the univariate case no longer enjoy this advantage in the multivariate case. Since it is the multivariate case that is of interest these distances lose their special appeal. Intuitively a distance

like the KS distance does not appear to be as good a distance measure as those involving integration over the whole space. It is also more difficult to compute in parametric situations than some of the integral relations. The Samuels-Bachi distance suffers a similar computational disadvantage. From the theoretical point of view distances based on the likelihood ratio appear to have some desirable properties (for example Theorem 2.4.1). As has already been noted these distances are based on pdf's rather than cdf's. The tendency, therefore, exists for these distances to more reliably indicate changes in pdf's rather than cdf's, and it is probably true that we are more interested in detecting changes in pdf's rather than cdf's, although this is certainly a rather subjective question.

Of the distances based on likelihood ratios the Bhattcharyya distance seems to have been gaining in favor. The prime reason for this seems to be the apparent close relation between probability of error and Bhattacharyya distance, as well as the relative ease of computing Bhattacharyya distance in theoretical problems. Other properties of the Bhattacharyya distance which enhance its prestige as a distance measure have been pointed out by Lainiotis[36] and Stein[37]. Another property of considerable theoretical utility is the close relation between the Bhattacharyya distance (or affinity) and the Jeffreys-Matusita distance (Equation 2.4.7). In the minimum distance decision framework

decisions made on the basis of the Bhattacharyya distance, Jeffreys-Matusita distance or affinity all yield identical results, and consequently have identical probability of error. The Jeffreys-Matusita distance is, however, a metric in a much larger class of distribution (see Table 2.4.2). This means that theoretical derivations regarding probability of error can be made using the metric properties of the Jeffreys-Matusita distance in this larger class, and the results are applicable if classification is effected using Bhattacharyya distance or affinity as well. This property has been used extensviely by Matusita.

Based on the general information presented above, and lacking experimental evidence to the contrary, the Bhattacharyya distance appears to be a reasonable choice for many problems. An important aspect of the experimental work to be described is to obtain the experimental evidence as to the comparative performance of a number of distance measures in minimum distance classification of multispectral data-imagery.

2.5 On Minimum Distance Classification

In this section we discuss work that has previously been done on the problem formulated in Section 2.3. Most of the work on minimum distance methods has been done by Matusita[38-45] and Wolfowitz.[46,47,48,49] Wolfowitz's work is primarily concerned with estimation, while much of Matusita's work deals with the decision problem. Contributions have also been made by Gupta,[50] Cacoullous,[51,52] and Srivastava.[53]

In dealing with minimum distance decision rules a common requirement is to insist that by using arbitrarily large samples, the probability of error can be made arbitrarily small. This concept is similar to the concept of consistency in estimation and prompts the following definition.

Definition 2.5.1

The minimum distance decision rule $D_{MD}(\underline{Y})$ is <u>consistent</u> in $\Omega = [\Omega^{(1)}, \Omega^{(2)}, \ldots, \Omega^{(k)}]$ with respect to the distance $d(\cdot,\cdot)$ and the estimator $\sim$ if for any $F \epsilon \Omega$ and any $i = 1,..,k$

$$\lim_{\text{All Sample Sizes} \to \infty} P(D_{MD}(\underline{Y}) = a_i \mid F \epsilon \Omega^{(i)}) = 1 \qquad 2.5.1$$

Where Ω is some family of q-variate cdf's and $\underline{Y}$ contains all samples used to obtain the sample-based cdf's used in the decision rule, including the sample to be classified. Note that $P(\cdot)$ is simply the probability of correctly classifying a random sample from the ith class.

If the above property holds uniformly for all $F \epsilon \Omega$, then the decision rule is <u>uniformly consistent</u> in Ω with respect to the distance $d(\cdot,\cdot)$ and the estimator $\sim$.

Note that consistency is defined with respect to both a distance, and an estimator for a given set of distributions. This is necessary because a change in either the

distance measure or estimator could conceivably make it impossible to make the probability of error arbitrarily small by increasing sample sizes for some distributions in the set.

We will also wish to use the concept of consistency of a distance function.

Definition 2.5.2

A distance function d between cdf's is said to be consistent in Ω with respect to the estimator $\sim$, if for an arbitrary cdf $F \epsilon \Omega$ and every $\epsilon > 0$

$$\lim_{N \to \infty} P(d(\tilde{F}_N, F) > \epsilon | F) = 0 \qquad 2.5.2$$

Where Ω is some family of q-variate cdf's, and $\tilde{F}_N$ is a sample based estimate of F based on a random sample of size N from F.

If the above condition holds uniformly for all $F \epsilon \Omega$, then the distance is uniformly consistent in Ω with respect to the estimator $\sim$.

For the nonparametric case where the distribution functions are unknown and each class can be represented by a single distinct function (i.e., problem Type II case (c)) Gupta[50] has shown that the minimum distance rule is consistent (uniformly consistent) in Ω, with respect to distance d and emperic cdf's, provided d is a metric that is consistent (uniformly consistent) in Ω, with respect to emperic cdf's. Approximately the same conclusion was apparently

reached independently by Matusita[44] who showed that for Type II case (c) problems, the minimum distance rule is consistent in Ω, with respect to a distance d and emperic cdf's, provided d is a metric that is either consistent or uniformly consistent in Ω with respect to emperic cdf's. Both Gupta and Matusita assume that d is a metric in $\Omega \cup \overline{\Omega}$ where $\overline{\Omega}$ is the set of emperic cdf's corresponding to Ω. Matusita has also shown that his result holds if the class distributions are known. (i.e., problem Type I case (c)). Under these circumstances he points out that the space in which d must be a metric can be somewhat smaller because distances between emperic cdf's are not involved in the decision procedure.

Matusita also points out that for the nonparametric case with finitely many subclasses (i.e., problems Type I, II, case (b))* no additional problems arise and that the results of the previous paragraph are still valid provided the subclasses are distinct (i.e., $d(\Omega^{(i)}, \Omega^{(j)})>0$, $i \neq j$; $i,j = 1,2,\ldots,k$) and d is a metric in $\Omega \cup \overline{\Omega}$. The reason this is true is because under the stated condition each subclass can be viewed as a separate class in the proof.

For the case of known but infinitely many subclasses Matusita shows that the minimum distance rule is consistent in Ω, with respect to a distance d and emperic

*Excluding the case where random samples from Ω are used.

cdf's, provided d is a metric that is uniformly consistent in Ω with respect to emperic cdf's. Actually this result had essentially been obtained earlier by Hoeffding and Wolfowitz[54] who were concerned with distinguishability of sets of distributions. Hoeffding and Wolfowitz assume two sets of distribution A_1 and A_2 are distinguishable in a class τ of tests if there exists a test in τ for which the probability of incorrectly classifying a random sample from a distribution in $A_1 \cup A_2$ can be made arbitrarily small. One class of tests they consider is the class of tests for which the maximum sample size is less than infinity. They call this set of tests τ_3 and define distributions which are distinguishable in τ_3 to be finitely distinguishable. It is apparent that τ_3 included the minimum distance rule. Hoeffding and Wolfowitz show that the sets A_1 and A_2 are finitely distinguishable (i.e., sufficient condition) if

$$d(A_1, A_2) > 0 \qquad 2.5.3$$

where d is uniformly consistent in $A_1 \cup A_2$ with respect to emperic cdf's. They prove this result by showing that the minimum distance rule, which is in τ_3, possess this property. Interestingly enough, the sufficient condition for finite distinguishability is also a necessary condition, subject to relatively weak restrictions on the set of distributions involved.

It is important to note that Hoeffding and Wolfowitz

assume that d has all the properties of a metric except that $d(F,G) = 0$ does not imply $F = G$ (i.e., metric property (b) (2) need not hold). It appears that Matusita and Gupta nowhere use this property of a metric in their proofs.

In some cases of infinitely many subclasses per class, the approach of Matusita, Susika, and Hudimoto[40] can be used to reduce the complexity of the problem. They assume that there exists boundary distributions $F_o^{(i)}$, $F_o^{(j)}$, for any two $\Omega^{(i)}$, $\Omega^{(j)}$ such that

$$d(F_o^{(i)}, \Omega^{(i)}) = 0,\ d(F_o^{(j)}, \Omega^{(j)}) = 0,\ d(F_o^{(i)}, F_o^{(j)}) > 0 \qquad 2.5.4$$

$$d(F_o^{(i)}, \Omega^{(j)}) \leq d(\Omega^{(i)}, \Omega^{(j)}),\ d(F_o^{(j)}, \Omega^{(i)}) \leq d(\Omega^{(i)}, \Omega^{(j)})$$

If these conditions are satisfied then the set of distributions for each class can be relaced by its boundary distribution; that is, the problem reduces to the situation where each class is represented by a single cdf.

For the parametric case the only paper known is apparently that of Matusita.[45] This paper deals with the two class problem where each class is represented by a single multivariate normal cdf. Various <u>apriori</u> assumptions regarding means and covariances are considered including the general case of unequal and unknown means and covariances. Matusita showed that for the case in question, the minimum distance rule is consistent if the Jeffreys-Matusita distance (or related affinity) and parametrically estimated cdf's are used.

Knowing that the minimum distance rule is consistent is certainly useful. From a practical point of view, it is of equal, or possibly even of greater importance, to know how great the probability of error is for a given sample size in a given situation. It is possible to show that a lower bound on the probability of correct classification depends only on probabilities of the following form

$$f_d(N,\varepsilon,F) = P(d(\tilde{F}_N,F)<\varepsilon|F) \qquad 2.5.5$$

In fact to verify (uniform) consistency in Ω with respect to d and $\sim$ it is only necessary to show that for arbitrary $F\varepsilon\Omega$ $P(\cdot)$ can (uniformly) be made arbitrarily small. If the probabilities can be evaluated or bounded from above in terms of N, then a lower bound can be obtained for the probability of correct classification in a given situation in terms of N. Both Gupta and Matusita have utilized this idea in deriving expressions for the lower bound on the probability of correct classification for the particular problems they considered. Note that the desired probabilities depend on d as well as N, ε and F. For the case where F is discrete, a number of useful inequalities[42] for $f_d(N,\varepsilon,F)$ are available if d is the Jeffreys-Matusita distance.

Apparently not very much is known about the optimum properties of the minimum distance decision rule. The admissibility of the minimum distance rule has been investigated only for the Mahalanobis distance. This, of course,

implies the assumption of normal cdf's where all classes have identical covariance matrices. When the class means are either known or unknown and common covariance is known, Cacoullos[51,52] proved the admissibility of the minimum distance rule in a restricted class of procedures. Srivastava[53] gave an admissible rule for the case where the means and common covariance are unknown. For the two class problem this rule reduces to the minimum distance rule. Both Cacoullos and Srivastava used a zero-one loss function.

CHAPTER 3

THEORETICAL RESULTS

In this chapter we present some theoretical results pertaining to distance measures and minimum distance classification. Although all the results presented concern some aspect of distance measures, or minimum distance classification, their subject matter is rather diverse. Consequently, it seems most appropriate to present each topic individually.

There are essentially three themes underlying the theoretical results. The first is the relationship between distance measures and probability of error in vector classifiers. Sections 3.1 and 3.2 are concerned with this theme. In Section 3.1 we establish a relationship between probability of error and a certain class of distance functions. Section 3.2 deals with a new separability measure defined in terms of random samples and considers some implications of this distance measure regarding probability of error in vector classifiers. The second theme is the relationship between minimum distance classification and other classification rules. This is the basis of Section 3.3 and 3.4 in which we establish certain relationships between minimum distance, nearest neighbor and maximum likelihood classification. The third theme concerns probability of error in minimum distance

classification, and is developed for a simple case in Section 3.5.

As mentioned in Chapter 1 the basic purpose of the theory developed is to provide guidance in conducting experiments and interpreting their results. This is achieved by considering simple situations which give insight into the complex situations of practical interest.

3.1 Probability of Error and a Class of Distance Measures Involving the Likelihood Ratio

Our objective is to prove Theorem 2.4.1 which we will not restate. We use the same notation as in Section 2.4. The proof rests on a theorem of Blackwell's[55] which we state in terms of convex rather than concave functions.

Theorem 3.1.1 (Blackwell)

$P_e(\beta,p) \leq P_e(\theta,p)$ for all p if and only if

$$E_{(g,\beta)}[C(L(\underline{x}|\underline{\beta})] \geq E_{(g,\theta)}[C(L(\underline{x}|\underline{\theta}))]$$

for all continuous convex functions C. Where $P_e(\beta, p)$ is the probability of error using parameter set $[\underline{\beta}^{(f)},\underline{\beta}^{(g)}]$ and prior probabilities $p=[p_f,p_g]$, $E_{(g,\beta)}$ is the expectation with respect to g using parameter set $[\underline{\beta}^{(f)},\underline{\beta}^{(g)}]$ and $L(\underline{x}|\underline{\beta})$ is the likelihood ratio using parameter set $[\underline{\beta}^{(f)},\underline{\beta}^{(g)}]$. $E_{(g,\theta)}$ and $L(\underline{x}|\underline{\theta})$ are defined in a similar manner.

Proof of Theorem

It is apparent from Blackwell's theorem that

$P_e(\beta,p) \leq P_e(\theta,p)$ for all p if and only if $I(E_{(g,\beta)}[C(L(\underline{x}|\underline{\beta}))]) \geq I(E_{(g,\theta)}[C(L(\underline{x}|\underline{\theta}))])$ for all continuous convex functions C, and all strictly increasing real functions of a real variable I. Negating the last statement we have: There exists some p such that $P_e(\beta, p) > P_e(\theta,p)$ if and only if there exists some C and I such that $I(E_{(g,\beta)}[C(L(\underline{x}|\underline{\beta}))]) < I(E_{(g,\theta)}[C(L(\underline{x}|\underline{\theta}))])$ or equivalently there exists some p such that $P_e(\beta,p) > P_e(\theta,p)$ if and only if there exists some $d'_\beta(F,G) < d'_\theta(F,G)$. This follows directly from the definition of d'. The last statement includes Theorem 2.4.1.

3.2 A Separability Measure, Dimensionality and Probability of Error

Much of the theory of pattern recognition is predicated on the underlying assumption that the observation space is a vector space of fixed dimension q. This approach enables the vast, powerful and well developed theory of vector spaces to be applied to the problem. Any pattern recognition journal will testify at a glance to the fruitfulness of this approach.

Problems in which the number of dimensions are variable do not readily fit the vector space approach. Consequently, it is not surprising that results dealing with the interrelationship between dimensionality and other factors, such as sample size and probability of error, are rather sparce. Understanding such relationships is of considerable importance in pattern recognition and the result

we present is in the spirit of fostering such understanding.

For some time it has been known that in a classification problem, in which estimation is involved, the probability of error may exhibit a minimum as a function of observation space dimensionality. That is, classification accuracy may actually decrease when another feature is added. The results of Estes[56], Allais[57], Hughes[58], Abend et al[59], and Kanal and Chandrasekaran[60] provide some insight as to why this occurs. The result we present provides further insight into this phenomenon.

We will consider a two class normal problem in which the covariance matrix Σ for each class is identical and of the form

$$\Sigma = \sigma^2 I \qquad 3.2.1$$

where I is the q dimensional identity matrix. Let $\underline{\eta}$ be the q dimensional vector with all components equal. That is,

$$\underline{\eta} = (\eta_1, \eta_2, .., \eta_q) \quad \text{with } \eta_i = \mu \quad i = 1,2,..,q. \qquad 3.2.2$$

We will assume that $\underline{\mu}^{(1)}$ the mean for class 1 is

$$\underline{\mu}^{(1)} = \underline{\eta} \qquad 3.2.3$$

and that the $\underline{\mu}^{(2)}$, the mean for class 2 is

$$\underline{\mu}^{(2)} = -\underline{\eta} \qquad 3.2.4$$

Consequently, the distance between class means is

$$2\xi = |\mu^{(1)} - \mu^{(2)}| = \sqrt{q}\ (2\mu). \qquad 3.2.5$$

The above assumptions are just as general as assuming the two densities have identical covariance matrices and arbitrary means. This follows because by an affine transformation (i.e., linear transformation plus translation) two densities with identical covariance matrices and arbitrary mean vectors can be put in the assumed form.

For the simple two class model described a separability measure is presently defined in terms of random samples from each class. This distance measure involves the ratio of the expected value of the average pairwise distance between vectors within each class (intra-sample distance) and the expected value of the average pairwise distance between vectors from the two classes (inter-sample distance). The expectation involved is with respect to all possible random samples of a given size. The next section is devoted to obtaining the required expectations.

3.2.1 Expected Value of the Average Intra- and Inter-Sample Distance

Let $\underline{X}_1^{(1)}, \underline{X}_2^{(1)}, \ldots, \underline{X}_{N_1}^{(1)}$ be a random sample of size N_1 for class 1. That is the $\underline{X}_i^{(1)}$'s are independent identically distributed random variables according to the density $N(\underline{\mu}^{(1)}, \Sigma)$. Similarity let $\underline{X}_1^{(2)}, \underline{X}_2^{(2)}, \ldots, \underline{X}_{N_2}^{(2)}$ be a random sample of size N_2 for class 2 from the distribution $N(\underline{\mu}^{(2)}, \Sigma)$. Note that because of the assumed form of the covariance matrix, not only are the $\underline{X}_i$'s independent but the q components of each $\underline{X}_i$ are also independent.

Consider now the average intra-sample distance $D_w^{(i)}(N_i,q)$ for class i defined as

$$D_w^{(i)}(N_i,q) = \frac{1}{n(N_i)} \sum_{j=1}^{N_i-1} \sum_{k=j+1}^{N_i} D_{jk}^{(i,i)} \quad i = 1,2 \qquad 3.2.1.1$$

where $D_{jk}^{(i,i)}$ is the Euclidean distance between $\underline{X}_j^{(i)}$ and $\underline{X}_k^{(i)}$ and $n(N_i)$ is the number of terms in the summation. That is $D_w^{(i)}(N_i,q)$ represents the average pairwise distance between all vectors in the random sample of size N_i for class i.

If we draw a number of random samples of size N_i for class i, we would expect to get a different value of $D_w^{(i)}(N_i,q)$ each time. That is, over all possible random samples that can be drawn for class i, $D_w^{(i)}(N_i,q)$ is a random variable. The expected value of this random variable over all possible random samples is

$$E(D_w^{(i)}(N_i,q)) = \frac{1}{n(N_i)} \sum_{j=1}^{N_i-1} \sum_{k=j+1}^{N_i} E(D_{jk}^{(i,i)}) \quad i = 1,2 \qquad 3.2.1.2$$

For fixed i the random variables $D_{jk}^{(i,i)}$ all have the same distribution for $j = 1,2,\ldots,N_i$; $k = j+1, j+2,\ldots,N_i$. This follows since each $D_{jk}^{(i,i)}$ represents the Euclidean distance between two random vectors with identical distributions. Furthermore, since class 1 and class 2 differ only in location, and the difference of vectors from identical distributions does not depend on location, it follows that the $D_{jk}^{(i,i)}$ have the same distribution regardless of class index i. If we write $R_w(q)$ for $E(D_w^{(i)}(N_i,q))$ and let D* be a

random variable distributed as the identically distributed random variables $D_{jk}^{(i,i)}$, then noting that 3.2.1.2 contains excatly $n(N_i)$ terms we have

$$R_w(q) = E(D^*) \qquad 3.2.1.3$$

Note that $D^* = |\underline{X}^* - \underline{Y}^*|$, where $\underline{X}^*$ and $\underline{Y}^*$ are independent random vectors with the identical distributions $N(\underline{\mu}', \sigma^2 I)$, where $\underline{\mu}'$ is arbitrary. The notation $R_w(q)$ reflects the fact that this quantity depends only on q and is independent of sample size and class index.

In Appendix B Section B.1 we show that if $\underline{X}^* \sim N(\underline{\mu}', \sigma^2 I)$ and $Y^* \sim N(\underline{\mu}', \sigma^2 I)$ then

$$R_w(q) = 2\sigma \frac{\Gamma(\frac{q+1}{2})}{\Gamma(\frac{q}{2})} \qquad q = 1,2,\ldots \qquad 3.2.1.4$$

where $\Gamma(x)$ is the Gamma function defined by

$$\Gamma(x) = \int_0^\infty e^{-t} t^{x-1}\, dt \qquad 3.2.1.5$$

By analogy to 3.2.1.1 we define an average sample distance $D_B(N_1, N_2, q)$ as

$$D_B(N_1, N_2, q) = \frac{1}{n(N_1, N_2)} \sum_{j=1}^{N_1} \sum_{k=1}^{N_2} D_{jk}^{(1,2)} \qquad 3.2.1.6$$

where $D_{jk}^{(1,2)}$ is the Euclidean distance between $\underline{X}_j^{(1)}$ and $\underline{X}_k^{(2)}$ and $n(N_1, N_2)$ is the number of terms in the summation. That is $D_B(N_1, N_2, q)$ represents the average pairwise distance

between all vector pairs, with one vector chosen from class 1, the other from class 2.

Taking the expectations with respect to all random samples we have

$$E(D_B(N_1,N_2,q)) = \frac{1}{n(N_1N_2)} \sum_{j=1}^{N_1} \sum_{k=1}^{N_2} E(D_{jk}^{(1,2)}) \qquad 3.2.1.7$$

By arguments similar to those presented in connection with 3.2.1.2 the distribution of $D_{jk}^{(1,2)}$ is the same for all $j = 1,2,\ldots,N_1$; $k = 1,2,\ldots,N_2$. Let $R_B(q) = E(D_B(N_1,N_2,q)$ and let D** be a random variable distributed as the identically distributed random variables $D_{jk}^{(1,2)}$ then noting that the summation in 3.2.1.6 contains exactly $n(N_1,N_2)$ terms we have

$$R_B(q) = E(D^{**}) \qquad 3.2.1.8$$

Note that $D^{**} = |\underline{X}^{**} - \underline{Y}^{**}|$, where $\underline{X}^{**} \sim N(\underline{\eta},\sigma^2 I)$ and $\underline{Y}^{**} \sim N(-\underline{\eta},\sigma^2 I)$. Again the notation reflects the fact that R_B depends only on q and is independent of N_1 and N_2.

Let us define a signal-to-noise ratio (S/N ratio) S as the square root of the Mahalanobis distance between the density functions for class 1 and class 2. That is,

$$S = [(\underline{\mu}^{(1)} - \underline{\mu}^{(2)})^t \Sigma^{-1} (\underline{\mu}^{(1)} - \underline{\mu}^{(2)})]^{1/2} \qquad 3.2.1.9$$

which for our case reduces to

$$S = \frac{2\xi}{\sigma} = \frac{\sqrt{q}\ (2\mu)}{\sigma} \qquad 3.2.1.10$$

Note that for the simple case under consideration the S/N ratio is simply the distance between the means

divided by the common standard deviation. In Appendix B Section B.2 we show that (writing $R_B(S,q)$ for $R_B(q)$)

$$R_B(S,q) = 2\sigma\frac{\Gamma(\frac{q+1}{2})}{\Gamma(\frac{q}{2})} e^{-(S/2)^2} \Phi\ (\frac{q+1}{2},\ \frac{q}{2},\ (S/2)^2) \quad q = 1,2,\ldots; \qquad 3.2.1.11$$

where $\Phi(a,b,x)$ is the degenerate confluent hypergeometric function defined by the series

$$\Phi(a,b,x) = 1 + \frac{a}{b}\frac{x}{1!} + \frac{a}{b}\frac{(a+1)}{(b+1)}\frac{x^2}{2!} + \cdots \qquad 3.2.1.12$$

If the signal-to-noise ratio is zero, then

$$R_B(0,q) = 2\sigma\frac{\Gamma(\frac{q+1}{2})}{\Gamma(\frac{q}{2})} \qquad 3.2.1.13$$

which is identical to $R_w(q)$.

In Fig. 3.2.1.1 we have plotted the expected value of the average inter-sample distance $R_B(S,q)$ as a function of dimensionality with signal-to-noise ratio as a parameter. By virtue of 3.2.1.13 the S = 0 curve is also a plot of the expected value of the average intra-sample distance.

Qualitatively the quantity $R_w(q)$ is a measure of how tight the distribution in class 1 and 2 are, while $R_B(S,q)$ is a measure of how far apart the two classes are. It is, therefore, reasonable for these quantities to be independent of sample size. The interrelationship between R_w and R_B together with a qualitative concept of these quantities prompts the definition of a measure of seperability

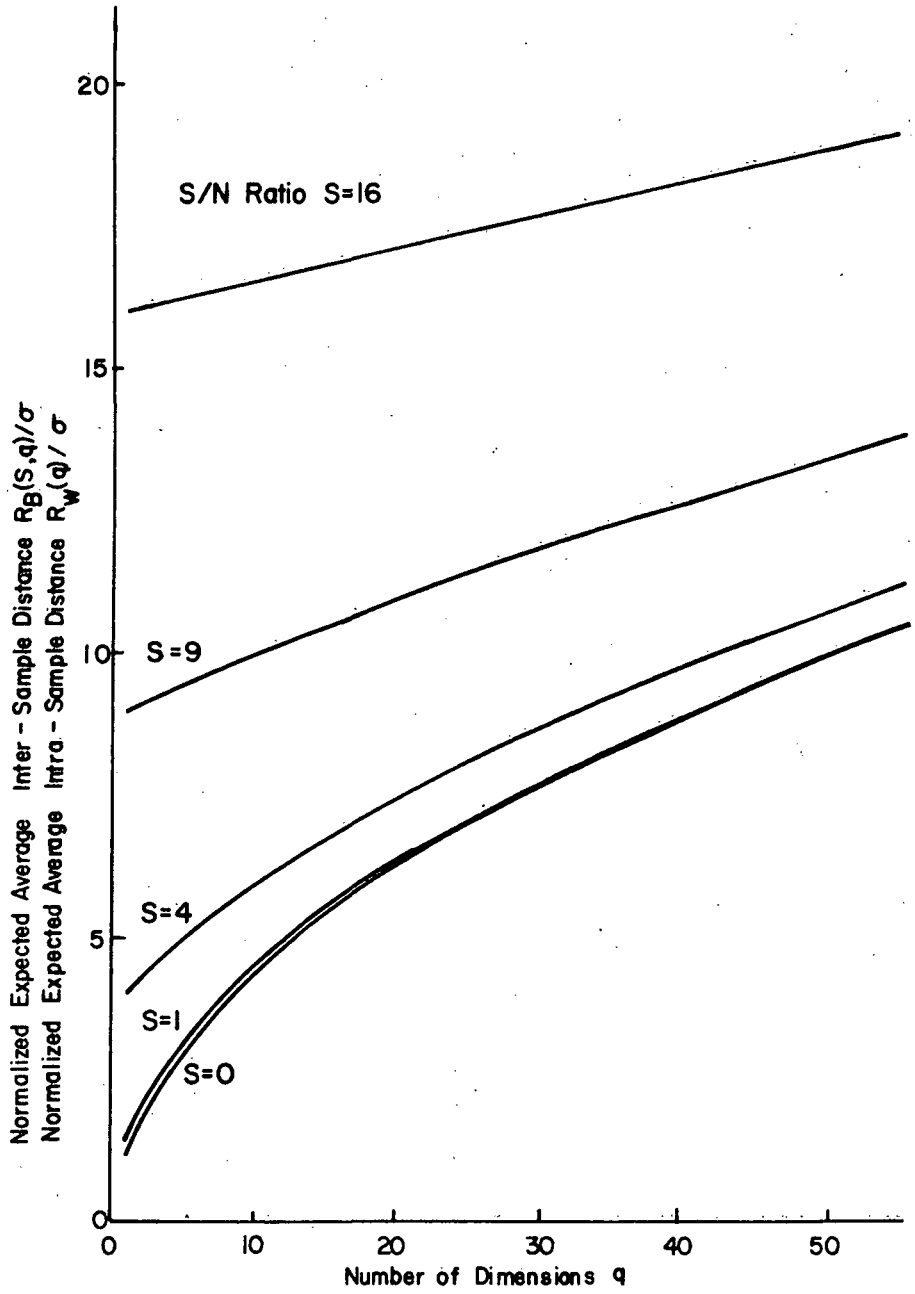


Figure 3.2.1.1 Normalized Expected Average Intra- and Inter-Sample Distance as a Function of Dimensionality.

R(S,q) between the two classes as

$$R(S,q) = \frac{R_B(S,q)}{R_W(q)} = e^{-(S/2)^2} \Phi\left(\frac{q+1}{2}, \frac{q}{2}, (S/2)^2\right) \qquad 3.2.1.14$$

Utilizing the identity

$$\Phi(a,b,x) = e^x \Phi(b-a, b, -x) \qquad 3.2.1.15$$

which is known as Kummers identity, an alternate form for R(S,q) results, namely,

$$R(S,q) = \Phi\left(-\frac{1}{2}, \frac{q}{2}, -(S/2)^2\right) \qquad 3.2.1.16$$

In series form this is the alternating series

$$R(S,q) = 1 + \frac{1}{q}\frac{(S/2)^2}{1!} - \frac{1}{q(q+2)}\frac{(S/2)^4}{2!} + \frac{(1)(3)}{q(q+2)(q+4)}\frac{(S/2)^6}{3!} - \ldots \qquad 3.2.1.17$$

In Fig. 3.2.1.2 R(S,q) is plotted as a function of dimensionality with S/N ratio as a parameter.

It follows from Eq. 3.2.1.17 that regardless of S/N ratio

$$\lim_{q \to \infty} R(S,q) = 1 \qquad 3.2.1.18$$

This fact is also rather evident from Fig. 3.2.1.2. Consider the significance of Eq. 3.2.1.18. Assume for convenience that σ is a constant. Then for fixed S/N ratio the distance between class means is also fixed by virtue of the definition of S/N ratio. Equation 3.2.1.18 states that in the

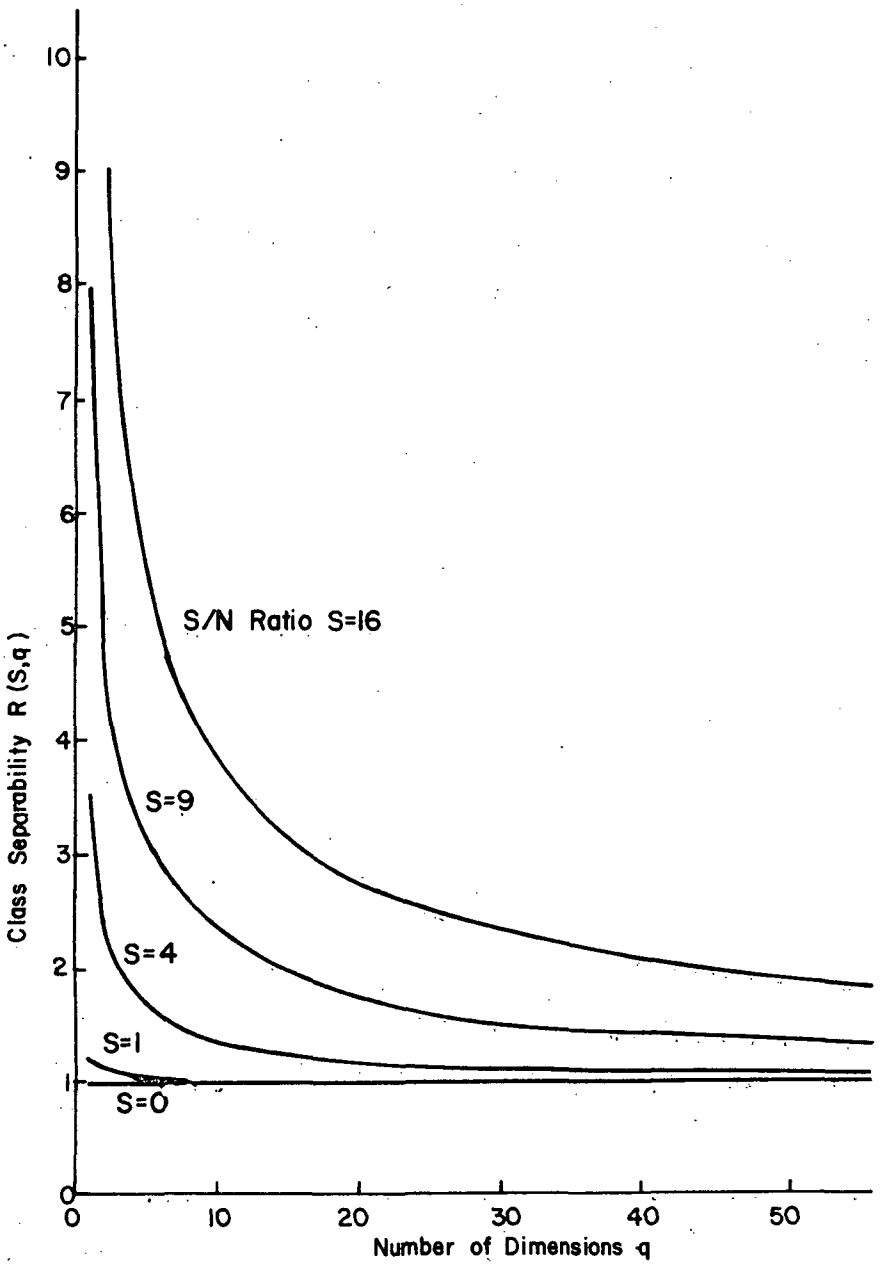


Figure 3.2.1.2 Class Separability vs Dimensionality for Constant S/N Ratio.

limit, as the dimensionality becomes very large, the vectors in class 1 are on the average just as close to vectors in class 2 as they are to vectors in class 1. This means that if one could view the clusters of vectors associated with each class in q dimensional space they would progressively become less and less distinct clusters as the dimensionality is increased.

3.2.2 Classification and Probability of Error

We now present what are essentially some well known results regarding probability of error for vector classifiers for the problem being considered. First we establish that if no estimation is involved then the average probability of error is independent of dimensionality. In the case where estimation of the means is involved we qualitatively discuss how an increase in dimensionality can, in a particular instance, increase the probability of error, and further suggest that on the average we should expect such an increase. We also suggest such an increase would be expected from considering the behavior of the separability measure R.

If the common covariance matrix and class means are known, then it is well known that for equal priors and a zero-one loss function, the minimum risk decision rule for classifying an unknown vector $\underline{X}^{(u)}$ into one of the two classes is the maximum likelihood decision rule.[61] This rule assigns $\underline{X}^{(u)}$ to the class whose density function is

largest at $\underline{X}^{(u)}$. This rule partitions the observation space into two disjoint regions by a hyperplane. The hyperplane passes through

$$\underline{\mu}_M = 1/2\ (\underline{\mu}^{(1)}+\underline{\mu}^{(2)}) \qquad 3.2.2.1$$

and is perpendicular to

$$\Delta\underline{\mu} = \underline{\mu}^{(1)}-\underline{\mu}^{(2)} \qquad 3.2.2.2$$

In this case the probability of error P_E is independent of the number of dimensions q and is given by

$$P_E = 1-Q\ (\frac{\xi}{\sigma}) = 1 - Erf(\frac{\xi}{\sqrt{2}\sigma}) \qquad 3.2.2.3$$

where Q(x) is the probability integral

$$Q(x) = \frac{1}{\sqrt{2\pi}} \int_{-x}^{x} e^{-1/2\ t^2} dt \qquad 3.2.2.4$$

and Erf(x) is the error function

$$Erf(x) = \frac{2}{\sqrt{\pi}} \int_{o}^{x} e^{-t^2} dt \qquad 3.2.2.5$$

To show 3.2.2.3 is valid we consider the rotated coordinate system with axes x_1', x_2', ..., x'_q centered at $\underline{\mu}_M$ with the positive x_1' axis oriented along the vector $\Delta\underline{\mu}$. Let $\underline{X}'$ be the unknown vector in this coordination system. Since the separating hyperplane is orthogonal to the x_1' axis the only component of the transformed unknown vector that enters into the decision rule is the first (i.e., X_1'). Now in the transformed coordinate system

$X_1' \sim N(\xi,\sigma^2)$ if class 1 is active and

$X_1' \sim N(-\xi,\sigma^2)$ if class 2 is active

Consequently, 3.2.2.3 follows.

Now consider the case where the common covariance matrix is known but the mean vectors are unknown. We will not derive an expression for the probability of error for this case, but only make some general observations. Since the class means are unknown they must be estimated. Let $\hat{\underline{\mu}}^{(1)}$ and $\hat{\underline{\mu}}^{(2)}$ be the estimated mean vectors for class 1 and class 2 respectively. For convenience assume that each estimate is based on a sample of size N. If the sample mean is used as the estimator, then

$$\hat{\underline{\mu}}^{(i)} = \frac{1}{N} \sum_{j=1}^{N} \underline{X}_j^{(i)} \qquad i = 1,2. \qquad 3.2.2.6$$

Since the $\hat{\underline{\mu}}^{(i)}$ are a sum of independent gaussian random variables it follows that $\hat{\underline{\mu}}^{(i)} \sim N(\underline{\mu}^{(i)}, \frac{\sigma^2}{N} I)$ $i = 1,2$

For a decision rule we use the maximum likelihood rule with the class means replaced by their estimates. As before this rule partitions the observation space into disjoint regions associated with class 1 and class 2. Since the covariance matrices are equal the partitioning surface is a hyperplane orthogonal to

$$\hat{\Delta\underline{\mu}} = \hat{\underline{\mu}}^{(1)} - \hat{\underline{\mu}}^{(2)} \qquad 3.2.2.7$$

which passes through the point $\hat{\underline{\mu}}_M$, where

$$\hat{\underline{\mu}}_M = \frac{1}{2}(\hat{\underline{\mu}}^{(1)} + \hat{\underline{\mu}}^{(2)}) \qquad 3.2.2.8$$

Note that $\hat{\underline{\mu}}_M \sim N(\underline{\mu}_M, \frac{\sigma^2}{2N} I)$.

Since $\hat{\underline{\mu}}^{(1)}$ and $\hat{\underline{\mu}}^{(2)}$ are random variables the partitioning hyperplane is random in location and orientation. The probability of error $P_E(N,q)$ is consequently a random variable since it depends on the partitioning hyperplane. We observe that the expected value of $P_E(N,q)$ over all possible samples must be larger than the probability of error for the case where the means as well as the common covariance are known. This follows since any hyperplane must yield a probability of error that is at least as large as the probability of error for the optimum hyperplane.

With regard to varying the dimensionality the following observations can be made as the dimensionality decreases from 2 to 1. First note that the probability of deciding a vector came from class 2 when class 1 is active (i.e., $P(2|1)$) depends only on Σ and the perpendicular distance $d^{(1)}$ between $\underline{\mu}^{(1)}$ and the separating hyperplane. The smaller the distance $d^{(1)}$ the larger is $P(2|1)$. A similar statement applies to $P(1|2)$ and $d^{(2)}$. Consider now an arbitrary realization of the random variable $\hat{\underline{\mu}}_M$. The "best" possible hyperplane for the observed value of $\hat{\underline{\mu}}_M$ is the hyperplane perpendicular to $\Delta\underline{\mu}$; but the probability that the separating hyperplane which is perpendicular to $\Delta\hat{\underline{\mu}}$

coincides with the "best" hyperplane is zero, since a continuum of possible separating hyperplanes pass through $\hat{\underline{\mu}}_M$. Suppose now that for every realization of $\hat{\underline{\mu}}_M$ the "best" possible hyperplane is used as the discriminant surface rather than the hyperplane orthogonal to $\Delta\underline{\mu}$. It is clear that on the average, over all possible realizations of $\hat{\underline{\mu}}_M$, this procedure reduces the probability of error. But the collection of the "best" hyperplanes for the two dimensional case are precisely the collection of hyperplanes used in the one dimensional case. Furthermore, the "probability" of selecting a particular hyperplane from this collection is precisely the same in the two cases. This follows since the distribution of $\hat{\underline{\mu}}_M$ projected on the vector $\Delta\underline{\mu}$ for the two dimensional case is identical to the distribution of $\hat{\underline{\mu}}_M$ for one dimension. It, therefore, follows that the average probability of error increases as the dimensionality is increased from 1 to 2. Actually the above argument can be extended to the case where the dimensionality is increased from q to q + 1 dimensions. Consequently, the average probability of correct classification is a monotonically decreasing function of dimensionality.

Returning now to the separability measure R we note that it is also a monotonically decreasing function of dimensionality, just as is the average probability of correct classification. It is not known how closely R is related to the average probability of correct classification.

On the basis of the behavior of R with dimensionality for fixed S/N ratio one would expect that the probability of error would increase with dimensionality. An alternate point of view is that as the dimensionality is increased the estimated location of the separating hyperplane must improve, or else the probability of error will increase because the random samples become less distinct.

3.2.3 Separability for S/N Ratio a Function of Dimensionality

Experimentally it is usually true that the probability of error decreases with increasing dimensionality, at least for low values of q. We attribute this to the fact that the signal-to-noise ratio is usually a rapidly increasing function of dimensionality for low values of q, rather than a constant as was assumed in the previous section. The increasing S/N ratio tends to override the effect of increase in dimensionality. In the absence of an exact analysis for the average probability of error, it is not possible to investigate the interrelationship between S/N ratio, probability of error, and dimensionality. We can, however, investigate such a interrelationship for our separability criterion R since we can incorporate in R a signal-to-noise ratio which varies in some manner with q. One reasonable assumption might be to assume a constant signal-to-noise ratio per dimension, rather than a constant overall signal to noise ratio. By signal-to-noise ratio per

dimension we mean the quantity.

$$S_d = \frac{\mu_j^{(1)} - \mu_j^{(2)}}{\sigma} = \frac{2\mu}{\sigma} \quad j = 1,2,\ldots,q \qquad 3.2.3.1$$

Note that by 3.2.1.10

$$S = \sqrt{q}\, S_d \qquad 3.2.3.2$$

We can use this value for S in the expression for R(S,q) and determine R(S,q) as a function of q for various fixed values of S_d. For this situation

$$R(S,q) = \Phi(-\tfrac{1}{2}, \tfrac{1}{q}, -q(S_d/2)^2) \qquad 3.2.3.3$$

Expressed in series form 3.2.3.3 becomes

$$R(S,q) = 1 + \frac{1}{1!}(S_d/2)^2 - \frac{q}{(q+2)2!}(S_d/2)^4 + \frac{1\cdot 3\, q^2}{(q+2)(q+4)3!}$$
$$(S_d/2)^6 - \ldots \qquad 3.2.3.4$$

Figure 3.2.3.1 is a plot of 3.2.3.3 with signal-to-noise ratio per dimension as a parameter. It may immediately be noted that for the range of the q considered R(S,q) given by Fig. 3.2.3.3 decreases very slowly with q except for low values of q. In Appendix B Section B.3 the limit of 3.2.3.4 as $q \to \infty$ is examined. The result obtained is that

$$\lim_{\gamma \to \infty} R(S,q) = 1 + \frac{1}{1!}(S_d/2)^2 - \frac{1}{2!}(S_d/2)^4 + \frac{1\cdot 3}{3!}(S_d/2)^6 - \ldots \qquad 3.2.3.5$$

This series converges only if $S_d \leq \sqrt{2}$. For $S_d > \sqrt{2}$ the series oscillates since successive terms ultimately become larger and larger. Although 3.2.3.4 is not well behaved for

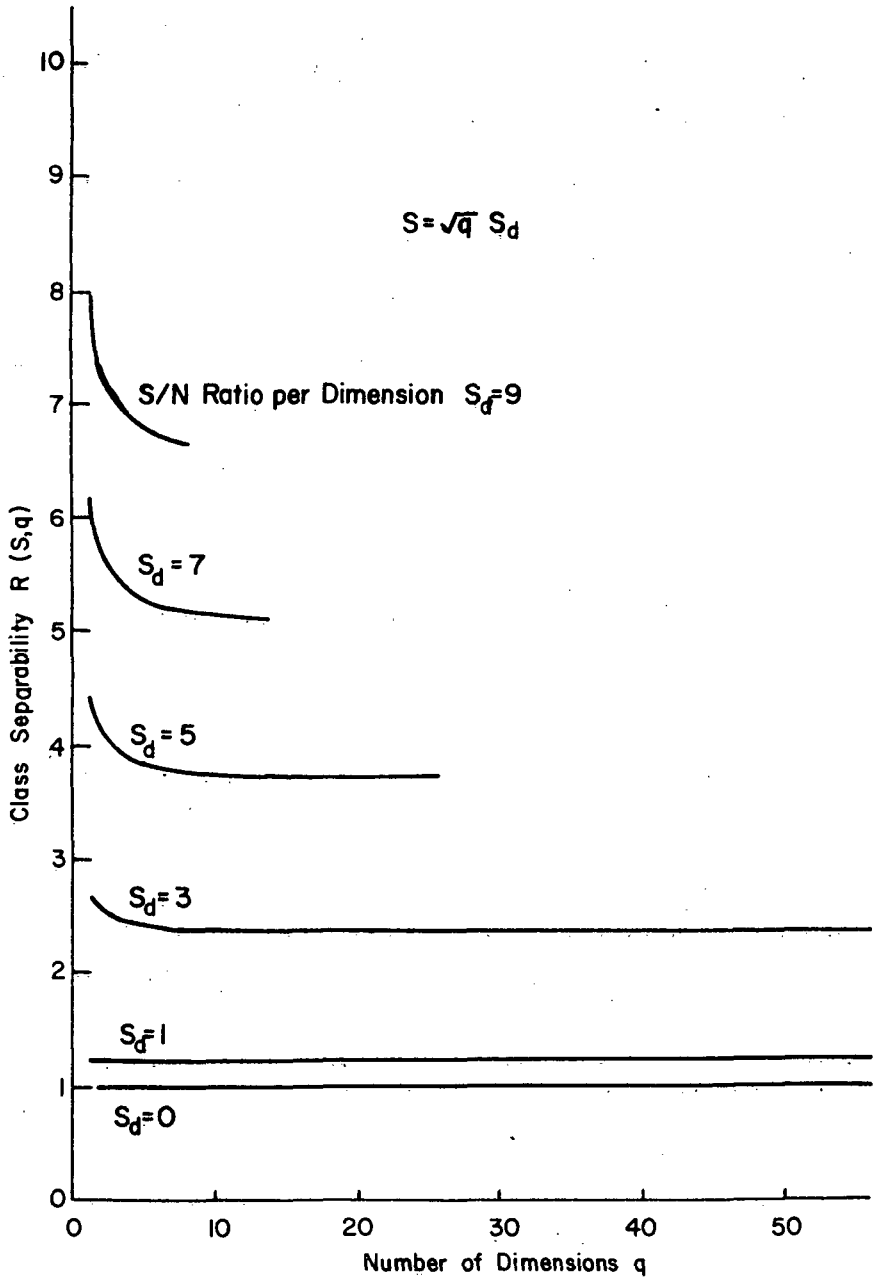


Figure 3.2.3.1 Class Separability vs Dimensionality for Constant S/N Ratio Per Dimension.

infinite q, for fixed q it does converge for all S_d. Consequently no problems are encountered in evaluating the series.

In practice it is probably unrealistic to assume that the total signal-to-noise ratio can be increased indefinitely by adding more and more dimensions as is implied by a constant signal to noise ratio per dimension. Perhaps a more reasonable assumption is to assume that there is some limiting signal to noise ratio S_L. One possible choice is an exponential variation of S with q. That is S is assumed to be of the form

$$S = S_L(1 - e^{\frac{q}{\tau}}) \qquad 3.2.3.6$$

The constant τ reflects how rapidly S approaches its limiting value S_L as a function of q.

Using 3.2.3.6 as S in the expression for R(S,q) the value of R(S,q) has been determined as a function of q for various values of S_L for $\tau = 5$. These results are plotted in Fig. 3.2.3.2. The most interesting factor about these curves is that they exhibit a maximum suggesting that the separability first increases and then decreases with increasing q. The limiting behavior for increasing q is the same as for fixed signal-to-noise ratio.

Two basic observations can be made regarding the development of the separability measure R. The first and most important is that it is based on the expected average

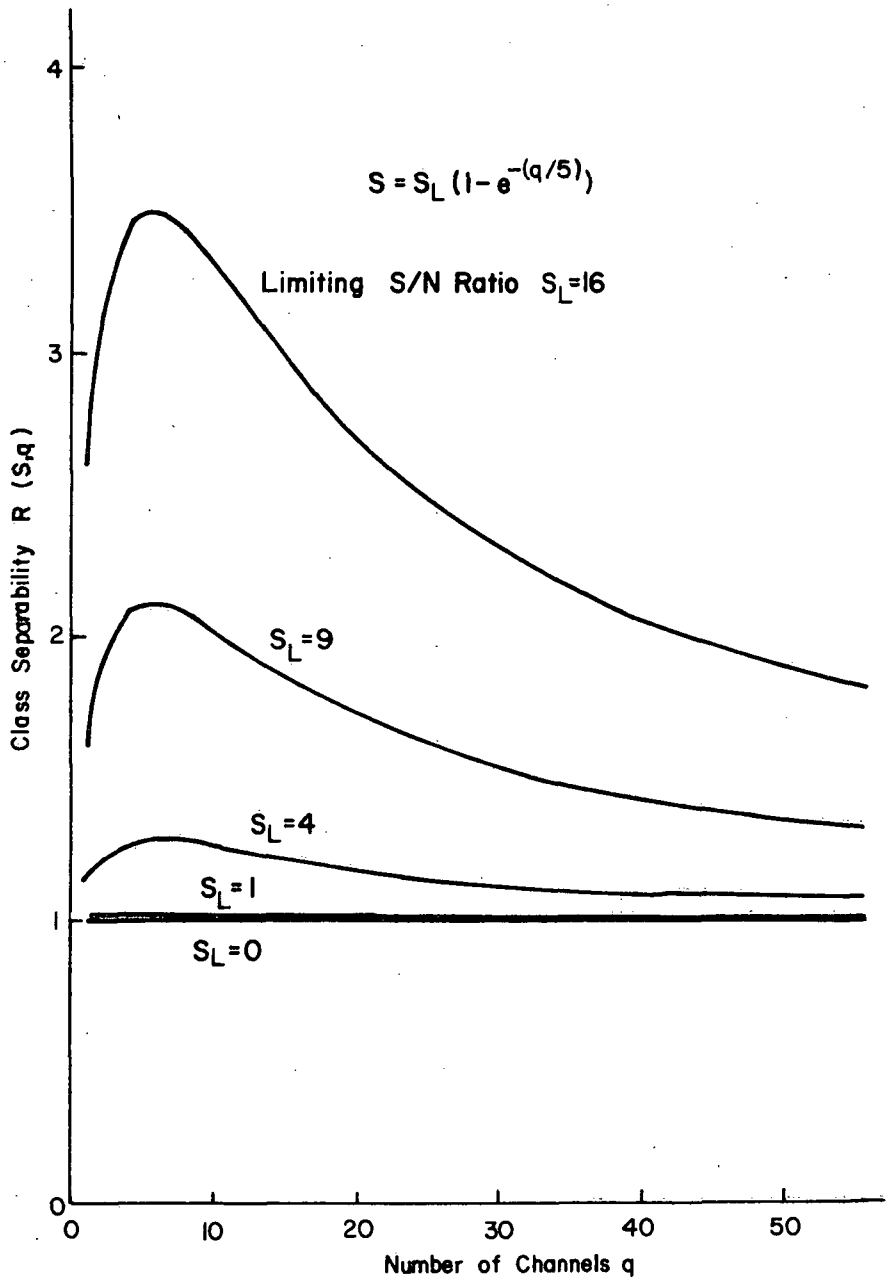


Figure 3.2.3.2 Class Separability vs Dimensionality for a Saturating S/N Ratio.

pairwise distances between all vector pairs, where the vector pairs originate from one (for the intra-class distance) or two (for the inter-class distance) random samples. Thus in essence the separability measure is completely nonparametric and in no way depends on the normal assumption. The normal assumption is made to simplify computations. The second important observation concerns the definition of signal-to-noise ratio and the assumed functional relationship between dimensionality and signal-to-noise ratio. The specific form here does depend on the normal assumption in that signal-to-noise ratio is defined in terms of the Mahalanobis distance. This dependence could be removed by defining signal-to-noise ratio in terms of a more general distance. For example, if we used the Bhattacharyya distance, which reduces to the Mahalanobis distance for the case considered, then the normal assumption could be removed. We make these comments since we are interested in extrapolating to more complex cases and it appears to us that the general behavior of the separability measure R is in fact not dependent on the underlying densities, at least for fairly well behaved densities. In particular for constant and saturating signal to noise ratios one would expect R to approach 1 as q approached infinity for most densities. Also one would expect that R could exhibit a maximum regardless of the densities involved provided the signal to noise ratio saturates with dimensionality.

While no direct relation between R and probability of error has been established we believe that R provides some insight into the mechanism by which dimensionality affects probability of error in a classification problem involving estimation. In particular, the decrease in R with dimensionality for fixed signal to noise ratio suggests that for this case the estimated location of the discriminant surfaces used in classification must improve with dimensionality or probability of error will increase.

3.3 A Relationship Between Maximum Likelihood and Minimum Distance Decision Rules

It is well known[61] that to classify a random vector $\underline{Y}$ drawn from one of k known populations, the expected loss (i.e., risk) is minimized provided we decide $\underline{Y}$ belongs to class m if

$$\sum_{i=1}^{k} p_i L(m,i) f_y^{(i)}(\underline{Y}) = \min_{j=1,..,k} \left[\sum_{i=1}^{k} p_i L(j,i) f_y^{(i)}(\underline{Y}) \right] \quad 3.3.1a$$

where p_i is the prior probability that $\underline{Y}$ belongs to class i, $L(j,i)$ is the loss incurred in deciding $\underline{Y}$ belongs to class j, when it was drawn from class i, and $f_y^{(i)}(\underline{y})$ is the known probability density for class i. In case 3.3.1a results in ties these can be broken in an arbitrary manner provided the probabilities of ties is zero.

For the zero-one loss function (i.e., $L(i,j) = 0 \quad i = j$; $L(i,j) = 1, \ i \neq j$) and equal priors 3.3.1a reduces to decide $\underline{y}$ belongs to class m if

$$f_y^{(m)}(\underline{Y}) = \max_{j=1,\ldots,k} f_y^{(j)}(\underline{Y}) \qquad 3.3.1b$$

If $\underline{Y} = (\underline{X}_1, \underline{X}_2, \ldots, \underline{X}_N)$ where the $\underline{X}$'s ϵE^q constitute a random sample of size N from $f(\underline{x})$ then 3.3.1b is equivalent to decide class m active if

$$\prod_{i=1}^{N} f^{(m)}(\underline{X}_i) = \max_{j=1,\ldots,k} \prod_{i=1}^{N} f^{(j)}(\underline{X}_i) \qquad 3.3.1c$$

where $f^{(m)}(\underline{x})$ is the q dimensional density for class m.

If the class densities are not known it is common to replace the unknown densities above by appropriate sample-based estimates. Thus for 3.3.1a we have decided class m active if

$$\sum_{i=1}^{k} p_i L(m,i) \tilde{f}_y^{(i)}(\underline{Y}) = \min_{j=1,\ldots,k} \sum_{i=1}^{k} p_i L(j,i) \tilde{f}_y^{(i)}(\underline{Y}) \qquad 3.3.2a$$

and for 3.3.1b decide class m active if

$$\tilde{f}_y^{(m)}(\underline{Y}) = \max_{j=1,\ldots,k} \tilde{f}_y^{(j)}(\underline{Y}) \qquad 3.3.2b$$

and for 3.3.1c decide class m active if

$$\prod_{i=1}^{N} \tilde{f}^{(m)}(\underline{X}_i) = \max_{j=1,\ldots,k} \prod_{i=1}^{N} \tilde{f}^{(j)}(\underline{X}_i) \qquad 3.3.2c$$

In 3.3.2 $\tilde{f}_y^{(j)}(\underline{y})$ and $\tilde{f}^{(j)}(\underline{x})$ are the sample-based estimated densities for $f_y^{(j)}(\underline{y})$ and $f^{(j)}(\underline{x})$ respectively j = 1, 2, ...k.

The relationship that is established between minimum distance and maximum likelihhod classification in essence asserts that if density histograms are used to estimate the densities, and KL numbers are used as the distance measure in the minimum distance rule; then excluding ties,

both classification rules produce identical results. This relationship is now stated more precisely.

Statement of Relationship Between Minimum Distance and Maximum Likelihood Classification

Let $f^{(j)}(\underline{x})$ be the pdf for class $j = 1,2,\ldots,k$ and $F^{(j)}(\underline{x})$ the corresponding cdf. Let $\underline{X}_i^{(j)}$ $i = 1,2,\ldots,N_j$ be a random sample of size N_j from $f^{(j)}(\underline{x})$. Let $\underline{X}_i^{(u)}$ $i = 1,2,\ldots,N$ be a random sample from $f^{(u)}(\underline{x})$ where u is an unknown integer between 1 and k. Further let $\dot{D}_{ML}$ be the maximum likelihood decision rule which decides u = m (i.e., unknown random sample belongs to class m) in case

$$\prod_{i=1}^{N} \dot{f}^{(m)}(\underline{X}_i) = \max_{j=1,\ldots,k} \prod_{i=1}^{N} \dot{f}^{(j)}(\underline{X}_i) \qquad 3.3.3$$

and let $\dot{D}_{MD}$ be the minimum distance decision rule which decides u = m in case

$$d(\dot{F}^{(u)},\dot{F}^{(m)}) = \min_{j=1,\ldots,k} d(\dot{F}^{(u)},\dot{F}^{(j)}) \qquad 3.3.4$$

where the distance d(F,G) between arbitrary densities F and G, with corresponding pdf's and f and g, is the KL number of density f for g given by

$$L_{fg} = \int_{-\infty}^{\infty} \mathrm{Ln}\,\frac{f(\underline{x})}{g(\underline{x})}\, f(\underline{x})\, d\underline{x} \qquad 3.3.5$$

and the $\cdot$ indicates density histograms are used as estimators.

Then the relationship established is that,

excluding ties, the maximum likelihood decision rule 3.3.3 and the minimum distance decision rule 3.3.4 make the same decisions.

It is relatively simple to prove the above relationship but first a few comments regarding the assumed behavior of 3.3.5 in regions were one or both of the densities involved are zero. If in E^q there exists a finite region where $g(\underline{x})$ is zero but $f(\underline{x})$ is not zero then L_{fg} is infinite. The integral over a region where $f(\underline{x})$ is zero, but $g(\underline{x})$ is not zero, is assumed to be zero. This is justified by noting that for arbitrary finite c

$$\underset{t \to \infty}{\text{Limit}} \; t \, \text{Ln}(ct) = \hat{0} \qquad 3.3.6$$

The integral over regions where both densities are zero is taken to be zero, because such region should not influence the distance between distributions.

It is important to note that in order for the KL number of density histogram $\dot{f}^{(u)}$ for $\dot{f}^{(j)}$ to be finite the bins occupied by $\dot{f}^{(u)}$ must be a subset of those occupied by $\dot{f}^{(j)}$. In most practical minimum distance classification situations infinite KL numbers would probably occur so frequently that an unknown density would often be an infinite distance from all classes. Modifications to the definition of KL numbers would probably be necessary to utilize this approach in a practical classification scheme. A somewhat

similar situation prevails with regard to the maximum likelihood rule where $\prod_{i=1}^{N} \dot{f}^{(j)}(\underline{X}_i) = 0$ unless all $\underline{X}_i^{(u)}$'s fall in the bins where $\dot{f}^{(j)}$ is not zero. Again some modifications would probably be necessary in a practical situation. In both minimum distance and maximum likelihood classifications the modifications would be aimed at alleviating the situation were disagreement in a few bins can completely dominate the result. While the behavior described above is of considerable practical importance it does not affect any theoretical investigation.

The stated relationship between minimum distance and maximum likelihood classification will now be proven. Taking logarithms of both sided of 3.3.3 we have

$$\sum_{i=1}^{N} \mathrm{Ln}(\dot{f}^{(m)}(\underline{X}_i)) = \underset{j=1,..,k}{\mathrm{Max}} \sum_{i=1}^{N} \mathrm{Ln}(\dot{f}^{(j)}(\underline{X}_i)) \qquad 3.3.7$$

In 3.3.7 the summation is over all vectors in the unknown sample. This can be written as the summation over the bins occupied by the unknown sample. Let $k_i^{(u)}$ be the number of vectors from the unknown sample that fall in the ith bin of the unknown density histogram and let N_b be the number of nonempty bins in the density histogram of the unknown sample, and let $\dot{f}^{(j)}(i)$ be the estimate for the density of the j'th class, in the i'th bin of the unknown density histogram. Then 3.3.7 becomes

$$\sum_{i=1}^{N_b} k_i^{(u)} \mathrm{Ln}(\dot{f}^{(m)}(i)) = \underset{j=1,..,k}{\mathrm{Max}} \sum_{i=1}^{N_b} k_i^{(u)} \mathrm{Ln}(\dot{f}^{(j)}(i)). \qquad 3.3.8$$

If b^N is the bin volume then dividing both sides of 3.3.8 by Nb^N and recognizing that

$$\dot{f}^{(u)}(i) = \frac{k_i^{(u)}}{Nb^N} \qquad 3.3.9$$

we have

$$\sum_{i=1}^{N_b} \dot{f}^{(u)}(i) \mathrm{Ln}(\dot{f}^{(m)}(i)) = \max_{j=1,\ldots,k} \sum_{i=1}^{N_b} \dot{f}^{(u)}(i)\ \mathrm{Ln}(\dot{f}^{(j)}(i)) \qquad 3.3.10$$

Multiplying 3.3.10 by minus one changes the Max operation to a Min operation and then adding the constant $\sum_{i=1}^{N_b} \dot{f}^{(u)}(i)$ Ln $(\dot{f}^{(u)}(i))$ to both sides yields the decision rule to announce m = u in case

$$\sum_{i=1}^{N_b} \dot{f}^{(u)}(i) \mathrm{Ln}\left(\frac{\dot{f}^{(u)}(i)}{\dot{f}^{(m)}(i)}\right) = \min_{j=1\ldots,k} \sum_{i=1}^{N_b} \dot{f}^{(u)}(i) \mathrm{Ln}\left(\frac{\dot{f}^{(u)}(i)}{\dot{f}^{(j)}(i)}\right) \qquad 3.3.11$$

But this is precisely the minimum distance decision rule using density histograms as density estimators and KL numbers, of the unknown density for the class density, as the distance measure. Thus the stated relation between minimum distance and maximum likelihood has been established.

3.4 On the Equivalence of the Minimum Distance and Nearest Neighbor Decision Rules

By the nearest neighbor rule[3] we mean a non-parametric decision procedure which classifies an unknown vector $\underline{X} \in E^S$ into the category of its nearest neighbor in

terms of some metric in E^s. Actually a number of variations of the nearest neighbor rule are in existence.[3,4,5,6] The type of equivalence we establish is such that each of the "nearest neighbor" rules has an equivalent "minimum distance" analog.

We will concern ourselves only with the case where Ω is a parametric family which can be characterized by s real parameters. There are several reasons why equivalence between minimum distance and nearest neighbor rules would be useful. Perhaps the most important is that theoretical results available for nearest neighbor rules would be directly applicable to our problem. Another equally important consideration is the fact that this equivalence enables us to choose reasonable metrics in the parameter space.

By parameter space we of course mean the space whose coordinate axes are defined by the parameters of the family of densities involved. For example, for the univariate normal family the parameter space is two dimensional, as two parameters are required to define a univariate normal probability density function. These two parameters are the mean and variance (or standard deviation) of the density. The axes of this two dimensional parameter space correspond to these two parameters. Every univariate normal density is represented by a single point in this parameter space. The location of the point corresponds to the mean and variance of the density in question. For example the

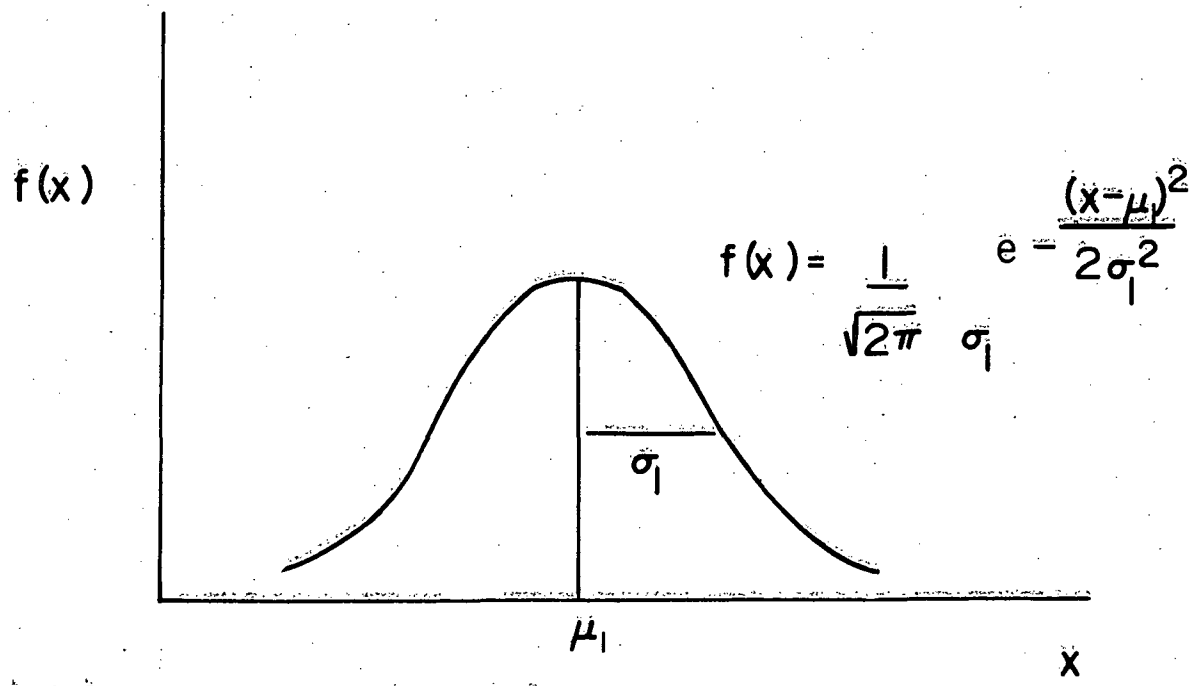
density in Fig. 3.4.1 is represented by the point z in Fig. 3.4.2.

No one would argue against the proposition that in a parametric problem characterizable in E^S, one could use a nearest neighbor decision rule in E^S. For example, to classify univariate normal distribution functions we could use a nearest neighbor rule in the parameter space depicted in Fig. 3.4.2. The choice of metric, however, presents a dilemma. Should the mean and variance be given equal weight in calculating distance or not? That is, should we or should we not use the Euclidean metric. Clearly a method of choosing a metric is required. The equivalence established enables us to choose a metric in the space of distribution functions which in turn generates a metric in the parameter space. In the space of distribution functions, metrics are available which are known to have some good theoretical properties. For example, Bhattacharyya distance is known to have the property of Theorem 2.4.1.

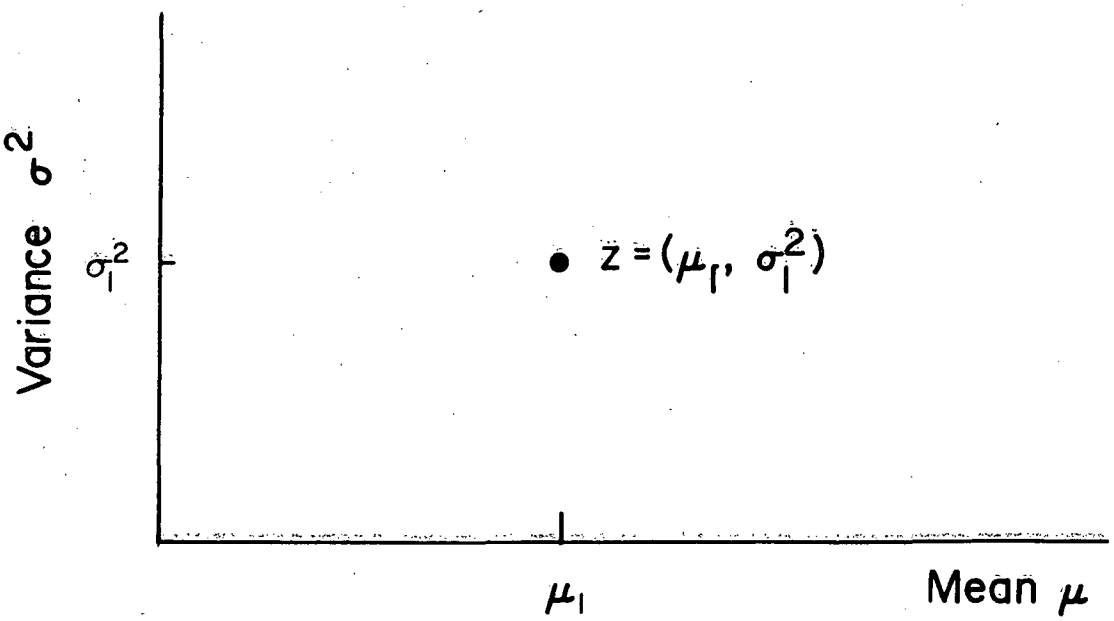
We now prove the following theorem involving the equivalence of minimim distance and nearest neighbor rules.

Theorem 3.4.1

Let Ω be a parametric family such that there exists a one to one correspondence between $F(\underline{x}|\underline{\theta})\varepsilon\Omega$ and $\underline{\theta}\varepsilon S\subseteq E^S$. Here $\underline{\theta}$ is the parameter vector characterizing F. Let $F(\underline{x}|\underline{\alpha})$ and $F(\underline{x}|\underline{\beta})$ be arbitrary elements of Ω with parameter $\underline{\alpha}$ and $\underline{\beta}$ respectively.



3.4.1 A Univariate Normal Density.



3.4.2 Parameter Space Representation of a Univariate Normal Density.

Consider a metric δ in Ω. Since Ω is a parametric family we can view δ as some function δ^* of the parameters. That is $\delta(F(\underline{x}|\underline{\alpha}), F(\underline{x}|\underline{\beta})) = \delta^*(\underline{\alpha},\underline{\beta})$. The theorem asserts that δ^* is a metric in S.

Proof of Theorem 3.4.1

The proof is very simple since we need only show that δ^* satisfies the metric properties in S. That is we need to show for arbitrary $\underline{u}$, $\underline{v}$, $\underline{w} \in S$ that

(a) $\delta^*(\underline{u},\underline{v}) \geq 0$

(b) $\delta^*(\underline{u},\underline{v}) = 0$ if and only if $\underline{u} = \underline{v}$ 3.4.1

(c) $\delta^*(\underline{u},\underline{v}) = \delta^*(\underline{v},\underline{u})$

(d) $\delta^*(\underline{u},\underline{v}) + \delta^*(\underline{v},\underline{w}) \geq \delta^*(\underline{u},\underline{w})$

To prove part (a) we note that because of the one to one correspondence between elements of S and Ω for arbitrary $\underline{u}$, $\underline{v} \in S$ there exists cdf's $F(\underline{x}|\underline{u})$, $F(\underline{x}|\underline{v})$ in Ω with parameters $\underline{u}$ and $\underline{v}$ respectively. By the definition of δ^* we have $\delta^*(\underline{u},\underline{v}) = \delta(F(\underline{x}|\underline{u}), F(\underline{x}|\underline{v}))$ but $\delta(F(\underline{x}|\underline{u}), F(\underline{x}|v)) \geq 0$ since δ is a metric in Ω. Therefore, $\delta^*(\underline{u},\underline{v}) \geq 0$ for arbitrary $\underline{u}$, $\underline{v} \in S$. Proofs for parts (b), (c), and (d) follow in analogous fashion.

Corollary 3.4.1

If δ only satisfies some subset of the metric axioms in Ω, then δ^* satisfies the same subset of metric asioms in S. In particular, a distance d in Ω generates a distance d* in S.

3.5 Minimum Distance Rule and Expected Probability of Error--Two Class Problem

Although the theoretical solution for the probability of error for most realistic, multispectral analysis problems does not appear tractable, it is instructive to consider grossly simplified situations which can be solved analytically. Such examples do provide some insight into more complex situations and are invaluable in guiding and interpreting experiments.

3.5.1 General Two Class Parametric Problem--Known Distributions

We consider a two class parametric problem in which the distributions are known and each class has infinitely many subclasses (Type I, case (a)). We will assume that even though all the distributions are known only a random subset, selected according to the parameter space distribution $H^{(i)}$, will be used to represent each class. The objective of this approach is to gain insight into the practical case where the distributions are unknown, without introducing the mathematical complexity that results when sample based estimates are used. The results should be approximately valid for the case where consistent estimators are used and a large number of vectors are available for estimating each density.

Let $H^{(i)}$ be the distribution over the parameter space for class i; i = 1,2. Let the set of distributions $\Lambda^{(i)}$ selected to represent class i be

$$\Lambda^{(i)} = \Omega_{M_i}^{(i)} = [{}_1F^{(i)}, {}_2F^{(i)}, \ldots, {}_{M_i}F^{(i)}] \quad i = 1,2 \qquad 3.5.1.1$$

Here the "training distributions" ${}_kF^{(i)}$ are the cdf's obtained by selecting a random sample of size M_i from the parameter space distribution $H^{(i)}$. Note that i indexes the class while k indexes the subclass. The average probability of error for the two class case can be written as

$$P_E = p_1P_1 + p_2P_2 \qquad 3.5.1.2$$

where p_1, p_2 are the prior probabilities of class 1 and class 2 respectively; P_E is the total average probability of error, and P_i is the average probability of erroneously classifying a distribution into class i. The averaging to obtain P_E and P_i is with respect to all random training sets of size M_1 from $H^{(1)}$ and M_2 from $H^{(2)}$, and over all possible parameter space realizations of the random parameter vector $\underline{\Theta}$.

Let $P_i(\underline{\theta})$ be the average probability (over all random training sets) of misclassifying into class i a distribution F characterized by the fixed parameter vector $\underline{\theta}$. Then allowing for all possible $\underline{\theta}$ the average probability of misclassifying a random sample from class j is

$$P_i = \int_{-\infty}^{\infty} P_i(\underline{\theta})h^{(j)}(\underline{\theta})d\underline{\theta} \quad i,j = 1,2; \quad i \neq j \qquad 3.5.1.3$$

where $h^{(j)}(\underline{\theta})$ is the parameter space density of $\underline{\Theta}$ for class j.

As before let F be the unknown distribution characterized by the fixed parameter space vector $\underline{\theta}$. Define the random variables

$$ {}_kD^{(i)}(\underline{\theta}) = d(F, {}_kF^{(i)}) \quad k = 1,2,\ldots,M_i;\ i = 1,2 \qquad 3.5.1.4 $$

Note that ${}_kD^{(i)}(\underline{\theta})$ is the distance between the unknown distribution and the k'th subclass of the i'th class given that the unknown distribution is characterized by $\underline{\theta}$. Also note that for fixed i and $\underline{\theta}$ the ${}_kD^{(i)}(\underline{\theta})$ are k independent identically distributed random variables over all random sets of M_i distributions selected to represent class i. Let $G^{(i)}(\underline{u}|\underline{\theta})$ be the common cdf of ${}_kD^{(i)}(\underline{\theta})$ $k = 1,2,\ldots,M_i$, $i = 1,2$; and let $g^{(i)}(\underline{u}|\underline{\theta})$ be the common pdf. Define the random variables $U^{(i)}(\underline{\theta})$ as

$$ U^{(i)}(\underline{\theta}) = \text{Min}\,[{}_kD^{(i)}(\underline{\theta}) | k = 1,2,\ldots,M_i] \quad i = 1,2 \qquad 3.5.1.5 $$

For fixed i and $\underline{\theta}$ the random variable $U^{(i)}(\underline{\theta})$ is the first order statistic of the independent identically distributed random variables ${}_kD^{(i)}(\underline{\theta})$ $k = 1,2\ldots,M_i$. From the theory of order statistics the pdf for $U^{(i)}(\underline{\theta})$ is

$$ h^{(i)}(u|\underline{\theta}) = M_i[1 - G^{(i)}(u|\underline{\theta})]^{M_i-1}\, g^{(i)}(u|\theta) \quad i = 1,2. \qquad 3.5.1.6 $$

Assume now that the distribution F characterized by $\underline{\theta}$ originates from class 1. Then F is misclassified whenever $U^{(2)}(\underline{\theta}) < U^{(1)}(\underline{\theta})$, since then F is nearer to class 2 then class 1. Consequently, the average probability of

classifying F characterized by $\underline{\theta}$ into class 2, given class 1 is active is

$$P_2(\underline{\theta}) = P(U^{(2)}(\underline{\theta}) < U^{(1)}(\underline{\theta})) \qquad 3.5.1.7$$

$$= \int_{-\infty}^{\infty} \int_{-\infty}^{v} h(u,v|\underline{\theta})du\ dv$$

where $h(u,v|\underline{\theta})$ is the joint probability of $U^{(1)}(\underline{\theta})$ and $U^{(2)}(\underline{\theta})$. Now $U^{(1)}(\underline{\theta})$ and $U^{(2)}(\underline{\theta})$ are independent because they originate from independent random samples. Thus from 3.5.1.7

$$P_2(\underline{\theta}) = \int_{-\infty}^{\infty} h^{(2)}(v|\underline{\theta})d \int_{-\infty}^{v} h^{(1)}(u|\underline{\theta})du \qquad 3.5.1.8$$

where $h^{(1)}(u|\underline{\theta})$ and $h^{(2)}(v|\underline{\theta})$ are the marginal densities for $U^{(1)}(\underline{\theta})$ and $U^{(2)}(\underline{\theta})$ respectively as given by 3.5.1.6. Similarily

$$P_1(\underline{\theta}) = \int_{-\infty}^{\infty} h^{(1)}(u|\underline{\theta})du \int_{-\infty}^{u} h^{(2)}(v|\underline{\theta})dv \qquad 3.5.1.9$$

By substituting 3.5.1.6 in 3.5.1.8 and 3.5.1.9 $P_1(\underline{\theta})$ and $P_2(\underline{\theta})$ can be evaluated which via 3.5.1.3 and 3.5.1.2 yields P_E.

If parameter space symmetry exists such that $P_1(\underline{\theta}) = P_2(\underline{\theta})$ then regardless of the priors p_1 and p_2 from 3.5.1.2

$$P_E = P_2(\underline{\theta}) = P_1(\underline{\theta}) \qquad 3.5.1.10$$

for this case combining 3.5.1.6, 8, 3, and 2 yields

$$P_E = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \int_{-\infty}^{v} H^{(1)}(\underline{\theta})\{M_2[1 - G^{(2)}(v|\underline{\theta})]^{M_2-1} g^{(2)}(v|\underline{\theta})\}$$

$$\{M_1[1 - G^{(1)}(u|\underline{\theta})^{M_1-1} g^{(1)}(u|\underline{\theta})\ du\ dv\ d\underline{\theta} \qquad 3.5.1.11$$

A comment regarding the significance of 3.5.1.11 appears advisable. Note that to evaluate P_E the following distributions are required; the parameter space distribution, and the distribution of the first order statistics of the nearest neighbor to F (characterized by $\underline{\theta}$) for both class 1 and class 2. Provided it is reasonable to assume a parameter space distribution then in order to evaluate 3.5.1.11 all that is required are the appropriate first order statistics. Obtaining these statistics is, of course, not necessarily a trivial task.

3.5.2 Univariate Normal Case with Fixed and Equal Variances and Means Normally Distributed in the Parameter Space

In this case we assume that the i'th class (i = 1,2) contains an infinite number of univariate normal subclasses all with common variance σ^2, but whose means are distributed in the parameter space according to the normal distribution $h^{(i)}(\theta)$. That is the sets of states of nature $\Omega^{(i)}$ for the ith class are given by

$$\Omega^{(i)} = \{F|F\sim N(\mu,\sigma^2) \text{ where } \mu\sim N(m^{(i)},r^2)\}\ i = 1,2 \qquad 3.5.2.1$$

Note that this assumes that the parameter space densities are normal and that they differ only in location.

For a distance measure we use the Bhattacharyya distance. Recall that for the case under consideration (i.e., equal variances) the Divergence, Bhattacharyya distance, Kullback-Leibler numbers, and the Mahalanobis distance are all proportional. Our results, therefore, apply for any of these distance measures. For convenience we use the Bhattacharyya distance. If f, the pdf to be classified, has mean μ and variance σ^2 then

$$ {}_kD^{(1)}(\mu) = \frac{({}_k\mu^{(1)} - \mu)^2}{8\sigma^2} \qquad k = 1,2,\ldots,M_1 \qquad 3.5.2.2 $$

$$ {}_kD^{(2)}(\mu) = \frac{({}_k\mu^{(2)} - \mu)^2}{8\sigma^2} \qquad k = 1,2,\ldots,M_2 \qquad 3.5.2.3 $$

Where the ${}_k\mu^{(i)}$ are a random sample of size M_i from $h^{(i)}(\mu)$. Since ${}_k\mu^{(1)} \sim N(m^{(1)},r^2)$ it follows that for $k = 1,2,\ldots,M_1$

$\frac{8\sigma^2}{r^2}\,{}_kD^{(1)}(\mu) \sim NCX^2(1,(\frac{m^{(1)}-\mu^2}{r}))$ where $NCX^2(n,\beta^2)$ is the Non-central Chi-Square distribution with n degrees of freedom and noncentrality parameter β. The density for a $NCX^2(n,\beta^2)$ distribution is given by

$$ f(x) = e^{-\frac{1}{2}\beta^2} \sum_{k=0}^{\infty} \frac{1}{k!}(\tfrac{1}{2}\beta^2)^k \frac{(\frac{1}{2}x)^{\frac{1}{2}(n+2k-2)}\, e^{-\frac{1}{2}x}}{2\Gamma(\frac{1}{2}(n+2k))} \qquad 3.5.2.4 $$

where $\Gamma(\cdot)$ is the Gamma function. The corresponding cdf is,

$$ F(x) = e^{-\frac{1}{2}\beta^2} \sum_{k=0}^{\infty} \frac{1}{k!} (\tfrac{1}{2}\beta^2)^k \frac{\gamma(\frac{1}{2}(n+2k), \frac{1}{2}x)}{\Gamma(\frac{1}{2}(n+2k))} \qquad 3.5.2.5 $$

where $\gamma(\cdot, \cdot)$ is the incomplete Gamma function defined by

$$\gamma(a,x) = \sum_{n=0}^{\infty} \frac{(-1)^n x^{a+n}}{n!(a+n)} \qquad 3.5.2.6$$

Similarly

$$\frac{8\sigma^2}{r^2}\ {}_kD^{(2)}(\mu) \sim NCX^2\left(1, \left(\frac{m^{(2)} - \mu}{r}\right)^2\right)$$

Since parameter space symmetry exists such that $P_2(\underline{\theta}) = P_1(\underline{\theta})$ the average probability of error is given by 3.5.1.11 with

$\underline{\theta} = \mu$ and

$$g^{(i)}(u|\mu) = 2\lambda \exp(-\beta_1^2) \sum_{k=0}^{\infty} \frac{1}{k!} \beta_i^{2k} (\lambda u)^{k-\frac{1}{2}} \frac{\exp(\lambda u)}{2\Gamma(k+\frac{1}{2})}$$

$$i = 1,2, \qquad 3.5.2.7$$

$$G^{(i)}(u|\mu) = \exp(-\beta_i^2) \sum_{k=0}^{\infty} \frac{1}{k!} \beta_i^{2k} \frac{\gamma(k+\frac{1}{2}, \lambda u)}{\Gamma(k+\frac{1}{2})} \qquad i = 1,2 \qquad 3.5.2.8$$

$$h^{(1)}(\mu) = \frac{1}{\sqrt{2\pi}r} \exp\left(-\frac{1}{2}\left(\frac{m^{(1)} - \mu}{r}\right)^2\right), \qquad 3.5.2.9$$

where in 3.5.2.7 and 3.5.2.8

$$\beta_i^2 = \frac{1}{2}\left(\frac{m^{(i)} - \mu}{r}\right)^2 \quad i = 1,2 \quad \text{and} \quad \lambda = \frac{4\sigma^2}{r^2} \qquad 3.5.2.10$$

The above constitutes a complete theoretical solution for the case of means normally distributed in the parameter space. It is rather apparent that the practical evaluation of P_E for this case is by no means a trivial task. While it is certainly possible to evaluate P_E numerically it appears likely that other assumptions regarding the parameter

space distribution might yield simpler and just as meaningful results. Consequently in the next section the normal assumption for the parameter space distribution is abandoned in favor of means uniformly distributed in the parameter space. The theoretical results of this section were included to facilitate further investigation of normally distributed means should this prove desirable.

3.5.3 Univariate Normal Case with Fixed and Equal Variances and Means Uniformly Distributed in the Parameter Space

In this case the sets of states of nature are

$$\Omega^{(i)} = \{F \mid F \sim N(\mu,\sigma^2) \text{ where } \mu \sim U(a_i,b_i)\} \quad i = 1,2 \qquad 3.5.3.1$$

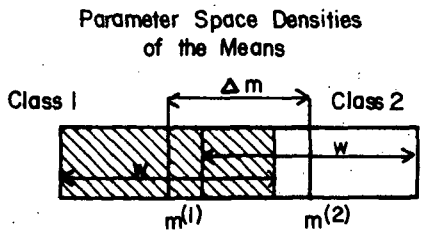
In addition to assuming that the distribution of the means for class 1 and class 2 are uniform it is also assumed that $U(a_1, b_1)$ and $U(a_2,b_2)$ differ only in location. That is, it is assumed that

$$a_1 - b_1 = a_2 - b_2 = w.$$

Assume also that $a_2 \geq a_1$. The case where a single distribution is selected to represent each class ($M_1 = M_2 = 1$) is considered first and the average probability of error as a function of the overlap of the parameter space densities determined. If $m^{(i)}$ is the mean of $h^{(i)}(\mu)$ (i.e., $U(a_i,b_i)$) $i = 1,2$ then define the normalized overlap γ as

$$\gamma = \frac{m^{(2)} - m^{(1)}}{w} = \frac{\Delta m}{w} \qquad 3.5.3.3$$

Fig. 3.5.3.1 depicts the situation.



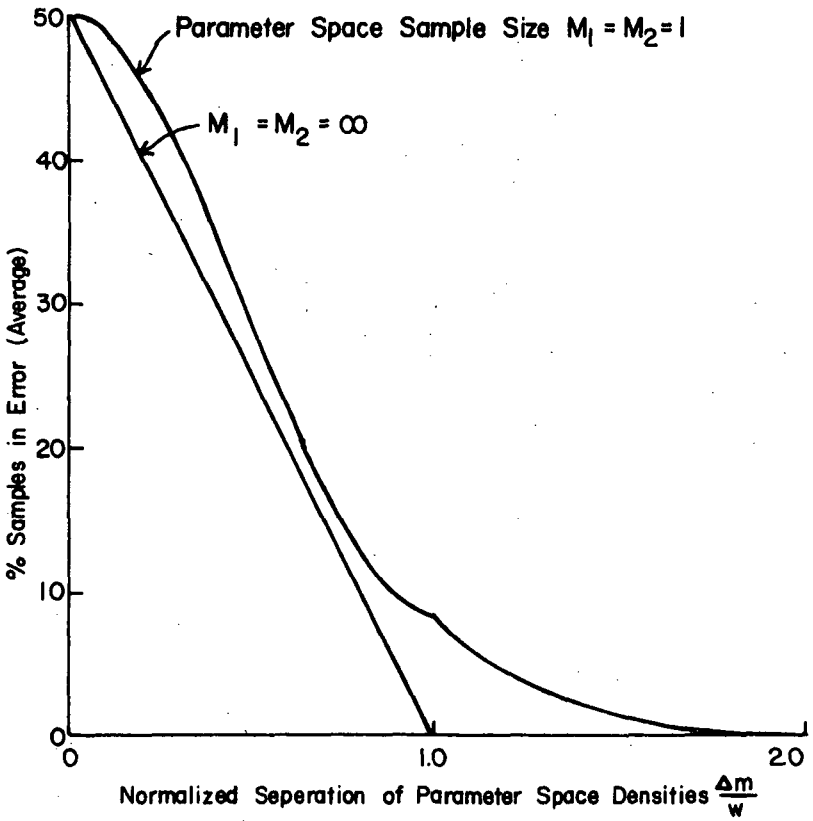


Figure 3.5.3.1 Average Classifier Error for Minimum Distance Classification. A Simple Normal Example.

The distance measure used is

$$ {}_kD^{(i)}(\mu) = |{}_k\mu^{(i)} - \mu| \qquad k = 1;\ i = 1,2 \qquad 3.5.3.4 $$

This distance measure is used because for the case under consideration it gives the same performance as the Bhattacharayya distance, or other distances proportional to the Bhattacharyya distance, but is somewhat simpler theoretically. The symmetry in the parameter space is again such that $P_2(\underline{\theta}) = P_1(\underline{\theta})$. Consequently setting $M_1 = M_2 = 1$ and $\underline{\theta} = \mu$ 3.5.1.11 reduces to

$$ P_E = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{v} h^{(1)}(\mu)g^{(2)}(v|\mu)g^{(1)}(u|\mu)du\ dv\ d\mu \qquad 3.5.3.5 $$

The densities $g^{(1)}$ and $g^{(2)}$ can be obtained by inspection. For example if $a_1 < \mu < \frac{1}{2}(a_1+b_1)$ then

$$ g^{(1)}(u|\mu) = \frac{2}{w} \qquad 0 < u \leq (\mu - a_1) \qquad 3.5.3.6 $$

$$ = \frac{1}{w} \qquad (\mu - a_1) < u < (b_1 - \mu) $$

Similarly $g^{(2)}$ can be readily obtained.

It is therefore a straightforward but time consuming task to evaluate 3.5.3.5. Particular care must be exercised to ensure that all discontinuities in $g^{(1)}$, $g^{(2)}$ and $h^{(1)}$ are properly handled. Carrying out the necessary computations the following results are obtained.

$$P_E(\gamma) = \frac{1}{12}(\gamma^2(10\gamma-15)+6) \qquad 0 \leq \gamma \leq 1 \qquad 3.5.3.7$$

$$= \frac{1}{12}(2-\gamma)^3 \qquad 1 \leq \gamma \leq 2$$

$$= 0 \qquad \gamma \geq 2$$

This equation is plotted in Figure 3.5.3.1.

In Fig. 3.5.3.1 we have also plotted the expected probability of error when each class is represented by a particular infinite set of distributions ($M_1 = M_2 = \infty$ curve). More specifically the set of distributions used to represent each class is all the possible distributions in that class. In this case it is easy to determine the average probability of error since only samples whose mean falls in the region where the parameter space densities overlap can be incorrectly classified. Any sample whose mean falls outside the region of overlap is correctly classified since it is some finite distance away from the incorrect class, and a distance of zero from the correct class. In the region of overlap the distance to the set of distributions representing each class is zero. We assume that these ties are broken in accordance with the relative probability of observing the given parameter value for each class. For the case under consideration assuming equal priors, half of the samples that fall in the overlap region will be incorrectly classified. Consequently we have immediately for infinite sample size:

$$P_E(\gamma) = \frac{1}{2}(1-\gamma) \qquad 0 \leq \gamma \leq 1 \qquad 3.5.3.8$$

$$= 0 \qquad \gamma \geq 1$$

The largest and smallest probability of error that can result when each class is represented by a single distribution is also of interest. These probabilities are easily obtained. For the case under consideration the minimum distance rule partitions the real axis into two parts. The partition point μ_M is given by

$$\mu_M = \frac{1}{2}({}_1\mu^{(1)} + {}_1\mu^{(2)}) \qquad 3.5.3.9$$

Unknown samples whose mean μ lies on the same side of μ_M as ${}_1\mu^{(i)}$ are assigned to class i, i = 1,2.
The values over which the partition point μ_M can range is

$$\frac{1}{2}(a_1+a_2) \leq \mu_M \leq \frac{1}{2}(b_1+b_2) \qquad 3.5.3.10$$

To determine the best and worst case for a given situation it is only necessary to examine all possible partitions in the permissable range and choose the best and the worst. Account must also be taken of the fact that if the parameter space densities overlap, then for partitions which fall in the range of overlap, ${}_1\mu^{(i)}$ i = 1,2 can lie on either side of the partition. For example the "minimum" and "maximum" insets in Fig. 3.5.3.2 shows both a "best" and a "worst" situation respectively for a given degree of overlap of the parameter space densities. Note that the "best" and "worst"

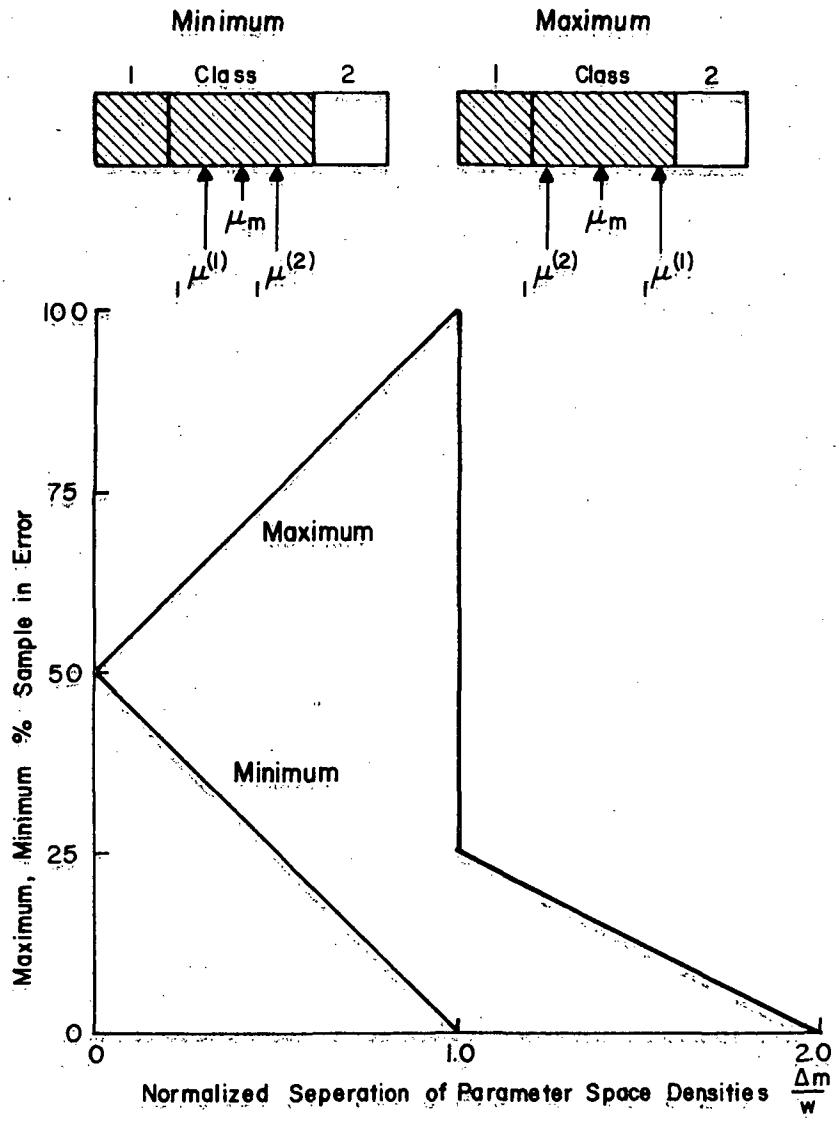


Figure 3.5.3.2 Minimum and Maximum Classifier Error for Minimum Distance Classification. A Simple Normal Example.

cases are not unique. In fact any training set which results in a partition point that falls in the region of overlap is either a "best" or "worst" case depending upon which of the situations depicted in the insets Fig. 3.5.3.2 pertains.

Proceeding in this manner it is easy to show that

$$\text{Min}(P_E(\gamma)) = \frac{1}{2}(1-\gamma) \qquad 0 \leq \gamma \leq 1 \qquad 3.5.3.11$$

$$= 0 \qquad \geq 1$$

and

$$\text{Max}\ (P_E(\gamma)) = \frac{1}{2}(1+\gamma) \qquad 0 \leq \gamma \leq 1 \qquad 3.5.3.12$$

$$= \frac{1}{4}(2-\gamma) \qquad 1 \leq \gamma \leq 2$$

$$= 0 \qquad \geq 2$$

These curves are plotted in Fig. 3.5.3.2. Note the abrupt drop in the maximum probability of error at $\gamma = 1$. This drop occurs since for $\gamma \geq 1$ it is no longer possible for the means of the training samples to fall on the "wrong" side of the partition μ_M.

The "best" and "worst" case curves shown in Fig. 3.5.3.2 have been derived on the basis that each class is represented by one distribution. A moments consideration shows that they are also valid if each class is represented by an infinite (even uncountably infinite) set of distributions. This follows since it is always possible that the means of every distribution chosen to represent class 1 falls below (above) the mean of every distribution

chosen to represent class 2 leading to the "best" ("worst") case curves depicted in Fig. 3.5.3.2. The likelihood of observing the best or worst cases of course decreases as the number of samples selected to represent each class increases.

A number of important factors emerge from the simple example considered. For convenience in referring to these factors in later sections they will be given a reference number.

Observation 1

If the parameter space densities overlap it is possible for the minimum distance method to perform very poorly.

Observation 2

The maximum, minimum and average performance for the case where each class is represented by all the densities in that case are identical. This follows since in this case the training distributions are always the same.

Observation 3

The average (which by virtue of observation 2 is also the "best") performance for the case where each class is represented by all the densities in that class, is only moderately better than the average performance achieved when each class is represented by a single density. This suggests that the very poor performance mentioned in observation 1 occurs rather infrequently. More importantly

it also suggests that in terms of average performance very little is gained by using many subclasses. What is gained by using many subclasses is a significant reduction in the probability of choosing a very poor training set, rather than a significant decrease in the average performance.

Observation 4

It is relatively easy to imagine situations where the overall performance (i.e. the overall probability of correctly classifying a unknown sample) changes drastically in either direction as the number of subclasses used to represent each class increases. For example consider increasing the number of distributions used to represent each class from 1 to 2. Let the minimum probability of error inset in Fig. 3.5.3.2 depict the situation when each class is represented by a single density. Let the densities used to represent each class in the maximum probability of error inset be the set of densities added to increase to 2 the number of distributions representing each class. It is obvious for this case that an increase in the number of subclasses causes a drastic decrease in overall performance. The situation described is a rather unlikely situation and changes would typically be much smaller, particularly for cases where each class is represented by a moderate number of distributions.

It is also easy to depict situations for which the performance by class (as opposed to overall performance)

changes drastically in either direction for one or both classes as the number of subclasses is increased. In fact drastic changes in class performance would appear to be more likely to occur than drastic changes in overall performance.

Observation 5

The discontinuity of the slope of the average probability of error curve in Fig. 3.5.3.1 for the $M_1 = M_2 = 1$ case at $\gamma = 1$ is due to the discontinuous behavior of the maximum probability of error in Fig. 3.5.3.2 at $\gamma = 1$.

It is necessary to remember that observations 1 to 5 pertain specifically to the particular case investigated. It is impossible to tell to what extent these observations carry over to more complex situations. The manner in which 1 to 4 occur means they will almost certainly have their counterpart in multiclass multidimensional problems.

CHAPTER 4

EXPERIMENTAL RESULTS

In this Chapter the experimental results obtained in the investigation of minimum distance classification and related problems are presented. To facilitate the description of the experiments performed it is desirable to devise a systematic method of describing an experiment. Not only does this simplify the description of an experiment but it also aids in clearly indicating the quantities that remain fixed throughout the experiment and those that are variable. In general we use the classification accuracy (or performance) in evaluating different procedures, distance measures, etc. For our purpose it is convenient to consider the performance to be a function of the three quantities listed at the top of Table 4.1; these are, the Training Procedure, Classifier Type, and Classifier Parameters. At present there is no need to be intimately concerned with the detailed breakdown of these three categories; it is sufficient to note that to describe an experiment it is only necessary to describe the three factors influencing performance.

Table 4.1 is not intended to be a comprehensive enumeration of all classifier possibilities, nor is it

Table 4.1 Method of Describing Experimental Problems

PERFORMANCE
TRAINING PROCEDURE
CLASSIFIER TYPE
CLASSIFIER PARAMETERS

TRAINING PROCEDURE	CLASSIFIER TYPE	CLASSIFIER PARAMETERS
Training Field Selection -Percent Training Acres by Class -Percent Standard Test Fields by Class	Vector (Maximum Likelihood) - Parametric(LARSYSSA)	LARSYSSA -Number of channels
Subclass Definition -Random -Every training field a subclass -All training fields combined (No subclasses) -Nonrandom -Parameter Space Clustering (GRPSAM) -Observation Space Clustering (NSCLAS)	Sample (Minimum Distance) - Parametric (PERFIELD) - Nonparametric (LARSYSDC)	PERFIELD -Distance measure -Number of channels -Number of vectors LARSYSDC -Distance measure -Number of channels -Number of vectors -Bin size
Feature Selection -Based on Average Divergence with Exhaustive Search of Feature Combination ($DIVG) -Based on Average Transformed Divergence with Sequential Search of Feature Combinations ($SEQDIVG)		

necessarily a method that is capable of describing all classifier problems. In fact only those Training Procedures, Classifier Types, and Classifier Parameters that are of direct concern in this work are listed. The sole purpose of the table is to facilitate description of the particular experiments performed. We will frequently refer to this table to assist in describing the organization of our work.

Classifiers are usually segregated into two broad categories, supervised and nonsupervised. A supervised classifier is characterized by the fact that it utilizes data of known classification as a basis for classifying unknown data. In particular before classification starts typical data for every class of interest is made available to the classifer. Such data is known as training data. In a nonsupervised classifier data may also be available to the classifier before classification commences, but the classification of this data is not known to the classifier.

Only supervised classifiers are used in this investigation. In such classifiers the process of extracting the information from the training data for subsequent use in the classification task is referred to as "training the classifier". Once the classifier has been trained it can be used to classify other data drawn from the classes for which it was trained. Such data is referred to as test data and the classification accuracy on such data is the test performance. It is, of course, also possible to classify the

training data itself. In this case the resultant correct classification is known as training performance.

For most experiments the performance is determined for both training and test data. The interpretation of results for training data is usually easier since the question of whether the training data was typical of the test data does not arise. In the final analysis, however, it is the performance on test data that is important.

Although the detailed subdivisions of Table 4.1 hint at the complexity of the classification problem for multispectral data-images a few additional comments seem appropriate. Even if the training procedure is entirely ignored the problems are still substantial. The number of main classes of interest can range up to 10 or more while the number of subclasses may be three or four times this number. The number of channels typically available is 13; a number that will undoubtedly increase in the future. While it is generally true that in the classification procedure itself very few classifications use all the available channels, it is equally true that the use of only one channel is very rare. Consequently, considering only the classifier (i.e., ignoring training) itself, the problem is still a multiclass, multidimensional problem, and very difficult to handle theoretically. Introduce the added complexities of different Training Procedures, various Classifier Parameters and also the difficulty in

establishing a mathematical model for multispectral data and it is clear that the best approach is an experimental approach.

The chapter commences with a description of the data used, and a discussion of the programs used to analyse the data. Some of the analysis programs were specifically written to carry out the experiments described, others were already available. One of the prime investigations concerns itself with the relative performance of different distance measures and how the number of subclasses affects performance. In situations where the desirable number of subclasses becomes impractically large, some method must be devised for combining subclasses that are most similar. Parameter space clustering is used as a method to achieve this goal for parametric problems. Since clustering in the parameter space is far from routine, considerable space is devoted to its evaluation, including its use in more convential vector by vector classifiers. Finally the effect of various parameters on performance is considered.

There is a certain experimental philosophy which pervades this work which should be clarified at the outset The philosophy is one of comparison. No real systematic attempt is made to adjust all pertinent variables in order to attain "the best" classification. Rather the philosophy is one of trying to establish which of several alternate procedures is most likely to yield the better classification,

without expending the time and energy required to greatly refine any of the classifications. Thus for example there is very little manipulation, purification, etc. of training sets to achieve the best possible classification. In short the emphasis is on relative performance under controlled conditions rather than absolute performance. The justification for this philosophy is that the scheme which provides the best relative performance should in the final analysis also provide the best absolute performance.

4.1 Description of the Experimental Data

In Chapter 1 we pointed out that we are concerned primarily with the classification of multispectral data-imagery. It is, therefore, natural to restrict the experimental investigation to such data. It is worthwhile to again emphasize that the techniques utilized are not restricted in this manner, although experimental conclusions must, of course, be interpreted in terms of the data on which the conclusions are based. Most of the multispectral data-imagery available at Purdue's Laboratory for Applications of Remote Sensing has been collected by an instrument known as a multispectral scanner.[62] We refer to such imagery as multispectral scanner imagery or multispectral scanner data. There is also a small amount of multispectral data-imagery that has been generated by digitizing photographs. Although for the purpose of the work herein there is no essential difference between the scanner and digitized

photographic data we shall only be concerned with the former.

A brief description of LARS multispectral scanner imagery and the scanner collection system appears pertinent. To obtain multispectral scanner imagery for a particular scene, the multispectral scanner is carried above the scene in question on an aerospace platform (presently an aircraft). The scanner is capable of simultaneously recording, on magnetic tape in analog form, the image of the scene below as seen through different spectral "windows". The manner in which this is achieved is briefly described. For each spectral band the electromagnetic radiation from an area on the ground is collected by an optical system in the scanner and focused onto a detector. The detector generates an electrical output which depends upon the radiation intensity in that wavelength band, and which after appropriate electronic processing is suitable for recording purposes. The area from which electromagnetic radiation is being collected is swept across the flight path of the aircraft by a rotating mirror arrangement in the scanner. At the same time the scanner is carried along the flight path by the forward motion of the aircraft. The combined motion results in a raster scan of the scene below. The scan lines generated in this manner are recorded on analog tape. Subsequent digitization results in a two dimensional array of measurement vectors in which the components of the vectors

correspond to the radiation intensity in the various spectral bands. After some processing the two dimensional array of measurement vectors is stored on a digital tape referred to as an Aircraft Data Storage Tape, which for our purposes constitutes the raw data. The area associated with the measurement vector will be referred to as an Image Resolution Element (IRE). Strictly speaking the spatial coordinates, or relative spatial coordinates designating the location of each IRE, could also be considered to be part of the measurement vector. However, since the coordinates are of a different nature than the spectral measurements their usage is different. In fact the spatial coordinates in the form of line and column numbers are used to reference the location of the measurement vectors on the Aircraft Data Storage Tape.

In selecting the particular multispectral scanner data to be utilized for the experimental investigation several factors were considered. By far the most important factor was that the data should be difficult to analyse. That is the data should contain some main classes that are difficult to seperate. It would be pointless to carry out an extensive investigation on data that is easily segregated into the classes of interest, since then apparently any advantage of minimum distance classification would be obscured. A second factor of considerable importance was that the data set should be of adequate size to provide

a realistic experimental test of the various procedures considered. A third factor that was considered was whether or not the data had previously been analysed by conventional techniques. Such analysis would enable a comparison of conventional and minimum distance techniques with a minimum of effort. To be most useful the conventional analysis should involve a relatively small number of main classes. The reason for this is that program restrictions of some existing analysis programs are such that the large number of subclasses anticipated for minimum distnace classification could only be accommodated if the number of main classes was relatively small.

The practice of utilizing existing programs whenever possible, in order to minimize the programming effort is logical and reasonable, as long as this does not place unrealistic restrictions on the experiments. Since many practical classifications do not require a large number of main classes focusing attention on such classifications was judged to be a reasonable restriction. An advantageous side effect of restricting the number of main classes is that results are somewhat simpler to interpret and much easier to report.

A final factor considered in selecting the multispectral scanner data to be examined experimentally was the desireability of having available several data sets that were similar, so that meaningful averages could be taken

over the data sets.

In light of the requirements outlined in the previous paragraphs the multispectral scanner data sets chosen for the experimental investigation were runs 70002200, 70002300 and 70002400. The data for these runs was collected* at an altitude of 3000 ft., between 9:45 and 10:45 a.m. E.D.T., on June 30, 1970, from flightlines 21, 23 and 24 respectively. The exact location and orientation of these flightlines, which are located in Tippecanoe County, Indiana, is shown in Fig. 4.1.1. The flightlines extend the 24 mile length from the north to the south end of the county and are roughly equally spaced in the east-west direction. Since the scanner geometry is such that at an altitude of 3000 feet the field of view is roughly 1 mile, the area covered by the three flightlines, approximately 72 square miles, is about 1/7 of the total area in the county. The scanner resolution and sampling rate are nominally three and six milliradians respectively. This means that at nadir the scanner "sees" a circle about 9 feet in diameter and that the spacing between adjacent IRE's is about 18 feet. Since the scanner resolution and sampling rate are independent of look angle the distance between adjacent IRE's is approximately 30% larger at the edge of the scanner's field of view with a corresponding change in the shape and area "seen" by the scanner. At the sampling rate indicated there are 220 samples across the width of a flightline and each flightline contains 5000 to 6000 lines. This

*Data collected by University of Michigan Scanner.

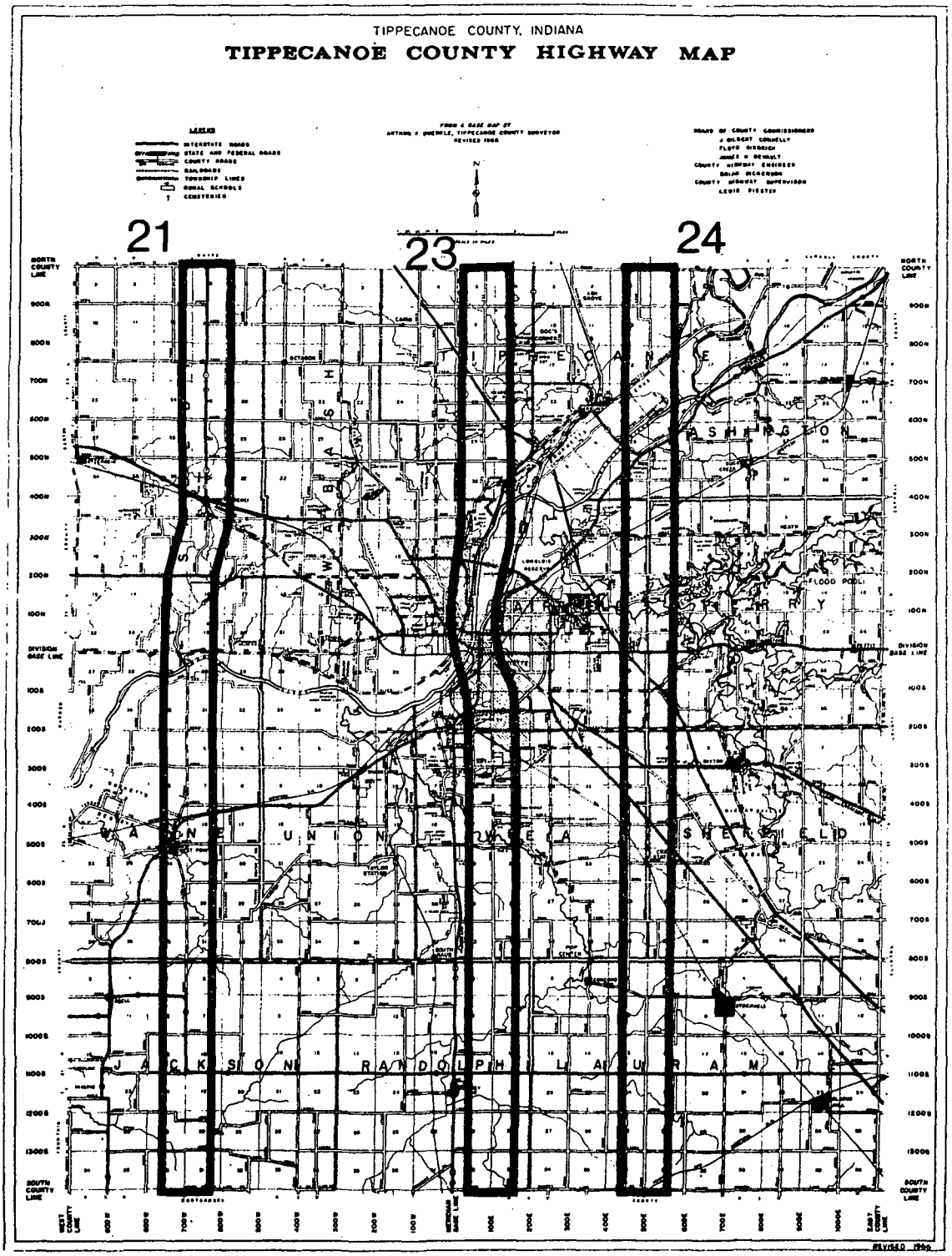


Figure 4.1.1 Location of Tippecanoe County Flightlines 21, 23, and 24.

means each flightline contains somewhat more than 10^6 IRE's.

The data from the flightlines selected met all the requirements stated above. A conventional analysis of this data had been carried out in connection with a crop yield study. In the yield study the main classes considered were wheat, corn, soybeans and other. Furthermore this analysis indicated that the corn and soybeans were not very separable, a situation that typifies data collected at this time of year.

Thirteen spectral bands of data were collected for each of the three runs being discussed. It is frequently convenient to refer to these spectral bands by channel number rather than specifically stating the wavelength bands involved. The correspondence between channel numbers and spectral bands is given in Table 4.1.1.

Of the approximately 10^6 IRE's in each flightline between 10% and 20% are typically used as test fields. There are a number of sets of test and training fields which are repeatedly used throughout the experiment. These are described in Appendix C which also contains the coordinates of the various fields. For continuity of the discussion it is adequate to recognize that the following decks are described: (1) Standard Test Field decks for flightlines 21, 23 and 24; these fields are used primarily for test purposes; (2) a field deck of Training Acres used primarily for training purposes, both in this study and the

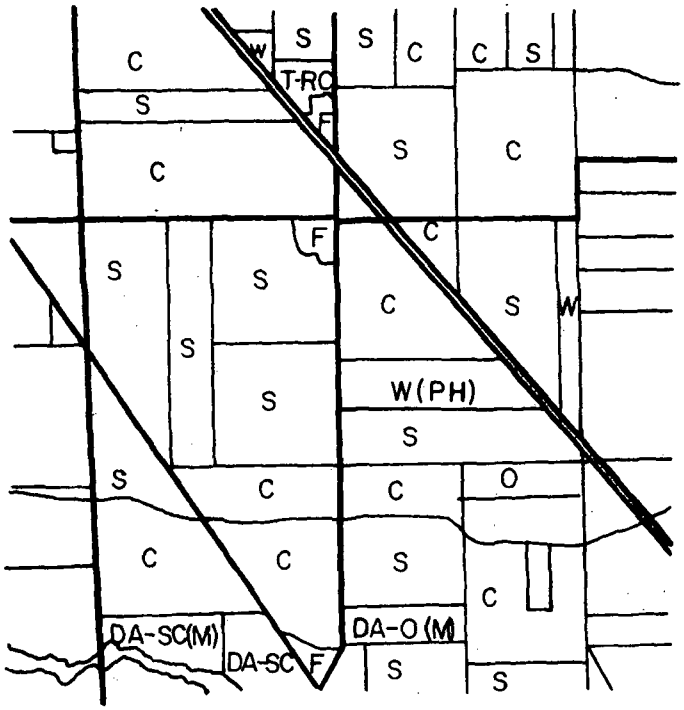
Table 4.1.1

Correspondence Between Channel Numbers and Spectral Bands

Channel Number	Spectral Band (Micrometers)
1	0.40 - 0.44
2	0.46 - 0.48
3	0.50 - 0.52
4	0.52 - 0.55
5	0.55 - 0.58
6	0.58 - 0.62
7	0.62 - 0.66
8	0.66 - 0.72
9	0.72 - 0.80
10	0.80 - 1.00
11	1.00 - 1.40
12	1.50 - 1.80
13	2.00 - 2.60

crop yield study; (3) a field deck of Flightline 21 Test Areas which are subareas within the Standard Test Fields for flightline 21 and are used as test fields.

A few comments regarding the type and extent of the ground cover at the time of the flights appear advisable. As already mentioned four principle ground cover categories are considered; wheat, corn, soybeans and other. Although the class other includes a considerable variety of ground cover most of the agricultural fields in this category are either small grains (other than wheat) or forage crops. There are also some bare soil and a number of diverted acre fields. Some natural categories such as trees and water are also included in this class. For most of the subcategories for the class other the ground cover is fairly complete, but the spectral properties of the ground cover are quite variable from field to field within a subcategory. Most of the wheat in the flightlines was mature and ready, or nearly ready, for harvest. In fact some portion of it had already been harvested. For corn and soybeans the crop canopy at flight time was such that a considerable fraction of the soil was not covered by vegetation when viewed from above. Some idea regarding the extent of the ground cover can be obtained from the color and color infra-red photographs shown in Fig. 4.1.3. Fig. 4.1.2 indicates the ground cover for the various fields. These photographs show a typical section of flightline 24 as it appeared on the day



KEY			
C	Corn	S	Soybeans
W	Wheat	RC	Red Clover
O	Oats	SC	Sweet Clover
T	Timothy	(M)	Mowed
F	Farm	(PH)	Part Harvested

Figure 4.1.2 "Ground Truth" for Figure 4.1.3.

Original in
Color

Original in
Color Infrared

Fig. 4.1.3. Color and Color Infrared Photographs of Part of Flightline 24.

of the flight. While the color photograph gives some indication of ground cover a much better indication can be obtained from the color infra-red photograph because of its property of portraying healthy green vegetation as bright red. Even the slightest amount of green vegetation is sufficient to give a reddish hue to a field. This point is adequately demonstrated by most of the soybean fields in Fig. 4.1.3. The green vegetation is barely observable on the color photo but shows up much better on the color IR. The ground cover for most corn fields in the area shown is considerably greater than for most soybean fields, however, there are exceptions. Notice the variability of the fields within one crop type even over the small region covered by the photographs. The difference between harvested and unharvested wheat is also of importance. Finally the fact that ground patterns show up quite distinctly in corn and soybean fields provides further evidence of the sparce ground cover in these fields.

4.2 Data Analysis Programs

A number of different programs were used in the analysis of the scanner data. The purpose of this section is to describe these programs. Some of the programs are analysis programs that are in general use at LARS and will be referred to collectively as LARS System Programs. Other programs were written specifically to investigate minimum distance classification and related problems.

The description given for each analysis program is a brief functional description. These brief descriptions are augmented by appropriate references for the LARS System Programs and by Appendix E for those programs written specifically to investigate minimum distance classification and related problems. While the brief functional descriptions are adequate for our purpose, the full capabilities of the programs can only be appreciated by examining the supplementary material.

There is a general philosophy that pervades LARS System Programs that can best be summed up by stating they are user oriented. A basic assumption is that the user should not be required to be very knowledgeable about computers or programming in order to use any of the LARS System Programs. This goal is in effect achieved by designing for each program what in essence is really a very simple language. The user selects program options and specifies program parameters by means of "control cards" written in this simple language. The principles of the language are very simple and remain fixed from program to program. Consequently it is very easy for the user to learn the language. In fact if the user has a reasonable understanding of the program's function, then the control cards seem to him to be a very natural and easy way of specifying the program options. For example a control card (whose location in the control card deck is arbitrary) containing CHANNEL

1, 2, 7 might mean that spectral channels 1, 2, and 7 are to be used in the program. Contrast this with the conventional situation where it would be necessary to remember the location of the channels card in the data deck as well as its format. A peripheral advantage of this approach is that program documentation tends to be simpler, since to describe the capabilities of a program it is only necessary to describe the function of each control card. Appendix D contains a brief description of the control card language. This description is included so that the "control card descriptions" of the programs in Appendix E can be understood.

Another aspect of the user orientation is that programs tend to be self documenting during execution. In other words sufficient information regarding program options, program parameters, etc., are listed on the printer, which together with a user supplied comment, enables the user to determine exactly what computations were carried out.

A final aspect of LARS System Programs, which is of importance to programmers rather than users, is that the program decks contain a sufficient number of comments to be substantially self-documenting.

The reason for dwelling on the philosophy of the LARS System Programs is that one is faced with the problem of whether or not this philosophy should be adopted for a research program. It is clear that to adopt such a

philosophy requires considerable additional programming, even though general purpose control card interpreting routines exist which lighten the programming burden somewhat. The biggest advantage in adopting this philosophy is that if the program proves to be of interest to a number of users it can be made available to them very quickly, and within a familiar framework. Another advantage is of course that the programs are also much easier to use during the research phase. The sole disadvantage is the additional programming time required.

Some of the programs specifically written for this investigation were written with the same philosophy as that underlying the LARS System Programs, except that the use of comments in the programs was not as consistent or liberal. On the other hand some programs were written without much regard to user convenience. On the basis of this experience it is our feeling that for research programs the user oriented approach is worthwhile provided there is a good possibility that a number of users will be interested in the program; or provided that during the research phase it is anticipated that the program will be used many times. If neither of these conditions is satisfied the additional programming effort is simply not justified.

4.2.1 LARSYSAA: A Parametric (normal) Maximum Likelihood Vector Classifier

The primary classification system presently used at LARS for classifying multispectral data-imagery is known

by the acronym LARSYSAA.[1,63,64] This is a supervised system in which it is assumed that the data for each class is drawn from a multivariate normal population, and classification of the unknown vectors is affected according to the maximum likelihood principle[65] on a vector by vector basis. The system is supervised[65] since samples (i.e., sets of measurement vectors) whose classification are known are used to train the classifier. Because of the Gaussian assumption, training simply amounts to utilizing the samples whose classification are known to estimate the mean vector and covariance matrix for each class. These estimated quantities are then used to compute the likelihood function upon which the classification decisions are based. Facilities exist in the system for selecting a good subset of the original spectral bands upon which to base the classification.[33] Such techniques are usually referred to as feature selection techniques. The particular feature selection technique used in LARSYSAA is based on Divergence or an exponentially saturating transformation of the Divergence.[66] The average transformed Divergence between all class pairs, or the average Divergence between all class pairs, is used as a measure of feature effectiveness. The capability to use the average transformed Divergence rather than just the average Divergence has only recently become available but at present it is the standard option unless the average Divergence is specifically requested.

LARSYSAA is organized into four processors. A statistics processor ($STAT), a feature selection processor ($DIVG), a classification processor ($CLASS) and a display processor ($DISP). The purpose of the statistics processor is to compute, list, store, and punch first and second order class statistics. It can also display histograms and spectral plots on the printer. Wherever approriate these operations can be carried out on either a class or field basis. The feature selection processor enables the "best" subset of features to be selected for a given set of classes. The classification processor classifies the vectors in a specified area in accordance with the maximum likelihood rule. The class to which every vector in the specified area is assigned together with the value of the likelihood function, is stored on a magnetic tape referred to as a Map Tape. Finally the display processor enables the classification to be displayed in map form on the line printer, and computes and lists performance tables. Except for the divergence processor the program is capable of accomodating up to 60 classes and up to 30 channels; although not necessarily simultaneously. The divergence processor, which is temporarily a stand alone program, can accomodate up to 30 classes and 18 channels.

The $DIVG processor in LARSYSAA requires a few additional comments. This processor is an optimum feature selection processor in the sense that it carries out a

comprehensive search of all feature combinations. Under certain circumstances the number of combinations becomes quite large and the processing time becomes exorbitant. This is for example the situation that prevails if the best k_b out of k_c channels are to be chosen and k_c is in the vicinity of 13 and k_b in the vicinity of 7. To alleviate this problem a modified suboptimum form of $DIVG, which we refer to as $SEQDIVG, was programmed. The $SEQDIVG processor differs from $DIVG only in that no comprehensive search of all feature combinations is performed, and in this sense it is suboptimum. The search procedure used is that features are added sequentially, one at a time, in such a manner that the addition of the next feature results in the greatest possible increase in the separability criterion. As in regular $DIVG the separability criterion is either the average transformed Divergence or average Divergence.

4.2.2 PERFIELD: A Parametric (normal) Minimum Distance Classifier

PERFIELD is a parametric minimum distance classifier based on the Jeffreys-Matusita distance*. Huang[67] did the initial work at LARS which led to the programming of this classifier. A statistics deck generated by the $STAT processor of LARSYSAA is used to define the classes for PERFIELD. Samples are classified one at a time.

*Strictly speaking PERFIELD is based on the Bhattacharyya distance but since the Bhattacharyya and JM distance produce identical classifications we consistently refer to the later since it is more convenient for our purpose.

They are defined by specifying a run number and the coordinates (i.e., line and column numbers) of a rectangular field in that run*. The vectors within the field constitute the sample to be classified. The classification is accomplished by retrieving the pertinent data from an Aircraft Data Storage Tape and carrying out the necessary computations. Details of the classification and performance tables are listed on the line printer. Since the completion of our experimental work PERFIELD has been added to LARSYSAA as a fifth processor.

In order to be able to perform minimum distance classifications for distances other than the JM distance, two modified versions of PERFIELD were programmed. The first used Divergence as the distance measure and the second used Kullback-Leibler numbers. Although there are really three distinct programs involved, it is convenient to treat them as a single program PERFIELD in which the distance measure is a program option.

---

*In Chapter 1 it was mentioned that a problem closely related to minimum distance classification is the problem of defining samples to be classified. It was also pointed out that one way of defining samples was through the use of closed boundaries. To implement such a technique it is highly desirable that the boundaries be located by computer on the basis of the spectral data. BOUND is a program that in part attains this goal in that it locates boundaries in multispectral scanner data. However, the boundaries are in general not closed and further development is needed before the method could be used to define samples for minimum distance classification. Appendix F contains a brief functional description of BOUND as well as pertinent references.

4.2.3. NSCLAS: An Observation Space Clustering Program

The purpose of NSCLAS is to group together, in the observation space, vectors which are similar. The measure of similarity used is Euclidean distance. In principle NSCLAS is similar to the ISODATA method of Ball and Hall.[68] The exact details of the clustering proceedure used in NSCLAS are identical to those of the clustering algorithm used by Wacker and Landgrebe to locate field boundaries.[69] Details about various clustering schemes can be found in the review papers by Ball[70] and Rolhf[71].

In essence NSCLAS provides the user with the capability of "classifying" a limited number of IRE's on a nonsupervised basis. It is a nonsupervised classification in that no training is involved. The user must identify the classes after clustering is completed.

To cluster a set of vectors the user designates the desired vectors by means of a deck of field coordinate cards. Vectors from the specified rectangular areas are read from Aircraft Data Storage Tapes and clustered into the number of classes specified by the user. Actually there is a rudimentary search procedure in NSCLAS, which at the users option attempts to establish the appropriate number of classes. In practice this procedure has not worked well for multispectral data-imagery of the earth's surface and in addition is very slow. Consequently, the search procedure option is seldom used with the user electing to specify the

number of classes instead.

After the vectors have been clustered into the required number of classes, maps depicting the areas clustered are displayed on the line printer. Tables containing the means and variances of each class as well as the pairwise separability between all class pairs are listed on the printer. The separability table is based on the Swain-Fu distance with the added assumption that the channels are independent.

Usually NSCLAS is used during the preliminary investigation of the data as an aid in defining classes and subclasses. To assist in this task the number of classes into which the vectors are clustered is frequently varied. The output maps generated by NSCLAS are invaluable aids in naming the classes and deciding on the correct number of classes. This is achieved by comparing the map with the "ground truth". The separability table is a valuable guide in defining spectrally separable classes.

4.2.4 GRPSAM: A Parameter Space (Normal) Clustering Program

Clustering is most commonly carried out in the observation space as opposed to the parameter space. The objective of observation space clustering is to group together observation space vectors that are in some sense similar. An example of an observation space clustering program is NSCLAS which has just been described. The objective of parameter space clustering is a little different.

In particular we wish to group together estimated density functions that are similar. Since we are assuming parametric densities this grouping can be done in the parameter space.

Initially the parameters characterizing the probability density function for each training sample are estimated and used to define points in the parameter space, one point for each sample. For the Normal case the parameters that must be estimated are of course the mean vector and covariance matrix for each sample. The hope is that in the parameter space training samples for a given main class would tend to group together at a number of points. Each such group represents a subclass. The objective is to find these groups by clustering in the parameter space.

A flow chart that is commonly used for clustering algorithms is that shown in Fig. 4.2.4.1. This flow chart is for example the basic flow chart for NSCLAS and also serves as a basis for the program to be discussed here. If clustering is done in the observation space, as in NSCLAS, then the objects to be clustered are observation space vectors. In the parameter space the objects to be clustered are points in the parameter space, or parameter space vectors, which in essence represent probability density functions.

A question that arises immediately when clustering

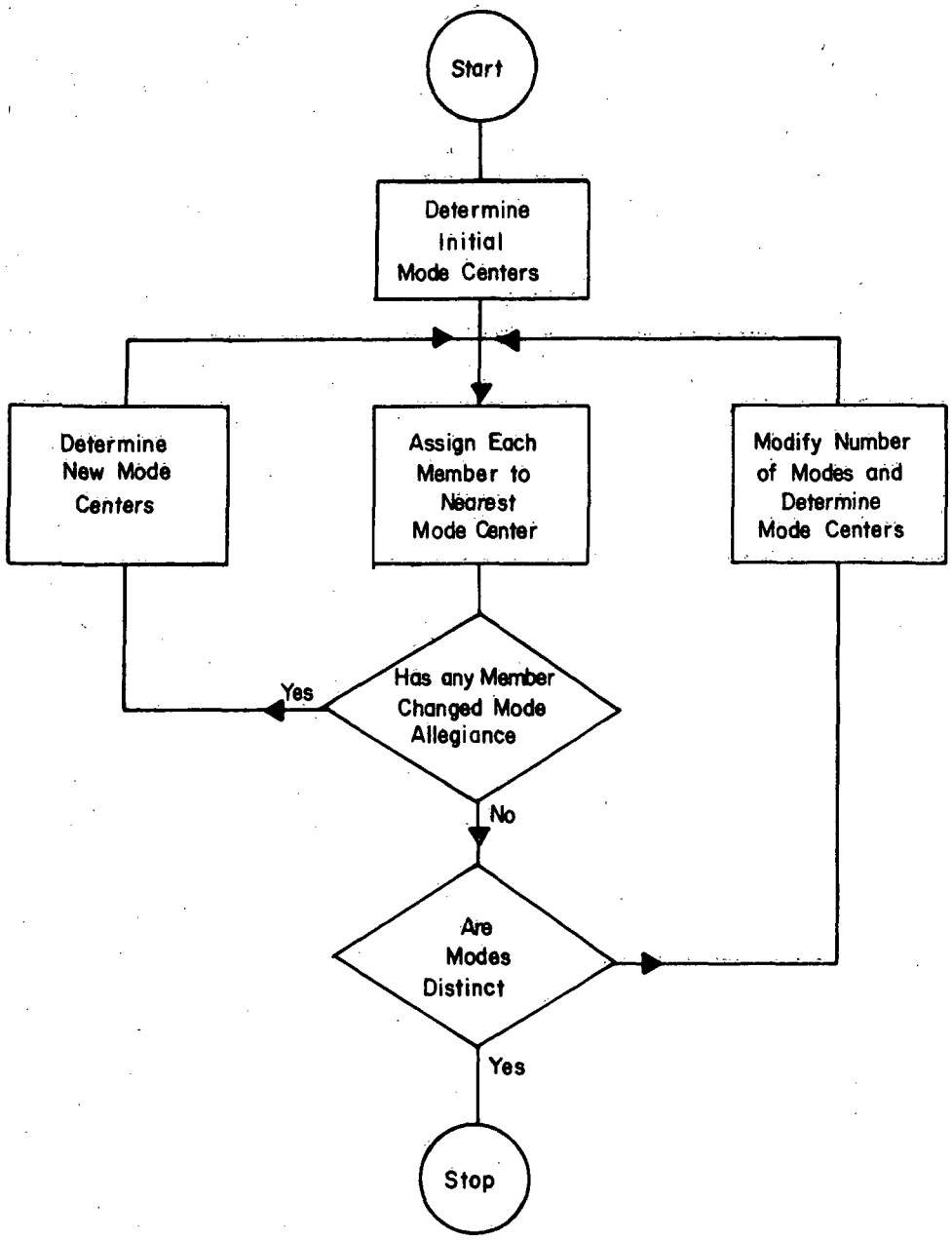


Figure 4.2.4.1 Flow Chart for Clustering.

in the parameter space is the problem of how to measure similarity or distance in this space. Is Euclidean distance a reasonable distance measure in the parameter space or should some other distance measure be used? Use of Euclidean distance for example implies that two univariate normal densities with equal variances and a difference of 1 in their means are just as far apart as two whose means are equal and whose variancesdiffer by 1. The problem of a parameter space distance is readily solved by recognizing that what is really required is a distance measure between density functions. In fact the problem is identical to the problem of choosing a parameter space distance for nearest neighbor classification considered in Section 3.4. Thus to compute the distance between two points in the parameter space we compute the distance between the densities associated with the two points, using one of the available distance measures. By virtue of Section 3.4 this can be viewed as computing the distance between points in the parameter space.

Another question that arises when clustering in the parameter space is that of grouping (i.e., the "determine new mode centers" block in Fig. 4.2.4.1). How does one group together the densities assigned to a mode center to arrive at a representative density or new mode center? In the observation space grouping is usually on the basis of an average of all the vectors in the group. Is this also a

reasonable way of grouping densities? Certainly such a grouping is vastly different from the grouping carried out in LARSYSAA where the statistics for a "grouped class" are based on the pooled vectors of all the samples that are to be grouped.

The previous paragraphs indicate that there are a number of unanswered questions regarding clustering in the parameter space. To answer some of these questions, and evaluate the usefulness of parameter space clustering of multi-spectral scanner data a program GRPSAM (for group samples) was written. The basic flow chart of the program, omitting minor details, is shown in Fig. 4.2.4.1. A discussion of each of the blocks in Fig. 4.2.4.1 is contained in the following paragraphs.

The input to GRPSAM, in addition to the control cards, consists of a statistics deck containing the first and second order statistics of all the samples to be grouped. The format of the statistics deck is the same as that generated by the $STAT processor in LARSYSAA.

The initial mode centers in the parameter space are simply chosen to coincide with the parameter space representation of some of the samples to be clustered. If 15 samples are to be clustered into 5 modes, then every third sample is chosen as an initial cluster center.

Within the clustering loop the assignment of any sample to the nearest mode center is on the basis of one of

four distance measures. The distance measure that can be selected are the Divergence, Bhattacharyya distance, Jeffreys-Matusita distance and Swain-Fu distance. Because of the interrelation between the B and JM distances, the clusters obtained using these two distance measures are identical. Both distances have been included to facilitate the comparison of the numerical output in the separability table with similar output from other programs where either distance may be used.

Four grouping methods are also provided. These are sample-, equal-large-sample-, average-, and product-grouping. In sample-grouping all the vectors used in estimating the densities assigned to a mode are pooled together and the mode mean and covariance are estimated from the pooled vectors. Equal-large-sample- grouping is identical to sample-grouping except it is assumed that all samples grouped contain the same number of vectors and that this number is large. In average-grouping the location of the mode center in the parameter space is simply the mean of all the points in the parameter space associated with that mode. For product-grouping the mode center is the Mth root of the product of the M densities associated with the mode. Appendix E Section E.1 contains more details on the grouping methods in GRPSAM including appropriate mathematical expressions to describe the grouping.

For the distinctness test on the flow chart (Fig. 4.2.4.1) the pairwise distance between all class pairs, using the distance measure selected for clustering, is computed. If the smallest of these pairwise distances exceeds a user specified threshold then the modes are considered to be distinct. If the modes are not distinct the number of modes is reduced by 1 and clustering is repeated. If the modes are distinct processing for that request is complete. The procedure just described is in essence a simple search procedure which can be utilized to attempt to establish the number of modes. It is identical to the procedure used in NSCLAS and has the same disadvantages described in conjunction with the discussion of that program.

The output from GRPSAM consists of a printout depicting the grouping arrived at by the program and if desired an output statistics deck which reflects this grouping is punched. In computing the output statistics the user has the option of utilizing either the grouping method that is selected for grouping in the clustering loop, or else utilizing sample-grouping. A separability table which gives the separation between all mode pairs for all four distance measures is also printed. The maximum, average and minimum pairwise sep ration for each distance measure is also shown in this table.

The different grouping methods available require further discussion. A rough idea of what the different

grouping options accomplish can be obtained by examining the univariate example shown in Fig. 4.2.4.2. Two normal densities which differ in mean and variance are shown as well as the densities that result if these two densities are grouped by the four available methods. Equal-large-sample- and average-grouping result in identical means but average-grouping leads to smaller variance for the grouped density. A still tighter grouped density results from product-grouping. In addition the mean is biased toward the mean of the sample density with smaller variance. Sample-grouping differs from the other three methods in that it takes into consideration the number of vectors used to estimate the parameters of the original densities. The resultant grouped density can be "anywhere between" the two original densities and is biased toward the estimated density based on the larger number of vectors. The equal-large-sample-grouping curve represents the "midrange" for sample-grouping, provided sample sizes are large.

The type of grouping choosen will usually affect the grouping of the samples and consequently the statistics for each mode. However, even if the grouping remains the same for the different grouping methods, the mode statistics for the different grouping methods are quite different. If relatively broad statistics are desired then sample- or equal-large-sample-grouping is most appropriate. To produce slightly tighter mode statistics average-grouping

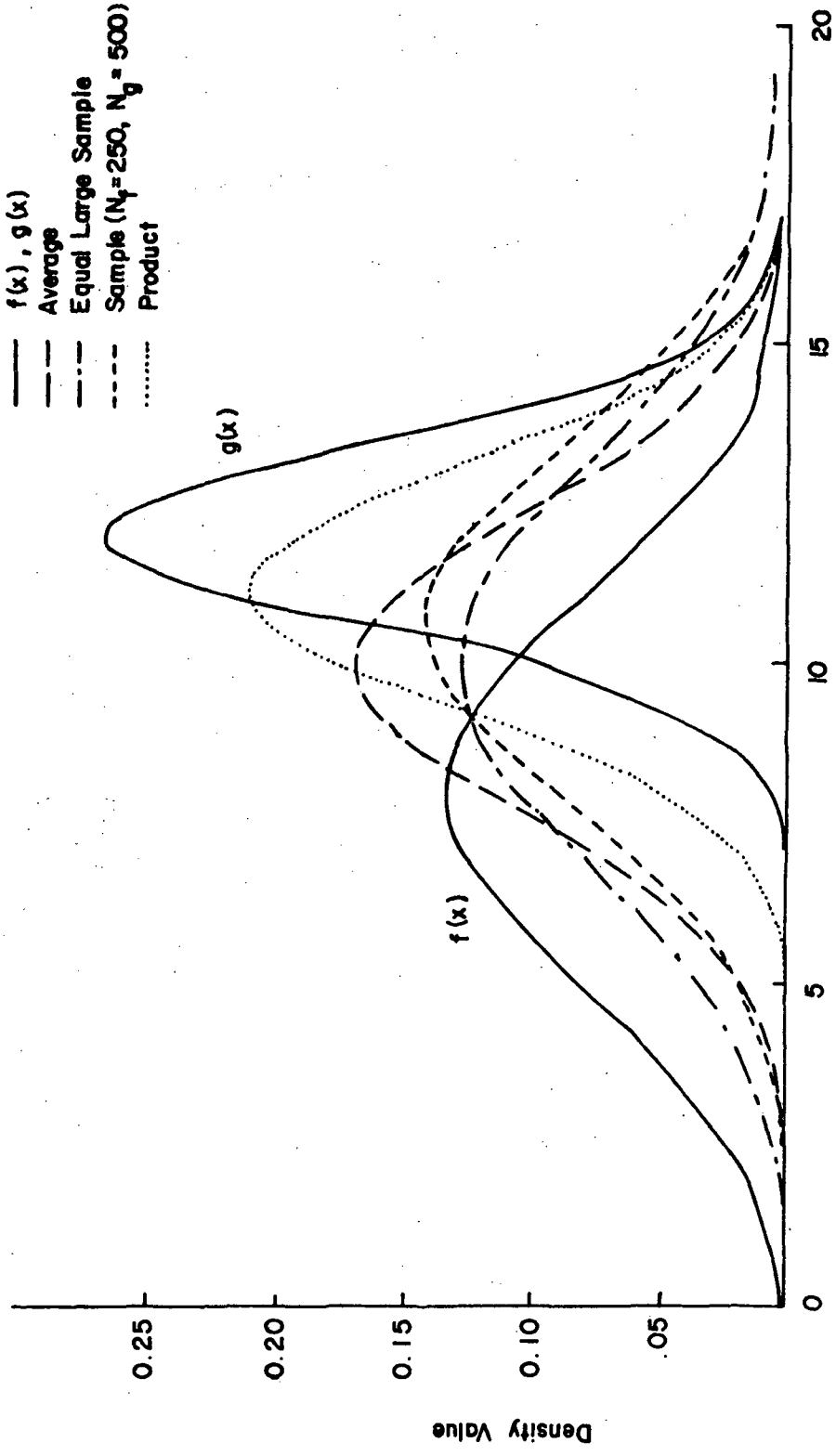


Figure 4.2.4.2 Comparison of Grouping Methods.

should be used. Product-grouping should be used if still tighter statistics are desired.

It is important to note that the statistics generated by GRPSAM can be generated by the $STAT processor of LARSYSAA only if sample-grouping is utilized in computing the statistics. Of course the field grouping used in LARSYSAA must be that arrived at by GRPSAM if identical statistics decks are to be produced. For vector by vector classifiers, such as LARSYSAA, it can be argued quite effectively that the only logical grouping is sample-grouping. For sample classification the situation is not as obvious. In particular one would expect that if a number of samples all with identical means and covariances are grouped, then the mean and covariance for the mode center should be the same as the mean and covariance for each sample. All four grouping methods except sample-grouping posses this property. For sample-grouping it is approximately true for large sample size.

Appendix E Section E.1 contains additional information about the program GRPSAM including a "Control Card description" of the program.

4.2.5 LARSYSDC: A Nonparametric Minimum Distance Classifier

LARSYSDC is a nonparametric minimum distance classifier based on the histogram approach of estimating pdf's and cdf's. Three different distance measures, namely the Kolmogorov-Smirnov, Kolmogorov-Variational and

Jefferies-Matusita distance can be used in the classifier. Only a brief functional description of LARSYSDC appears in the ensuing paragraphs. Appendix E Section E.2 considers in greater detail some aspects of the program, particularly the reasons for selecting histogram estimators and some of the problems associated with these estimators are discussed. A "control card description" of LARSYSDC is also given.

LARSYSDC is divided into three processors under the control of a monitor as shown in Fig. 4.2.5.1. The first processor is the nonparametric pdf processor ($NPDF) which computes density histograms, for the samples specified*, and stores them in a file on magnetic tape. The operation is performed for both the training and test samples, with different tapes used to store the training and test histograms. Storing both training and test histograms facilities classifying the same data with different distance measures. To generate a density histogram for a given sample two passes through the data, associated with that histogram, are necessary. This is a result of the method used to store histograms which is described in Appendix E Section E.2. The first pass essentially establishes the location of the data in E^q while the second pass generates the density histogram.

The second processor in LARSYSDC is the nonparametric cdf processor ($NCDF). This processor converts a

*There are two methods of specifying samples. These are described in Appendix E Section E.2.

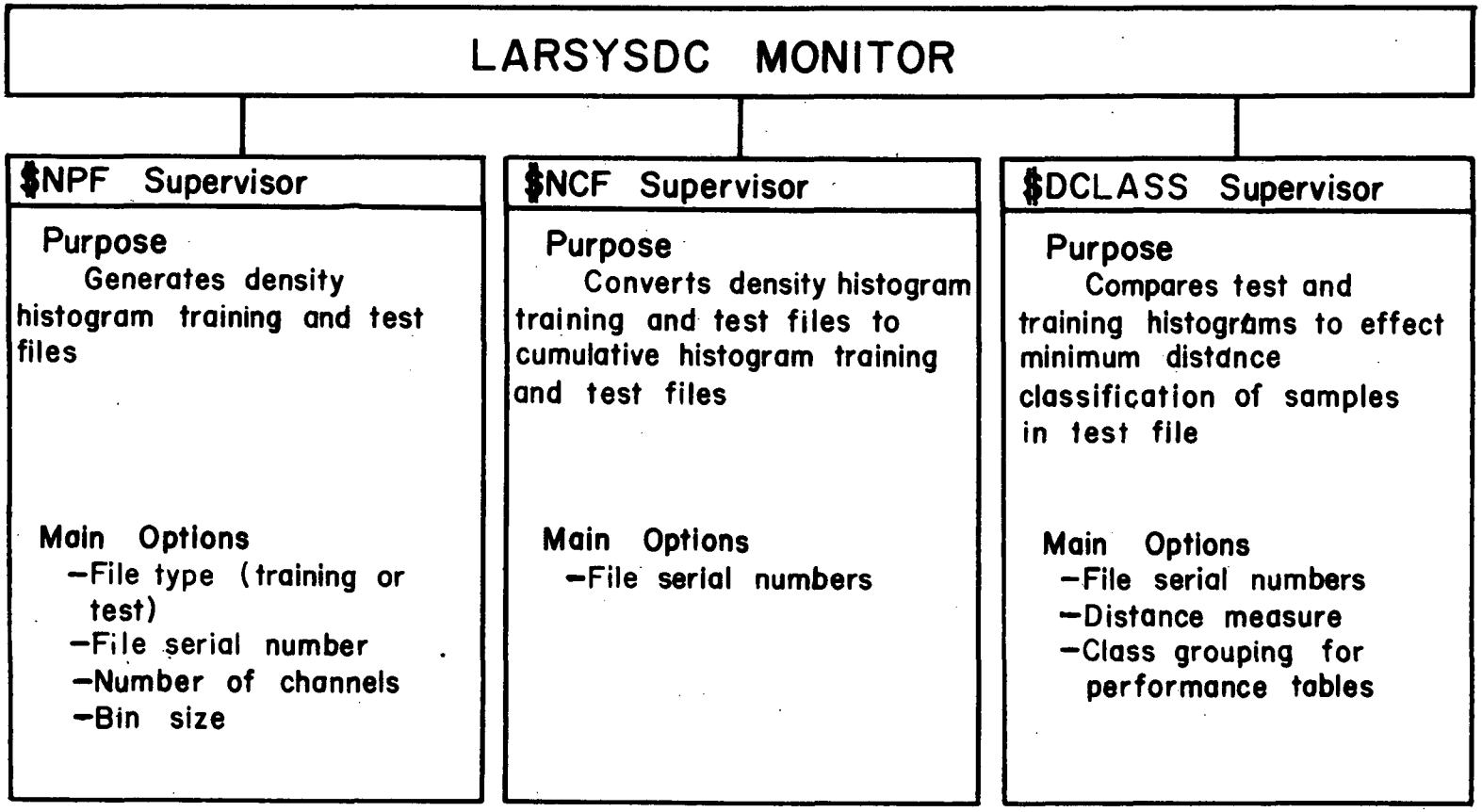


Figure 4.2.5-1 Organization of LARSYSDC

density histogram file to a cumulative histogram file and is used only for distances based on cdf's (i.e., KS distance). Usually the conversion process can be performed fairly quickly but if the number of bins in the density histogram is quite large the required time can be quite large.

The third processor in LARSYSDC is the classification processor ($DCLAS). This processor reads histograms from a file of test histograms and compares them with the training histograms in accordance with the selected distance measure, and lists the classification results. Actually the five nearest neighbors to the unknown density are listed. Performance tables are also printed. The test and training histograms used in the classification must be compatible as to type (i.e., density or cumulative), channels used, and bin size. To enable the largest possible number of channels to be used (i.e., biggest histograms) only two histograms are stored in core at a given time. This means that for each sample to be classified the training histograms must be read into core one at a time and the appropriate distance computations performed. To facilitate this procedure the training histograms are transferred from tape to disk at the start of a classification and then read from disk as required. At the users option the training histograms can be read repetitively from tape rather than disk. Although tape is

considerably faster it is much less reliable in that the excessive tape usage quickly causes frequent read errors to occur.

The selection of the distance measures available in LARSYSDC requires comment. The original intention was to consider most of the distance measures given in Table 2.4.2. Difficulties arise with some of these measures and consequently only the Jeffreys-Matusita, Kolmogorov-Variational and Kolmogorov-Smirnov distances were initially implemented. The classification results obtained with these distances, in addition to those in the parametric classifier PERFIELD suggested that the distance used is not very critical and consequently others were not implemented.

In any case the distances included in LARSYSDC are adequate to enable an investigation of most interesting problem areas. Thus the JM distance is one of the distances implemented in the parametric as well as the nonparametric classifier. This enables a comparison of parametric and nonparametric minimum distance classifiers. The KS distance is based on cdf's and illuminates some of the problems arising in utilizing distances based on cdf's.

The difficulties encountered with some distance measures, which were referred to in a previous paragraph, require discussion. The basic problem is that for some distance measures the distance between most estimated distribution is infinite when histogram estimators are used.

Practical difficulties of this general nature have already been pointed out in Section 3.3 for KL numbers. The Divergence presents an even greater problem in that the Divergence between two density histograms is infinite unless the bins in which the histograms are not zero are identical. A somewhat similar situation prevails for the Cramer-Von Mises distance. In this case the distance between most distributions is infinite unless the distributions are univariate. Recall that to compute the CV distance integration is carried out over all of E^q. This means that unless the two distributions involved approach each other rapidly enough as the independent variable approaches infinitely in most directions the CV distance will be infinite.

The above discussion does not mean that the distances listed could not be used in minimum distance classifiers based on histogram estimators. It does mean that some modification to the fundamental definition of the distance, such as restricting the region of integration, is necessary. Moreover, as already indicated, the results obtained eliminated the need to consider more distance measures.

There is one other problem regarding the implementation of minimum distance classifiers, which are based on histogram estimators, that must be discussed. This concerns the region of E^q over which operations must be carried out

in computing the distance. The basic definitions given in Table 2.4.2 imply that this is typically all of E^q. In practice the region can usually be reduced by virtue of the fact that density histograms are zero in much of E^q, while cumulative histograms contain regions where they are zero or one. This problem is considered in greater detail in Appendix E where we show that the number of bins involved is typically much smaller for the JM and KV distances than for the CV distance. Furthermore it is probably generally true that distances defined in terms of pdf's will usually involve smaller "search regions" than those defined in terms of cdf's. This of course directly affects computation time, which together with the larger time required to estimate cdf's places distances based on cdf's at a definite speed disadvantage in minimum distance classifiers using histogram estimators.

4.3 On Multispectral Scanner Data, Class Selection, and Training Field Selection

Since multispectral scanner data is to serve as the vehicle for the investigation of minimum distance classification a brief description of some of the problems encountered in classifying such data is the subject of this section. The discussion is directed primarily at classifying agricultural scenes since most of the experience has been with this type of data. Furthermore interest in sample classification schemes is greatest in this context.

In the agricultural setting the classes of interest are frequently the various types of ground cover (i.e., crops). These classes, and indeed in general, any classes that might be considered as possible classes in classifying multispectral data should possess the following two characteristics:

(a) Classes should be of practical utility. That is the classes defined should be of interest to some individual or group of individuals.

(b) Classes should be sufficiently seperable spectrally so that the established constraints on probability of error can be achieved.

Requirement (a) can be met without reference to the data and consequently fits nicely into a supervised system. Requirement (b) on the other hand requires that the data be examined and is essentially of an unsupervised nature. It is important to note the (a) and (b) may be conflicting requirements and that it may not be possible to satisfy them simultaneously. Frequently classes are defined (at least initially) on the basis of their practical utility and then tested for separability. If separability is poor, as evidenced by a large probability of error, a new set of classes is defined taking into account what has been learned about separability. It is also possible to devise a classification system that approaches the problem with the other initial premise. In such a system classes would

be defined on the basis of their separability. An attempt would then be made to associate the resultant classes with classes that have some practical utility. Defining classes on the basis of observation space clustering is such an approach. The ideal training procedure would effect a compromise between requirements (a) and (b) prior to the start of classification.

Another factor which must be born in mind when LARSYSAA and PERFIELD are used is that these programs are based on the Gaussian assumption. This, of course, does not mean that they cannot be used if the data is not Gaussian, but it does mean that performance predictions based on the Gaussian assumption are not applicable. In general one might expect reasonable performance if the data is unimodal and symmetrical. Unless classes are very separable multimodal classes tend to give rise to large probabilities of error and should be avoided.

With regard to the Gaussian assumption it appears that typically data from an individual field, regardless of crop type, is usually reasonably unimodal and symmetrical. The unimodality makes the Gaussian reasonable for an individual field. Occasionally individual fields do exhibit bimodality, but if field boundaries are chosen with care this is the exception rather than the rule. On the other hand, different fields of the same crop type frequently are sufficiently different spectrally so that the combined data

from two such fields exhibits distant bimodality. Under these circumstances in order that the Gaussian assumption is approximately satisfied, subclasses are usually defined for each main class (e.g., wheat 1, wheat 2, etc.), such that the distribution of each subclass is unimodal. Perhaps if training samples could be drawn from sufficient variety of fields for a given crop type a unimodal distribution would result for each main class and the definition of subclasses would not be necessary, even for a parametric classifier. The class distribution in this case would naturally be broader than the distribution of any of the subclasses of which it is composed. It is presently not known whether better classification is achieved by using many subclasses whose distribution are relatively narrow or using fewer subclasses with broader distributions, although the trend appears to be toward the definition of many subclasses.

From the above discussion it is apparent that the definition of subclasses is a problem of considerable importance in classifying multispectral scanner data. Consequently, the usual methods that are used to select subclasses will be briefly discussed.

(a) Histogramming Method - A large number of fields are histogrammed for each main class and the number of subclasses defined in the basis of visual examination of these histograms.

(b) Iterative Classification Method - The data is classified on the basis of one or more classes per crop type. Fields that are incorrectly classified are used to help establish subclasses.

(c) Divergence Method - Every possible training field for a given crop type is defined as a subclass. The Divergence computing capability of the feature selection algorithm ($DIVG) is then used to decide which of the subclasses are sufficiently alike so that they may be combined.

(d) Observation Space Clustering - Observation space vectors all belonging to the same main class are clustered into various number of modes and subclasses established on the basis of the mode separability.

(e) Composite Method - Some combination of (a), (b), (c) and (d).

All of these methods have disadvantages of one sort or another. The histogramming and iterative methods require considerable personal intervention and judgement and consequently, are quite slow. Furthermore, there appears to be no way in which the iterative method could be automated. The histogramming method could be automated by defining a suitable distance function between histograms. If this

were done, this method would very much resemble the Divergence method, except that it would appear to be inferior in that it depends only on the marginal distributions and ignores correlation effects. The Divergence method seems to be a useful approach. Utilizing LARSYSAA to implement this approach is somewhat akward in that the available software is used in a non-standard fashion; but this is not a fundamental problem. A further extension of the Divergence approach leads to parameter space clustering; in this situation the manual grouping is replaced by automatic grouping.

Observation space clustering is probably the most automated and "best" method of defining subclasses in general use at LARS. The rapidity with which this method gained acceptance clearly testifies to its usefulness. Normally, since the number of separable subclasses is unknown, it is necessary to cluster the data into various numbers of modes. This together with the large volume of computations that must be performed to cluster the data for each mode specification means that considerable computation time is involved. The method does have the distinct advantage that it readily leads to the definition of subclasses whose histograms are reasonably unimodal and symmetrical.

It is worthwhile noting that regardless of the manner in which classes and subclasses are defined,to obtain a classification with the parametric classifiers LARSYSAA and PERFIELD is usually an iterative process. It is unfortunate that this is so, since the iterative approach is

very time consuming. The crux of the problem is that the classifiers are supervised systems. Consequently, the assumptions that the number of classes are known apriori, and that training samples are available for each class are inherent in the classifiers. In practice these assumptions are simply not valid for a parametric classifier. One may know the number of main classes (i.e., classes of practical utility) but the number of subclasses required to reasonable satisfy separability requirements and the parametric assumptions are not known; and consequently, the total number of classes is unknown. There appears to be no simple solution to this problem for the parametric case. The use of clustering programs like NSCLAS and GRPSAM assists somewhat in alleviating this problem in that some idea about classifier performance can be obtained before proceeding to the classification stage. Ultimately, however, it is the classifier that decides the quality of the training and a certain amount of iterative classification appears unavoidable. In this regard care must be exercised to avoid the temptation of using test results to improve classifier performance. Such a procedure of necessity leads to optimistic results. Modifications to the training statistics must in most realistic situations be based on the training results only. Test fields serve the sole purpose of evaluation classifier performance. In a certain sense utilizing test results to improve classifier performance is equivalent to

the utilization of the test fields as training fields.

At first glance nonparametric classifiers appear to provide some advantage in that the definition of subclasses is no longer necessary, and in fact some favorable results have been obtained with such methods under very controlled conditions on exceedingly limited amount of data.[73] In terms of classifying a large volume of data it is not at all clear that nonparametric technique simplify the training. The problem of defining subclasses is simply replaced with the problems of selecting the samples to be included in the training set. Of course, nonparametric methods should not be overlooked but they do have a number of disadvantages. In general nonparametric methods tend to be slower and require more storage than parametric methods. This is in fact a very real problem if one considers classifying the vast amount of data that becomes available in the remote sensing of earth resources. Intuitively one feels that a simpler system will be achieved if reasonable results can be obtained and the parametric assumption maintained.

Another factor of considerable importance is that as flightlines become longer, the need for systems that have adaptive capabilities will increase. The reason for this is that the data almost certainly will not remain sufficiently uniform over a long flightline so that a single fixed set of training fields will suffice.

4.4 Experimental Evaluation of GRPSAM

In describing the program GRPSAM it was pointed out that a number of options existed with regard to the distance measure and grouping method used during clustering. In this section experiments designed to evaluate the various grouping methods and distance measures are described. The evaluation is accomplished by comparing the classification accuracy acheived on a fixed set of training and test fields, where the class statistics are generated by clustering the training fields with GRPSAM using various combinations of distance measures and grouping methods.

Before becoming involved in the details of these comparative classifications it is advisable to try and establish a "feeling" for the clustering properties of GRPSAM, as well as the distance measures utilized. Although observation space clustering is a technique in common usage this does not appear to be true for parameter space clustering. In addition the distances (in some cases metrics) used in parameter space clustering are rather complicated functions of the coordinates and it would be useful to obtain a deeper understanding of the "metric-properties" of the distances involved. For example it would be desirable to know if what the eye perceives as a cluster in a parameter space scatter plot still appears as a cluster in terms of a particular distance measure. After all, the distance measures used in the parameter space differ

considerably from the Euclidean metric to which the eye is attuned. Consequently, before comparing various distance measures and grouping methods we consider the distance measures involved in GRPSAM from a parameter space point of view.

4.4.1 "Metric-Properties" and Other Characteristics of Distance Measures used in GRPSAM

For the bivariate case the parameter space is five dimensional. Consequently any graphical aids in understanding the distance measures used in GRPSAM are essentially restricted to the univariate case. For this reason we focus attention on this case.

Perhaps the simplest technique for gaining some understanding of the "metric-properties" of the distances involved is to draw constant distance contours in the parameter space. Actually for the univariate case the expressions for JM Distance, Divergence, and SF distance can be normalized and a universal set of constant distance contours can be drawn on the resulting normalized axis. Let (μ_o,σ_o) be a point in the parameter space about which constant distance contours are drawn and let (μ,σ) be an arbitrary point at a fixed distance from (μ_o,σ_o). Then utilizing table 2.4.3 and defining the normalized mean μ_n as

$$\mu_n = (\mu-\mu_o)/\sigma_o \qquad 4.4.1.1$$

and the normalized standard deviation as

$$\sigma_n = \sigma/\sigma_o \qquad 4.4.1.2$$

we can write for the JM distance, the Divergence, and the SF distance respectively;[+]

$$M = \{2[1 - (\frac{2\sigma_n}{1+\sigma_n^2})^{1/2} \exp(\frac{-\mu_n^2}{4(1+\sigma_n^2)})]\}^{1/2} \qquad 4.4.1.3$$

$$J = \frac{1}{2}(\frac{1-\sigma_n^2}{\sigma_n^2}) + \frac{1}{2}(\frac{1+\sigma_n^2}{\sigma_n^2})\mu_n^2 \qquad 4.4.1.4$$

$$T = \frac{\mu_n^2}{\sqrt{3}(1+\sigma_n)} \qquad 4.4.1.5$$

Families of these equations are plotted in Fig. 4.4.1.1 with constant values of M, J, and T as a parameter. Constant distance contours for the Bhattacharyya distance are identical to those for the JM distance by virtue of 2.4.7, only the numerical value for the distance is different.

The constant distance contours for the JM distance and Divergence have some points of similarity in that they are closed and have an oval shape. The similarity is more pronounced for densities whose separation is small. For densities with large separation the differences become more pronounced and consequently the global properties for the two distances are quite different as we presently

[+]Recall that the mathematical symbols used to represent the JM distance, Divergence and SF distance are M, J and T respectively.

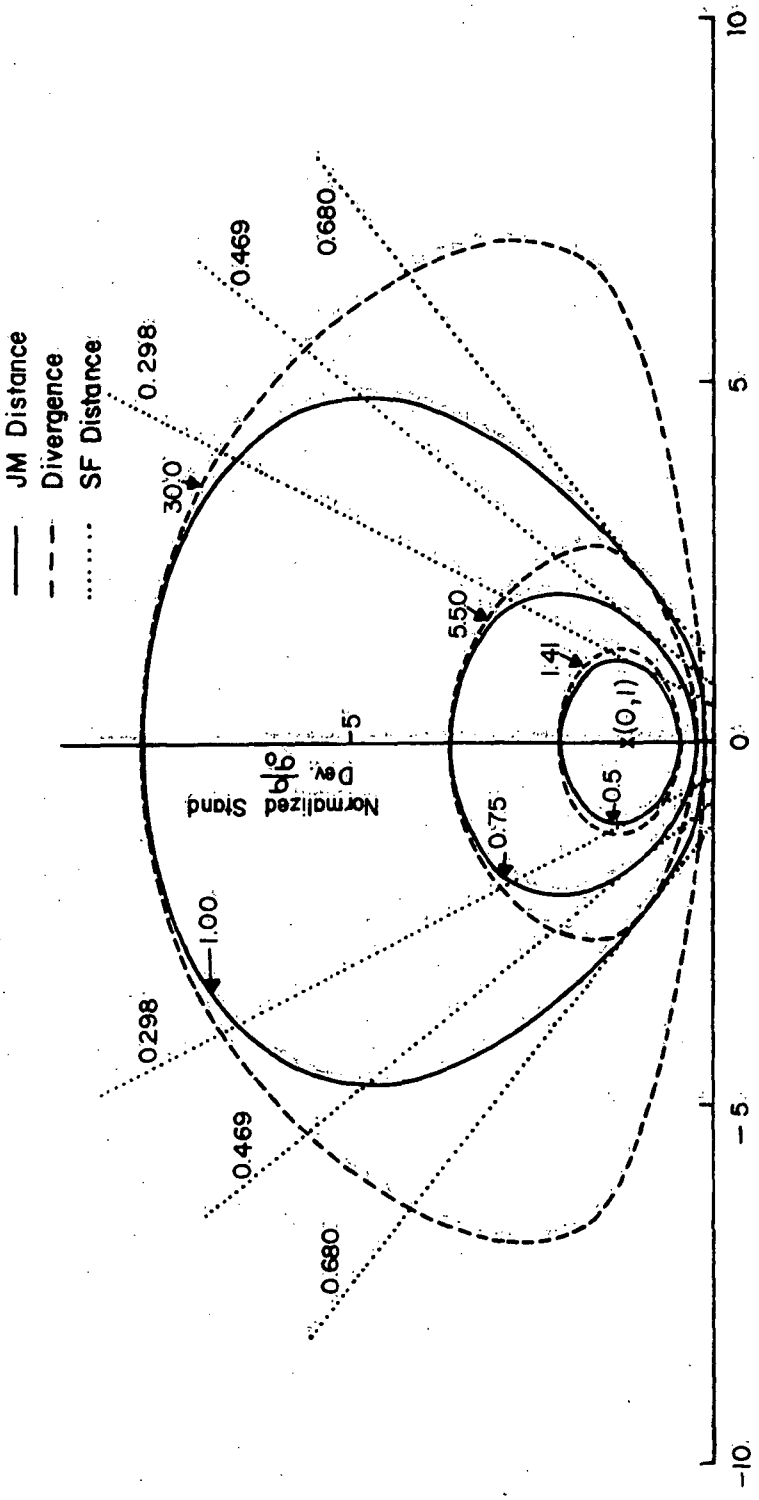


Figure 4.4.1.1 Constant Distance Contours for the Normalized Univariate Case for the JM distance, Divergence and SF distance.

demonstrate. The SF constant distance contours obviously differ considerably from those for the JM distance and Divergence.

Another way of demonstrating some of the "metric-properties" of a distance measure is to plot contours that are equi-distant from the two selected points (mode centers) in the parameter space. In fact equi-distance curves are more important than constant distance curves from the view point of clustering. It is of course true that equi-distant contours can be constructed by using constant distance contours, but the shape of the equi-distance curves is extremely difficult to visualize from the constant distnace contours. Subtle changes in the shape of the constant distacne curves can produce radical changes in the equi-distance controus. A good example of this is Fig. 4.4.1.2 where equi-distance contours for the three distances under consideration are shown. Note the difference between the equi-distance contours for JM distance and Divergence even through their constant distance curves were quite similar.

Normalization of equi-distance curves is not possible. This means that many examples like that shown in Fig. 4.4.1.2 must be considered before a good understanding of the "metric properties" of the distances can be obtained. Actually the curves Fig. 4.4.1.2 are fairly typical of the situation encountered for real multispectral scanner data. Typically in the vicinity of the mode centers

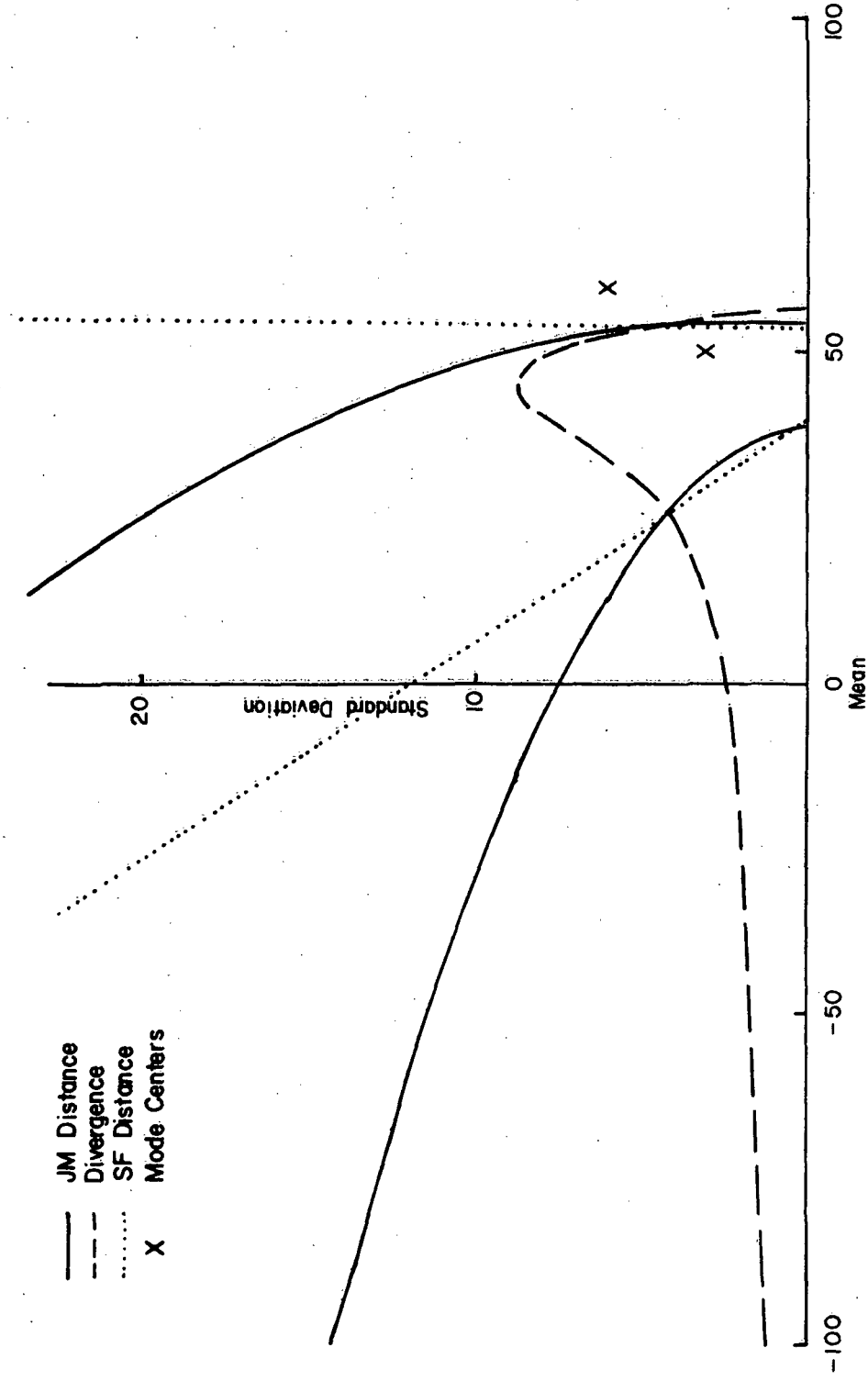


Figure 4.4.1.2 Global Partitions of the Parameter Space for Arbitrary Mode Centers Using JM Distance, Divergence and SF Distance.

the curves are all quite similar for the different distances and roughly at right angles to the mean axis. In regions of the parameter space that are remote from the mode centers the curves are drastically different. In practice this is of little consequence since typically there is no data in the remote regions. The fact that in the vicinity of the mode centers the curve are roughly orthogonal to the mean axis implies that the means of the mode centers have considerably greater influence in determining the partition surface than do the variances. Furthermore, the investigation of higher dimensional cases (by observing appropriate two dimensional cross plots) indicates that this situation also tends to prevail in higher dimensional cases.

The constant distance contours in Fig. 4.4.1.1 can be used to infer the existence of certain bounds involving the three distance measures under consideration. For example we note that the 5.50 constant Divergence curve appears to lie between the 0.75 and 1.00 constant JM distance curves. This implies that for a Divergence of 5.50 the JM distance is bounded above by 1.00 and below by 0.75, and in general suggests the existance of a upper and lower bound on the JM distance for a given Divergence.

The upper bound is quickly established because it is known for the multivariate normal case[23] that

$$J \geq 8B \qquad 4.4.1.6$$

This combined with 2.4.7 yields

$$M^2 \leq 2(1-e^{-J/8}) \qquad 4.4.1.7$$

It is interesting to note that this upper bound can be inferred directly from Fig. 4.4.1.1. Let $M(\mu_n,\sigma_n)$ and $J(\mu_n,\sigma_n)$ be the JM distance and Divergence as given by 4.4.1.3 and 4.4.1.4 respectively. A careful examination of the largest JM distance curve that just fits outside a given Divergence curve (e.g., JM distance equals 1.00 and Divergence equals 5.50) suggests that the mathematical property relating such curves is

$$M(|\mu_n|,1) = J(|\mu_n|,1). \qquad 4.4.1.8$$

That is, the upper bound appears to coincide with the case where both the constant JM distance and constant Divergence contours pass through the points $(\pm\mu_n,1)$ for arbitrary μ_n. It is readily verified that the slope of both contours passing through these points are identical lending further credence to the suggested relation. Using 4.4.1.8 in conjunction with the expressions for the JM distance and Divergence quickly leads to the upper bound given by 4.4.1.7.

A lower bound can also be inferred from Fig. 4.4.1.1. For this case the mathematical property that appears to relate the constant JM distance contour that just fits inside a given constant Divergence contour (e.g., the JM distance equals 0.75 and the Divergence equals 5.50 curves)

is

$$M(0,\sigma_n) = J(0,\sigma_n) \qquad 4.4.1.9$$

That is, the lower bound appears to coincide with the case where the constant Divergence and constant JM contours pass through the same points on the σ_n axis. Thus setting μ_n to zero in 4.4.1.3 and 4.4.1.4 and eliminating σ_n we obtain

$$M^2 \geq 2\left[1 - \left\{\frac{2(\sqrt{J/2} + \sqrt{J/2 + 1})}{(\sqrt{J/2} + \sqrt{J/2 + 1})^2 + 1}\right\}^{1/2}\right] \qquad 4.4.1.10$$

Utilizing the mathematical identity

$$\text{Sinh}^{-1}\sqrt{J/2} = \text{Ln}\,(\sqrt{J/2} + \sqrt{J/2+1}) \qquad 4.4.1.11$$

4.4.1.10 can be written as

$$M^2 \geq 2[1 - \text{Sech}^{1/2}(\text{Sinh}^{-1}\sqrt{J/2})]. \qquad 4.4.1.12$$

The derivation for the lower bound given by 4.4.1.12 is not rigorous and we have not been able to rigorously prove that it is correct. Experimental results have been obtained which suggest it is correct even in the multivariate case. These results are shown in Fig. 4.4.1.3 and 4.4.1.4 where scatter diagrams of the JM distance squared vs. Divergence are plotted for the univariate and the trivariate normal cases respectively. The upper and lower bounds given by equations 4.4.1.7 and 4.4.1.12 have also been plotted.

The data used for the scatter plots are data from 20 of the wheat Training Acres whose coordinates are given in Table C.4. Statistics were calculated for these

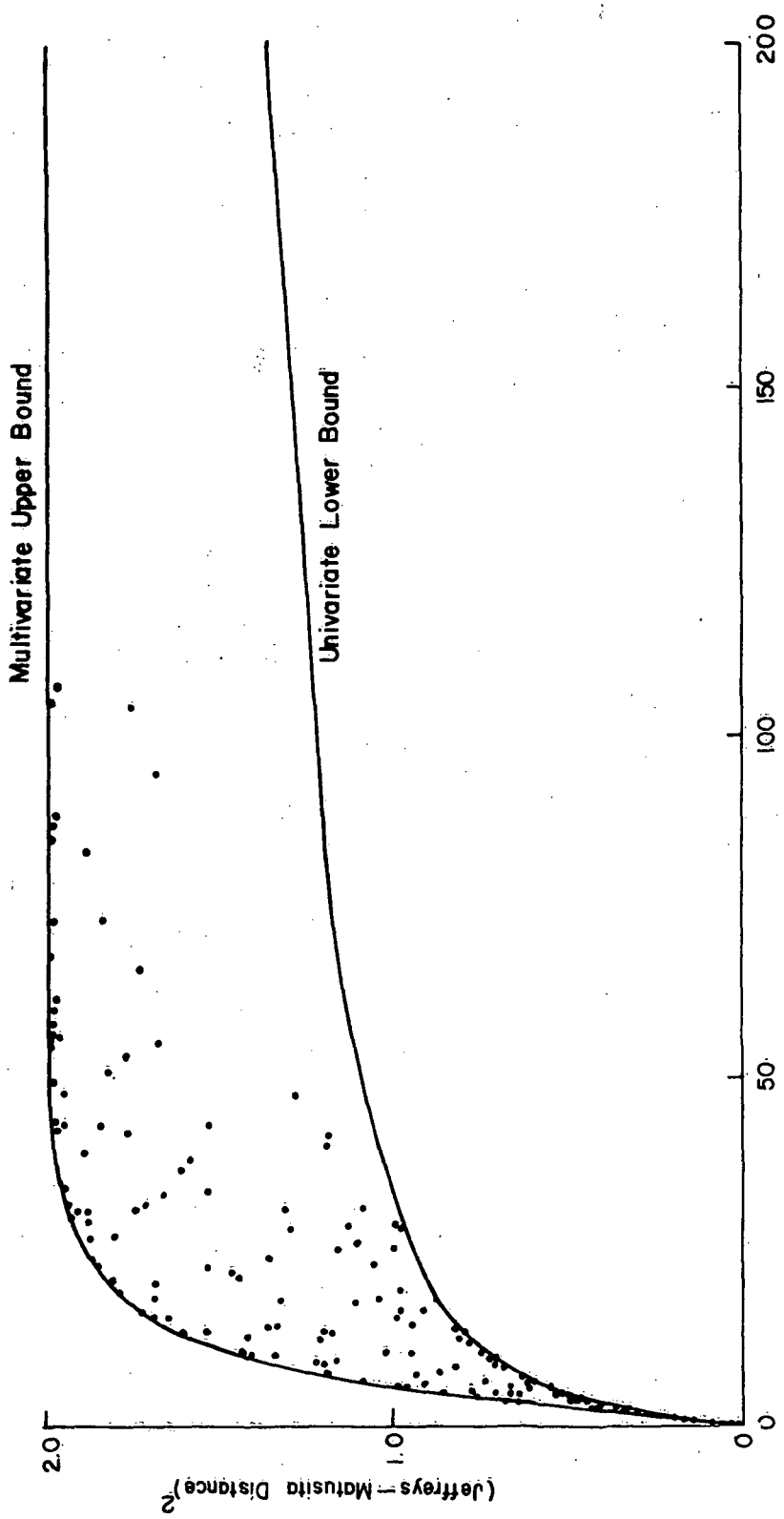


Figure 4.4.1.1.3 Upper and Lower Bounds for JM Distance as a Function of Divergence. Univariate Normal Case.

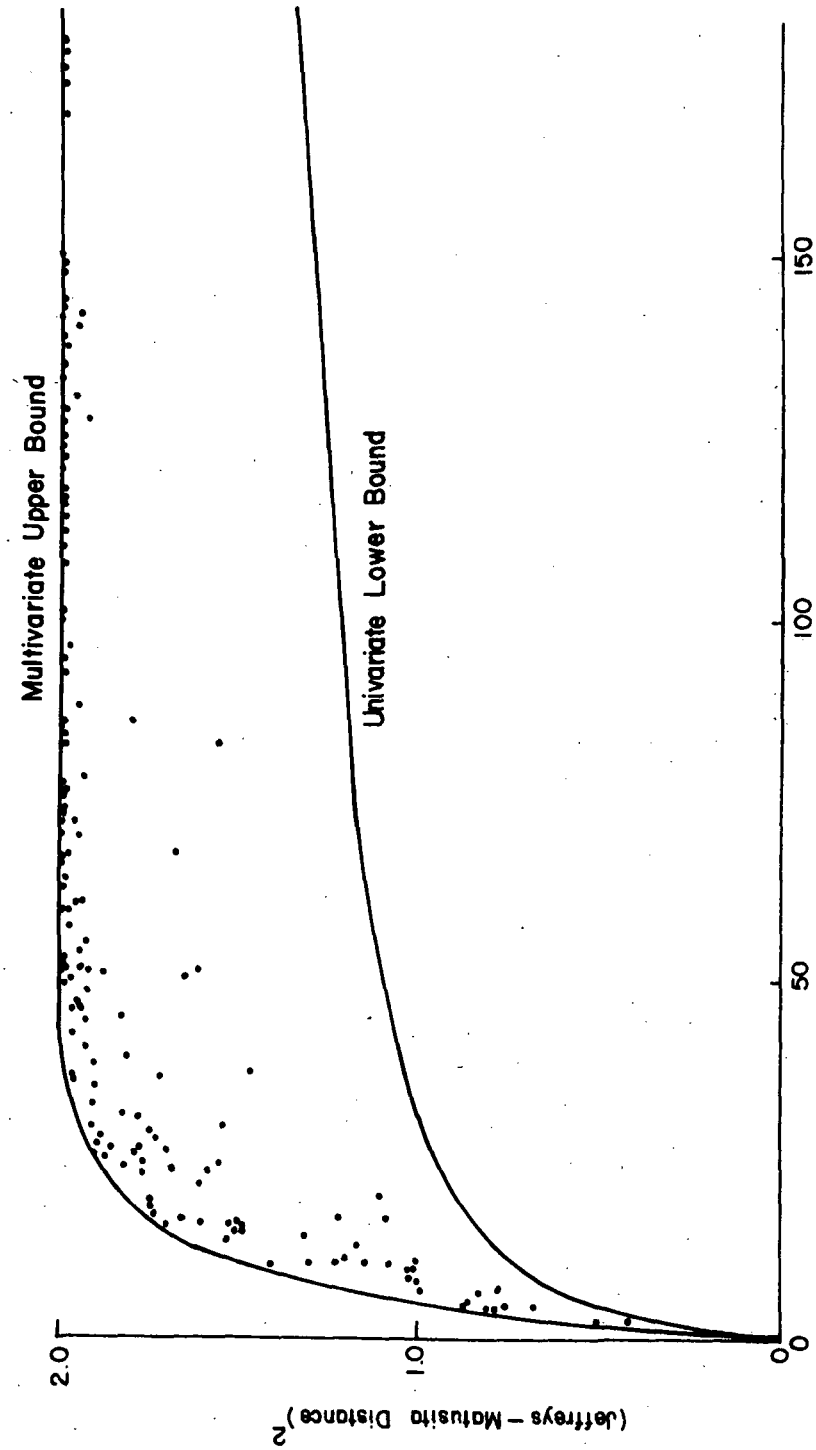


Figure 4.4.1.4 Upper and Lower Bounds for JM Distance as a Function of Divergence. Trivariate Normal Case.

acres using LARSYSAA and then GRPSAM was used to compute the pairwise JM distance and pairwise Divergence between all 20 wheat acre densities. In particular by setting the number of modes equal to the number of acres, the separability table in GRPSAM contains the pairwise separation between the densities of all acre pairs based on the channels selected. All the data so obtained for both the trivariate and univariate case fell between the bounds depicted. For the trivariate case all data was considerably above the lower bound. In fact as the number of dimensions increases the points tend to become more and more concentrated near the upper bound. Whether this is due to an increase in the lower bound or simply due to a general increase in separability as the dimensionality increases is not known; but it is believed to be due to the latter factor. In any case Swain et al. [66] have utilized this property in feature selection. They observed experimentally that the average (over class pairs) JM distance provided better feature selection capabilities than the average Divergence, but was computationally more complex. By utilizing the upper bound in Fig. 4.4.1.4 as a "transformed Divergence", they were able to retain the computational simplicity of the Divergence and attain performance approaching that achieved with the JM distance. Since for a reasonable number of dimensions most of the points are near the upper bound the choice of the upper bound as a transforming relationship

between Divergence and JM distance is quite reasonable.

The constant distance contours of Fig. 4.4.1.1 also suggest a bound on the SF distance. In particular one would expect that for a given Divergence the SF distance should have a lower bound of zero, and that an upper bound should also exist. In fact by procedures similar to those discussed for the JM distance the relation

$$T \leq \sqrt{J/12} \qquad 4.4.1.13$$

is obtained as an inferred upper bound for the univariate normal case. This result has also been rigorously derived. The derivation is given in Appendix A Section A.2 where for arbitrary dimensionality q we show that

$$T \leq \sqrt{J/4(q+2)} \qquad 4.4.1.14$$

In Fig. 4.4.1.5 and 4.4.1.6 we show scatter plots of the SF distance vs Divergence for the univariate and trivariate normal cases respectively. These plots are based on the same data and are obtained in the same manner as the JM distance plots previously described. In all cases the data conforms with the derived bounds. The most striking characteristic of these graphs is the decrease in the upper bound as the dimensionality increases in accordance with 4.4.1.14. This means that unless the Divergence increases sufficiently rapidly with dimensionality the SF distance between distributions will in the limit decrease as

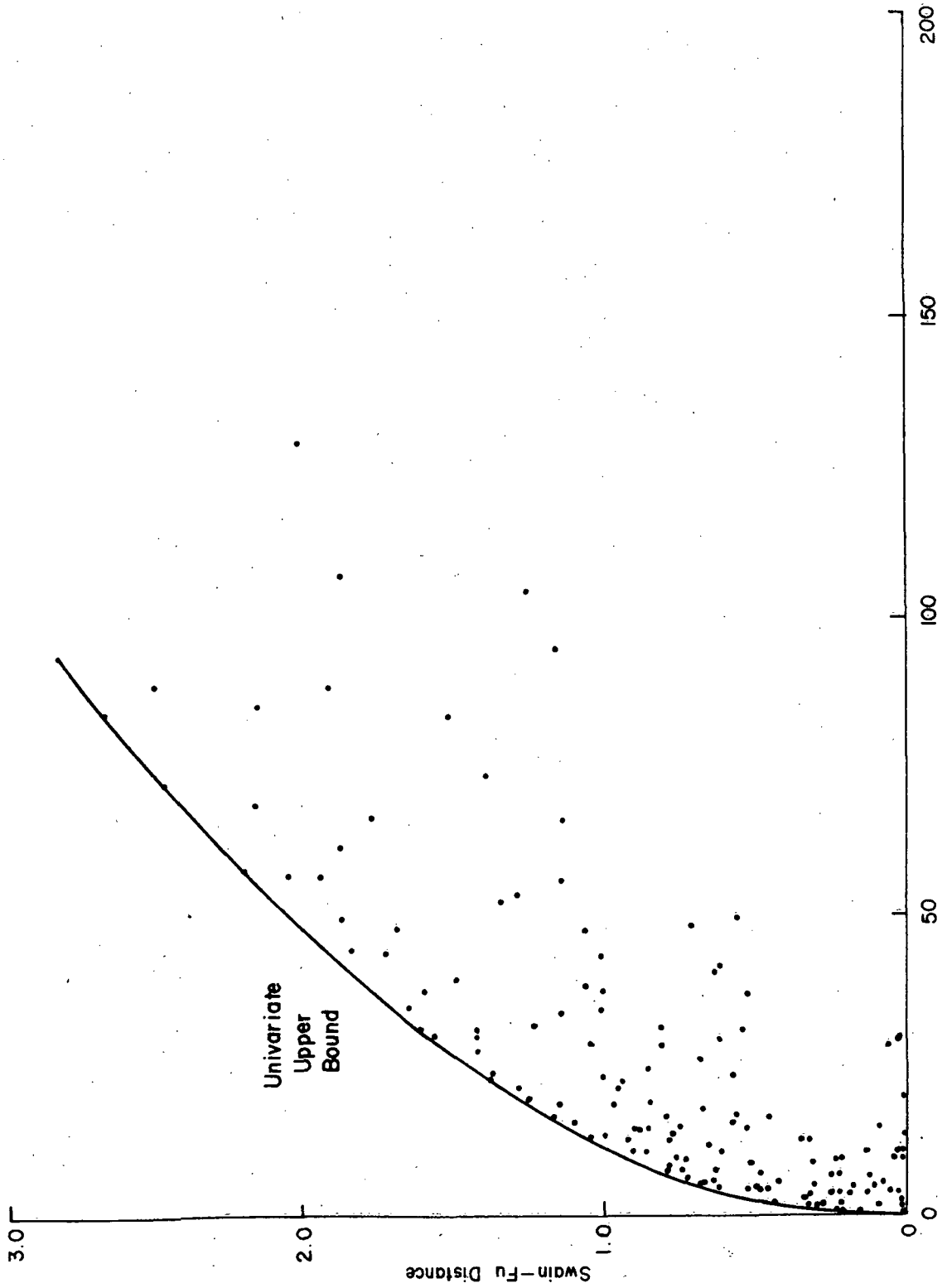


Figure 4.4.1.5 Upper Bound for SF Distance as a Function of Divergence. Univariate Normal Case.

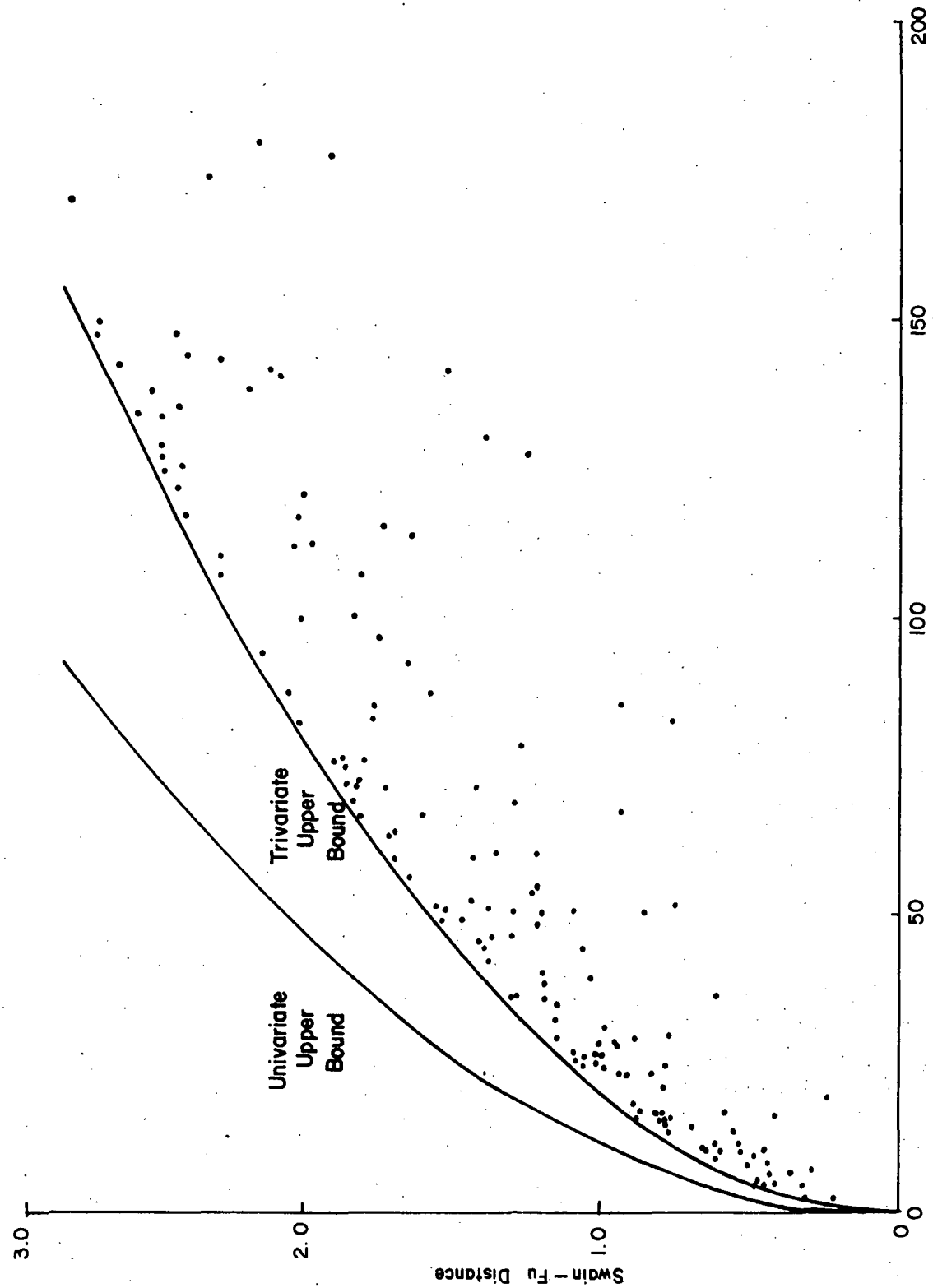


Figure 4.4.1.6 Upper Bound for SF Distance as a Function of Divergence. Trivariate Normal Case.

dimensionality increases.

The manner in which the Divergence, JM distance and SF distance vary with dimensionality is a matter of considerable practical importance since these distances may be used to assess class separability and in feature selection. The question that arises is whether or not a given numerical distance should be interpreted in the same manner regardless of dimensionality. To shed some light on this question GRPSAM was utilized to calculate the average pairwise distance over all class pairs between the parametrically estimated densities of 20 of the wheat Training Acres for the JM distance, Divergence and the SF distance. The results are plotted in Fig. 4.4.1.7. While these results were computed for one particular data set the gross characteristics undoubtedly apply to most sets of multispectral scanner data.

The manner in which the average distances in Fig. 4.4.1.7 varies with dimensionality depends very much on the distance measure involved. Perhaps the most interesting variation is that of the average SF distance which first increases and then decreases as extra dimensions are added. The behavior is similar to the behavior of the separability measure of Section 3.2 for a saturating S/N ratio. In fact the behavior of the SF distance can be interpreted in terms of that result. Thus the increase with dimensionality of the average distance between pairs of vectors in a class means that the ellipsoids of concentration must get larger

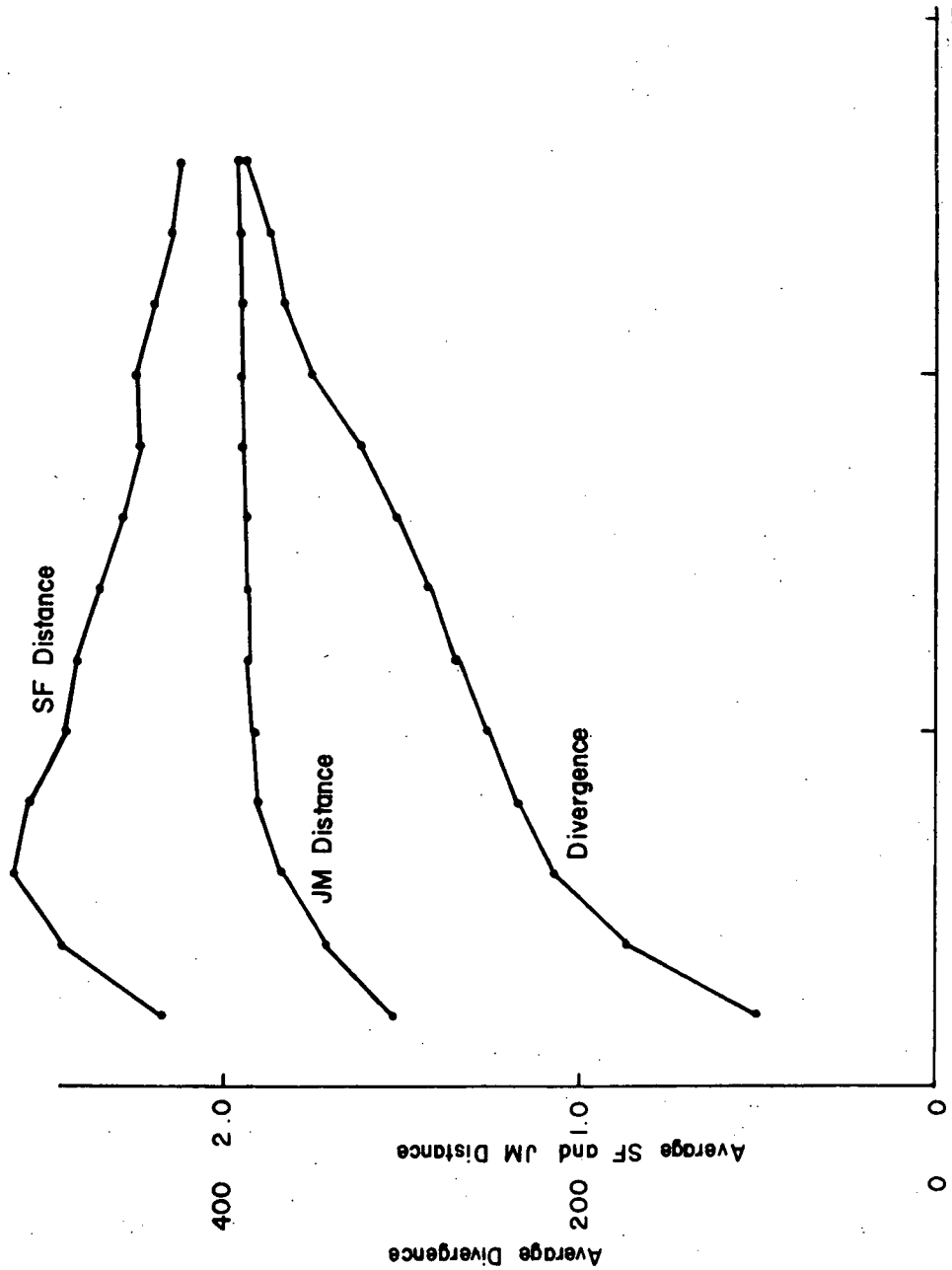


Figure 4.4.1.7 An Example of Average Class Separation as a Function of Dimensionality

as the dimensonality increases. Recalling that the Swain Fu distance is the ratio of the distance between mode centers, to the sum of the distances from the mode centers to the ellipsoids of concentration along the line joining the mode centers, it is obvious that unless the distance between the mode centers (essentially S/N ratios) increases rapidly enough with dimensionality the SF distance must decrease. The average separability curve for the SF distance also lends credence to the earlier contention that the basic results of Section 3.2. are essentially independent of the restrictive assumptions of that section, this follows because the wheat acre densities certainly did not obey the restrictive assumptions of Section 3.2.

The behavior of the average Divergence and average JM distance in Fig. 4.4.1.7 are also of interest. The average Divergence continues to increase as the dimensionality increases while the average JM distance saturates. The saturation of the average JM distance is easy to explain in that no pairwise JM distance can exceed 2. The shape of the JM distance curve is generally similar to that obtained for probability of correct recognition. This in our opinion is a rather desirable property in feature selection and other applications. The properties of the JM distance, Divergence and SF distance depicted in Fig. 4.4.1.7 in no way restricts the use of these distance measures. If these distance measured are used in a situation where the number of dimensions is variable then the results of this section are essential if

misinterpretation is to be avoided.

4.4.2 An Example of Parameter Space Clustering of Multispectral Scanner Data

In this section an example of clustering in the parameter space is presented. The wheat Training Acres listed in Table C.4 are selected for this example. Statistics were obtained for each of the 59 wheat Training Acres and these were then clustered in the parameter space using various number of channels and each of three grouping methods. The three grouping methods used are sample-, average-, and product-grouping. Equal-large-sample-grouping is not considered as all acres were of equal size and moderately large (121 vectors). Thus the results for sample- and equal-large-sample-grouping would be very similar.

Figure 4.4.2.1 shows the parameter space groupings arrived at when only channel 11 is used to group the data, with Divergence as the distance measure. Results are shown for each of the three grouping methods. The mode centers obtained are indicated by X's and the letters S, A, and P are used to indicate sample-, average-, and product-grouping respectively. Once the mode centers are known then equidistant contours can be constructed as described in the previous section. Such contours are shown for each of the grouping methods. These curves partition the parameter space into disjoint regions associated with each mode.

There are a number of observations that can be made with regard to Fig. 4.4.2.1 which we list numerically.

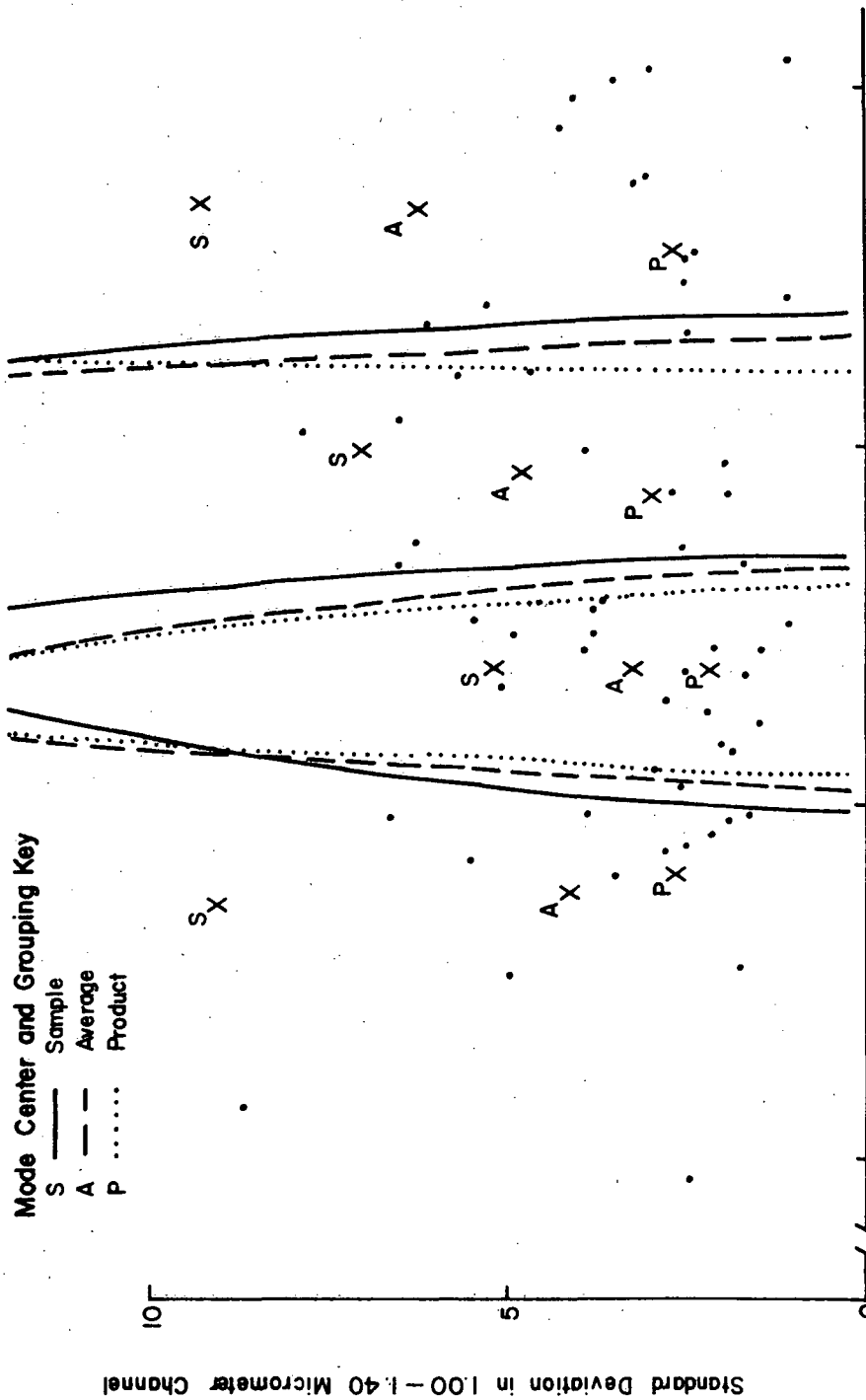


Figure 4.4.2.1 Parameter Space Clustering of Wheat Training Acres Using Divergence and Channel 11.

Observation 1

There is considerable variability in the data in that on a scale that ranges between 0 and 255 the means for the wheat Training Acres ranged between approximately 100 and 175. One is tempted to attribute this variation to the fact that harvested as well as unharvested fields are included in the acres. While it is true that the harvested acres are on the higher end of the 100 to 175 range they also are spread out over at least half that range. Furthermore, there are unharvested fields whose mean is near 175. Obviously there are other important factors. A close inspection of the data indicates that geometry (i.e., relation of sun and field to the scanner) is a very important factor. In fact it appears to be the most important single factor contributing to the spread of the data in Fig. 4.4.2.1. To verify this contention in a statistical sense is beyond the scope of this investigation.

Observation 2

The partitions are roughly orthogonal to the mean axis. This is in accordance with the results of the previous section.

Observation 3

The data does not appear to have any distinct clusters even when the "metric properties" of the Divergence are taken into consideration (i.e., partitions roughly at right angles to the mean axis). This is disappointing in that one

would hope that at least harvested and unharvested wheat would tend to be rather distinct. The influence of geometry and other factors appears to be great enough to obscure such clusters at least in this particular channel.

Observation 4

The mode centers change considerably when the method of grouping is changed. The changes are largely changes in the standard deviation rather than changes in the means; with progressively tighter mode centers as grouping goes from S to A to P. Since the partitions in the vicinity of the mode centers are controlled primarily by the means, the partitioning of the space is not greatly influenced by the grouping method, at least in the vicinity of the data.

Fig. 4.4.2.2 shows the grouping arrived at when two channels (11 and 12) are used to cluster the wheat Training Acres. The curves shown are simply for the purpose of indicating the grouping and are not equi-distance curves. In fact since the parameter space is five dimensional an equi-distance "contour" is in fact a five dimensional surface and cannot be shown as a single contour on a two dimensional projection. Note that the grouping of the fields is the same for sample- and average-grouping. The mode centers are however located at different points in the parameter space even though this is not true for the particular projection of the parameter space shown in Fig. 4.4.2.2. (i.e., covariance matrices differ).

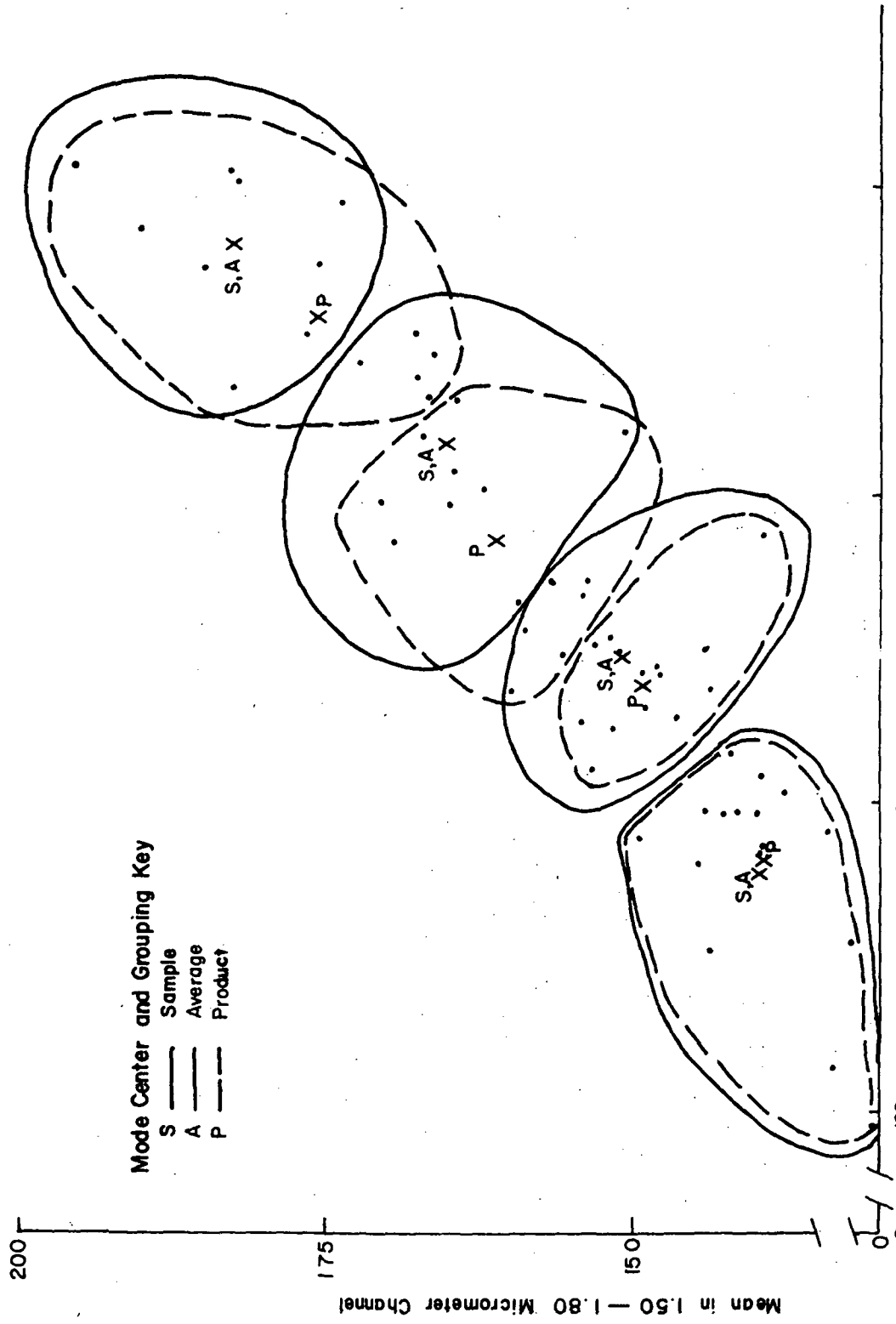


Figure 4.4.2.2 Parameter Space Clustering of Wheat Training Acres Using Divergence and Channels 11 and 12.

Since it appears that the elements of the covariance matrix are not of great significance in determining the clusters obtained it should be possible to roughly visualize clusters when the parameter space is projected onto the axis of the means as shown in Fig. 4.4.2.2. This tends to be true although the various curves in Fig. 4.4.2.2 tend to obscure any clusters that the eye might perceive. If only the data points in Fig. 4.4.2.2 are plotted, and visually grouped into four groups, the resultant groups are very similar to those achieved by GRPSAM with sample- and average-grouping.

The experiments required to obtain Fig. 4.4.2.1 and Fig. 4.4.2.2 were repeated both for the JM Distance and the SF distance. For the one channel case the partitioning curves for both the JM distance and SF distance tended to be more nearly orthogonal to the axis of the means and not as curved. The curves for the SF distance showed greater variability with grouping method than those for the Divergence while the JM distance curves showed less variability.

Finally all thirteen channels were used to cluster the wheat Training Acres using the JM distance and product-grouping. The grouping was considerably different from that obtained when only one or two channels were used. There were 4, 23, 12 and 20 acres in subclasses 1 to 4 respectively. LARSYSAA was used to obtain 13 channel histograms for these subclasses. These are shown in Fig. 4.4.2.3. Subclass 1 is very multimodal. In fact all of the 4 fields are distinctly

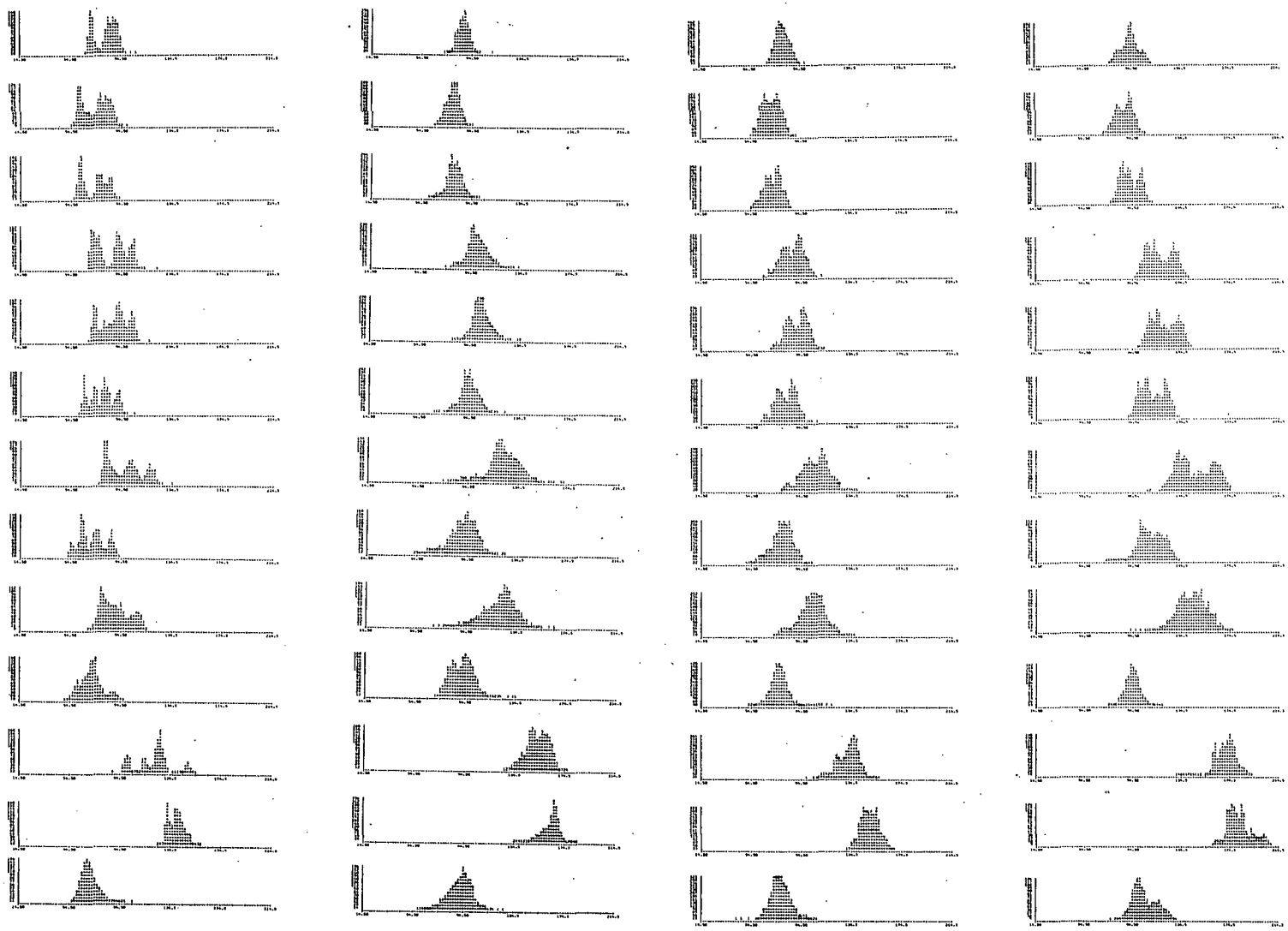


Fig. 4.4.2.3 Histograms for Wheat Training Acres Clustered by GRPSAM (13 Channels, JM Distance and P Grouping). Subclasses 1 Through 4 From Left to Right. Channels 1 Through 13 From Top to Bottom.

visable in some channels. Some of the other classes exhibit some bimodality in some channels. It is apparent that if all traces of multimodality were to be removed the number of subclasses would have to be increased considerably.

It is worth emphasizing at this point that the examples presented above are for the sole purpose of obtaining a deeper understanding of the distances and grouping methods considered. These examples do not form the basis of judging the value of a distance measure or grouping method. In summary there are two principle results. The first is the relative insensitivity of the distance measures to the covariance matrix; the second is that because of this insensitivity the mode centers obtained by different grouping methods differ largely only in covariance matrix.

4.4.3 Evaluation of Grouping Methods and Comparison of Maximum Likelihood and Minimum Distance Classification

Now that the preliminaries of parameter space clustering have been discussed the main problem of Section 4.4, namely, that of evaluating GRPSAM, can be considered.

Previously it was mentioned that the criterion to be used in comparing procedures, etc., is to compare the experimentally observed error rates for the procedures, etc., under consideration. This means that an experiment must be designed in which the various parameters of interest can be varied and their effect on classification accuracy determined. In particular, the distance measures and grouping method are

specific parameters of interest. Apart from evaluating different distance measures and grouping methods the value of parameter space clustering as a technique to assist in subclass definition for vector by vector and sample classification schemes is of prime importance.

It seems advisable to clarify the conditions under which parameter space clustering should be useful. We do this in terms of an agricultural example. It is of course clear that parameter space clustering is a parametric technique (in our case Gaussian). In the agricultural case if some care is exercised in defining training field boundaries it is usually possible to obtain reasonably homogeneous samples. In terms of subclass definition this means that the number of subclasses is at most equal to the number of training fields and classifications could be performed on this basis. In terms of processing time it is of course essential to reduce the number of subclasses to the lowest practical number. Thus if two training fields are spectrally identical it is surely desirable to treat them as one subclass. It is in this context that GRPSAM should be of assistance in that potential subclasses can be combined as long as all subclasses remain spectrally separable.

The factors discussed in the previous two paragraphs formed the basis of devising an experiment to evaluate GRPSAM, and to determine the relative value of the different distance measures and grouping methods. As mentioned earlier

a crop yield study had been carried out utilizing June'70 multispectral scanner data from flight lines 21, 23 and 24, and that in this study the randomly selected Training Acres of Table C.4 had been used for training. The test fields used for the yield study are the Standard Test Fields of Tables C.1, C.2 and C.3.

Part of the objective of the yield study was to use the Training Acres, which were selected on a random basis from all three flight lines, to generate one set of statistics suitable for classifying all three flight lines into four main classes. Yield predictions were then based on these classifications. The main classes considered were wheat, corn, soybeans and other.

It is apparent that by classifying the flight lines used in the yield study with both PERFIELD and LARSYSAA, using subclasses defined by GRPSAM, an evaluation of GRPSAM as an aid in subclass definition is possible. By performing such classifications for various distance measures, and grouping methods, the effect of these parameters can be determined. Finally by comparing the LARSYSAA classifications obtained in the yield study with those of the present study it is possible to reach some conclusions regarding the relative performance of parameter and observation space clustering. Such a comparison is legitimate since the objectives and constraints of the two approaches are essentially the same. Actually the constraints of the

present study are slightly different in that some slight modifications of the training set is necessary. Some of the acres on the yield study were in fact only partial acres. In this study it was decided not to use any partial acres because every acre is originally treated as a possible subclass, and it was felt the number of vectors in most partial acres is too small for the estimation of 13 channel statistics. In fact the number of vectors in a full acre (121) is marginal. Also since GRPSAM required statistics for each acre, and since LARSYSAA can only handle a maximum of 60 classes, some of the 65 full wheat acres were discarded. Consequently, for this study the Training Acres consisted of 59 wheat acres, 44 corn acres, 23 soybean acres and 46 other acres. This set differs slightly, though not significantly, from the set used in the yield study.

To achieve the objective of evaluating GRPSAM the original intention was to carry out PERFIELD and LARSYSAA classification of all three flight lines on the basis of statistics obtained by clustering the Training Acres with each distance measure (i.e., Divergence, JM distance, and SF distance) and each grouping method (i.e., Sample Average and Product) available in GRPSAM. These intentions were modified during the course of the experiment as a consequence of some of the experimental results. Specifically two changes were made. The SF distance was dropped from consideration and a fourth grouping method was added. The rational behind these changes is described in the sequel.

The SF distance was dropped from consideration because in comparison with the JM distance and Divergence it was exceedingly slow computationally. The implementation of the SF distance in GRPSAM is essentially based on the expressions given by Swain and Fu[29] as contained in Appendix A. This form is simply not competitive timewise with the JM distance and Divergence. The alternative form derived in Appendix A and given in Table 2.4.3 is competitive but unfortunately was not known at the time the experiment was performed. By the time the alternative expression for the SF distance was derived a considerable body of data had been collected which suggested that in practice the choice of distance is not exceedingly critical, consequently, no attempt was made to perform the SF portion of the experiment.

With regard to the added grouping method partial experimental results suggested that a grouping method, which had not originally been included in GRPSAM, might yield better performance (i.e., classification accuracy). GRPSAM was modified to include this grouping method. Specifically the experimental evidence suggested that during clustering the mode centers should be "tight" whereas once the grouping has been established the samples should be combined using a grouping method that leads to broader statistics. The extreme approach, within the limits of the grouping methods provided in GRPSAM, would be to compute the final statistics using sample-grouping on the basis of the clusters obtained

with product-grouping. We refer to this grouping method as product-sample-grouping (PS grouping). To facilitate the investigation of this grouping method GRPSAM was modified so that PS grouping could be specified. Average-sample-grouping was also provided at the same time but has not been used. Note the statistics generated by GRPSAM for PS grouping, are identical to those obtained when LARSYSAA is used to compute statistics on the basis of the fields grouping arrived at by GRPSAM using product-grouping.

As a consequence of the modifications just mentioned the experimental results we described involve two distance measures (Divergence and JM distance) and four grouping methods (Sample, Average, Product and Product-Sample). The procedures followed and the various options selected are shown in flow chart form in Fig. 4.4.3.1. The organization of this flow chart is based on the method of describing experiments given in Table 4.1.

The first task in conducting the experiment is the task of determining the number of subclasses. The procedure followed is to use GRPSAM with the JM distance and sample-grouping to cluster the acres for each class individually into subclasses. Using only the even numbered channels the fields for each main class are clustered into each of 2, 3, 4,...,10 subclasses. The separability tables are then examined with the objective of determining the "best" number of subclasses for each class. Both minimum

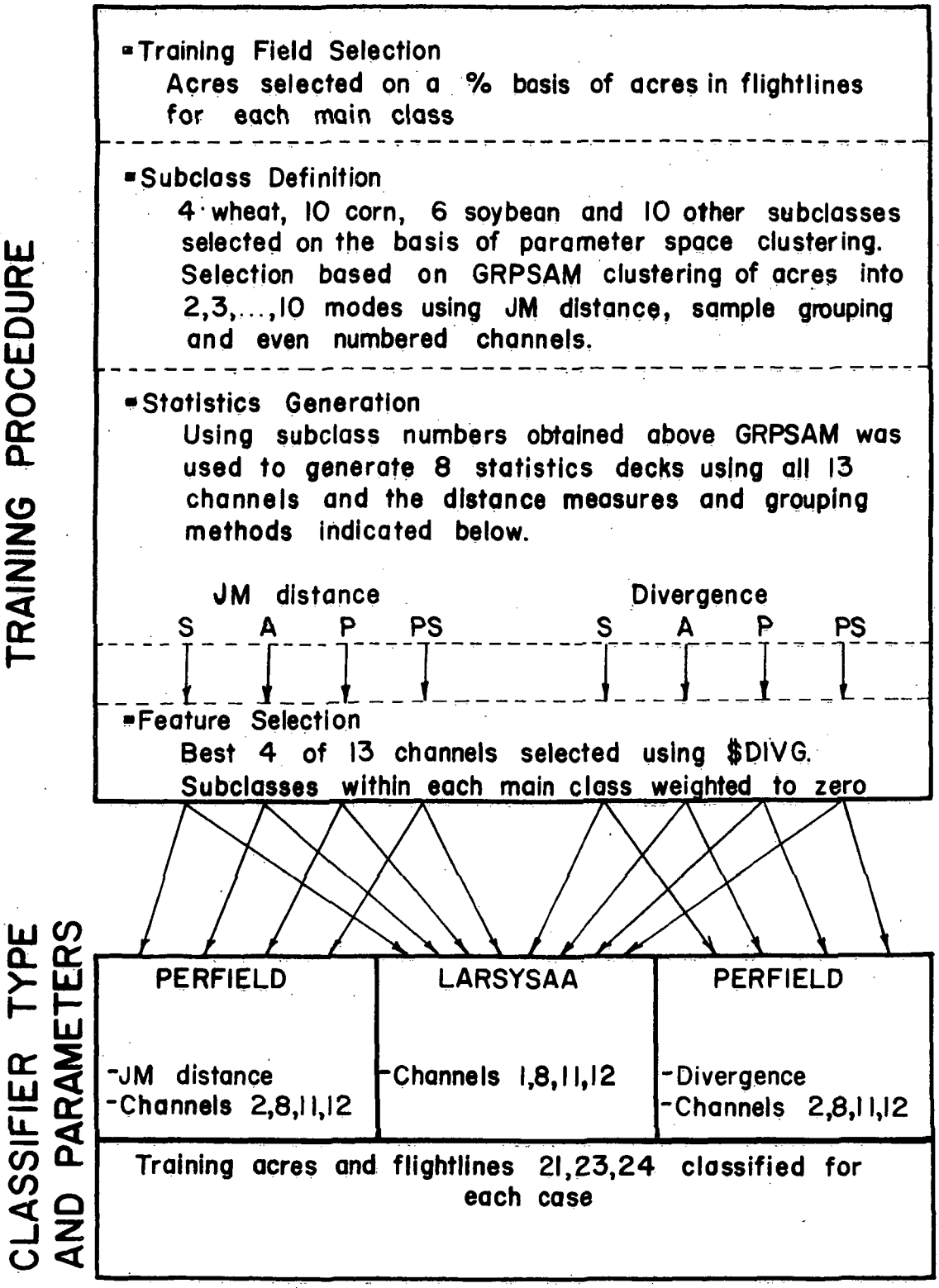


Figure 4.4.3.1 Flow Chart Showing Organization of Experimental Procedure for Evaluating GRPSAM.

pairwise separability and average pairwise separability are examined in an attempt to establish the "best" number of subclasses. Unfortunately neither of these indicators seems to give a clear indication of the appropriate number of subclasses. To demonstrate the problem the minimum pairwise separability, and average pairwise separability are plotted in Fig. 4.4.3.2 as a function of the number of modes. Although these indicators do not give a decisive answer regarding the best number of subclasses, they are of some value in selecting the number of subclasses. Other factors must also be considered. For example, since wheat would be expected to be fairly separable from other vegetation the number of wheat subclasses need not be too large. Considering such factors and recalling that the maximum number of subclasses that PERFIELD can handle is 30 it was decided to use 4, 10, 6 and 10 subclasses of wheat, corn, soybeans and other respectively.

Note that from Fig. 4.4.3.1 only one distance measure and one grouping method are involved in defining the number of subclasses. Since apparently no real indication as to the number of subclasses results from the method described, it appeared that no purpose would be served to repeat this work for various distance measures and grouping methods. Furthermore, for comparative purposes it is not essential anyway. In essence the question reduces to one of finding the best grouping method and distance measure

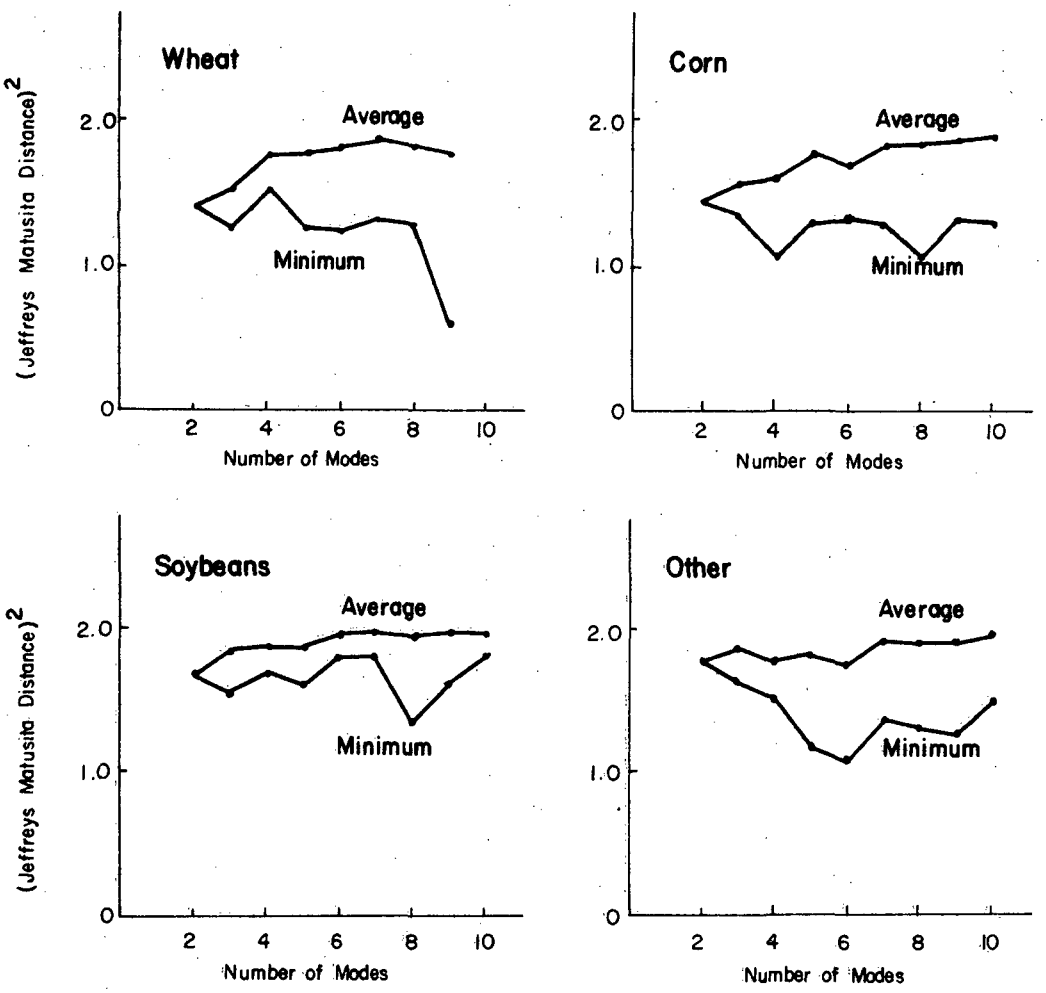


Figure 4.4.3.2 Average and Minimum Class Separability vs Number of Modes for Clusters Obtained with GRPSAM Using JM Distance and Sample Grouping.

given the number of subclasses.

With the number of subclasses established GRPSAM is used to cluster the samples and generate a statistic deck for each combination of distance and grouping method using all 13 channels. It is important to recall that each main class is clustered individually. This means for example that samples from corn and soybeans are never clustered simultaneously. It also means that for each combination of distance measure and grouping method four statistics decks are generated; one for each main class. These decks are merged into a single statistics deck suitable for use in LARSYSAA and PERFIELD. All 6 field groupings acheived in this manner are indicated in Appendix C Table C.4. There are only 6 rather than 8 groupings as P and PS grouping always result in the same field grouping.

In the next step each merged statistics deck is processed by the LARSYSAA feature selection processor $DIVG with the objective of selecting the best 4 of the 13 channels for classification purposes. The decision to use four channels was based on the fact that four channels were used in the yield study. To enable comparison of results four channels were also used in the present study. In utilizing $DIVG the weights between all subclasses in a class were set to zero. Consequently the divergence between subclasses within a class does not affect the feature selection process.

The Training Procedure used for this experiment will also be used for a number of other experiments. A concise way of referring to this particular training procedure is required. Using the method of describing an experiment outlined in Table 4.1 we note that to describe a Training Procedure it is necessary to indicate the training fields, describe the subclass selection procedure, and describe the feature selection procedure. The method used is indicated by an example. Thus, JM-PS ($DIVG) training means that subclasses were defined with the aid of GRPSAM using the JM distance and PS grouping; and that feature selection was on the basis of $DIVG. The training fields are understood to be the Training Acres and the number of subclasses are understood to be 4, 10, 6 and 10 for wheat, corn, soybeans, and other respectively. Neither of these last two factors are reflected in the notation as both factors remain fixed in all the work reported.

To keep the number of variables that effect performance as small as possible, it is obviously desirable to utilize the same channels for all classifications, provided this is at all reasonable. There was no one feature set that was clearly the best in all cases, but there were a number of sets that consistently showed up very well so that any one of about 4 or 5 features sets could have been used for our purpose. In all of the eight cases essentially all of the more optimum feature sets contained channels 8,

11, 12. The fourth channel tended to vary with the particular statistics deck with channels 1, 2, 4, and 5 frequently showing up very well. Typically one would expect performance to vary only slightly if three channels are held fixed and the fourth channel is chosen from amongst the more optimum remaining channels. For this reason selecting one set of channels for all classifications was judged to be a reasonable procedure. Channel 2 was chosen as the 4th channel because the minimum pairwise Divergence was frequently higher for channel 2 than for the other competing channels.

Using channels 2, 8, 11, 12 the necessary classification as indicated in Fig. 4.4.3.1 were performed. The results of these classifications are shown in Fig. 4.4.3.3 to Fig. 4.4.3.6. The overall training performance is shown in Fig. 4.4.3.3 while Fig. 4.4.3.4 displays the training performance by class. The test results, which represent an average over three flight lines, are shown in Fig.'s 4.4.3.5 and 4.4.3.6 for overall test performance and test performance by class respectively. The classifications were, of course, carried out using both PERFIELD and LARSYSAA respectively. In the Figures the terms sample classifier and vector classifier identify the PERFIELD and LARSYSAA results respectively. The distance measure used to group the training fields is also shown in these figures. For the PERFIELD classifications the same distance measure was used

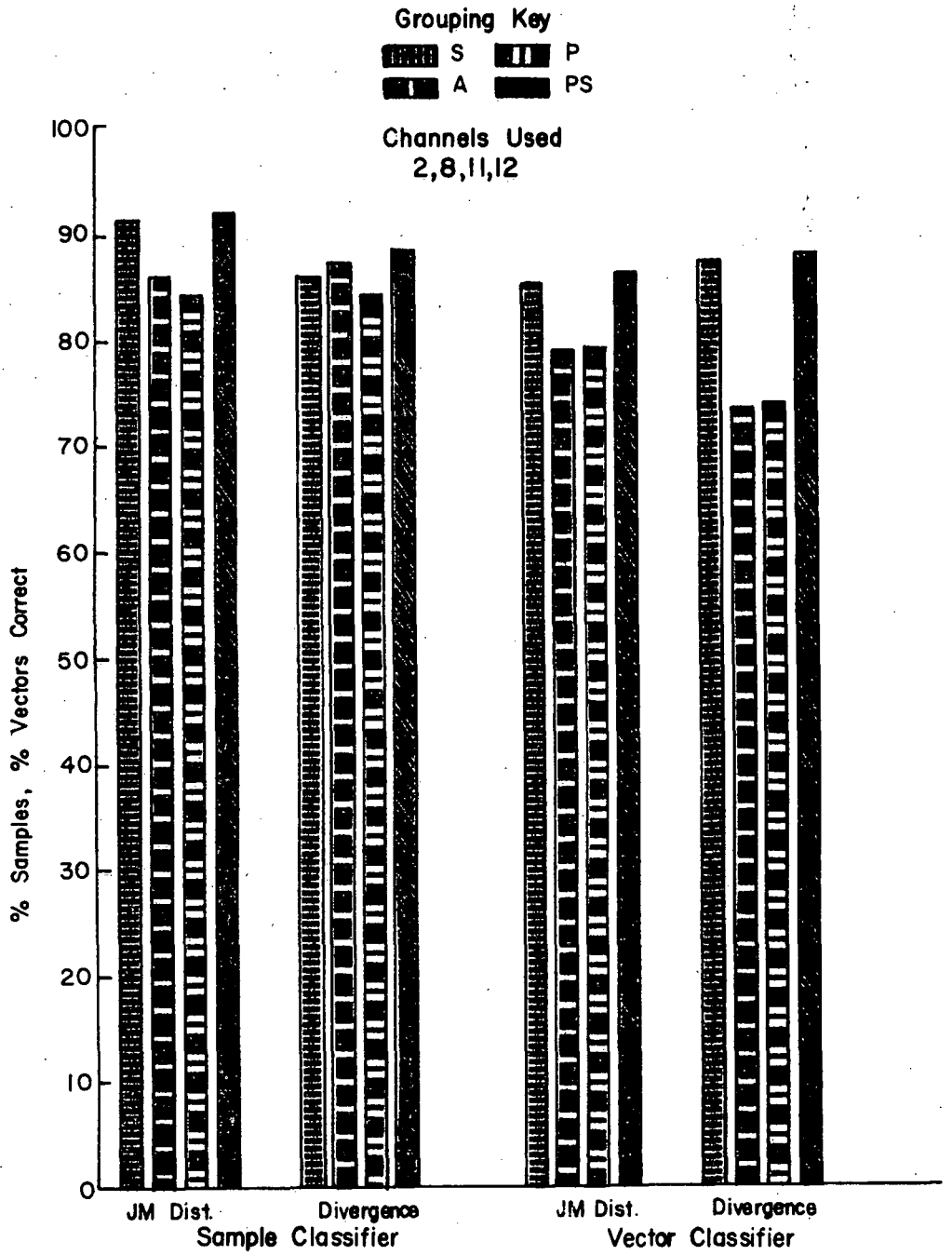


Figure 4.4.3.3 Effect of Grouping Method, Distance Measure, and Classifier Type on Overall Training Performance.

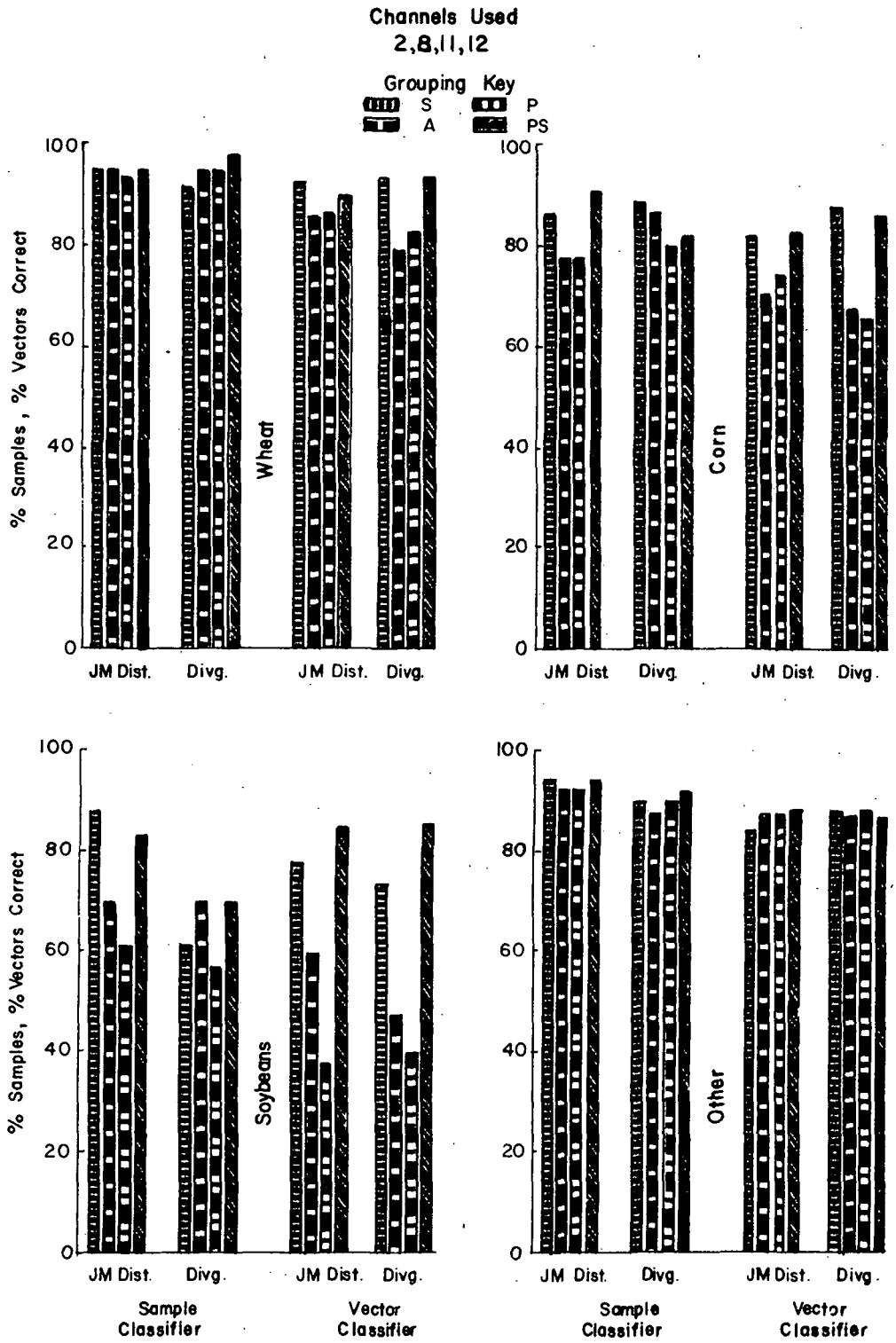


Figure 4.4.3.4 Effect of Grouping Method, Distance Measure and Classifier Type on Training Performance by Class.

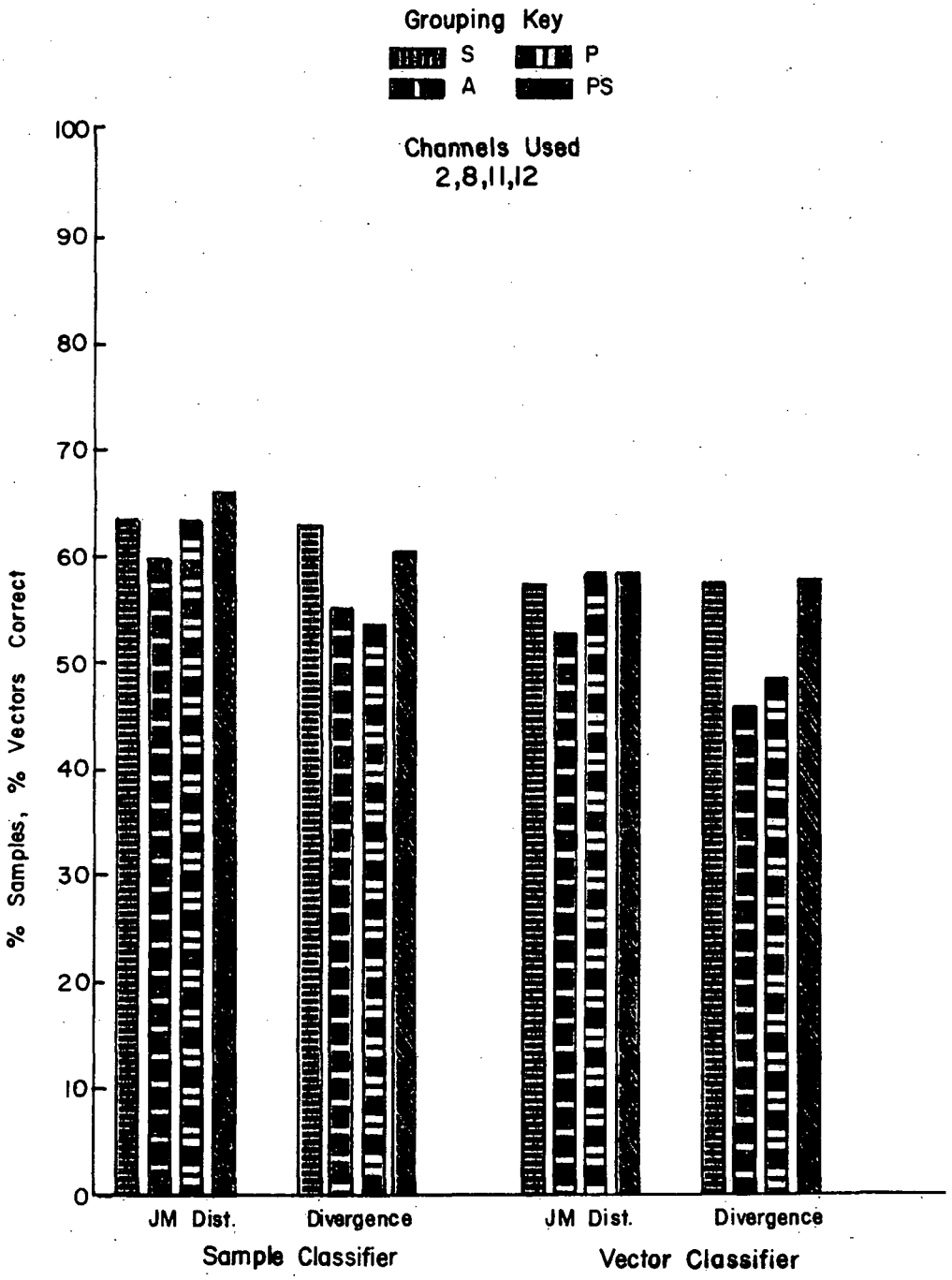


Figure 4.4.3.5 Effect of Grouping Method, Distance Measure, and Classifier Type on Average Overall Test Performance.

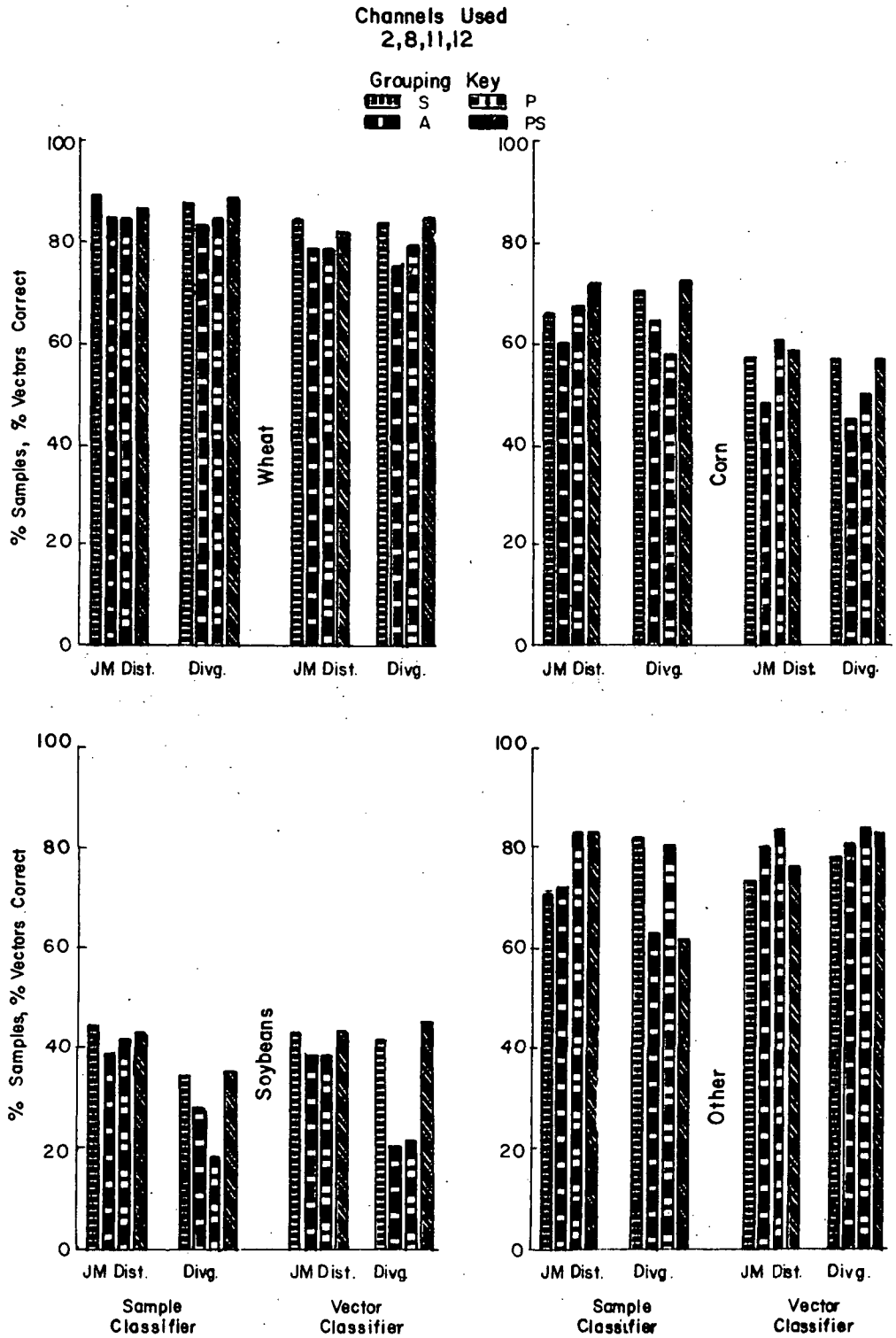


Figure 4.4.3.6 Effect of Grouping Method, Distance Measure, and Classifier Type on Average Test Performance by Class.

to classify the data as was originally used to group the training fields. The PERFIELD results therefore show the relative value of utilizing JM distance in the whole system (i.e., clustering and classification) as opposed to the Divergence. In both these systems feature selection is based on the Divergence. Consequently, whatever bias existed in the experiment should favor the Divergence. Mention must also be made of the fact that the performance of LARSYSAA is given in terms of % vectors correct while that for PERFIELD is in terms of % samples correct.

A comparison of the LARSYSAA results obtained in the yield study, were observation space clustering was used, with those of the present study using parameter space clustering is given in Table 4.4.3.1. The parameter space results are those obtained with the JM distance and sample grouping. The channels used in the yield study were 1, 8, 11, 12 compared with 2, 8, 11, 12 for the present study.

In comparing the experimental results the emphasis is placed on the overall performance rather than the performance by class. The most important reason for doing this is because of the fact that it provides one single number for comparing different classifications. There is also a tendency for the overall performance by class to be "better behaved" than the class performance. Thus if the performance of one class goes up drastically at the expense of another class this effect issmoothed out in the overall performance. While most of the conclusions are based on the

Table 4.4.3.1

Comparison of LARSYSAA Results for Training by Observation and Parameter Space Clustering

CLASS	PARAMETER SPACE (Threshold = 0.5%, Channels 2, 8, 11, 12)			OBSERVATION SPACE (Threshold 0.5%, Channels 1, 8, 11, 12)		
	Number of Subclasses	% Vectors Correct Train	Test	Number of Subclasses	% Vectors Correct Train	Test
Wheat	4	94.9	89.6	7	>98.0	86.3
Corn	10	86.4	66.4	9	>98.0	61.1
Soy-beans	6	87.0	44.4	9	>98.0	40.9
Other	10	93.5	70.7	9	>98.0	69.9
Overall	30	91.3	63.4	33	>98.0	57.3

overall performance we do not ignore the performance by class entirely and comment on some interesting anomolies. The class performance is also included for the sake of completeness. On the basis of the overall training and overall test performance the following observations can be made.

Observation 1

On the basis of average overall performance sample-grouping is usually superior to either average- or product-grouping by a few to about 12%. In those cases where product- or average-grouping are superior to sample-grouping their superiority is only a few percent.

Product-sample-grouping usually performs slightly better than sample-grouping but its advantage appears slight (1 or 2%). In an operational system considering the intuitive statistical appeal of sample-grouping, coupled with educational and interpretational problems that arise if a multitude of grouping methods are used, and noting that vector classifiers naturally use sample-grouping; it is recommended that sample-grouping be utilized as the grouping method for parameter space clustering.

Observation 2

The grouping method used appears to have a greater influence over the performance of LARSYSAA than PERFIELD. This is readily explained. Recall from the wheat acre clustering example that the grouping method

affected primarily the subclass variance with minor effects on the means. Thus regardless of grouping method the mode means are roughly the same and only the covariances differ. Classifying samples with PERFIELD, with a distance that is likewise rather insensitive to the covariance matrix, suggests that the grouping method used will not drastically affect PERFIELD performance. In LARSYSAA the discriminant surfaces can be drastically affected by the covariance matrix implying a greater sensitivity to grouping method. That the statistics are much too tight when average- and product-grouping are used can alsobe demonstrated by using a threshold in LARSYSAA. By this we mean a vector is not classified (i.e., thresholded) unless the likelihood function exceeds some predetermined number. This number is computed so that a specified percentage of vectors from a normal distribution are thresholded rather than classified. The number of points thresholded for a very light threshold (theoretically 0.5%) are of the order of 0%, 25%, 50% and 0% for S, A, P and PS grouping respectively. This suggests that average- and product-grouping produce statistics that are much tighter than the distribution of the actual vectors drawn from that class.

Observation 3

For a given grouping method the performance of the JM distance is generally slightly better than the Divergence (by up to about 10%). This tends to be true for all grouping methods for both LARSYSAA and PERFIELD and for both training and test results. The sole exceptions are that the Divergence shows up better in LARSYSAA for P and PS grouping. On the basis of these results the JM distance appears to be slightly better for clustering than the Divergence. Recall that because of feature selection a bias in favor of Divergence might have been expected.

Observation 4

The performance for PERFIELD (% Samples correct) for a given set of statistics was typically 5 to 10% greater than the performance of LARSYSAA (% vectors correct) based on the same statistics. This is a smaller improvement than had been anticipated but can be understood in the light of the following two examples. The first example indicates the basis for expecting a large improvement, while the second suggests why the anticipated improvement is not realized.

In the first example consider a two class problem in which each class is represented by a single distribution function. If the distributions are sufficiently separable, such that LARSYSAA makes essentially no errors, then essentially no improvement results when PERFIELD is used.

If the two distributions are almost identical then the LARSYSAA error is in the vicinity of 50%, but for sufficiently large samples PERFIELD makes essentially no errors. It is on the basis of this result that one expects a dramatic improvement in PERFIELD performance over LARSYSAA.

In the second example consider the case discussed in Section 3.5.3 where the classes are Gaussian (with equal variance) but the means are distributed uniformly in the parameter space. For convenience assume that each class is represented by all the distributions in that class. For large separation between the parameter space densities both PERFIELD and LARSYSAA are essentially error free. For small separation of the parameter space densities (i.e., considerable overlap) assuming that ties are broken in accordance with the prior class probabilities, it is easily seen that the probability of error for LARSYSAA is about 50%. This is precisely the same as for PERFIELD. Thus in this example, for either very large or very small separation between the parameter space densities, PERFIELD offers little advantage over LARSYSAA. We summarize this discussion by stating that for data that is very easy or very difficult to analyse PERFIELD appears to offer little advantage in classification accuracy over LARSYSAA. It is data of intermediate difficulty for which the potential for increased classification accuracy is greatest.

It is important to note that a similar situation prevails in evaluating the merit of different classification

Parameters as well as different Training Procedures. Thus for example if classification accuracies are very high or very low the advantages of any particular parameter or procedure will tend to be obscured.

Observation 5

The training performance is very much greater than the test performance. This suggests that the training fields are not too representative of the test fields. Since the training fields were distributed over all the flight lines it is difficult to see how a more representative set could be chosen.

Observation 6

In performance by class the classification accuracy for the class soybeans was lowest. Usually the majority of the confusion was between corn and soybeans although some confusion also existed between other, and corn and soybeans. It is possible that the number of soybean subclasses should have been somewhat larger.

Observation 7

From Table 4.4.3.1 it is apparent that parameter space clustering is a useful technique. Although the training set classification was considerably better using observation space clustering the overall test performance (samples for PERFIELD, vectors for LARSYSAA) was 6% poorer and improvement was shown in every class. The fact that parameter space clustering is probably faster makes it that much more appealing.

Note, however, if homogeneous fields can not be defined then parameter space clustering is not applicable; but observation space clustering is not affected.

4.5 Experimental Comparison of Distance Measures

The previous section contains a comparative evaluation of the Divergence and JM distance in parameter space clustering. The evaluation of the relative merits of the two distances is based on the performance of minimum distance classifiers, which are trained on the basis of the clustering results. The same distance is used in both clustering and classification. As a consequence of this approach the results can also be viewed as a comparison of two classification systems; one based on the Divergence the other on the JM distance. They do not directly give a comparative evaluation as to which distance would perform better in only the classification phase of a minimum distance classification system, since in the experiments described training was purposely biased, supposedly in favor of the distance used in the classifier. Such bias must be avoided if the comparison is to involve the classifier only. Furthermore, the systems were compared only in the parametric case.

The question of comparing various parametric and nonparametric distance measures in the classification phase is the main topic of this section. This comparison is effectively treated in Sections 4.5.1 and 4.5.2 which respectively consider the case of many subclasses and the case

of no subclasses.

The thrust of Section 4.5.3 is slightly different. The objective of that section is to compare two methods of defining subclasses. The first method is based on random selection of training fields, which we refer to as random training while the second involves the clustering of randomly selected fields which we refer to as nonrandom training.

As before results are presented for both average overall performance and average performance by class. In interpreting the results the emphasis is again placed on average overall performance rather than average performance by class. Only test results are presented. This is largely a consequence of the fact that the training method used in Section 4.5.1 ensures that training performance is 100%. While this is not true of Section 4.5.2 or Section 4.5.3 no attempt was made to obtain the training performance for these sections.

4.5.1 Random Training Field Selection - Each Training Field Treated as a Subclass

It is convenient to describe the experimental procedure in terms of the method summarized in Table 4.1. It is apparent that to accomplish our goal of an unbiased comparison of distance measures, a fixed Training Procedure which is in no way biased in favor of any distance measure, must be used to train the classifier. The relative value of any distance measure is then established by considering the classification accuracy achieved with that distance measure.

Both the parametric minimum distance classifier PERFIELD and the nonparametric implementation LARSYSDC are used. By utilizing both PERFIELD and LARSYSDC five different distance measures can be studied and one of these can be studied in both parametric and nonparametric form. The distance measures involved are KL numbers, Divergence and JM distance in PERFIELD; KS distance, KV distance and JM distance in LARSYSDC.

To remove bias in favor of any distance measure from the Training Procedure the training fields are selected at random and the classification channels are fixed and specified apriori. In this way no known bias is introduced either in training or feature selection. Because of the random training field selection classification accuracy will be high for some classifications and low for others; in other words the fact that performance is a random variable will show up with greater clarity than is typical. One way of comparing such classifications is to perform a number of similar classifications under similar conditions, and use average correct classification as the performance index. This is the procedure adopted. The Standard Test Fields of flightlines 21, 23, and 24 provide the three sets of data on which the average performance is based. One would perhaps prefer to have a larger number of data sets over which to take averages, but it is difficult to obtain suitable data sets and the computation time rapidly becomes prohibitive.

The detailed Training Procedure adopted was to randomly select a set of training fields from the Standard Test Fields for that flightline. This was done on a "percentage basis by class" to ensure that each main class is represented and treated in a similar manner. By selecting the training fields on a percentage basis by class we mean that for a given flightline the same percentage of the Standard Test Fields for each of the classes wheat, corn, soybeans and other are used as training fields for that flightline. The classification channels were arbitrarily selected to be 1, 8, and 11.

The above approach is also ideal for studying the effect of varying the relative size of the training set. With this objective in mind three classifications are performed for each flight with the training set respectively comprising a nominal 5%, 10%, and 20% of the Standard Test Fields in that flightline. Table C.1, C.2, and C.3 which list the Standard Test Fields for flightlines 21, 23 and 24 also show the fields selected as training fields for these flightlines for each of 5%, 10%, and 20% training. Note that the fields used for 10% training are chosen so that they contain the 5% training fields. Similarily the 20% training fields contain the 10% training fields. The fields in the Standard Test Field decks that are not selected as training fields are used as test fields.

As already mentioned all classifications are based on channels 1, 8, and 11. The reason for using 3 rather than

the more commonly used 4 channels, is because the use of more than 3 channels produced some histograms that contained more bins than could be handled by LARSYSDC for the bin size used (5). In fact some difficulty is even encountered with nonrandom training (Section 4.5.3) for this bin size when only 3 channels are used. Although the bin size of 5 was arbitrarily selected it appears to be a reasonable value based on typical histograms of multispectral scanner data. Furthermore in Section 4.6.3 this choice is experimentally shown to be reasonable.

The average overall test performance and the average test performance by class is given in Fig. 4.5.1.1 and Fig. 4.5.1.2.* Recall that in interpreting the results the emphasis is placed on the average overall test performance. Table 4.5.1.1 contains the experimentally observed standard deviation in the overall test performance.

Table 4.5.1.1

Standard Deviation in Overall Test Performance. Random Training with Subclasses

% Training	Standard Deviation for Parametric Distances	Standard Deviation for Nonparametric Distances
5	6.53	3.31
10	5.87	3.90
20	2.32	4.44

*For convenience in these and subsequent Figures, the abscissa is shown as "% Samples as Training". It should be recognized that this percentage is roughly based on 100% equals 175 fields (cf Appendix C).

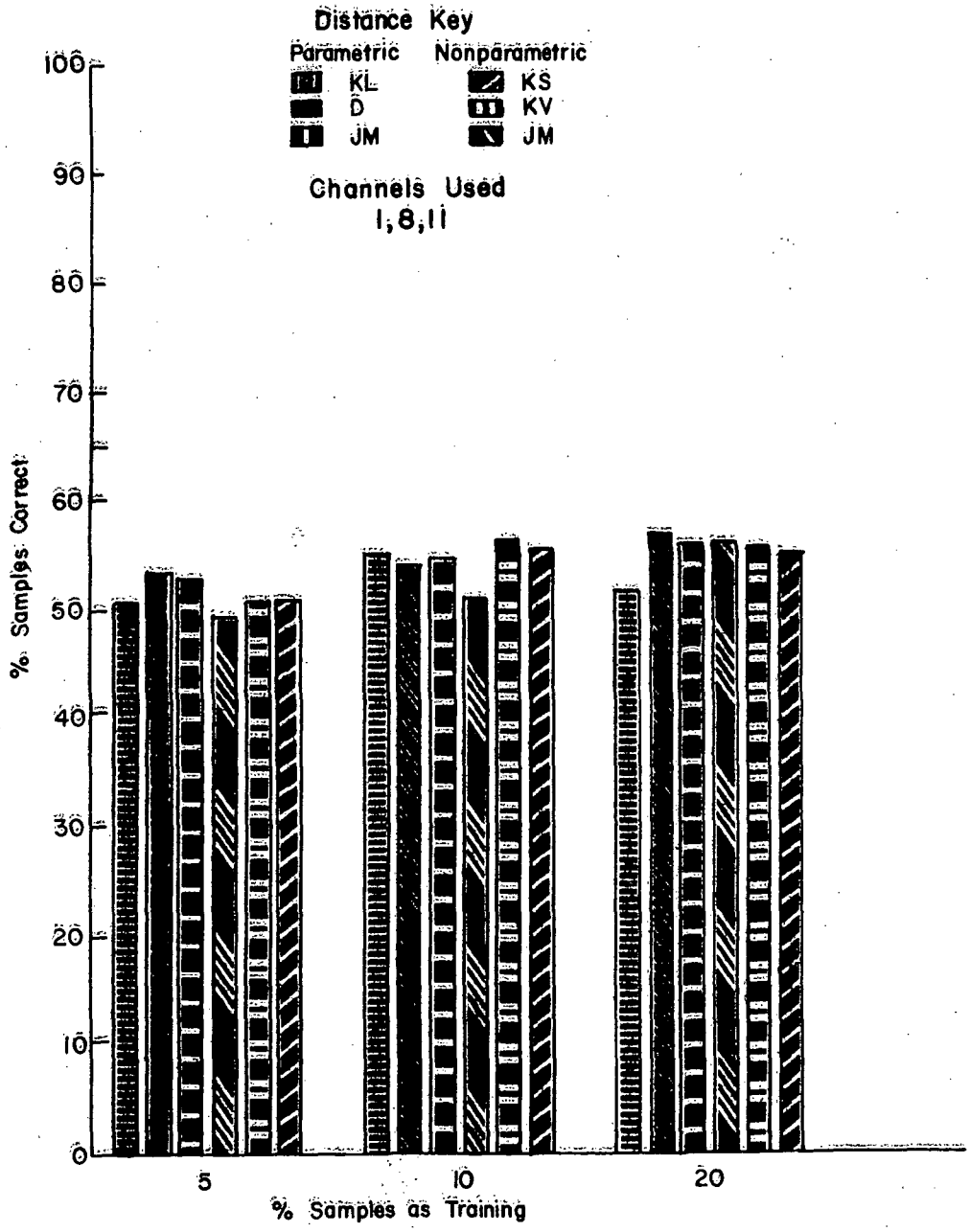


Figure 4.5.1.1 Average Overall Test Performance for Various Distance Measures. Random Training with Subclasses.

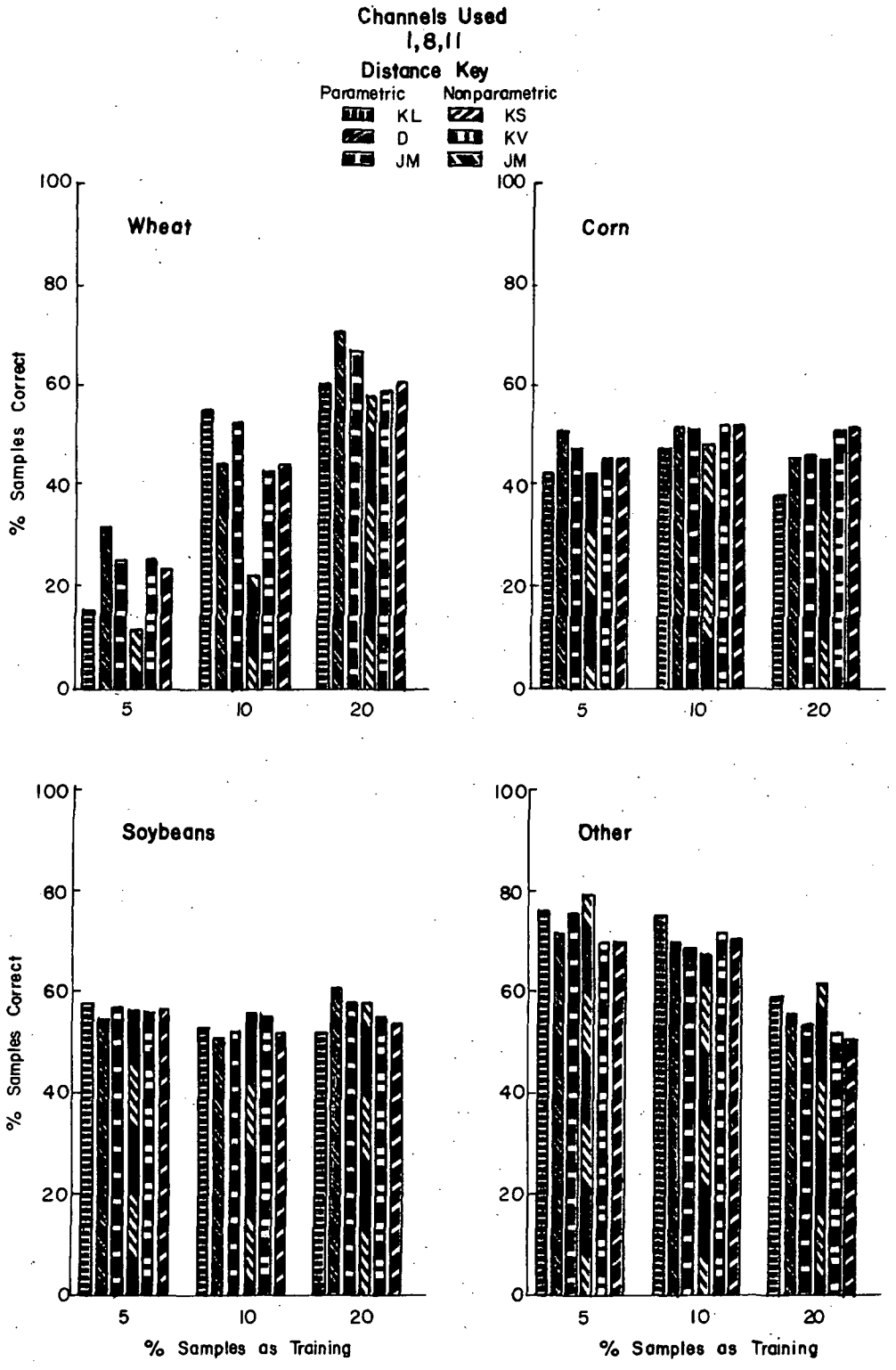


Figure 4.5.1.2 Average Test Performance by Class for Various Distance Measures. Random Training with Subclasses.

Table 4.5.1.1 is of considerable interest since it gives some indication of how radically the average test performance fluctuates. Notice that not all distances are considered separately in this Table. While the variances were originally computed individually for each distance measure, the data indicated it was reasonable to combine all the nonparametric and all the parametric distances into separate groups; especially since the intended use is primarily qualitative rather than quantitative. The advantage of this is that 9 rather than 3 classifications are used to estimate each variance, resulting in a "better" estimate.

With the aid of Fig. 4.5.1.1, Fig, 4.5.1.2 and Table 4.5.1.1 the following observations emerge.

Observation 1

Average performance is not drastically effected by the choice of distance measure. In fact on the basis of Table 4.5.1.1 it is quite likely that the variations that do show up are simply statistical variations.

Observation 2

The parametric and nonparametric classifiers using the JM distance have essentially the same average performance. This result is to be expected provided the training and test samples are reasonably Gaussian. Since each field is treated separately this condition will tend to exist.

Observation 3

Increasing the training percentage from 5% to 20% results in only a slight increase in average performance. This behavior is similar to the behavior of the simple two class univariate example considered in Section 3.5.3. There the average performance also improved only slightly as the number of subclasses increased. Thus this situation apparently carries over to the many class multivariate problem. In Section 3.5.3 it was suggested that increasing the number of subclasses is of greater importance in reducing the variance of the performance than in actually improving the performance itself. By using many subclasses one is more likely to get results near the average than if the number of subclasses is small. Table 4.5.1.1 demonstrates this property for the parametric distances. The nonparametric distances actually show a slight increase in standard deviation with an increase in the percentage of fields used as training. This behavior is largely due to the anomoulous behavior of the KS distance whose variance for 5% training was much less than for 20% training. The KV and JM distance behaved in a more normal fashion. Even so the variability in performance for nonparametric distances does not appear to be as sensitive to the number of subclasses as is the variability in performance for parametric distances. There is no known explanation for this behavior.

Observation 4

In classifying an individual flightline there were numerous instances where increasing the number of subclasses resulted in significantly poorer overall performance. This effect also prevails in a few instances even when average overall performance is considered, although in view of the standard deviations in Table 4.5.1.1, and the very slight change in average overall performance with percent training, the decrease would not appear to be statistically significant.

In light of the results of the simple two class univariate example considered in Section 3.5.3 it is not surprising that performance for an individual flight line can deteriorate when the number of subclasses is increased. (cf Section 3.5.3 Observation 4). Apparently the behavior of the many class multivariate problem is in this respect similar to the two class univariate problem. In terms of the results of Section 3.5.3 a decrease in average overall performance is not expected. As already mentioned the decrease observed for some distance measures appears to be due to statistical variation but could conceivably also be a consequence of the inadequacy of the model in Section 3.5.3.

Observation 5

The performance by class graphs (Fig. 4.5.1.2) contain a few items of interest. The main features of these graphs is that the number of subclasses increase the corn and soybean results remain essentially constant, the wheat

results improve and those for the class other deteriorate, particularily in increasing from 10% to 20% training. Behavior of this type if a single flight line is involved can again be readily explained in terms of the two class univariate problem of Section 3.5.3 (cf Section 3.5.3 Observation 4). That this behavior should occur on the average is a little more difficult to explain. While a number of explanations in terms of parameter space densities are possible the most likely one occurs only in problems involving 3 or more classes. This explanation naturally has no counterpart in the two class problem of Section 3.5.3. Explanation of the observed behavior for two class problems with different parameter space densities is also possible.

Consider the following 3 class univariate example which explains how an increase in average performance can occur in one class, while that for the other two classes remain essentially unchanged. Similar examples can also be devised to explain decreases in average performance. Assume that the 3 parameter space distributions are all uniform and that the parameter space density for class 1 is identical to that for class 2, while the parameter space density for class 3 is just barely disjoint from the class 1 and class 2 densities. It is clear that if the number of subclasses for each class is very large then on the average essentially all samples from class 3 will be correctly identified, while only about 1/2 of class 1 and class 2 samples will be correctly

identified. If the number of subclasses is reduced until class 3 is represented by only one density, while the number of densities representing class 1 and 2 are still quite large, then on the average the number of class 3 samples correctly identified will have decreased considerably, while for class 1 and class 2 there will essentially be no change. This example should make it clear that in a multiclass problem, an increase in percentage of fields used as training, may improve the performance for one class without a significant change in the performance of other classes. This example also makes it fairly clear that by appropriate adjustment of parameter space densities almost any variation of average class performance with increase in the number of subclasses is possible.

The classes corn, soybeans and wheat behave somewhat like the classes 1, 2 and 3 respectively in the above example. Thus the parameter space densities for corn and soybeans show considerable overlap while the parameter space density for wheat is somewhat disjoint. Furthermore, the relative abundance of corn, soybean, and wheat fields means that corn and soybeans are always represented by a considerably larger number of subclasses than wheat.

The above example is, therefore, a plausible explanation for the behavior of the wheat performance graphs of Fig. 4.5.1.2. A similar explanation could be devised for the class other but there is some doubt as to the correctness

of this interpretation for the class other. Due to extenuating circumstances it is likely that the decrease in average performance for the class other, with increase in subclasses, is not actually real but that the decrease is simply due to a rather drastic statistical fluctuation. This problem does not arise for the class wheat since the performance for every distance measure and every flightline showed an increase in performance.

The decrease in performance for the class other as training increases from 10% to 20% is largely due to the collapse in performance for flightline 23. For this flightline the performance for the class other decreases from the vicinity of 70% to the vicinity of 30%. Flightlines 21 and 24 do not exhibit this behavior and the results for these flightlines is virtually unchanged as the training fields increase from 10% to 20%. Since flightline 23 contains a rather small number of test fields for the class other it is actually the misclassification of a relatively small number of fields that is responsible for the decrease in class other when training is increased from 10% to 20%.

4.5.2 Random Training Field Selection - No Subclasses

The experimental procedure for this section is identical with that of Section 4.5.1 except that instead of treating each field as a subclass all the randomly selected fields for each main class are combined. Thus each class is represented by a single distribution function. Classifications

are again performed of flightlines 21, 23 and 24 using 5%, 10% and 20% of the Standard Test Fields as training. The average overall test performance and the average performance by class are given in Fig. 4.5.2.1 and Fig. 4.5.2.2 respectively. Table 4.5.2.1 shows the variance in the overall performance where parametric and nonparametric distances have again been grouped.

Table 4.5.2.1

Standard Deviation in Overall Test Performance.
Random Training with No Subclasses

% Training	Standard Deviation for Parametric Distances	Standard Deviation for Nonparametric Distances
5	4.42	3.11
10	6.12	1.60
20	8.94	2.98

When each field is treated as a subclass then the classes tend to be unimodal and symmetrical and the Gaussian assumption should be reasonably valid. Consequently, nonparametric methods have no particular advantage in this setting. By combining all the training fields into one subclass the class distributions will almost surely be multimodal and the normal assumption would not be very valid. One would anticipate that in this situation the nonparametric classifier LARSYSDC would be a better classifier than the parametric classifier PERFIELD. It was essentially this contention that prompted the investigation described in this

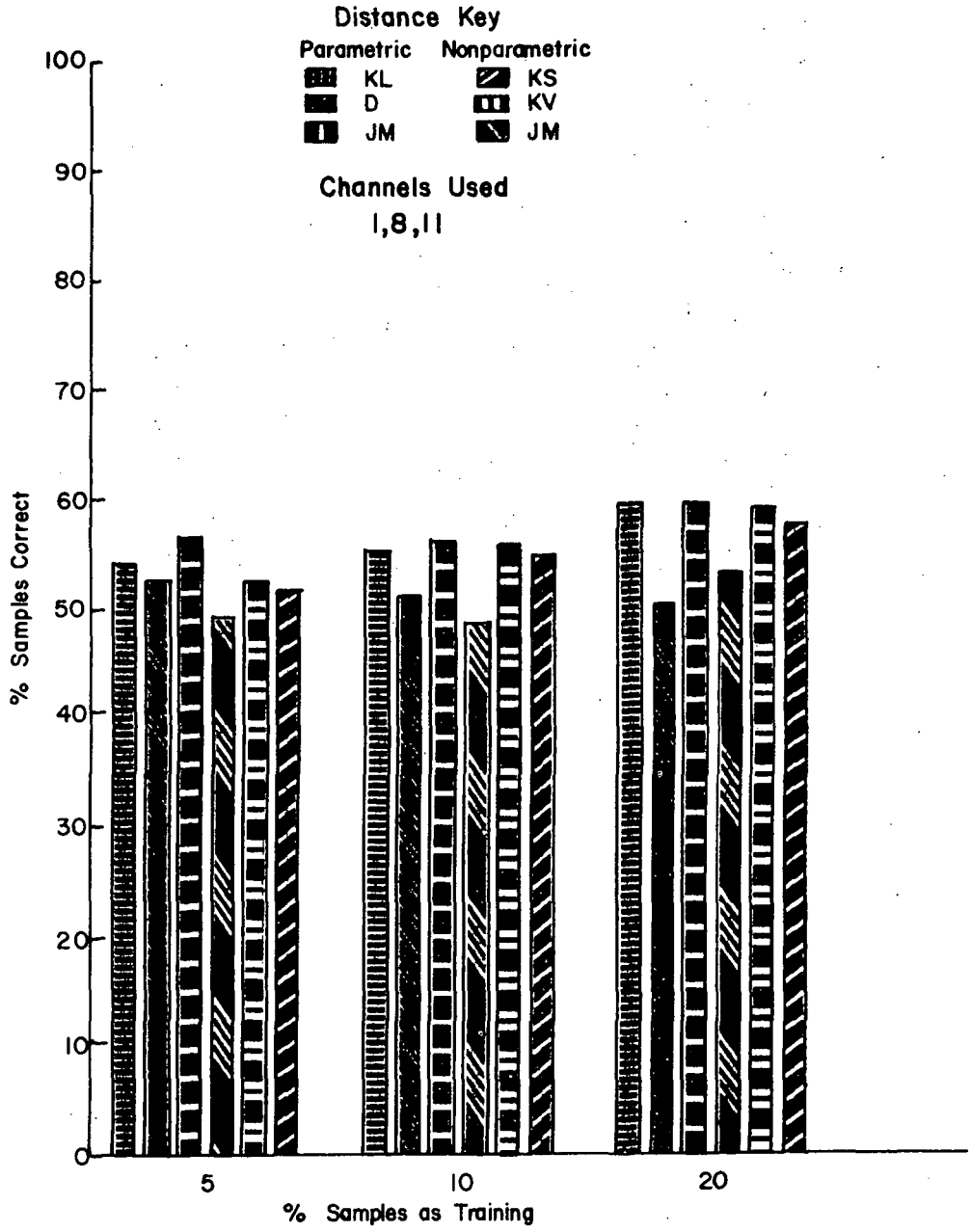


Figure 4.5.2.1 Average Overall Test Performance for Various Distance Measures. Random Training with No Subclasses.

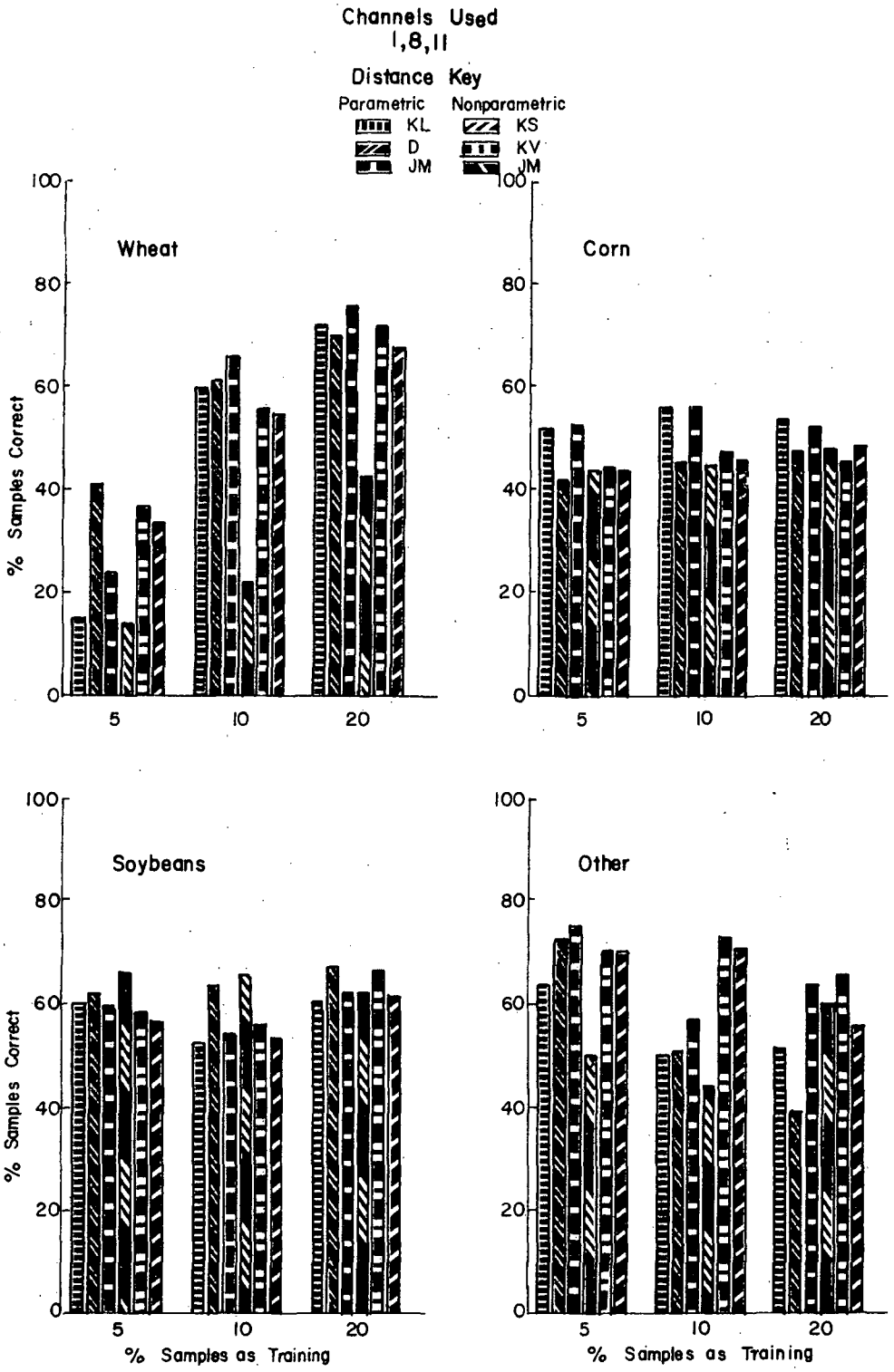


Figure 4.5.2.2 Average Test Performance by Class for Various Distance Measures. Random Training with No Subclasses.

section. The extent and manner in which these expectations agree with the experimental results is somewhat different than anticipated.

Based on Fig. 4.5.2.1, Fig. 4.5.2.2 and Table 4.5.2.1 we make the following observations.

Observation 1

Within the limits of statistical fluctuations as suggested by Table 4.5.2.1 the average performance of all distance measures is roughly equivalent although, the Divergence and KS distances appear to perform somewhat poorer than the other distances.

Observation 2

In terms of average performance the parametric classifier using the JM distance does just as well as the nonparametric version using the JM distance. The typical variance in performance is, however, much greater for the parametric than the nonparametric classifier (Table 4.5.2.1). Furthermore, the variance in performance for the parametric classifier increases as the percentage training increases while for the nonparametric classifier this quantity remains reasonably fixed. These factors are important from a classification viewpoint. They mean in effect that in performing a single classification one is more likely to obtain reasonable results with the nonparametric classifier and that for the parametric classifier the results become more erratic as the number of fields grouped together increases. If the

results for many classifications are to be averaged then the parametric classifier does just as well on the average as the nonparametric classifier.

Because of the multimodal nature of the class distributions one might expect that on the average the nonparametric classifier would do better than the parametric classifier. The basic fallacy in this reasoning is that although the class distributions are multimodal the samples to be classified are essentially unimodal. In other words the distribution of any sample to be classified is not really based on a random sample from the distribution of any class. Instead it simply tends to account for one of the modes in the class distribution. Furthermore, there is no apparent way of rectifying this situation within the constraints of minimum distance classification.

We can summarize the results as follows. For the parametric classifier better results are obtained if many subclasses are used. The result is not better in terms of performance averaged over many flightlines but in terms of the variability in performance from flightline to flightline. For the nonparametric classifier results with many and no subclasses are comparable. Therefore it is certainly advantageous to use no subclasses since computations increase directly with the number of classes.

Observation 3

Increasing the training percentage from 5% to 20% results in only a slight increase in average performance. This behavior is similar to the behavior observed when subclasses are permitted and can be explained in a similar manner (cf, Section 4.5.1 Observation 3).

Observation 4

Increasing the number of subclasses for a given distance measure quite often results in a significant decrease in performance for the classification of any flightline, and occasionally results in a small (probably not significant) decrease in the performance averaged over the three flight lines. This result is similar to the behavior observed when subclasses are permitted and can be explained in a similar manner (cf, Section 4.5.1 Observation 4).

Observation 5

The performance by class is qualitatively similar to that observed in the case where subclasses are used, except that the disparity between different distance measures is sometimes greater. In particular the KS distance appears to perform poorly. The reason for this is unknown.

4.5.3 Training Fields Grouped by Parameter Space Clustering

The objective of this section is to compare the random training procedures in the previous two sections with a training procedure based on parameter space clustering which we refer to as nonrandom training, more precisely it is

really the subclass definition that is nonrandom. In particular results using the training procedure used in evaluating GRPSAM are compared with the results of random training with each field treated as a subclass (20% training). In terms of the method of describing experiments given in Table 4.1 we are studying the effect of two Training Procedures with the distance measure as a Classification Parameter. Both the parametric and nonparametric implementations of the minimum distance classifier are again considered.

It is possible to view the case of nonrandom training as a logical extention of the case of random training where each training field is treated as a subclass. If the number of training fields is larger than the number of subclasses the system can handle, then it is logical to search for ways of combining subclasses that are sufficiently alike. Clustering in the parameter space serves this purpose. The training fields that were clustered with GRPSAM using the JM distance and PS grouping were the Training Acres of Table C.4. As before the Standard Test Fields of flightlines 21, 23, and 24 were classified with all the distance measures available in both PERFIELD and LARSYSDC using channels 1, 8 and 11. The results of these classifications together with the results obtained for 20% random training are compared in Fig. 4.5.3.1 and Fig. 4.5.3.2. The first figure compares the average overall test performance while the second compares the average test performance by class. The variance

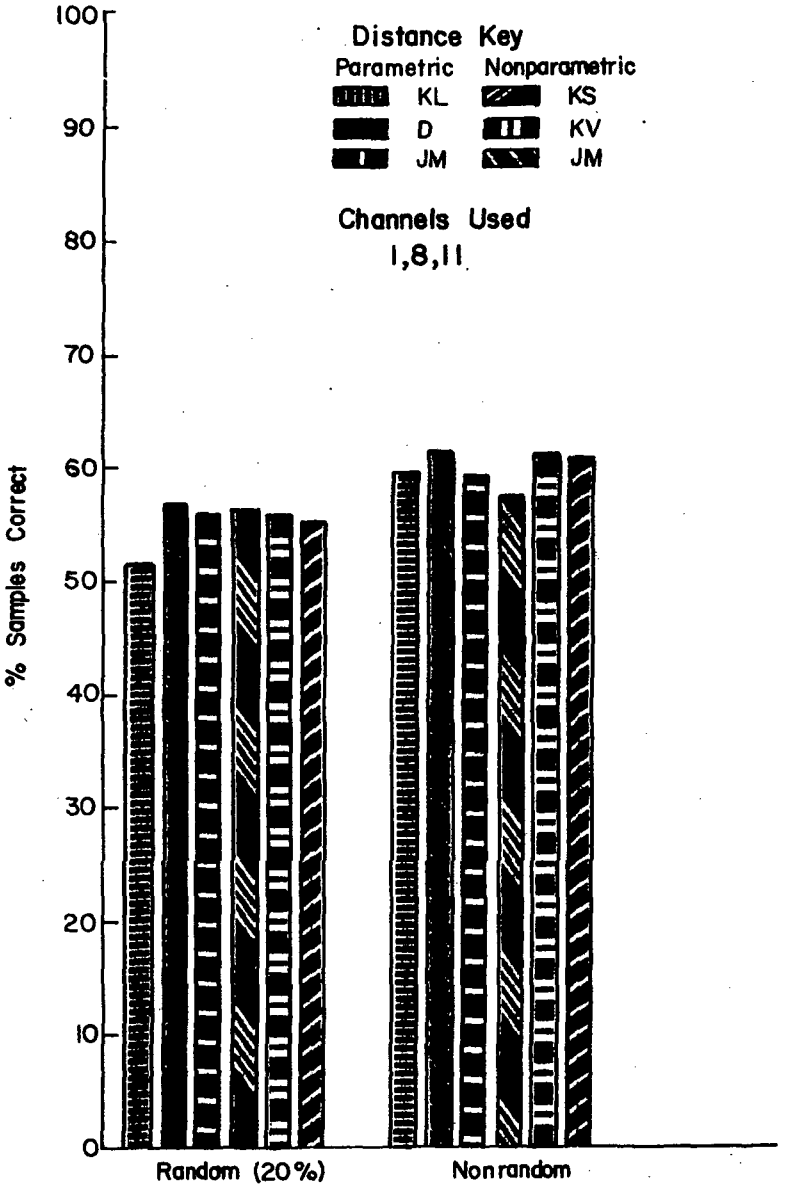


Figure 4.5.3.1 Average Overall Test Performance for Various Distance Measures. Random and Nonrandom Training.

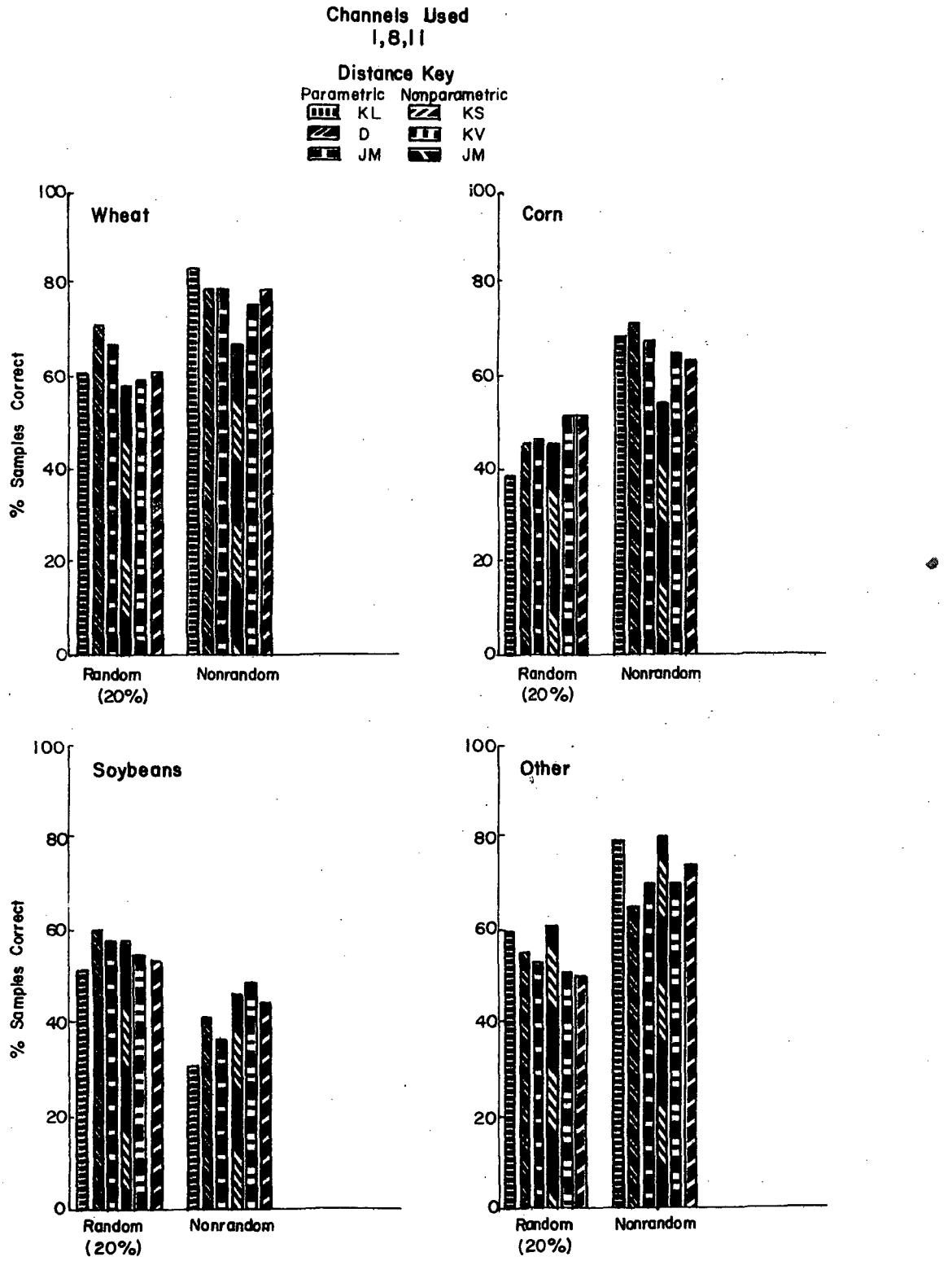


Figure 4.5.3.2 Average Test Performance by Class for Various Distance Measures. Random and Nonrandom Training.

in the average test performance for nonrandom training was 4.30 and 4.41 for the parametric and nonparametric distances respectively. The histogram bin size used in LARSYSDC was 10 for the nonrandom training results and 5 for the random training results. This difference was necessitated by the fact that for a bin size of 5 some of the training classes for nonrandom training contained more than the maximum allowable number of bins as determined by programming constraints. On the basis of Fig. 4.5.3.1 and Fig. 4.5.3.2 we make the following observations.

Observation 1

Again no particular distance measure appears to have any advantage. This was previously observed for random training and is also true for nonrandom training.

Observation 2

Average overall performance for nonrandom training is slightly better than for random training. This is perhaps to be expected since in effect a training set drawn from a larger number of fields was used. The Training Acres were, of course, also chosen on a percentage basis by class but the percentage varied from class to class with wheat being sampled much more densely than corn, soybeans and other.

In interpreting the difference between random and nonrandom Training two factors must be considered. For random training all test fields were physically disjoint from the training fields. In nonrandom training many of the

Training Acres are in fact contained within the Test Fields. This would tend to increase the nonrandom training performance. Offsetting this effect is the fact that the bin size for nonrandom training is larger which would tend to favor random training.

Observation 3

The average performance by class again shows greater variability from distance measure to distance measure than the average overall performance. Nonrandom training shows up favorably for all classes except soybeans where random training was superior. As mentioned previously in connection with the results on the evaluation of GRPSAM it is possible that the number of subclasses for soybeans should have been somewhat larger.

4.6 Effects of Some Parameters on Performance

It is of considerable interest to know how some of the Classifier Parameters affect performance. Our purpose in this section is to investigate some of the more important parameters. In terms of the method of describing problem summarized in Table 4.1 we focus our attention on determining the effect on classification accuracy of the Classifier Parameters listed in that table.

Table 4.6.1 contains a summary of the experiments performed. This table indicates not only the nature of the various studies but also depicts the range of the parameter studied and lists the section number in which the results

Table 4.6.1

Classification Parameters Studied

Study	Section	Training Method*	Classifier	Classifier Parameters and Range**
Number of Channels	4.5.1	JM-PS($SEQDIVG)	PERFIELD	Number of Channels (1 thru 13), Distance (KL,D,JM), Sample Size (121)
		D-PS($SEQDIVG)	PERFIELD	Number of Channels (1 thru 13), Distance (KL,D,JM), Sample Size (121)
		JM-PS($SEQDIVG)	LARSYSDC	Number of Channels (1,2,3), Distance (KS,KV,JM), Sample Size (121), Bin Size (5)
Sample Size	4.5.2	JM-PS($SEQDIVG)	PERFIELD	Number of Channels (2,3), Distance (JM), Sample Size (4,6,16,36,66,121)
		JM-PS($SEQDIVG)	LARSYSDC	Number of Channels (2,3), Distance (JM), Sample Size (4,6,16,36,66,121), Bin Size (5)
Bin Size	4.5.3	JM-PS($SEQDIVG)	LARSYSDC	Number of Channels (1,2,3), Distance (KS,KV,JM), Sample Size (121), Bin Size (1,5,10, 20,30)

*JM-PS($SEQDIVG)
Training Field Selection - Percent Acres by Class (i.e., Training Acres)
Subclass Definition - Nonrandom
- JM-PS clustering by class with GRPSAM (4 wheat, 10 corn, 6 soybeans, and 10 other subclasses)
Feature Selection - Based on Average Transformed Divergence with Sequential Search of Feature Combinations ($SEQDIVG)

D-PS($SEQDIVG) - As for JM-PS($SEQDIVG) except Subclass Definition based on D-PS clustering rather than JM-PS clustering.

** In a number of instances the variables were not varied through the full range for all distances specified.

are described. It is convenient to describe the portions of the experimental procedures that are common to all the studies in this section relegating to the appropriate subsections those procedures that apply only to that subsection.

To study the effect of different parameters, it is of course necessary to fix the Classifier Type and Training Procedure and then vary the Classifier Parameter of interest. The only Classifier Types considered are the minimum distance classifiers PERFIELD and LARSYSDC. The training procedure is based on clustering the Training Acres using either the Divergence or JM distance with PS grouping on a class by class basis; feature selection is via $SEQDIVG (i.e. JM-PS($SEQDIVG) or D-PS($SEQDIVG) training). The Classifier Parameters studied are number of channels, bin size and the number of vectors used to estimate the test histograms (i.e., sample size). Again wherever appropriate the various distances in LARSYSDC and PERFIELD are compared.

Results in all cases are given for both training and test fields. The training fields used are the Training Acres listed in Table C.4. The test fields are derived from the flightline 21 Test Areas given in Table C.5. Rather than list the actual test decks used we describe instead the method of deriving the test decks from the flightline 21 Test Areas. The reason for this approach is that in the sample size study 12 different decks are used. Half of

these decks are derived from the flightline 21 Test Areas and half of them from the Training Acres. It is simpler to describe the method of generating these "derived fields" than to list all the decks. To generate a derived field from an original field it is necessary to specify the number of vectors the derived field must contain. The line and column intervals of the derived field are then adjusted so that the vectors in the derived field are spread out as much as possible over the original field. For example a derived test field containing four vectors would contain the four vectors located on the corners of the original field. The objective of this rather involved procedure is to ensure that the vectors in the derived field are as independent as possible within the constraint that they must be contained in the original field.

For all the studies except the sample size study there are 121 vectors in each training and test field. The number 121 was chosen because this represents all the vectors in a Training Acre. Since flightline 21 Test Areas contain up to 900 vectors the procedure described above was used to select the 121 vectors from each Test Area to generate a derived test field. In the sample size study the same procedure was used to select "training"[+] and test fields

---

[+]The full Training Acres were in all cases used for training purposes. The word "training" is used to designate test fields derived from the Training Acres.

from the Training Acres and flightline 21 Test Areas respectively.

A comment regarding the graphs in this Section appears advisable. For most of the graphs the independent variable is discrete. For convenience in reading the graphs experimental points have been joined by straightline segments, but these segments do not have meaning except for integer values and then only those integer values that were experimentally investigated (c f Table 4.6.1).

4.6.1 Number of Channels

In discussing the experiments performed to determine the effect of dimensionality on classification accuracy it is convenient to segregate the experiments into two categories. The segregation is on the basis of Classifier Type (i.e., parametric vs nonparametric).

With reference to Table 4.6.1 it is apparent that for the parametric case classifications were performed for the three available distance measures (KL number, Divergence and JM distance) in PERFIELD. The number of channels was varied from 1 thru 13 for each of the three distance measures. Both D-PS($SEQDIVG) and JM-PS($SEQDIVG) Training Procedures were used.

In other words two sets of statistics were processed by the $SEQDIVG processor corresponding to the output from GRPSAM for JM-PS and D-PS clustering. The channels sequences obtained for these two cases were 11, 12, 8, 5,

10, 1, 2, 7, 13, 3, 9, 6, 4 and 11, 12, 8, 1, 5, 10, 2, 13, 7, 3, 9, 6, 4 and JM-PS and D-PS clustering respectively. These two sequences are really quite similar with difference occuring only near the middle of the sequence.

For the nonparametric case classifications were performed for the three distance measures in LARSYSDC (KS, KV, and JM distances). Results were obtained for the JM-PS ($SEQDIVG) Training Procedure only and the number of channels was varied between 1 and 3 except for the KS distance where no 3 channel results were obtained.

The results of the number of channels study appear in Figs. 4.6.1.1 through Fig. 4.6.1.12 with the parametric results occupying the first eight figures and the nonparametric results in the last four. Fig. 4.6.1.1 through Fig. 4.6.1.4 contain the training and test results for the parametric case where JM-PS clustering is used in the Training Procedure while Fig. 4.6.1.5 through Fig. 4.6.1.8 present similar results for the case where D-PS clustering is used. In each case overall test and training performance together with test and training performance by class account for the four figures. A similar set of figures for the nonparametric case accounts for the four nonparametric figures. For comparison purposes the performance of the parametric distance measures has also been included on the nonparametric graphs.

It is worthwhile remarking that apart from the training results presented in connection with the evaluation of

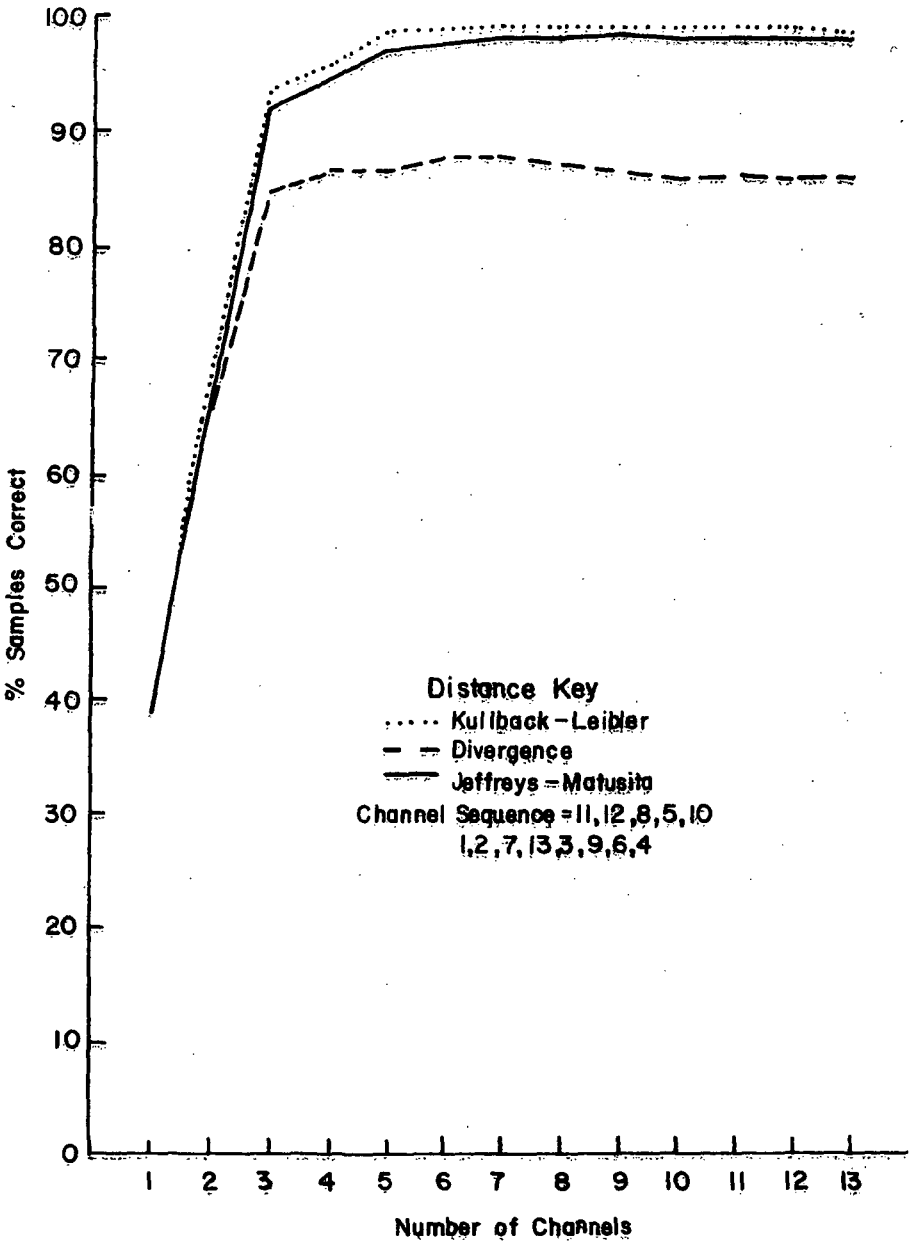


Figure 4.6.1.1 Overall Training Performance vs Number of Channels for Parametrically Implemented Distance Measures, JM-PS($SEQDIVG) Training.

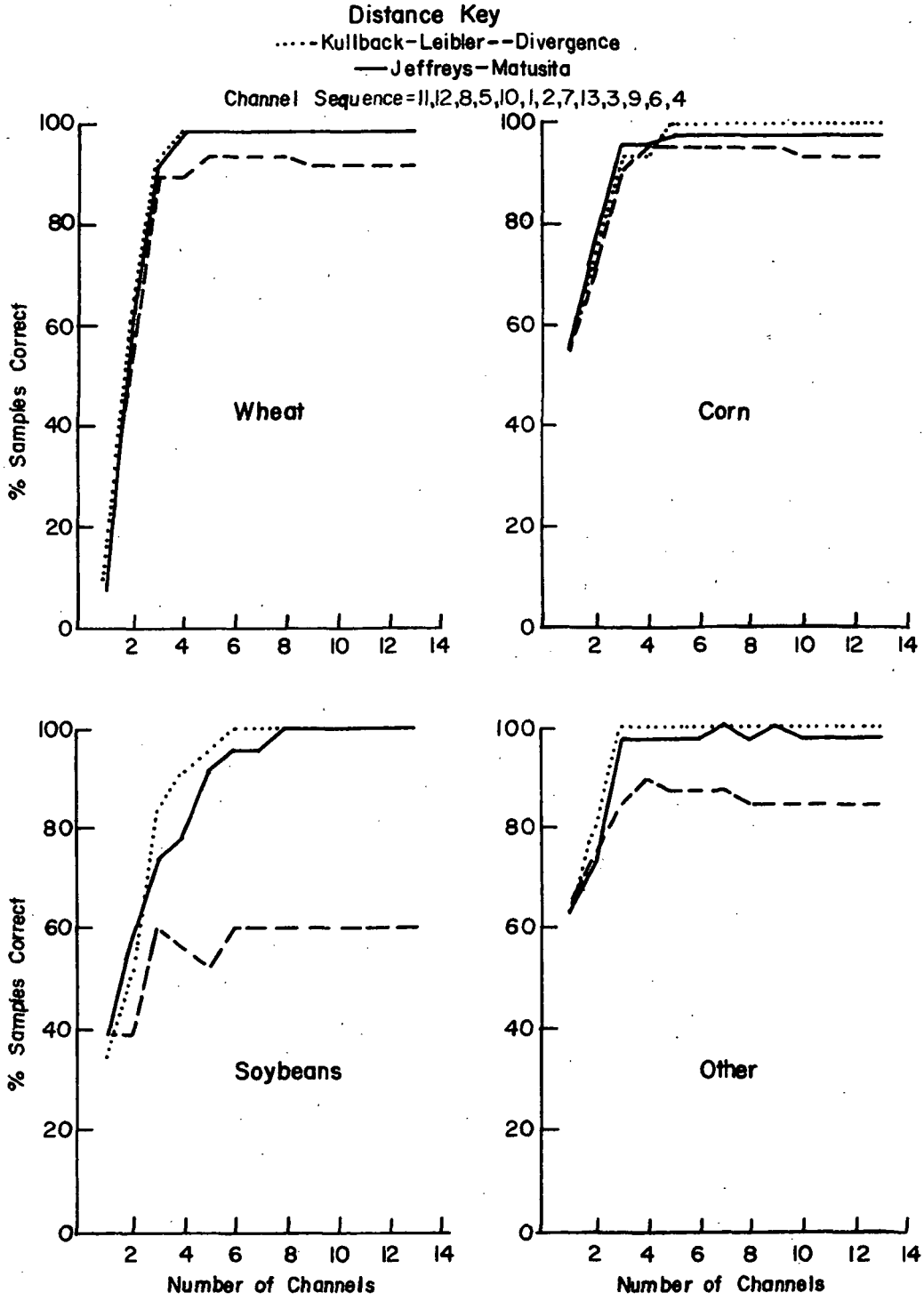


Figure 4.6.1.2 Training Performance by Class vs Number of Channels for Parametrically Implemented Distance Measures. JM-PS($SEQDIVG) Training.

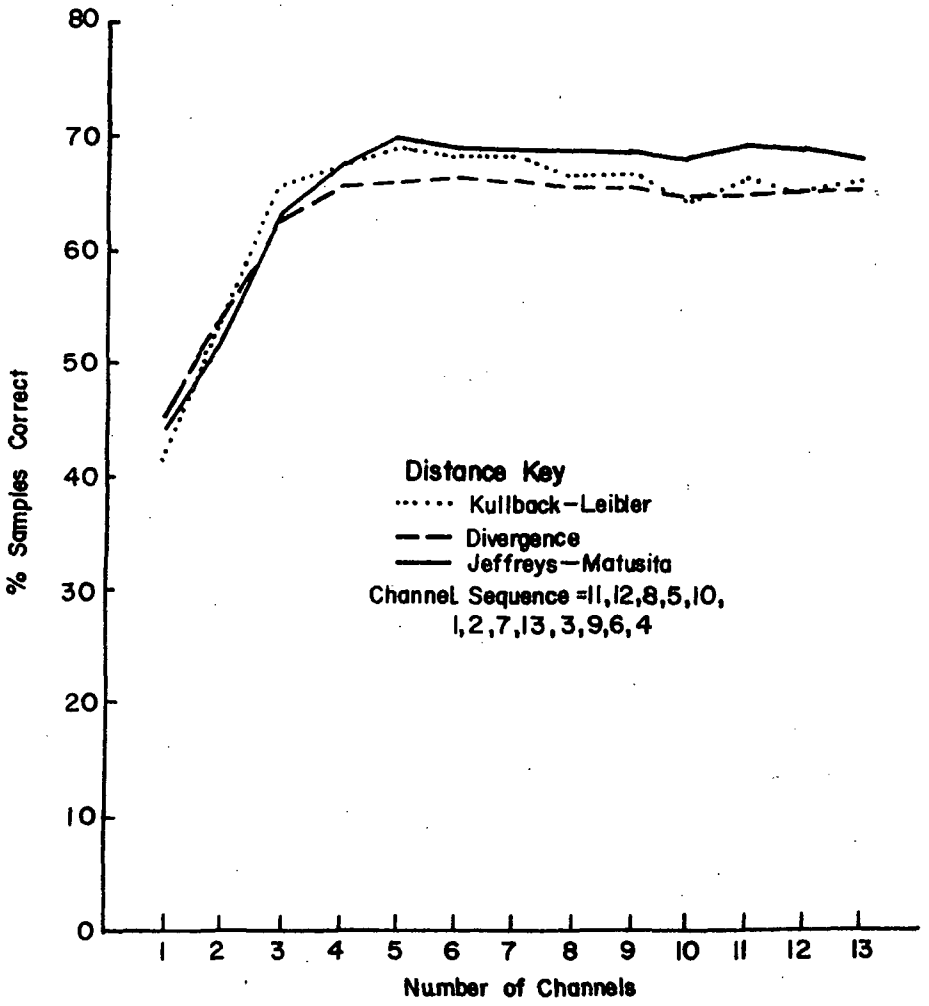


Figure 4.6.1.3 Overall Test Performance vs Number of Channels for Parametrically Implemented Distance Measures. JM-PS($SEQDIVG) Training.

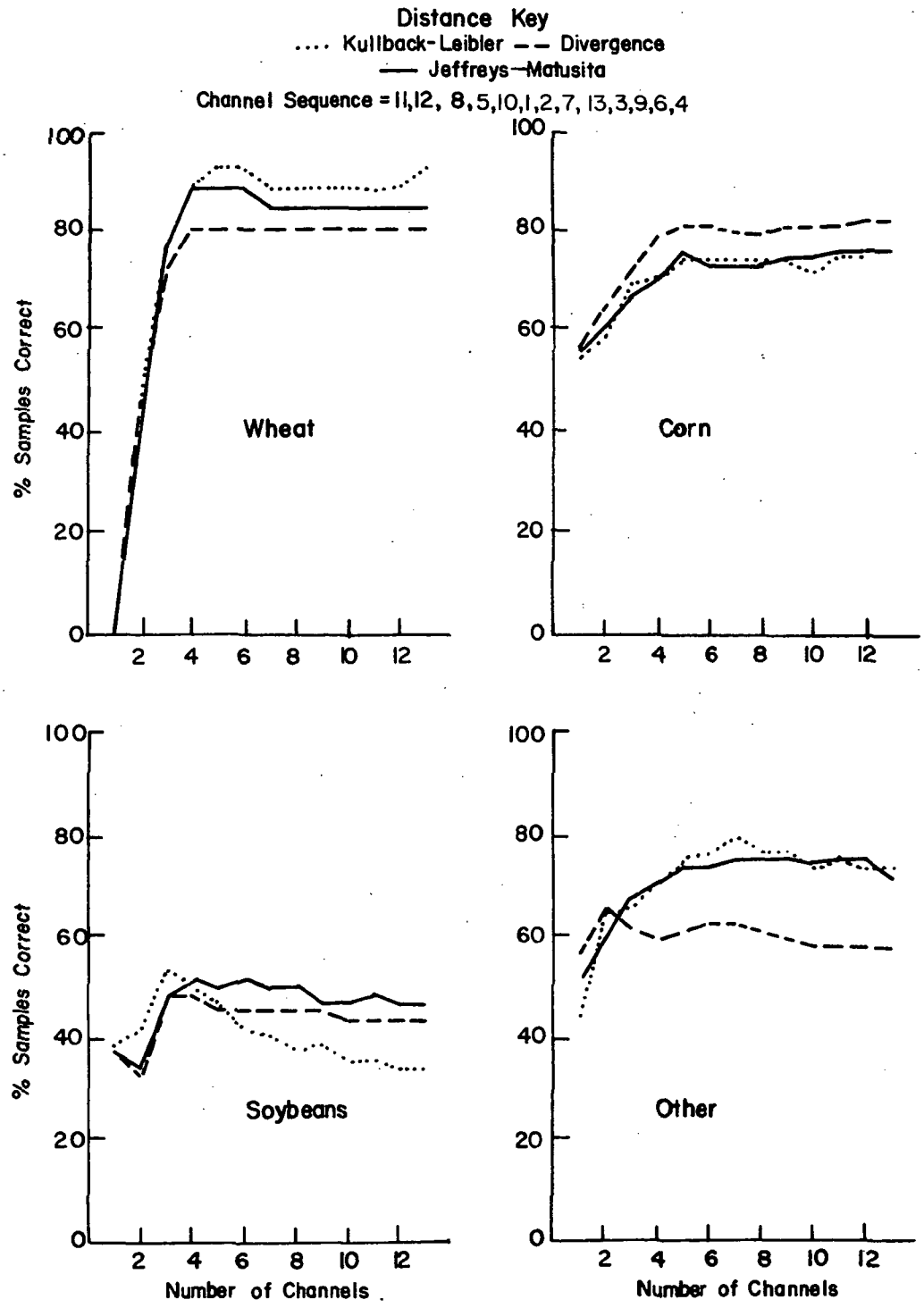


Figure 4.6.1.4 Test Performance by Class vs Number of Channels for Parametrically Implemented Distance Measures. JM-PS($SEQDIVG) Training.

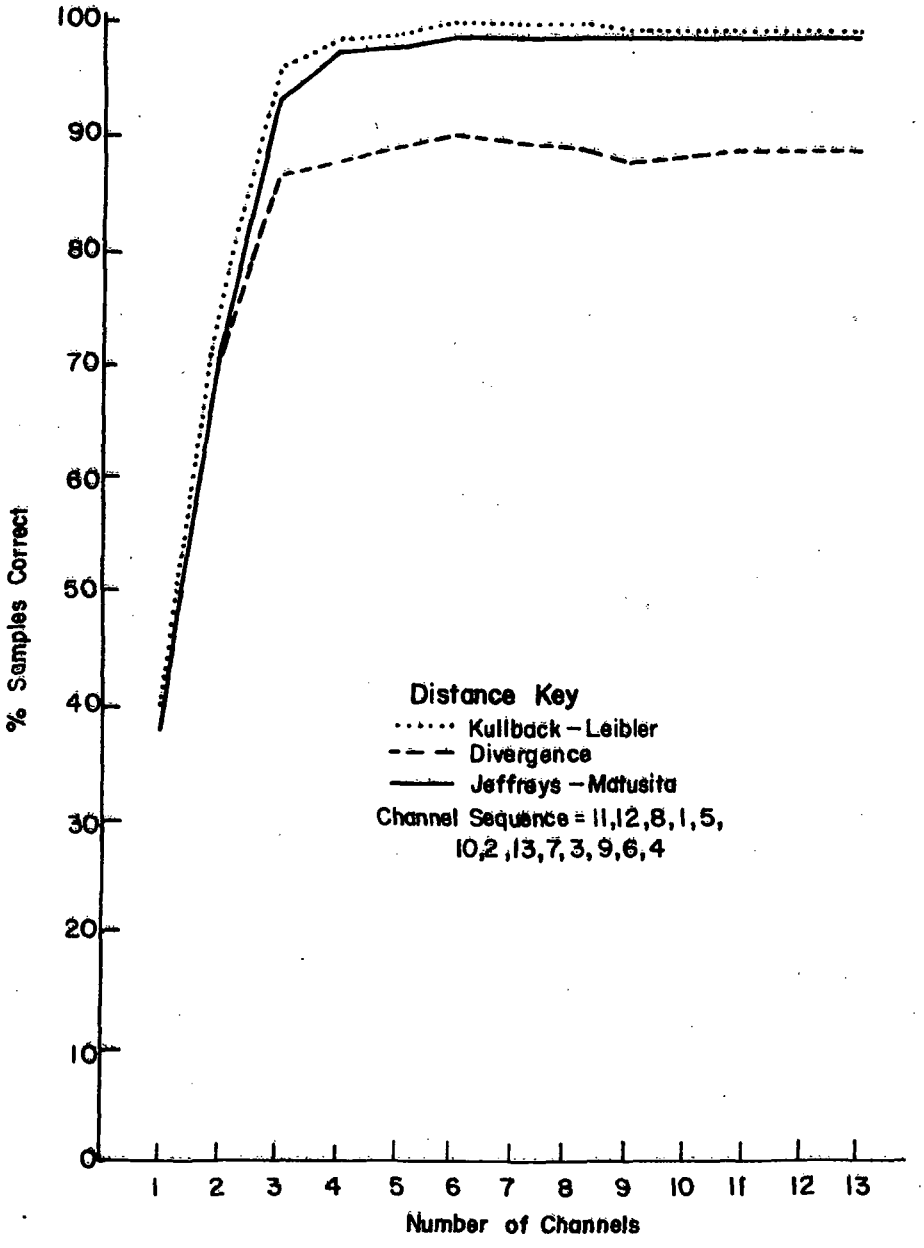


Figure 4.6.1.5 Overall Training Performance vs Number of Channels for Parametrically Implemented Distance Measures. D-PS($SEQDIVG) Training.

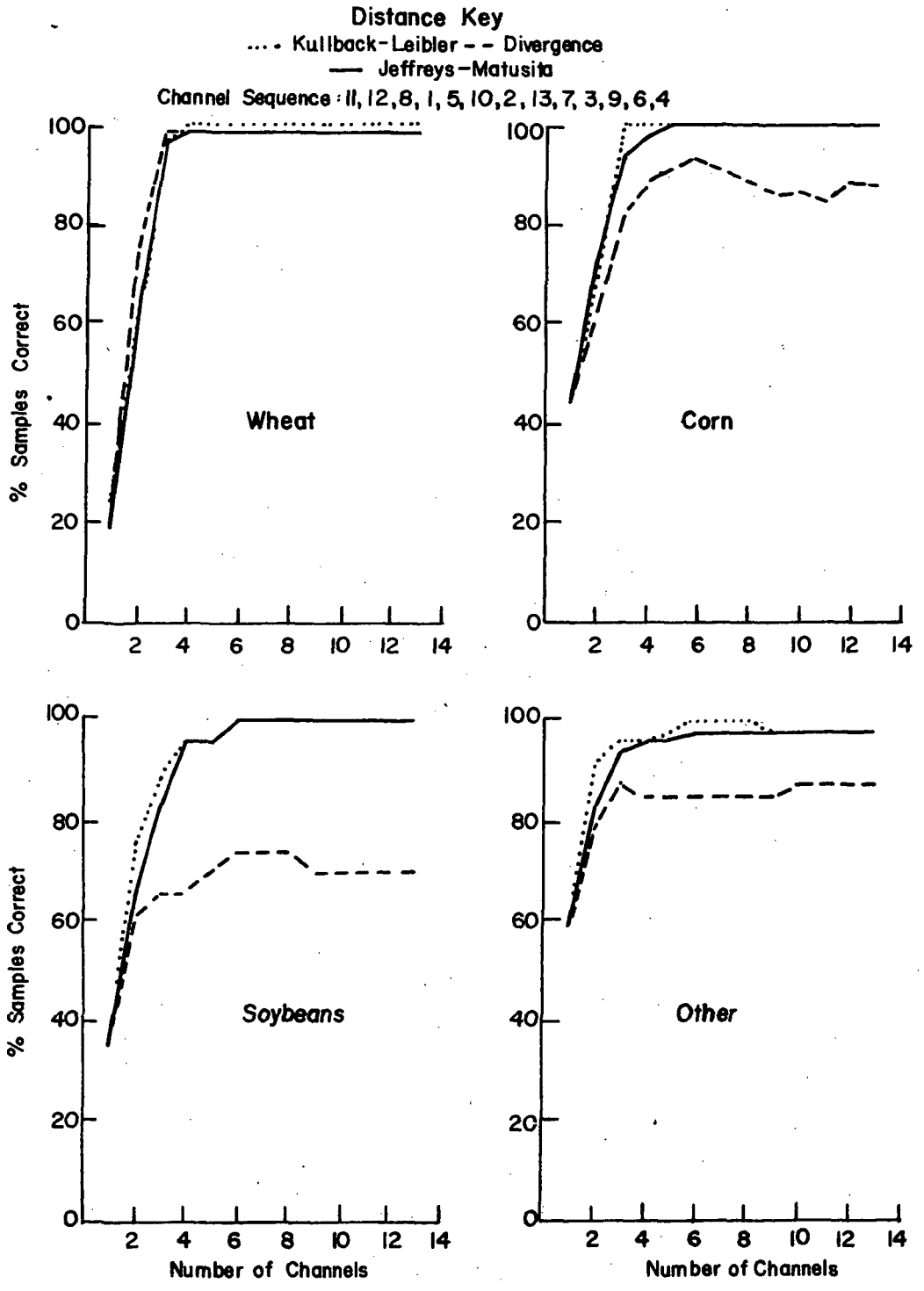


Figure 4.6.1.6 Training Performance by Class vs Number of Channels for Parametrically Implemented Distance Measures. D-PS($SEQDIVG) Training.

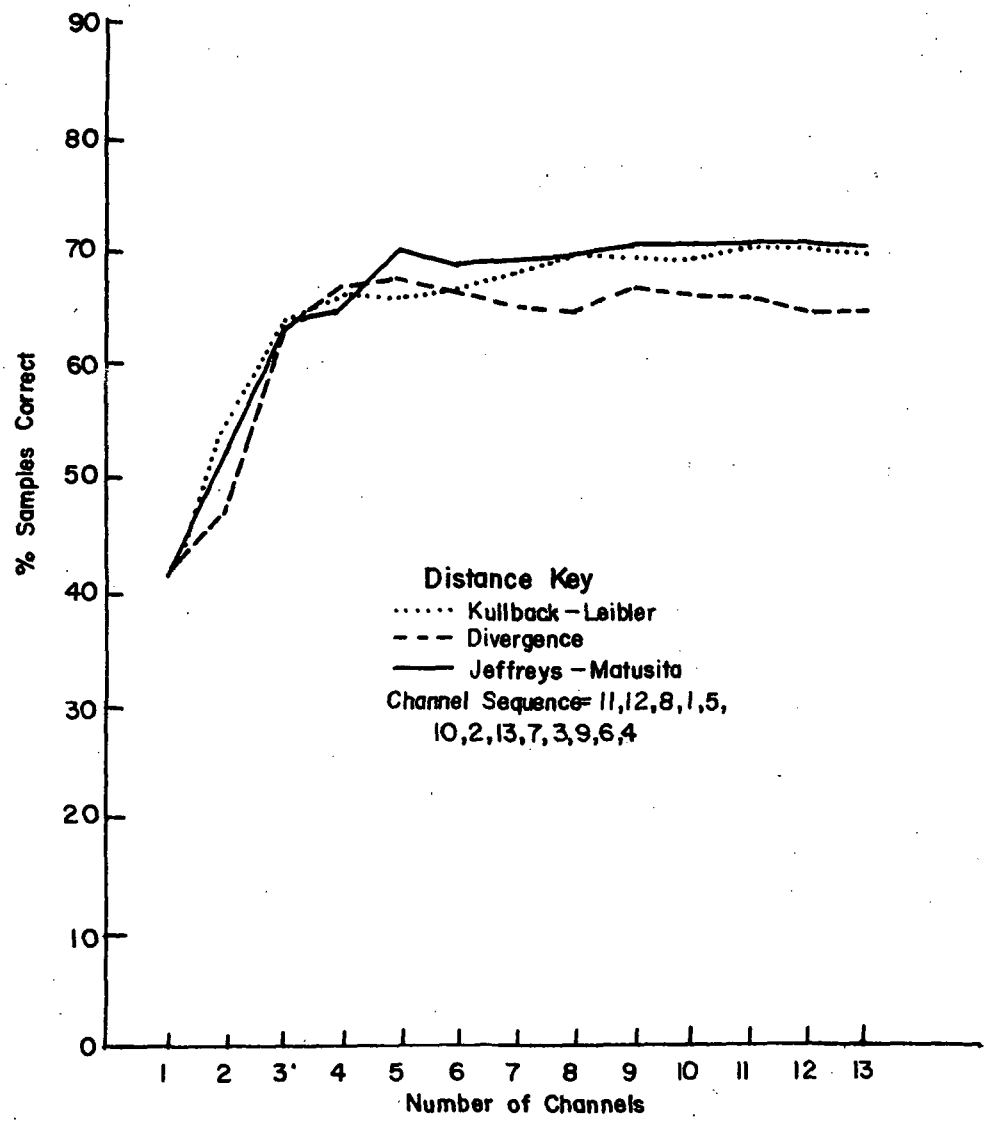


Figure 4.6.1.7 Overall Test Performance vs Number of Channels for Parametrically Implemented Distance Measures. D-PS($SEQDIVG) Training.

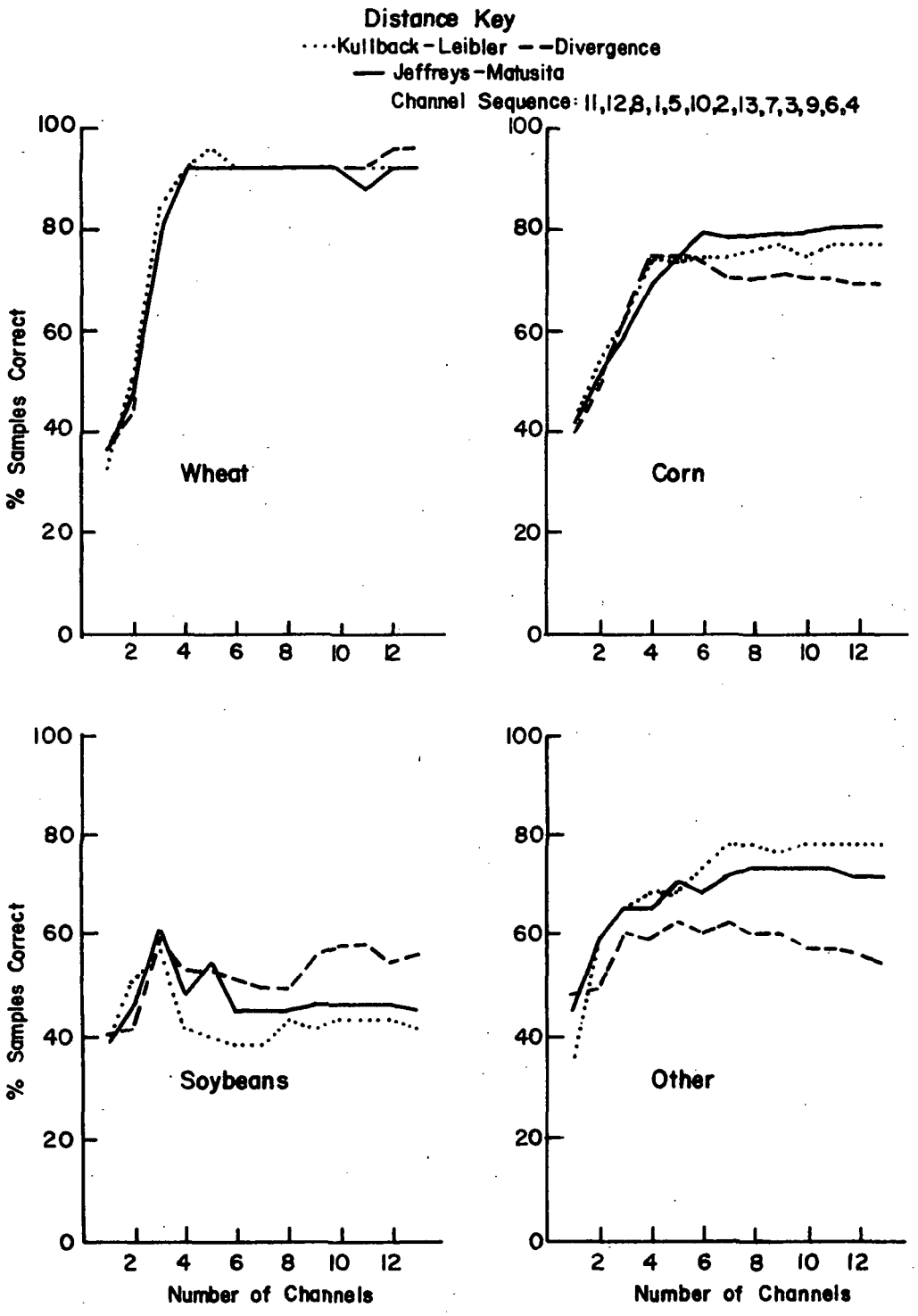


Figure 4.6.1.8 Test Performance by Class vs Number of Channels for Parametrically Implemented Distance Measures. D-PS($SEQDIVG) Training.

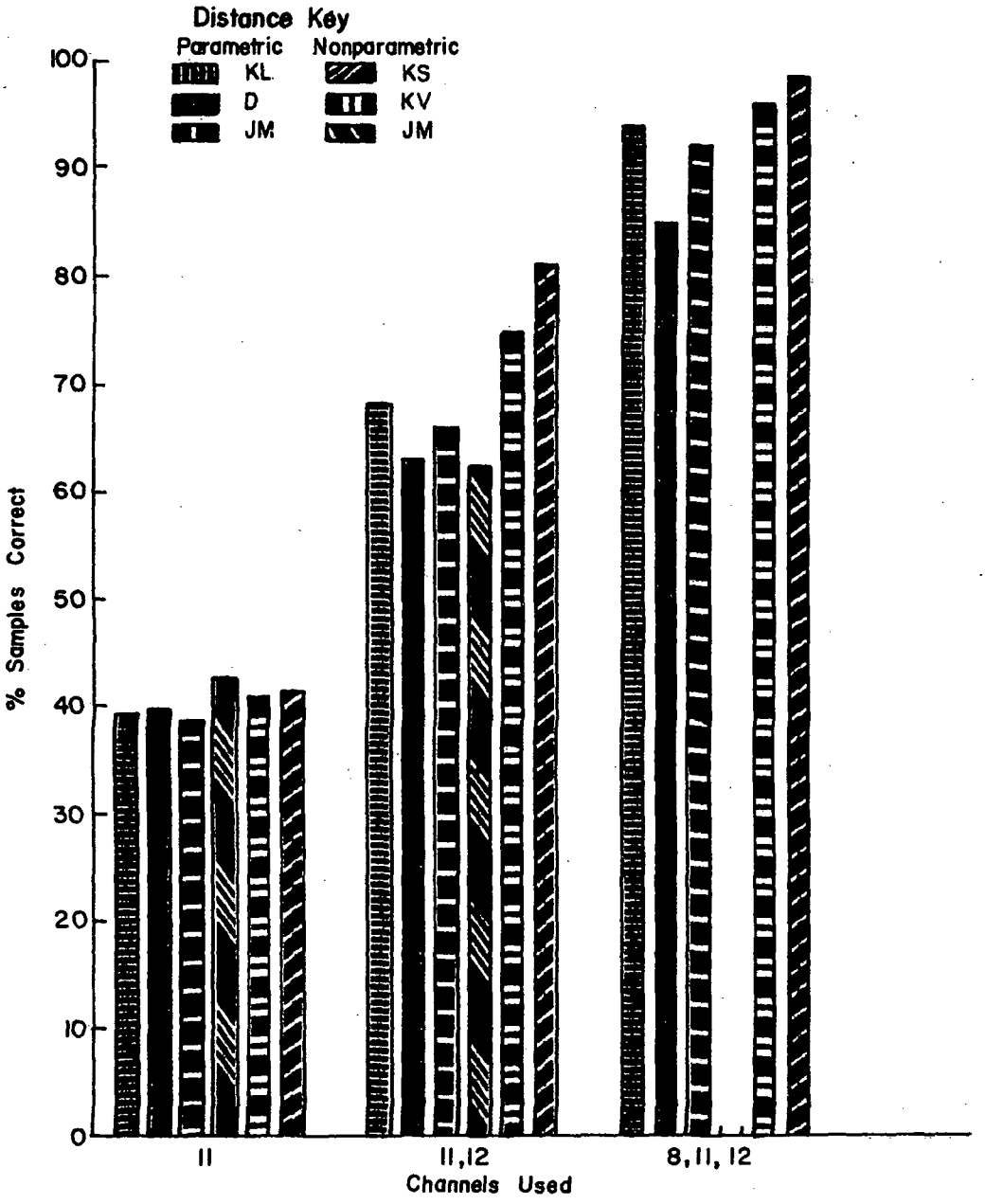


Figure 4.6.1.9 Overall Training Performance vs Number of Channels for Nonparametrically Implemented Distance Measures. JM-PS($SEQDIVG) Training.

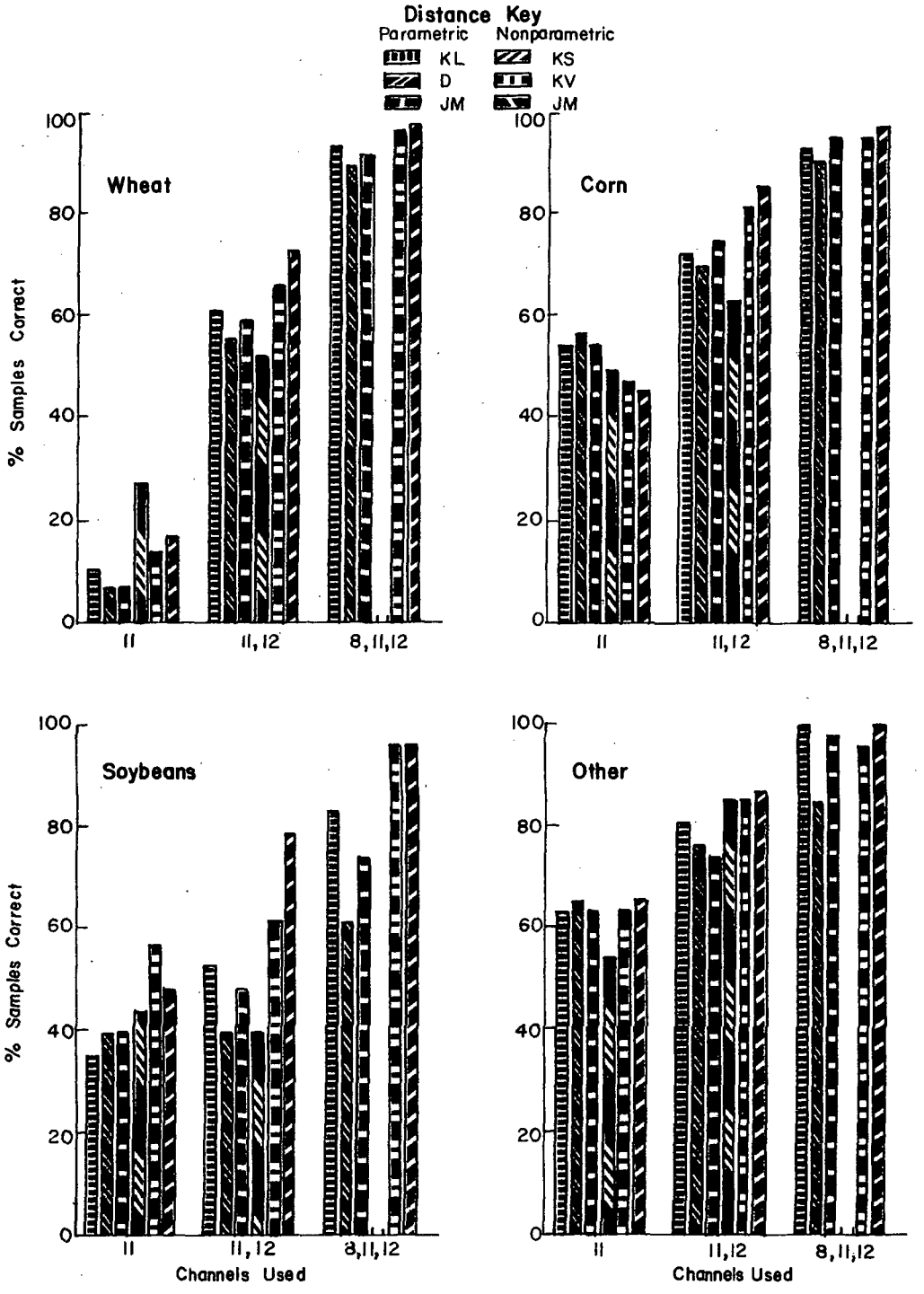


Figure 4.6.1.10 *Training Performance by Class vs Number of Channels for Nonparametrically Implemented Distance Measures. JM-PS($SEQDIVG) Training.*

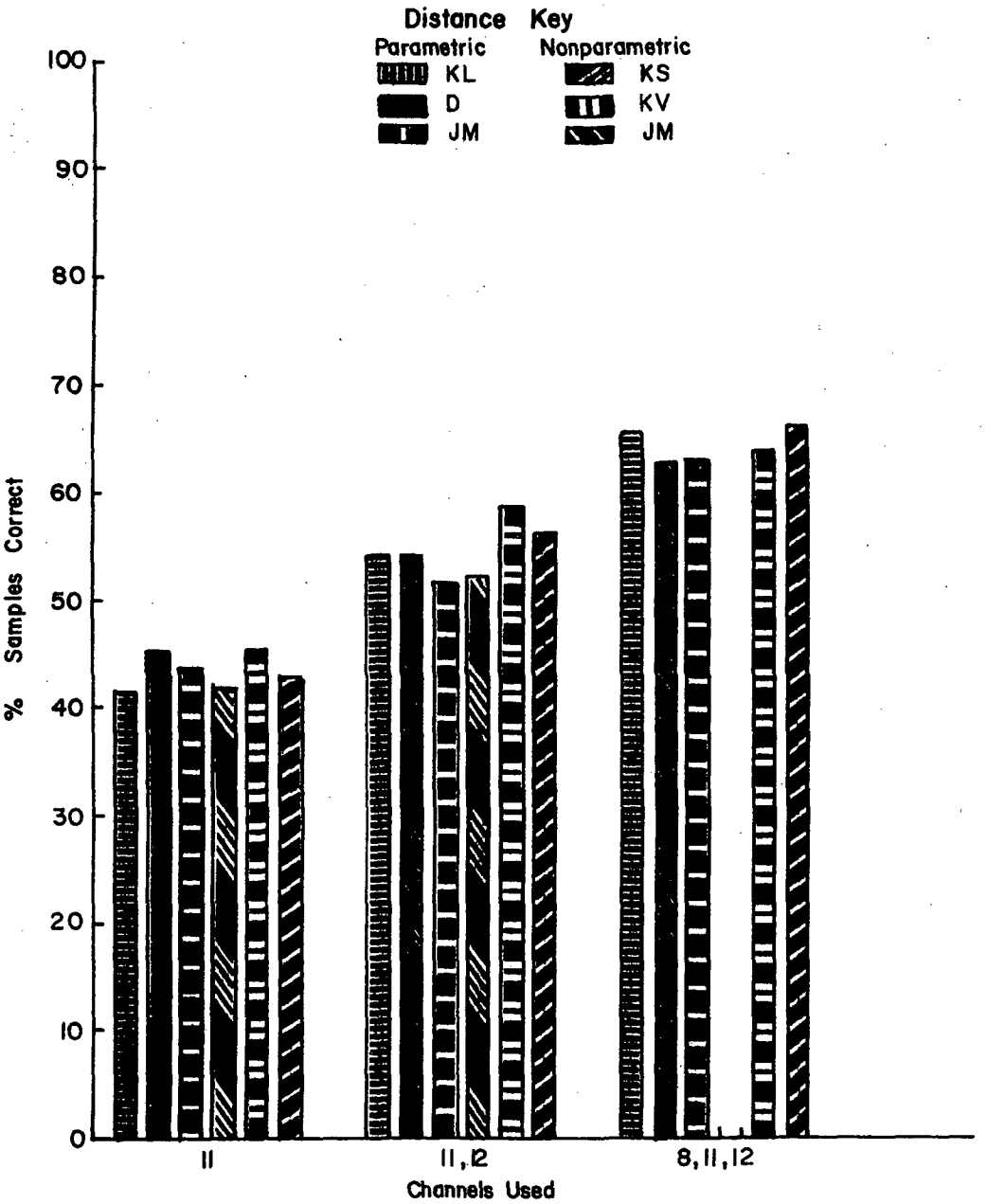


Figure 4.6.1.11 Overall Test Performance vs Number of Channels for Nonparametrically Implemented Distance Measures. JM-PS($SEQDIVG) Training.

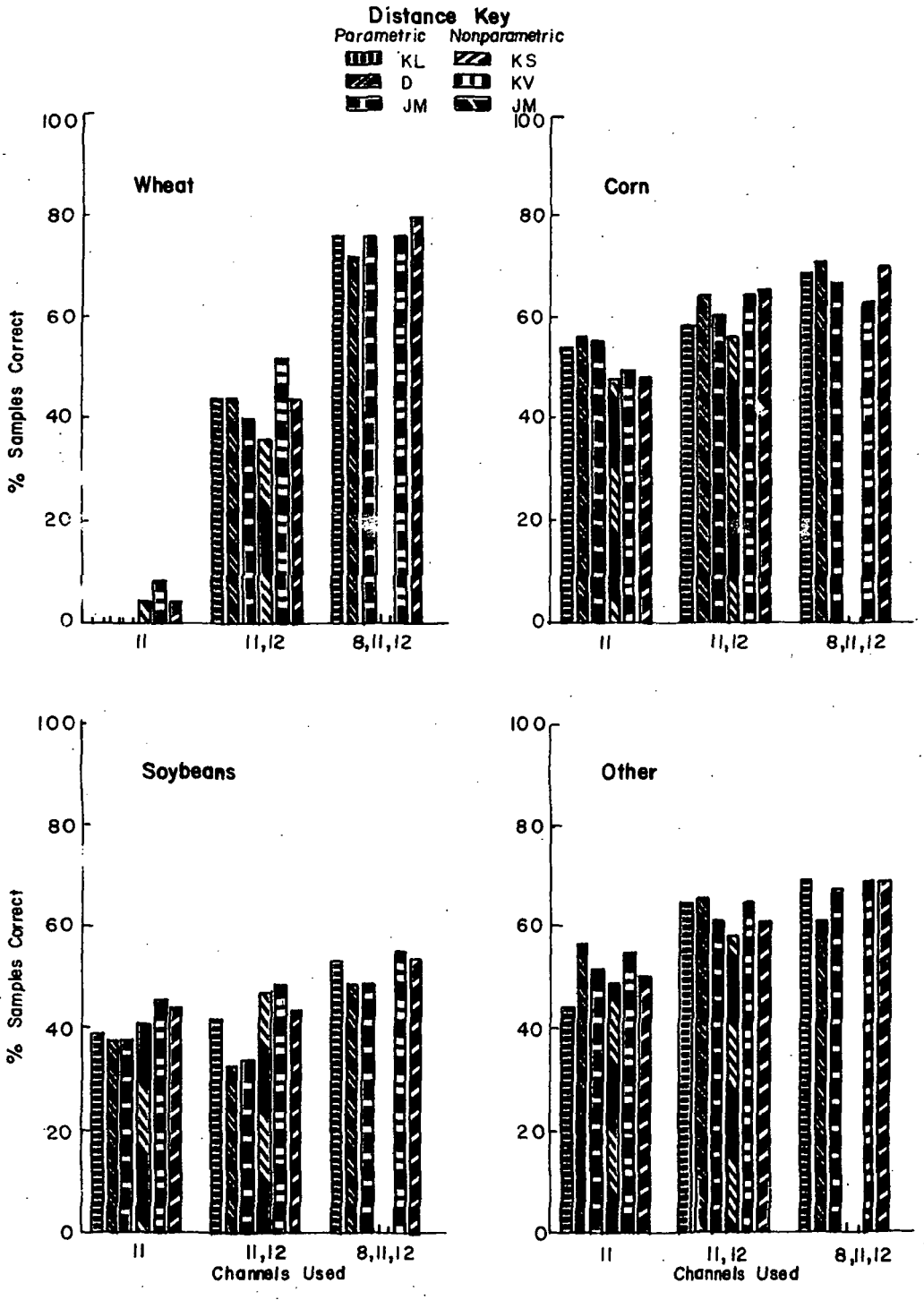


Figure 4.6.1.12 Test Performance by Class vs Number of Channels for Nonparametrically Implemented Distance Measures, JM-PS($SEQDIVG) Training.

GRPSAM no other training results have so far been presented. This must be borne in mind in interpreting the present results as training and test results tend to differ.

On the basis of Fig. 4.6.1.1 through Fig. 4.6.1.12 we make the following observations.

Observation 1

The overall performance increases rapidly as the number of channels increases and saturates in the vicinity of four or five channels (Figs. 4.6.1.1, 4.6.1.3, 4.6.1.5 and 4.6.1.7). The training performance curves saturate somewhat more rapidly than the test performance curves. In this respect minimum distance classification behaves in essentially the same manner as maximum likelihood classification[33] (i.e., LARSYSAA). It is worth noting the similarity between the performance curves and the plot of average JM distance as a function of dimensionality in Fig. 4.4.1.7.

Observation 2

On the basis of overall test performance, the performance of all distance measures is approximately the same (Figs. 4.6.1.3, 4.6.1.7, and 4.6.1.11). The same is, however, <u>not</u> true for training performance where in the parametric case the JM distance and KL numbers perform considerably better than the Divergence (Fig. 4.6.1.1 and Fig. 4.6.1.5), especially when the number of channels is large. Furthermore, in the nonparametric case the KV distance and nonparametric JM distance perform marginally better than KL

KL numbers or the parametric JM distance (Fig. 4.6.1.9). The basic difference between training and test fields is, of course, the fact that there is no guarantee that the training fields are really representative of the test fields. The evidence therefore, seems fairly conclusive that if training is truly representative of the sample to be classified then the particular distance measure used is important. Under these circumstances the nonparametric JM distance also appears to perform better than the parametric JM distance. The last statement is based largely on the 2 channel results since for 3 channels the performance is too large for any distance to show any significant advantage and in the 1 channel case it is too small (cf, Section 4.4.3 Observation 4).

Observation 3

Regardless of whether the JM distance or Divergence is used to cluster the Training Acres, the overall performance for PERFIELD using the JM distance is better than when the Divergence is used. This is also generally true for the performance by class. This is rather unexpected. One would certainly expect that the distance measure used in clustering the data would have a distinct advantage in classification. Since this does not occur the logical conclusion is that the JM distance is a better distance measure than the Divergence. At least this is true for the training data involved. As noted in Observation 2 there

is only a hint of this superiority in the test results.

The performance for KL numbers for training fields is very near that of the JM distance but usually slightly better. For test fields the two distances perform roughly the same. It is interesting to speculate why KL numbers seem to perform slightly better than any other parametric distance considered. And why the Divergence, a symmetrized form of KL numbers, does not perform nearly as well. Perhaps on the basis of the theoretical relationship that exists between maximum likelihood classification and minimum distance classification using KL number this results is not too surprising. Recall that the main factor that distinguishes KL numbers from the other distance measures is that it is not symmetrical with respect to the densities involved. This is probably significant since classification is not entirely a symmetric procedure. Intuitively assigning a field to a class makes more sense than assigning a class to a field. Expressing in words what the KL number represents provides further insight. Thus the KL number of the field for the class is the mean information of discrimination of the field for the class.[33] Intuitively, this rather than the converse (or some mixture), is a logical basis for classifying a field.

Observation 4

The performance by class results reflect fairly closely the overall performance except that as usual the behavior of the class results is more variable. There do not

appear to be any distinguishing features that require comment.

4.6.2 Number of Vectors in the Test Sample

It is of considerable interest to establish how large the test sample must be to enable a minimum distance classifier to achieve reasonable performance. In parametric (normal) problems a commonly used rule of thumb states that at least 10q vectors should be used to get a "good" estimate of a q dimensional covariance matrix. In nonparametric problems no such rule is known but it is usually implied that a large number of vectors are required to adequately estimate a nonparametric density. It is the objective of this section to establish guide lines on the sample size required to achieve reasonable performance in the parametric classifier PERFIELD and the nonparametric classifier LARSYSAA. We only concern ourselves with the test samples and essentially assume that the number of vectors used to estimate the training distribution is large enough so that good estimates are obtained. This fact must be borne in mind in interpreting the results. In other words the question to which an answer is sought is not how many vectors are in general required to adequately estimate a distribution, but rather what is the minimum number of vectors required to estimate a test sample distribution in order that the performance of a minimum distance classifier will not deteriorate. The answer will, of course, depend on the data and again we restrict

our consideration to typical multispectral scanner data.

The experiment devised to explore this problem is the sample size study described in Table 4.6.1. The training method used was the JM-PS($SEQDIVG) method described earlier. Experiments were performed for 2 channels (11, 12) as well as three channels (8, 11, 12). Classifications were performed with both PERFIELD and LARSYSDC using the only distance implemented in both classifiers (i.e., JM distance).

Both "training" and test results are presented. Fig. 4.6.2.1 and Fig. 4.6.2.2 contain the graphs depicting the overall "training" performance and "training" performance by class respectively. Fig. 4.6.2.3 and Fig. 4.6.2.4 contain the corresponding test results. Since the number of vectors used to estimate the distributions of the sample to be classified is the quantity being varied the "training" performance curves are in fact based on a subset of the vectors in the Training Acres rather than all of the vectors as is usually the case for determining training performance. More specifically to obtain the "training" performance curves rather than use all the vectors in an acre to estimate the distribution for that acre for classification purposes, only the appropriate number of vectors from the acre are selected for estimation purposes. Of course, all the vectors in the acre still form the basis for estimating the training distribution. Similarily to obtain the test performance curves the appropriate number of vectors are selected from the Flightline 21 Test Areas. The method of selecting the vectors

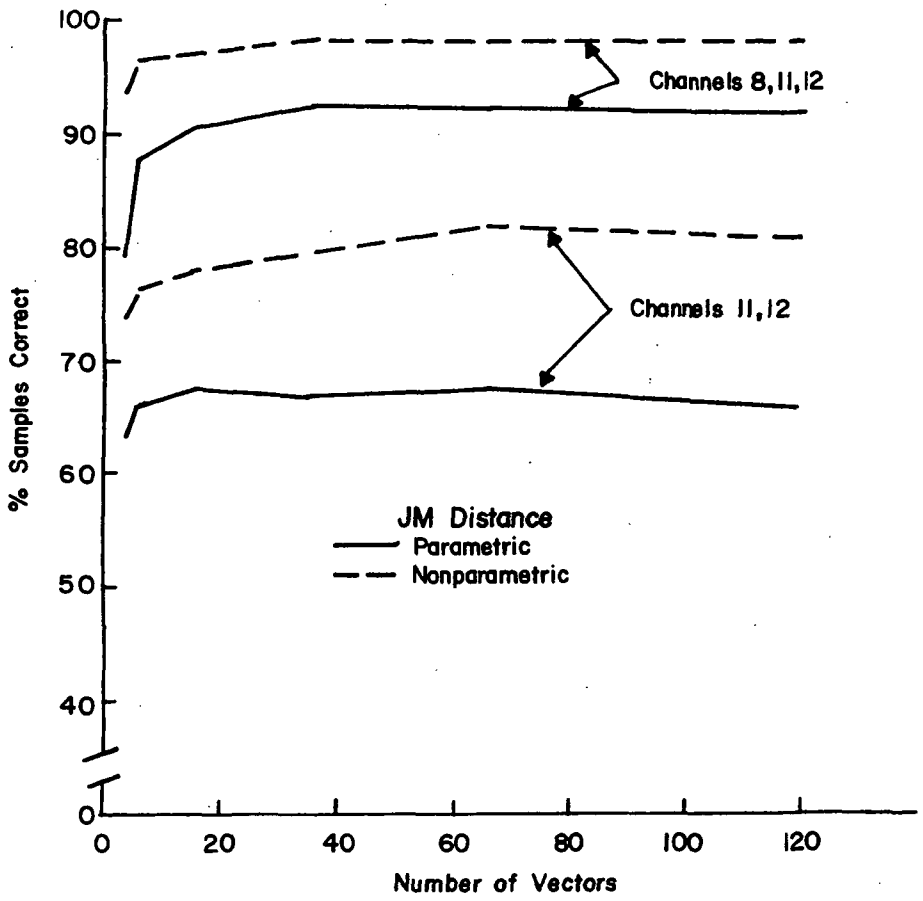


Figure 4.6.2.1 Effect of Sample Size on Overall Training Performance.

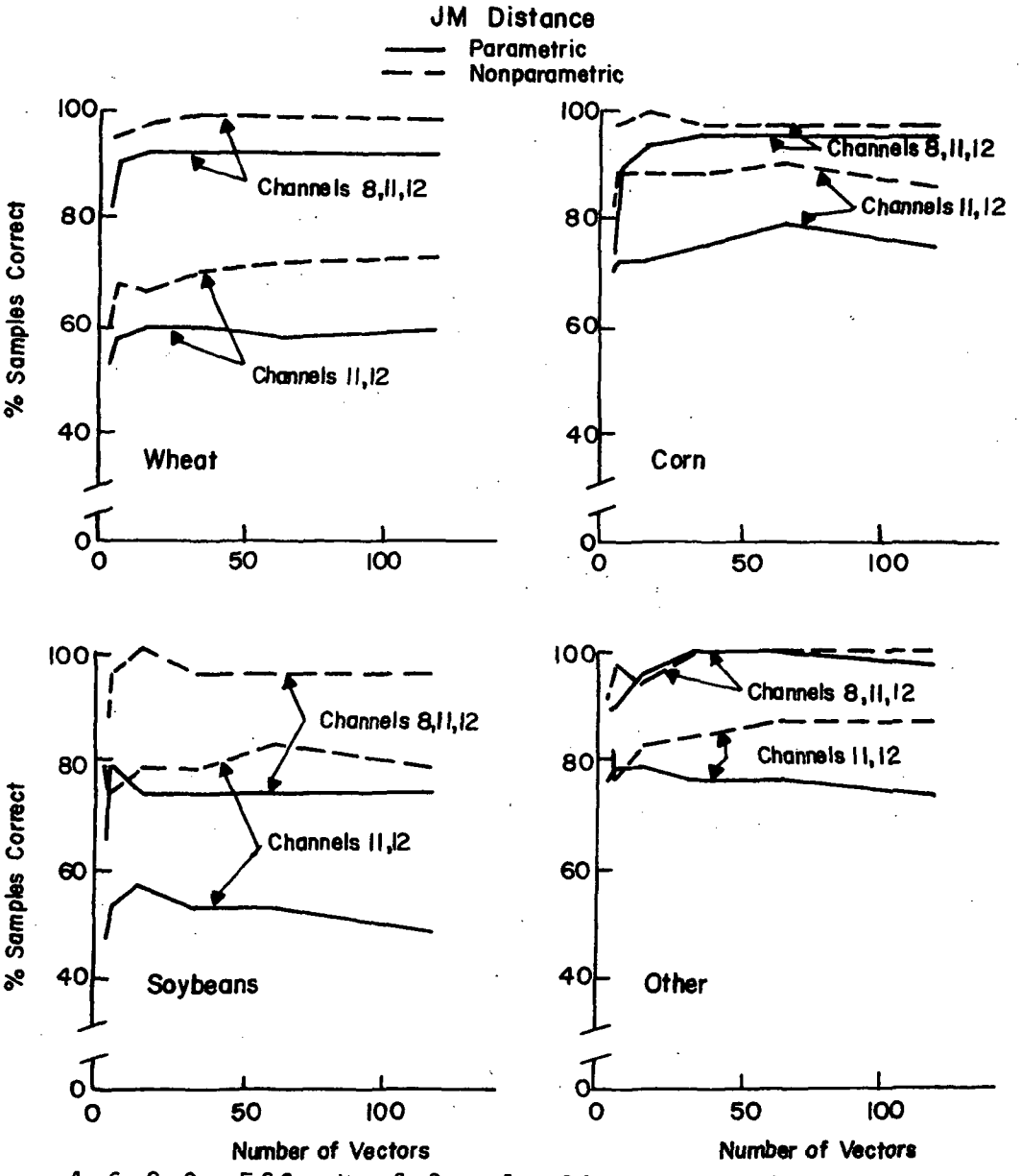


Figure 4.6.2.2 Effect of Sample Size on Training Performance by Class.

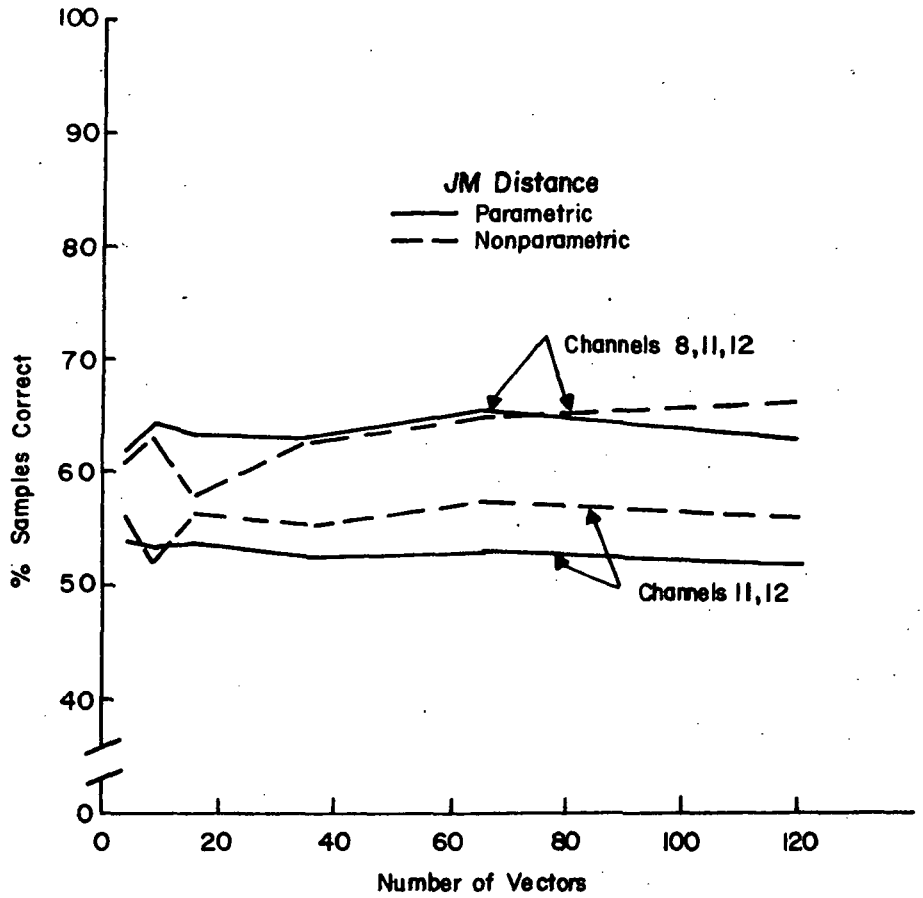


Figure 4.6.2.3 Effect of Sample Size on Overall Test Performance.

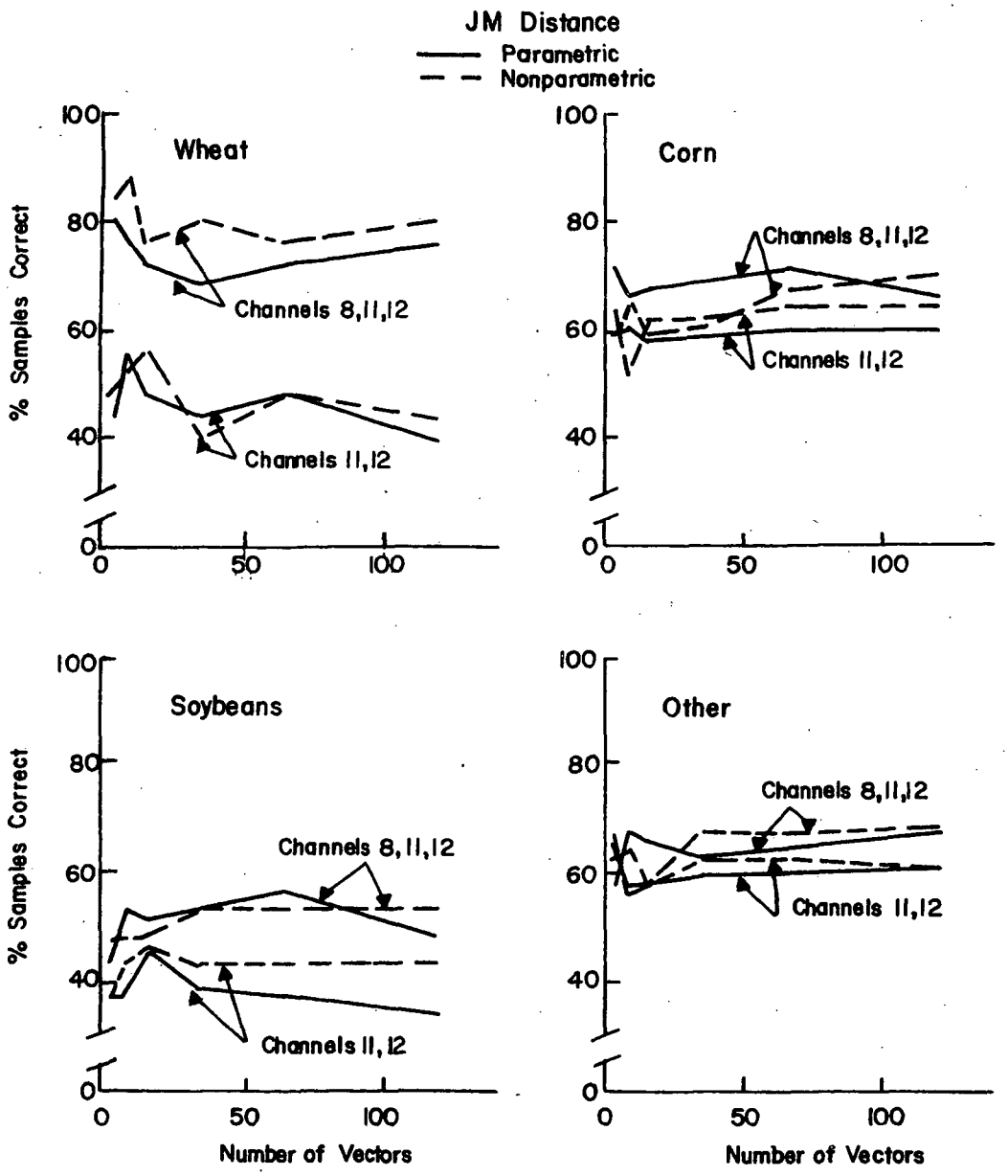


Figure 4.6.2.4 Effect of Sample Size on Test Performance by Class.

for estimating distribution of the sample to be classified is the same for both training and test results. This method has been described in Section 4.6. In essence the vectors are selected to be spread out as much as possible over the area from which they are chosen.

On the basis of the results presented in Fig. 4.6.2.1 to Fig. 4.6.2.4 the following facts emerge.

Observation 1

The overall training performance definitely decreases as the sample size decreases but the sample size must be extremely small before the decrease is significant. The overall test performance does not exhibit as definite a trend. Instead it seems simply to become somewhat erratic as the samples size decreases. In any case it appears that the use of 10q vectors is adequate to estimate the distribution of the samples to be classified for both PERFIELD and LARSYSDC.

Observation 2

There is absolutely no indication that the number of vectors required to adequately estimate a density histogram for classification purposes need by any larger than the number required to obtain the corresponding parametrically estimated density.

On the basis of this results it appears likely that in general the number of vectors considered necessary to adequately represent a density histogram is over estimated.

In fact it appears likely that in a situation where a parametric description is reasonable, the number of vectors required to adequately estimate a density histogram need be no greater than the number required to adequately estimate the parameters. It appears that for reasonably well behaved densities the number of vectors required for nonparametric estimation purposes is quite reasonable and not as large as is typically implied.

Observation 3

It is interesting to consider what happens if the sample size is reduced until only one vector is available from the field to be classified. In this situation the parametric classifier PERFIELD cannot classify the sample since the covariance matrix cannot be estimated. It is trivial to show that the nonparametric classifier, using either the KV or JM distance, becomes a maximum likelihood vector classifier in which density histograms are used to estimate the class distributions. Thus as the test sample size is reduced to its lower limit LARSYSDC (with JM or KV distance) becomes a vector by vector classifier of a rather desireable type. Considering that the performance of a parametric maximum likelihood classifier (LARSYSAA) is only slightly less than the parametric minimum distance classifier PERFIELD (see section 4.4.3 Observation 4). It is clear that the performance of LARSYSDC will typically not drop a large amount when the sample size is decreased. This result also

suggests that usually the minimum distance classifier, based on density histograms, will perform better than the maximum likelihood classifier based on density histogram. This follows because the limiting form of the minimum distance classifier is the maximum likelihood classifier.

Observation 4

The nonparametric JM distance yields a higher classification accuracy on the Training Acres than the parametric JM distance. Not only is this true for the overall performance but it is also true for each class individually. This behavior is similar to that observed in the number of channels study and would in fact be expected on the basis of that study (cf, Section 4.6.1, Observation 2). As in the number of channels study the possible superiority of the nonparametric technique is essentially not evident in the test results. While the classification accuracy on test fields is slightly larger for the nonparametric case the difference is slight.

4.6.3 Bin Size

A parameter of considerable significance in LARSYSDC is the bin size. Certainly if the bin size is too large, small differences between densities will be obscured and performance will deteriorate. On the other hand a small bin size implies longer computation times and possibly poorer estimates as well; since if the bin size is very small, then the number of bins is very large and more vectors are needed

to adequately estimate the distribution.

The objective of this section is to determine what effect bin size has on performance. Actually it is probably the ratio of number of vectors to the bin volume that is the important parameter but since the number of vectors is fixed at 121, bin size can be considered directly.

With reference to the bin size study portion of Table 4.6.1 we note the training is again based on JM-PS clustering of the Training Acres with the number of subclasses as established in the evaluation of GRPSAM (i.e, JM-PS($SEQDIVG) training). Naturally only the LARSYSDC classifier is involved since PERFIELD does not use density histograms. Classifications are performed for 1 Channel (11), 2 Channels (11, 12) and 3 Channels (8, 11, 12) for each of the distance measures available in LARSYSDC.

Fig. 4.6.3.1 thru Fig. 4.6.3.8 contain the experimental results. Basically results were obtained for bin sizes of 1, 5, 10, 20 and 30 with some exceptions necessitated either by exceedingly large histograms or by difficulties in converting large pdf's to cdf's. These exceptions are apparent from the figures and will not be enumerated.

On the basis Fig. 4.6.3.1 thru Fig. 4.6.3.8 we make the following.

Observation 1

The overall test performance is remarkably insensitive to bin size while the overall <u>training</u> performance

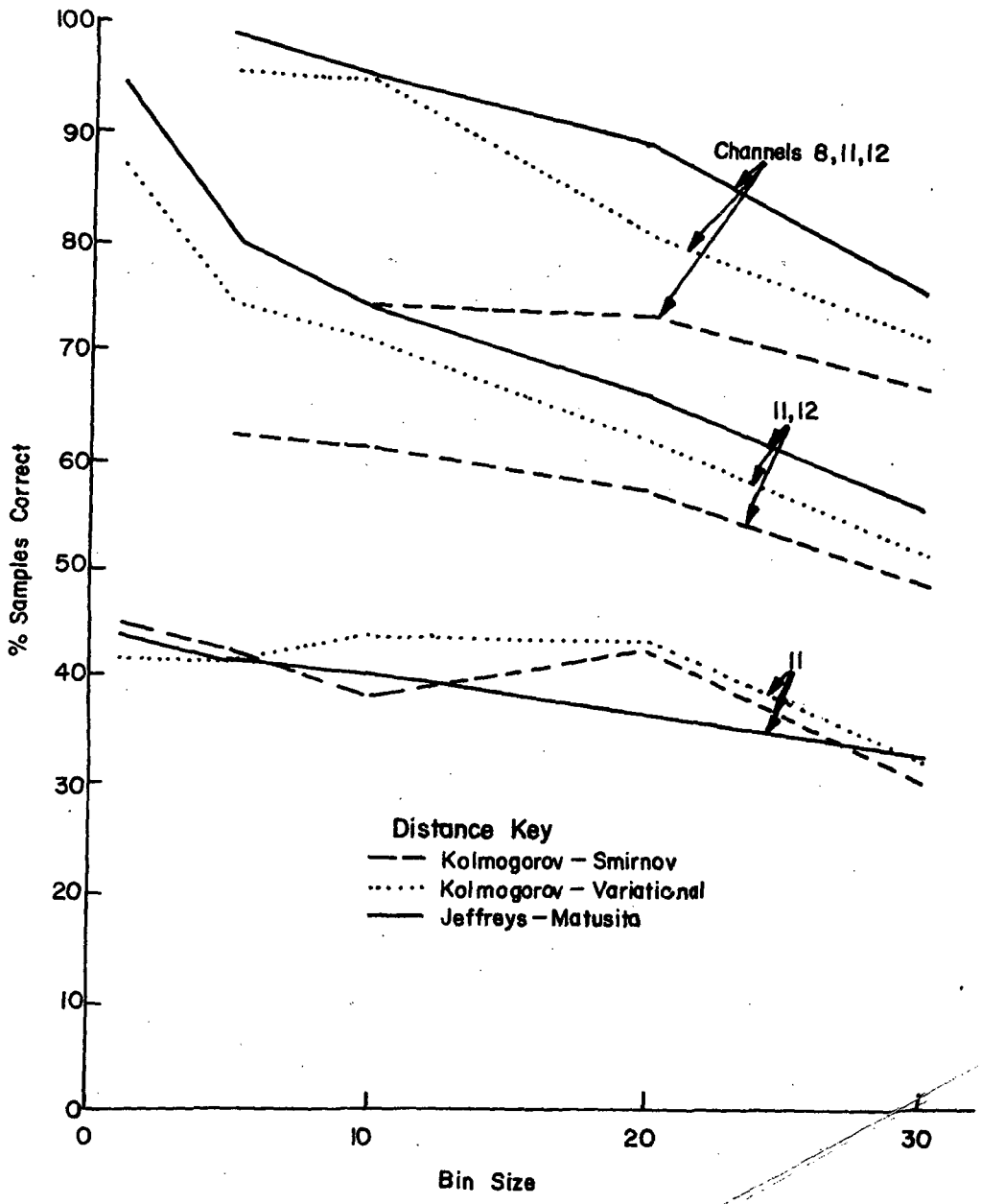


Figure 4.6.3.1 Effect of Bin Size on Overall Training Performance.

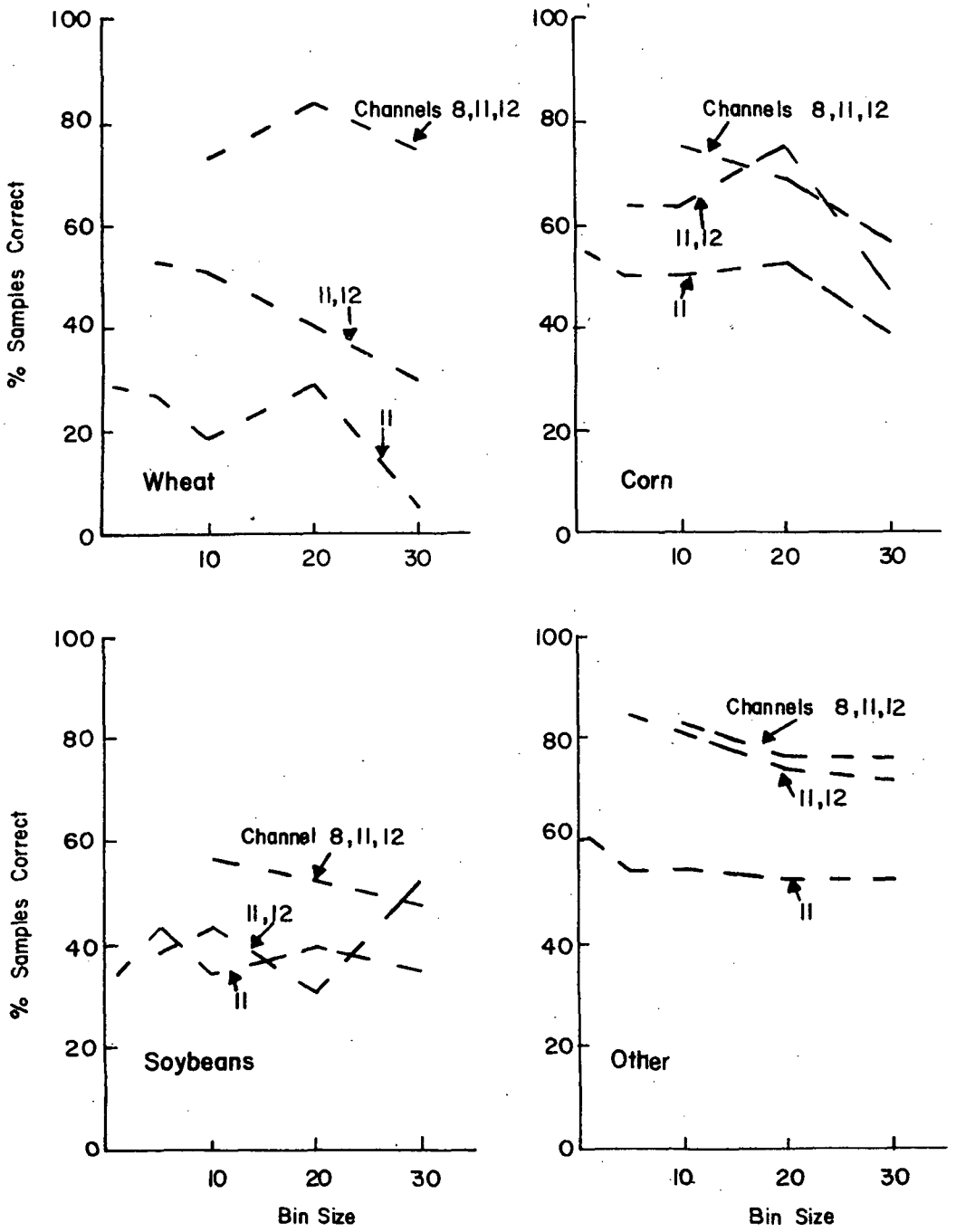


Figure 4.6.3.2 Effect of Bin Size on Training Performance by Class for Kolmogorov-Smirnov Distance.

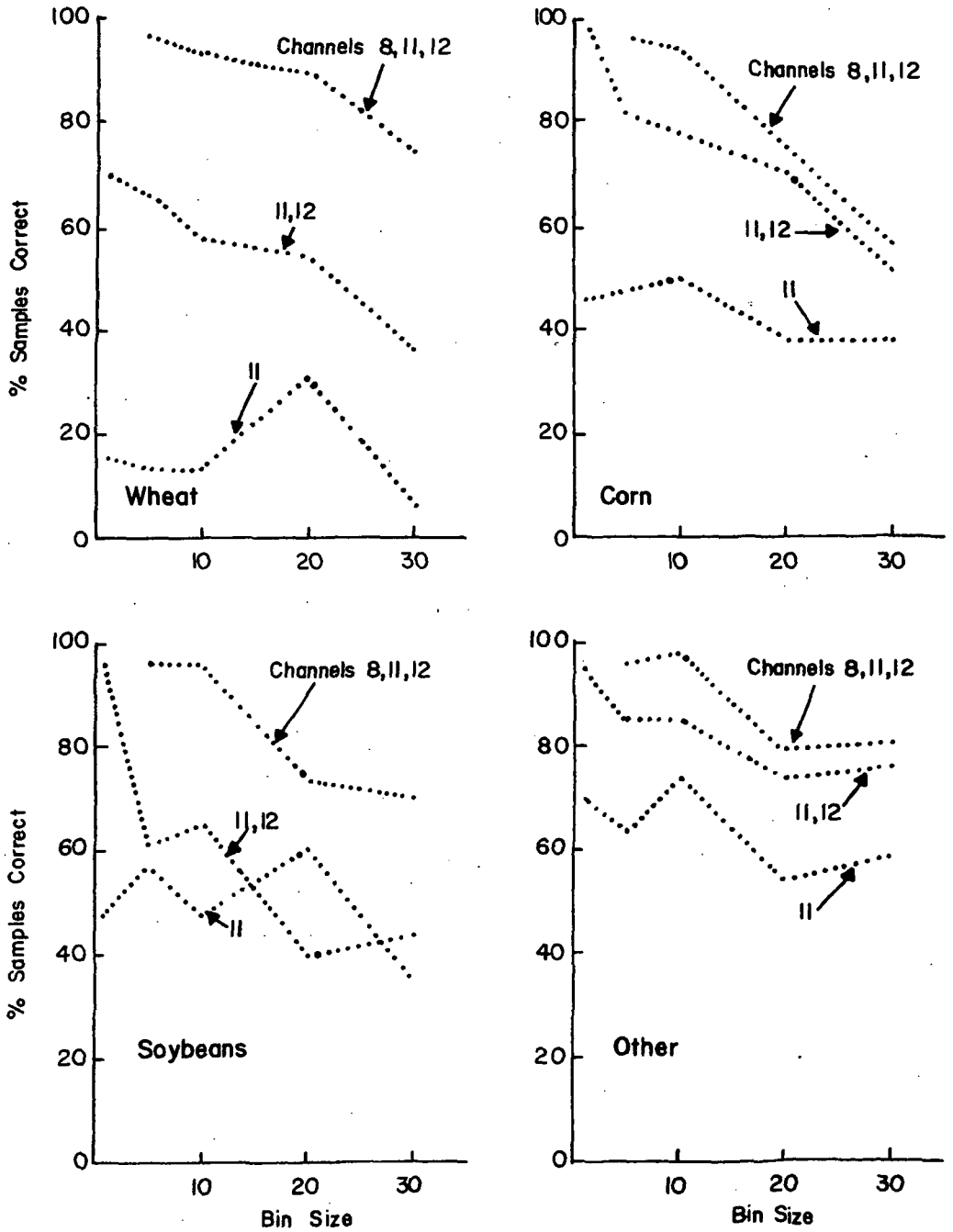


Figure 4.6.3.3 Effect of Bin Size on Training Performance by Class for Kolmogorov-Variational Distance.

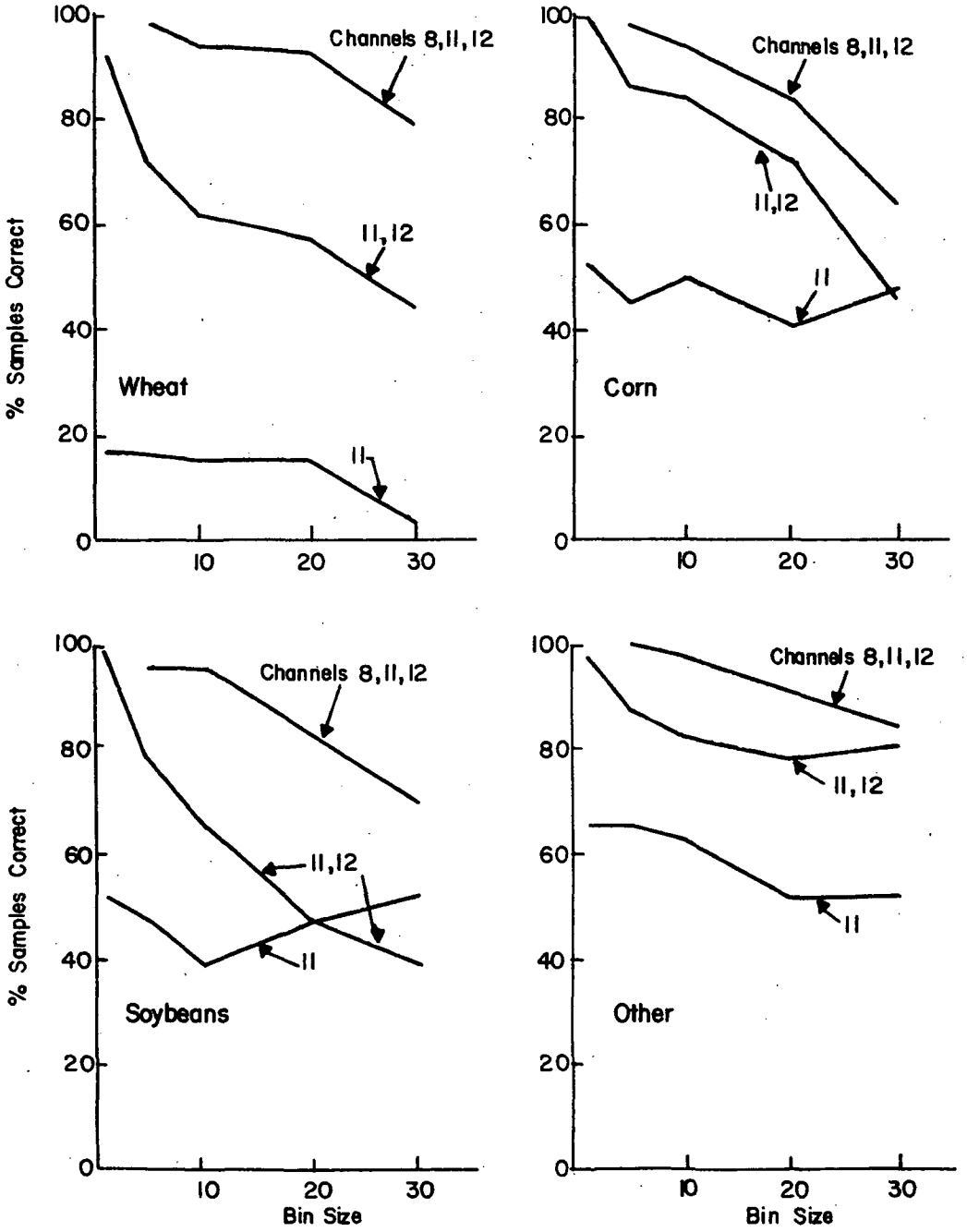


Figure 4.6.3.4 Effect of Bin Size on Training Performance by Class for Jeffreys-Matusita Distance.

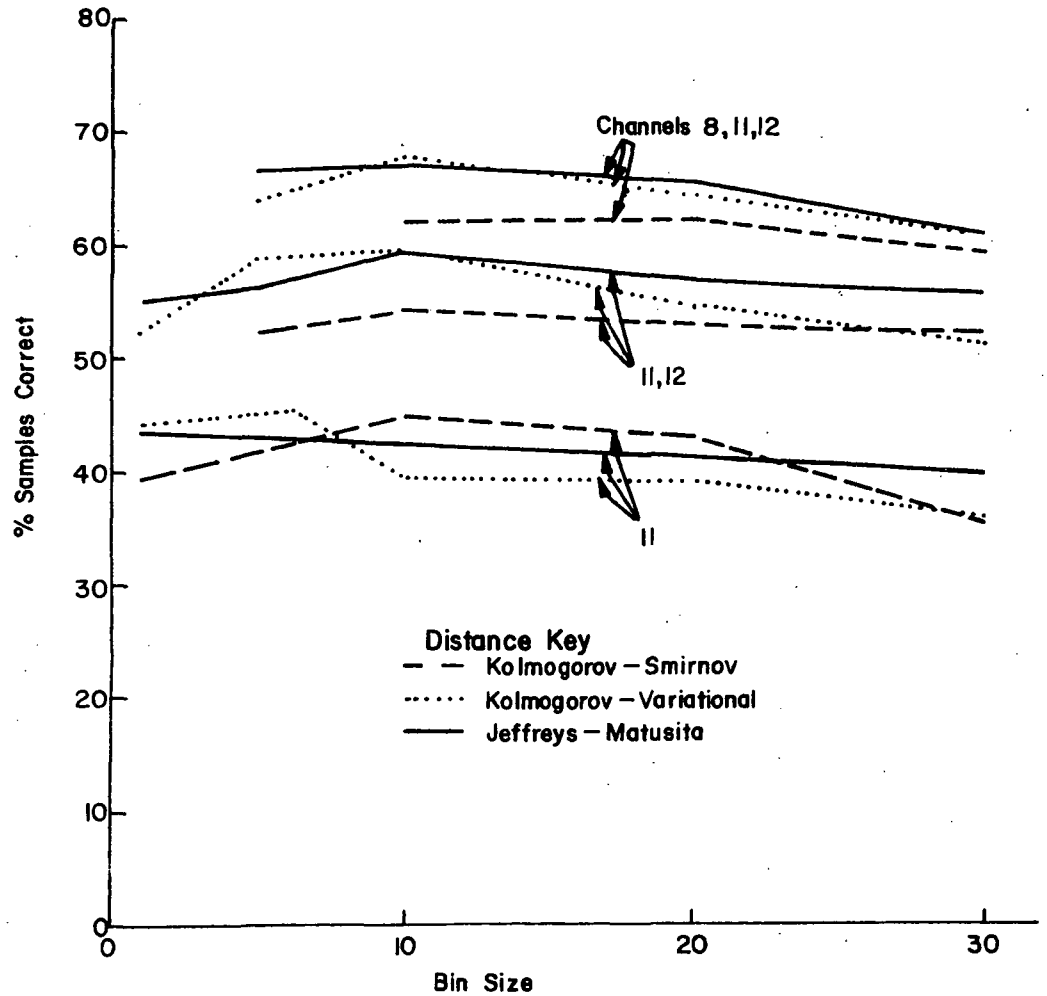


Figure 4.6.3.5 Effect of Bin Size on Overall Test Performance.

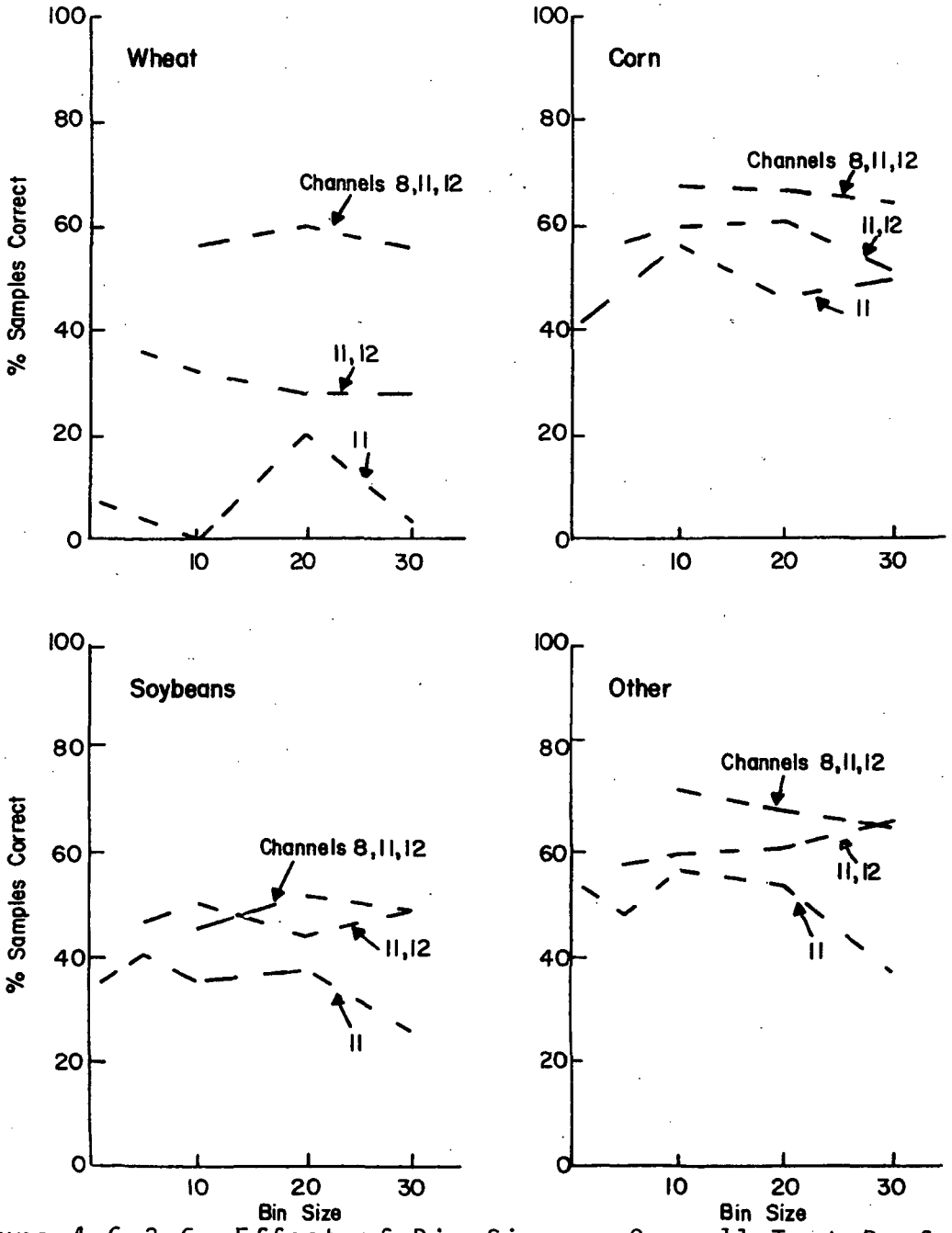


Figure 4.6.3.6 Effect of Bin Size on Overall Test Performance by Class for Kolmogorov-Smirnov Distance.

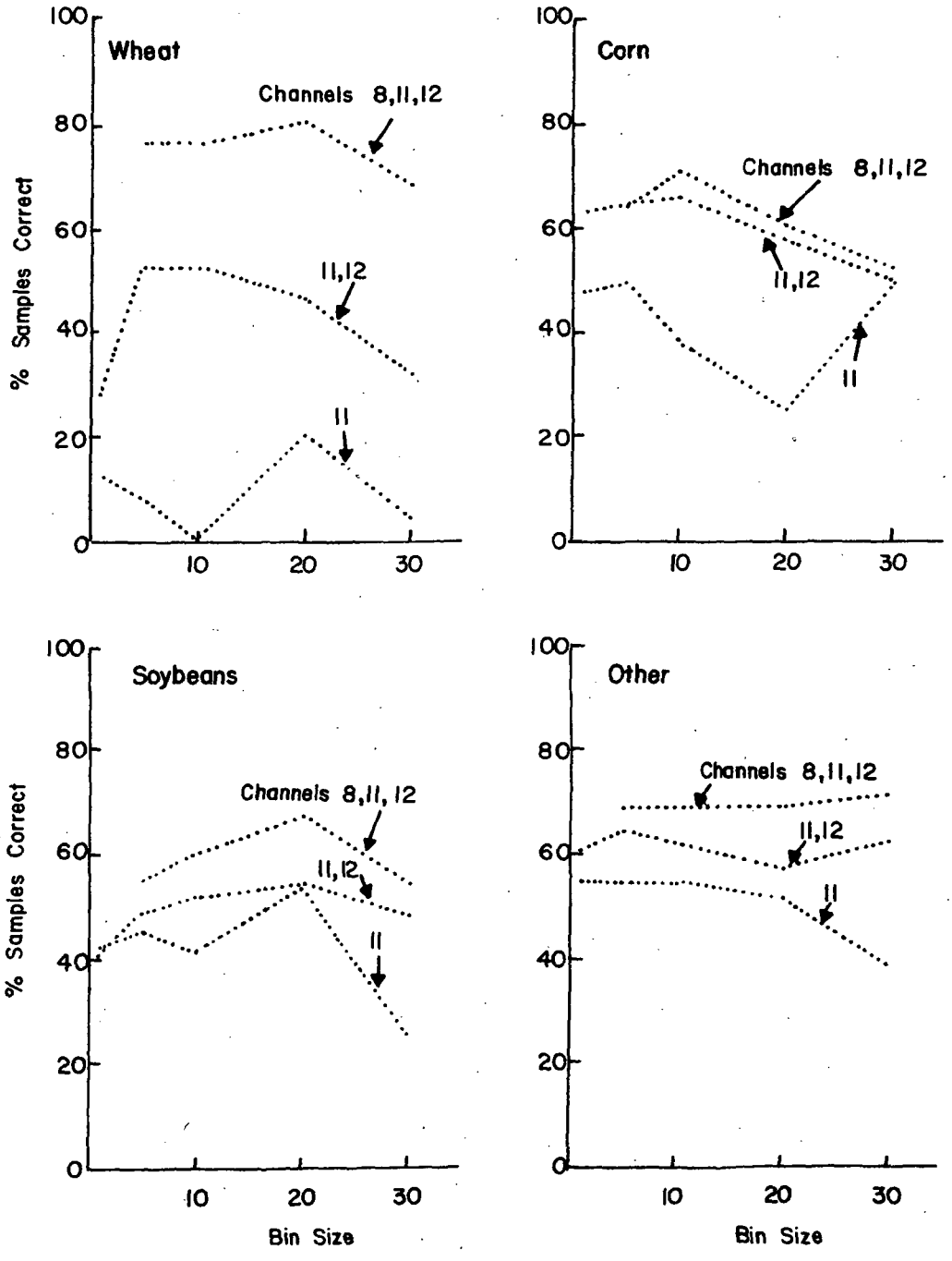


Figure 4.6.3.7 Effect of Bin Size on Overall Test Performance by Class for Kolmogorov-Variational Distance.

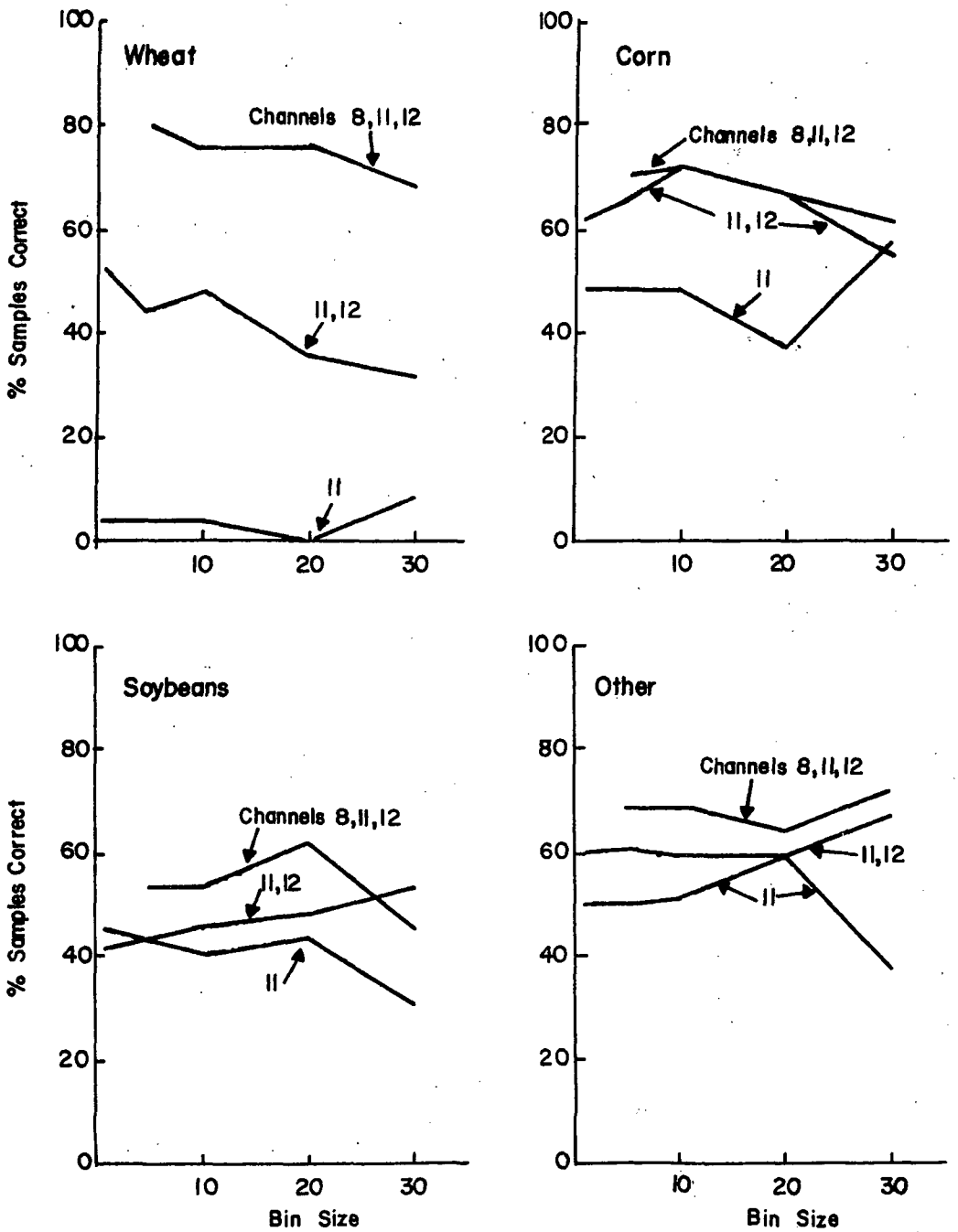


Figure 4.6.3.8 Effect of Bin Size on Overall Test Performance by Class for Jeffreys-Matusita Distance.

exhibits a greater sensitivity; at least this is true for the two and three channel classifications. As noted in the number of channels study for the one channel case performance seems to be so poor, and the parameter space densities overlapped to such an extent, that single channel results provide little useful information regarding the superiority of any parameter studied.

Observation 2

The overall test performance suggests that there is perhaps an optimum bin size in that the test performance seems to decrease slightly for very small as well as for large bin size. The training performance continues to improve as the bin size decreases. Because of the limited number of results the evidence is not too conclusive but the apparent different behavior for test and training is not necessarily contradictory as the following argument demonstrates. To simplify the explanation and possibly exaggerate the effect, suppose the multispectral scanner data is real (as opposed to integer) data and that the bin size is chosen small enough so that every nonempty bin for <u>both</u> test and training density histograms contained only one vector. Then the JM distance between two distributions depends only on the ratio of the number of coincident nonempty bins from the two distributions to the maximum number of possible coincident bins. In other words it is only the spatial distribution of bins that is important. The true shape of the

distribution (i.e., bimodal, etc.) has no direct influence on the classification except to the extent that this influences the location of the nonempty bins. Since the training histograms are derived from the Training Acres the spatial correlation between the two is quite large. In fact every nonempty bin for any particular Training Acre to be classified coincides with a nonempty bin in the subclass to which that acre should be assigned. Only if the histograms for two or more subclasses overlap over the whole region occupied by the Training Acre can the Training Acre be incorrectly classified. This condition does not prevail for test fields where conceivably the general shape of the densities is of greater importance to correct classification then the spatial distribution of nonempty bins. It is not known if this is the correct explanation of the above phenomena but the information in Table 4.6.3.1 tends to support this explanation. This table gives the average over both test and training histograms for the data involved of the average number of vectors per non empty bin for various combinations of channels and bin size of interest.

Table 4.6.3.1

Average Number of Vectors Per Nonempty Bin

Bin Size	Number of Channels	Histogram Type	Average Number of Vector Per Nonempty Bin
1	2	Train	2.26
1	2	Test	1.66
5	2	Train	16.38
5	2	Test	9.42
5	3	Train	5.73
5	3	Test	4.77

Observation 3

The improvement in performance with decreasing bin size is not as great for the KS distance as for the KV and JM distances. This is particularily true for training results and appears to be true for test results. In fact, the percentage of training samples correctly classfied when the KS distance is used falls considerably below the percentage classified correctly by the KV and JM distances.

Observation 4

The behavior of the performance by class curves for both test and training results is quite erratic although the general trends observed in the overall performance curves are also present in the performance by class curves.

CHAPTER 5
CONCLUSIONS

Scattered throughout the various sections are numerous "Observations" most of which in essence are really conclusions with discussions pertaining to the conclusions. In the current chapter the more significant "Observations" are collected from their diverse locations and presented in a unified manner. In general the conclusions presented are based on experimental results obtained with a particular set of data and strictly speaking the conclusions are really only valid for that data. It is, of course, extrapolation of these conclusions to other data sets that is of interest. We believe that such extrapolation is valid for most multispectral scanner data, at least as long as it bears a reasonable similarity to the particular data studied. In fact the wording of the conclusions is based on the assumption that this is the case. Of course, we recognize that multispectral scanner data sets will be encountered for which not all of the conclusions will be valid.

Some of the conclusions are based on averages over three similar flightlines. Others are based on a single flightline. Obviously the conclusions based on the average of three flightlines should be more reliable than those based

on a single flightline. However, even if only one flightline is involved the amount of data upon which the conclusions are based is always quite substantial. In all cases the experimental investigations involved problems that in terms of number of classes and number of subclasses are quite realistic.

Probably the most significant conclusion is that for the training methods employed the test performance that can be achieved with minimum distance classifiers is "essentially" independent of the distance measures considered, or on whether the implementation of the classifier is based on parametrically estimated densities or density histograms. The word "essentially" has been inserted because the non-parametric classifier using the JM distance gave "hints" of superiority even for test data but the variability of the results is sufficiently large that many more classifications would be necessary to establish if this distances had some small advantage.

In contrast the training performance is significantly influenced by the distance measure, and whether or not the classifier is implemented parametrically. More specifically the nonparametric implementation utilizing the JM distance gave the best performance on test results. In the parametric case the JM distance also performed well with KL numbers doing slightly (but probably not significantly) better.

A feature common to all the classifications performed in this study, as well as those in the crop yield study, is the disparity in classification accuracy of test and training data. Test performance is typically of the order of 25% below training performance. Also the behavior of the test data does not entirely mirror the behavior of the training data. Apparently the training data and/or Training Procedure results in subclasses that are not really representative of the true data.

Considering simultaneously the test results, training results and the nonrepresentativeness of the training data the implications seem fairly clear. Until training techniques are developed which ensure that the training data is truly representative of the test data the choice of distance in a minimum distance classifier is not critical, and the extra complexity of a nonparametric classifier is not warranted.

Although a nonparametric minimum distance classifier based on density histograms at present does not offer any advantage in classification accuracy over a parametric classifier, it does have two advantages that should be mentioned. The first is that if random training is used subclasses can be eliminated without paying any penalty in either average performance or variability in performance. This is not true for the parametric minimum distance classifier where elimination of subclasses leads to a great

increase in the variability of performance, though apparently not a significant loss in average performance. Since computation time is directly related to the number of subclasses this is an important advantage of the nonparametric approach. It is, however, probably true that a parametric (normal) classifier with an adequate number of subclasses will still be competetive in terms of computation time and storage with a nonparametric classifier without subclasses. The second advantage of a nonparametric minimum distance classifier based on density histograms is that as the sample size is reduced it becomes a maximum likelihood vector classifier, provided an appropriate distance measure is used. As a maximum likelihood classifier it should, with proper programming, be relatively fast.

The main disadvantages of the nonparametric classifier LARSYSDC are the large storage requirements and relatively slow speed. Actually the storage problem can be alleviated considerably from that encountered in LARSYSDC by storing only nonempty histogram bins and the bin index. It is the storage of too many empty bins in LARSYSDC that creates the main problem. The facility to use a subset of channels from a given statistics deck is an exceedingly important capability of parametric normal classifiers. Perhaps a method could be devised to select a subset of channels for a stored multidimensional histogram but the complexity of such a method would certainly greatly exceed

the analogous procedure for the parametric normal case.

The nature of the problem of choosing a distance measure is substantially different than the nature of the parametric vs nonparametric question. The recommendation against nonparametric minimum distance classifiers is primarily based on the inability to significantly improve test accuracy with such a classifier even though it is slower and more complex. The added complexity means that for a given core storage the capabilities of a nonparametric system, in terms of number of classes and number of channels, would be considerably below the capabilities of a parametric system. With regard to the choice of distance a different situation prevails. The distance measure has only a minor impact on the complexity of the classifier and on its capabilities, (i.e., number of classes, number of channels, etc.) except possibly speed. Consequently, if a distance measure exhibits even a slight superiority it is a natural choice provided it is not unreasonably slow. On the basis of this investigation our choice for a distance measure for minimum distance classification, from amongst those distances considered, would be the JM distance. This choice applies to both the parametric and nonparametric classifiers. KL numbers are a close second choice for the parametric case. The choice of JM distance depends on three factors. (1) There is some evidence to suggest that the JM distance is superior to the other distance measures (i.e., training results) and in no case does the JM distance show up substantially inferior to

any other distance. (2) The behavior of the JM distance as a function of dimensinality for multispectral scanner data tends to resemble the behavior of the probability of correct classification. (3) Theoretically it is among the simplest of the distances to compute and has the important theoretical property of being a metric in a large space of distribution functions.

Generally as expected the classification accuracy for minimum distance classification is greater than for maximum likelihood vector classification. For the data studied the advantage of minimum distance over maximum likelihood is not very great. This we attribute to the general inseparability of the classes for the data classified and in fact suggest (but do nor verify) that for the extreme cases of very high and very low class separability minimum distance classification will afford little if any improvement in classification accuracy over maximum likelihood vector classification. The greatest potential for increased classification accuracy appears to be for data in which the classes are moderately seperable. It is probably important to mention that in the experiments performed no great care was exercised to ensure that the data in a sample was reasonably homogeneous except that each sample originated from a physical field. Thus a fair number of samples exhibited some bimodality. Greater care in this regard would probably increase performance somewhat. Offsetting this

potential increase is the fact that in a realistic system fieldswould have to be defined automatically which might in fact result in poorer field definition than was actually used.

With regard to sample definition it is important to note the definition of samples by observation space clustering should work quite well. We base this statement primarily on our experience with BOUND and NSCLAS and on the experimentally observed fact that in minimum distance classification the test sample size need not be very large to ensure reasonable performance. The reason this latter factor is so important is that for a minimum distance classification scheme based on sample definition by observation space clustering to be at all competitive timewise with other classification schemes, it is essential that the clustering time be reasonably small. This is only possible if the number of vectors clustered simultaneously remains small. The relatively good performance of minimum distance classifiers for small sample sizes makes this possible. An incidental advantageous by product of using observation space clustering to define samples in a parametric classifier is that such samples tend to be unimodal and symmetrical.

Parameter space clustering was shown to be a useful technique in the process of defining subclasses. Thus as a result of parameter space clustering the classification accuracy of flightlines 21, 23 and 24 was improved slightly

from that previously obtained for these flightlines with observation space clustering. With regard to "best" distance measure for GRPSAM the JM distance appears superior to the Divergence. The grouping method that gave the best results was product-sample-grouping with sample-grouping a very close second. In view of the small difference between PS and S grouping and the inherent statistical appeal of sample-grouping, sample-grouping is recommended for any LARS System Program or other operational programs.

The behavior of sample classification accuracy with dimensionality for minimum distance classifiers resembles the vector classification accuracy of maximum likelihood classifiers. Both typically saturate around 4 channels.

On the basis if Test performance the bin size study for LARSYSDC indicates that under the condition of the experiment (i.e., 2 or 3 channels and 121 vectors per sample), a bin size of 5 to 10 is reasonable. For training results a bin size of one appears to give the best performance but this is believed to be due to a phenomena which typically only occurs for training samples.

In concluding it should be mentioned that no comparative computation times have been given. The fact that the experiments involved a number of different programs, two computer systems (one in a time sharing mode) and the inherent dependence of processing time on the Classification

Parameters and on the manner in which the data is stored (i.e., data retrieval time is by no means negligible) makes it virtually impossible to give meaningful comparative times. Suffice it to say that to classify a typical flight-line time would be measured in fractions of an hour to hours on the IBM 360 System Model 44, and that PERFIELD is the fastest classifier, followed by LARSYSDC and LARSYSAA in that order.

LIST OF REFERENCES

LIST OF REFERENCES

1. K. S. Fu, D. A. Landgrebe, and T. L. Phillips, "Information Processing of Remotely Sensed Agricultural Data," Proc. IEEE, Vol. 57, pp. 639-654, April 1969.

2. D. A. Landgrebe, "Systems Approach to the Use of Remote Sensing", LARS Information Note 041571, Purdue University, Lafayette, Indiana, April, 1971.

3. T. M. Cover and P. E. Hart, "Nearest Neighbor Pattern Classification," IEEE Trans. on Information Theory, IT-13, pp. 21-27, January 1967.

4. D. W. Peterson, "Some Convergence Properties of a Nearest Neighbor Decision Rule," IEEE Trans. on Information Theory, Vol. IT-16, pp. 26-31, January 1970.

5. T. M. Cover, "Rates of Convergence for Nearest Neighbor Classification," Proc. 1st Ann. Hawaii Conf. on Systems Theory, January 1968.

6. E. A. Patrick, and F. P. Fischer II, "K-Nearest Neighbor Rules," School of Electrical Engineering, Purdue University, Lafayette Indiana, Tech. Report TR-EE10-34.

7. E. Fix and J. L. Hodges, Jr., "Discriminatory Analysis, Nonparametric Discrimination: Consistency Properties, "USAF School of Aviation Medicine, Randolph Field, Texas, Project No. 21-49-004, Report No. 4, February, 1951.

8. F. Rosenblatt, Principles of Neurodynamics, Perceptions and the Theory of Brain Mechanisms, Spartan Books, Washington, D.C., 1962.

9. P. Whittle, "On the Smoothing of Probability Density Functions," J. Roy. Statist. Soc. Ser. V, Vol. 20, pp. 334-343, 1968.

10. E. Parzen, "On Estimation of Probability Density Function and Mode," Ann. Math. Stat., Vol. 33, pp. 1065-1076, 1962.

11. T. Cacoullos, "Estimation of a Multivariate Density," Ann. Inst. Stat. Math. (Tokyo), Vol. 18, No. 2, pp. 179-189, 1966.

12. Z. W. Birbaum, "Distribution Free Tests of Fit for Continuous Distribution Functions," Ann. Math. Stat., Vol. 24, pp. 1-8, 1953.

13. E. Samuel and R. Bachi, "Measures of Distances of Distribution Functions and Some Applications," Metron, Vol. 23, pp. 83-122, December 1964.

14. E. L. Lehmann, "Significance Level and Power," Ann. Math. Stat., Vol. 29, pp. 1167-1176, December 1958.

15. H. Cramer, "On the Composition of Elementary Erros," Skand. Aktuarietids, Vol. 11, pp. 13-74 and 141-180, 1928.

16. R. Von Mises, "Wahrscheinlichkeitsrechnung," Leipzig-Wien, 1931.

17. D. A. Darling, "The Kolmogorov-Smirnov, Cramer-Von Mises Tests," Ann. Math. Stat., Vol. 28, pp. 823-838, December 1957.

18. W. Sahler, "A Survey on Distribution-Free Statistics Based on Distances Between Distribution Functions," Metrika, Vol. 13, pp. 149-169, 1968.

19. A. N. Kolmogorov, "Sulla Determinazione Empirica Di Une Legge Di Distribuzione," Giorn, dell'Insit. degli att., Vol. 4, pp. 83-91, 1933.

20. N. V. Smirnov, "On the Estimation of the Discrepancy Between Empirical Curves of Distribution for Two Independent Samples," Bull. Math. Univ. Moscow, Vol. 2, pp. 3-14, 1939.

21. H. Jeffreys, "An Invariant for the Prior Probability in Estimation Problems," Proc. Roy. Soc. A., Vol. 186, pp. 454-461, 1946.

22. H. Jeffreys, "Theory of Probability," Oxford University Press, 1948.

23. T. Kailath, "The Divergence and Bhattacharyya Distance Measures in Signal Selection," IEEE Trans. on Comm. Tech., Vol. COM-15, pp. 52-60, February, 1967.

24. A. Bhattacharyya, "On a Measure of Divergence Between Two Statistical Populations Defined by Their Probability Distributions," Bull. Calcutta Math. Soc., Vol. 35, pp. 99-109, 1943.

25. K. Matusita, "On the Theory of Statistical Decision Functions," Ann Institit. Stat. Math. (Tokyo), Vol. 3, pp. 17-35, 1951.

26. B. P. Adhikari and D. D. Joshi, "Distance Discrimination et Resume Exhaustif," Pbls, Inst. Stat., Vol. 5, pp. 57-74, 1956.

27. C. H. Kraft, "Some Conditions for Consistency and Uniform Consistency of Statistical Procedures," University of California Publications in Statistics, 1955.

28. S. Kullback and R. A. Leibler, "On Information and Sufficiency," Ann. Math. Stat., Vol. 22, pp. 79-86, 1951.

29. P. H. Swain and K. S. Fu, "Nonparametric and Linguistic Approaches to Pattern Recognition," LARS Information Note 051970, Purdue University, Lafayette, Indiana. June 1970.

30. P. C. Mahalanobis, "Analysis of Race Mixture in Bengal," J. Asiat. Soc. (India), Vol. 23, pp. 301-310, 1925.

31. P. C. Mahalonobis, "On the Generalized Distance in Statistics," Proc. Nat'l. Inst. Sci. (India), Vol. 12, pp. 49-55, 1936.

32. J. Keifer and J. Wolfowitz, "Consistency of the Maximum Likelihood Estimator in the Presence of Infinitely Many Incidental Parameters," Ann. Math. Stat., Vol. 27, pp. 887-906, 1956.

33. K. S. Fu and P. J. Min, "On Feature Selection in Multiclass Pattern Recognition," Tech. Report TR-EE68-17, Purdue University, Lafayette, Indiana, July 1968.

34. S. Karlin and R. N. Bradt, "On the Design and Comparison of Dichotomous Experiments," Ann. Math. Stat., Vol. 27, pp. 390-409, 1956.

35. S. M. Ali and S. D. Sivey, "A General Class of Coefficients of Divergence of one Distribution From Another," J. Roy. Stat. Soc., Ser. B, Vol. 28 pp. 131-142, 1966.

36. D. G. Lainiotis, "On a General Relationship Between Estimation, Detection, and the Bhattacharyya Coefficient," IEEE Trans. on Information Theory, Vol. IT-15, pp. 504-505, July 1969.

37. C. Stein, "Approximations of Improper Prior Measures by Prior Probability Measures," Dept. of Statistics, Stanford University, Stanford, California, Tech. Report 12, 1964.

38. K. Matusita, "On the Theory of Statistical Decision Functions," Ann. Inst. Stat. Math. (Tokyo), Vol. 3, pp. 17-35, 1951.

39. K. Matusita, "On Estimation by the Minimum Distance Method," Ann. Inst. Stat. Math. (Tokyo), Vol. 5, pp. 59-65, 1954.

40. K. Matusita, Y. Suzuki, and H. Hudimoto, "On Testing Statistical Hypothesis," Ann. Inst. Stat. Math. (Tokyo), Vol. 6, pp. 133-141, 1954.

41. K. Matusita and H. Akaike, "Decision Rules Based on the Distance for the Problems of Independence Invariance and Two Samples," Ann. Inst. Stat. Math., Vol. 7, pp. 67-80, 1956.

42. K. Matusita and M. Motoo, "On the Fundamental Theorem for the Decision Rule Based on Distance $||\ ||$," Ann. Inst. Stat. Math., Vol. 7, pp. 137-142, 1956.

43. K. Matusita, "Decision Rule Based on the Distance for the Classification Problem," Ann. Inst. Stat. Math. (Tokyo), Vol. 8, pp. 67-70, 1956.

44. K. Matusita, "Distance and Decision Rules," Ann. Inst. Math. (Tokyo), Vol. 16, pp. 305-315, 1964.

45. K. Matusita, "Classification Based on Distance in Multivariate Gaussian Case," Proc. 5th Berkeley Symposium on Math. Stat. and Prob., Vol. 1, pp. 299-304, 1967.

46. J. Wolfowitz, "Consistent Estimations of the Parameters in a Linear Structural Relationship," Skand. Aktuarietids, pp. 132-151, 1952.

47. J. Wolfowitz, "Estimation by the Minimum Distance Method," Ann. Inst. Stat. Math. (Tokyo), Vol. 5, -p. 9-23, 1953.

48. J. Wolfowitz, "Estimation by the Minimum Distance Method in Nonparametric Difference Equations," Ann. Math. Stat., Vol. 25, pp. 203-217, 1954.

49. J. Wolfowitz, "The Minimum Distance Method," Ann. Math. Stat, Vol. 28, pp. 75-88, 1957.

50. S. Das Gupta, "Nonparametric Classification Rules," Sankhyā, Indian Jour. of Stat., Series A, Vol. 26, pp. 4-30, 1964.

51. T. Cacoullos, "Comparing Mahalanobis Distance I: Comparing Distances between k Known Populations and Another Unknown," Sankhyā, Indian Jour. Stat., Series A, Vol. 27, pp. 1-22, March 1965.

52. T. Cacoullos, "Comparing Mahalanobis Distances II: Bayes Procedures When the Mean Vector are Unknown," Sankhya, Indian Jour. Stat., Series A, Vol. 27, pp. 23-32, March 1965.

53. M. S. Srivastava, "Comparing Distances Between Multivariate Populations - The Problem of Minimum Distance," Ann. Math. Stat., Vol. 38, pp. 550-556, April 1967.

54. W. Hoeffding and J. Wolfowitz, "Distinguisability of Sets of Distributions," Ann. Math. Stat., Vol. 29, pp. 700-718, September 1958.

55. Blackwell, "Comparison of Experiments," Proc. 2nd Berkeley Symposium on Probability and Statistics, Berkeley, California, University of California Press, Vol. 1, pp. 93-102, 1951.

56. S. E. Estes, "Measurement Selection for Linear Discriminants Used in Pattern Classification," IBM Corporation, San Jose, Calif., Research Report RJ-331, April 1956.

57. D. C. Allias, "The Selection of Measurements for Prediction," Stanford Electronics Lab., Stanford, Calif., Tech. Report 6103-9, November 1964.

58. G. F. Hughes, "On the Mean Accuracy of Statistical Pattern Recognizers," IEEE Trans. Inform. Theory, Vol. IT-14, pp. 55-63, Jan. 1968.

59. K. Abend, T. J. Harley, Jr., B. Chandrasekaran, and G.F. Hughes, "Comments on 'On the Mean Accuracy of Statistical Pattern Recognizers'", *IEEE Trans, Inform. Theory*, Vol. iT-15, pp. 420-423, May 1969.

60. L, Kanal and B. Chandrasekaran, "On Dimensionality and Sample Size in Statistical Pattern Classification," Proc. 1968 Nat. Electronics Conf., pp. 2-7.

61. T. W. Anderson, "An Introduction to Multivariate Statistical Analysis," John Wiley & Sons, 1957.

62. R. A. Holmes and R. B. MacDonald, "The Physical Basis of System Design for Remote Sensing in Agriculture," Proc. IEEE, Vol. 57, pp. 629-639, April 1969.

63. "Remote Multispectral Sensing in Agriculture," Vol. 3 (Annual Report), Laboratory for Agricultural Remote Sensing, Purdue University, Lafayette, Indiana, 1968.

64. D. A. Landgrebe and LARS Staff, "LARSYSAA, A Processing System for Airborne Earth Resources Data," LARS Information Note 091968, Purdue University, Lafayette, Indiana September 1968.

65. K. S. Fu, "Sequential Methods in Pattern Recognition and Machine Learning," New York Academic Press, 1968.

66. P. H. Swain, T. V. Robertson, and A. G. Wacker, "Comparison of the Divergence and B-Distance in Feature Selection," LARS Information Note 020871, Purdue University, Lafayette, Indiana, February 1971.

67. T. Huang, "Per Field Classifier for Agricultural Applications," LARS Information Note 060569, Purdue University, Lafayette, Indiana, June 1969.

68. G. H. Ball and D. J. Hall, "ISODATA, A Novel Method of Data Analysis and Pattern Classification Stanford Research Institute, Melno Park, Calif., pp. 1-16.

69. A. G. Wacker and D. A. Landgrebe, "Boundaries in Multispectral Imagery by Clustering," 1970 IEEE Symposium on Adaptive Processes (9th) Decision and Control, pp. X14.1-X14.8, December 1970.

70. G. H. Ball, "Data Analysis in the Social Sciences: What About the Details," IEEE Proc. Fall Joint Comp. Conf., pp. 533-559.

71. F. J. Rohlf, "Adaptive Hierarchical Clustering Schemes," Systematic Zoology, Vol. 19, No. 1, pp. 58-82, March 1970.

72. P. Reddy, P. A. Wintz and D. A. Landgrebe, "A Linear Transformation for Data Compression and Feature Selection in Multispectral Imagery" LARS Information Note 072071, Purdue University, Lafayette, Indiana, July 1971.

73. E. G. Henrichon, "On Nonparametric Methods for Pattern Recognition," Ph.D. Thesis, Purdue University, Lafayette, Indiana, 1969.

APPENDICES

Appendix A

Some Results on the Swain-Fu Distance

A.1 Alternate Form of Swain-Fu Distance

For distribution $F^{(1)}$ and $F^{(2)}$ with means $\underline{\mu}^{(1)}$ and $\underline{\mu}^{(2)}$ and nonsingular covariances $\Sigma^{(1)}$ and $\Sigma^{(2)}$ the Swain-Fu distance is given by[29]

$$T = \frac{|\underline{\mu}^{(1)} - \underline{\mu}^{(2)}|}{D_1 + D_2} \qquad \text{A.1.1}$$

where

$$D_k = \left\{\frac{|\underline{\mu}^{(1)} - \underline{\mu}^{(2)}|^2 \det(\Sigma^{(k)})^{(q+2)}}{\sum_{i=1}^{q}\sum_{j=1}^{q}(\Sigma_{ij}^{(k)})(\mu_i^{(1)} - \mu_i^{(2)})(\mu_j^{(1)} - \mu_j^{(2)})}\right\}^{1/2} \quad k = 1,2 \qquad \text{A.1.2}$$

and $\Sigma_{ij}^{(k)}$ is the ijth cofactor of $\Sigma^{(k)}$ $k = 1,2$. Since $\det(\Sigma^{(k)}) \neq 0$ dividing numerator and denominator by this quantity we can show by direct expansion that an alternate form of D_k is

$$D_k = \left\{\frac{|\underline{\mu}^{(1)} - \underline{\mu}^{(2)}|^{2(q+2)}}{\mathrm{tr}\left\{\frac{\mathrm{Adj}(\Sigma^{(k)})}{\det(\Sigma^{(k)})}(\underline{\mu}^{(1)} - \underline{\mu}^{(2)})(\underline{\mu}^{(1)} - \underline{\mu}^{(2)})^t\right\}}\right\}^{1/2} \quad k = 1,2 \qquad \text{A.1.3}$$

Where $\mathrm{Adj}(\Sigma^{(k)})$ is the adjoint of $\Sigma^{(k)}$ and tr is the trace. From the definition of the adjoint A.1.3 can also be written as

$$D_k = \left\{\frac{|\underline{\mu}^{(1)} - \underline{\mu}^{(2)}|^{2(q+2)}}{\mathrm{tr}\{(\Sigma^{(k)})^{-1}(\underline{\mu}^{(1)} - \underline{\mu}^{(2)})(\underline{\mu}^{(1)} - \underline{\mu}^{(2)})^t\}}\right\}^{1/2} \qquad \text{A.1.4}$$

Note that D_k is indeterminant if $\underline{\mu}^{(1)}$ and $\underline{\mu}^{(2)}$ are equal. The reason for this is that the direction of the line joining $\underline{\mu}^{(1)}$ and $\underline{\mu}^{(2)}$ is not defined. The distance from $\underline{\mu}^{(k)}$ to the ellipsoid of concentration is, however, not zero regardless of the direction, since $\Sigma^{(k)}$ is not singular. Consequently, from A.1.1 the Swain-Fu distance between classes with equal means is zero. Consequently, we can write

$$T = 0 \qquad \underline{\mu}^{(1)} = \underline{\mu}^{(2)}$$

$$T = \frac{\sqrt{c_1}\sqrt{c_2}}{\sqrt{c}_1+\sqrt{c}_2}\,(q+2)^{-1/2} \qquad \underline{\mu}^{(1)} \neq \underline{\mu}^{(2)} \qquad \text{A.1.5}$$

where $c_i = \mathrm{tr}\{(\Sigma^{(k)})^{-1}\,(\underline{\mu}^{(1)}-\underline{\mu}^{(2)})(\underline{\mu}^{(1)}-\underline{\mu}^{(2)})^t\}$

From A.1.5 it follows that T is invariate under linear transformations because the trace is invariate under linear transformations. Note also that c_1 and c_2 are positive by virtue of the fact that D_1 and D_2 are positive when $\underline{\mu}^{(1)} \neq \underline{\mu}^{(2)}$.

A.2 Upper Bound on SF Distance for Given Divergence

We derive an expression for the upper bound on the SF distance for a given Divergence. We need only consider the case where the means are not equal, since otherwise regardless of the divergence the SF distance is zero, which is certainly not the upper bound. From A.1.5 we can write

$$(q+2)T^2 = \frac{1}{C^2} \qquad \text{A.2.1}$$

where $$C^2 = \frac{1}{c_1} + \frac{2}{\sqrt{c_1 c_2}} + \frac{1}{c_2}$$

Now since the geometric mean of two positive numbers is less than or equal to their arithmetic mean, it follows that

$$(c_1+c_2)C^2 = 6 + \frac{c_2}{c_1} + \frac{c_1}{c_2}. \qquad \text{A.2.2}$$

Direct minimization of the right hand side of A.2.2 with respect to c_2/c_1 yields

$$C^2 \geq \frac{8}{c_1+c_2}. \qquad \text{A.2.3}$$

Combining A.2.3 and A.2.1 we have

$$(q+2)T^2 \leq \frac{c_1+c_2}{8}. \qquad \text{A.2.4}$$

But from the definition of c_1, c_2 and Divergence

$$c_1 + c_2 = 2J - tr\{[\Sigma^{(1)}-\Sigma^{(2)}][(\Sigma^{(2)})^{-1} - (\Sigma^{(1)})^{-1}]\} \qquad \text{A.2.5}$$

$$\leq 2J, \qquad \text{A.2.6}$$

where the last inequality follows because the tr $\{\cdot\}$ is greater than or equal to zero. This is readily seen by considering diagonal covariance matrices, which by virtue of the invariance of the trace under linear transformations is equivalent to the general case. Finally combining A.2.5 and A.2.4 we have

$$T \leq \sqrt{\frac{J}{4(q+2)}} \qquad \text{A.2.7}$$

Appendix B

Miscellaneous Result Pertaining to the Separability Measure R

B.1 Expected Value of D*

By definition

$$D^* = \left(\sum_{j-1}^{q} (X_j^* - Y_j^*)^2 \right)^{1/2} \qquad \text{B.1.1}$$

where $\underline{X}^*$, $\underline{Y}^* \sim N(\underline{\mu}', \sigma^2 I)$. Let

$$d_j^* = X_j^* - Y_j^* \qquad \text{B.1.2}$$

then

$$D^{*2} = \sum_{j=1}^{q} (X_j^* - Y_j^*)^2 = \sum_{j=1}^{q} d_j^{*2} \qquad \text{B.1.3}$$

Now $X_j^* \sim N(\mu', \sigma^2)$ and $Y_j^* \sim N(\mu', \sigma^2)$, therefore $d_j^* \sim N(0, 2\sigma^2)$ and $d_j^*/(\sqrt{2}\sigma) \sim N(0,1)$. Furthermore, the d_j^* are independent since $\underline{X}^*$ and $\underline{Y}^*$ are independent vectors. Consequently $Z = D^*/(2\sigma^2)$ is the sum of the square of q independent $N(0,1)$ random variables and consequently has the Chi-Square distribution with q degrees of freedom. Now

$$E(D^*) = \sqrt{2}\sigma E(\sqrt{Z}) \qquad \text{B.1.4}$$

$$= \sqrt{2}\sigma \int_0^{\infty} \sqrt{z} \, \frac{1}{\Gamma(q/2)2^{/2}} \, z^{\frac{1}{2}q-1} \, e^{z/2} \, dz$$

This be direct computation yields

$$E(D^*) = 2\sigma \frac{\Gamma(\frac{q+1}{q})}{\Gamma(\frac{q}{2})} \qquad \text{B.1.5}$$

B.2 Expected Value of D**

By definition

$$D^{**} = \left(\sum_{j=1}^{q} (X_j^{**} - Y_j^{**})^2 \right)^{1/2} \qquad \text{B.2.1}$$

where $X^{**} \sim N(\underline{\eta}, \sigma^2 I)$ and $Y^{**} \sim N(-\underline{\eta}, \sigma^2 I)$. Let

$$d_j^{**} = X_j^{**} - Y_j^{**} \qquad \text{B.2.2}$$

then

$$D^{**2} = \sum_{j=1}^{q} (X_j^{**} - Y_j^{**})^2 = \sum_{j=1}^{q} d_j^{**2} \qquad \text{B.2.3}$$

Now $X_j^{**} \sim N(\mu, \sigma^2)$ and $Y_j^{**} \sim N(-\mu, \sigma^2)$, therefore $d_j^{**} \sim N(2\mu, 2\sigma^2)$ and $d_j^{**}/(\sqrt{2}\sigma) \sim N(2\mu, 1)$. Furthermore the d_j^{**} are independent since $\underline{X}^{**}$ and $\underline{Y}^{**}$ are independent vectors. Therefore $Z = D^{**2}/(2\sigma^2)$ is the sum of q independent $N(2\mu, 1)$ variables and consequently has the Noncentral Chi-Square distribution with parameters q and $2q\mu^2/\sigma^2 = (S/\sqrt{2})^2$ (i.e., $NCX^2(q, (S/\sqrt{2})^2)$ with pdf.

$$f(z) = e^{-\frac{1}{2}(S/2)^2} \sum_{r=0}^{\infty} \frac{1}{r!} (S^2/4)^r f_{q+2r}(z) \qquad \text{B.2.4}$$

where $f_{q+2r}(z)$ is the Chi-Square density with q+2r degrees of freedom. This can be put in a more convenient form

$$f(z) = \frac{1}{2} e^{-\frac{1}{4}(S^2+2z)} (2z/S^2)^{\frac{1}{4}(q-2)} I_{\frac{1}{2}q-1}(S\sqrt{z}/2) \qquad \text{B.2.5}$$

where $I_\nu(x)$ is the modified Bessels function

$$I_\nu(x) = \sum_{r=0}^{\infty} \frac{(x/2)^{\nu+2r}}{r!\Gamma(\nu+1+r)} \qquad \text{B.2.6}$$

Now

$$E(D^{**}) = \sqrt{2}\sigma E(\sqrt{Z}) \qquad \text{B.2.7}$$

$$= \sqrt{2}\sigma \int_0^{\infty} \sqrt{z}\, \frac{1}{2}\, e^{-\frac{1}{4}(S^2+2z)} (2z/S^2)^{\frac{1}{4}(q-2)} I_{\frac{1}{2}q-1}(S\sqrt{z}/\sqrt{2})\, dz$$

Using integral tables this yields

$$E(D^{**}) = 2\sigma \frac{\Gamma(\frac{q+1}{2})}{\Gamma(\frac{q}{2})} e^{-(S/2)^2} \Phi(\tfrac{q+1}{2}, \tfrac{q}{2}, (S/2)^2) \qquad \text{B.2.8}$$

B.3 Limiting Form of R(S,q)

From Eq . 3.2.3.3 R(S,q) is given by

$$R(S,q) = 1 + \frac{1}{1!}(S_d/2)^2 + \sum_{n=2}^{\infty} (-1)^{n+1} \frac{1.3.5 \ldots (2n-3)\, q^n}{q(q+2)(q+4)\ldots(q+2n-2)n!}$$

$$(S_d/2)^{2n} \qquad \text{B.3.1}$$

Since this is a power series in $(S_d/2)^2$, the limit of the sum as the dimensionality approaches infinity, is the sum of the limits and hence

B.3.2

$$\lim_{q\to\infty} R(S,q) = 1 + \frac{1}{1!}(S_d/2)^2 + \sum_{n=2}^{\infty} \frac{(-1)^{n+1} 1.3.5\ldots(2n-3)}{n!} (S_d/2)^{2n}$$

Let the nth term in B.3.2 be t_n. Then since B.3.2 is an alternating series it converges only if

$$\lim_{n\to\infty} |t_n| = 0 \qquad \text{B.3.3}$$

But

$$\lim_{n\to\infty} |t_n| = \lim_{n\to\infty} \frac{1.3.5\ldots(2n-3)}{n!} (S_d/2)^{2n}$$

$$= \lim_{n\to\infty} \frac{(n)(n+1)(2n)!}{2^{n-2}(2n)(2n-1)(2n-2)n!n!} (S_d/2)^{2n} \qquad \text{B.3.4}$$

Using Stirlings factorial formula for large N

$$\lim_{n\to\infty} |t_n| = \lim_{n\to\infty} \frac{\sqrt{4n\pi}\ (2n/e)^{2n}}{2^n(2n-1)\ 2n\pi(n/e)^{2n}} (S_d/2)^{2n} \qquad \text{B.3.5}$$

$$= \lim_{n\to\infty} \frac{1}{2^n(2n-1)\sqrt{n\pi}} (S_d/2)^{2n}$$

This limit is zero only if $S_d \leq \sqrt{2}$.

Page Intentionally Left Blank

Pgs - 288 - 342

APPENDICES C,
" D.
" E.
"

Appendices C, Description of Test and Training Field Decks; D, Control Card Language; and E, Program Descriptions have been omitted in this printing to conserve space. They may be purchased, beginning February, 1972 from University Microfilms, 300 N. Zeeb Road, Ann Arbor, Michigan 48106.

Appendix F

BOUND: A Boundary Tracing Program

The principle upon which the program BOUND is based is clustering in the observation space. The scene under investigation is partitioned into square regions called "Boundary Cells" such that the union of the Boundary Cells is the whole scene (except for the narrow border). Each Boundary Cell consists of a square array of image resolution elements (IRE's). Boundaries are found seperately for each Boundary Cell and the union of these boundaries constitutes the boundaries for the scene.

To locate the boundaries for a given Boundary Cell a clustering algorithm is used to effect a nonsupervised classification of the vectors that originate from IRE's in an area slightly larger (to provide some overlap) than a Boundary Cell. This results in a spatial "Clustered Array" in which each IRE is represented by the group number (i.e., class number) to which it has been assigned. The "Clustered Array" is scanned in both directions and a boundary is assumed to exist whenever k (user specified) or more IRE's on each side of the boundary belong to a different class. This definition of a boundary provides for some spatial smoothing but necessitates the overlap and narrow border mentioned above.

Experimentally it is found that for the 12 to 13 channel multispectral scanner data presently available, a reasonable compromise between performance and computation time is achieved by using 3 or 4 channels of data, a Boundary Cell size of about 5 x 5 IRE's and by setting k equal to two.[69] It is probably not coincidental that principal-component analysis of multispectral scanner data suggests that 3 or 4 principal components are sufficient to represent similar data with small mean squared error.[72]

unclassified

Security Classification

DOCUMENT CONTROL DATA – R & D

(Security classification of title, body of abstract and indexing annotation must be entered when the overall report is classified)

1. ORIGINATING ACTIVITY *(Corporate author)*
School of Electrical Engineering
Purdue University

2a. REPORT SECURITY CLASSIFICATION
unclassified

2b. GROUP

3. REPORT TITLE
Minimum Distance Approach to Classification

4. DESCRIPTIVE NOTES *(Type of report and inclusive dates)*
scientific report

5. AUTHOR(S) *(First name, middle initial, last name)*
Arthur G. Wacker
David A. Landgrebe

6. REPORT DATE
October 1, 1971

7a. TOTAL NO. OF PAGES
306

7b. NO. OF REFS

8a. CONTRACT OR GRANT NO.
NASA Grant NGR 15-005-112

b. PROJECT NO.

c.

d.

9a. ORIGINATOR'S REPORT NUMBER(S)
TR-EE 71-37

9b. OTHER REPORT NO(S) *(Any other numbers that may be assigned this report)*
LARS Information Note #100171

10. DISTRIBUTION STATEMENT
unlimited

11. SUPPLEMENTARY NOTES

12. SPONSORING MILITARY ACTIVITY
National Aeronautics and Space Administration

13. ABSTRACT

This report investigates minimum distance classification. Results of a thorough literature survey for suitable distance measures are presented. Both theoretical and experimental investigations based on multispectral scanner data are reported.

Theoretically, it is shown that minimum distance classification is equivalent to a form of maximum likelihood sample classification, and that for a parametric case, it is equivalent to nearest-neighbor classification in the parameter space. A separability measure defined in terms of random samples provides insight into some experimentally observed effects of dimensionality.

The experimental investigation of minimum distance classification is based on a supervised parametric (normal) minimum distance classifier PERFIELD and a supervised nonparametric minimum distance classifier LARSYSDC.

The principal experimental results are: 1) the Jeffreys-Matusita distance is a good, general purpose distance measure; 2) the minimum distance classification accuracy averaged 5 to 10% better; than in the equivalent case using a per vector approach; no distance measure was consistently superior for classifying test samples. The nonparametric classifier was also not consistently superior to the parametric classifier PERFIELD in these circumstances; 4)classifier accuracy can be improved only slightly by using more than four channels for this data.

DD FORM 1 NOV 65 1473

-over-

unclassified

Security Classification

14 KEY WORDS	LINK A		LINK B		LINK C	
	ROLE	WT	ROLE	WT	ROLE	WT
13. Abstract (continued) Results regarding training set size and bin size are also given.						