N73-26615
CR-128988

Johnson Space Center
Earth Observations Division
Science and Application Directorate

# Theory and Analysis of Statistical Discriminant Techniques as Applied to Remote Sensing Data

## FINAL TECHNICAL REPORT

Department of Mathematics and Statistics
Texas Tech University
Lubbock, Texas 79409

May 1973

ANNUAL RESEARCH REPORT

# THEORY AND ANALYSIS OF STATISTICAL DISCRIMINANT TECHNIQUES AS APPLIED TO REMOTE SENSING DATA

Patrick L. Odell
Principal Investigator
The University of Texas at Dallas
Dallas, Texas 75230

## Acknowledgments

Most of this report covers work done under NASA Manned Spacecraft Center, Houston, Texas, Contract NAS-12775, to Texas Tech University, Lubbock, Texas, from 1 June 1972 to 31 May 1973. The research work described here has been accomplished in collaboration with Dr. J. P. Basu and Dr. R. S. Chhikara, research associates; and Dr. T. Boullion, Dr. B. Duran and Dr. H. Gray of the Texas Tech University Statistics faculty and their graduate students: W. A. Coberly, J. Tubbs and Mike Pore.

## Preface

Recently it has been deemed important that remote sensing data analysis technology be developed so that earth resources programs initiated by several agencies, both in the United States and abroad, can be utilized effectively to assess fully available earth resources. Techniques when judiciously used could provide answers to certain important and pertinent problems facing mankind. Under the technology of remote sensing the multichannel scanning devices are employed for securing earth information and identifying response pattern for each natural (earth) resource as completely and reliably as possible. This, however, gives rise to multivariate data and hence, handling of the large amounts of data requires a careful consideration of both the underlying physical properties and the data analysis techniques. The foremost problem after a region has been scanned from above by using airborne data scanning devices is that of recognition of different earth resources in the region. This emphasizes the importance of developing certain classification procedures that can meet any urgent demand of identification of the scanned ground.

The greater part of the work presented in this report is addressed to the classification problem from a statistical viewpoint as applied to the remote sensing technology. It is easy to visualize that a remote sensing data will generally inherent a high degree of variability. This suggests being careful in associating a stochastic model to the underlying earth resources. The empirical approach for determining a model may not be reliable because of uncertainity in identifying the resource of an observation obtained via remote sensors. This has led investigators to assume the usual law of errors, and thereby, to consider the multivariate Gaussian probability model for the earth resource classes. Though the Gaussian model has many virtues, it is somewhat an arbitrary choice, and so a careful screening of the data is called for prior to associating any such model with the underlying classes. This and further consideration of computational advantages has led us to suggest alternative models which we have called normed exponential densities in one of the given reports. The use of these density models which are appropriate in several physical situations provides an exact solution for the probabilities of classifications (i.e. probability of misclassification and probability of correct classification) associated with the Bayes discriminant procedure even when the covariance matrices are unequal (a property not enjoyed in the normal density case). The computational difficulties that one faces in a complex situation such as remote sensing can be reduced to a certain degree if suggested normed densities are used as class models.

In another report the problem of finding the probability that a random instantaneous field of views of a multispectral scanning device consists of areas across class boundaries is discussed. Based upon an empirical study an estimate of this probability for our example was approximately .4. Such an amount of contamination is significant and it points out another potential irregularity that any remotely sensed data may have.

Some of our reports deal with the problem of estimating probabilities of misclassification for the Bayes procedure as applied to Gaussian distributed classes, a well-studied problem in the classification theory. Both theoretical and empirical work has been done toward the investigation of the estimation problem. Further, the relationship between sample-size, feature-size, and Euclidean distance measure between classes was evaluated using a Monte Carol study. These results indicate that if the ratio of a sample-size to a feature-size becomes small, the probability of misclassification is increased accordingly.

Also, we have obtained the minimum variance unbiased estimates (MVUE) of the probabilities of misclassification for the case of univariate normal probability models for the classes. Other ad hoc procedures previously given in the literature such as table look-up technique, have also been studied and certain modifications are suggested for making these more useful in remote sensing application.

One report deals with clustering using dynamic programming. The problem of dimensionality of the observation vector has been treated from the point of both computation and reduction. The reduction of dimensionality has been related to the probability of misclassification under the Bayes disciminant procedure. It is pointed out that such basis of probability of misclassification for reduction is not possible by using Karhunen-Loeve expansion method; and Wilks' dispersion techniques is recommended for the purpose of reducing dimensionality as it involves smaller risk in losing information concerning separability of the classes.

The present work is a continuation of the research being conducted by us on the subject of remote sensing data analysis techniques. Besides the classification and feature selection problems, other statistical investigations such as the determination of sample size and the accuracy of estimates have been discussed in one of the reports. For the problem of estimating proportions for different categories of objects a model has been developed which takes into account the uncertainty that exists in classifying an object measured by remote sensor. More related problems are being considered for the sampling scheme in obtaining sample observations from any remotely sensed data and deriving estimates of the actual amount or size of underlying classes.

Our reports as listed in the table of contents contain results deserving of publication in professional journals. Publications based upon the results of the following two reports have been accepted and are expected to appear soon in the respective journals:

(a) Discriminant analysis using certain normed exponential densities, (Journal of Pattern Recognition, 1973)

(b) On the table look-up in discriminate analysis, (Journal of Statistical Computation and Simulation, 1973)

Other reports submitted for publication and being reviewed are:

(c) An empirical study of classification by thresholding, (IEEE Transactions on Computers)

(d) A space application of an extension of the buffon needle problem, (Journal of American Statistical Association)

(e) Concerning dimension reduction in discriminate analysis, (IEEE, Information Theory)

(f) Effect of intraclass correlation among training samples on the linear discrimination procedure, (Journal of Pattern Recognition)

(g) Estimation of proportion of objects and determination of training sample-size in a remote sensing application, (Journal of American Statistical Association).

# TABLE OF CONTENTS

# An Empirical Study of Classification by Thresholding[1]

by

J. Tubbs, B. S. Duran, T. L. Boullion, and P. L. Odell[2]

## 1. Introduction

Consider m populations $\pi_1, \pi_2, \ldots, \pi_m$ and suppose each individual in the union of these populations possesses p common observable characteristics $c_1, c_2, \ldots, c_p$. The observed values of an individual are denoted by $x = (x_1, \ldots, x_p)^T$, where $x_j$ denotes the observed value of $c_j$. Let $p_1(x), p_2(x), \ldots, p_m(x)$ denote m known multivariate probability density functions of the p-dimensional observation vector x and $q_1, q_2, \ldots, q_m$ be known a priori probabilities that an individual, I, be selected from a population $\pi_1, \pi_2, \ldots, \pi_m$, respectively.

The classical discriminate analysis problem consists of formulating a technique for assigning an individual I selected at random from $\bigcup_{i=1}^{m} \pi_i$ into one of the m populations. There have been various techniques proposed for solving the problem, of which the Bayesian solution is optimal, in the sense that it minimizes the expected cost of misclassification.

In various applications of discriminate analysis, for example in the analysis of remote sensing data [6], the amount of computation time is immense. Thus it is desirable to develop a tech-

[1] This research was supported in part by NASA Manned Spacecraft Center under Contract NAS 9 - 12775.

[2] Presently with the University of Texas at Dallas.

nique which either reduces the number of calculations by reducing the dimension of the problem or which judiciously assigns individual I to its "most likely" population, while maintaining approximately the optimality of the Bayesian procedure. In this paper we investigate a discrimination procedure of the latter type, which is called <u>classification</u> <u>by</u> <u>thresholding</u> as formulated by Minter and Hallum [3].

Before considering the thresholding procedure we first review the maximum likelihood classification technique. Suppose the observed value $x = (x_1, x_2, \ldots, x_p)^T$ is to be classified into one of m populations $\pi_1, \pi_2, \ldots, \pi_m$. Assuming $q_i = q$, $i = 1, 2, \ldots, m$, the maximum likelihood procedure is to assign x to $\pi_j$ if $p_j(x) > p_i(x)$ for all $i \neq j$. For unequal a priori probabilities $q_1, q_2, \ldots, q_m$ the procedure becomes that of assigning x to $\pi_j$ if $q_j p_j(x) > q_i p_i(x)$ for all $i \neq j$.

In the thresholding procedure the maximum likelihood procedure has been reformulated in the following manner:

1. Obtain an observation x to be classified.
2. Select a density $p_i(x)$ and evaluate it at $x = X$ where the index i is such that $q_i > q_j$ for all $j \neq i$.
3. Compare $p_i(x)$ with threshold $T_i$ where

   $$T_i = \max_{j \neq i} t_{ij}$$

   If $p_i(x) > T_i$, classify X into population $\pi_i$ and go to step 1, otherwise go to step 4.
4. If $p_i(x) < T_i$ go to step 2 and select another density function (next largest $q_i$). If all the density functions

have been evaluated go to step 5.

5. Assign X using the maximum likelihood decision rule.

One would expect the computation time to be reduced by using the thresholding technique since the thresholds can be computed prior to the discrimination operation and need to be computed only once.

Consider the following example and figure to clarify the thresholding procedure.

Example 1.1. Let $m = 3$, $p = 1$, $q_1 > q_3 > q_2$ and suppose $T_1$, $T_2$, and $T_3$ are known. Since $q_1 = \max \{q_1, q_2, q_3\}$, X is "most likely" to be a member of $\pi_1$. Hence we first compute $q_1p_1(x)$ and compare it with $T_1$. If $q_1p_1(x) > T_1$, X is assigned to $\pi_1$; otherwise, compute $q_3p_3(x)$ and compare it with $T_3$. If X is not assigned to $\pi_3$, i.e., $q_3p_3(x) < T_3$, then the maximum likelihood procedure is used. If the classification is done according to the maximum likelihood procedure, then the actual classification will take longer than the usual Bayesian procedure, due to the extra time involved in thresholding. However, if X is classified in $\pi_1$ considerable amount of time has been saved since only one density function has had to be evaluated.

In this paper we compare the thresholding and maximum likelihood procedures for four choices of three populations.

## 2. Comparison of Thresholding and Maximum Likelihood Procedures

Let $p_1(x), p_2(x), \ldots, p_m(x)$ be m multivariate normal probability density functions and let $q_1, q_2, \ldots, q_m$ be m a priori probabilities corresponding to m normal populations $\pi_1, \pi_2, \ldots, \pi_m$. Then

$$q_i p_i(x) = \frac{q_i \exp[-1/2(x - \hat{\mu}_i)^T \hat{\Sigma}_i^{-1}(x - \hat{\mu}_i)]}{(2\pi)^{p/2} |\Sigma_i|^{1/2}}$$

where $\hat{\mu}_i$ is the estimate of the mean vector, $\hat{\Sigma}_i$ is the estimate of the covariance matrix for population $\pi_i$, and p is the number of characteristics measured.

If $t_{ij}$ denotes the threshold for populations $\pi_i$ and $\pi_j$ then

$$t_{ij} = \begin{cases} \min\{q_i p_i(\hat{\mu}_i),\ q_j p_j(\hat{\mu}_j)\}, \text{ if } & q_i p_i(\hat{\mu}_i) - q_j p_j(\hat{\mu}_i) \text{ and } \\ & q_i p_i(\hat{\mu}_j) - q_j p_j(\hat{\mu}_j) \text{ have same sign.} \\ \max\limits_{x \in E} (q_i p_i(x) = q_j p_j(x)), \text{ otherwise,} & \end{cases}$$

where $E = \{x\colon q_i p_i(x) = q_j p_j(x)\}$. Minter and Hallum [3] have developed a procedure by which the set E can be determined. They define $t_{ij}$ to be

$$t_{ij} = \begin{cases} \min\{\ln \frac{q_i^2}{|\hat{\sum}_i|}, \ln \frac{q_j^2}{|\hat{\sum}_j|}\}, & \text{if } q_ip_i(\hat{\mu}_i) - q_jp_j(\hat{\mu}_i) \text{ and } \\ & q_ip_i(\hat{\mu}_j) - q_jp_j(\hat{\mu}_i) \text{ have} \\ & \text{the same sign} \\ \min\{\ln |\hat{\sum}_i| - 2 \ln q_i + (x^{(k)} - \hat{\mu}_i)^T \hat{\sum}_i^{-1}(x^{(k)} - \hat{\mu}_i) , & \\ & \text{otherwise} \end{cases}$$

where the $x^{(k)}$ are "candidates" determined by their procedure. Then the threshold for population $\pi_i$ is

$$T_i = \min_{j \neq i} t_{ij} \quad , \quad j = 1,2,\ldots,m,$$

where the classification is carried out as before except x is classified in $\pi_i$ if

$$\ln |\hat{\sum}_i| - 2 \ln q_i + (x - \hat{\mu}_i)^T \hat{\sum}_i^{-1}(x - \hat{\mu}_i) < T_i.$$

## 3. A Monte Carlo Evaluation

For the Monte Carlo simulation m = 3, p = 3, and n = 100. The 100 samples from each population were generated using the multivariate normal random generator described in [4]. Four trials were considered in each of which thresholding was compared with the classical Bayes technique when the parameters are unknown and must be estimated. The same set of covariance matrices $\Sigma_1$, $\Sigma_2$, and $\Sigma_3$ were used for all four trials. The four trials then differed only in "separation" among the mean vectors. The following covariance matrices were used in each of the four trials:

$$\Sigma_1 = \begin{pmatrix} 400 & -240 & -200 \\ -240 & 400 & 360 \\ -200 & 360 & 400 \end{pmatrix} ,$$

$$\Sigma_2 = \begin{pmatrix} 400 & 240 & -200 \\ 240 & 400 & -360 \\ -200 & -360 & 400 \end{pmatrix} ,$$

$$\Sigma_3 = \begin{pmatrix} 400 & -240 & 200 \\ -240 & 400 & -360 \\ 200 & -360 & 400 \end{pmatrix} .$$

The mean vectors for each trial were

Trial 1: $\mu_1 = (125, 150, 175)^T$,

$\mu_2 = (150, 175, 125)^T$,

$\mu_3 = (175, 125, 150)^T$,

Trial 2: $\mu_1 = (115, 150, 185)^T$,

$\mu_2 = (150, 185, 115)^T$,

$\mu_3 = (185, 115, 150)^T$.

Trial 3: $\mu_1 = (100, 150, 200)^T$,

$\mu_2 = (150, 200, 100)^T$,

$\mu_3 = (200, 100, 150)^T$.

Trial 4: $\mu_1 = (50, 150, 250)^T$,

$\mu_2 = (150, 200, 50)^T$,

$\mu_3 = (200, 50, 150)^T$.

The results of the comparison of thresholding and Bayes procedure for each of the four trials is given in Table 1. The time of 650 (.01 seconds) in the last row of the table was obtained only for trial 1. However, since the only difference among the four trials was the choice of mean vectors, one would expect about 650 (.01 seconds) for trials 2, 3, and 4, also.

The a priori probabilities used were $q_1 = .5$, $q_2 = .3$, and $q_3 = .2$. The thresholding classification procedure was carried out by classifying 100 observations from $\pi_1$ (say), then 100 observations from $\pi_2$, and finally 100 observations from $\pi_3$. This is not the usual manner in which the classification should be carried out, however, this method of classifying the 300 observations will yield the minimum time for classification for thresholding. This allows us to assess the "best" that can be done by thresholding. For example, thresholding took 4.37 seconds as compared with 6.50 seconds for the Bayes procedure in trial 4 in Table 1. In reality, however, thresholding would yield a value greater than 4.37 seconds whereas the Bayes procedure would still take 6.50 seconds.

In remote sensing applications, such as in per field classification [2], the measurement vectors corresponding to a "small" area consisting of adjacent resolution cells, would tend to be from the same population. Thus one obtains a set of observations from one population, followed by a set from another population, etc. The thresholding procedure would perform better (timewise) for data observed in this fashion than for data taken in the usual manner. The situation described in the previous paragraph is an extreme case of data observed in a remote sensing application.

## 4. Feasibility of Thresholding as a Discriminate Technique

It is evident from our empirical study that there are situations in which thresholding is a feasible procedure, and situations when one would be better off using the classical Bayes technique. According to Table 1 thresholding is a "better" procedure for the situation in trials 3 and 4. However, in trials 3 and 4 the separation among the populations is sufficiently

TABLE 1

| | Trial 1 | Trial 2 | Trial 3 | Trial 4 |
|---|---|---|---|---|
| Classified in $_1$ | 95 | 99 | 100 | 100 |
| Classified in $_2$ | 94 | 100 | 100 | 100 |
| Classified in $_3$ | 89 | 100 | 100 | 100 |
| Misclassified | 22 | 0 | 0 | 0 |
| Number of times classified using Bayes | 159 | 78 | 16 | 0 |
| Time using Thresholding* (in .01 sec.) | 707 | 545 | 463 | 437 |
| Time using Bayes*(in .01 sec.) | 650 | 650 | 650 | 650 |

* includes input-output time of 2 seconds

large to yield a small probability of misclassification, in which case one could use some other procedure such as the "table look-up" technique, [1], [5]. The table look-up technique has been shown to require smaller computer time than the classical Bayes procedure [5].

In trial 1 the "closeness" of the populations forced the time for thresholding to be higher than for the Bayes technique. In this case 159 out of the 300 observations were classified according to the Bayes technique. Additional time was required for thresholding on the remaining 141 observations. Also, at least one of the density functions $p_1(x)$, $p_2(x)$, and $p_3(x)$ had to be evaluated for each of these 141 observations.

The following example illustrates the relationship between separation among the populations and the feasibility of the thresholding technique.

Example 4.1. Let $m = 3$, $p = 1$, and suppose $q_1$, $q_2$, and $q_3$ are known (see Figure 2). By the thresholding technique $x$ is classified in $\pi_i$ if $x \varepsilon R_i'$, $i = 1,2,3$. Thus the probability that $x$ is classified using thresholding is the probability that $x \varepsilon \bigcup_{i=1}^{3} R_i'$. Now extending the above argument to the general case we have

$$P_r \text{ (x classified using thresholding)} = \sum_{i=1}^{m} P_r(x \varepsilon R_i')$$

where $R_i' = \{x\colon (x - \hat{\mu}_i)^T \hat{\sum}_i^{-1}(x - \hat{\mu}_i) \leq Q_i\}$,

$Q_i = (y_i - \hat{\mu}_i)^T \hat{\sum}_i^{-1}(y_i - \hat{\mu}_i)$, and $y_i$ is such that $q_i p_i(y_i) = T_i$, for $i = 1,2,\ldots,m$. Since $(x - \mu)^T \sum^{-1}(x - \mu)$ has a chi-square distribution with p degrees of freedom the probabilities $P_r(\acute{x} \varepsilon R_i')$, i = 1,2,...,m, are easily obtained. These probabilities can be used to determine the number of times one can expect to classify according to the thresholding technique.

The results in Table 1 indicate that141, 222, 284, and 300 observations were classified by thresholding for trials 1, 2, 3, and 4, respectively. The corresponding expected number of times thresholding would be used are 130 , 228, 287, and 300 for trials 1, 2, 3, and 4, respectively.

## 5. Concluding Remarks

From the similation study of this paper it appears that there are situations in which the thresholding technique might be optimal in terms of time required for classification. One such situation is when the number of populations m and the number of observations to be classified are large. For example, in a remote sensing application the number of observations to be classified may be extremely large.

The number of populations in our study was m = 3. It seems reasonable to expect the thresholding technique to be useful, for example, when m = 10 and the average number of density functions evaluated for each classification is 5 (say). This depends, of course, on the degree of separation among the 10 populations. Final emphasis should be placed on the fact that if there is a high degree of closeness among the m populations then one should use the classical Bayes or some other technique. In summary, thresholding could prove useful when there are many populations with moderate separation among them.

## 6. References

[1] Eppler, W. G., Helmke, C. A. and Evans, R. H., "Table Look-up Approach to Pattern Recognition", Proceedings of 7th International Symposium on Remote Sensing of the Environment, The University of Michigan, May 1971.

[2] Huang, T., "Per Field Classifier for Agricultural Applications", LARS Information Note 060569, Purdue University, Lafayette, Indiana, June 1969.

[3] Minter, T. C. and Hallum, C. R., "Macimum Likelihood Classification by Thresholding", Lockheed Electronics Co., Inc., Houston, Texas, for NASA Manned Spacecraft Center, Houston, Texas, LEC/HASD No. 640-TR-114, June 1972.

[4] Newman, T. G. and Odell P. L., The Generation of Random Variates, Griffin's Statistical Monographs and Courses, No. 29, (1971), pp. 37-44.

[5] Odell, P. L., Duran, B. S. and Coberly, W. A., "On the Table Look-up in Discriminate Analysis", to appear in Journal of Statistical Computation and Similation, Vol. 2.

[6] Remote Sensing of Earch Resourses, NASA SP 7036, A Literature Survey with Indexes, September 1970.

# A SPACE APPLICATION OF AN EXTENSION OF THE BUFFON NEEDLE PROBLEM[1]

by

H. L. Gray and B. S. Duran

Texas Tech University

## 1. Introduction

A problem of some interest in the current space program is that of collecting data from the ground through airborne remote sensors and using these data to classify the type of ground cover or vegetation below. The data collected are often in the form of a measure of the light energy radiated in various bands of the light spectrum, from a small square on the ground. When the interest is in classifying the data as coming from a finite number of populations, such as in field classification of an agricultural area, the problem is a multiple decision problem. Since this is a standard problem and an optimal solution is known only in the case of conditions which, here, are unrealistic, a number of solutions have been posed [1]. Regardless of which solution is utilized, its success is of course a function of the amount of noise in the data. This noise, although a function of many variables is strongly influenced by the altitude of the satelite or other collecting device.

[1] This research was supported in part by NASA Manned Spacecraft Center under Contract NAS9-12775.

One of the more direct influences of the altitude on the data is the fact that the diagonal of the data square is a monotonic increasing function of the altitude and consequently so is that component of the noise which is due to the size of the square. In many instances then the question of primary interest is whether or not the data square is too large, i.e., whether or not certain types of remote sensing are feasible from altitudes as great as those of a satelite. Thus if the data square is a one mile square and the average field size of interest is a 1/2 mile square then no data will be obtained which is strictly from the population of interest. If those dimensions are reversed, it is still quite likely that very little data of interest will be collected. Briefly, the question is, for a square of a given size, how much data is likely to be obtained from the population of interest and how much will be a conglomerate of several populations?

Although it is not a complete answer to the problem, a measure which conveys essentially the information needed to decide the feasibility question is the probability, $P_o$, that a square dropped at random on a "map" will land in the interior of a "region". Put in terms more suggestive of the title of this paper, and equally informative, the measure of interest is the probability, P, that a square randomly placed on a map will cross a boundary line of one of the pieces which form the map. The purpose of this paper is to determine this latter probability for maps which are "quilts" (to be defined later) and to give some empirical evidence to show that the results can also be used for maps which are not "quilts".

## 2. Main Result

Since a map can be formed by any collection of geometrical shapes it is impossible to find the probability, P, (defined above) for all possible maps. However it is usually possible (at least for agricultural fields) to approximate a given map by a finite collection of rectangles having at least one side in common. Moreover, as we shall see, an analytical expression for P when our map is such a collection of rectangles which we will henceforth refer to as a quilt, can be obtained. Given any map our approach will therefore be to approximate it by a quilt, calculate the probability of crossing the boundary lines of the quilt, and approximate P by this probability. We now make the following definitions.

[CQ] = Event that a square dropped at random on a quilt, Q, will cross a boundary line.

$[R_i]$ = Event that the center of a square dropped at random will land in rectangle $R_i$, where $\bigcup_{i=1}^{n} R_i$ defines the quilt, Q.

$A_i$ = Area of $R_i$.

A = Area of the quilt.

$2w_i$ = width of $R_i$.

$2h_i$ = length of $R_i$.

With this notation we can now write

$$(1) \qquad P[CQ] = \sum_{i=1}^{n} P[CQ|R_i]\ P[R_i] = \frac{1}{A} \sum_{i=1}^{n} P[CQ|R_i]A_i .$$

A typical fall of the square center in a rectangle of dimensions 2w by 2h is displayed in Figure 1.

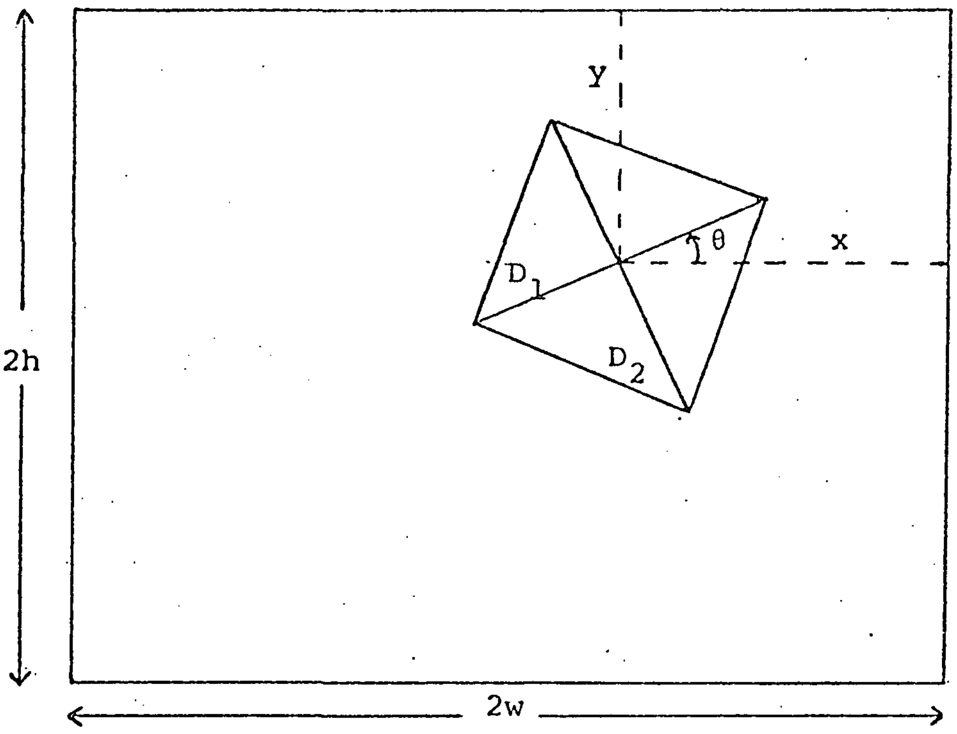


Figure 1.

In Figure 1. x and y denote the minimum distances from the center of the square to the vertical and horizontal edges, respectively, and θ denotes the smallest angle that a diagonal makes with the horizontal. We assume that the length, 2d, of the diagonal of the square is such that $d < w$ and $d < h$. This is a convenience and could be removed but realistically it does not seem necessary to do so.

In determining the probability of the square intersecting the quilt we first note that the square crosses the quilt if and only if at least one of the diagonals crosses it. Thus it is sufficient to determine the probability that at least one diagonal crosses the quilt.

We will let $D_1$ and $D_2$ denote the diagonals that form the smallest and largest positive angles, respectively, with the horizontal. Further we assume the quantities X, Y, and $\theta$ defined by figure 1 are independent uniform random variables over $(0,w)$, $(0,h)$ and $(0,\pi/2)$, respectively.

In determining $P[CQ|R_i]$ we first consider the events

$[D_jC|R_i]$ = event that diagonal $D_j$ crosses the quilt, given the center of the square is in $R_i$.

$[D_jC_iV]$ = event that diagonal $D_j$ crosses a vertical line given the center of the square is in $R_i$.

$[D_jC_iH]$ = event that diagonal $D_j$ crosses a horizontal line, given the center of the square is in $R_i$.

$[D_jC_iV_o]$ = event that diagonal $D_j$ crosses a vertical line only, given the center of the square is in $R_i$.

$[D_jC_iH_o]$ = event that diagonal $D_j$ crosses a horizontal line only, given the center of the square is in $R_i$.

$[D_jC_iVH]$ = event that diagonal $D_j$ crosses a horizontal and vertical line given the center of the square is in $R_i$.

In view of these definitions we then have

$$P[CQ|R_i] = P[D_1C|R_i] + P[D_2C|R_i] - P[(D_1C|R_i) \cap (D_2C|R_i)]$$

$$= 2P[D_1C|R_i] - P[(D_1C|R_i) \cap (D_2C|R_i)]$$

$$= 2\{P[D_1C_iV] + P[D_1C_iH] - P[(D_1C_iV) \cap (D_1C_iH)]\}$$

$$- P[(D_1C|R_i) \cap (D_2C|R_i)].$$

However,

$$P[D_1C_iV] = P(X < d\cos\theta) = \frac{2}{w_i\pi}\int_0^{\pi/2}\int_0^{d\cos\theta} dxd\theta = \frac{2d}{\pi w_i},$$

$$P[D_1C_iH] = P(Y < d\sin\theta) = \frac{2d}{\pi h_i},$$

and

$$P[(D_1C_iV) \cap (D_1C_iH)] = \frac{d^2}{\pi h_i w_i}.$$

Therefore,

$$P[D_1C|R_i] = \frac{2d}{\pi}\left(\frac{1}{w_i} + \frac{1}{h_i} - \frac{d}{2h_iw_i}\right). \tag{2}$$

To determine $P[D_1C|R_i) \cap (D_2C|R_i)]$ we observe that $D_1$ can cross a vertical line only, a horizontal line only, or both types of lines. Similarly for the diagonal $D_2$. Consequently, the event $[D_1C|R_i] \cap [D_2C|R_i]$ can be decomposed into a complete system of nine events, *i.e.*,

$$[D_1C|R_i] \cap [D_2C|R_i] = [(D_1C_iV_o) \cap (D_2C_iV_o)] \cup [(D_1C_iV_o) \cap (D_2C_iH_o]$$

$$\cup [(D_1C_iV_o) \cap (D_2C_iVH] \cup [(D_1C_iH_o) \cap (D_2C_iV_o)]$$

$$\cup [(D_1C_iH_o) \cap (D_2C_iH_o] \cup [(D_1C_iH_o) \cap (D_2C_iVH)]$$

$$\cup [(D_1C_iVH) \cap (D_2C_iV_o)] \cup [(D_1C_iVH) \cap (D_2C_iH_o)]$$

$$\cup [(D_1C_iVH) \cap D_2C_iVH] .$$

Using this decomposition we obtain

$$P[(D_1C_iV_o) \cap (D_2C_iV_o)] = P[(X \leq d \cos \theta, Y \leq d \sin \theta)$$

$$(X \leq d \cos \theta, Y \leq d \cos \theta)]$$

$$= \frac{2d}{\pi h_i w_i} [(2 - \sqrt{2})h_i - \frac{d}{2}] \quad .$$

Similarly,

$$P[(D_1C_iH_o) \cap (D_2C_iH_o)] = \frac{2d}{\pi h_i w_i} [(2 - \sqrt{2})w_i - \frac{d}{2}] ,$$

$$P[(D_1C_iVH) \cap (D_2C_iVH)] = \frac{d^2}{\pi h_i w_i} [\frac{\pi}{2} - 1],$$

$$P[(D_1C_iV_o) \cap (D_2C_iH_o)] = P[(D_1C_iH_o) \cap (D_1C_iV_o)]$$

$$= \frac{d^2}{\pi h_i w_i} [\frac{\pi}{2} - 1],$$

$$P[(D_1C_iVH) \cap (D_2C_iV_o)] = P[(D_1C_iV_o) \cap (D_2C_iVH)]$$

$$= \frac{d^2}{\pi h_i w_i} [1 - \frac{\pi}{4}],$$

$$P[(D_1C_iVH) \cap (D_2C_iH_o)] = P[(D_1C_iH_o) \cap (D_2C_iVH)$$

$$= \frac{d^2}{\pi h_i w_i} [1 - \frac{\pi}{4}].$$

Collecting all these results we obtain

$$\begin{aligned} P[CQ|R_i] &= \frac{2d}{\pi} [\frac{\sqrt{2}}{w_i} + \frac{\sqrt{2}}{h_i} - \frac{d}{2h_iw_i} - \frac{\pi d}{4h_iw_i}] \\ &= \frac{2d}{\pi h_i w_i} [\sqrt{2}(w_i + h_i) - \frac{d}{2}(1 + \frac{\pi}{2})]. \end{aligned} \tag{3}$$

By (1) and (3) the probability that the square crosses a quilt consisting of n patches is given by

$$\begin{aligned} P[CQ] &= \sum_{i=1}^{n} P[CQ|R_i] \; (A_i/A) \\ &= \frac{2d}{\pi A} \sum_{i=1}^{n} \{\frac{\sqrt{2}(w_i+h_i) - \frac{d}{2}(1 + \frac{\pi}{2})}{h_iw_i}\} \; 4w_ih_i \end{aligned} \tag{4}$$

$$(4) \qquad = \frac{8d}{\pi A} \sum_{i=1}^{n} \{\sqrt{2}(w_i + h_i) - \frac{d}{2}(1 + \frac{\pi}{2})\}$$

since $A_i = 4w_ih_i$. If we let $P_i$ denote the perimenter of the $i^{th}$ patch then the probability (4) may be written as

$$(5) \qquad P[CQ] = \frac{2d}{\pi A} \{\sqrt{2} \sum_{i=1}^{n} P_i - 2nd\ (1 + \pi/2)\}.$$

## 3. Examples and Applications

It is of interest, and in fact quite simple, to verify the probability given by (4) empirically. For a particular rectangular region consisting of n rectangles, all one needs to do is choose a rectangle (at random) from the n rectangles and then choose x, y, and $\theta$ in $(0,w)$, $(0,h)$, and $(0,\pi/2)$ by means of a random number generator. With this information one can then determine whether the square, of fixed diagonal length, crosses the rectangle boundaries. If this procedure is repeated m times then an estimate of P[CQ] is given by

$$\hat{P} = m_1/m,$$

where $m_1$ denotes the number of times that the square crosses a boundary.

For $n = 12$ and $d = 1$ (see Figure 2) $\hat{P} = 0.7030$ was obtained using $m = 1000$. The actual value of P[CQ], by (4), is $P[CQ] = 0.7039$.

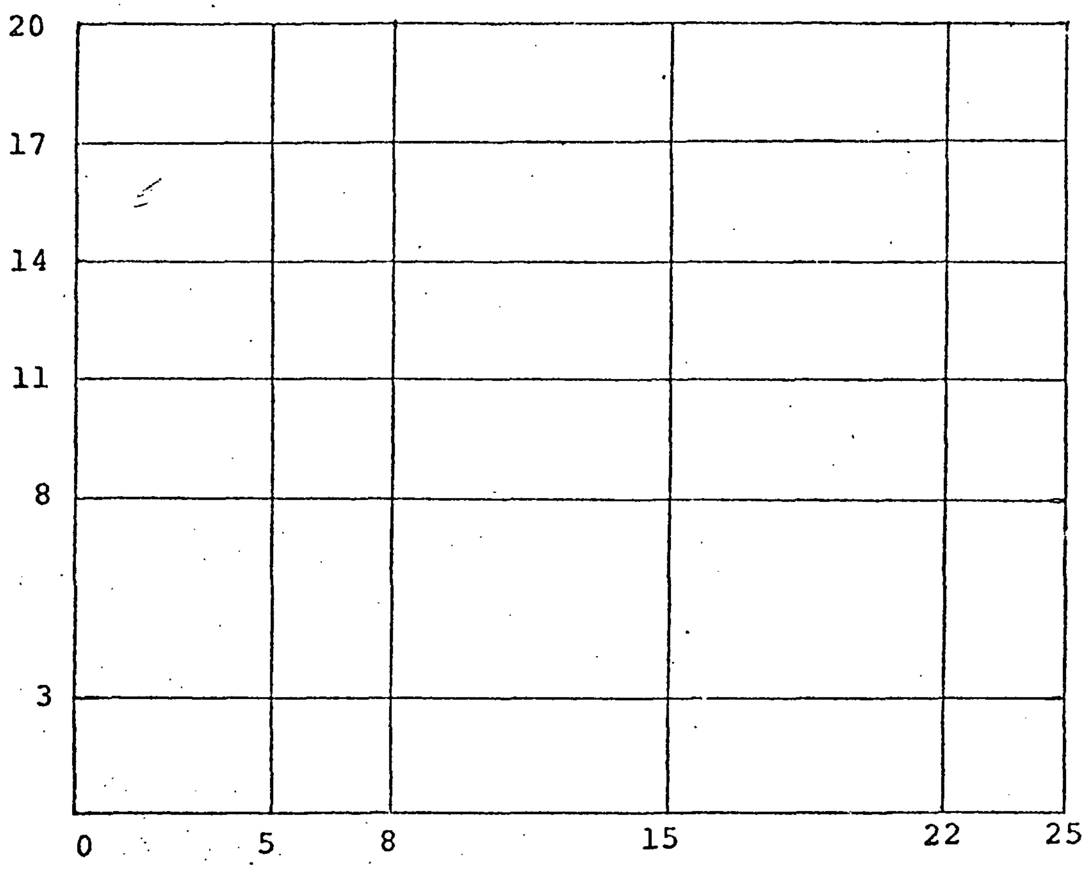


Figure 2.

As we previously stated, the probability that a square dropped at random on a map intersects at least one boundary of the map can be estimated by the probability in (4). To demonstrate this an aerial photo of an agricultural area was considered. The map was first "covered" with an approximating quilt and then equation (4) was used to find the probability that the square crossed the quilt. This probability was then used to approximate P. The accuracy of such an estimate of course depends on how well the quilt fits the map. However, it is possible to check the accuracy in this instance by generating points on the approximating quilt as described above and then by means of an overlay, note those instances where the quilt was crossed but not the map and visa versa.

Figure 3 contains a particular map including an approximating quilt. The dimensions of the quilt are 20 cm. by 19.6 cm. and the map scale is 1 mi. = 10.4 cm. The diagonal of the square was 2d = 0.85 cm. For m = 100 we obtained, empirically, 37 drops which crossed quilt and map boundaries, 3 which crossed map boundaries only, and 2 which crossed quilt boundaries only. We thus have the empirical results $\hat{P}[CQ] = .39$ and $\hat{P}[CM] = .40$. The actual value of P[CQ] by (4) is P[CQ] = .324. The large difference between $\hat{P}[CQ]$ and P[CQ] is probably due to the small number of simulations (m = 100). However, the important observation is the closeness of $\hat{P}[CQ]$ and $\hat{P}[CM]$, and hence, these results indicate that P[CQ] may be used to approximate P[CM].

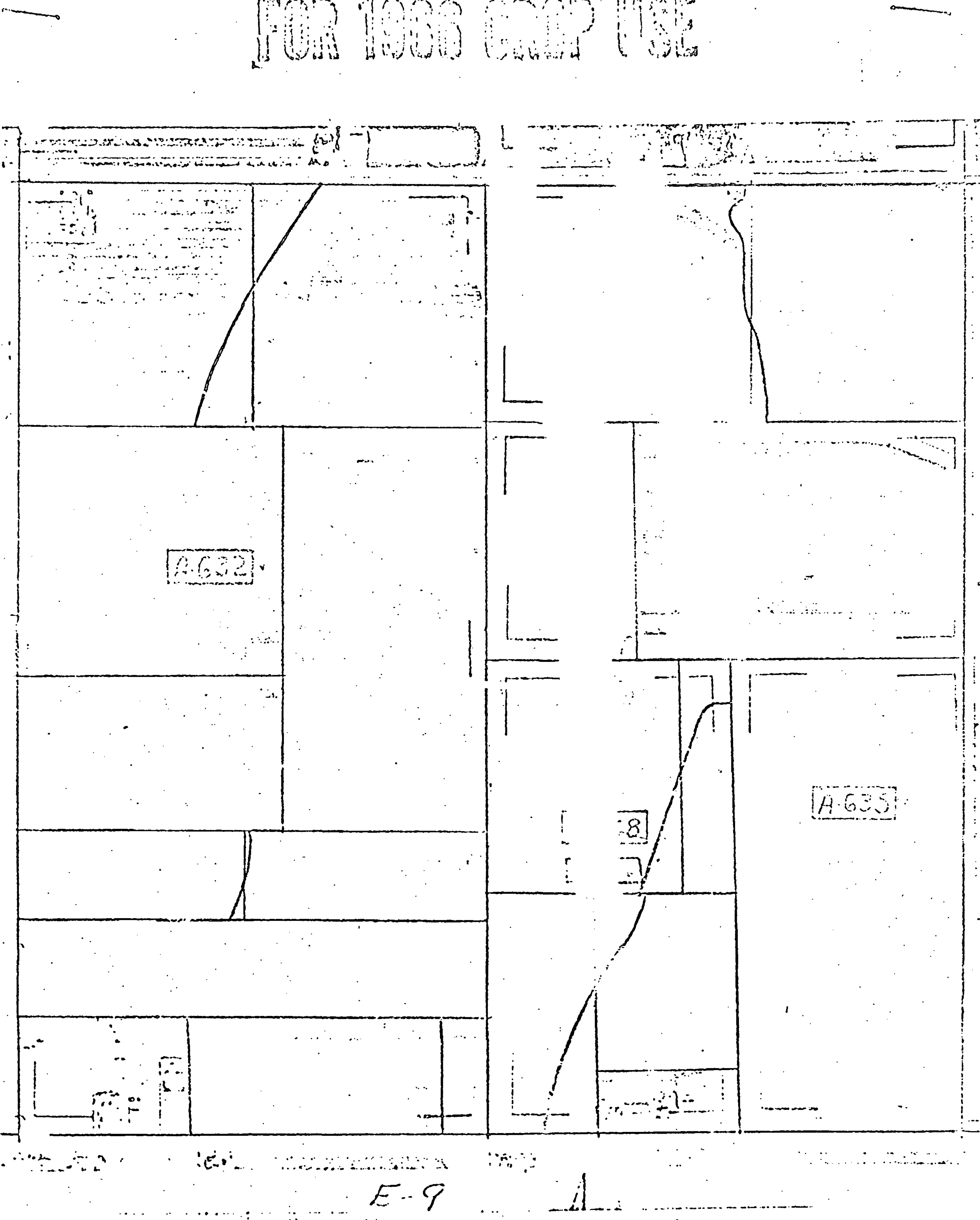


Figure 3. Agricultural map including quilt overlay.

## 4. References

[1] Huang, Teddy, "Per Field Classifier for Agricultural Applications," LARS Information Note 060569, Purdue University, Lafayette, Indiana, June, 1969.

# CONCERNING DIMENSION REDUCTION IN DISCRIMINATE ANALYSIS[1]

by

J. P. Basu[2]

and

P. L. Odell[3]

[1] This research was supported (in part) by NASA Manned Spacecraft Center under Contract NAS 9-12775

[2] Texas Tech University, Lubbock, Texas 79409

[3] University of Texas, Dallas, Texas 75080

# CONCERNING DIMENSION REDUCTION IN DISCRIMINATE ANALYSIS

## 1. Introduction

Let $\pi_1, \pi_2, \ldots, \pi_m$ be m distinct populations with their individuals having p common observable characteristics. The samples obtained as observations on these characteristics are then vectors from p-dimensional real Euclidean space, the probability density functions $p_i(x)$ of the populations are p-dimensional density functions. When p is large compared to m, or large irrespective of magnitude of m, then we face an undesirable situation (from computational point of view) where statistical analysis involves (perhaps unnecessarily) large dimension of data vector. One way to avoid this situation is to compress the data before starting the actual statistical analysis by a "suitable linear transformation," to be referred to as compression matrix later, of the form

$$Y = CX$$

where X is a p × 1 sample vector and C is a real r × p matrix satisfying certain conditions. This in statistical literature is referred to as "dimension reduction" technique and in the engineering literature as feature selection. There are at least two different approaches to the problem of selection of C, of which one due to Wilks is based on his concept of scatter and the other is based on so called Karhunen-Loève expansion. Both methods are modifications of principal component analysis. In this paper we compare the two methods and show how they are related to total misclassification probability consideration when the populations are all normal with equal covariance matrices.

## 2. Wilks' Technique for Dimension Reduction [5]

Let $\{X_i^{(\alpha)}\}(i=1,2,\ldots,N_\alpha;\ \alpha=1,2,\ldots,m)$ be m sets of samples from m populations $\pi_1,\ldots,\pi_m$. Then the within scatter matrix $S_W$, between scatter matrix $S_B$ and the total scatter matrix $S_T$ are defined to be

$$S_W = \sum_{\alpha=1}^{m} S_\alpha = \sum_{\alpha=1}^{m} \sum_{i=1}^{N} (X_i^{(\alpha)} - \overline{X}^{(\alpha)})\ (X_i^{(\alpha)} - \overline{X}^{(\alpha)})^T , \tag{2.1}$$

$$S_B = \sum_{\alpha=1}^{m} N_\alpha\ (\overline{X}^{(\alpha)} - \overline{X})\ (\overline{X}^{(\alpha)} - \overline{X})^T, \tag{2.2}$$

$$S_T = \sum_{\alpha=1}^{m} \sum_{i=1}^{N_\alpha} (X_i^{(\alpha)} - \overline{X})\ (X_i^{(\alpha)} - \overline{X})^T \tag{2.3}$$

where

$$S_\alpha = \sum_{i=1}^{N_\alpha} (X_i^{(\alpha)} - \overline{X}^{(\alpha)})\ (X_i^{(\alpha)} - \overline{X}^{(\alpha)})^T,\quad \overline{X}^{(\alpha)} = \sum_{i=1}^{N_\alpha} X_i^{(\alpha)}/N_\alpha, \tag{2.4}$$

and

$$\overline{X} = \sum_{\alpha=1}^{m} \sum_{i=1}^{N} X_i^{(\alpha)}/(N_1+N_2+\ldots+N_m) . \tag{2.5}$$

It is easy to show that $S_T=S_B+S_W$. Let us denote these matrices by $S_X(x)$, $S_B(x)$ and $S_T(x)$ respectively, so that the corresponding matrices of the transformed samples $\underset{r\times 1}{Y} = \underset{r\times p}{C}\ \underset{p\times 1}{X}$ of dimension r are denoted by $S_W(y)$, $S_B(y)$ and $S_T(t)$. It is evident that

$$S_W(y) = C\ S_W(x)C^T,\ S_B(y) = C\ S_B(x)C^T,\ S_T(y) = C\ S_T(x)C^T . \tag{2.6}$$

The Wilks' technique selects a $r \times p$ real matrix C such that the r columns of $C^T$ are orthogonal $p \times 1$ vectors and the ratio

$$|S_W(y)|/|S_T(y)|$$

is minimized for a fixed value K of $|S_W(y)|$.

Using Lagrange multiplier method it is not difficult to show that such C will satisfy the following relations.

$$|S_B(x) - \lambda S_W(x) )\ C^T| = 0 \tag{2.7}$$

and

$$C\ S_w(x)\ C^T = K\ . \tag{2.8}$$

Equation (2.7) has a nonzero solution if and only if

$$|S_B(x) - \lambda S_W(x)| = 0 \tag{2.9}$$

It is well known (Rao [5] p. 37) that since $S_B(x)$ and $S_W(x)$ are two p × p symmetric matrices of which $S_W(x)$ is positive definite (with probability one), then there exists a nonsingular matrix P such that

$$P^T S_W(x) P = I_p \text{ and } P^T S_B(x) P = L \tag{2.10}$$

where $I_p$ is the p × p unit matrix and L is the diagonal matrix,

$$\Gamma = \text{diag}\ \{\lambda_1, \ldots, \lambda_p\}\ .$$

Without loss of generality it can be assumed that $\lambda_1 > \lambda 2 > \ldots > \lambda_p$ . $P^T P \neq 0$, the values of $\lambda$ for which $|S_B(x) - \lambda S_W(x)|$ vanishes are identical with those for which $|P^T S_B(x) P - \lambda P^T S_W(x) P| = |L - \lambda I|$ vanishes. The nonzero roots of equation (2.9) are thus the nonzero elcments among $\lambda_1, \ldots, \lambda_p$. The number of such nonzero elements $\lambda_1, \ldots, \lambda_p$ is exactly equal to the rank of $S_B(x)$, which in its turn is equal to the dimension of the affine subspace spanned by the points $\overline{X}^{(1)}, \ldots, \overline{X}^{(m)}$, which is (m-1) with probability one.

If r = (m-1), that is if X is to be projected to a (m-1) dimensional subspace, then the (m-1) rows $C_1, \ldots, C_{(m-1)}$ of C will be chosen as orthogonal vectors such that

$$(S_B(X) - \lambda_j S_W(X))\ C_j^T = 0 \quad (j=1, \ldots, (m-1)\ , \tag{2.11}$$

that is, $C_j^T$ is the eigen vector associated with the eigen value $\lambda_j$. The vectors $C_1^T,\ldots,C_{m-1}^T$ are then scaled so that $|C\ S_W(x)C^T| = K$. The value of $|S_W(y)|/|S_T(y)|$ is then a minimum for the above choice of C and

$$\min \frac{|S_W(y)|}{|S_T(y)|} = \frac{|C\ S_W(x)C^T|}{|C\ S_T(x)C^T|} = \frac{1}{(1+\lambda_1)\ldots(1+\lambda_{m-1})} = \frac{|S_W(x)|}{|S_T(x)|}\ . \tag{2.12}$$

Thus if $r = (m-1)$, the above choice of C not only minimizes $|S_W(y)|/|S_T(y)|$, subject to the condition $|S_W(y)|= K$, but it conserves the ratio $|S_W|/|S_T|$ also.

When $r < (m-1)$, we select $C_1^T,\ldots,C_r^T$ as the eigen vectors corresponding to the r largest eigen values $\lambda_1,\ldots,\lambda_r$ such that

$$(S_B(x) - \lambda_j S_W(x)C_j^T = 0\ \ (j=1,\ldots,r) \tag{2.13}$$

and

$$|C\ S_W(x)C^T| = K. \tag{2.14}$$

In this case

$$\min\ |S_W(y)|/|S_T(y)| = 1/[(1+\lambda_1)\ \ldots\ (1+\lambda_r)] \neq |S_W(x)|/|S_T(x)|.$$

It should be noted that the λ's are random variables. When $r = (m-1)$, the transformation $Y = C\ X$ for the above choice of C is actually a projection onto the subspace spanned by the vectors $\overline{X}^{(2)} - \overline{X}^{(1)},\ldots,\overline{X}^{(m)} - \overline{X}^{(1)}$.

## 3. Karhunen-Loève Expansion Technique [2,3]

Let $\{X_i^{(\alpha)}\}(i=1,\ldots,N_\alpha;\ \alpha=1,\ldots,m)$ be m sets of samples (p × 1 vectors) from m populations. The covariance matrix S of the grand sample is then

$$\hat{S} = S_T/(N_1+\ldots+N_m)$$

where $S_T$ is the total scatter matrix. The generalized Karhunen-Loève expansion theorem (due to Chien and Fu[2]) states that every sample vector $X_i^{(\alpha)}$ can be represented as

$$X_i^{(\alpha)} = \sum_{j=1}^{p} a_{ij}^{(\alpha)} \phi_j ,$$

where $$a_{ij}^{(\alpha)} = \phi_j^T X_i^{(\alpha)} ,$$

and $\phi_1,\ldots,\phi_p$ are eigen vectors of $\hat{S}$. Let the corresponding eigen values be $\lambda_1,\ldots,\lambda_p$, which without loss of generality can be assumed to be ordered, viz, $\lambda_1 \; \lambda_2 \ldots \lambda_p$. The compression matrix $\underset{r\times p}{C}$ for compressing the observation to r dimensions is then given by

$$\underset{p\times r}{C^T} = [\underset{p\times 1}{\phi_1} \ldots \underset{p\times 1}{\phi_r}] .$$

The motivation of such choice of C lies in the fact that the mean squared error between $X_i^{(\alpha)}$ and $C\ X_i^{(\alpha)}$

$$||X_i^{(\alpha)} - C\ X_i^{(\alpha)}|| = \lambda_{r+1}+\ldots+\lambda_p$$

is minimum. But this choice of C has no association (in general) with misclassification probability or separability of classes. Nothing can be said about the information regarding separability of classes, since it is not apparent what this transformation does to the between scatter matrix $S_B$.

## 4. Probability of Misclassification and Dimension Reduction

Let the probability densities $p_i(x)$ be known to be normal $N_p(\mu_i,\Sigma)$ $(i=1,\ldots,m)$. Let $x = (x_1,\ldots,x_p)^T$ denote a set of observations on the p(common) characteristics of an individual $I(x)$, $q_i$ the a priori probability of selecting an individual from $\pi_i$, $C(j|i)$ the cost of misclassifying an individual from $\pi_i$ as being from $\pi_j$. Let $R = (R_1,\ldots,R_m)$ be a partition of the p-dimension at Euclidean space $E_p$ into m regions defining a classification procedure such that $I(x) \varepsilon \pi_i$ whenever $x \varepsilon R_i$. Then the expected misclassification cost for this procedure is given by

$$Q_p(x,R) = \sum_{i=1}^{m} \sum_{\substack{j=1 \\ j\neq i}}^{m} \int_{R_j} [q_i p_i(x) \; C(j|i)] \, dx \quad . \tag{4.1}$$

If $C(j|i) = 1$ for all $j \neq i$ and $q_i = 1/m$ for all $i$, then

$$Q_p(x,R) = \sum_{i=1}^{m} \sum_{\substack{j=1 \\ j\neq i}}^{m} 1/m \int_{R_j} p_i(x) \, dx$$

$$= \sum_{i=1}^{m} 1/m(1-\int_{R_i} p_i(x)dx)=1-\sum_{i=1}^{m} \int_{R_i} p_i(x)dx \; . \tag{4.2}$$

It is well known (Anderson [1] p 148) that the regions of classification $R^* = (R_1^*,\ldots,R_m^*)$ that minimize the expected cost $Q_p(x,R)$ (or maximize the sum of probabilities of proper classification $\sum_i \int_{R_i} p_i(x)dx$, in this case) are defined by

$$R_j^* = \{x:U_{jk}(x)>0 \text{ for all } k\neq j\}(j=1,\ldots,m),$$

where
$$U_{jk}(x) = [x - 1/2(\mu_j+\mu_k)]^T \sum^{-1} (\mu_j-\mu_k). \tag{4.3}$$

and
$$Q_p(x,R^*) = 1 - \sum_{j=1}^{m} \int_{R_j^*} p_j(x) \, dx \; . \tag{4.4}$$

As in dimension reduction technique, let us compress the $p \times 1$ observation vectors by a linear transformation $\underset{r\times 1}{Y} = \underset{r\times p}{C} \; \underset{p\times 1}{X}$ before starting classification. The probability density functions $p_i(y)$ of the transformed populations $\pi_i(i=1,\ldots,m)$ are then given by

$$\hat{p}_i(y) = N_r(C\mu_i, C\Sigma C^T).$$

The classification regions in this r dimensional space, $\hat{R}^*=(\hat{R}_1^*,\ldots,\hat{R}_m)$, giving minimum expected misclassification cost (or maximum sum of probabilities of proper classification, in this case) are now given by

$$\hat{R}_j^* = \{\hat{u}_j k(y) > 0 \text{ for all } k \neq j\} , \tag{4.5}$$

where $$\hat{U}_j k(y) = [y - 1/2\ C(\mu_j+\mu_k)]^T (C\Sigma C^T)^{-1}\ C(\mu_j-\mu_k)$$

$$= [x - 1/2\ (\mu_j+\mu_k)]^T C^T (C\Sigma C^T)^{-1} C(\mu_j-\mu_k), \tag{4.6}$$

and $$Q_r(y,\hat{R}^*) = 1 - \sum_{j=1}^{m} \int_{\hat{R}_j} \hat{p}_j(y)dy . \tag{4.7}$$

Our problem is to find C such that $Q_p(x,R^*)$, that is, (in this case) probabilities of proper classification or misclassification are not changed. This is possible if

$$\int_{R_j^*} p_j(x)dx = \int_{\hat{R}_j^*} p_j^{\hat{}}\ (y)\ dy . \tag{4.8}$$

Let $\Gamma$ denote the projection of $E_p$ onto a r-dimensional subspace S given by $y = Cx$. By definition $\hat{R}_j^* = \Gamma(R_j^*)$ and $\hat{p}_j(y)$ is the density of the probability measure $P_r$ obtained as the projection onto this r-dimensional space of the probability measure $P_p$ with density $p_j(x)$. Then it follows from (4.5) that the probabilities of proper classification or misclassification are preserved if

(1) $P_p$ can be expressed as the product of two measures, viz $P_p=P_r \times P_{p-r}$,

(2) $R_j^*$, for each j=1,...,m can be expressed as the cartesian product of two sets, viz $R_j^* = \hat{R}_j^*\ \ E_p\text{-}S = P(R_j^*) \times E_p\text{-}S$ ($E_p$-S denoting subspace orthogonal to S).

The first condition can always be met. But, as we will show, the second condition can be met *only if* $\Gamma$ is a projection onto the subspace M spanned by the difference of the mean vectors, viz. $\mu_2-\mu_1,\ldots,\mu_m-\mu_1$. Without loss of generality we can assume $\mu_1=0$. Then M becomes the subspace spanned by $\mu_2,\mu_3,\ldots,\mu_m$. Thus, we need to show that condition (2) is met if S = M.

Let the dimension of M, the space spanned by $\mu_2,\ldots,\mu_m$ (assuming $\mu_1 = 0$), be k; C be a k×p matrix such that $y = Cx\varepsilon M$ for $x\varepsilon E_p$ and $\hat{C}$ a p-k×p matrix such that $z = \hat{C}x\ E_p-M$. Then $C\hat{C}^T = \phi$. If $X \sim N_p(\mu;\Sigma)$ then $Y = CX \sim N_k (C\mu,C\varepsilon C^T)$, $Z = \hat{C}X \sim N_{p-k} (\phi, \hat{C}\Sigma\hat{C}^T)$ and Y and Z are independent. Let

$$D = \begin{bmatrix} C \\ \hat{C} \end{bmatrix}$$

so that

$$\begin{bmatrix} y \\ z \end{bmatrix} = \begin{bmatrix} C \\ \hat{C} \end{bmatrix} x = D\,x \qquad (4.9)$$

and

$$x = D^{-1} \begin{bmatrix} y \\ z \end{bmatrix} . \qquad (4.10)$$

Then, since $\hat{C}\mu_j = 0$ for all j, we have

$$U_{jk}(x) = [x - 1/2(\mu_j+\mu_k)]^T \textstyle\sum^{-1} (\mu_j-\mu_k)$$

$$= [y^T - 1/2\ C(\mu_j+\mu_k)^T\ z^T](D^{-1})^T \textstyle\sum^{-1}\ D^{-1}[C(\mu_j-\mu_k)\phi]$$

$$= [y^T - 1/2\ C(\mu_j+\mu_k)^T\ z^T](D\Sigma D^T)^{-1}\ [C(\mu_j-\mu_k)\ \phi]$$

$$= [y^T - 1/2\ C(\mu_j+\mu_k)^T]\ (C\Sigma C^T)^{-1}\ C(\mu_j-\mu_k) \qquad (4.11)$$

$$= [x - 1/2(\mu_j+\mu_k)]^T\ C^T(C\Sigma C^T)^{-1}C(\mu_j-\mu_k) \qquad (4.12)$$

Thus for each j (j = 1,...,m) we can write $R_j^*$ as

$$R_j^* = \bigcap_{k\neq j} \{x: u_{jk}(x) \quad 0\}$$

$$= \bigcap_{k\neq j} \{y: [y^T - 1/2\ C(\mu_j+\mu_k)^T](C\Sigma C^T)^{-1}C(\mu_j-\mu_k) > 0\} \times (E_p-m)$$

$$= R_j^* \times (E_p-M) \ . \tag{4.13}$$

We thus have the following theorem.

Theorem 1. Let C be a k × p matrix projecting $E_p$ orthogonally to the subspace M spanned by the vectors $\mu_2-\mu_1,\ldots,\mu_m-\mu_1$, k being the dimension of M. Then the maximum sum of probabilities of proper classification (hence, for equal a priori probabilities and misclassification cost, the expected misclassification cost) is not altered even when the classification is done on the basis of the compressed observation $y = Cx$.

Now let us further restrict C so that $C_i\Sigma C_j^T = 0$, for $i \neq j$ and $C_i, C_j$ representing the $i^{th}$ and $j^{th}$ row of C respectively. With such C, $C\Sigma C^T$ will be a diagonal matrix. Now let us define the between and within scatter matrices for these m populations respectively by

$$S_B = \sum_{i=1}^{m} (\mu_i-\overline{\mu})(\mu_i-\overline{\mu})^T \ , \tag{4.14}$$

where $$\overline{\mu} = \sum_{i=1}^{m} \mu_i/m$$

and $$S_W = m\Sigma \ . \tag{4.15}$$

Let us further restrict C such that $C_i S_B C_j T = 0$ for $i \neq j$. Then both $C\ S_B C^T$ and $C\ S_W C^T$ are diagonal matrices. Now if we choose $\lambda_1 > \lambda_2 \ldots > \lambda_k > 0$ such that

$$C_j S_B C_j^T = \lambda_j C_j S_W C_j^T ,$$

Then

$$C(S_B - \lambda S_W)C^T = 0 . \tag{4.16}$$

Conversely, if C is chosen as solutions of

$$(S_B - \lambda S_W)C^T = 0$$

C will project $E_p$ orthogonally to M. (4.17)

Thus we have the following theorem.

Theorem 2. Let C be a k × p matrix satisfying the equation

$$(S_B - \lambda S_W)C^T = 0,$$

where $S_B$ and $S_W$ are defined as in (4.14) and (4.15). Then the maximum sum of probabilities of proper classification (or expected misclassification cost, in this case) is not altered if classification is made on the basis of compressed observation $y = Cx$.

Now if $\underset{r\times p}{C}$ is not a projection onto the subspace spanned by $\mu_2 - \mu_1$, $\mu_m - \mu_1$, then $\hat{C}(\mu_j - \mu_k) \neq \phi$ and hence

$$U_{jk}(x) = [y^T - 1/2C(\mu_j + \mu_k)^T]\ (C\Sigma C^T)^{-1}\ C(\mu_j - \mu_k)$$

$$+ [Z^T - 1/2\hat{C}(\mu_j + \mu_k)^T](C\Sigma C^T)^{-1}\ \hat{C}(\mu_j - \mu_k) .$$

Thus we see that in this case, $R_j^*$ cannot be expressed in the form $\hat{R}_j^* \times E_p - M$, where $\hat{R}_j^*$ is a subset of the r dimensional subspace, the range of C.

## 5. Comparison of Wilk's and Karhunen-Loève Technique: An Example.

Let us assume that we have same number (N) of samples from each of the m populations. Then we may note that $S_B/N$ in 2.3 is an estimate of $\Sigma = \sum_{i=1}^{m} (\mu_i-\bar{\mu})(\mu_i-\bar{\mu})^T$ in (4.14) and $S_B/N$ in (2.1) is an estimate of $m\Sigma$. Thus equation (4.17) is obtained from (2.7) if we replace $S_B$ and $S_W$ by $\sum_{i=1}^{m} (\mu_i-\bar{\mu})(\mu_i-\bar{\mu})^T$ and $m\Sigma$ respectively. This leads us to expect intuitively that misclassification probabilities are preserved under Wilk's technique when C is m-1 × p. We now give an example for comparison of performance of Karhunen-Loève expansion technique and Wilks' technique.

Example. Let $p_i(x) = N_3(\mu_i,\Sigma)$ $(i=1,2)$. As there are 2 populations, the means are collinear. Let $\mu_1 = [0,0,0]^T$, $\mu_2 = [0,0,2]$ and

$$\Sigma = \begin{bmatrix} 6 & 0 & 0 \\ 0 & 5 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

Then $\sum_{i=1}^{2} (\mu_i-\bar{\mu})(\mu_i-\bar{\mu})^T = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 2 \end{bmatrix}$

and hence $\hat{S}$ of Karhunen-Loève expansion is given by

$$\hat{S} = \sum_{i=1}^{2} (\mu_i-\bar{\mu})(\mu_i-\bar{\mu})^T + 2\Sigma$$

$$= \begin{bmatrix} 12 & 0 & 0 \\ 0 & 10 & 0 \\ 0 & 0 & 5 \end{bmatrix}.$$

The maximum eigen value is 12 and corresponding eigen vector $\phi$ is given by $\phi_1^T = (1,0,0)$. Thus $C = (1,0,0)$. $Y = C X = X_1$. But there is no discriminant information in Y, since $Y \sim N`(0,12)$ given $I(x) \varepsilon \pi_1$ and $Y \sim N\ (0,12)$ given $I(x) \varepsilon\pi_2$ .

In case of Wilks' technique,

$$S_B - \lambda m\sum = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 2 \end{bmatrix} -2\lambda \begin{bmatrix} 6 & 0 & 0 \\ 0 & 5 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

$$= \begin{bmatrix} -12\lambda & 0 & 0 \\ 0 & -10\lambda & 0 \\ 0 & 0 & 2-2\lambda \end{bmatrix}$$

The nonzero root of $|S_B - \lambda m\Sigma| = 0$ is $\lambda=1$. The eigen vector corresponding to the eigen value $\lambda=1$ is $\phi_1 = [0,0,1]^T$ and thus $C = (0,0,1)$. $Y = CX$ and $Y \sim N(0,1)$ if $I(x)$ , and $Y \sim N\ (2,1)$ if $I(x)\varepsilon\pi_2$. This shows that there do exist discriminant information in Y.

## 6. Concluding Remarks

Among the engineers [2,3], a popular technique of "dimension reduction" or "feature selection" is one based on so called Karhunen-Loeve expansion. We do not recommend this method for selecting the compression matrix C for the following reasons.

(1) There may be loss of information concerning the separability of classes. We have encountered an extreme case in the example of section 5, where after compression, two populations have become identical.

(2) There is no apparent relation between this method and misclassification probability. It is not apparent what the compression matrix C selected by this method does to the misclassification probability.

We rather recommend Wilks' technique for selecting the compression matrix. We have noted in sections 4 and 5 how the selection of C by Wilks' technique is related to misclassification probability and why it may be expected to preserve the misclassification probability. Besides, as the compression matrix C selected by Wilk's technique maximizes the between scatter matrix, it may be expected that there will be no loss in information concerning the separability of classes.

## References

1. Anderson, T. W. (1958, An Introduction to Multivariate Statistical Analysis, Wiley, New York.

2. Chien, Y. T. and Fu, K. S. (1967), "On the Generalized Karhunen-Loeve Expansion." IEEE Trans. Inf. Theory 13, 518-520.

3. Meisel, W. S. (1972), Computer-Oriented Approach to Pattern Recognition, Academic Press, New York.

4. Rao, C. R. (1965), Linear Statistical Inference and Its Applications, Wiley, New York.

5. Wilks, S. S. (1967), "Multidimensional Statistical Scatter" in Collected Works of S. S. Wilk, Wiley, New York.

# Discriminant Analysis Using Certain Normed Exponential Densities.[1]

R. S. Chhikara[2]

and

P. L. Odell[3]

## 1. Introduction

The statistical discriminant analysis technique is a relatively old one [1], [2]; yet in recent years there has been renewed interest generated primarily by the desire to develop automatic statistical recognition systems; both analog and digital [3], [4]. Electronic and optic scientists have developed multichannel spectral measuring devices [5] which when attached to aircrafts or space crafts take enormous amounts of data whose speedy reduction depends on rapid repeated performance of a discriminant algorithm using high speed computing machines.

The purpose of this paper is to discuss the Bayes discriminant analysis using certain normed exponential probability densities as models and to provide ways to reduce computations that are performed for discriminant analysis in the remote sensing application [6], [7]. For clarity and completeness we

[1] This research was supported by NASA Manned Spacecraft Center under Contract NAS 9-12775.

[2] Texas Tech University, Lubbock, Texas 79409.

[3] University of Texas, Dallas, Texas 75080.

will rewiew briefly the classical statistical discriminant problem. Let I denote an individual belonging to one of m distinct populations. Assume that each member of the union of the m populations possesses a finite set of observable common characteristics or features which we denote by $C = (C_1,C_2,\ldots,C_p)^T$ whose observed values are denoted by $x = (x_1,x_2,\ldots,x_p)^T$ such that $x_j$ is the observed value of the characteristic $C_j$, $j = 1,2,\ldots,p$. If one assumes that the characteristics $C = (C_1,C_2,\ldots,C_p)^T$ are selected a priori, the discriminant problem can be summarized as follows:

<u>The Bayesian Discriminant Problem</u>. Let $\Pi_1,\Pi_2,\ldots,\Pi_m$ denote m distinct populations whose known multivariate probability density functions of the p-dimensional measurement random vector x are denoted by $p_1(x)$, $p_2(x),\ldots, p_m(x)$, respectively. Let $q_1,q_2,\ldots,q_m$ be the known a priori probabilities that an individual, I, be selected from a population $\Pi_1,\Pi_2,\ldots,\Pi_m$, respectively, Let $C(i|j)$ be the cost of misclassifying an individual from population $\Pi_j$ as being from population $\Pi_i$ such that

$$\begin{aligned} C(i|j) &> 0 \qquad i \neq j \qquad i,j = 1,2,\ldots,m \\ &= 0 \qquad i = j \qquad i,j = 1,2,\ldots,m \,. \end{aligned} \tag{1.1}$$

Given the $p \times 1$ measurement vector x made on the characteristics of an individual, I, selected at random from the union of the populations $\Pi_1,\ldots,\Pi_m$, the problem is to formulate a decision rule R which miminizes the expected cost of misclassification for assigning I to one of the populations $\Pi_i$, $i = 1,\ldots,m$.

Let $R = (R_1, R_2, \ldots, R_m)$ denote an exhaustive partitioning of the Euclidean p-dimensional space into m mutually exclusive subsets such that if the observation vector x belongs to $R_i$, then we assign the individual, I, which generated the observation vector x to the population $\Pi_i$. Note that

$$L(R) = \sum_{i=1}^{m} q_i \sum_{\substack{j=1 \\ j\neq i}}^{m} C(j|i)P(j|i) \tag{1.2}$$

is the expected cost of misclassification associated with an individual, I, where

$$P(j|i) = \int_{R_j} p_i(x)dx \tag{1.3}$$

is the probability that x belongs to $R_j$ given that the individual I, is from $\Pi_i$. Clearly, there exists many partitions R, such that L(R) in (1.2) is minimized. The following theorem proved in [1] summarizes the Bayesian approach for computing the optimal procedure (partition) R.

Theorem 1. The procedure R, that minimizes the expected cost of misclassification (1.2) is defined by assigning x to $R_k$ if

$$\sum_{\substack{i=1 \\ i\neq k}}^{m} q_i p_i(x) C(k|i) \leq \sum_{\substack{i=1 \\ i\neq j}}^{m} q_i p_i(x) C(j|i) \tag{1.4}$$

$$j = 1,2,\ldots,m.$$

Corollary 1.1. If $C(i|j) = C$ for all i and j such that $i \neq j$, then (1.4) reduces to

$$\sum_{\substack{i=1 \\ i\neq k}}^{m} q_i p_i(x) \leq \sum_{\substack{i=1 \\ i\neq j}}^{m} q_i p_i(x), \quad j = 1,\ldots,m \tag{1.5}$$

which is an ordering of the probabilities of misclassification, which is in turn equivalent to

$$q_k p_k(x) = \max_{i=1,\ldots,m} \{q_i p_i(x)\} . \tag{1.6}$$

If further $q_i = q_j$ for all i and j = 1,2,...,m then (1.6) reduces to

$$p_k(x) = \max_{i=1,\ldots,m} \{p_i(x)\} , \tag{1.7}$$

the maximum likelihood solution of the discriminant problem.

It is important to note that one must know a great amount in order to apply Theorem 1. Unfortunately, there are many cases in practice in which the a priori probabilities $q_i$, i = 1,2,...,m are unknown. If $C(j|i)$ are unknown or not assumed equal, then the problem is not very tractable, hence in most applications where $C(i|j)$ are not known they are tacitly assumed equal for all $i \neq j$. If $q_1,\ldots,q_m$ are unknown, one may assume that $q_i = q_j$ for all i, j which implies that (1.4) is void of the $q_i$'s. Another approach which requires a previously performed sampling task would be to estimate $q_i$, i = 1,2,...,m and then approximate the Bayes solution to the problem with $q_i = \hat{q}_i$, i = 1,2,...,m, in (1.4) where $\hat{q}_i$ is an estimate of $q_i$. In the remote sensing application which interests us here, the assumptions $C(i|j) = C$ and $q_i = 1/m$ for all $i \neq j$ and all i = 1,2,...,m, respectively, are not in many cases unrealistic. The remote sensing problem can be summarized as follows:

The Remote Sensing Problem. Let an image (or scene) be a rectangular region with r rows (scan lines) and c columns (number of resolution elements per scan line) consisting of rc resolution elements (individuals). Each cell (individual, I) generates a p × 1 measurement vector $X_{ij} = \{x_{hij}\}$ where $h = 1,2,\ldots,p$, $i = 1,2,\ldots,r$, and $j = 1,2,\ldots,c$. In order to recognize a single scene, one must perform rc discriminating tasks "as effectively as possible" (if the scene is classified point by point). For details, refer to [8], [9] and [10].

The problem is conceptually a repeated application of multivariate discriminant analysis outlined earlier in which, if necessary, estimates of parameters are substituted for parameter values. That is, if $p_i(x)$, $i = 1,2,\ldots,m$, denotes the unknown probability density function of a p × 1 measurement vector X associated with the ith population $\Pi_i$, $i = 1,2,\ldots,m$, due to the unknown mean vector $\mu_i$ and covariance matrix $\sum_i$, an estimate of $p_i(x)$ is

$$\hat{p}_i(x) = p_i(x,\hat{\mu}_i,\hat{\sum}_i)$$

where $\hat{\sum}_i$ and $\hat{\mu}_i$ are estimates of $\mu_i$ and $\sum_i$.

Suppose X = x denotes an observation from an individual we wish to classify. In one of the simplest cases where normality holds and $q_i = 1/m$ and $C(i|j) = C$ for all $i \neq j$, the optimal solution is given by (1.6); or equivalently, the individual $I = I(x)$ who generated the observation x is assigned to $\Pi_k$ if

$$\ln \hat{p}_k(x) = \max_{i=1,2,\ldots,m} \{\ln \hat{p}_i(x)\} \qquad (1.8)$$

where

$$\ln \hat{p}_i(x) = K_i - 1/2(x - \hat{\mu}_i)^T \hat{\Sigma}_i^{-1}(x - \hat{\mu}_i)$$

and

$$K_i = -p/2 \ln 2\Pi - 1/2 \ln \left|\hat{\Sigma}_i\right| ,$$

and in order to evaluate (1.6) and (1.8) one must always be able to evaluate the quadratic form

$$\hat{Q}_i = (x - \hat{\mu}_i)^T \hat{\Sigma}_i^{-1}(x - \hat{\mu}_i)$$

for $i = 1,2,\ldots,m$.

## 2. On Eliminating the Normality Assumption

One notes that a computational problem in the form of evaluating a quadratic is associated with the normality assumption. Experience has shown that reasonable to excellent results in the form of minimal expected costs of misclassification are obtained when the normality assumption is made; hence gives empirical experience to support its value even though there exists cases in which one can reject statistically that the data is normal.

Since the normality assumption is an arbitrary choice of a model, a natural suggestion is to replace the assumption with a selection of an alternative multivariate probability density model which can realistically describe the density of the measurements and whose likelihood values are faster to compute. However, it is reasonable to conjecture that if one incorporates correlation in any multivariate model it will necessarily imply quadratic terms to be evaluated. Nevertheless the following example gives us some experience in our attempt to formulate a multivariate density not unlike the multivariate normal but eliminates the quadratic form.

Example 2.1. For a random vector x, let $E[x] = \mu$ and $E[(x-\mu)(x-\mu)^T] = \sum$. Since $\sum^{-1}$ is positive definite, then the quadratic form can be expressed as

$$Q_2 = (x-\mu)^T B^T B(x-\mu)$$

where B is such that

$$\sum^{-1} = B^T B \tag{2.1}$$

and is a unique lower triangular matrix. Now define

$$Y = B(x-\mu) = (Y_1, Y_2, \ldots, Y_p)^T \tag{2.3}$$

Then $$Q_2 = Y^T Y = \sum Y_i^2 \tag{2.4}$$

and $E[Y] = \phi$, a null vector, and $E[YY^T] = I$, a n×n unit matrix.

Let us denote $||Y||_r^r = \sum_{i=1}^{p} |Y_i|^r$, $0<r<\infty$. Then $Q_2 = ||Y||_2^2$, the squared Euclidean distance of x from $\mu$ weighted by the matrix $\sum^{-1}$. For $r=\infty$, denote $||Y||_\infty = \max_{k=1,2,\ldots,p} (|Y_k|)$. Note that $||Y||_r^r$ and $||Y||_\infty$ are different measures of the weighted distance of x from $\mu$. However, it is much faster to compute $||Y||_1$ and $||Y||_\infty$ than $||Y||_2^2$ as these eliminate the quadratic form. We will elaborate on this aspect in section 5.

Next, as we discussed in [17], the evaluation of probabilities of classification $P(j|i)$, i and j = 1,2,...,m, defined in (1.3) is a difficult task when $p_i(x)$ are assumed to be normal, and in particular, a theoretical solution is completely out of order if covariance matrices $\sum_i$, i = 1,2,...,m are not assumed equal. This is because $P(j|i)$ will involve multivariate normal probability integrals over arbitrary domains described by quadratic functions. This has led investigators in the past either to restrict their discussion to a highly simplified form of mean vectors $\mu_i$ $(i=1,2,\ldots,m)$ and $\sum_i = \sum$ for all $i = 1,2,\ldots,m$ or to seek refuge in a computer algorithm based upon approximations which may be far from being accurate and even expensive due to a large number of repeated computations.

With these considerations in mind, we propose the use of certain normed exponential densities given in the next section for the Bayes discriminant analysis. These densities lead to minimum number of computations, piecewise linear discriminant functions when there exists inequality among the covariance matrices (a property not attained under normality when unequal covariance matrices are assumed) and a theoretical solution regarding evaluation of probabilities of classification.

## 3. Normed Exponential Density Functions

In order to enlarge the class of density functions, we first define the r-normed exponential density.

Definition 1: For $0<r<\infty$, $f^{(r)}(y)$ is the r-norm~~al~~ed exponential density function of a random vector $Y = (Y_1,\ldots,Y_p)^T$ if

$$f^{(r)}(y) = Ke^{-c||y||_r^r} \qquad , c>0 \qquad (3.1)$$

where

$$K = [2\int_0^\infty e^{-cu^r} du]^{-p}$$

and c is determined so that $E[YY^T] = I$.

Definition 2: For $r = \infty$, the maximum normed exponential density function $f^{(\infty)}(y)$ of a random vector Y is given by

$$f^{(\infty)}(y) = Ke^{-c||y||_\infty} \quad , \qquad c>0 \qquad (3.2)$$

where

$$c = \left[\frac{2^p(p+2)!}{3}\right]^{1/(p+2)} \text{ and } \quad K = \frac{1}{p!}\left[\frac{(p+2)!}{12}\right]^{p/(p+2)}$$

(Again c was determined so that $E[YY^T] = I$)

Note that $f^{(r)}(y)$ and $f^{(\infty)}(y)$ are symmetrical about $y = 0$ and cover a wide range of multivariate density functions. For $p = 1$, $c = \sqrt{2}$ and $K = 2^{-p/2}$, the density function is given by

$$f^{(1)}(y) = (1/2^{p/2}) \exp(-\sqrt{2}\sum_1^p |y_k|) \quad , \quad \begin{array}{l} -\infty<y_k<\infty \\ k = 1,2,\ldots,p \end{array} \quad (3.3)$$

which is the multivariate analog of the double exponential density and can be interpreted as the likelihood function when a set of p observations are sampled from the univariate population with p.d.f.

$$p(y) = 1/\sqrt{2}\, e^{-\sqrt{2}|y|} \quad , \qquad -\infty<y<\infty$$

For $p = 2$, $c = 1/2$ and $K = (2\pi)^{-p/2}$ the 2-normed density function is

$$f^{(2)}(y) = (1/2\pi)^{p/2} \exp(-1/2 \sum_1^p y_k^2) \qquad (3.4)$$

which is the multivariate normal density with mean vector $\phi$ and covariance matrix I. Observe that $f^{(2)}(y)$ is less peaked at $y = 0$ as compared to $f^{(1)}(y)$ but more peaked when it is compared to $f^{(\infty)}(y)$.

Though $f^{(r)}(y)$ in (3.1) leads to several other density functions, here we will primarily be concerned with $f^{(1)}(y)$ and $f^{(\infty)}(y)$ which are suitable as models in various physical situations. In the next section, we discuss the problem of discrimination using these density functions and provide examples for better comprehension.

## 4. Bayes Discriminant Procedures

Consider the populations with the probability density functions $p_i^{(1)}(x)$ or $p_i^{(\infty)}(x)$ which can be obtained from $f^{(1)}(y)$ in (3.3) or $f^{(\infty)}(y)$ in (3.2) for the random vector $X = (x_1, \ldots, x_p)^T$,

$$X = B_i^{-1}Y + \mu_i \qquad , \quad i = 1,2,\ldots,m. \quad (4.1)$$

Accordingly, we have

$$p_i^{(1)}(x) = \frac{|B_i|}{2^{p/2}} e^{-\sqrt{2}\sum_1^p |B_{ik}(x-\mu_i)|} \qquad , \qquad (4.2)$$

$$i = 1,2,\ldots,m$$

where $B_{ik}$ is the kth row of the matrix $B_i$, that is

$$B_i = \begin{bmatrix} B_{i1} \\ \vdots \\ B_{ip} \end{bmatrix}$$

where each of $B_{ik}$, $k = 1,2,\ldots,p$, is a $1\times p$ vector. Observe that the determinant $|B_i| = |\Sigma_i^{-1}|^{1/2}$.

Similarly,

$$p_i^{(\infty)}(x) = \frac{1}{p!}\left[\frac{(p+2)!}{12}\right]^{p/(p+2)} |B_i| e^{-c \max_{k=1,2,\ldots,n} (|B_{ik}(x-\mu_i)|)} \qquad (4.3)$$

$$i = 1,2,\ldots,m$$

where

$$c = \left[\frac{2^p (p+2)!}{3}\right]^{1/(p+2)}$$

### 4.1. Populations with density functions $p_i^{(1)}(x)$, $i = 1,2,\ldots,m$.

For the sake of simplicity, let us assume equal costs of misclassification. For given a priori probabliities $q_1,q_2,\ldots,q_m$, it follows from [1. p. 142-143] that the Bayes discriminant regions $R_1,R_2,\ldots,R_m$ with respect to density functions $p_i^{(1)}(x)$, $i = 1,2,\ldots,m$ are given by

$$R_j = \{x: \sum_{k=1}^{p} |B_{ik}(x-\mu_i)| - \sum_{k=1}^{p} |B_{jk}(x-\mu_j)| \geq \frac{1}{\sqrt{2}} \log \frac{q_i}{q_j} \quad i=1,2,\ldots,m, \; i\neq j, \; j=1,2,\ldots,m.\} \tag{4.4}$$

Observe that the discriminant boundaries for regions $R_1,R_2,\ldots,R_m$ are piecewise linear and can be distinctly found for given $\mu$'s and $\sum$'s. Since the integration of a density function $p_i^{(1)}(x)$ over any domain of linear (rectangular) planes can easily be evaluated, one should be able to obtain the probability of correctly classifying an observation x from population $\pi_i$,

$$p(i|i) = \int_{R_i} p_i^{(1)}(x)dx \quad , \quad i = 1,2,\ldots,m.$$

And the probabilities of misclassifying an observation x from $\pi_i$ to another population $\pi_j$,

$$p(j|i) = \int_{R_j} p_i^{(1)}(x)dx \quad , \quad j = 1,2,\ldots,m, \; j \neq i$$

However, a Bayes region $R_j$ could be characterized by as many as $2^{2p}$ different piecewise linear functions. This is because of so many different possibilities that exist for the values in determining any inequality in (4.4). If $q_1=q_2=\ldots=q_m$, the Bayes regions are given by

$$R_j = \{x: \sum_{k=1}^{p} | B_{ik}(x-\mu_i)| \geq \sum_{k=1}^{p} | B_{jk}(x-\mu_j)| , \quad i=1,2,\ldots,m, \; i\neq j\} \quad j=1,2,\ldots,m. \tag{4.5}$$

We now give an example to illustrate the algorithm involved in obtaining the Bayes regions and $P(j|i)$.

Example 4.1: Suppose we have two populations $\pi_1$ and $\pi_2$ with mean vectors $\mu_1 = (1,-2)^T$, $\mu_2 = (-1,1)^T$ and covariance matrices

$$\Sigma_1 = \begin{bmatrix} 4 & 0 \\ 0 & 9 \end{bmatrix} , \qquad \Sigma_2 = \begin{bmatrix} 9 & 0 \\ 0 & 16 \end{bmatrix}$$

respectively. Then, due to (2.1)

$$B_1 = \begin{bmatrix} 1/2 & 0 \\ 0 & 1/3 \end{bmatrix} \quad \text{and} \quad B_2 = \begin{bmatrix} 1/3 & 0 \\ 0 & 1/4 \end{bmatrix}$$

For the sake of simplicity, assume $q_1 = q_2$. Since

$$B_1(x-\mu_1) = \begin{bmatrix} \frac{x_1-1}{2} \\ \frac{x_2+2}{3} \end{bmatrix} \quad \text{and} \quad B_2(x-\mu_2) = \begin{bmatrix} \frac{x_1+1}{3} \\ \frac{x_2-1}{4} \end{bmatrix}$$

the two density functions are

$$p_1^{(1)}(x) = \frac{1}{12} e^{-\sqrt{2}[1/2|x_1-1|+1/3|x_2+2|]}$$

$$p_2^{(1)}(x) = \frac{1}{24} e^{-\sqrt{2}[1/3|x_1+1|+1/4|x_2-1|]}$$

and the Bayes discriminant regions are determined by

$$R_1 = \{x: \ 1/2|x_1-1|+1/3|x_2+2| \leq 1/3|x_1+1|+1/4|x_2-1|\}$$

and

$$R_2 = \{x: \ 1/2|x_1-1|+1/3|x_2+2| > 1/3|x_1+1|+ 1/4|x_2-1|\}.$$

Due to absolute sign, the two-dimensional Euclidean space is partitioned into 9 different regions and the Bayes discriminant regions are given by

$$R_1 = \underset{K=1}{\cup} R_{1K}$$

where

$$\begin{aligned}
R_{11} &= \{x:\ 2x_1+x_2+1\leq 0,\ x_1>1,\ x_2>1\} \\
R_{12} &= \{x:\ 10x_1-x_2-13\geq 0,\ -1<x_1<1,\ x_2>1\} \\
R_{13} &= \{x:\ 2x_1-x_2-21\geq 0,\ x_1<-1,\ x_2>1\} \\
R_{14} &= \{x:\ 2x_1-7x_2-15\geq 0,\ x_1<-1,\ -2<x_2<1\} \\
R_{15} &= \{x:\ 10x_1-7x_2-7\geq 0,\ -1<x_1<1,\ -2<x_2<1\} \\
R_{16} &= \{x:\ 2x_1+7x_2-5\leq 0,\ x_1>1,\ -2<x_2<1\} \\
R_{17} &= \{x:\ 2x_1-x_2-21\leq 0,\ x_1>1,\ x_2<-2\} \\
R_{18} &= \{x:\ 10x_1+x_2+9\geq 0,\ -1<x_1<1,\ x_2<-2\} \\
R_{19} &= \{x:\ 2x_1+x_2+1\geq 0,\ x_1<-1,\ x_2<-2\},
\end{aligned}$$

and $R_2$ is $R_1^c$, the complement of $R_1$. Since $R_{11}$, $R_{12}$, $R_{13}$, $R_{14}$ and $R_{19}$ are empty sets,

$$R_1 = R_{15} \cup R_{16} \cup R_{17} \cup R_{18}$$

One can sketch $R_1$ and $R_2$ as in Figure 3.1 below.

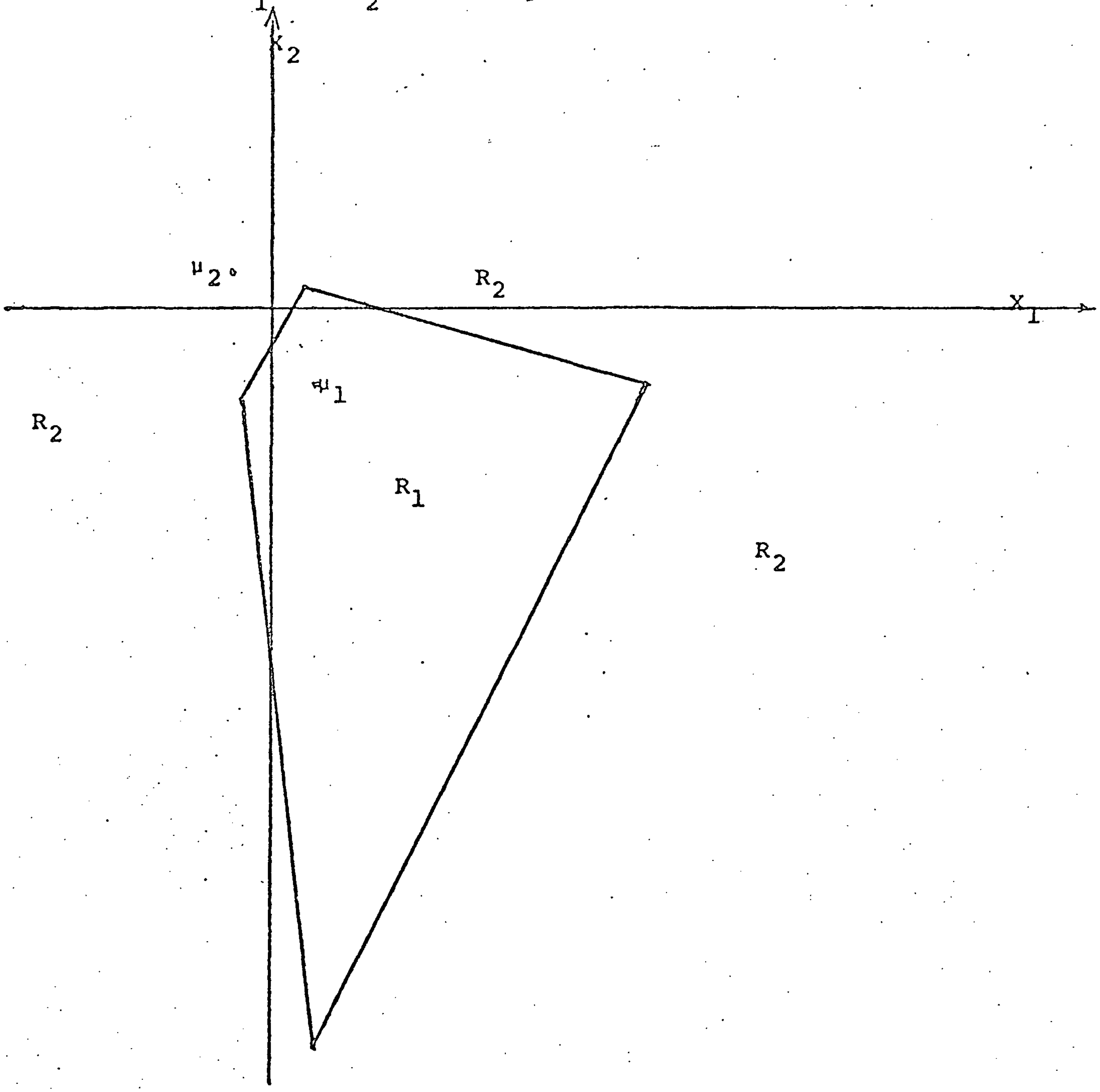


Figure 3.1: Bayes discriminant regions $R_1$ and $R_2$ for populations $\pi_1$ and $\pi_2$.

One can now evaluate the probabilities of correct classification and the probabilities of misclassification. For example, the probability of correct classifying an observation from $\pi_1$ is given by

$$
\begin{aligned}
P(1|1) &= \int_{R_1} p_1^{(1)}(x)\,dx \\
&= 1/12 \int_{-1}^{1}\int_{-2}^{10/7x_1-1} e^{1/\sqrt{2}(x_1-1)-\sqrt{2}/3(x_2+2)}\,dx_2dx_1 \\
&+ 1/12\int_{1}^{\infty}\int_{-2}^{(5-2x_1)/7} e^{-1/\sqrt{2}(x_1-1)-\sqrt{2}/3(x_2+2)}\,dx_2dx_1 \\
&+ 1/12\int_{1}^{\infty}\int_{2x-21}^{-2} e^{-1/\sqrt{2}(x_1-1)+\sqrt{2}/3(x_2+2)}\,dx_2dx_1 \\
&+ 1/12\int_{-1}^{1}\int_{-10x_1-9}^{-2} e^{1/\sqrt{2}(x_1-1)+\sqrt{2}/3(x_2+2)}\,dx_2dx_1 \\
&= 1/4(1-e^{-\sqrt{2}}-e^{-17\sqrt{2}/21}+e^{-6\sqrt{2}/7})+ \text{remaining three terms.}
\end{aligned}
$$

Similarly, the probability of misclassifying an observation from $\pi_2$ is given by

$$
\begin{aligned}
P(1|2) &= \int_{R_1} p_2^{(1)}(x)\,dx \\
&= 1/12 \int_{-1}^{1}\int_{-2}^{10/7x_1-1} e^{-\sqrt{2}/3(x_1+1)+1/2\sqrt{2}(x_2-1)}\,dx_2dx_1 \\
&+ 1/12\int_{1}^{\infty}\int_{-2}^{(5-2x_1)/7} e^{-\sqrt{2}/3(x_1+1)+1/2\sqrt{2}(x_2-1)}\,dx_2dx_1 \\
&+ 1/12\int_{1}^{\infty}\int_{2x_1-21}^{-2} e^{-\sqrt{2}/3(x_1+1)+1/2\sqrt{2}(x_2-1)}\,dx_2dx_1 \\
&+ 1/12\int_{-1}^{1}\int_{-10x_1-9}^{-2} e^{-\sqrt{2}/3(x_1+1)+1/2\sqrt{2}(x_2-1)}\,dx_2dx_1
\end{aligned}
$$

which can be easily evaluated. In a similar way one can find $P(2|2)$, the probability of correctly classifying an observation from $\pi_2$ and $P(2|1)$, the probability of misclassifying an observation from $\pi_1$.

In example 4.1, we had zero covariance in the covariance matrices. In the following example we consider random vectors whose components are correlated, that is covariances are not zero.

Example 4.2  Suppose there are populations $\pi_1$ and $\pi_2$ with mean vectors $\mu_1 = (2,1)^T$, $\mu_2 = (-1,-2)^T$ and covariance matrices

$$\Sigma_1 = \begin{bmatrix} 13/4 & -9/2 \\ -9/2 & 9 \end{bmatrix} , \quad \Sigma_2 = \begin{bmatrix} 100/9 & 32/3 \\ 32/3 & 16 \end{bmatrix} .$$

Then due to (2.1)

$$B_1 = \begin{bmatrix} 1 & 1/2 \\ 0 & 1/3 \end{bmatrix} \quad \text{and} \quad B_2 = \begin{bmatrix} 1/2 & -1/3 \\ 0 & 1/4 \end{bmatrix} ,$$

and so

$$B_1(x-\mu_1) = \begin{bmatrix} \frac{2x_1+x_2-5}{2} \\ \frac{x_2-1}{3} \end{bmatrix} , \quad B_2(x-\mu_2) = \begin{bmatrix} \frac{3x_1-2x_2-1}{6} \\ \frac{x_2+2}{4} \end{bmatrix} .$$

Accordingly, the normed density functions associated with $\pi_1$ and $\pi_2$ are given by

$$p_1^{(1)}(x) = 1/6 \; e^{-1/3\sqrt{2}(3|2x_1+x_2-5|+ 2|x_2-1 )}$$

and

$$p_2^{(1)}(x) = 1/16 \; e^{-1/6\sqrt{2}(2|3x_1-2x_2-1|+3|x_2+2|)}$$

Let $q_1 = \frac{3}{11}$, $q_2 = \frac{8}{11}$ . Then from (4.4), the Bayes discriminant regions are obtained as

$$R_1 = \{x: \; 2|3x_1-2x_2-1|+3|x_2+2|\geq 6|2x_1+x_2-5|+4|x_2-1|\}$$

and $R_2$ is the complement of $R_1$.

By simplying the inequality in $R_1$, we obtain

$$R_1 = \overset{8}{\underset{K=1}{U}} R_{1K}$$

where

$$R_{11} = \{x:\ \ 6x_1+11x_2-38\le 0,\ 2x_1+x_2-5\ge 0,\ 3x_1-2x_2-1\ge 0,\ x_2\ge 1\}$$
$$R_{12} = \{x:\ \ 2x_1+x_2-10\le 0,\ 2x_1+x_2-5\ge 0,\ 3x_1-2x_2-1\ge 0,\ -2\le x_2\le 1\}$$
$$R_{13} = \{x:\ \ 2x_1+3x_2-6\le 0,\ 2x_1+x_2-5\ge 0,\ 3x_1-2x_2-1\ge 0,\ x_2\le -2\}$$
$$R_{14} = \{x:\ \ 6x_1+x_2-14\le 0,\ 2x_1+x_2-5\ge 0,\ 3x_1-2x_2-1\le 0,\ x_2\ge 1\}$$
$$R_{15} = \{x:\ \ 2x_1+3x_2-6\ge 0,\ 2x_1+x_2-5\le 0,\ 3x_1-2x_2-1\le 0,\ x_2\ge 1\}$$
$$R_{16} = \{x:\ \ 18x_1+x_2-22\ge 0,\ 2x_1+x_2-5\le 0,\ 3x_1-2x_2-1\ge 0,\ x_2>1\}$$
$$R_{17} = \{x:\ \ 6x_1+3x_2-10\ge 0,\ 2x_1+x_2-5\le 0,\ 3x_1-2x_2-1\ge 0,\ -2\le x_2\le 1\}$$
$$R_{18} = \{x:\ \ 6x_1+x_2-14\ge 0,\ 2x_1+x_2-5\le 0,\ 3x_1-2x_2-1\ge 0, x_2\le -2\}$$

Infact, we can write $R_1 = \overset{4}{\underset{K=1}{U}} A_K$ where

$$A_1 = R_{11} U\ R_{16}$$
$$= \{x:\ \ 6x_1+11x_2-38\le 0,\ 3x_1-2x_2-1\ge 0,\ 18x_1+x_2-22\ge 0,\ x_2\ge 1\}$$

$$A_2 = R_{14} U\ R_{15}$$
$$= \{x:\ \ 6x_1+x_2-14\le 0,\ 3x_1-2x_2-1\le 0,\ 2x_1+3x_2-6\ge 0,\ x_2\ge 1\}$$

$$A_3 = R_{12} U\ R_{17}$$
$$= \{x:\ \ 2x_1+x_2-10\le 0,\ 3x_1-2x_2-1\ge 0,\ 6x_1+3x_2-10\ge 0,\ -2\le x_2\le 1\}$$

$$A_4 = R_{13} U\ R_{18}$$
$$= \{x:\ \ 2x_1+3x_2-6\le 0,\ 3x_1-2x_2-1\ge 0,\ 6x_1+x_2-14\ge 0,\ x_2\le -2\}$$

A sketch of $R_1$ and $R_2$ is given in figure 3.2 below.

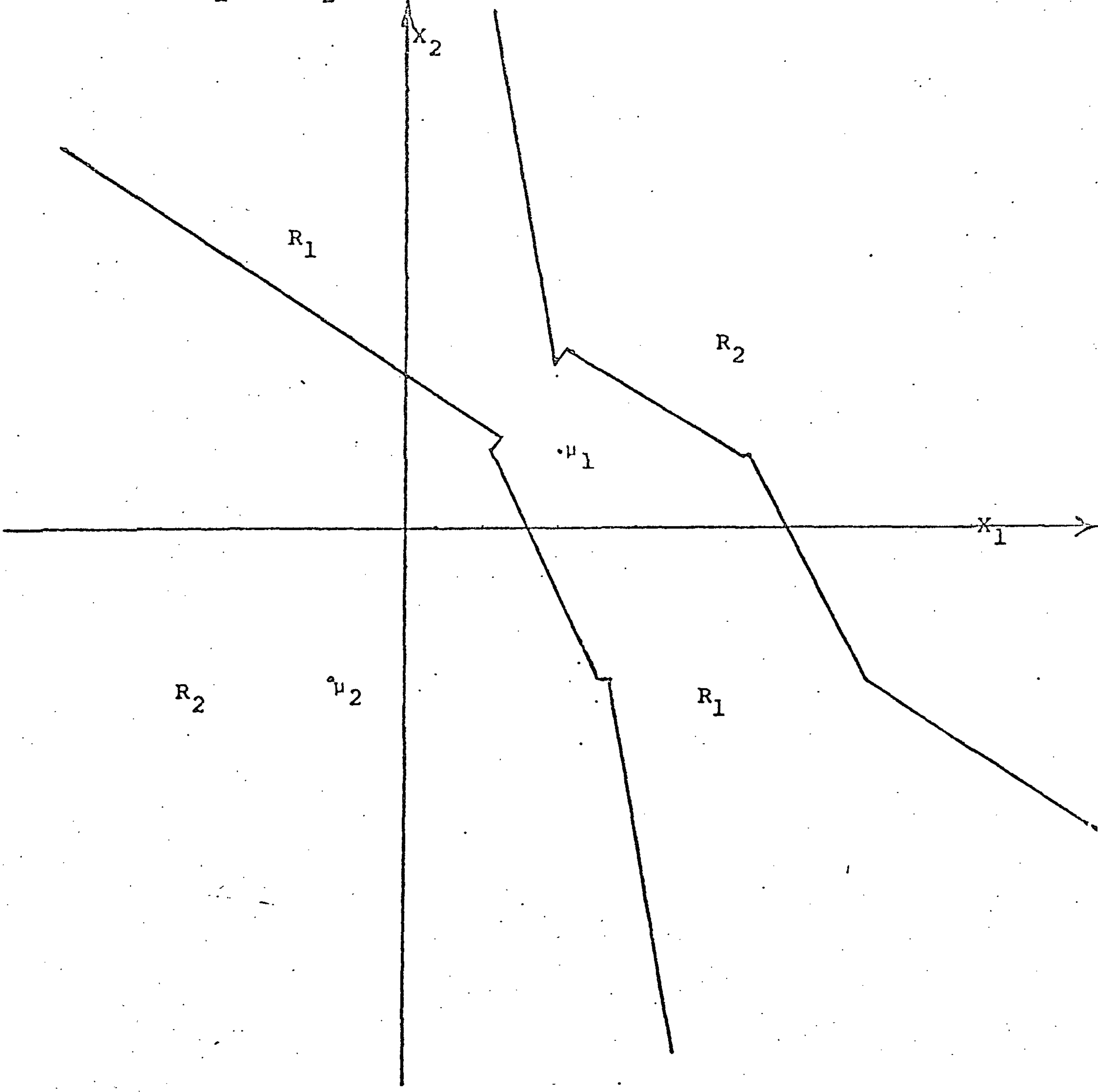


Figure 3.2: Bayes discriminant regions $R_1$ and $R_2$ for populations $\pi_1$ and $\pi_2$.

Though each of the regions $R_1$ and $R_2$ consists of more than one subregion, these are piecewise linear and probabilities of classification $P(i|i)$ and $P(i|j)$, $i$ and $j = 1,2$, can be easily evaluated.

4.2. Populations with density functions $p_i^{(\infty)}(x)$, $i = 1,2,\ldots,m$.

For given a priori probabilities $q_1,q_2,\ldots,q_m$, the Bayes discriminant regions are given by

$$R_j = \{x: \max_{K=1,2,\ldots,p}(|B_{iK}(x-\mu_i)|) - \max_{K=1,2,\ldots,p}(|B_{jK}(x-\mu_j)|)$$
$$\geq \left(\frac{3}{2^p(n+2)!}\right)^{1/(p+2)} \log \frac{q_i}{q_j},\ i=1,2,\ldots,m;\ i\neq j\}, \qquad (4.6)$$
$$j=1,2,\ldots,m.$$

If $q_1=q_2=\cdots=q_m$, (4.6) is reduced to

$$R_j = \{x: \max_{K=1,2,\ldots,p}(|B_{iK}(x-\mu_i)|)\geq \max_{K=1,2,\ldots,p}(|B_{jK}(x-\mu_j)|),$$
$$i=1,2,\ldots,m;\ i\neq j\}, \qquad (4.7)$$
$$j=1,2,\ldots,m.$$

Again these discriminant regions are described by piecewise linear functions and once determined these lead to an exact evaluation of probabilities of classification $P(i|i)$ and $P(j|i)$, $i$ and $j=1,2,\ldots,m$.

Example 4.3. Consider the populations in Example 4.2 with the density functions

$$p_1^{(\infty)}(x) = \frac{1}{3\sqrt{2}}\, e^{-1/3\,\sqrt[4]{2}\,\max(3|2x_1+x_2-5|,\ 2|x_2-1|)}$$

and

$$p_2^{(\infty)}(x) = \frac{1}{8\sqrt{2}}\, e^{-1/6\,\sqrt[4]{2}\,\max(2|3x_1-2x_2-1|,\ 3|x_2+2|)}$$

For $q_1=q_2$, the Bayes discriminant regions are

$$R_1 = \{x:\ \max(2|3x_1-2x_2-1|,\ 3|x_2+2|) \geq \max(6|2x_1+x_2-5|, 4|x_2-1|)\}$$

and $R_2$ is the complement of $R_1$. After the possibilities for the inequality are considered we obtain

$$R_1 = \bigcup_{K=1}^{4} A_K$$

where

$$A_1 = \{x:\ 2|3x_1-2x_2-1| \geq 6|2x_1+x_2-5|,\ 2|3x_1-2x_2-1| \geq 3|x_2+2|,\ 6|2x_1+x_2-5| \geq 4|x_2-1|\}$$

$$A_2 = \{x:\ 2|3x_1-2x_2-1| \geq 4|x_2-1|,\ 2|3x_1-2x_2-1| \geq 3|x_2+2|,\ 6|2x_1+x_2-5| \leq 4|x_2-1|\}$$

$$A_3 = \{x:\ 3|x_2+2| \geq 6|2x_1+x_2-5|,\ 2|3x_1-2x_2-1| \leq 3|x_2+2|,\ 6|2x_1+x_2-5| \geq 4|x_2-1|\}$$

$$A_4 = \{x:\ 3|x_2+2| \geq 4|x_2-1|,\ 2|3x_1-2x_2-1| \leq 3|x_2+2|,\ 6|2x_1+x_2-5| \leq 4|x_2-1|\}.$$

By solving the inequalities in $A_1, A_2, A_3$ and $A_4$ for different possible cases, we obtain $R_1$ and $R_2$ as shown in figure 3.3.

We omit the evaluation of probabilities of classification and let the interested readers carry out these computations.

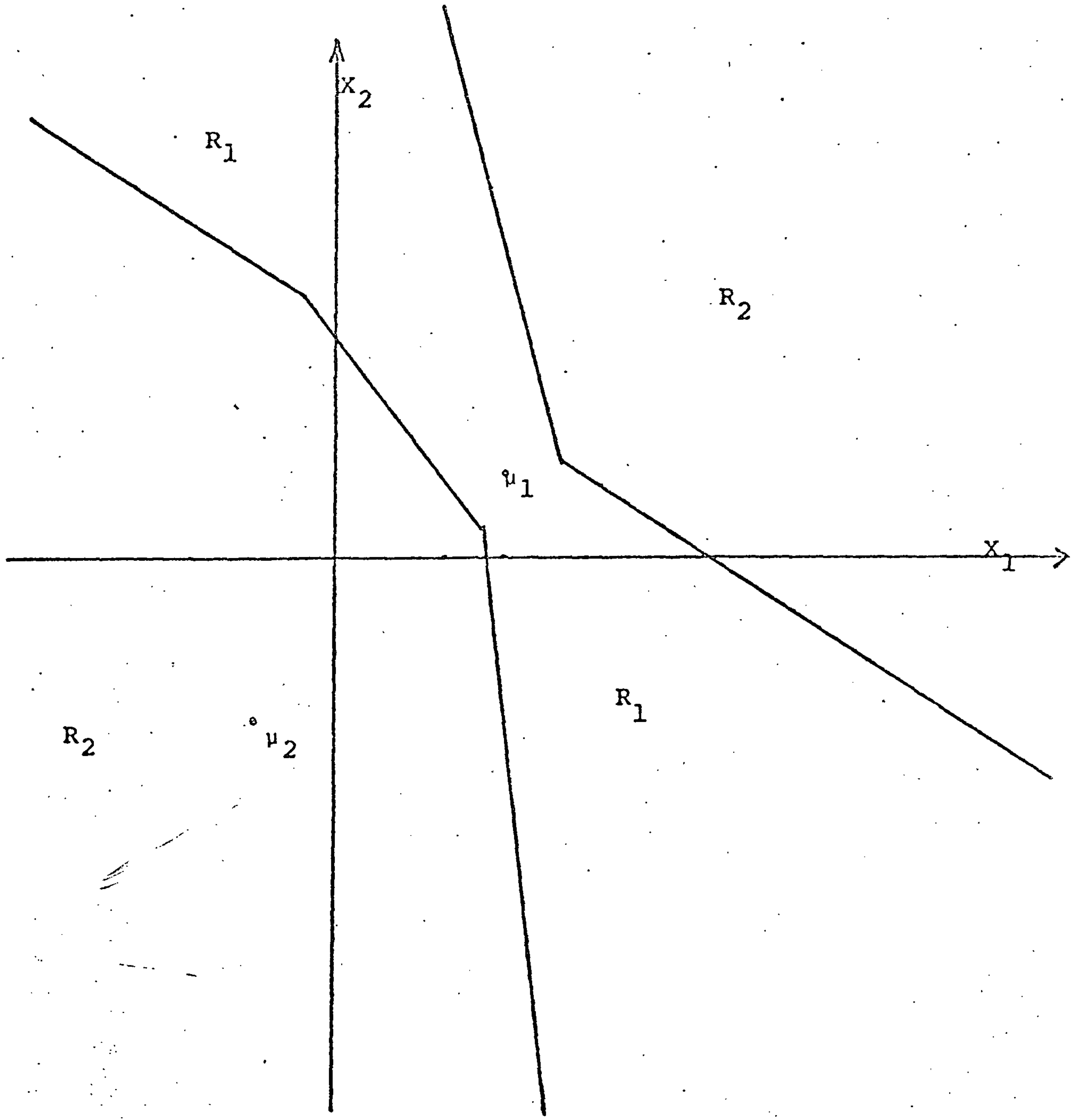


Figure 3.3: Bayes discriminant regions $R_1$ and $R_2$ for populations $\pi_1$ and $\pi_2$.

## 5. Computational Aspects

Apart from a theoretical solution that the problem of classification has, the number of computations involved in the algorithm are also reduced when the density functions $p_i^{(1)}(x)$ and $p_i^{(\infty)}(x)$ are considered instead of $p_i^{(2)}(x)$, $i = 1,2,\ldots,m$, as population models. Its explanation derives from the elimination process for the quadratic form

$$Q_2 = (x - \hat{\mu}_i)^T \hat{\Sigma}_i^{-1}(x - \hat{\mu}_i)$$

that needs to be computed when finding an estimate of $p_i^{(2)}(x)$, $i = 1,2,\ldots,m$. To give an idea of computations associated with the evaluation of $Q_2$, consider the following examples.

Example 5.1. Let $p = 3$, then a quadratic form $Q = X^TAX$ can be written as

$$Q = [x_1, x_2, x_3] \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}$$

$$= [x_1a_{11} + x_2a_{21} + x_3a_{31}, x_1a_{12} + x_2a_{22} + x_3a_{32}, x_1a_{13} + x_2a_{23} + x_3a_{33}] \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}$$

$$= [b_1x_1 + b_2x_2 + b_3x_3]$$

where

$$b_i = x_1a_{1i} + x_2a_{2i} + x_3a_{3i}.$$

In the first multiplication of $X^TA$ requires 3 multiplications and $3 - 1$ additions performed 3 times to obtain the vector $b = [b_1, b_2, b_3]$. The second multiplication bX requires 3 multiplications and $3 - 1$ additions. Hence the total number of multiplications is $3 \cdot 3+3 = 3^2+3$ and the total number of additions are $3(3 - 1) + (3 - 1) = (3 + 1)(3 - 1) = 3^2 - 1$. Inductively one can deduce the formulas, $p^2 + p$ multiplications and $p^2 - 1$ additions are required to compute a value for the quadratic Q.

<u>Example 5.2</u>. Let $p = 3$, $A = \{a_{ij}\}$ be symmetric then

$$Q = [x_1\ x_2\ x_3] \begin{bmatrix} a_{11} & 0 & 0 \\ 2a_{21} & 2a_{31} & 0 \\ 2a_{31} & 2a_{32} & a_{33} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}$$

$$= [x_1a_{11} + 2x_2a_{21} + 2x_3a_{31},\ x_2a_{22} + 2x_3a_{32},\ x_3a_{33}] \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}$$

$$= b_1'x_1 + b_2'x_2 + b_3'x_3$$

where

$$b_1' = x_1a_{11} + 2x_2a_{21} + 2x_3a_{31}$$

$$b'_2 = x_2a_{22} + 2x_3a_{32}$$

$$b'_3 = x_3a_{33} .$$

In the first multiplication of $X^T\sum^{-1}$ requires $3 + (3 - 1) + (3 - 2) = 3 + 2 + 1$ multiplications and $(3 - 1) + (3 - 2)$ additions to obtain the vector $b' = [b'_1, b'_2, b'_3]$. The second multiplication $b'X$ requires 3 multiplications and $3 - 1$ additions. Hence, the total number of multiplications is $3 + (3 - 1) + (3 - 2) + 3$; and the total number of additions is $(3 - 1) + (3 - 2) + (3 - 1)$. Inductively one can deduce the formulas $p + (p - 1)...+ 1 + p = (p(p+1))/2 + p = (p^2+3p)/2$ multiplications and $(p - 1) + (p - 2) + ... + 1 + (p - 1) = (p - 1)/2 + p - 1 = (p^2 + p - 2)/2$ additions. Note that if one takes advantage of the symmetry property of the matrix $\sum$, then the savings $\Delta_M$ and $\Delta_A$ in the number of multiplications and additions are given by

$$\Delta_M = (p^2 - p)/2$$

$$\Delta_A = (p^2 - p)/2 .$$

Also, note that when $x_i = (x - \hat{\mu}_i)$ then one must add an additional $p$ additions; hence in order to compute $Q_2$ in which symmetry is exploited the number of multiplications remains $(p^2 + 3p)/2$, but the number of additions is increased by $p$ additional additions making a total of $p + (p^2 + p - 2)/2 = (p^2 + 3p - 2)/2$ additions.

Recall that the value of $m$ is the number of populations from one of which the individual generating the measurement X might come.

Hence to perform the computation to accomplish a single discriminant task will require $(p^2 + 3p)/2$ multiplications and $(p^2 + 3p - 2)/2$ additions for each $Q_i$, $i = 1,2,\ldots,m$, when $A$ is symmetric. The decision to classify follows after ordering $m$ positive numbers, $p_i(x)$, $i = 1,2,\ldots,m$. Then

$$t_0(p,m) = m$$

$$t_M(p,m) = m(p^2 + 3p)/2$$

and

$$t_A(p,m) = m(p^2 + 3p - 2)/2$$

where $t_0(p,m)$, $t_M(p,m)$, and $t_A(p,m)$; the number of orderings, multiplications, and additions (per resolution element), respectively. Since these operations must be repeated a very large number of times they should be performed using the computer assembly language upon that computer being used instead of a general language such as FORTRAN, etc. These values for known applications are not extraordinarily large, yet when put into a remote sensing application total time becomes significantly large.

If we denote by $T_0$, $T_M$, and $T_A$ the total number of orderings, multiplications and additions, respectively, per image, then

$$T_0(p,m,r,c) = mrc$$

$$T_M(p,m,r,c) = mrc(p^2 + 3p)/2$$

and

$$T_A(p,m,r,c) \;=\; mrc(p^2 + 3p - 2)/2$$

To illustrate what size of values $T_0$, $T_M$, and $T_A$ can take, consider the following "real" data where $r = 10^3$, $c = 200$, $p = 5$, and $m = 10$, then

$$T_0 = 2 \cdot 10^6$$

$$T_M = 4 \cdot 10^7$$

$$T_A = 3.8 \cdot 10^7$$

Clearly these are large numbers of operations to be performed, and when one realizes that if this image represents only approximately a 2 mile by 10 mile strip of the earth's surface, and that it is proposed by space scientists that complete earth surveys be performed by remote sensing techniques, the size of the computation task indeed is large.

Clearly, it becomes important to investigate schemes which will reduce significantly the size of the remote sensing problem. Since r, c and m are not in the strictest sense arbitrary, there appears only one parameter, the value of p, which might be reduced. Through techniques call characteristic selection [11], [12], [13] and data compression [14] one can reduce the value of p and hopefully maintain approximately the optimality of the Bayes Procedure for discriminating. A second technique developed heuristically [15] by computer scientists have proved successful in several empirical cases and can be considered a close approximation to a Bayes or optimal procedure. This technique is one which has "traded off" floating point addition and multiplication for an integer addition, in a table look up computer operation, thereby reducing the time for computing from 2 units to .066 units in one empirical example [14]. This technique has come to be known as the table look-up discriminant technique.

Further savings in computing operations can be achieved by using the new models proposed in this paper. In the case of $p_i^{(1)}(x)$ in (4.2), one needs to evaluate the linear form

$$Q_1 = \|B_i(x - \hat{\mu}_i)\|_1$$

It can be seen that this requires $p(p + 1)/2 = (p^2 + p)/2$ multiplications and $p + p(p - 1)/2 + (p - 1) = (p^2 + 3p - 2)/2$ additions. When compared to computations involved for $Q_2$, there is a saving of p multiplications but no saving in the number of additions. In most cases any such saving may not be of significance. But we should check to see if in the remote sensing application the saving will be significant. If we denote $T_0'$, $T_M'$, and $T_A'$ the total number of orderings, multiplications and additions, respectively, per image, when (4.2) is used, then

$$T_0'(p,m,r,c) = mrc$$

$$T_M'(p,m,r,c) = mrc\,(p^2 + p)/2$$

$$T'_A(p,m,r,c) \quad = \quad mrc\ [(p^2 + 3p - 2)/2]$$

In our example where $r = 10^3$, $c = 200$, $p = 5$, and $m = 10$, the numbers are

$$\begin{array}{lllllll}
T'_0 & = & 2 \cdot 10^6 & & & = & 2 \cdot 10^6 & = & T_0 \\
T'_M & = & 30 \cdot 10^6 & = & 3.0 \cdot 10^7 & < & 4 \cdot 10^7 & = & T_M \\
T'_A & = & 38 \cdot 10^6 & = & 3.8 \cdot 10^7 & = & 3.8 \cdot 10^7 & = & T_A \ .
\end{array}$$

Note that there exists a savings of 3 to 4 in multiplications by using the probability density function as defined by (4.2) instead of the normal probability density functions in this example.

Next, for the density function $p_i^{(\infty)}(x)$ in (4.3), an evaluation of

$$Q_\infty = \|B_i(x - \hat{\mu}_i)\|_\infty$$

is needed. What this will do compared to $Q_1$ is that (p-1) additions will be eliminated at the cost of an ordering operation to determine

$$\max_{i=1,2...,p} \{|B_{ik}(x - \hat{\mu}_i)|\}$$

and will not effect any change in the number of multiplications. If $T''_0$, $T''_M$, and $T''_A$ denote the total number of orderings, multiplications and additions, respectively, per image, when (4.3) is used, then

$$T''_0\ (p,m,r,c) = mrcp$$

$$T''_M\ (p,m,r,c) = mrc(p^2 + p)/2$$

$$T''_A\ (p,m,r,c) = mrc(p^2 + p)/2$$

For the numerical example where $r = 10^3$, $c = 200$, $p = 5$, and $m = 10$, we have

$$T''_0 = 10^7 > 2 \cdot 10^6 = T_0$$

$$T''_M = 3.0 \cdot 10^7 < 4 \cdot 10^7 = T_M$$

$$T''_A = 3.0 \cdot 10^7 < 3.8 \cdot 10^7 = T_A$$

This leads to a saving of approximately 3 to 4 in total number of computing operations, but has increased the number of orderings by a multiple of 5.

It may be observed from these examples that the non-zero covariances in the covariance matrices $\sum_i$, $i = 1,2,\ldots,m$, are the sources of our computational problems. It certainly would be desirable to select those characteristics which will be uncorrelated and yet discriminate well.

## 6. Concluding Remarks

There are several facets of our discussion in this paper. First it may be noted that we have considered transformed random variables obtained by linear combinations of components of a random vector. The linear combinations depend upon the covariance matrix of the random vector and are therefore not arbitrary. It is sometime desirable to have the data transformed in some suitable way so that a consequent analysis becomes more relevant and useful. For example, the technique of transforming variables is well exploited in regression analysis for making regression more nearly linear and, possibly, random variables distributed more nearly normal. Next, in his discussion on the problem of principal components, Anderson [1, chapter 11] has cited many advantages that the linear transformation of random variables has. It is therefore our hope that this paper furthers such a cause and that the consideration of the proposed normed exponential density functions leads to some kind of break in the "stalemate" which the Bayes classification problem has reached in the case of normal density functions with unequal covariance matrices regarding the evaluation of probabilities of classification.

Though we have discussed only the problem of discriminant analysis with respect to the normed exponential density functions as models, the idea is sufficiently general and perhaps other problems of the multivariate analysis theory can be treated and solved by using these density functions.

## Bibliography

[1] Anderson, T. W., An Introduction to Multivariate Statistical Analysis, John Wilsy & Sons, Inc., New York, 1958.

[2] Von Mises, R. "On the classification of observation data into distinct groups," Annals of Mathematical Statistics, Vol. 16, No. 1, (1945), pp. 68-73.

[3] Nagy, G., "State of the art of pattern recognition," Proceeding of IEEE, Vol. 56, No. 5, May 1968, pp. 836-862.

[4] Fu, K. S., Langreke, D. E., and Phillips, T. L., "Information processing of remotely sensed agricultural data," Proceedings of the IEEE, Vol. 57, No. 4, April, 1969, pp. 639-653.

[5] Remote Sensing of Earth Resources, NASA SP 7036, A Literature Survey with Indexes, September 1970.

[6] Proceedings of 6th International Symposium on Remote Sensing of the Environment, University of Michigan, May , 1970.

[7] Proceedings of 7th International Symposium on Remote Sensing of the Environment, University of Michigan, May, 1971.

[8] Holmes, R. A., and MacDonald, R. B., "The physical basis of system design for remote sensing in agriculture," Proceedings IEEE, Vol. 57, April 1969, pp. 629-639.

[9] Landgrebe, D. A. and LARS Staff, "LARSYAA, A Processing system for airborne earth resource data," LARS Information Note 091968, Purdue University, Lafayette, Indiana, September, 1969.

[10] Landgrebe, D. A., and Phillips, T. L., "A multichannel image data handling system for agricultural remote sensing," Proceeding of Seminar on Computerized Image Handling Techniques, Washington, D. C., June 1967, pp. XIT-1 to 10.

[11] Marill, T. and Green, D. M., "On the effectiveness of receptors in recognition system," IEEE Trans. Information Theory, IT-9, January 1963.

[12] Lewis, P. M. "The characteristic selection problem in recognition systems," IRE Transaction on Information Theory, IT-8, February 1962, pp. 171-178.

[13] Levine, M. D. "Feature selection: a survey," Proceedings of IEEE, Vol. 57, No. 8, August 1969, pp. 1391-1408.

[14] Eppler, W. G., Helmke, C. A., and Evans, R. H., "Table look-up approach to pattern recognition," Proceedings of 7th International Symposium on Remote Sensing of the Environment, The University of Michigan, May 1971.

[15] Okamoto, M., "An asymptotic expansion for the distribution of the linear discriminate function," Annals of Mathematical Statistics, Vol. 34, p. 1286.

[16] Odell P. L., "Computational Problems Associated with Performing Discriminate Analysis in a remote Sensing Application", Mimeographed report, Texas Center For Research, June 1972.

[17] Odell, P. L. and Chhikara, R. S., "Diagnostic Procedures for Simulation Evaluation of Some Discrimination Routines in Remote Sensing", Mimeographed report, Texas Center For Research, July, 1972.

# DYNAMIC PROGRAMMING AND CLUSTER ANALYSIS

B. S. Duran
Texas Tech University

P. L. Odell
University of Texas at Dallas

# ABSTRACT

This report discusses dynamic programming and cluster analysis. A dynamic programming technique which yields an optimal partition is motivated and discussed and its relevance to data of the magnitude of remote sensing data is noted.

# DYNAMIC PROGRAMMING AND CLUSTER ANALYSIS

B. S. Duran[1]
Texas Tech University

P. L. Odell[2]
University of Texas at Dallas

## 1. Introduction.

The utilization of discriminant analysis in the remote sensing application has been widely discussed; for example see [6] and [7]. The more basic subject of cluster analysis has also been discussed in relation to remote sensing data [2], [4], [5]. Cluster analysis is more basic in that the number of classes (populations) is not assumed known but is determined, in general, as part of the solution.

A particular cluster analysis technique used in the remote sensing data situation is Ball and Hall's [1] well known ISODATA technique. This procedure is well documented and its use, including various modifications of the original procedure, are discussed in [5]. The ISODATA procedure is an iterative procedure which has received very wide acceptance.

The cluster problem can be viewed as that of partitioning objects into m subsets such that objects within each subset are

[1] This research was supported (in part) by NASA Manned Spacecraft Center under Contract NAS9-12775.

[2] This research was supported (in part) by NASA Manned Spacecraft Center under Contract NAS9-11925.

"similar" and subjects between subsets are "dissimilar". The objective in solving the cluster problem is then to determine the optimal partitioning such that a certain criterion of homogeneity within clusters is satisfied. One way of accomplishing this objective is by complete enumeration, i.e. examine the homogeneity criterion for all possible partitions into m clusters and choose that one which is optimal. Unfortunately, the method of complete enumeration is in general impractical, even for small values of n and m.

One alternative to the complete enumeration technique is to utilize some of the techniques popularly called dynamic programming techniques in an attempt to reduce the amount of computation but yet converge on the optimal solution. Many techniques, such as hierarchichal techniques, search for the optimal solution in a class of subsets (clusters) and the optimal solution over the whole class of clusters is not guaranteed. The aim here is to motivate and discuss a dynamic programming scheme of Jensen [3].

## 2. Application of Dynamic Programming to the Cluster Problem.

In this section we consider the problem of partitioning a set of 6 objects into 3 subsets when the distance between two objects is the Eulidean metric or the criterion is the minimization of Within Groups Sum of Squares (WGSS).

Recall that WGSS is given by

$$W = \text{tr} \sum_{j=1}^{m} S_j = \sum_{j=1}^{m} W_j$$

where $S_j$ denotes the n × n scatter matrix for the $j^{th}$ cluster and tr $S_j = W_j$. Equivalently, we have

$$(1) \qquad W = \sum_{\ell=1}^{m} (\frac{1}{2n_\ell} \sum_{i=1}^{n_\ell} \sum_{j=1}^{n_\ell} d^2(X_i, X_j)) = \sum_{\ell=1}^{m} (\frac{1}{2} \sum_{i=1}^{n_\ell} \sum_{j=1}^{n_\ell} d^2_{ij})$$

where $d^2(X_i, X_j) = (X_i - X_j)^T (X_i - X_j)$.

The purpose of a dynamic programming scheme is to systematically search for groupings which yield minimum values of the quantity W, eliminating those groupings which do not yield minimum values of W and also those that are redundant.

We now discuss the problem of partitioning n = 6 objects into m = 3 subsets by complete enumeration. This will serve to motivate a dynamic programming scheme for the cluster problem given by Jensen [3].

The total number of ways of partitioning 6 objects into 3 subsets is given by

$$S(6,3) = \frac{1}{3!} \sum_{k=0}^{3} (-1)^k \binom{3}{k} (3-k)^6$$

$$= 90.$$

The 90 clustering alternatives can be classified according to their <u>distribution</u> <u>forms</u> [3]. The three distribution forms in this case are denoted by

(i) {4} {1} {1},

(ii) {3} {2} {1},

(iii) {2} {2} {2},

where each of the components in a distribution form {i} denotes the number, i, of objects in the corresponding cluster. The components of a distribution form will always be written in descending order. In our example there are 90 clustering alternatives but only 3 distribution forms. In general the number of distribution forms is substantially smaller than the number of clustering alternatives.

There are $\frac{\binom{6}{4}\binom{2}{1}}{2} = 15$ clustering alternatives corresponding to the distribution form {4}, {1}, {1}; $\binom{6}{3}\binom{3}{2} = 60$ clustering alternatives corresponding to {3}, {2}, {1}; and $\binom{6}{2}\binom{4}{2}\binom{2}{2}/3! = 15$ clustering alternatives corresponding to {2}, {2}, {2}. The clustering alternatives corresponding to each distribution form are now listed.

Distribution Form {4}, {1}, {1}:

(1, 2, 3, 4), (5), (6)

(1, 2, 3, 5), (4), (6)

(1, 2, 5, 4), (3), (6)

(1, 5, 3, 4), (2), (6)

(5, 2, 3, 4), (1), (6)

(1, 2, 3, 6), (5), (4)

(1, 2, 6, 4), (5), (3)

(1, 6, 3, 4), (5), (2)

(6, 2, 3, 4), (5), (1)

(1, 2, 5, 6), (3), (4)

(1, 5, 6, 4), (2), (3)

(5, 6, 3, 4), (1), (2)

(5, 2, 3, 6), (1), (4)

(1, 5, 3, 6), (2), (4)

(5, 2, 6, 4), (1), (3)

Distribution Form {3}, {2}, {1}:

(1, 2, 3), (4, 5), (6)

(1, 2, 3), (4, 6), (5)

(1, 2, 3), (5, 6), (4)

(1, 2, 4), (3, 5), (6)

(1, 2, 4), (3, 6), (5)

(1, 2, 4), (5, 6), (3)

(1, 2, 5), (4, 3), (6)

(1, 2, 5), (4, 6), (3)

(1, 2, 5), (6, 3), (4)

(1, 2, 6), (4, 5), (3)

(1, 2, 6), (3, 5), (4)

(1, 2, 6), (3, 4), (5)

(1, 4, 3), (2, 5), (6)

(1, 4, 3), (2, 6), (5)

(1, 4, 3), (5, 6), (2)

(1, 5, 3), (4, 2), (6)

(1, 4, 5), (2, 3), (6)

(1, 4, 5), (2, 6), (3)

(1, 4, 5), (3, 6), (2)

(1, 4, 6), (2, 3), (5)

(1, 4, 6), (2, 5), (3)

(1, 4, 6), (3, 5), (2)

(1, 5, 6), (2, 3), (4)

(1, 5, 6), (2, 4), (3)

(1, 5, 6), (3, 4), (2)

(2, 4, 5), (1, 3), (6)

(2, 4, 5), (1, 6), (3)

(2, 4, 5), (3, 6), (1)

(2, 4, 6), (1, 3), (5)

(2, 4, 6), (1, 5), (3)

(2, 4, 6), (3, 5), (1)

(2, 5, 6), (1, 3), (4)

(1, 5, 3), (4, 6), (2)

(1, 5, 3), (2, 6), (4)

(1, 6, 3), (4, 5), (2)

(1, 6, 3), (4, 2), (5)

(1, 6, 3), (2, 5), (4)

(4, 2, 3), (1, 5), (6)

(4, 2, 3), (1, 6), (5)

(4, 2, 3), (5, 6), (1)

(5, 2, 3), (4, 1), (6)

(5, 2, 3), (4, 6), (1)

(5, 2, 3), (1, 6), (4)

(6, 2, 3), (4, 5), (1)

(6, 2, 3), (4, 1), (5)

(6, 2, 3), (1, 5), (4)

(2, 5, 6), (1, 4), (3)

(2, 5, 6), (3, 4), (1)

(3, 4, 5), (1, 2), (6)

(3, 4, 5), (1, 6), (2)

(3, 4, 5), (2, 6), (1)

(3, 4, 6), (1, 2), (5)

(3, 4, 6), (1, 5), (2)

(3, 4, 6), (2, 5), (1)

(3, 5, 6), (1, 2), (4)

(3, 5, 6), (1, 4), (2)

(3, 5, 6), (2, 4), (1)

(4, 5, 6), (1, 2), (3)

(4, 5, 6), (1, 3), (2)

(4, 5, 6), (2, 3), (1)

Distribution Form {2}, {2}, {2}:

(1, 2), (3, 4), (5, 6)

(1, 2), (3, 5), (4, 6)

(1, 2), (3, 6), (4, 5)

(1, 3), (2, 4), (5, 6)

(1, 3), (2, 5), (4, 6)

(1, 3), (2, 6), (4, 5)

(1, 4), (2, 3), (5, 6)

(1, 4), (2, 5), (3, 6)

(1, 4), (2, 6), (3, 5)

(1, 5), (3, 4), (2, 6)

(1, 5), (3, 2), (4, 6)

(1, 5), (3, 6), (2, 4)

(1, 6), (3, 4), (5, 2)

(1, 6), (3, 5), (4, 2)

(1, 6), (3, 2), (4, 5)

Under complete enumeration the objective function W(WGSS) would need to be evaluated for each of the 90 clustering alternatives given above and that clustering alternative chosen for which W is a minimum. One notes from the list of clustering alternatives that under complete enumeration the WGSS would be calculated more than once for some of the clusters, for example, the cluster (1, 2, 3).

A dynamic programming scheme applied to the cluster problem is a scheme which works for the optimum grouping in stages such that at each stage the objective function is computed in such a way that redundant calculations inherent in the complete enumeration procedure are eliminated. In this way, the optimal solution will be attained in stages. The dynamic programming approach will require large amounts of rapid access storage.

The above example can be put into the framework of a dynamic programming solution as follows. The clustering alternatives are first classified according to their distribution forms. Recall that the distribution form components are listed in descending order. At the first stage the objective function for each cluster corresponding to the first distribution form component is evaluated and saved. At the second stage the objective function for the clusters corresponding to the first two components of the distribution forms is evaluated using all information from the first stage, that is, the within sum of squares is not recomputed for any cluster but "carried over" from the first stage.

For a discussion of the dynamic programming approach consider Table 3.1. The second column gives the clusters corresponding to the first component of the distribution forms, that is, the clusters available for the first stage. The number of clusters for the first stage is $\binom{6}{4} + \binom{6}{3} + \binom{6}{2} = 50$. The function W will be computed for each of the fifty clusters in stage 1. At the second stage we will have 2 clusters corresponding to the first two components of the distribution forms, that is, we will have clusters of size {4} and {1}, {3} and {2}, or {2} and {2}. Thus the total number of objects at stage 2 will be 5 or 4. The number of ways of obtaining 5 objects is given by $\binom{6}{4}\binom{2}{1} + \binom{6}{3}\binom{3}{2} = 90$. The number of ways of obtaining 4 objects is $\binom{6}{2}\binom{4}{2} + \binom{6}{3}\binom{3}{1} = 150$. The total number of ways of obtaining objects in stage 2 is therefore 240. However, there are $\frac{1}{2}\binom{6}{2}\binom{4}{2} = 45$ ways to form clusters giving rise to distribution form components {2} {2} at stage 2, that is, there are 45 redundancies. Also, for components {3} {1} at the second stage it will be necessary to add 2 entities at the third stage giving rise to distribution form {3} {1} {2} which is ultimately equivalent to form {3} {2} {1}. Thus, the total number of ways of obtaining entities for the second stage is

$$\binom{6}{4}\binom{2}{1} + \binom{6}{3}\binom{3}{2} + \frac{1}{2}\binom{6}{2}\binom{4}{2} = 30 + 60 + 45 = 135$$

TABLE 3.1

| Stage 0 | Stage 1 | Stage 2 | Stage 3 |
|---|---|---|---|
| | 1. (1, 2, 3, 4) | | |
| | 2. (1, 2, 3, 5) | | |
| | 3. (1, 2, 5, 4) | | |
| | 4. (1, 5, 3, 4) | | |
| | 5. (5, 2, 3, 4) | | |
| | 6. (1, 2, 3, 6) | | |
| | 7. (1, 2, 6, 4) | | |
| | 8. (1, 6, 3, 4) | | |
| | 9. (6, 2, 3, 4) | | |
| | 10. (1, 2, 5, 6) | | |
| | 11. (1, 5, 6, 4) | | |
| | 12. (5, 6, 3, 4) | | |
| | 13. (5, 2, 3, 4) | | |
| | 14. (1, 5, 3, 6) | | |
| | 15. (5, 2, 6, 4) | 1. (1, 2, 3, 4, 5) | |
| | 16. (1, 2, 3) | 2. (1, 2, 3, 4, 6) | |
| | 17. (1, 2, 4) | 3. (1, 2, 3, 5, 6) | |
| | 18. (1, 2, 5) | 4. (1, 2, 4, 5, 6) | |
| | 19. (1, 2, 6) | 5. (1, 2, 3, 5, 6) | |
| | 20. (1, 3, 4) | 6. (2, 3, 4, 5, 6) | |
| | 21. (1, 3, 5) | 7. (1, 2, 3, 4) | |
| | 22. (1, 3, 6) | 8. (1, 2, 3, 5) | |
| | 23. (1, 4, 5) | 9. (1, 2, 3, 6) | |
| | 24. (1, 4, 6) | 10. (1, 2, 4, 5) | |
| 1. ( ) | 25. (1, 5, 6) | 11. (1, 2, 4, 6) | 1. (1,2,3,4,5,6) |
| | 26. (2, 3, 4) | 12. (1, 2, 5, 6) | |
| | 27. (2, 3, 5) | 13. (1, 3, 4, 5) | |
| | 28. (2, 3, 6) | 14. (1, 3, 4, 6) | |
| | 29. (2, 4, 5) | 15. (1, 3, 5, 6) | |
| | 30. (2, 4, 6) | 16. (1, 4, 5, 6) | |
| | 31. (2, 5, 6) | 17. (2, 3, 4, 5) | |
| | 32. (3, 4, 5) | 18. (2, 3, 4, 6) | |
| | 33. (3, 4, 6) | 19. (2, 3, 5, 6) | |
| | 34. (3, 5, 6) | 20. (2, 4, 5, 6) | |
| | 35. (4, 5, 6) | 21. (3, 4, 5, 6) | |
| | 36. (1, 2) | | |
| | 37. (1, 3) | | |
| | 38. (1, 4) | | |
| | 39. (1, 5) | | |
| | 40. (1, 6) | | |
| | 41. (2, 3) | | |
| | 42. (2, 4) | | |
| | 43, (2, 5) | | |
| | 44. (2, 6) | | |
| | 45. (3, 4) | | |
| | 46. (3, 5) | | |
| | 47. (3, 6) | | |
| | 48. (4, 5) | | |
| | 49. (4, 6) | | |
| | 50. (5, 6) | | |

a reduction of 105.

The number of distinct sets containing either 4 or 5 objects for stage 2 is $\binom{6}{5} + \binom{6}{4} = 21$. These are listed under stage 2 in the table 3.1 and are called states. Thus there are 21 states in stage 2. There were 50 states in stage 1. Five of the 135 feasible ways of obtaining states in stage 2 are indicated in Table 3.1.

The final stage of the process is stage 3. The final stage will result in 3 clusters. There is only one state in the final stage, the one involving all six objects. The number of ways of arriving at the six objects in the final state is

$$\binom{6}{5}\binom{1}{1} + \binom{6}{4}\binom{2}{2} = 6 + 15 = 21,$$

that is, there are 21 <u>feasible arcs</u> from stage 2 to stage 3.

For the example with n = 6 and m = 3 there are a total of 135 + 21 = 156 feasible arcs. If one includes the number of initial states then there are 156 + 50 = 206 feasible arcs. Each feasible arc results in what is called a <u>transitional calculation</u> defined by

$$(2) \qquad T(g_k) = \frac{1}{n_k} \sum_{i<j\varepsilon g_k} d_{ij}^2$$

where $g_k$ denotes a group of $n_k$ objects and $d_{ij}$ the distance between $X_i$ and $X_j$.

The total enumeration procedure involves 90 clustering alternatives and 3 transitional calculations for each alternative resulting in a total of 270 transitional calculations. The dynamic programming approach involves 206 or 64 fewer transitional calculations.

Under dynamic programming suppose there exists a state, at some stage k, containing objects $X_1,\ldots,X_q$, $q \leq n$. The dynamic programming procedure stores in memory the optimal way to partition the q objects in k nonempty and mutually exclusive subsets. In later stages in which the q objects are partitioned into k subsets it is not necessary to recompute all feasible ways of performing the partitioning.

As an illustration consider our example with n = 6, m = 3.

Table 3.2

| Alternative | Transitional Calculations |
|---|---|
| 1 | T(1, 2) + T(3, 4) + T(5, 6) |
| 2 | T(1, 3) + T(2, 4) + T(5, 6) |
| 3 | T(1, 4) + T(2, 3) + T(5, 6) |

Recall that when n = 6 and m = 3 there are S(6, 3) = 90 unique clustering alternatives available. Three of these are listed in Table 3.2. Under complete enumeration 9 transitional computations would be required for these 3 alternatives. Under dynamic programming it would take 6 transitional computations for the optimal partition of (1, 2, 3, 4) into two groups of size 2. The optimal partition, say T(1, 3) + T(2, 4) = $W_2$(1, 2, 3, 4) is recorded in memory so that only one additional computation is required to determine $W_2(1, 2, 3, 4) + T(5, 6)$. For these 3 alternatives dynamic programming has eliminated $9 - 7 = 2$ redundant calculations. Actually, as n and m are increased the number of redundant arcs that are eliminated is substantial, however, relative to the total number of transitional calculations the difference may not be so great.

### 3. Jensen's Dynamic Programming Model.

There is no standard mathematical formulation for the dynamic programming problem. This is in contrast to the linear programming problem for which there does exist a precise standard formulation. The equations and formulas pertinent to a dynamic programming problem depend on the particular situation at hand. The problem is usually reduced to a recursive relationship or equation which reflects the multiple interrelated decisions inherent in the dynamic programming procedure and which result in the final "optimal" result.

Jensen's dynamic programming formulation is given in terms of the recursive equation

$$(3) \qquad W_k(z) = \begin{cases} 0 & \text{if } k = 0, \\ \min\limits_{y} [T(z-y) + W_{k-1}(y)], & \text{if } k = 1, 2, \ldots, m_0. \end{cases}$$

where

m ≡ number of disjoint and non-empty subsets into which the n objects are to be partitioned,

$k \equiv$ index or stage variable,

$m_0 \equiv m$ if $n \geq m$, and $n - m$ if $n < m$,

$z \equiv$ state variable representing a given set of objects at stage k,

$y \equiv$ state variable representing a given set of objects at stage k-1,

$z - y \equiv$ subset of all objects contained in z but not in y,

$T(x-y) \equiv$ is the "transition cost" of the objects in the cluster of objects in (z-y).

The variables y and z represent 2 states (sets of objects) in stages k-1 and k, respectively. The difference z-y represents those objects contained in the stage k state but not in the stage k-1 state. T(z-y) then represents the "transition cost" or WGSS for those objects which are combined with the stage k-1 state objects and $W_k(z) = \min_y [T(z-y) + W_{k-1}(y)]$ gives the minimum value for WGSS in partitioning the objects represented by z into k disjoint and nonempty subsets. It will be seen that the use of formula (3) calls for a substantial amount of bookkeeping. Recall from section 2, e.g. (2) that if $g_i$ denotes a cluster of $n_i$ objects then the transition cost $T(g_i)$ is given by

$$T(g_i) = \frac{1}{n_i} \sum_{k<j\varepsilon g_i} d_{kj}^2$$

which is actually the WGSS for cluster $g_i$.

Note that the number of stages is $m_0 = m$ if $n \geq m$, and $n - m$ if $n < 2m$. The reason for this is that if $n < 2m$ there must always be at least $n - m + 1$ single-object clusters. The transition cost T for a single-object cluster is 0 so that single object clusters add nothing to W. Consequently, the process may be terminated at stage $m_0$ and all remaining clusters are assumed to be single-object clusters. Also, in computing $W_k(z)$ it should be emphasized that the objects corresponding to any state in stage k consist of objects contained in some set corresponding to some state y of stage k - 1 and objects contained in another set represented by z - y.

As an example to illustrate the notions involved in the recursive equation (3) consider state 37 of stage 1 and state 15 of stage 2 when n = 6 and m = 3 (Table 3.1). In this case y represents the objects (1, 3) in stage 37 of stage 1, z represents the objects (1, 3, 5, 6) in state 15 of stage 2, and z - y represents the objects (5, 6). The "transition cost" from stage 37 to state 15 is then

$$T(z - y) = T(5, 6) = d_{56}^2$$

The transition cost from state 37 in stage 1 to state 1 in stage 2 would be

$$T(z - y) = T(2, 4, 5) = \frac{d_{24}^2 + d_{25}^2 + d_{45}^2}{3} .$$

At the first stage the dynamic programming algorithm considers the evaluation of $W_1(z)$ for a given set of clusters. In this case

$$W_1(z) = \min_{y} [T(z-y) + W_0(y)] = T(z),$$

where z represents a given set of objects. The quantity $W_1(z)$ is computed for each of the possible clusters at the first stage. The maximum number of objects available for a cluster in the first stage, denoted by max (1), is given by

$$\max (1) \equiv n - m + 1,$$

that is, the largest cluster has n - m + 1 objects in which case the remaining clusters would be single-object clusters. The minimum number of objects, denoted by min (1), in a cluster in stage 1 is

$$\min (1) = n/m$$

if n is an even multiple of m, and

$$\min (1) = \begin{cases} [n/m] + 1, & \text{for } 1 \leq n - m[n/m], \\ n - (m-1)[n/m], & \text{for } n - m[n/m] < 1 \leq m, \end{cases}$$

when n is not an even multiple of m, where [n/m] denotes the largest integer $\leq$ n/m. The total number of clusters available for the first stage, denoted by NS(1) is given by

$$(4) \qquad NS(1) = \sum_{j=\min(1)}^{\max(1)} \binom{n}{j} \quad .$$

The first stage of the algorithm consists of computing the quantity T(z) for each of the NS(1) possible clusters.

In general the maximum number of objects in any one state in stage k is equal to the maximum sum of distribution form components from stages 1 through k inclusive. The minimum number of states is the minimum sum of the distribution form components. For max(k) and min(k) we have

$$(5) \qquad \max(k) = n - m + k \quad .$$

and

$$(6) \qquad \min(k) = k[n/m]$$

if n is an even multiple of m. If n is not an even multiple of m we have

$$(7) \qquad \min(k) = \begin{cases} ([n/m] + 1)k, \text{ for } 1 \leq k \leq n - m[n/m] \\ \\ n - (m - k)[n/m], \text{ for } n - m[n/m] < k \leq m. \end{cases}$$

The number of states available for stage k is given by

$$(8) \qquad NS(k) = \begin{cases} 1 & \text{for } k = 0 \\ \displaystyle\sum_{j=\min(k)}^{\max(k)} \binom{n}{j} & \text{for } k = 1,2,\ldots,m_0. \end{cases}$$

The total number of states available in the dynamic programming formulation is thus given by

$$(9) \qquad \sum_{k=0}^{m_0} NS(k) \quad .$$

A very important quantity in the formulation is the total number of values for $W_k(z)$ in going from stage k-1 to stage k, that is, the number of ways of forming a state in stage k. States in successive stages are connected by arcs. Two states, in stages k-1 and k, are connected by a feasible arc if the objects in the state in stage k consist of objects in the state in stage k-1. That is a feasible arc cannot exist between a state in stage k-1 and a state in stage k if an object contained in the stage k-1 state is not contained in the stage k state for $2 \leq k \leq m_0$.

In the dynamic programming algorithm the total number of feasible arcs is given by

$$(10) \qquad TFA \equiv NS(1) + \sum_{k=1}^{m_0-1} TA(k)$$

where TA(k) represents the total number of feasible arcs between stage k and stage k+1 for k = 1, 2,...,$m_0$. The value of TA(k) is given by

$$\text{(11)} \qquad TA(k) = \sum_{j=\min(k)}^{\max(k)} \; \sum_{i=1}^{\max(k+1)-\min(k)} FA(j,i) \;,$$

where

$$\text{(12)} \qquad FA(j,i) = \begin{cases} \binom{n}{j}\binom{n-j}{i} & \text{if } \min(k+1) \leq i + j \leq \max(k+1) \\ 0 & \text{otherwise} \end{cases}$$

In (11) and (12), i denotes the number of objects among a class of (<u>feasible</u>) <u>states</u> at stage k. There are $\left|\begin{matrix} n \\ i \end{matrix}\right|$ such states containing i objects, since $\left|\begin{matrix} n \\ i \end{matrix}\right|$ is the number of subsets of size i. The quantity j denotes the number of objects to be combined with the i objects to form a new state at stage k+1. Obviously we must have min(k+1) $\leq$ i + j $\leq$ max(k+1) for a state of size i + j to exist at stage k+1. If i + j satisfies the required condition, then there are $\binom{n-i}{j}$ sets of size n-i that may be added to the i objects j at a time.

Jensen gives a way of computing the efficiency of dynamic programming relative to complete enumeration. Efficiency is defined as the ratio of the total number of transitional calculations under dynamic programming to the corresponding number of calculations under complete enumeration. Alternatively, the numerator can be taken to be the total number of feasible arcs.

In either case the dynamic programming procedure is quite efficient. However, the dynamic programming procedure requires more computer memory and consequently slow-access storage could make it less useful than complete enumeration. In any event for large n and m one might be better off using some other technique such as ISODATA or hierarchial procedures.

In order to illustrate Jensen's formulation consider the example with n = 6 and m = 3. In this case n = 2m so we need to consider n - m = 6 - 3 = 3 stages, i.e. $m_0 = 3$. Furthermore,

$$\max(1) = n - m + 1 = 4$$
$$\min(1) = ([6/3])1 = 2$$
$$\max(2) = n - m + 2 = 5$$
$$\min(2) = ([6/3])2 = 4$$
$$\max(3) = n - m + 3 = 6$$
$$\min(2) = ([6/3])3 = 6$$

as can be seen from Table 3.1

The total number of states in stages 0, 1, 2, and 3 are given by

$$NS(0) = 1$$
$$NS(1) = \binom{6}{4} + \binom{6}{3} + \binom{6}{2} = 50$$
$$NS(2) = \binom{6}{4} + \binom{6}{5} = 21$$
$$NS(3) = \binom{6}{6} = 1$$

These states are listed in Table 3.1. The total number of states is thus 73. This figure agrees with Table 3.1 if NS(0) $\equiv$ 1.

From equation (12) we have, for k=1,

$$FA(3,1) = \binom{6}{3} \binom{3}{1} = 60$$

$$FA(3,2) = \binom{6}{3} \binom{3}{2} = 60$$

$$FA(4,1) = \binom{6}{4} \binom{2}{1} = 30$$

$$FA(2,2) = \binom{6}{2} \binom{4}{2} = 90$$

$$FA(3,3) = FA(4,2) = 0$$

and the total numbe. of feasible arcs between stages 1 and 2 is

$$TA(1) = 240.$$

Similarly for k=2 we have

$$TA(2) = \binom{6}{4} \binom{2}{2} + \binom{6}{5} \binom{1}{1} = 21 .$$

Thus the total number of feasible arcs in our example is, by (10),

$$TFA = NS(1) + \sum_{k=1}^{2} TA(k) = 50 + 240 + 21 = 311.$$

From section 1 it was seen that half of the stages in stage 2 corresponding to distribution form components {2}, {2} are redundant and that the 60 arcs corresponding to components {3} {1} ultimately lead to the form {3} {1} {2} which is equivalent to {3} {2} {1}. Thus in the reduced formulation the number of feasible arcs, denoted by NA,

$$NA = 50 + 135 + 21 = 206.$$

The number of feasible arcs, between stages k and k+1, elimination is given by

$$NA(k) = \sum_{i=\min(k)}^{\max(k)} \sum_{j=1}^{\max(k+1)-\min(k)} A(i,j)$$

where

$$A(i,j) = \begin{cases} \binom{n}{i}\binom{n-i}{j} & \text{if } i \neq j \\ \frac{1}{2}\binom{n}{i}\binom{n-i}{j} & \text{if } i = j \\ 0 & \text{otherwise,} \end{cases} \quad \text{and} \quad \left\{ \begin{array}{l} \min(k+1) \leq i+j \leq \max(k+1) \\ (m-k)j + i \geq n \end{array} \right.$$

The total number of arcs in the reduced formulation is given by

$$(13) \qquad NA = NS(1) + \sum_{k=1}^{m_0-1} NA(k).$$

It can be verified that (13) yields 206.

The maximum number of feasible arcs that must be evaluated in the dynamic programming formulation is then 206.

To illustrate how the dynamic programming algorithm operates let p = 2 and let the six objects be (1,1), (3,4), (5,5), (4,4), (1,2), and (5,6) or

$$X = \begin{pmatrix} 1 & 3 & 5 & 4 & 1 & 5 \\ 1 & 4 & 5 & 4 & 2 & 6 \end{pmatrix} .$$

The squared distances are then

$$d_{12}^2 = 13, \quad d_{13}^2 = 32, \quad d_{14}^2 = 18, \quad d_{15}^2 = 1, \; d_{16}^2 = 41,$$

$$d_{23}^2 = 5, \quad d_{24}^2 = 1, \quad d_{25}^2 = 8, \quad d_{26}^2 = 8, \quad d_{34}^2 = 2,$$

$$d_{35}^2 = 25, \quad d_{36}^2 = 1, \quad d_{45}^2 = 13, \quad d_{46}^2 = 5, \quad d_{56}^2 = 32.$$

According to the dynamic programming algorithm we would have:

<u>Stage 0</u>: $W_0(0) = 0.$

<u>Stage 1</u>: Compute $W_1(z) = T(z-y) + W_0(y) = T(z) + W(0) = T(z)$ for each set of objects in stage 1. For example

$$W_1(1, 2, 3, 4) = T(1, 2, 3, 4) + W_0(0)$$

$$= \frac{(d_{12}^2 + d_{13}^2 + d_{14}^2 + d_{23}^2 + d_{24}^2 + d_{34}^2}{4}$$

$$= 17.75$$

There are 50 such values for stage 1.

<u>Stage 2</u>: Compute $W_2(z) = \min\limits_{y}\{T(z-y) + w_1(y)\}$ for each set of objects in stage 2. For example

$$\begin{aligned}W_2(1,\ 2,\ 3,\ 4) = \min\ \{&T(5) + W_1(1,2,3,4),\ T(4) + W_1(1,2,3,5),\\ &T(3) + W_1(1,2,4,5),\ T(2) + W_1(1,3,4,5),\\ &T(1) + W_1(2,3,4,5),\ T(1,2) + W_1(1,2,3),\\ &T(1,3) + W_1(2,4,5),\ T(1,4) + W_1(2,3,5),\\ &T(1,5) + W_1(2,3,4),\ T(2,3) + W_1(1,4,5),\\ &T(2,4) + W_1(1,3,5),\ T(2,5) + W_1(1,3,4),\\ &T(3,4) + W_1(1,2,5),\ T(3,5) + W_1(1,4,5),\\ &T(4,5) + W_1(1,2,3)\}.\end{aligned}$$

<u>Stage 3</u>: Compute $W_3(z) = \min\limits_{y}\{T(z-y) + W_2(y)\}$ for each set of objects in stage 3. In this stage z represents the one set of objects (1,2,3,4,5,6). There are 21 feasible arcs between states in stage 2 and states in stage 3. Thus, we would choose the minimum of 21 values. As an example one of these 21 values is

$$T(2,4) + W_2(1,3,5,6)$$

corresponding to state number 15 (see Table 3.1) in which case y corresponds to the set (1,3,5,6), z corresponds to the set (1,2,3,4,5,6) and z-y corresponds to the set (2,4).

The results of the dynamic programming procedure are the clusters (1,1) and (1,2); (3,4) and (4,4); and (5,5) and (5,6) with distribution form {2}, {2}, {2}. The minimum value for W is

$$W_3(1,2,3,4,5,6) = 1.5.$$

The results are displayed in Figure 1.

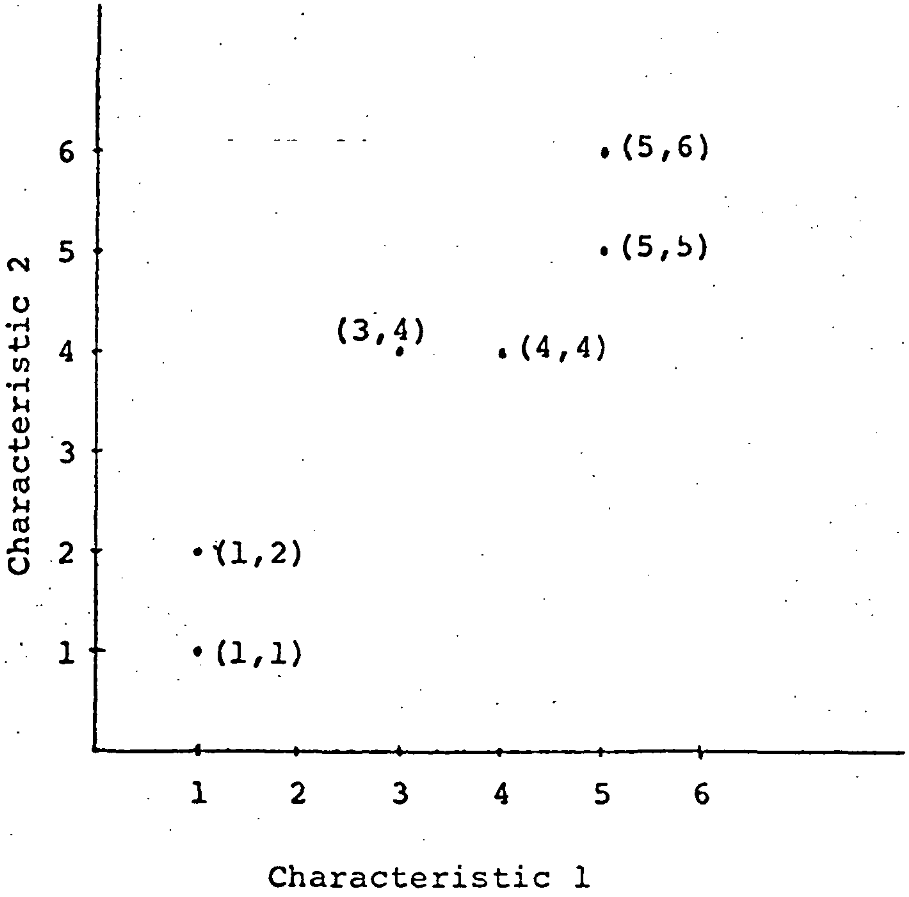


Figure 1. Graph of n=6 objects

## 4. Concluding Remarks.

It is apparent that the dynamic programming technique discussed in this report will not prove useful in a remote sensing data situation in view of the large magnitude of such data. The technique discussed herein does, however, yield an optimal partition. It appears that the search for a useful dynamic programming technique will yield one which strives for a local optimum at each stage of the process.

## References.

[1] Ball, G. H. and Hall, D. J., "ISODATA, an Iterative Method of Multivariate Analysis of Pattern Classification", *Proceedings of the International Communication Conference*, Philadelphia, Pa., June, 1966.

[2] Holley, W. A., "Description and Users' Guide for the IBM 360/44 ISODATA Program", *Lockheed Electronics Co., Inc.*, HASD, Houston, Texas, Tech. Rep. 640-TR-030, Sept., 1971.

[3] Jensen, R. E., "A Dynamic Programming Algorithm for Cluster Analysis", *Operations Research*, 12 (November-December, 1969), 1034-57.

[4] Kan, E. P. F., "ISODATA: Thresholds for Splitting Clusters", *Lockheed Electronics Co., Inc.*, HASD, Houston, Texas, Tech. Rep. 640-TR-058, Jan., 1972.

[5] Kan, E. P. F. and Holley, W. A., "More on Clustering Techniques With Final Recommendations on ISODATA", *Lockheed Electronics Co., Inc.*, HASD, Houston, Texas, Tech. Rep. 640-TR-112, May, 1972.

[6] Odell, P. L., "Computational Problems Associated With Performing Discriminate Analysis in a Remote Sensing Application", Unpublished manuscript, June, 1972.

[7] Odell, P. L. and Duran, B. S., "On the Table Look-Up in Discriminate Analysis", To appear in Journal of Statistical Computation and Simulation.

# EFFECT OF INTRACLASS CORRELATION AMONG TRAINING SAMPLES ON THE LINEAR DISCRIMINATION PROCEDURE[1]

by

J. P. Basu[2]

and

P. L. Odell[3]

[1] This research was supported (in part) by NASA Manned Spacecraft Center under Contract NAS 9-12775.

[2] Texas Tech University, Lubbock, Texas 79409

[3] The University of Texas at Dallas, Dallas, Texas 75080

## 1. Introduction

When misclassification costs are equal and prior probabilities are equal, for classifying an individual I(x), observation on whose p characteristics is p × 1 vector x, into one of two normal populations $\pi_1$ and $\pi_2$ with densities $N_p(\mu_i,\sum)$, (i=1,2), the Bayes procedure that minimizes the total misclassification probability partitions the p-dimensional real Euclidean space $E_p$ into two regions $R_1$ and $R_2$ given by

$$R_2^C = R_1 = \{x:(\mu_1-\mu_2)^T \sum^{-1} [x-1/2(\mu_1+\mu_2)] \geq 0 \quad . \tag{1}$$

The misclassification probabilities are known (Anderson [1] p 136) to be given by

$$\begin{aligned} P(2|1) &= P(X\varepsilon R_2|I(X)\varepsilon\pi_1) = \Phi(-\Delta/2) \\ P(1|2) &= P(X\varepsilon R_1|I(X)\varepsilon\pi_2) = \Phi(-\Delta/2) \end{aligned} \tag{2}$$

where $\Delta^2 = (\mu_1-\mu_2)^T \sum^{-1}(\mu_1-\mu_2)$, Mahalanobis distance between $\pi_1$ and $\pi_2$ and

$\Phi(x) = \int_{-\infty}^{x} \exp(-t^2/2)dt/\sqrt{2\pi}$.

When the parameters are unknown, they are estimated from the training samples from each population. In practice, the true values of the parameters occurring in (1) are taken to be the value of their corresponding estimates obtained from the training samples of large size. These estimates are obtained on the assumption that the samples are independent. But in reality, especially when the data are obtained by remote sensing technique, the samples are never independent. Often the assumption of independence may at best be approximately valid. So, it will be perhaps rational to assume the samples to be equicorrelated, that is, all pairs of these samples to have the same correlation. In this paper we investigate the effect of such intraclass correlation on the misclassification probabilities of linear discrimination function.

## 2. Basic Concepts

The random vectors $X_1,\ldots,X_n$ are said to be equicorrelated (Basu, Odell and Lewis [ 2]) if

(1) $D(X_i) = E[X_i-EX_i)(X_i-EX_i)^T] = \sum$, a symmetric matrix, for all i,

and (2) $\text{Cov}(X_i,X_j) = E[(X_i-EX_i)(X_j-EX_j)^T] = R$, a symmetric matrix, for all $i \neq j$. If $X_1,\ldots,X_n$ are equicorrelated random vectors, then the dispersion matrix V of their joint distribution is given by

$$V = \begin{bmatrix} \sum & R & \cdots & R \\ R & \sum & & R \\ \vdots & \vdots & & \vdots \\ R & R & \cdots & \sum \end{bmatrix} = I_n \otimes (\sum - r) + E_n \otimes R \tag{3}$$

where A $\otimes$ B denotes the Kronecker product of the matrices A and B, $I_n$ is the n × n identity matrix and $E_n$ the n × n matrix all of whose elements are 1.

The random vectors $X_1,\ldots,X_n$ are said to be simply equicorrelated if

(1) $D(X_i) = \sum$, a symmetric matrix, for all i,

and (2) $\text{Cov}(X_i,X_j) = \rho\sum$, $\rho$ being a scalar constant for all $i \neq j$.

If $X_1,\ldots,X_n$ are simply equicorrelated, then the dispersion matrix V of their joint distribution is given by

$$V = [(1-\rho)I_n + E_n] \otimes \sum \,. \tag{4}$$

Obviously V in (4) has been obtained from (3) by substituting $\rho\sum$ for R.

Let us define the np × 1 random vector X as

$$X = \begin{bmatrix} X_1 \\ \vdots \\ X_n \end{bmatrix} \tag{5}$$

and suppose that X has a multivariate normal distribution with mean and variance given by

$$EX = e_n \otimes \mu \text{ and } D(X) = V = I_n \otimes (\Sigma - r) + E_n \otimes R \tag{6}$$

where $e_n$ is the n × 1 vector, all of whose components are 1. Also, let B be the n × n orthogonal matrix given by

$$B = \begin{bmatrix} \frac{1}{\sqrt{n}} & \frac{1}{\sqrt{n}} & \frac{1}{\sqrt{n}} & \frac{1}{\sqrt{n}} \cdots & \frac{1}{\sqrt{n}} \\ \frac{1}{\sqrt{2}} & \frac{-1}{\sqrt{2}} & 0 & 0 \cdots & 0 \\ \frac{1}{\sqrt{6}} & \frac{1}{\sqrt{6}} & \frac{-2}{\sqrt{6}} & 0 \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \frac{1}{\sqrt{n(n-1)}} & \frac{1}{\sqrt{n(n-1)}} & \frac{1}{\sqrt{n(n-1)}} & \cdots & \frac{-(n-1)}{\sqrt{n(n-1)}} \end{bmatrix} \tag{7}$$

Then the p × 1 random vectors $Z_1, Z_2, \ldots, Z_n$, the n components of the np × 1 random vector Z given by

$$Z = (B \otimes I_p)X \tag{8}$$

are independent, since

$$\begin{aligned} DZ &= (B \otimes I_p)\, DX\, (B \otimes I_p)^T \\ &= (B \otimes I_p)[I_n \otimes (\Sigma - R) + E_n \otimes R](B^T \otimes I_p) \\ &= \begin{bmatrix} \Sigma + (n-1)R & \phi & \cdot\cdot & \phi \\ \phi & \Sigma - R & \cdot\cdot & \phi \\ \vdots & \vdots & & \vdots \\ \phi & \phi & \cdot\cdot & \Sigma - R \end{bmatrix} \end{aligned} \tag{9}$$

Also, $EZ = (B \otimes I_p)EX = \begin{bmatrix} \sqrt{n}\mu \\ \phi \\ \vdots \\ \phi \end{bmatrix}$ . (10)

Thus, for all i $(2 \leq i \leq n)$ $Z_i \sim N_p(\phi, \Sigma - R)$, that is, if $X_1, \ldots, X_n$ are equicorrelated samples from a $N_p(\mu, \Sigma)$ population such that their joint distribution is given by (3), then $Z_2, \ldots, Z_n$ are independent samples from $N_p(\phi, \Sigma - r)$ population. The maximum likelihood estimator of $\Sigma$-R is given by

$$\sum_{i=2}^{n} Z_i Z_i^T/(n-1) \tag{11}$$

Since $B \otimes I_p$ is an orthogonal transformation, it is well known (Anderson [ 1] p. 52, Lemma 3.3.1) that

$$\sum_{i=1}^{n} X_i X_i^T = \sum_{i=1}^{n} Z_i Z_i^T \tag{12}$$

Also from (7) and (8) it is evident that $Z_1 = \sqrt{n}\,\overline{X}$. Therefore

$$\begin{aligned} \sum_{i=2}^{n} Z_i Z_i^T &= \sum_{i=1}^{n} X_i X_i^T - n\,\overline{XX}^T \\ &= \sum_{i=1}^{N} (X_i - \overline{X})(X_i - \overline{X})^T . \end{aligned} \tag{13}$$

Thus when the samples $X_1,\dots,X_n$ are equicorrelated

$$\sum_{i=1}^{n}(X_i-\overline{X})(X_i-\overline{X})^T/(n-1) \qquad (14)$$

is an unbiased maximum likelihood estimator of $\sum$-R, but not of $\sum$.

## 3. Effect of Intraclass Correlation Among Training Samples

Let $X_1,\dots,X_n$ and $Y_1,\dots,Y_n$ be training samples from the populations $N_p(\mu_1,\sum)$ and $N_p(\mu_2,\sum)$ respectively. In practice, when Bayes classification regions (1) are defined on the basis of the training samples, $\mu_1,\mu_2$ and $\sum$ being taken respectively as $\overline{X}$, $\overline{Y}$ and $[\sum_{i=1}^{n}(X_i-\overline{X})(X_i-\overline{X})^T + \sum_{i=1}^{n}(Y_i-\overline{Y})(Y_i-\overline{Y})^T]/2(n-1)$. When the training samples are really independent, for large values of n, the misclassification probabilities of Bayes procedure are given by (2). When the training samples are equicorrelated such that the dispersion matrices $V_x$ and $V_y$ of their respective joint distribution is given by

$$V_X = I_N \times (\sum-R) + E_n \times R$$

and

$$V_Y = I_N \times (\sum-R) + E_n \times R,$$

for large n $\overline{X}$ and $\overline{Y}$ still gives estimates of $\mu_1$ and $\mu_2$; but

$$S = [\sum_{i=1}^{n}(X_i-\overline{X})(X_i-\overline{X})^T + \sum_{i=1}^{n}(Y_i-\overline{Y})(Y_i-\overline{Y})^T]/2(n-1) \qquad (15)$$

fails to provide a good estimate of $\sum$, it then provides a good estimate of $(\sum-r)$ instead.

If the training samples are equicorrelated and inadvertently S is used in place of $\sum$ in (1), then for large n the regions $R_1$ and $R_2$ become the regions

$$\hat{R}_2{}^C = \hat{R}_1 = \{x: (\mu_1-\mu_2)^T (\Sigma-R)^{-1} [x-1/2(\mu_1+\mu_2)] \geq 0 \quad . \tag{16}$$

If we write

$$\hat{W}(X) = (\mu_1-\mu_2)^T(\Sigma-R)^{-1}[X-1/2(\mu_1+\mu_2)] \ , \tag{17}$$

then the new misclassification probabilities are given by

$$P(2|1) = P(\hat{W}(X) < 0|I(X) \varepsilon \pi_1) \tag{18}$$

and

$$P(1|2) = P(\hat{W}(X) \geq 0|I(X) \varepsilon \pi_2) \ . \tag{19}$$

Now, $\hat{W}(X)$ is distributed normally with mean given by

$$E\hat{W}(X) = 1/2(\mu_1-\mu_2)^T(\Sigma-R)^{-1}(\mu_1-\mu_2) = \hat{\Delta}^2/2 \text{ if } I(X) \ \varepsilon \ \pi_1$$

and

$$E\hat{W}(X) = -1/2(\mu_1-\mu_2)^T(\Sigma-R)^{-1}(\mu_1-\mu_2) = -\hat{\Delta}^2/2 \text{ if } I(X) \ \varepsilon \ \pi_2$$

and variance under either hypothesis given by

$$\text{Var } \hat{W}(X) = (\mu_1-\mu_2)^T(\Sigma-R)^{-1}\Sigma(\Sigma-R)^{-1}(\mu_1-\mu_2). \tag{20}$$

Obviously,

$$P(2|1) = P(1|2) = \Phi(-\hat{\Delta}^2/2\sqrt{\text{Var}\hat{W}} \ ). \tag{21}$$

<u>Case 1</u>. <u>$R = \rho\Sigma$</u>, that is, <u>training samples are simply equicorrelated</u>.

$$\hat{\Delta}^2 = (1/2) \ (\mu_1-\mu_2)^T \ \Sigma^{-1}(\mu_1-\mu_2)/(1-\rho) = \Delta^2/2(1-\rho)$$

$$\text{Var } \hat{W} = (\mu_1-\mu_2)^T \ \Sigma^{-1}(\mu_1-\mu_2)/(1-\rho)^2 = \Delta^2/(1-\rho)^2$$

So, $\hat{\Delta}^2/2\sqrt{\text{Var}\hat{W}} = \Delta/2$.

Therefore, $P(1|2) = P(2|1) = \Phi(-\Delta/2)$ .

Thus when the training samples are simply equicorrelated and yet the Bayes regions are constructed on the inadvertent assumption of independence, the misclassification probabilities do not change.

Case 2. Training samples are equicorrelated, $R \neq \rho\Sigma$

We consider a numerical example to illustrate how the misclassification probabilities are changed when the training samples are equicorrelated and yet Bayes regions are defined with the inadvertent assumption of independence.

Example. Let $\pi_1$ and $\pi_2$ be two 3 dimensional normal population $N_3$ $(\mu_i, \Sigma)$ where

$$\mu_1 = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}, \quad \mu_2 = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix} \quad \text{and } \Sigma = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 4 & 1 \\ 1 & 1 & 2 \end{bmatrix},$$

Also let the training samples be equicorrelated, such that for both population

$$R = \begin{bmatrix} 0.2 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 0.4 \end{bmatrix}$$

Then $\Delta^2 = (\mu_1-\mu_2)^T \Sigma^{-1}(\mu_1-\mu_2) = 7/3,$

$\Delta/2 = 0.7638$

$\hat{\Delta}^2 = (\mu_1-\mu_2)^T (\Sigma-R)^{-1} (\mu_1-\mu_2) = 13.75$

$\text{Var}\hat{W} = (\mu_1-\mu_2)^T (\Sigma-R)^{-1} \Sigma(\Sigma-R)^{-1} (\mu_1-\mu_2) = 81.25$

$\hat{\Delta}^2/2\sqrt{\text{Var}\hat{W}} = 0.7627$

Therefore, the misclassification probabilities for

actual Bayes procedure: $\phi(-\Delta/2) = \phi(-0.7638)$

uncorrected Bayes procedure: $\phi(-\hat{\Delta}^2/2\sqrt{\text{Var}W}) = \phi(-0.7627)$.

Thus the misclassification probabilities increases when training samples are equicorrelated and yet Bayes regions are defined with the inadvertent assumption of independence.

## References

1. Anderson, T. W. (1958). An Introduction to Multivariate Statistical Analysis, Wiley, New York.

2. Basu, J. P., Odell, P. L., and Lewis, T. O. The Effects of Intraclass Correlation on Certain Significance Tests When Sampling From Multivariate Normal Populations. Accepted for publication in Communication in Statistics.

# ESTIMATION OF PROPORTION OF OBJECTS AND DETERMINATION OF TRAINING SAMPLE-SIZE IN A REMOTE SENSING APPLICATION[1]

R. S. Chhikara[2] and P. L. Odell[3]

[1] This research was supported by NASA Manned Spacecraft Center under Contract #NAS9-12775

[2] Texas Tech University, Lubbock, Texas 79409

[3] The University of Texas at Dallas, Dallas, Texas 75080

# ESTIMATION OF PROPORTION OF OBJECTS AND DETERMINATION OF TRAINING SAMPLE-SIZE IN A REMOTE SENSING APPLICATION

## 1. Introduction

The multichannel spectral measuring devices that are used as remote sensors fail to observe any vegetation, flora, etc. grown underneath timber on the earth surface. Suppose the latter is observable via a spectral measuring device and is, however, identifiable with certain uncertainity. If we know how the amount of vegetation/flora is associated with different types of timber, an evaluation of the former over a large track of land covered by forest, etc. can be made easily by the remote sensing technique.

However, as in most cases, the true parametric values such as the probability of correct identification, amount of vegetation/flora corresponding to various types of timber are unknown quantities. Hence a study of the problem first requires estimates of the unknown parameters on the basis of samples of both timber and vegetation from the ground. In the present report this estimation problem is being considered in its general form and our approach constitutes a two-stage sampling process where at the first stage samples consist of individuals called primary units, and at the second stage samples consist of categorized elements called subunits (e.g., timber and vegetation types at first and second stages, respectively, in the above example). A formal formulation of the problem is stated as follows:

Let $\Pi_i$, $i=1,2,\ldots,m$ be m different classes and every individual from these classes be characterized by p common observable features so that a measurement vector $X=(X_1,X_2,\ldots,X_p)^T$ is associated with an individual I(X) from each class. Next, associated with these individuals let there be another kind of elements categorized into k groups with proportions $p_{ij}$, $j=1,2,\ldots,k$ for each i. Further we assume that at least one element (subunit) is associated with each individual (primary unit) from every class. On the basis of an observation X, the associated primary unit I(X) may be misclassified, and let $P(i|k)$ denote the probability of misclassifying I(X) into $\pi_i$ when it belongs to $\pi_k$ and $P(i|i)$ denote the probability of correctly classifying I(X) into its class $\pi_i$. Then, given an observation X, the expected proportion of jth category subunits associated with primary unit I(X) from $\pi_i$ is given by

$$e_{ij} = \sum_{t=1}^{m} p_{tj} \, P(t|i) \qquad (1)$$

$$i=1,2,\ldots,m \text{ and } j=1,2,\ldots,k \, .$$

Note that for any fixed i, $\sum_{j=1}^{k} e_{ij} = 1$ and $e_{ij} = p_{ij}$, $j=1,2,\ldots,k$, if and only if $P(i|i) = 1$. But the later condition is an ideal one and often is not achievable. However, an effort should be made to separate out the underlying classes maximum possible, and thereby to obtain maximum possible values for $P(i|i)$, $i=1,2,\ldots,m$, so that $p_{ij}$'s can be ascertained with minimum possible error. Otherwise, the evaluation of $p_{ij}$ provided by $e_{ij}$ can be very misleading.

For an estimate of $e_{ij}$, one needs to obtain estimates for $P(t|i)$, $p_{ij}$, t, $i=1,2,\ldots,m$ and $j=1,2,\ldots,k$. For that, samples of primary units and of subunits are required from each class. Below in section 2 we outline a sampling procedure and introduce some of the notations being used later on. Our main results are obtained in section 3 and section 4 where we will discuss the interval estimation of $e_{ij}$'s and the determination of training sample size so that for a given probability an estimate allows only a specified amount of deviation about each $e_{ij}$.

## 2. Notations and Sampling Procedure

Without loss of generality, let there be two classes $\pi_1$ and $\pi_2$. Further, suppose the measurement vector X is distributed multivariate normal with mean $\mu$ if $I(X) \varepsilon \pi_1$ and mean $\nu$ if $I(X) \varepsilon \pi_2$ and has variance-covariance matrix $\sum$ for both classes. Then it follows by maximum likelihood principle [1] that $P(1|2) = P(2|1) = \Phi(-\Delta/\sqrt{2})$, where

$$\Delta^2 = (\mu-\nu)^T \sum^{-1} (\mu-\nu)$$

and

$$\Phi(a) = (1/\sqrt{2\pi})\int_{-\infty}^{a} \exp(-y^2/2)\, dy .$$

In case, $\mu$, $\nu$ and $\sum$ are known, $P(1|2)$ and $P(2|1)$ will be known. So in order to estimate $e_{1j}$ and $e_{2j}$, one only needs to estimate $p_{1j}$ and $p_{2j}$, $j=1, 2,\ldots,k$. This will be achieved by sampling $N_1$ primary units randomly from $\pi_1$ and $N_2$ from $\pi_2$ and then determining separately the observed proportions of k categories of subunits associated with these $N_1$ and $N_2$ sampled primary units.

When $\mu$, $\nu$ and $\sum$ are partially or completely unknown, $\Delta$ will be unknown and so also $P(1|2)$ and $P(2|1)$. Then for estimating any $e_{ij}$ it will require two estimates, one for $p_{ij}$ and the other for $P(1|2)$ and $P(2|1)$. The sampling procedure in that case will be to select randomly $M_1$ and $M_2$ primary units from $\pi_1$ and $\pi_2$, respectively. The observations for these selected units will be utilized to estimate $\Delta$ and thereby $P(1|2)$ and $P(2|1)$. Next, $N_1$ out of $M_1$ and $N_2$ out of $M_2$ primary units are again randomly selected and these $N_1$ and $N_2$ units are used similarly to the previous case in finding estimates for $p_{ij}$, $i=1,2$ and $j=1,2,\dots,k$.

## 3. Interval Estimation for $e_{ij}$'s

### 3.1. $\mu$, $\nu$ and $\sum$ all known

Let $n_{ijs}$ denote the number of the jth category subunits associated with sth primary unit randomly selected from $\pi_i$. Also, denote

$$n_{ij} = \sum_{s=1}^{N_i} n_{ijs} \quad \text{and} \quad n_i = \sum_{j=1}^{k} n_{ij}, \qquad j=1,2,\dots,k ,$$

for $N_i$ randomly selected primary units from $\pi_i$, $i=1,2$. Then $\hat{p}_{ij} = n_{ij}/n_i$ is an unbiased estimate of $p_{ij}$, and so also

$$\hat{e}_{ij} = \sum_{t=1}^{2} \hat{p}_{tj} P(t|i) \tag{2}$$

of $e_{ij}$, $i=1,2$ and $j=1,2,\dots,k$. This can be easily seen because the sample-size $n_i$ of subunits being a direct consequence of the sampled primary units $N_i$ can be recognized fixed for a specified value of $N_i$ and $P(t|i)$'s are known quantities. So $E[\hat{e}_{ij}] = e_{ij}$ and

$$\mathrm{Var}\ (\hat{e}_{ij}) = \sum_{t=1}^{2} [P(t|i)]^2\ p_{tj}(1-p_{tj})/n_t \tag{3a}$$

$$\mathrm{Cov}\ (\hat{e}_{ij},\hat{e}_{ij'}) = -\sum_{t=1}^{2} [P(t|i)]^2\ p_{tj}\ p_{tj'}/n_t\ ,\quad j\neq j' \tag{3b}$$

$i=1,2$ and $j=1,2,\ldots k$.

By the large sample theory, asymptotically the random vector $\hat{e}_i = (\hat{e}_{i1}, \hat{e}_{i2}, \ldots, \hat{e}_{ik})^T$ has multivariate normal distribution with mean $e_i = (e_{i1}, e_{i2}, \ldots, e_{ik})^T$ and covariance matrix E consisting of elements in (3a) and (3b). So a 100(1-α)% confidence region for $e_i$ is approximately given by the ellipsoid of points $e_i$'s satisfying

$$(\hat{e}_{ij}-e_{ij})^T\ E^{-1}\ (\hat{e}_{ij}-e_{ij}) \leq \chi^2_\alpha\ (p-1) \tag{4}$$

where $\hat{E}$ is an estimate of E obtained by replacing $p_{ij}$'s by their estimates $\hat{p}_{ij}$'s and $\chi^2_\alpha$ (p-1) is the 100(1-α)% quantile for $\chi^2$ variate with (p-1) degrees of freedom. Considering the coordinate-wise projection, this yields the simultaneous confidence intervals

$$\hat{e}_{ij} \pm [\chi^2_\alpha\ (p-1) \sum_{t=1}^{2} (P(t|i))^2\ \hat{p}_{tj}\ (1-\hat{p}_{tj})/n_t]^{1/2} \tag{5}$$

for $e_{ij}$, $j=1,2,\ldots,k$ and $i-1,2$.

## 3.2. Not all of $\mu$, $\nu$ and $\sum$ known

Here we need to estimate both $p_{ij}$'s and $P(t|i)$'s in (1) for an estimate of $e_{ij}$, $i=1,2$ and $j=1,2,\ldots,k$. Since

$$P(t|i) = \begin{cases} 1-\Phi(-\Delta/2)\ , & t=i \\ \Phi(-\Delta/2)\ , & t\neq i \end{cases} \tag{6}$$

with $i=1,2$, the estimation of any $P(t|i)$ amounts to estimating the quantity $\Phi(-\Delta/2)$. For the later an estimate is given by $\Phi(-\hat{\Delta}/2)$ where $\hat{\Delta}$ is the maximum likelihood estimate of $\Delta$ based upon samples observations $X_1,X_2,\ldots,X_{M_1}$ of $M_1$ randomly selected primary units from $\pi_1$ and observations $Y_1,Y_2,\ldots,Y_{M_2}$ of $M_2$ randomly selected primary units from $\pi_2$. Since $\Phi(-\hat{\Delta}/2)$ is a consistent estimate of $\Phi(-\Delta/2)$, due to (6) it leads to consistent estimates of $P(t|i)$, $i=1,2$ and $t=1,2$, denoted by $\hat{P}(t|i)$. Next, as in the previous case, for an estimate of $p_{ij}$, let $\hat{p}_{ij}$ be the estimate obtained from subunits associated with $N_1$ and $N_2$ randomly selected primary units from $M_1$ and $M_2$ respectively. Then the estimates

$$\hat{e}_{ij} = \sum_{t=1}^{2} \hat{p}_{tj}\ \hat{P}(t|i) \tag{7}$$

for $e_{ij}$, $i=1,2$ and $j=1,2,\ldots,k$, are consistent.

For our purpose of finding an asymptotic simultaneous confidence intervals for $e_{ij}$'s, it is suffice to find the mean square errors (MSE) for these estimates. After considering the two estimates $\hat{p}_{tj}$ and $\hat{P}(t|i)$ stochastically independent so that

$$\mathrm{Var}\ (\hat{p}_{ij}\ \hat{P}(t|i)) = [E(\hat{p}_{tj})]^2[\mathrm{Var}(\hat{P}(t|i)] + [E(\hat{P}(t|i)]^2\mathrm{Var}(\hat{p}_{tj})$$

$$+ \mathrm{Var}(\hat{p}_{tj})\ [\mathrm{Var}(\hat{P}(t|i))]\ ,$$

one can easily deduce the MSE of $\hat{e}_{ij}$, $i=1,2$ and $j=1,2,\ldots,k$.

Since $\mathrm{Var}\ (\hat{P}(1|i)) = \mathrm{Var}\ (\hat{P}(2|i)) = \mathrm{Var}\ (\Phi(-\hat{\Delta}/2))$, $i=1,2$, we obtain

$$\mathrm{Var}(\hat{e}_{1j}) = \mathrm{Var}(\hat{p}_{1j}\hat{P}(1|1))+\mathrm{Var}(\hat{P}_{2j}\hat{P}(2|1))+2\ \mathrm{Cov}(\hat{p}_{1j}\hat{P}(1|1),\hat{p}_{2j}\hat{P}(2|1)$$

$$=\left[\frac{3p_{1j}(1-p_{1j})}{n_1} + \frac{3p_{2j}(1-p_{2j})}{n_2} + p_{1j}^2+p_{2j}^2-2p_{1j}p_{2j}\right]\mathrm{Var}(\Phi(-\frac{\hat{\Delta}}{2}))$$

$$+\ [1-E(\Phi(-\frac{\hat{\Delta}}{2}))]^2\,\frac{p_{1j}(1-p_{1j})}{n_1} + [E(\Phi(-\frac{\hat{\Delta}}{2}))]^2\,\frac{p_{2j}(1-p_{2j})}{n_2}\ . \qquad (8)$$

Since it is difficult to evaluate $E[\Phi(-\frac{\hat{\Delta}}{2})]$ and $\mathrm{Var}(\Phi(-\frac{\hat{\Delta}}{2}))$, we here consider the mean square error of $\hat{e}_{ij}$ given by

$$\mathrm{MSE}(\hat{e}_{ij}) = \left[\frac{3p_{1j}(1-p_{1j})}{n_1} + \frac{3p_{2j}(1-p_{2j})}{n_2} + (p_{1j}-p_{2j})^2\right]\mathrm{MSE}(\Phi(-\frac{\hat{\Delta}}{2}))$$

$$+\ (1-\Phi(-\frac{\Delta}{2}))^2\,\frac{p_{1j}(1-p_{1j})}{n_1} + (\Phi(-\frac{\Delta}{2}))^2\frac{p_{2j}(1-p_{2j})}{n_2}\ . \qquad (8a)$$

Similarly,

$$\mathrm{MSE}(\hat{e}_{2j}) = \left[\frac{3p_{1j}(1-p_{1j})}{n_1} + \frac{3p_{2j}(1-p_{2j})}{n_2} + (p_{1j}-p_{2j})^2\right]\mathrm{MSE}(\phi(-\frac{\hat{\Delta}}{2})]$$

$$+ (\phi(-\frac{\Delta}{2}))^2 \frac{p_{1j}(1-p_{1j})}{n_1} + (1-\phi(-\frac{\Delta}{2}))^2 \frac{p_{2j}(1-p_{2j})}{n_2} \qquad (8b)$$

We denote $\mathrm{MSE}(\hat{e}_{ij}) = s_{ij}$ and let its estimate $\hat{s}_{ij}$ be obtained by replacing unknown quantities by their estimates.

Now as in the previous case, an approximate 100(1-α)% simultaneous confidence intervals for $e_{ij}$, i=1,2 and j=1,2,...,k are given by

$$\hat{e}_{ij} \pm [\hat{s}_{ij}\chi^2_\alpha(p-1)]^{1/2} \qquad (9)$$

i=1,2 and j=1,2,...,k, respectively.

The two particular cases of interest are (i) $\mu,\nu$ are unknown and $\sum$ is known and (ii) $\mu,\nu$ and $\sum$ are all unknown. For (i), the maximum likelihood estimate of $\Delta^2$ is given by

$$\hat{\Delta}^2 = (\overline{X}-\overline{Y})^T \textstyle\sum^{-1} (\overline{X}-\overline{Y}) \qquad (10)$$

and for (ii), this estimate is

$$\hat{\Delta}^2 = (\overline{X}-\overline{Y})^T S^{-1} (\overline{X}-\overline{Y}) \tag{11}$$

where $\overline{X} = \sum_1^{N_1} X_i/N_1$ , $\overline{Y} = \sum_1^{N_2} Y_i/N_2$ and

$$(N_1+N_2-2)S = \sum_1^{N_1} (X_i-\overline{X})(X_i-\overline{X})^T + \sum_1^{N_2} (Y_i-\overline{Y})(Y_i-\overline{Y})^T .$$

## 4. Sample Size

Presently our concern is to determine the sample size so that only a specified amount of error for $e_{ij}$'s is allowed by their estimates with a given probability. In specific terms the problem is to find $(n_1,n_2)$ and consequently $(N_1,N_2)$ so that $\hat{e}_{ij}$ fall simultaneously in intervals given by $e_{ij} \pm r_{ij}$, i=1,2 and j=1,2,...,k, with probability $(1-\alpha)$. However, this is equivalent to obtaining $(n_1,n_2)$ when the length of a confidence interval for $e_{ij}$ with confidence level 100 $(1-\alpha)$% is given. Hence, based upon the discussion in section 3, an asymptotic solution for the sample size is available from equations (5) and (9) for the two cases considered above.

Suppose $\mu,\nu$ and $\sum$ are known. Then using equation (5), an asymptotic sample size $(n_1,n_2)$ is the solution of

$$\sum_{t=1}^{2} [P(t|i)]^2 \hat{p}_{tj}(1-\hat{p}_{tj})/n_t = \gamma_{ij}^2/\chi_\alpha^2(p-1) , \tag{12}$$

i=1,2 for any j. After simplifying (12), we obtain

$$n_1 = C \hat{p}_{ij}(1-\hat{p}_{ij})/([\gamma_{ij}P(2|2)]^2 - [\gamma_{2j}P(1|2)]^2) \tag{13}$$

and

$$n_2 \doteq C\, \hat{p}_{2j}(1-\hat{p}_{2j})/([\gamma_{2j}P(1|1)]^2 - [\gamma_{1j}P(1|2)]^2) \tag{14}$$

where

$$C = ([P(1|1)\ P(2|2)]^2 - [P(1|2)\ P(2|1)]^2)\ \chi^2_\alpha(p-1)\ .$$

Note that for each j one finds $(n_1,n_2)$ from (13) and (14). So by taking the maximum of these solutions for each of $n_1$ and $n_2$ separately a determination of sample size is obtained. However, by a judicious choice of $\gamma_{ij}$, $i=1,2$ and $j=1,2,\ldots,k$, (13) and (14) may yield the same value for all solutions of $n_1$ and so also for $n_2$. Then any such common solution $(n_1,n_2)$ will be the desired sample size.

When at least one of $\mu,\nu$ and $\sum$ is unknown, a similar asymptotic determination of sample size is obtained from a solution of

$$\hat{s}_{ij} = \gamma^2_{ij}/\chi^2_\alpha(p-1) \tag{15}$$

$i=1,2$ for any j. Denoting MSE $(\phi(-\frac{\hat{\Delta}}{2})) = s_0$ and its estimate by $\hat{s}_0$, it follows from (8a), (8b) and (15) that

$$[3\hat{s}_0+(1-\phi(-\frac{\hat{\Delta}}{2}))^2]\ \frac{\hat{p}_{1j}(1-\hat{p}_{1j})}{n_1} + [3\hat{s}_0+(\phi(-\frac{\hat{\Delta}}{2}))^2]\ \frac{\hat{p}_{1j}(1-\hat{p}_{2j})}{n_2}$$

$$= (\bar{\hat{p}}_{1j}-\hat{p}_{2j})^2\hat{s}_0 + \frac{\gamma^2_{1j}}{\chi^2_\alpha(p-1)}$$

and

$$[3\hat{s}_0+(\phi(-\tfrac{\hat{\Delta}}{2}))^2]\,\frac{\hat{p}_{1j}(1-\hat{p}_{1j})}{n_1}+[3\hat{s}_0+(1-\phi(-\tfrac{\hat{\Delta}}{2}))^2]\,\frac{\hat{p}_{2j}(1-\hat{p}_{2j})}{n_2}$$

$$=(\hat{p}_{1j}-\hat{p}_{2j})^2\hat{s}_0+\frac{\gamma_{2j}^2}{\chi_\alpha^2(p-1)}\quad .$$

Then for a determination of $(n_1,n_2)$, we have

$$n_1=\frac{(a^2-b^2)\ \hat{p}_{1j}\ (1-\hat{p}_{1j})\ \chi_\alpha^2(p-1)}{(\hat{p}_{1j}-\hat{p}_{2j})^2\ (1-2\Phi(-\frac{\hat{\Delta}}{2}))\ \hat{s}_0\ \chi_\alpha^2(p-1)+a\gamma_{1j}^2-b\gamma_{2j}^2}\tag{16}$$

and

$$n_2=\frac{(a^2-b^2)\ \hat{p}_{2j}\ (1-\hat{p}_{2j})\ \chi_\alpha^2(p-1)}{(\hat{p}_{1j}-\hat{p}_{2j})^2\ (1-2\Phi(-\frac{\hat{\Delta}}{2}))\ \hat{s}_0\ \chi_\alpha^2(p-1)+a\gamma_{2j}^2-b\gamma_{1j}^2}\tag{17}$$

where

$$a=3\hat{s}_0+(1-\Phi(-\hat{\Delta}|2))^2$$

$$b=3\hat{s}_0+(\Phi(-\hat{\Delta}|2))\quad .$$

Once again, $n_1$ and $n_2$ are determined by taking the maximum value among such solutions of $n_1$ and $n_2$ respectively for $j=1,2,\ldots,k$ or by having a common solution derived from a judicious choice of $\gamma_{ij}$, $i=1,2$ and $j=1,2,\ldots,k$.

## 5. Univariate Case

In order to provide a specific and also somewhat interesting result, we specialize the problem to the case where the random measurement X is univariate having normal distribution with mean $\mu_1$ if $I(X) \varepsilon \pi_1$ and $\mu_2$ if $I(X) \varepsilon \pi_2$ and with variance $\sigma^2$ in both cases. Instead of considering maximum likelihood estimates $\hat{e}_{ij}$'s given in section 3, we want to find an unbiased estimate for $e_{ij}$, i=1,2 and j=1,2,...,k. Since any $\hat{p}_{ij}$ and $\hat{P}(i|k)$ are stochastically independent and $\hat{p}_{ij} = n_{ij}/n_i$ is the minimum variance unbiased estimate (MVUE) of $p_{ij}$, a similar estimate of P(i|j) in (7) leads to the MVUE of $e_{ij}$, i=1,2 and j=1,2,...,k.

In order to find the MVUE of any P(t|i), it is suffice to find the similar estimate of $\Phi(-\frac{\Delta}{2})$ where $\Delta = |\mu_1-\mu_2|/\sigma$. Without loss of generality let $\mu_1 > \mu_2$. It follows by theorem 1 in Ellison (1964) that $(2U-1)\sqrt{\nu}\,S$ has normal distribution with mean 0 and variance $\sigma^2$ where U independent of

$S^2 = \sum_1^{M_1} (X_i - \bar{X})^2 + \sum_1^{M_2} (Y_i-\bar{Y})^2/(M_1+M_2-2)$ is distributed as $\beta(\frac{\nu-1}{2},\frac{\nu-1}{2})$, beta distribut

$\nu=M_1+M_2-2$. Then the random variable

$$\frac{1}{2}\,[(2U-1)S\sqrt{\nu(4-(\frac{1}{M_1}+\frac{1}{M_2})}\;+(\bar{X}-\bar{Y})]$$

has the normal distribution with mean $(\mu_1-\mu_2)/2$ and variance $\sigma^2$. Accordingly, we observe that

$$\Phi(-\frac{\Delta}{2})=\text{Prob }\{(2U-1)S\;\sqrt{\nu[4-(\frac{1}{M_1}+\frac{1}{M_2})]}\;+(\bar{X}-\bar{Y}) \leq 0\}$$

$$=E[\text{Prob}\{(2U-1)s\;\sqrt{\nu[4-\frac{1}{M_1}+\frac{1}{M_2})]}\;+(\bar{x}-\bar{y}) \leq 0|\bar{x},\bar{y},s\}] \qquad (18)$$

where E stands for expectation with respect to variates $\overline{X}$, $\overline{Y}$ and S. Now from (18), the MVUE of $\phi(-\frac{\Delta}{2})$ is

$$\text{Prob}\ \{(2U-1)\ s\ \sqrt{\nu[4-(\frac{1}{M_1}+\frac{1}{M_2})]}+(\overline{x}-\overline{y})\leq 0|\overline{x},\overline{y},s\}\ .$$

Thus the MVUE of $\phi(-\frac{\Delta}{2})$ is

$$\text{Prob}\ \{U\leq\frac{1}{2}-(\overline{x}-\overline{y})/2s\ \sqrt{\nu[4-(\frac{1}{M_1}+\frac{1}{M_2})]}\ \ |\ \overline{x},\overline{y},s\} \tag{19}$$

where U has $\beta(\frac{\nu-1}{2},\frac{\nu-1}{2})$ distribution. Since extensive incomplete-beta integral tables are available, (19) can be easily evaluated for any given values of $\overline{x},\overline{y}$ and s obtained from sample observations on $M_1$ and $M_2$ individuals from $\pi_1$ and $\pi_2$ respectively.

Denote

$$\omega=\frac{1}{2}-\frac{\overline{x}-\overline{y}}{2s\ \sqrt{\nu(4-(\frac{1}{M_1}+\frac{1}{M_2}))}}$$

Then

$$\hat{P}(t|i)=\begin{cases}1-\hat{\phi}(-\frac{\Delta}{2})\ , & t=i\\ \\ \hat{\phi}\ (-\frac{\Delta}{2})\ , & t\neq i\end{cases} \tag{20}$$

where

$$\hat{\phi}\ (-\frac{\Delta}{2})=\frac{1}{B(\frac{\nu-1}{2}\ \frac{\nu-1}{2})}\int_0^{\omega}\ [u(1-u)]^{(\nu-3)/2}\ du \tag{21}$$

Since the use of (20) leads to the MVUE of $e_{ij}$ given by

$$\hat{e}_{ij} = \sum_{1}^{2} \hat{p}_{tj}\, \hat{P}(t|i) ,$$

it is convenient to find $\mathrm{var}(\hat{e}_{ij})$. This easily follows from (8) if $\mathrm{var}(\hat{\phi}(-\frac{\Delta}{2})$ is evaluated. For that we only need to evaluate $E[(\hat{\Phi}(-\frac{\Delta}{2}))^2]$ because we already know $E[\hat{\Phi}(-\frac{\Delta}{2})] = \Phi(-\frac{\Delta}{2})$. Observing that $(\overline{X}-\overline{Y})\sqrt{\nu}/S$ is a non-central t variable with $\nu$ degrees of freedom, denoted by $t(\Delta,\nu)$, we have

$$E[(\hat{\Phi}(-\frac{\Delta}{2}))^2] = E[(\mathrm{Prob}\ \{U \leq \frac{1}{2} - (\overline{x}-\overline{y})/2s\ \sqrt{\nu[4-(\frac{1}{M_1}+\frac{1}{M_2})]}\ \ |\ \overline{x},\overline{y},s\})^2]$$

$$= E[(\mathrm{Prob}\ \{2U-1 \leq -t(\Delta,\nu)/\nu\ \sqrt{[4-(\frac{1}{M_1}+\frac{1}{M_2})]}\ \ |t\})^2]$$

$$= E[\mathrm{Prob}\ \{\max(2U_1-1,\ 2U_2-1) \leq -t(\Delta,\nu)/\nu\ \sqrt{[4-(\frac{1}{M_1}+\frac{1}{M_2})]}\ \ |t\}] \qquad (22)$$

where $U_1$ and $U_2$ are two independent random variables, each distributed as $\beta(\frac{\nu-1}{2},\frac{\nu-1}{2})$. If we let $W = \max\ (2U_1-1,\ 2U_2-1)$, the density function of W is

$$f(\omega) = \frac{1}{2^{\nu-3}B(\frac{\nu-1}{2},\frac{\nu-1}{2})}\ (1+\omega)^{(\nu-3)/2}(1-\omega)^{(\nu-3)/2}\ I_{(1+\omega)/2}\ (\frac{\nu-1}{2},\frac{\nu-1}{2}), \qquad -1\leq\omega\leq1$$

where $I_x(a,b)$ stands for the incomplete-beta integral, and

$$F(\omega)=\text{Prob }\{W \leq \omega\}= \int_{-1}^{\omega} f(\omega)\ d\omega$$

$$= \frac{2}{B(\frac{\nu-1}{2},\frac{\nu-1}{2})} \int_{0}^{(1+\omega)/2} y^{\frac{\nu-3}{2}} (1-y)^{\frac{\nu-3}{2}} I_y(\frac{\nu-1}{2},\frac{\nu-1}{2})\ dy$$

$$= \frac{4}{(\nu-1)B^2(\frac{\nu-1}{2},\frac{\nu-1}{2})} \int_{0}^{(1+\omega)/2} y^{\nu-2}(1-y)^{\nu-2}\ \{1+ \sum_{n=0}^{\infty} \frac{B(\frac{\nu+1}{2},n+1)}{B(\nu-1,n+1)}\ y^{n+1}\}dy$$

$$= \frac{4}{(\nu-1)B^2(\frac{\nu-1}{2},\frac{\nu-1}{2})} \Big[B(\nu-1,\nu-1)\ I_{\frac{1+\omega}{2}}\ (\nu-1,\nu-1)$$

$$+ \sum_{n=0}^{\infty} \frac{B(\frac{\nu+1}{2},n+1)}{B(\nu-1,n+1)}\ B(n+\nu,\nu-1)\ I_{\frac{1+\omega}{2}}\ (n+\nu,\nu-1)\Big]$$

Then from (22)

$$E[(\hat{\Phi}(-\frac{\Delta}{2}))^2] = \int_{-\infty}^{\infty} F(-t/\nu\sqrt{[4-(\frac{1}{M_1}+\frac{1}{M_2})]})\quad g(t;\Delta,\nu)\ dt \tag{23}$$

where $g(t;\Delta,\nu)$ is a non-central density function involving the non-centrality parameter $\Delta$ and $\nu$ degrees of freedom. As it is somewhat difficult to evaluate the right side exactly, it can be easily computed numerically for a given value of $\Delta$ and $\nu$.

At present we are seeking an estimate of $\text{var}(\hat{e}_{ij})$ so as to be able to evaluate the standard errors $s_{ij}$, $i=1,2$ and $j=1,2,\ldots,k$. For that purpose $E[(\hat{\Phi}(-\frac{\Delta}{2}))^2]$ can be estimated by evaluating the right side of (23) numerically after considering $\Delta$ and $\nu$ given by $\hat{\Delta}$ and $(M_1+M_2-2)$ respectively; and similarly $\hat{\Phi}(-\frac{\Delta}{2})$ can be used replacing $E(\hat{\Phi}(-\frac{\Delta}{2}))$. Further, replacing $p_{ij}$ by its estimate $\hat{p}_{ij}$, one thus obtains the s.e. $s_{ij}$, i=1,2 and j=1,2,...k, and then from (16) and (17) the sample size $(n_1,n_2)$ is determined after replacing $\hat{s}_0$ by the estimate of Var $(\hat{\Phi}(-\frac{\Delta}{2}))$.

## 6. Remote Sensing Application

The previous discussion on estimation and determination of sample size though treated in general has primarily been motivated by our need of finding desirable estimates for the expected proportion of various types of vegetation/flora over a certain region covered by different types of timber using remote sensing techniques. But the analogy seems to exist in many other cases dealing with multispectral sensor data because it is not unlikely for different types of objects to be within the instantaneous field of views of a multispectral scanning device. Hence, the above discussion can be applied to ascertain the contribution of each type of objects making up a resolution element that gives rise to an observation obtained by a remote sensor. The analogy may briefly be outlined as follows:

Without loss of generality let there be two classes of resolution elements. The measurements on these resolution elements in each class are supposed to be normally distributed and on the basis of a measurement the resolution element may be misclassified. Further, let there be k different categories of objects that might be associated with the resolution elements in each class. With this set-up one can now obtain estimates of the expected proportions of the specified categories of objects from (2) or (7) depending upon the knowledge about the underlying normal distributions. Furthermore, by considering the number of objects selected according to (13) and (14) or (16) and (17) as the case may be, one can actually obtain the desirable estimates which approximately allows a specified amount of error about the true expected proportions for the object types with a desired amount of probability.

## References

[1] Anderson, T. W. (1958), An Introduction to Multivariate Statistical Analysis, John Wiley and Sons

[2] Cochran, W. (1968),"Errors of Measurements in Statistics," Technometrics, Vol. 10, No. 4, pp. 637-666

[3] Ellison, B. E. (1964), "Two Theorems for Inferences about the Normal Distributions with Applications in Acceptance Sampling," JASA, pp. 89-95

[4] Miller, R. G., Jr. (1966) Simultaneous Statistical Inference, McGraw-Hill

# ON ADJUSTING REMOTE SENSING DATA USING A RADIATION TRANSFER MODEL[1]

P. L. Odell and T. L. Boullion

Texas Tech University

[1] Research was supported in part under NASA Contract #NAS 9-12775.

# ABSTRACT

A computing technique for adjusting remote sensing spectral data for environmental effects is formulated. The technique is essentially invariant with respect to the atmospheric model used in the paper; hence, it can be replaced with a better model.

# ON ADJUSTING REMOTE SENSING DATA USING A RADIATION TRANSFER MODEL

## I. Introduction

In a recent report [1] a radiation-transfer model was developed to predict the apparent radiance, L, of a ground target, as it is observed by a multispectral scanner

$$L = \frac{\rho}{\pi} ET + L_p \;, \qquad (1)$$

where $\rho$ = the diffuse reflectance of the target material

$E$ = the irradiance at the target

$T$ = the atmospheric transmittance from the target to the multispectral scanner,

$L_p$ = the path radiance.

The following parameters are used to describe the conditions of observation:

1. $\tau(h)$ = optical depth of atmosphere at altitude h of the sensor.
2. $\mu$ = cosine of zenith or nadir angle.
3. $(\phi - \phi_0)$ = angle between scan direction and sun's azimuth.
4. $\theta$ = nadir scan angle.
5. $\theta_0$ = zenith angle of sun.
6. V = visibility (visual range) of the atmosphere, or some

better estimate of the distribution of haze and scattering particles present.

7. $\rho_{avg}$ = average diffuse reflectance of the scene that contributes to the path radiance scattered into the sensor's field of view at wavelength $\lambda$.

8. $\eta$ = anisotropy parameter.

9. $\tau_0$ = total optical depth.

10. $E_0$ = Solar irradiance on a surface where normal lies in the direction given by $\mu_0$ and $\phi_0$.

In order to show the functional dependencies involved in equation (1), it can be rewritten as

$$L[\rho,\theta,(\phi - \phi_0),\tau(h),\theta_0,V,\rho_{avg},\lambda]$$

$$= \frac{\rho(m,\lambda)}{\pi} E[\tau(h),\theta_0,V,\rho_{avg},\lambda]\ T[\theta,\tau(h),V,\lambda]$$

$$+ L_p[\theta,(\phi - \phi_0),\tau(h),\theta_0,V,\rho_{avg},\lambda].$$

## II. Main Content

### 2.1 When Observations are known to be generated by same target.

We want to consider the radiation-transfer model statistically as a covariance analysis model with covariates given by 1 through 10 and obtain the best estimate for E, the irradiance at the target, once the observations L have been adjusted for the atmospheric conditions. For ease of presentation, the covariates will be denoted by $\beta_1,\ldots,\beta_{10}$.

The model in equation (1) can be written as

$$L_i = g_i(\beta_1,\ldots,\beta_{10};E) + e_i, \ i = 1,2,\ldots,N \tag{2}$$

where the $e_i$ are non-observable, uncorrelated random errors each with mean zero and variance $\sigma^2$.

The problem is to obtain estimates of $\beta_1,\ldots,\beta_{10}$ which minimize

$$Q = [L - G]^T [L - G]$$

where $L = (L_1,\ldots,L_N)^T$ and $G = [g_1,\ldots,g_N]^T$.

Direct search techniques [2] have been used with success. One simply searches judiciously for the value of $\beta = (\beta_1,\ldots,\beta_{10})^T = \beta_{LS}$ which minimizes Q, by approximating $\beta_{LS}$ with an a priori value, say $\beta_k$, and then iteratively computing $\beta_{k+1}$ so that Q approaches a minimum.

Hartley [3] suggests a method in which he solves for $\beta_{LS}$ by solving the non-linear system of equations $\partial Q/\partial\beta_i = 0$ which is a necessary condition for Q to be minimal. By selecting $\beta_k$ a priori and letting G be approximately

$$G(L;\beta) \cong G(L;\beta_k) + \frac{\partial G}{\partial \beta}\bigg|_{\beta=\beta_k} (\beta_{k+1} - \beta_k)$$

one can solve for $\beta_{k+1}$, and on iterating the sequence $\{\beta_k\}$

converges in many cases to $\beta_{LS}$.

Walling [4] and Nelson [5] have developed techniques which take advantage of the linearity in the non-linear function G. Instead of approximating G* in a truncated Taylor's expansion, they approximate G by

$$G(L;\beta) \cong G(L;\beta_k) + \frac{\partial G}{\partial \beta}\bigg|_{\beta_k} (\beta_{k+1} - \beta_k)$$

$$\frac{\partial G(L;\beta)}{\partial \beta} \cong \frac{\partial G}{\partial \beta}\bigg|_{\beta=\beta_k}$$

$$\theta(\beta) \cong \theta(\beta_k)$$

where $\theta(\beta) = (\theta_1(\beta), \theta_2(\beta), \ldots, \theta_r(\beta))^T$ is a (non-random) parameter vector and where $G^* = A\theta(\beta_k) + C(L;\beta)$, and A is a known matrix of constants.

Then an a priori estimate $\beta_k$ leads to $\beta_{k+1}$ which when iterated gives a sequence $\{\beta_k\}$ which converges in many cases more rapidly to $\beta_{LS}$.

Comparison of Hartley's technique, Walling's technique, and Nelson's technique can be found in [5] and [6].

## 2.2 <u>When observations are not known to be generated by the same target</u>.

In the previous section a technique for estimating $\beta_i$, i = 1,2,...,10 and $E_i$ was briefly described. If $\hat{E}$ is the estimate for E the irradiance of the target, then one can use this value to perform discriminate analysis. However, this is indeed a special case and is not a realistic analysis for the remote sensing application. In most cases one is almost never sure if the $i^{th}$ and $j^{th}$ observations, $X_i$ and $X_j$, are from the same class until after the discriminate task has been performed. The symbol E must be replaced with the $E_i$ in the model described in (2), that is

$$L_i = g_i(\beta_1,\beta_2,\ldots,\beta_{10};E_i) + e_i \qquad i = 1,2,\ldots,N. \tag{3}$$

Note that in (3), there are N equations and N + 10 unknowns, $\beta_1,\ldots,\beta_{10},E_1,\ldots,E_N$, an undetermined system. The estimates for $E_1,E_2,\ldots,E_N$ are the values we seek to base our discriminate task upon, since the values of these estimates would be void of any modeled environmental effects. That is

$$\hat{E}_i = h_i(\hat{\beta}_1,\hat{\beta}_2,\ldots,\hat{\beta}_{10},L_i) \tag{4}$$

One can solve for $\hat{E}_i$ if values of $\hat{\beta}_1,\hat{\beta}_2,\ldots,\hat{\beta}_{10}$ and $L_i$ are available. Hopefully, this is the case in the remote sensing application.

However, in the case in which some values from the set $\{\hat{\beta}_1,\hat{\beta}_2,\ldots,\hat{\beta}_{10}\}$ are unknown, one can iteratively estimate $E_i$'s. This can be done by the following scheme:

(a) Discriminate using the non-adjusted measurement $L_i$ as $E_i$.

(b) Collect those resolution cells I(x) such that $x \in R_j$, then use these elements to estimate $\beta_1,\beta_2,\ldots,\beta_{10}$ and $E = E(\pi_j)$ as in section 2.1.

Since if $x \varepsilon R_j$, one assumes that $E_i = E(\pi_j) = E$, the radiance of an individual from $\pi_j$. The symbol $N_j$ denotes the number of elements I(x) assigned to $\pi_j$.

(c) for each j there exists a set

$$\hat{\beta}_1(j),\hat{\beta}_2(j),\ldots,\hat{\beta}_{10}(j) \quad j = 1,2,\ldots,m.$$

These are combined to get a better estimate

$$\hat{\beta}_i = \sum_{j=1}^{m} a_j \beta(j)$$

such that

$$a_j = N_j / \sum_{i=1}^{m} N_i$$

(d) Using $\hat{\beta}_i$, $i = 1,2,\ldots,10$, compute $\hat{E}_i$ in the equation

$$L_i = g(\hat{\beta}_1,\ldots,\hat{\beta}_{10};\hat{E}_i)$$

and use $\hat{E}_i$ to perform the discrimination.

## III. Concluding Remarks

It is important to note that our purpose here is to determine any problem areas in adjusting data for environmental factors and formulate a computation scheme to perform the adjustment and not select, evaluate, formulate, or modify an environmental or atmospheric model. Those whose expertise covers the topic of modeling an atmosphere should select the "best" model. The computation procedure suggested here is essentially model invariant.

## IV. References

1. Malila, W. A., Crane, R. B., Omarzu, C. A., and Turner, R. E. Studies of Spectral Discrimination, Willow Run Laboratories, May, 1971, NASA CR-WRL 31650-22-T.

2. Wilde, D. J., Optimum Seeking Methods, Prentice-Hall, Englewood Cliffs, New Jersey (1964).

3. Hartley, H. O., "The Modified Gauss-Newton Method for the fitting of Non-Linear Regression Functions by Least Squares", Technometrics, 3, No. 2 (1961), pp. 269-280.

4. Walling, D. D., "Non-Linear Least Squares Curve Fitting When Some Parameters are Linear", Texas J. of Science 20, #2 (1968), pp. 119-123.

5. Nelson, D. L., "Numerical Methods for the Solution of Non-Linear Least Squares Problems", Masters Thesis, Texas Tech University, Lubbock, Texas (1969).

6. Lewis, T. O. and Odell, P. L., Estimation in Linear Models, Prentice Hall (1971).

7. Reeser, W. K., "A Feasibility Study in the Use of a Sun Photometer in Gathering Aerosol Optical Depth Data for PREPS," Lockheed Electronics Co. Technical Report LEC/HASD No. 640-TR-086 March, 1972.

8. __________, "Aerosol Size Distribution Detection for PREPS Using Simple Processing Techniques" Lockheed Electronics Co. Report LEC/HASD 640-TR-091, March, 1972.

# ON ESTIMATING THE PROBABILITY OF MISCLASSIFICATION[1]

by

B. S. Duran, H. L. Gray, J. Tubbs, and T. L. Boullion
Texas Tech University

## 1. Introduction

The problem of discrimination has long been a problem of intense interest (see for example [2], [9]). Given m populations $\pi_1, \pi_2, \ldots, \pi_m$, the problem is that of classifying a p×1 vector observation or an individual I(x) corresponding to x, as belonging to one of the m populations. Anderson ([2], chapter 6) discusses the classification problem when the populations $\pi_1, \pi_2, \ldots, \pi_m$, are multivariate normal with equal covariance matrices. In the classification problem when m = 2 and $\Sigma_1 = \Sigma_2 = \Sigma$, the discriminant function is given by

$$(1) \quad U = x^T\Sigma^{-1}(\mu^{(1)} - \mu^{(2)}) - \frac{1}{2}(\mu^{(1)} + \mu^{(2)})\Sigma^{-1}(\mu^{(1)} - \mu^{(2)}).$$

However, if $\mu^{(1)}, \mu^{(2)}$, and $\Sigma$ are unknown then a reasonable discriminant function to use is

$$(2) \quad V = x^TS^{-1}(\bar{x}^{(1)} - \bar{x}^{(2)}) - \frac{1}{2}(\bar{x}^{(1)} + \bar{x}^{(2)})S^{-1}(\bar{x}^{(1)} - \bar{x}^{(2)})$$

where $\bar{x}^{(1)} = \sum_{j=1}^{n_1} x_j^{(1)}/n_1$, $\bar{x}^{(2)} = \sum_{j=1}^{n_2} x_j^{(2)}/n_2$, and

$$(n_1 + n_2 - 2)S = \sum_{j=1}^{n_1} (x_j^{(1)} - \bar{x}^{(1)})(x_j^{(1)} - \bar{x}^{(1)})^T$$

$$+ \sum_{j=1}^{n_2} (x_j^{(2)} - \bar{x}^{(2)})(x_j^{(2)} - \bar{x}^{(2)})^T.$$

[1] This research was partially supported by NASA Manned Spacecraft Center under Contract NAS 9-12775.

The distribution of U is the normal distribution $N(\frac{1}{2}\alpha,\alpha)$ if x is distributed $N(\mu^{(1)},\Sigma)$ or $N(-\frac{1}{2}\alpha,\alpha)$ if x is distributed $N(\mu^{(2)},\Sigma)$, where $\alpha = (\mu^{(1)} - \mu^{(2)})\Sigma^{-1}(\mu^{(1)} - \mu^{(2)})$. The discriminant function (2) is a special case of a class of statistics considered by Wald [10]. Further work concerning the distribution of V has been done by Anderson [1], Sitgreaves [8], and Kabe [6].

Wald [10] actually considered a class of statistics of which the statistic $x^T S^{-1}(\bar{x}^{(1)} - \bar{x}^{(2)})$ is a special case. For large values of $n_1$ and $n_2$ this appears to be a reasonable statistic for classification purposes. The distribution of the above statistic is given by Wald in terms of three quantities, say, $m_1$, $m_2$, $m_3$, and an expected value which he does not evaluate.

According to Sitgreaves [8] the statistic V may be written as

$$V = ay_1^T A^{-1} y_2 + by_2^T A^{-1} y_2$$

where a and b are known scalars, $y_1$ and $y_2$ are p-dimensional normal variates with $E(y_1) = \xi_1$ and $E(y_2) = \xi_2$, and A is a p×p symmetric matrix having a Wishart distribution with $n = n_1+n_2$ degrees of freedom. Furthermore $y_1$, $y_2$ , and A are independently distributed with common covariance matrix $\Sigma$.

Anderson [1] obtained the distribution of $m_1$, $m_2$, $m_3$ explicitly, including the evaluation of the expected value in Wald's result, for the special case when $\xi_1$ is proportional to $\xi_2$.

Sitgreaves [8] gave the distribution of $m_1$, $m_2$, $m_3$ when $\xi_1$ is proportional to $\xi_2$ and included the normalizing constant of the distribution which was not obtained by either Wald or Anderson.

Kabe [6] obtained a further extension by finding the distribution of $m_1$, $m_2$, $m_3$ without assuming the proportionality of $\xi_1$ and $\xi_2$. However, his result is not in closed form and appears to be very awkward to work with.

Of primary concern in the classification problem is the probability $P(i|j)$, $i \neq j$, of misclassifying an individual $I(x)$ from population j in population i. The complexity of the distribution of V makes it virtually impossible to compute $P(i|j)$. An available option is to evaluate an estimate, $\hat{P}(i|j)$, by the Monte Carlo technique. This is the topic that concerns us in this paper.

## 2. Description of the method

Suppose $x_1^{(1)}, x_2^{(1)}, \ldots, x_{n_1}^{(1)}$ and $x_1^{(2)}, x_2^{(2)}, \ldots, x_{n_2}^{(2)}$ denote two independent samples from two normal populations $\pi_1$ and $\pi_2$ with mean vectors $\mu^{(1)}$ and $\mu^{(2)}$, respectively, and common covariance matrix $\Sigma$. If $n_1 = n_2$ then the distribution of V if x is from $\pi_1$ is the same as the distribution of -V if x is from $\pi_2$ (see [2], p. 135). The statistic U also has this property. For large values of $n_1$ and $n_2$ the probability of classifying an observation from $\pi_2$ in $\pi_1$ is approximately

$$(3) \qquad \int_0^\infty \frac{1}{\sqrt{2\pi\alpha}} e^{-(x+\alpha/2)^2/2\alpha} dx$$

since x is asymptotically distributed $N(-\alpha/2,\alpha)$ when it comes from $\pi_2$. Similarly $P(2|1)$ is approximated by

$$\int_{-\infty}^{0} \frac{1}{\sqrt{2\pi\alpha}} e^{-(x-\alpha/2)^2/2\alpha}\, dx. \tag{4}$$

Hence, if x is classified in $\pi_1$ when $V \geq 0$ and in $\pi_2$ otherwise, $P(2|1) = P(1|2)$.

The probability $P(2|1)$ can be estimated by Monte Carlo methods for small ( and moderately large) values of $n_1 = n_2 = n$. The process involves first sampling n training samples from each of two multivariate normal populations which differ only in the means, i.e., $\mu^{(1)} \neq \mu^{(2)}$. These samples are then used to compute $S^{-1}$, $\bar{x}^{(1)}$, and $\bar{x}^{(2)}$. A sample of size m is generated from $N(\mu^{(1)},\Sigma)$ and m values of V are calculated. The probability $P(2|1)$ can then be estimated by

$$\hat{P}(2|1) = k/m$$

where k is the number of values of V that are negative. This process is repeated r times and the average of these r values of $\hat{P}(2|1)$ is used as a final estimate of $P(2|1)$. It is of interest to examine this probability for various values of $n_1 = n_2$, $\alpha$, and p, keeping in mind that the large sample value of $P(2|1)$ is given by (4). Values of $\alpha$ are obtained by keeping $\Sigma$ fixed and varying the values of $\mu^{(1)}$ and $\mu^{(2)}$ (or $\mu^{(2)} - \mu^{(1)}$).

## 3. Monte Carlo Results

Estimates of $P(2|1)$ were generated for various values of $n = n_1 = n_2$, $\alpha$, and p, where $\alpha = (\mu^{(1)} - \mu^{(2)})\Sigma^{-1}(\mu^{(1)} - \mu^{(2)})$

reflects the separation of the two populations. The covariance matrices for $\pi_1$ and $\pi_2$ were $\Sigma_1 = \Sigma_2 = \Sigma = I$. (No loss of generality is incurred by using I for the common covariance matrix since there exists an orthogonal transformation followed by a linear non-singular transformation that yields I.)

The data was generated by means of the normal random generator described in [7]. The estimates of $P(2|1)$ where simulated for various choices of two populations by fixing the covariance matrix and varying $\alpha$. The values $\alpha$ considered were $\alpha = 1,2,3,4,5,10,12,20,25$. The training sample sizes considered were $n = 5,10,15,20,50,100$. For each choice of $\alpha$, n, and p, the values for m and r were 50 and 50 respectively. Each observation was classified into $\pi_1$ if $V \geq 0$ and into $\pi_2$, otherwise.

Table 1 gives the estimate $\hat{P}(2|1)$ as a function of n and $\alpha$ for various values of p. Note that for every $\alpha$, $\hat{P}(2|1) \doteq P(2|1)$ when $n = 50$ or $100$, where $P(2|1)$ is the asymptotic probability of misclassification given by (4).

Figures I-IV reflect the fact that $P(2|1)$ is a decreasing function of $\alpha$ (for fixed n and p). Figures V-VII reflect the fact that for fixed $\alpha$ and p the probability $P(2|1)$ is a decreasing function of n. Also for fixed n, indications are that $P(2|1)$ is an increasing function of p.

The value of $\alpha$ is of primary concern since in a given situation the values of p and n are generally known. In an actual situation such as in a remote sensing application [5] it may be desirable to know approximately how many training samples are necessary so that classification based on these training samples will incur an error of misclassification not exceeding certain preassigned bounds. In a remote sensing application one could estimate $\alpha$ from the data and thus get an estimate of $P(2|1)$ from Table 1.

The quantity p is also of importance. Its significance in relation to $P(2|1)$ has already been observed in Table 1 and Figures V-VII. In agricultural applications of remote sensing data analysis a popular value appears to be $p = 4$, [3], [4].

## TABLE 1

Estimated Probabilities of Misclassification, $\hat{P}(2|1)$, for Values of $\alpha$, n and p.

| | | n | | | | | | |
|---|---|---|---|---|---|---|---|---|
| α | p | 5 | 10 | 15 | 20 | 50 | 100 | ∞ |
| 1.0 | 3 | 0.4192 | 0.3672 | 0.3568 | 0.3432 | 0.3252 | 0.3110 | .3085 |
| | 6 | 0.3960 | 0.3824 | 0.3632 | 0.3520 | 0.3200 | 0.3208 | |
| | 9 | 0.4176 | 0.4052 | 0.3884 | 0.3556 | 0.3350 | 0.3150 | |
| | 12 | | 0.4084 | 0.3775 | 0.3960 | 0.3524 | 0.3290 | |
| 2.0 | 3 | 0.3292 | 0.2768 | 0.2724 | 0.2600 | 0.2552 | 0.2460 | .238 |
| | 6 | 0.3416 | 0.3224 | 0.2940 | 0.2796 | 0.2444 | 0.2337 | |
| | 9 | 0.3324 | 0.3304 | 0.3176 | 0.3000 | 0.2572 | 0.2670 | |
| | 12 | | 0.3584 | 0.3224 | 0.3392 | 0.2820 | 0.2730 | |
| 3.0 | 3 | 0.2732 | 0.2336 | 0.2344 | 0.2116 | 0.2040 | 0.1875 | .192 |
| | 6 | 0.2792 | 0.2736 | 0.2428 | 0.2300 | 0.2072 | 0.2050 | |
| | 9 | 0.2932 | 0.2888 | 0.2596 | 0.2464 | 0.2072 | 0.2020 | |
| | 12 | | 0.3088 | 0.2924 | 0.2884 | 0.2268 | 0.2050 | |
| 4.0 | 3 | 0.2396 | 0.2004 | 0.1868 | 0.1856 | 0.1724 | 0.1620 | .158 |
| | 6 | 0.2452 | 0.2356 | 0.2084 | 0.1828 | 0.1672 | 0.2048 | |
| | 9 | 0.2700 | 0.2632 | 0.2352 | 0.2284 | 0.1728' | 0.1970 | |
| | 12 | | 0.2780 | 0.2490 | 0.2552 | 0.1916 | 0.1560 | |
| 6.0 | 3 | 0.1872 | 0.1472 | 0.1364 | 0.1228 | 0.1268 | 0.1102 | .111 |
| | 6 | 0.2276 | 0.1652 | 0.1520 | 0.1440 | 0.1080 | 0.1522 | |
| | 9 | 0.2352 | 0.2156 | 0.1824 | 0.1572 | 0.1224 | 0.1290 | |
| | 12 | | 0.2260 | 0.1884 | 0.1844 | 0.1328 | 0.1040 | |
| 10.0 | 3 | 0.1288 | 0.0804 | 0.0744 | 0.0728 | 0.0728 | 0.0572 | .057 |
| | 6 | 0.1216 | 0.1120 | 0.0864 | 0.0716 | 0.0604 | 0.0770 | |
| | 9 | 0.1540 | 0.1420 | 0.1104 | 0.1008 | 0.0644 | 0.0630 | |
| | 12 | | 0.1752 | 0.1352 | 0.1100 | 0.0748 | 0.0600 | |
| 12.0 | 3 | 0.0868 | 0.0556 | 0.0560 | 0.0460 | 0.0440 | 0.0414 | .041 |
| | 6 | 0.0884 | 0.0916 | 0.0664 | 0.0580 | 0.0464 | 0.0632 | |
| | 9 | 0.1368 | 0.1196 | 0.0844 | 0.0656 | 0.0584 | 0.0540 | |
| | 12 | | 0.1580 | 0.0968 | 0.0856 | 0.0516 | 0.0540 | |
| 20.0 | 3 | 0.0528 | 0.0212 | 0.0192 | 0.0188 | 0.0156 | 0.0124 | .012 |
| | 6 | 0.0460 | 0.0428 | 0.0268 | 0.0188 | 0.0196 | 0.0171 | |
| | 9 | 0.0768 | 0.0596 | 0.0332 | 0.0300 | 0.0176 | 0.0120 | |
| | 12 | | 0.0784 | 0.0308 | 0.0420 | 0.0156 | 0.0140 | |
| 25.0 | 3 | 0.0272 | 0.0140 | 0.0140 | 0.0112 | 0.0072 | 0.0054 | .006 |
| | 6 | 0.0336 | 0.0248 | 0.0132 | 0.0104 | 0.0064 | 0.0060 | |
| | 9 | 0.0588 | 0.0436 | 0.0232 | 0.0152 | 0.0076 | 0.0040 | |
| | 12 | | 0.0508 | 0.0244 | 0.0256 | 0.0100 | 0.0060 | |

Figure I

Probability of misclassification versus α when p = 3.

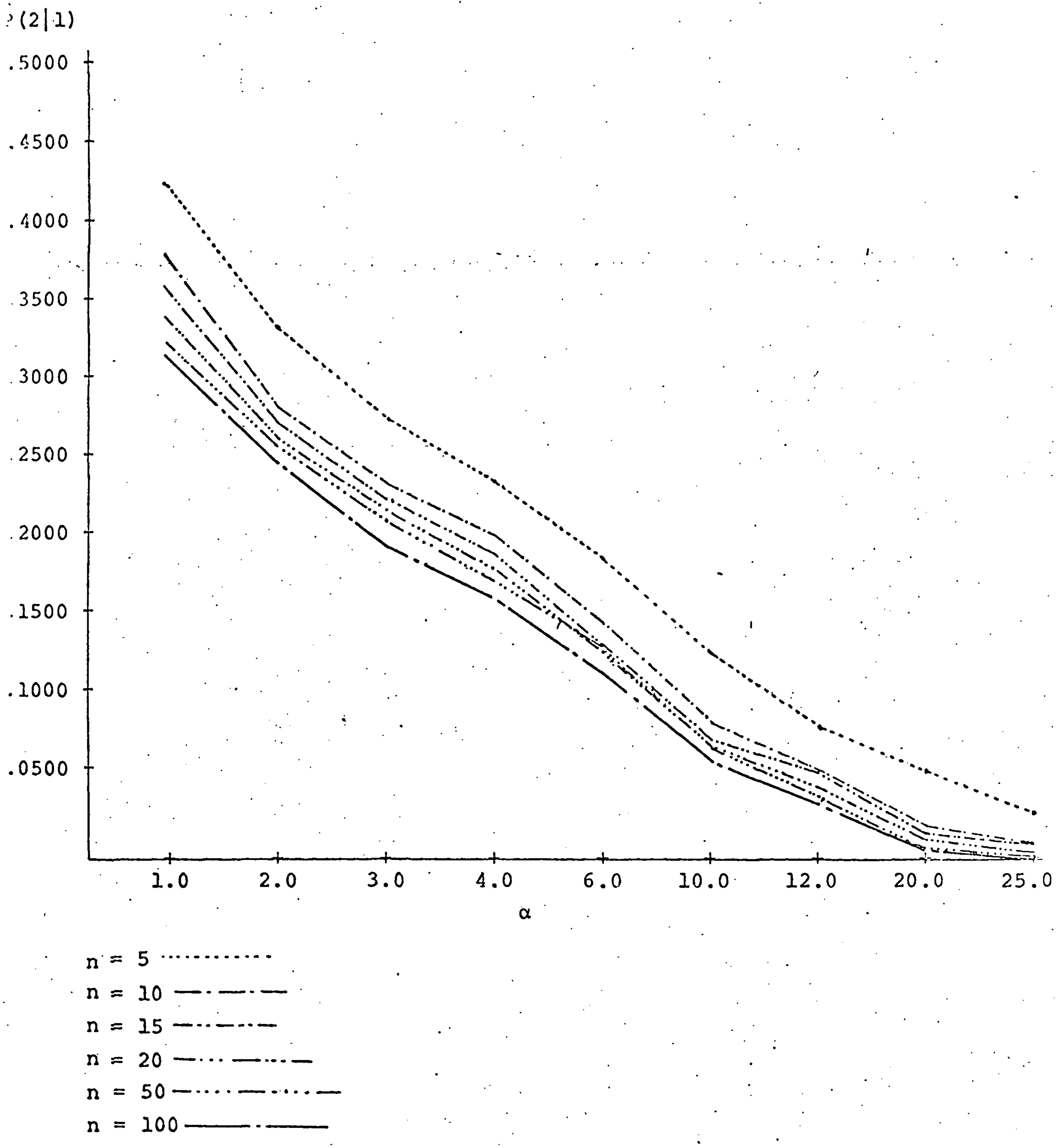


Figure II

Probability of misclassification versus α when p = 6.

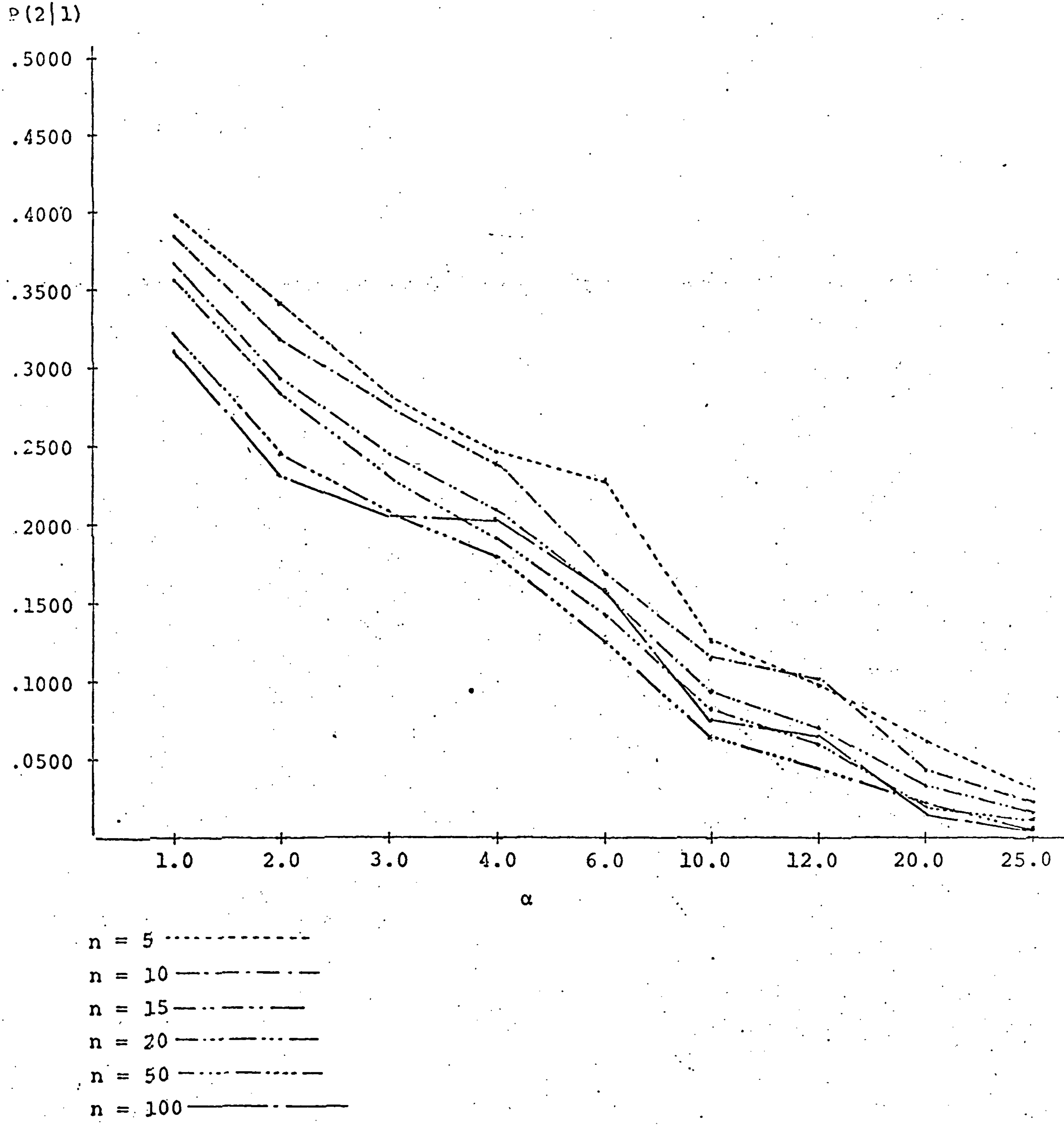


Figure III
Probability of misclassification versus α when p = 9.

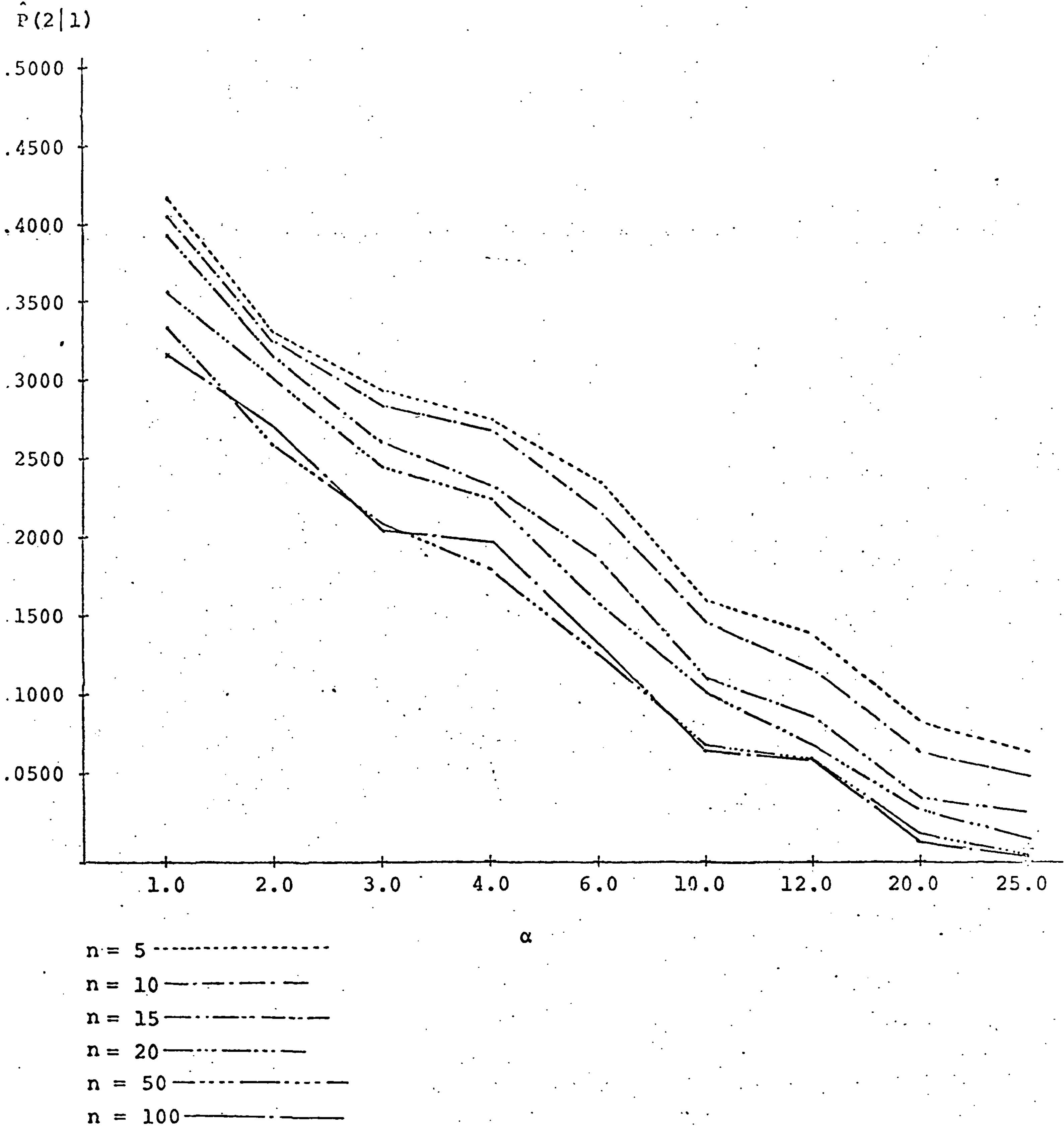


Figure IV

Probability of misclassification versus α when p = 12.

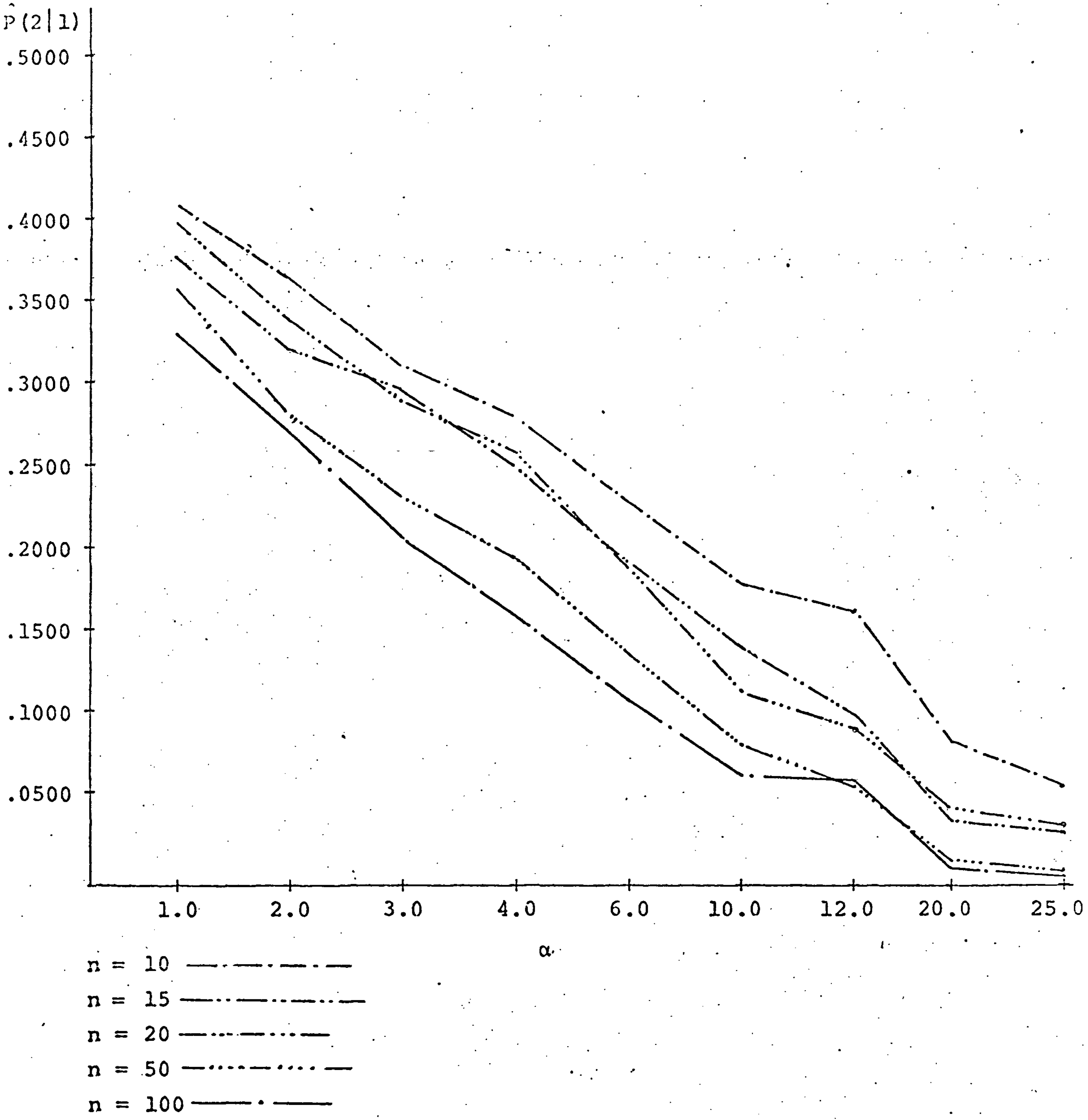


Figure V

Probability of misclassification versus n when α = 1.0.

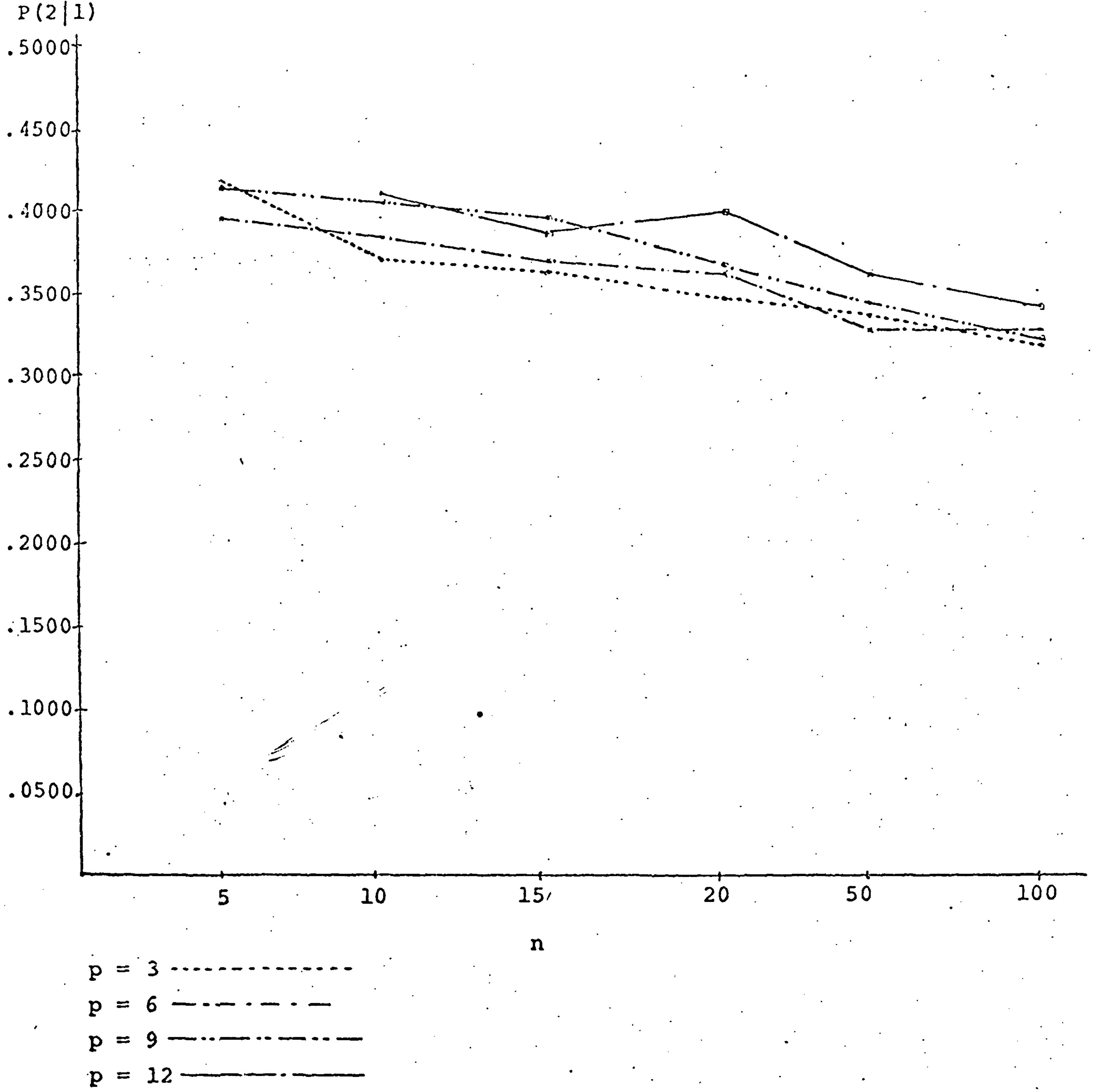


Figure VI

Probability of misclassification versus n when $\alpha = 6$.

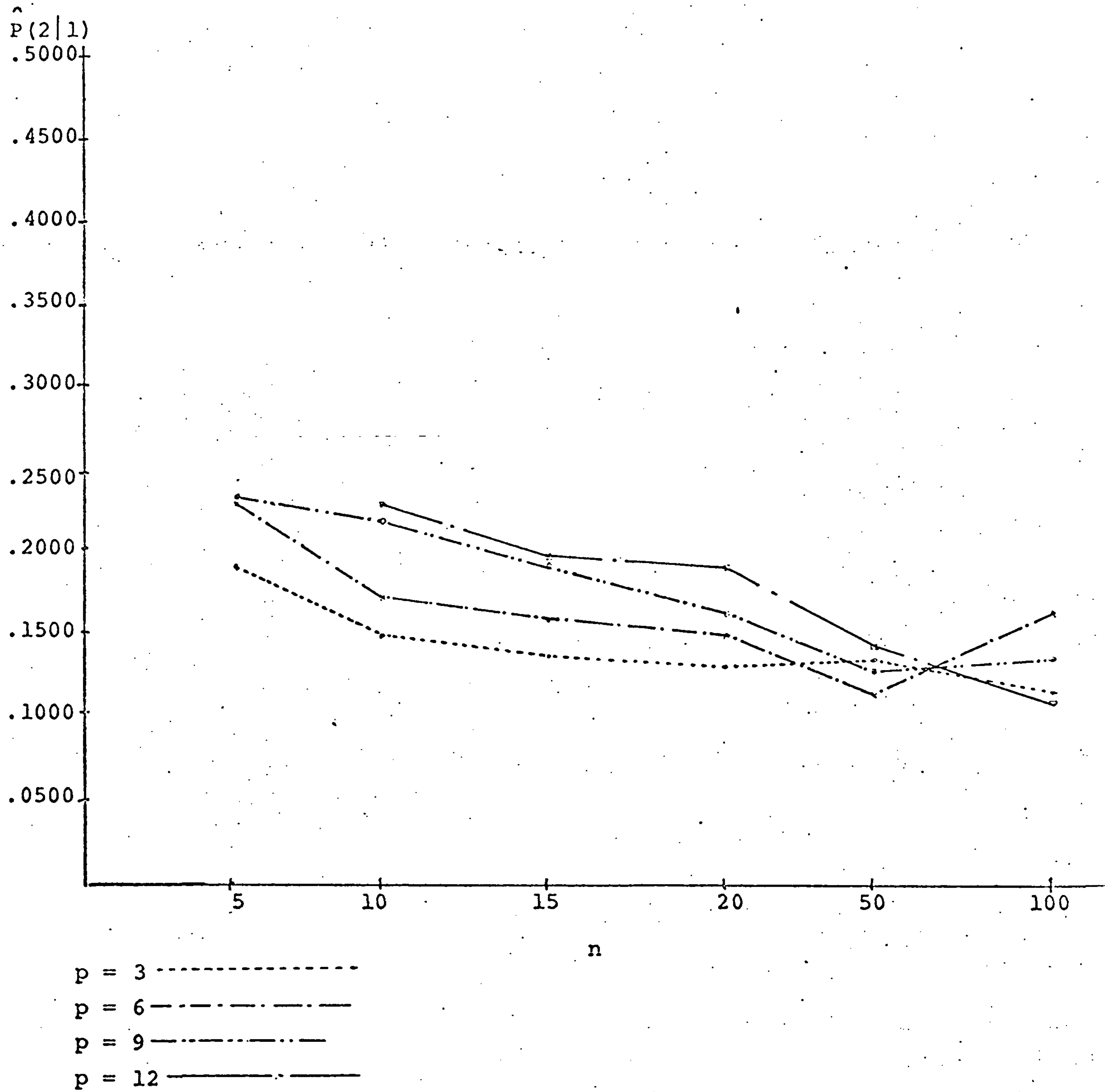
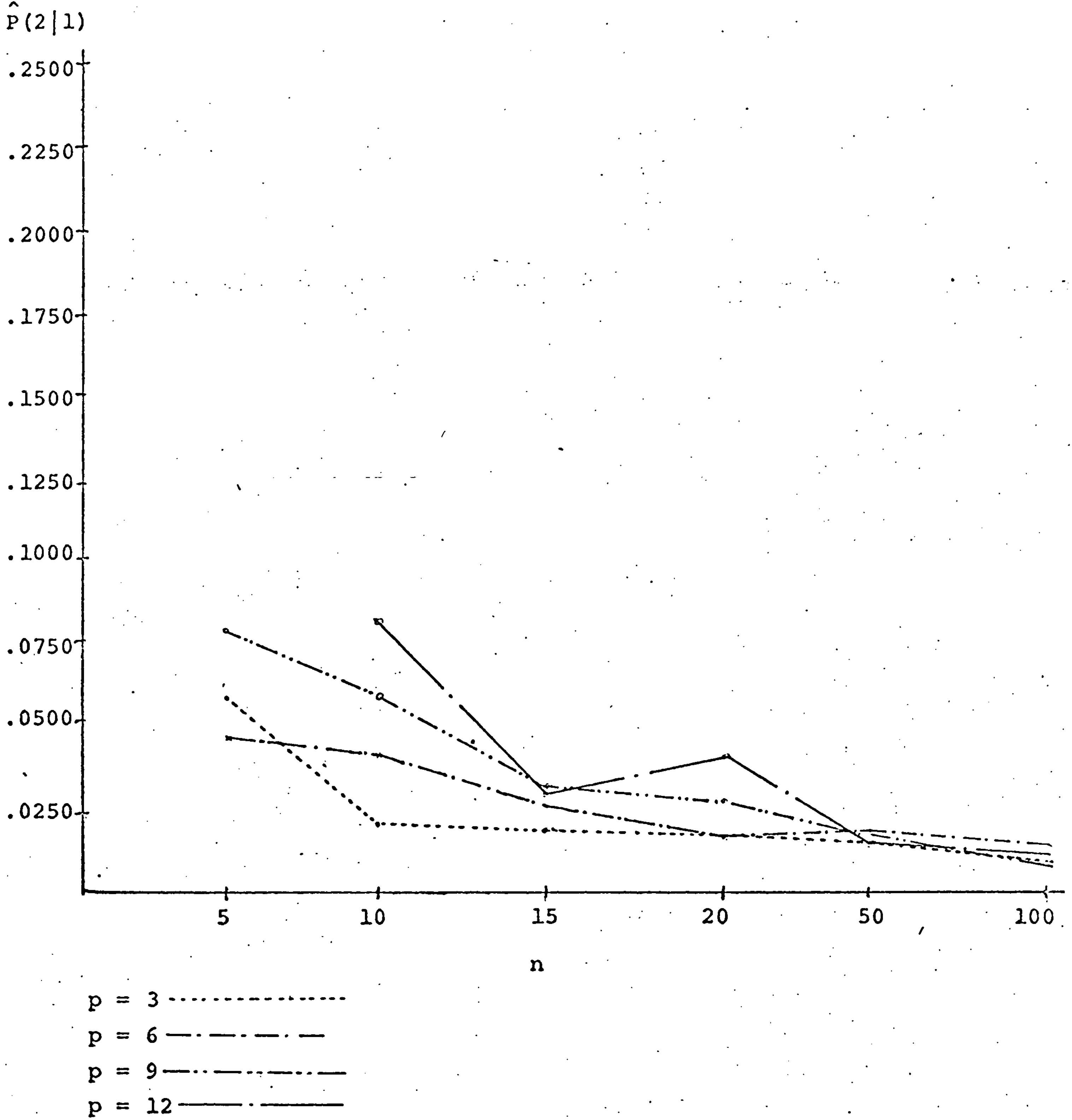


Figure VII

Probability of misclassification versus n when $\alpha$ = 20.



## REFERENCES

[1] Anderson, T. W., "Classification by Multivariate Analysis," Psychometrika, Vol. 16 (1951), pp. 31-50.

[2] ______________, An Introduction to Multivariate Statistical Analysis, John Wiley and Sons, Inc., New York (1958).

[3] Eppler, W. G., Helmke, C. A., and Evans, R. H., "Table look-up approach to pattern recognition," Proceedings of 7th International Symposium on Remote Sensing of the Environment, The University of Michigan, May, 1971.

[4] Fu, K. S., Landgrebe, D. A., and Phillips, T. L., "Information processing of remotely sensed agricultural data," Proc. IEEE, Vol. 57, No. 4 (April 1969), pp. 639-653.

[5] Huang, T., "Per field classifier for agricultural applications," LARS Information Note 060569, Purdue University, Lafayette, Indiana, June, 1969.

[6] Kabe, D. G., "Some Results on the Distribution of Two Random Matrices Used in Classification Procedures," Ann. Math. Statist., Vol. 33 (1962), pp. 181-185.

[7] Newman, T. G. and Odell, P. L., The Generation of Random Variates, Griffins Statistical Monographs and Courses, No. 29, (1971), pp. 37-44.

[8] Sitgreaves, R., "On the Distribution of Two Random Matrices Used in Classification Procedures," Ann Math. Statist., Vol. 23 (1952), pp. 263-270.

[9] Von Mises, R., "On the classification of observation data into distinct groups," Ann. Math. Statist., Vol. 16 (1945), pp. 68-73.

[10] Wald, A., "On a Statistical Problem Arising in the Classification of an Individual Into One of Two Groups," Ann. Math. Statist., Vol. 15 (1944), pp. 145-162.

# ON THE TABLE LOOK-UP IN DISCRIMINATE ANALYSIS

P. L. Odell[1], B. S. Duran[2], and W. A. Coberly[3]

Texas Tech University

1. Research supported in part by NASA Contract #NAS9-11925.

2. Research supported in part by NASA Contract #NAS9-12775.

3. Research supported in part by Texas State Organized Research Money.

## ABSTRACT

The discriminate analysis problem is discussed briefly. An analytic formulation of the so-called Eppler (Table Look-up) algorithm is given along with a modification which equates the algorithm with the classical Bayes procedure. Simulation results comparing several discriminate analysis techniques are given.

# ON THE TABLE LOOK-UP IN DISCRIMINATE ANALYSIS

P. L. Odell, B. S. Duran, and W. A. Coberly
Texas Tech University

## 1. Introduction

Consider m populations $\pi_1, \pi_2, \ldots, \pi_m$ and suppose each individual in the union of these populations possesses p common observable characteristics $c_1, c_2, \ldots, c_p$. The observed values of an individual are denoted by $x = (x_1, x_2, \ldots, x_p)^T$, where $x_j$ denotes the observed value of $c_j$. Let $p_1(x), p_2(x), \ldots, p_m(x)$ denote m known multivariate probability density functions of the p-dimensional observation vector x and $q_1, q_2, \ldots q_m$ be the known a priori probabilities that an individual, I, be selected from a population $\pi_1, \pi_2, \ldots, \pi_m$, respectively.

The classical discriminate analysis problem consists of formulating or developing a technique for assigning an individual selected at random from $\bigcup_{i=1}^{m} \pi_i$ into one of the m populations. There have been various techniques proposed for solving the problem, of which the Bayesian solution is optimal, in the sense that it minimizes the expected cost of misclassification.

In various applications of discriminate analysis, for example in the analysis of remote sensing data [1], the amount of computation involved is immense. Thus it seems desirable to either develop new techniques, modify existing ones, or to decrease the dimensions of the problem with the hope of maintaining approximately the optimality of the classical Bayes procedure. The dimensions of the problem can be decreased by means of characteristic selection [8] and/or data compression [12] techniques. These techniques allow for reducing the value of p. Since the number of populations, m, is not arbitrary the only parameter which can be reduced is p, the number of characteristics.

The characteristic selection procedure calls for selecting from the set of p characteristics q, $q \leq p$, characteristics, say $c_{i_1}, c_{i_2}, \ldots, c_{i_q}$, which are "best" with respect to identifying individuals from the populations $\pi_1, \pi_2, \ldots, \pi_m$. It is important to note that complete enumeration of all possible choices of characteristics is practically impossible since the number of ways one can select q characteristics for $1 \leq q \leq p$ is

$$2^p - 1 = \sum_{q=1}^{p} \frac{p!}{q!(p-q)!} \quad .$$

If one lets the compression matrix B be a $q \times p$ matrix with q ones in positions $(i_j, i_j)$, $j = 1, 2, \ldots, q$, then $Y = Bx$ is simply the vector $T = (x_{i_1}, x_{i_2}, \ldots, x_{i_q})^T$. Thus characteristic selection is a special case of data compression.

Wilks [12] discusses a special type of data compression whereby k + 1 p-dimensional samples are projected into a q-dimensional space, $q < p$, in such a manner that the k+1 projected samples are reasonably well separated. The projection is actually carried out so that the pooled-sample scatter is as large as possible relative to the within (total) - sample scatter. For example, one might desire to project three 3-dimensional sample points into a 2-dimensional or a 1-dimensional space. This is actually done in section 6 where various discrimination procedures are evaluated by Monte Carlo simulation.

A general approach in solving the discriminate problem is to define a distance between two populations, say D(i, j; c) where $c = (c_{i_1}, c_{i_2}, \ldots, c_{i_q})$, and then select a c such that the minimal distance between any two populations $\pi_i$ and $\pi_j$ is maximized. One such distance function is divergence [6], [9] defined by

$$D(i,j) = \int_{-\infty}^{\infty} [p_i(x) - p_j(x)] \ln [p_i(x) / p_j(x)] dx$$

which is of course an arbitrary choice for a distance function [5], but is being used in at lease one large computer program for reducing remote sensing data [7]. It is not well known [2], [3] just how distance or divergence is related to misclassification, except in the case when the covariance matrices are equal. However, one is compelled intuitively to believe that the expected cost of misclassification should decrease with increasing pairwise distances between populations.

Another technique, although developed heuristically [4], has proved successful in reducing the amount of computation involved in the solution of the discriminate problem. This technique, called the table look-up technique, is an approximation to the Bayes partition solution. In this paper we show that the table look-up technique can be modified so that it is a "closer" approximation to the Bayes procedure. The table look-up technique "trades off" floating point addition and multiplication for integer or fixed point addition in a table look up computer operations, thereby reducing the computing time from 2 units to 0.066 units in at least one empirical example [4].

In the classical Bayes procedure the probability density function $p_i(x)$ has to be evaluated for each observation vector x. The procedure discussed in this paper eliminates the need for computing $p_i(x)$ for each observation vector x. The table look-up technique also utilizes a different set of characteristics from the p characteristics, for testing the membership of an individual in different populations. This concept was developed by Eppler, Helmke, and Evans [4] and they have shown empirically that their version in the form of a computing algorithm leads to a significant decrease in computer time.

We now consider the analytic development of the table look-up technique.

## 2. Analytic Development of the Table Look-up Technique

Let q be chosen a priori and p be known. Let $c = (c_1, c_2, \ldots, c_q)^T$ denote the q characteristics selected from the larger set of p characteristics which maximize the minimal distance between each pair of populations $\pi_i$ and $\pi_j$, $i = 1, 2, \ldots, m$; $j = 1, 2, \ldots, m$, $i \neq j$, with respect to an a priori chosen distance function $D(i,j)$. Let $x_i$ denote the scalar measurement made on the characteristic $c_i$ such that

$$a_i \leq x_i \leq a_i + n_i d, \quad i = 1, 2, \ldots, q,$$

where $a_i$ is known and $x_i$ can take on only those values $a_i + jd$, $j = 0, 1, 2, \ldots, n_i$. For this choice of $x_i$'s the measurement space $S_q$ will contain $\prod_{i=1}^{q} (n_i + 1)$ points. The measurement space $S_q$, which is a set of lattice points consisting of all possible measurement points can be written as

$$(1) \qquad S_q = (a_1, a_1 + n_1 d) \otimes (a_2, a_2 + n_2 d) \otimes \cdots \otimes (a_q, a_q + n_q d).$$

In the case considered by Eppler, Helmke, and Evans, $a_i = 0$, $d = 1$, and $n_i = 255$ for all $i = 1, 2, \ldots, m$, motivated by the units and manner in which the $x_i$'s were measured. The number of points in $S_q$ when $a_i = 0$ and $n_i = 255$ for all $i$, is $256^q$, a very large number.

Consider the region

$$\hat{R}_i = \{x;\ p_i(x) = \max\ \{p_j(x)\} \text{ and } p_i(x) \geq \max\ (T_j)\}$$

where each $T_j$ is an arbitrarily chosen threshold value. One may make a normality assumption and select $T_i$ such that

$$(2) \qquad P\ \{(x-\mu_i)^T \textstyle\sum_i^{-1}(x-\mu_i) \leq C_\alpha \mid I(x)\ \varepsilon\ \Pi_i\} = 1-\alpha$$

where $1-\alpha$ is selected much as one would select a confidence coefficient in determining confidence intervals in statistical estimation. Statement (2) may be written equivalently as

$$P\ \{\frac{1}{(2\pi)^{p/2}|\sum_i|^{1/2}}\ e^{-\frac{1}{2}(x-\mu_i)^T \sum_i^{-1} (x-\mu_i)}$$

$$\geq \frac{1}{(2\pi)^{p/2}|\sum_i|^{1/2}}\ e^{-\frac{1}{2}C_\alpha} | I(x)\ \varepsilon\ \pi_i\} = 1-\alpha.$$

The threshold value $T_i$ is then given by

$$T_i = \frac{1}{(2\pi)^{p/2}|\sum_i|^{1/2}} e^{-\frac{1}{2} C_\alpha}.$$

Since $x \sim N(\mu_i, \sum_i)$, then $\chi^2_{(p)} = (x-\mu_i)^T \sum_i^{-1}(x-\mu_i)$ is distributed chi-square with p degrees of freedom. The value of $C_\alpha$ may simply be read from a $\chi^2$ table [10] from which the value $T_i$ may be computed.

Let $\hat{R}_0 = \{x;\ p_i(x) \leq T = \max_j \{T_j\},\ i = 1, 2,\ldots,m\}$ be the region of no decision, that is, the region in which the information contained in the measurement vector x gives very little or no discriminate information. The region $\hat{R}_0$ is not unlike the no decision region in classical statistical sequential testing [10]. If R is the p-dimensional space $S_p$, then

$$S_p = \hat{R}_0 \cup \hat{R}_1 \cup \cdots \hat{R}_m.$$

Let S(x;R) denote a storing transformation (storing operation) defined as follows

$$S:\ x \rightarrow i \text{ if } x \in \hat{R}_i.$$

The table look-up technique is based on pre-storing in fast random access core memory the prescribed region $\hat{R}_i$, as i, for all points in $S_p$. That is, every vector x defined by

$$\hat{R}_i = \{x;\ (x-\mu_i)^T \textstyle\sum_i^{-1} (x-\mu_i) \leq c_\alpha\}$$

and whose components $x_j \in \{a_j, a_j + d,\ldots,a_j + n_j d\}$, $j = 1, 2,\ldots,p$

is placed in a table with x corresponding to i. The value of i is stored in the "x" location. Thus when x is measured, its location is looked up and if a value of i is found then I(x) is classified into population $\pi_i$. The table look-up technique replaces the calculation in the classical discriminate technique with a retrieval operation for each observation x. The savings in time is then the difference in time to retrieve the population classification and calculation and ordering in the classical technique.

It is important to note that except for the introduction of the region of no decision $\hat{R}_0$, the table look-up technique is a new (different) computational technique for performing the Bayes Algorithm. By selecting $T = \max\{T_i\}$ sufficiently small, $R_0 = \emptyset$, the empty set, and the table look-up technique is simply a clever way to perform the Bayes Algorithm. This last statement is substantiated in the next section.

## 3. Comparison of the Table Look-up Technique with Bayes Algorithm

Let the p-dimensional Euclidean space R be partitioned into the Bayesian discriminate partition $R = (R_1, R_2, \ldots, R_m)$, where $R_k$ is defined

$$R_k = \{x;\ r_k = \min_{1 \leq i \leq m} \{r_i\}\ \}$$

where

$$r_i = \sum_{\substack{j=1 \\ i \neq j}}^{m} q_j p_j(x)\ C(i|j),$$

and $C(i|j)$ denotes the cost of classifying an observation from $\pi_j$ as coming from $\pi_i$. The probability of proper classification is given by

$$P(i|i;\ R) = \int_{R_i} p_i(x)\ dx.$$

Let $\hat{R}_i$ be a subset of $R_i$, that is $\hat{R}_i \subset R_i$, $i = 1, 2,\ldots m$, and define

$$\hat{R}_0 = R - \bigcup_{i=1}^{m} \hat{R}_i,$$

to be the no decision region. Then

$$P(i|i,\ \hat{R}) = \int_{\hat{R}_i} p_i(x)\ dx \leq \int_{R_i} p_i(x)\ dx = P(i|i,\ R)$$

and the probability of not making a decision is $P\ (x \in \hat{R}_0;\ \hat{R})$. Note that if $R = \hat{R}$, then $P\ (x \in \hat{R}_0;\ R) = 0$ since $\hat{R}_0 = \emptyset$, the empty set.

Let $S\ (x;\ R)$ be a storing operation such that

$$S\ (x;\ \hat{R}) : x \to i \quad \text{if} \quad x \in \hat{R}_i.$$

When an observation $x$ is taken on an individual $I(x)$ one searches through the storage for the range of $S\ (x;\ \hat{R})$ which is equivalent to determing the integer $i$ for which $x \in \hat{R}_i$. If $x \in \hat{R}_i$ or equivalently, if $i$ is stored in the "x" location in storage, then we assign $I(x)$ to population $\pi_i$.

Since there exists a continuum of x's in the interval $a \leq x \leq b$, the memory requirements are infinite. However, due to the manner in which the data is taken, x takes on only a finite number of vector values. That is, there exists only a finite number of values that each $x_i$ in the vector $x = (x_1, x_2, \ldots, x_p)$ can take on. The possible values for each $x_i$ are given by

$$x_i = a_i + jd;\ j = 0,\ 1,\ 2, \ldots, n_i.$$

In the remote sensing application for example, $a_i = 0$ for $i = 1, 2, \ldots, m$, $d = 1$, and $n_i = 255$, for all i. Hence, the number of storage locations required is $256^p$. This figure is very large indeed, even for small values of p. However, there are ways of reducing this number substantially. Comments on this item and other feasible and practical aspects of the Table look-up technique are discussed in section 7.

The foregoing results are summarized in the following theorem and corollary.

THEOREM: Let $R = (R_1, R_2, \ldots, R_m)$ be a Bayes partition and $\hat{R} = (\hat{R}_0, \hat{R}_1, \ldots, \hat{R}_m)$ be any other partition such that $\hat{R}_i \subset R_i$, $i = 1, 2, \ldots, m$. Then $P\ (i|i, \hat{R}) \leq P\ (i|i, R)$ if $C\ (i|j) = C$ for all $i \neq j$.

COROLLARY: Let $\hat{R}_0$ tend to the empty set and $\hat{R}_i$ tend to $R_i$. Then $\hat{R}$ tends to the Bayes partition R.

The problem in using the table look-up technique reduces to that of selecting the partition $\hat{R} = (\hat{R}_0, \hat{R}_1, \ldots, \hat{R}_m)$ which minimizes computer time and storage requirements but yet approximates the optimality of the Bayes partition sufficiently closely. Thus the problem is to select that partition $\hat{R}$ which maximizes computer efficiency.

| | TABLE LOOK-UP USING BEST 4 CHARACTERISTICS OUT OF NINE FOR EACH POPULATION | BAYES TECHNIQUE USING FOUR BEST CHARACTERISTICS OUT OF NINE FOR ALL POPULATIONS | BAYES TECHNIQUE USING SIX BEST CHARACTERISTICS OUT OF NINE FOR ALL POPULATIONS |
|---|---|---|---|
| TIME TO CLASSIFY A 222-SAMPLE LINE | 0.066 SEC | 2.0 SEC | 4.0 SEC |
| ACCURACY | CORRECT 92.4%<br>UNDECIDED 3.2%<br>INCORRECT 4.4% | CORRECT 93.1%<br>UNDECIDED 0.7%<br>INCORRECT 6.2% | CORRECT 95.0%<br>UNDECIDED 0.0%<br>INCORRECT 5.0% |
| ARITHMETIC OPERATIONS REQUIRED BY ALGORITHM | INTEGER ADDITION | FLOATING-POINT ADD AND MULTIPLY | FLOATING-POINT ADD AND MULTIPLY |

Table 1. Comparison Between Table Look-up and Bayes Approaches.

## 4. Modification of the Table Look-up Procedure.

There exists at least one competing algorithm to the table look-up algorithm. We consider one such algorithm which is suggested in an attempt to minimize storage requirements but yet retain the desirable properties of the table look-up technique. Let $q \leq p$ denote the number of characteristics to be used in a discriminate analysis. The following two alternatives are available in choosing q.

(1) Select the q "best" characteristics for the union of the m populations and perform a table look-up algorithm using measurements on these characteristics.

(2) Select the q "best" characteristics for each population and for each such choice perform a table look-up algorithm.

Eppler, Helmke, and Evans [4] have given some computational comparisons between the Bayes and Table look-up techniques [see Table 1]. They selected the four best characteristics from a set of nine using real data from the Purdue experiment and found that they were able to decrease computer time by a factor of 32 to 1. However, storage requirements apparently remained a problem so they introduced a scaling transformation to produce a coarse set of lattice points which they called "pointer scale". Arguments are given in [4] to assure us that little is lost by modifying the algorithm to include a coarse lattice. These arguments seem reasonable but one should remember that primitive (i.e. $R = \hat{R}$) table look-up implies relative large storage requirements.

As an alternative procedure one may consider the following modification of the table look-up procedure. Let $R = (R_1, R_2, \ldots, R_m)$ be the optimal Bayes partition and $R = (\hat{R}_0, \hat{R}_1, \ldots, \hat{R}_m)$ be any table look-up partition. Let

$$\hat{R}_0 = \hat{R}_{10} \cup \hat{R}_{20} \cup \cdots \cup \hat{R}_{m0}$$

where

$$\hat{R}_{i0} = \hat{R}_0 \cap R_i$$

is the intersection of the $i^{th}$ Bayes region $R_i$ and $\hat{R}_0$. If $R_i$ is such that $\hat{R}_i \subset R_i$ for all i, then

$$P\ (i|i;\ R) \geq P\ (i|i;\ \hat{R}).$$

Let us select as $\tilde{R}_i$ the largest p-dimensional rectangle in

$R_i$, with planar boundaries parallel to the coordinate planes, which contains as much of $\hat{R}_i$ as possible, including the center of $\hat{R}_i$. Then for $a_i \leq x \leq b_i$, $i = 1, 2, \ldots, m$, the probability of proper classification is

$$P\ (i|i,\ \tilde{R}) = \int_{a_{i1}}^{b_{i1}} \cdots \int_{a_{ip}}^{b_{ip}} p_i(x)\ dx$$

which one wishes to be such that the approximation error

$$P\ (i|i;\ R) - P\ (i|i;\ \tilde{R}) = e_i, \tag{3}$$

is small.

Now, since the planar sides (bounds) of $\tilde{R}_i$ are parallel to the coordinate planes of the p-dimensional space one needs only to find those bounds such that if $x \varepsilon \tilde{R}_i$ then $I(x)$ is assigned to $\pi_i$.

Let the bounds of $R_i$ be given as a $a_i \leq x \leq b_i$ for $i = 1, 2, \ldots, m$. Then a modification of the table look-up procedure which uses the rectangles $\tilde{R}_1, \tilde{R}_2, \ldots \tilde{R}_m$ as an approximation to the Bayes Partition may be summarized in the following algorithm.

Step 1. If the observation vector is such that $a_i \leq x \leq b_i$, then $I(x)$ is assigned to $\Pi_i$.

Step 2. If $x < a_i$ or $x > b_i$ then replace i with $i' \neq i$ and go to Step 1.

Step 3. Repeat the algorithm m times and if $x \not\varepsilon \tilde{R}_i$ for $i = 1, 2, \ldots, m$ then assign $I(x)$ to $\tilde{R}_0$, the no decision region.

This modification of the table look-up algorithm can also be employed by using the two alternatives in choosing the q "best" characteristics from the p characteristics.

If the errors $e_i$ in (3) are not sufficiently small, then the approximation could be improved by choosing two disjoint rectangles $\tilde{R}_{i1} = \tilde{R}_i$ and $\tilde{R}_{i2}$ such that $\tilde{R}_{i1} \cup \tilde{R}_{i2} = \bar{R}_i$ contains more of $\hat{R}_i$ than did $\tilde{R}_i$ and

$$R_i = \tilde{R}_{i0} \cup \tilde{R}_{i1} \cup \tilde{R}_{12}.$$

An algorithm similar to the one above would hold for the case of two rectangles $\tilde{R}_{i1}$ and $\tilde{R}_{i2}$. In fact, a union of p-dimensional rectangles could be used as an approximation to $R_i$ and the appropriate algorithm could be formulated. However, the amount of increase in classification accuracy might be so small as to not warrant such a venture.

The algorithm above places an individual I(x) in the no decision region if $x \not\in \tilde{R}_i$ for $i = 1, 2, \ldots, m$. The results of Table 1 indicate that the table look-up places 3.2% of the cases in the no decision region for that particular example. For any observation falling in the no decision region the Bayes procedure could be used. Thus all individuals would be classified and the procedure involved would be as optimal as the Bayes procedure and the computer time involved would be less than the time required for the Bayes procedure alone..

A last item to note about the modification for the table look-up technique is that in using the p-dimensional rectangle approach, the need to store values of i for each x is eliminated and replaced with an ordering procedure. That is if $a_i \leq x \leq b_i$ then $x \in \tilde{R}_i \subset R_i$ and I(x) is assigned to $\pi_i$.

## 5. Evaluation of Various Discriminate Techniques

In the application of discriminate analysis to large sets of data, such as in the remote sensing application, it is extremely important that one is able to select the "best" available procedure. The ideal situation would be to have an optimal procedure that can be performed in the least amount of time. However, this is never the case.

A methodology for ranking existing discriminate techniques is lacking; however, we will attempt to rank several techniques which have been mentioned and/or discussed in the previous sections. The suggested rankings involving these techniques will be obtained merely by how one would expect them to perform on the basis of the way they are defined. For example, a table look-up procedure takes less time to perform than a Bayes procedure; however, the Bayes procedure is more accurate. In the next section several of these techniques will be examined by means of Monte Carlo simulation. One can then check to see how those results bear out some of the results in this section.

The evaluations in this section are in reference to (1) accuracy, (2) computing speed, and (3) storage requirements. The discriminate techniques that will be considered are:

$T_1$ : A Bayes algorithm using data compressed by means of Wilks concepts [12]

$T_2$: A Table look-up technique using the same p characteristics for all populations $\pi_i$.

$T_3$: A Table look-up technique using the best q (q < p) characteristics for all populations. See (1) of section 4.

$T_4$: A Table look-up using the best q (q < p) characteristics for each population $\pi_j$. See (2) of section 4.

$T_5$: A p-dimensional rectangle approximation using the same p characteristics for all populations.

$T_6$: A q-dimensional rectangular approximation using the best q characteristics for all populations.

$T_7$: A q-dimensional rectangular approximation using the best q characteristics for each population $\pi_j$.

$T_8(j)$: Let $j = 2,\ldots,7$ and $T_8(j)$ denotes the $T_j$ algorithm with the modification that a classical Bayes procedure is performed if $x \varepsilon \hat{R}_0$.

$T_9$: Classical Bayes algorithm in which $p_i(x)$, $i = 1, 2,\ldots,m$ are known.

$T_{10}$: Classical Bayes algorithm in which $p_i(x)$, $i = 1, 2,\ldots,m$ are assumed normal with unknown parameters $\mu_i$ and $\sum_i$.

$T_{11}$: Classical Bayes algorithm in which $p_i(x)$, $i = 1, 2,\ldots,m$ are unknown and must be estimated "nonparametrically".

There are other techniques but we will restrict ourselves to these. The Bayes technique using the best q (q < p) characteristics for all populations is not included, however, it is compared with the Table look-up in Table 1. The Bayes solution is given by $T_9$, $T_8(2)$, and $T_8(5)$ when the probability density functions $p_i(x)$, $i=1,2,\ldots,m$ are known. Hence, it is meaningless to ask which is the more accurate. However, the difference in characteristic selection implies that $T_8(2) \geq T_8(3)$, $T_8(4) \geq T_8(3)$, $T_8(7) \geq T_8(6)$, and $T_8(5) \geq T_8(6)$ where the symbol "$\geq$" means "is as accurate as". If one can determine the fact that $\tilde{R}_i \subset \hat{R}_i$ then one can say that the table look-up technique is as accurate as the p-dimensional rectangular approximation technique. In this case $T_8(2) \geq T_8(5) \geq T_8(6)$

and $T_8(4) \geq T_8(7)$.

Other orderings with respect to accuracy are:

(a) $T_8(4) \geq T_4$,

(b) $T_2 \geq T_3$,

(c) $T(7) \geq T_7$,

(d) $T_5 \geq T_6$,

(e) $T_7 \geq T_6$,

(f) $T_8(6) \geq T_6$,

(g) $T_8(5) \geq T_5$,

(h) $T_8(3) \geq T_3$,

(i) $T_8(2) \geq T_2$,

(j) $T_9 \geq T_{10} > T_{11}$ when $p_i(x)$, $i = 1, 2, \ldots, m$ are known and normal

(k) $T_{10} \geq T_{11}$ when $p_i(x)$, $i = 1, 2, \ldots, m$ are normal with unknown parameters,

(l) $T_{10}$ and $T_{11}$ cannot be compared when $p_i(x)$, $i = 1, 2, \ldots, m$ are not known since the accuracy will depend on how far the $p_i(x)$, $i = 1, 2, \ldots, m$ are from being normal.

(m) $T_1$ is uniformly less accurate when compared to every other member of the list.

Following are several orderings with respect to computing speed. In this case "$\geq$" means "takes no more time than".

(a) $T_2 \geq T_3 \geq T_4 \geq T_8(4) \geq T_1 \geq \{T_9, T_{10}, T_{11}\}$,

(b) $T_8(2) \geq T_8(3) \geq T_8(4) \geq T_1 \geq \{T_9, T_{10}, T_{11}\}$,

(c) $T_3 \geq T_8(3) \geq T_8(4) \geq T_1 \geq \{T_9, T_{10}, T_{11}\}$,

(d) $T_5 \geq T_6 \geq T_7 \geq T_8(7) \geq T_1 \geq \{T_9, T_{10}, T_{11}\}$,

(e) $T_8(5) \geq T_8(6) \geq T_8\ (7) \geq T_1 \geq \{T_9, T_{10}, T_{11}\}$,

(f) $T_6 \geq T_8(6)$,

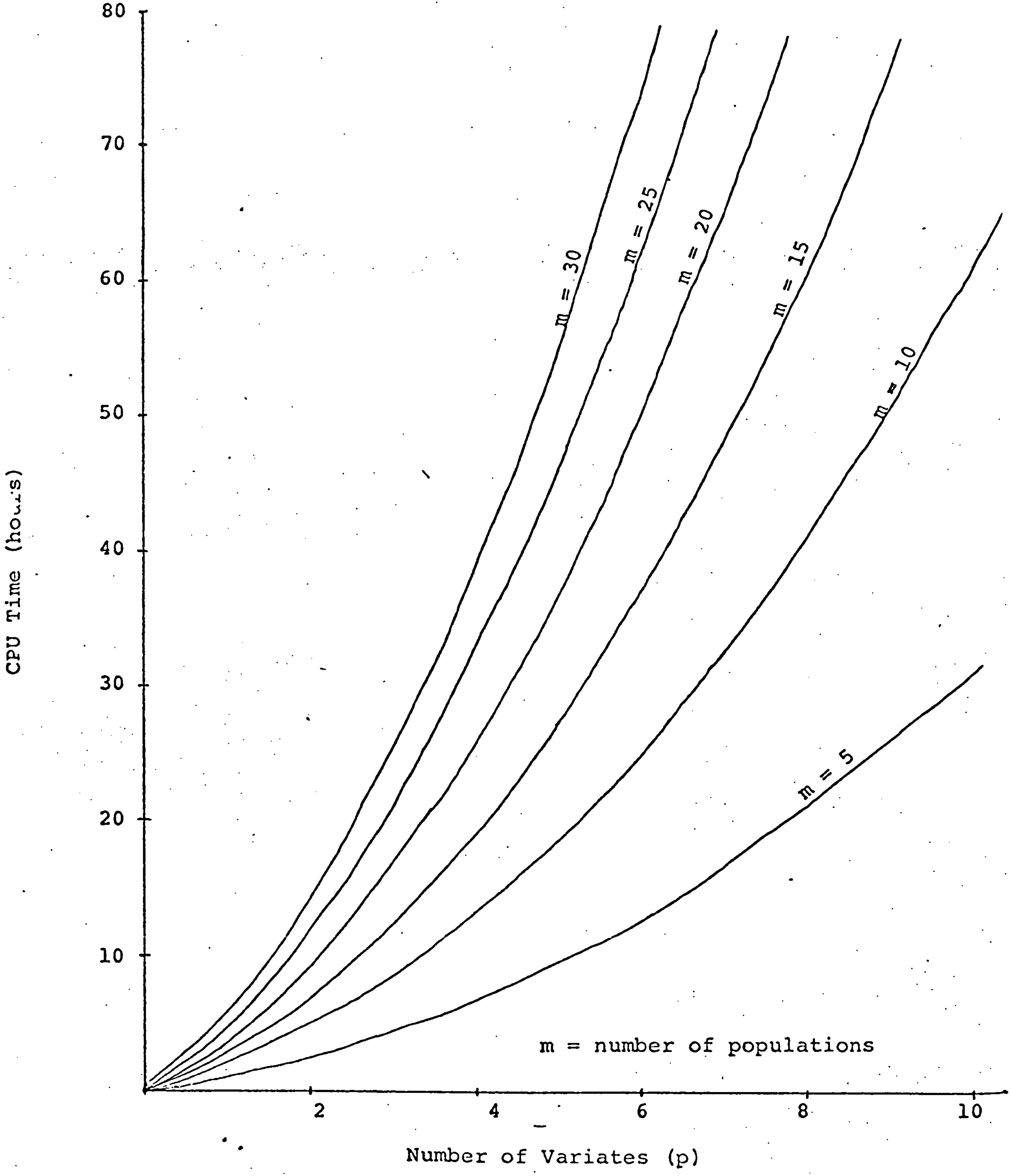
(g) whether $T_2 \geq T_5$ depends on the speed required to "compare" with the speed required to "look up".

A preliminary analysis was performed by Data Processing Branch and Systems Engineering Branch personnel, National Aeronautics and Space Administration, Manned Spacecraft Center in conjunction with Control Data Corporation, to estimate the Computer Processing Unit time required to perform pattern recognition using the Purdue University LARS (Laboratory for Applications of Remote Sensing) classification technique. The LARS technique assumes the data from each class is drawn from a multivariate normal population and the classification is done according to the maximum likelihood rule. That is, an observation x is classified in population k if $\hat{p}_k(x) = \max_i \{\hat{p}_i(x)\}$, k = 1, 2,...,m, where the multivariate density functions $p_i(x)$, i = 1, 2,...,m are evaluated using estimates of the mean vector and covariance matrix for each class. Thus the LARS technique is the classical Bayes technique $T_{10}$ with equal priors, that is, $q_i = 1/m$, i = 1, 2,...,m. The results of the analysis are summarized in Figure 1. The graph in Figure 1 presents the CYBER 73-14 time required to classify $16 \cdot 10^6$ picture elements in a remote sensing data situation. The graph gives the time required to classify elements given the member of classes (populations) to be separated and the number of channels (variates) to be used in the classification process.

In a remote sensing data situation the data is obtained in the form of an image (or scene) which is a rectangular region consisting of r rows (scan lines) and c columns (number of resolution elements or cells per scan line). Each cell generates a p-variate observation. Thus to recognize a scene one must perform rc

Figure 1.

CPU Time as a Function of the Number of Variates (p)



discriminate tasks, i.e. one must classify rc observations. The graph in Figure 1 gives the time required to classify $16 \cdot 10^6$ observations (elements), where, for example, $r = 4 \cdot 10^3$ and $c= 4 \cdot 10^3$. Data of this magnitude is quite common in a remote sensing data situation.

## 6. A Monte Carlo Evaluation.

The techniques evaluated by Monte Carlo Techniques are:

1. ($T_1$) A Bayes algorithm using Wilks concepts [12].
2. ($T_2$) A table look-up technique.
3. ($T_3$) A table look-up technique.
4. ($T_5$) A p-dimensional rectangular technique.
5. ($T_6$) A q-dimensional rectangular technique.
6. ($T_9$) The classical Bayes technique assuming known parameters.
7. ($T_{10}$) The classical Bayes technique using estimated parameters.

The seven techniques listed above have been taken from the list in section 5. For the Monte Carlo simulation we took m = 3, p = 3, and n = 100 samples from each population were generated using the multivariate normal random generator in [11]. Three separate trials were conducted corresponding to three different sets of multivariate normal populations. Following are the vector means and covariance matrices for each trial.

Trial I:

$$\mu_1 = (150,\ 200,\ 100)^T,$$
$$\mu_2 = (100,\ 150,\ 200)^T,$$
$$\mu_3 = (200,\ 100,\ 150)^T,$$

and

$$\Sigma_1 = \Sigma_2 = \Sigma_3 = 625\ I.$$

Trial II:

$$\mu_1 = (150,\ 200,\ 100)^T,$$
$$\mu_2 = (100,\ 150,\ 200)^T,$$
$$\mu_3 = (200,\ 100,\ 150)^T,$$

$$\Sigma_1 = \Sigma_2 = \begin{pmatrix} 625 & 375 & 0 \\ 375 & 625 & 375 \\ 0 & 375 & 625 \end{pmatrix},$$

and

$$\Sigma_3 = \begin{pmatrix} 625 & -375 & 0 \\ -375 & 625 & -375 \\ 0 & -375 & 625 \end{pmatrix}.$$

Trial III:

$$\mu_1 = (125,\ 150,\ 175)^T,$$
$$\mu_2 = (150,\ 175,\ 125)^T,$$
$$\mu_3 = (175,\ 125,\ 150)^T,$$

$$\Sigma_1 = \begin{pmatrix} 400 & -240 & -200 \\ -240 & 400 & 360 \\ -200 & 360 & 400 \end{pmatrix},$$

$$\Sigma_2 = \begin{pmatrix} 400 & 240 & -200 \\ 240 & 400 & -360 \\ -200 & -360 & 400 \end{pmatrix},$$

and

$$\Sigma_3 = \begin{pmatrix} 400 & -240 & 200 \\ -240 & 400 & -360 \\ 200 & -360 & 400 \end{pmatrix}$$

The results of the discriminate analysis results are given in Table 2, 3, and 4. There were a total of 24 discriminate analyses performed. The techniques labeled 1 and 1a in the tables denote Bayes procedures with the data compressed by Wilks technique from 3 variates to 2 and 3 variates to 1, respectively. For technique number 5($T_6$) the q-dimensional rectangular technique was used for the choice of the best 2 variates from the 3 original variates.

Tables 2, 3, and 4 give the number of correct classifications in each population, the number of misclassifications, the number not classified, and the amount of time (in .01 sec.) in each analysis for each trial, respectively.

The orderings with regard to accuracy in section 6 are supported by the Monte Carlo results for the 7 techniques used.

However, all the orderings with regard to computer time are not, strictly speaking, supported. For example, section 6 has $T_2$ as taking no more computer time than $T_3 (T_2 \geq T_3)$. However, for all three trials $T_3$ took less time than $T_2$, although the difference was very small, and the computer ordering with respect to time could be due to certain factors such as language used (Fortran), programming procedures, and so on.

Table 2. Trial I.

| | Technique | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | 1 | 1a | 2 | 3 | 4 | 5 | 6 | 7 |
| Classified in $\pi_1$ | 99 | 100 | 97 | 97 | 79 | 91 | 99 | 100 |
| Classified in $\pi_2$ | 100 | 98 | 97 | 96 | 80 | 91 | 99 | 99 |
| Classified in $\pi_3$ | 98 | 99 | 98 | 96 | 83 | 94 | 100 | 100 |
| Misclassified | 3 | 3 | 2 | 5 | 0 | 2 | 2 | 1 |
| Not classified | 0 | 0 | 6 | 6 | 58 | 22 | 0 | 0 |
| Time in .01 sec.* | 542 | 508 | 288 | 278 | 287 | 278 | 750 | 718 |

* Includes Input-Output time of 2 seconds.

Table 3. Trial II.

|  | Technique |  |  |  |  |  |  |  |
|---|---|---|---|---|---|---|---|---|
|  | 1 | 1a | 2 | 3 | 4 | 5 | 6 | 7 |
| Classified in $\pi_1$ | 100 | 100 | 96 | 99 | 54 | 87 | 99 | 99 |
| Classified in $\pi_2$ | 99 | 98 | 94 | 98 | 56 | 86 | 100 | 100 |
| Classified in $\pi_3$ | 100 | 70 | 90 | 96 | 81 | 96 | 98 | 98 |
| Misclassified | 1 | 32 | 2 | 4 | 9 | 11 | 3 | 3 |
| Not classified | 0 | 0 | 18 | 3 | 94 | 29 | 0 | 0 |
| Time in .01 sec.* | 560 | 457 | 283 | 282 | 287 | 283 | 660 | 662 |

* Includes Input-Output time of 2 seconds.

Table 4. Trial III.

|  | Technique |  |  |  |  |  |  |  |
|---|---|---|---|---|---|---|---|---|
|  | 1 | 1a | 2 | 3 | 4 | 5 | 6 | 7 |
| Classified in $\pi_1$ | 94 | 69 | 94 | 90 | 25 | 85 | 95 | 95 |
| Classified in $\pi_2$ | 91 | 75 | 93 | 76 | 23 | 89 | 94 | 95 |
| Classified in $\pi_3$ | 84 | 84 | 88 | 83 | 19 | 80 | 89 | 90 |
| Misclassified | 31 | 72 | 12 | 39 | 1 | 34 | 22 | 20 |
| Not classified | 0 | 0 | 13 | 11 | 232 | 12 | 0 | 0 |
| Time in .01 sec.* | 578 | 463 | 287 | 280 | 282 | 302 | 650 | 668 |

* Includes Input-Output time of 2 seconds.

7. Feasibility and Practical Aspects of Table Look-Up.

There are various instances when the Table look-up ceases to be a practical technique. One such instance is when the number of values to be computed for the table exceeds the number of values to be classified. Three ways in which to reduce memory requirements for the Table look-up technique are

(1) reduce the number of variates p,

(2) store only regions of interest in the measurement space, and

(3) compress the regions of interest.

Item (3) involves storing the classification for several contiguous locations all in a single core memory location. This is the transformation discussed in [4] which produces a coarse set of lattice points and is called "pointer scale". In [4] it is seen how storage requirements for a table for one population are reduced from $256^2 = 65,536$ to 864 and from 864 to 144 by successively using (1), (2), and (3) above.

In our simulation study the table for each population consisted of a p-dimensional lattice cube having 12 points to a side. This called for $12^p = 12^3 = 1728$ storage locations. The total storage requirements in our case were then $m(12)^p = 3(12)^3 = 5184$ locations. In using the best 2 of 3 characteristics for each population the storage requirements were $m(12)^q = 3(12)^2 = 432$ locations. Each of the lattice points was classified by the Bayes procedure prior to classifying the 300 observations resulting in 5184 classifications ($p = 3$) or 432 classifications ($q = 2$). In these cases the number of classifications necessary to construct the table exceeds the number (300) of further classifications. This was done to get information regarding accuracy, speed, and storage requirements. In practice one could be faced with classifying data of the magnitude of $10^6$ observations, such as in remote sensing, in which case the Table look-up technique would be quite practical.

In summary, there are cases when the Table look-up would prove quite useful.

## 8. Concluding Remarks

From the evaluation in section 5 and the simulation results it appears that the table look-up technique has much to recommend it, especially if all p variates are used and if all observations x falling in the no decision region are classified according to the classical Bayes procedure. The procedure $T_8(2)$ had 6, 18, and 13 observations for trials I, II, and III, respectively, falling in the no decision region. Procedure $T_8(5)$ had 22, 29, and 12 observations falling in the no decision region. All these observations could have been classified according to the classical Bayes procedure and would have made the results as accurate.

If the number of observations falling in the no decision region is large, say 30% or more, then $T_8(2)$ and $T_8(5)$ would take considerably more time than $T_2$ and $T_5$. It would be useful in cases like these to use the p-dimensional rectangular approach where two or more rectangles in each $R_i$ are utilized.

# REFERENCES

[1] Remote Sensing of Earth Resources, NASA SP 7036, A Literature Survey with Indexes, September 1970.

[2] Chang, C. Y., "Review of Pattern Recognition Techniques Development in Remote Sensing Applications", Lockheed Electronics Co., Houston, Texas for NASA Manned Spacecraft Center, Houston Texas, LEC/HASD No. 640-TR-021, August 31, 1971.

[3] Chang, C. Y., "Divergence and Probability of Misclassification", Lockheed Electronics Co., Houston, Texas, for NASA Manned Spacecraft Center, Houston, Texas, September 30, 1971, LEC/HASD No. 640-TR-031, NASA EOD No. EOD1750.

[4] Eppler, W. G., Helmke, C. A., and Evans, R. H., "Table look-up approach to pattern recognition", Proceedings of 7th International Symposium on Remote Sensing of the Environment, the University of Michigan, May 1971.

[5] Kailath, T., "The divergence and Bhattacharyya distance measures in signal selection", IEEE transaction on Communication Technology, Vol. COM-15, (February 1967), pp. 52-60.

[6] Kullback, S., Information Theory and Statistics, John Wiley & Sons, Inc., London (1969), pp. 6-31.

[7] Landgrebe, D. A. and LARS Staff, "LARSYAA, A processing system for airborne earth resource data, LARS Information Note 091968, Purdue University, Lafayette, Indiana, September 1969.

[8] Levine, M. D. "Feature selection: a survey", Proceedings of IEEE, Vol. 57, No. 8, August 1969, pp. 1391-1408.

[9] Marill, T. and Green, D. M., "On the effectiveness of receptors in recognition system", IEEE Trans. Information Theory, IT-9, January 1963.

[10] Mood A., and Graybill, F. A., Introduction to the Theory of Statistics. McGraw-Hill Company, New York (1963), p. 432.

[11] Newman, T. G. and Odell, P. L., The Generation of Random Variates, Griffin's Statistical Monographs and Courses, No. 29, (1971), pp. 37-44.

[12] Wilks, S. S., Mathematical Statistics, John Wiley and Sons, Inc., (1963), pp. 540-592.