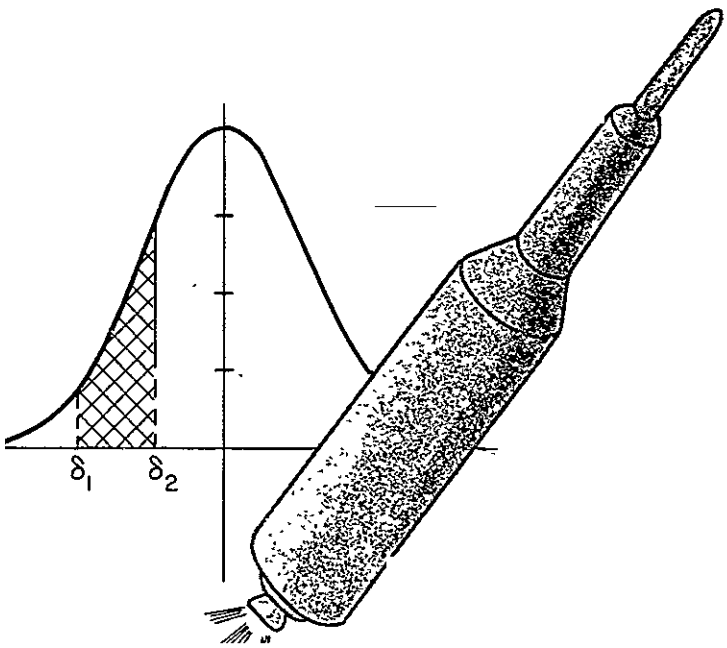
COMPUTING PROCEDURES IN STATISTICAL DISCRIMINATE ANALYSIS



TEXAS CENTER FOR RESEARCH
3100 PERRY LANE, AUSTIN, TEXAS

THREE LECTURES ON COMPUTING PROCEDURES IN STATISTICAL DISCRIMINATE ANALYSIS

by Patrick L. Odell
Professor of Mathematics and Statistics
Texas Tech University
Lubbock, Texas

July, 1969

Preface

In this short monograph, I have recorded in manuscript three lectures and a selected bibliography on statistical discriminate analysis. The lectures were given to scientists at NASA Manned Spacecraft Center in July 1969. Each of the lectures which are preliminary drafts of papers to be published contain information not found in the statistical literature. In the process several questions have been introduced that remain open at this date, and it remains an important task to resolve these.

The efforts of Dr. T. G. Newman, Texas Tech University, and Dr. Mike Speed of NASA Manned Spacecraft Center who contributed portions of their knowledge and experience to this study are gratefully acknowledged.

P. L. O.

July, 1969

TABLE OF CONTENTS

PRELIMINARY REMARKS

In Chapter 6 of his excellent text [1] Professor Anderson gives a clear and precise development of the theory of statistical discrimination. The definitions and theorems presented here are essentially those in that chapter and give a brief background for the lectures.

Let $\pi_1, \pi_2, \ldots, \pi_m$ denote m distinct population whose associated probability p-variate density functions are $p_1(x), p_2(x), \ldots, p_m(x)$, respectively. Let x be an observation from one of these populations. Let

$$C(i|j,R) \tag{1}$$

denote the cost of misclassification of an individual from population π_j as being from population π_i using the decision rule

$$R = (R_1, R_2, \ldots, R_m)$$

where R_i denotes a distinct p-dimensional region such that if x belongs to R_i, the individual is assigned to π_i. Also, we denote the a priori probabilities that an observation comes from population π_i as q_i. We note that

$$P(j|i,R) = \int_{R_j} p_i(x)\,dx \tag{2}$$

is the probability of classifying on observation x from π_i as being from π_j. The classification procedure, then, is to classify an observation as coming from π_j if it falls in R_j. The following theorem summarizes the procedure $R = (R_1, R_2, \ldots, R_m)$ such that the expected loss

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

is minimized. That procedure is usually called the Bayes procedure. The classification procedure, then, is to classify an observation as coming from π_j if it falls in R_j.

Theorem 1. If q_i is the a priori probability of drawing an observation from population π_i with density $p_i(x)$ $(i = 1,\ldots,m)$ and if the cost of misclassifying an observation from π_i as from π_j is $C(j|i)$, then the regions of classification, $R_1,\ldots,R_m$, that minimize the expected cost are defined by assigning x to R_k if

$$\sum_{\substack{i=1 \\ i\neq k}}^{m} q_i p_i(x)C(k|i) < \sum_{\substack{i=1 \\ i\neq j}}^{m} q_i p_i(x)C(j|i) \qquad (3)$$

$$(j = 1,\ldots,m,\ j \neq k).$$

[If (3) holds for all $j(j \neq k)$ except for h indices and the inequality is replaced by equality for those indices, then this point can be assigned to any of the $h + 1$ π's.] If the probability of equality between right-hand and left-hand sides of (3) is zero for each k and j under π_i (each i), then the minimizing procedure is unique except for sets of probability zero.

We note that in many cases the values of $q_1,\ldots,q_m$ are unknown. If this is true then one can define

$$\sum_{\substack{j=1 \\ j\neq i}}^{m} C(j|i)P(j|i,R) = r(i,R),$$

the conditional expected loss if the observation is from π_i. If

we choose a procedure $R = (R_1,\ldots,R_m)$ which minimizes the maximum conditional expected loss, this procedure is called the minimax procedure.

If $\alpha_i(R) = 1 - P(i|i,R)$ is the probability of making a wrong decision when using procedure R and sampling from π_i, then Von Mises [2] has shown that the set of q's, say $q_1^*,q_2^*,\ldots,q_3^*$, such that

$$\alpha_1(R^*) = \alpha_2(R^*) = \ldots = \alpha_m(R^*)$$

yields the minimax solution. The problem becomes one of searching for the set $q^* = \{q_1^*,\ldots,q_m^*\}$ and then computing the regions of classification $R_1^*,\ldots,R_m^*$ which follows from Theorem 1.

REFERENCES

Anderson, T.W., An Introduction to Multivariate Statistical Analysis, Wiley and Sons, 1958.

Von Mises, R.,"On the Classification of Observation Data into Distinct Groups", Annals of Math. Statistics. Vol. 16, No. 1, (1945), pp. 68-73.

Lecture

NON-PARAMETRIC DISCRIMINATION TECHNIQUES[1]

P. L. Odell[2], T. G. Newman[2] and M. Speed[3]

Introduction

Except for the work done by Kendall [1], [2] and Fix and Hodges [3], little has been reported concerning non-parametric techniques for performing statistical discrimination. By non-parametric techniques we mean only that the mathematical form of the multivariate probability density functions of the populations involved are unknown to the experimenter. It should be noted that we did not use the term distribution free which is used by many in a different sense, (See [2, p. 170] or [21, p. 15-17]).

For clarity and completeness we define what we mean by:

The Discrimination Problem, I. Let $\pi_1, \pi_2, \ldots, \pi_m$ denote m distinct p-variate populations whose multivariate probability density functions $p_1(x), p_2(x), \ldots, p_m(x)$ are known. Let $q_1, q_2, \ldots, q_m$ be the known a priori probabilities that a sample is selected from population $\pi_1, \pi_2, \ldots, \pi_m$, respectively. Let $C(i|j)$ be the cost of assigning an individual from population π_j to population π_i such that

$$C(i|j) > 0 \qquad i \neq j \quad i,j = 1,\ldots,m$$
$$= 0 \qquad i = j \quad i = 1,\ldots,m.$$

Given a sample generated by an individual selected at random from

---

1 This research was supported in part by NASA-MSC, under Contract NAS - 9-6963.

2 Texas Tech University, Lubbock, Texas.

3 Computation and Analysis Division, NASA-MSC, Houston, Texas

one of the populations, give a decision rule to assign the individual to one of the m populations.

A decision rule for assigning an individual to one of the populations π_i, $i = 1,2,\ldots,p$ which minimizes the expected cost of misclassification is known and discussed clearly in [4, p. 142-147]. The following theorem summarizes the technique:

Theorem 1. The regions of classification $R_1,\ldots,R_m$, that minimizes the expected cost of misclassification are defined by assigning x to R_k if

$$\sum_{\substack{i=1 \\ i\neq k}} q_i p_i(x) C(k|i) < \sum_{\substack{i=1 \\ i\neq j}} q_i p_i(x) C(j|i) \qquad (1)$$

$$j = 1,\ldots,m \qquad j \neq k.$$

We will use the notation $R = \{R_1,\ldots,R_m\}$ to be a set of disjoint but not necessarily connected regions whose union is the Euclidean p-space.

Clearly, one must know a great amount in order to apply Theorem 1. However, in practice most of the quantities $q_i, C(i|j)$, and $p_i(x)$ $i = 1,\ldots,m$ are actually estimates yet can be assumed known exactly so that Theorem 1 can be applied. References which discuss Problem I are [4], [5], and [6].

The Discriminate Problem II. Let the conditions in Problem I remain true, except that $q_i,\ldots,q_m$ are unknown. Given a sample selected at random from one of the populations give a decision rule to assign a sample x to a population.

Various reductions can be made if $q_1 = q_2 = \ldots = q_m$, and/or $C(i|j) = C(i'|j')$ for all i,i' and j,j' $i \neq j$, $i \neq j$.

Obviously (1) cannot be used to determine the regions R_1, $R_2,\ldots,R_m$ when $q_1,\ldots,q_m$ are unknown, hence another strategy for making decisions has been developed. If we have a region R_i for classifying x as from π_i, the probability or classifying the sample x as from π_j is

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

The expected loss for classifying x as being in π_j if the observation is from π_i is

$$C(j|i)\ P(j|i) = r_j(i,R)$$

where $R = (R_1,\ldots,R_m)$ again denotes the partition of the Euclidean p-space into m distinct regions of classification.

A principle that usually leads to a unique procedure is the so-called <u>minimax principle</u>. A procedure is minimax if the maximum expected loss r(i,R) is minimum for all possible values of q_1, $q_2,\ldots,q_m$.

Example 1. Let $R^{(1)} = (R_{11},R_{12},R_{13})$ and $R^{(2)} = (R_{21},R_{22},R_{23})$ be two partioning associated with two procedures for classifying. Let r_K denote the (3x3) matrix

$$r_K = \{r_j(i|R^{(K)})\} \qquad K = 1,2$$

$$= \begin{bmatrix} 0 & r_i(2|R^{(K)}) & r_i(3|R^{(K)}) \\ r_2(1|R^{(K)}) & 0 & r_2(3|R^{(K)}) \\ r_3(1|R^{(K)}) & r_3(2|K^{(K)}) & 0 \end{bmatrix}$$

Select that procedure $R^{(1)}$ and $R^{(2)}$ that minimizes the maximum expected loss as q_1, q_2, and q_3 range over all admissible values, that is $0 < q_i < 1$, $q_i > 0$ and $\sum_{i=1}^{3} q_i = 1$. The elements of each matrix is easily computed if the regions $R^{(1)}$ and $R^{(2)}$ are given. However, if one is searching for a procedure (or a partition) R which gives a minimax solution, it is not clear how one establishes that procedure. Comparisons of procedures once they are given is not difficult, but defining the region out of all possible regions can be difficult. [See 4, p. 142-147]. Solutions for m = 2 and 3 have been indicated. The reader is referred to [4, pp. 134-136] for a discussion of these cases and an indication of the problems associated with searching for a minimax solution. The solution in general is iterative and approximate.

Clearly, the problem of discrimination is ill-posed if the costs of misclassification are not known; however, there are many applications in which these costs can be assumed identical and further special results obtained when costs are unknown.

The Discrimination Problem III. Let the conditions defined in Problem I be true except that the set $\{p_1(x),\ldots,p_m(x)\}$ is not known. Instead let $p_i(x) \in F_i(x;\theta_i)$ where F_i is a known family of probability densities depending on a $k_i \times 1$ parameter vector θ_i, whose elements are unknown and must be estimated using a previously drawn sample $x_i = \{x_{i1},x_{i2},\ldots,x_{in}\}$ $i = 1,2,\ldots,n$. Given a sample selected at random from one of the populations give a decision rule to assign the sample x to one of the m populations.

As one might suspect, great amounts have been reported concerning this problem when for each i, F_i is the normal family with the covariance matrix held fixed either known or unknown. Generally, one requires that if F is the normal family with unknown mean vector μ and covariance matrix Σ for each population, then

$$P_i(x) = N(\mu^{(i)}, \Sigma^{(i)})$$

It is popular, but not necessary, to assume that

$$\Sigma^{(1)} = \Sigma^{(2)} = \ldots = \Sigma^{(m)},$$

the covariances to be the same. This assumption leads to the linear discriminate function. The analysis is summarized in [4, pp. 137-139]. One should be aware of the work done by Kabe [7] which is not referenced there.

Briefly, one proceeds as in Problem I with each $\mu^{(i)}$ and $\Sigma^{(i)}$ replaced by its maximum likelihood estimate obtained from the m previously drawn samples x_i $i = 1,2,\ldots,m$.

In this paper we are interested primarily in

The Discrimination Problem IV. Let the conditions in Problem I be true, except that there is no information concerning $p_i(x)$ $i = 2,\ldots,m$ except that they are continuous, finite and their moments exist. Given a sample from each of the m populations, devise a decision rule for assigning a sample value from one of these populations to a population of the set $\pi = \{\pi_1,\pi_2,\ldots,\pi_m\}$.

2. Kendall's Suggestions

Kendall has suggested two techniques; one he calls the convex

hull technique; the second he calls the order-statistic method. In his discussion Kendall discounts the convex hull technique and gives reasons why it may not be particularly useful. The authors refer the interested reader to [1] and [2] for these comments. Kendall's example in [2] indicates there may be some real value in using his second suggestion, the order-statistic method. This latter technique is compared with a new technique proposed in this paper.

Let $x^{(i)}$ be a p x 1 random vector whose probability density function $p_i(x)$ is unknown. Let

$$x_j^{(k)} = \{x_{ij}^{(k)}\},\ j = 1,2,\ldots,N_i \qquad \begin{matrix} i = 1,2,\ldots,p \\ k = 1,2,\ldots,m \end{matrix}$$

denote a random sample of size N_i from the k^{th} population. Define the interval for each (i,k)

$$I(i,k) = [\min_j x_{ij}^{(k)},\quad \max_j x_{il}^{(k)}]$$

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

$$j = 1,2,\ldots,N_i$$

Hence for each i there are m such intervals

$$I(i,1),\ I(i,2),\ldots,I(i,m).$$

Using set notation, let

$$D_{i1} = I(i,1) - \bigcup_{j' \neq 1}^{m} I(i,j')$$

$$D_{i2} = I(i,2) - \bigcup_{j' \neq 2}^{m} I(i,j')$$

$$D_{ij} = I(i,j) - \bigcup_{j' \neq j'}^{m} I(i,j'),$$

the set of points in $I(i,j)$ but not in the union of sets $I(i,j')$ $j \neq j$, $j' = 1,2,\ldots,m$. Let

$$x = \begin{bmatrix} x_1 \\ x_2 \\ \cdot \\ \cdot \\ \cdot \\ x_p \end{bmatrix}$$

be the sample from one of the m populations to be assigned. The decision rule proposed by Kendall is simply

a. if $x_{i_1} \in D_{i_1,j}$ for some j assign the sample to π_j.

b. if $x_{i_1} \in D_{i_1,j}$ for all j then consider x_{i_2}, where $i_2 \neq i_1$.

c. if $x_{i_1} \in D_{i_2,j}$ for some j, then assign the sample to π_j.

d. if $x_{i_2} \notin D_{i_2,j}$ for all j, then consider x_{i_3} where $i_3 \neq i_2$ or i_3.

e. finally if for every i and every j $x_i \notin D_{i,j}$ $j = 1,2,\ldots,m$ then x cannot be assigned and no decision is made.

The advantages cited by Kendall are as follows.

(1) The procedure is distribution-free.

(2) The procedure involves no arithmetic other than counting.

Relatively small amounts of data can be analyzed by hand. For large amounts it would be simple to write a program for an electronic computer.

(3) It shows which variables are the most important in the discrimination. In fact, it proceeds by using the variables in order of importance as measured by degree of overlap.

It is important to note if for all i, that x is assigned to π_j, that one is indeed confident that the assignment is correct. However, there exists no reason to believe that such a situation would exist for every sample taken. Hence, different decisions would result depending on the order of elements of the vector x used in discriminating. It is also clear that as the sample sizes $N_1,\ldots,N_m$ increase in magnitude, that the number of indecisions will increase. That is, the larger the initial sample sizes, the greater the confusion. Clearly, this appears strange to those who are used to techniques in which the more information one has the better one can estimate or discriminate.

3. An Alternative Technique

Assuming that the experimenter can sample from $\pi_1,\ldots,\pi_m$, we denote these samples by the sets

$$x^{(1)} = \{x_1^{(1)}, x_2^{(1)},\ldots,x_{N_1}^{(1)}\}$$

$$x^{(2)} = \{x_1^{(2)}, x_2^{(2)},\ldots,x_{N_2}^{(2)}\}$$

$$\vdots$$

$$x^{(m)} = \{x_1^{(m)}, x_2^{(m)},\ldots,x_{N_m}^{(m)}\}$$

where N_i $i = 1,2,\ldots,m$ denotes the sample sizes and each $x_j^{(k)}$ $j = 1,2,\ldots,N_j$ is a p x 1 vector of elements $x_{ij}^{(k)}$ $i = 1,2,\ldots,p$.

The k^{th} sample is used to estimate the k^{th} probability density function by an estimator discribed later in this section. If the set

$$\hat{p}(x) = \{\hat{p}_1(x), \hat{p}_2(x), \ldots, \hat{p}_m(x)\}$$

denotes the set of estimators of $p_1(x), p_2(x), \ldots, p_m(x)$ respectively, and x^* a sample from one of the populations, then the individual whose measurements yielded x^* is assigned to π_p if k is such that

$$\hat{p}_k(x^*) = \max_{1 \leq j \leq m} \hat{p}_j(x^*).$$

That is to say, we assign the individual to that population whose estimated likelihood is maximum.

Note that if $p_i(x)$ is a consistent estimator for $p_i(x)$ for every $i = 1,2,\ldots,m$, then, in the limit, we have the situation defined in Problem I whose solution is known to possess optimal properties with respect to minimizing expected costs of misclassification.

The estimator $p_i(x)$ for the probability density function $p_i(x)$ is defined by

$$\hat{p}_k(x) = \frac{1}{N_k} \sum_{j=1}^{N_k} W_h(x - X_j^{(k)}) \tag{2}$$

where $X_i^{(k)}$ is the i^{th} sample vector from the k^{th} population and

$$W_k(x - X_i^{(k)}) = C_p \prod_{i=1}^{p} \left[\frac{\sin(x_i - X_{ij}^{(k)})}{X_{ij}^{(k)}/2} \right]^4 \tag{3}$$

where $$x = \begin{bmatrix} x_1 \\ \vdots \\ x_p \end{bmatrix}$$

is the independent variable in $\hat{p}_k(x)$.

The constant C_p is selected in such a manner such that

$$\int_{-\infty}^{\infty} W(t)dt = 1$$

The estimator selected is then consistent in quadratic mean. A discussion of such estimators can be found in [8], [9], [10]. However, we will give an outline for a precise development of the estimator. Let

$$\delta(t_1, t_2, \ldots, t_p)$$

be a p-dimensional Dirac delta functional and $p(x_1, x_2, \ldots, x_p)$ a p-variate continuous probability density function, then

$$p(x_1, x_2, \ldots, x_p) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \delta(x_1 - t_1, \ldots, x_p - t_p) p(t_1, \ldots, t_p) \, dt_1 \ldots dt_p$$

which can be approximated by (read $\cong$ as "approximately equal to")

$$p(x_1, x_2, \ldots, x_p) \cong \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} W(x_1 - t_1, \ldots, x_p - t_p) p(t_1, \ldots, t_p) \, dt_1 \ldots dt_p$$

where W(sometimes called a window function) is an approximation for δ. Then

$$E[W(x_i - t_i), \ldots, (x_p - t_p)] \cong p(x_i).$$

Finally, the arithmetic mean of W's is approximately the best linear unbiased estimator for E[W], that is

$$\hat{p}(x_1,\ldots,x_p) = \frac{1}{N} \sum_{j=1}^{N} W(x_i - t_{ij},\ldots,x_p - t_{ij}).$$

We have selected W for each k to be (3).

It should be noted that there are other estimators for $\hat{p}_i(x)$ depending on the form of W. The right-hand side of (3) is only one of many approximations of the p-dimensional Dirac delta function.

In summary, one approaches the Problem IV just as one would in Problem I, except one replaces the probability density functions $p_1(x), p_2(x), \ldots, p_m(x)$ by their estimators $\hat{p}_1(x), \hat{p}_2(x), \ldots, \hat{p}_m(x)$ obtained by using (2).

4. A Comparison Of Techniques

Since the optimality properties of the alternative technique proposed in this paper cannot be easily studied analytically, it is at least convenient to study its properties using a simulation based on the technique popularly called "Monte Carlo" [11], [12]. The results of the simulation performed on a Univac 1108 at NASA Manned Spacecraft Center are presented and discussed in this section.

In the simulation we arbitrarily let m = 3, p = 2; that is, we considered the problem of assigning a (2 x 1) observation vector to one of three bivariate populations, π_1, π_2, and π_3. Each population is defined by a mean vector and a (2 x 2) covariance matrix. Table 1. briefly summarizes the input data to the simulation.

Case	i	N_i	Mean Vector	The Elements of the Covariance Matrix
	1	100	(0,0)	$\sigma_{11} = \sigma_{22} = 1, \sigma_{12} = 0$
1	2	100	(1,1)	$\sigma_{11} = \sigma_{22} = 1, \sigma_{12} = 0$
	3	100	(-1,1)	$\sigma_{11} = \sigma_{22} = 1, \sigma_{12} = 0$
	1	100	(2,1)	$\sigma_{11} = \sigma_{22} = 1, \sigma_{12} = 0$
2	2	100	(-.3,.6)	$\sigma_{11} = \sigma_{22} = 1, \sigma_{12} = 0$
	3	100	(1.5,.8)	$\sigma_{11} = \sigma_{22} = 1, \sigma_{12} = 0$
	1	100	(2,1)	$\sigma_{11} = 1.6, \sigma_{22} = 2.0, \sigma_{12} = .3$
3	2	100	(-.3,.6)	$\sigma_{11} = 4, \sigma_{22} = 1.2, \sigma_{12} = .5$
	3	100	(1.5,.8)	$\sigma_{11} = .9, \sigma_{22} = 4.0, \sigma_{12} = .6$
	1	100	(2,5)	$\sigma_{11} = 3, \sigma_{22} = 3, \sigma_{12} = .6$
4	2	100	(1,-1)	$\sigma_{11} = 1, \sigma_{22} = 2.5, \sigma_{12} = 0.2$
	3	100	(6,1)	$\sigma_{11} = 2, \sigma_{22} = 1.2, \sigma_{12} = 0.4$
	1	100	(2,5)	$\sigma_{11} = \sigma_{22} = 1, \sigma_{12} = 0$
5	2	100	(1,-1)	$\sigma_{11} = \sigma_{22} = 1, \sigma_{12} = 0$
	3	100	(6,1)	$\sigma_{11} = \sigma_{22} = 1, \sigma_{12} = 0$

Table 1. The Input Data for First Simulation

One hundred (2 x 1) samples from population, π_1 in each of the five cases defined in Table I, when the probability density was assumed to be bivariate normal with appropriate mean vectors and covariance matrices we generated using the usual Monte Carlo techniques. Three techniques were compared, those being

(i) Kendall's ranking technique

(ii) The Optimal Technique as defined by Theorem 1.

(iii) The Non-Parametric Technique using Estimated Probability Densities

The number of times a sample value from population π_1 was assigned to each population is listed along with the number of times where Kendall's technique led to a no-decision situation. These counts are summarized and presented in Table II.

CASE	TECHNIQUE	π_1	π_2	π_3	NO DECISION
1	Kendall	7	0	4	89
	Optimal	61	16	23	0
	Non-Parametric	57	18	25	0
2	Kendall	3	1	0	96
	Optimal	56	9	35	0
	Non-Parametric	60	9	31	0
3	Kendall	0	0	2	98
	Optimal	45	9	46	0
	Non-Parametric	48	9	43	0
4	Kendall	88	0	2	10
	Optimal	96	2	2	0
	Non-Parametric	94	4	2	0
5	Kendall	92	0	5	3
	Optimal	100	0	0	0
	Non-Parametric	99	1	0	0

Table II. The Number of Samples from π_1 Assigned to π_i $i = 1,2,3$ or No-Decision.

Clearly, Kendall's technique leads to what many, including these authors, would believe to be an unrealistic number of no-decision situations. The value of Kendall's technique can be questioned, especially when sample sizes are relatively large.

It is noted that the proposed non-parametric technique is indeed efficient when compared with the optimal technique. However,

as previously indicated one would expect Kendall's technique to be poor for relatively large sample sizes. A natural question arises. Is there a sample size such that the Kendall technique is more or equally efficient?

In order to study the effect of sample size on the probabilities of proper classification, we considered three populations π_1, π_2, π_3. The first population π_1 is bimodal and is the simple average of two univariate normals, that is

$$p(x) = \frac{1}{2}\left[p_1(x) + p_4(x)\right]$$

where $p_1(x)$ and $p_4(x)$ are bivariate normal with mean $\mu^{(1)}$ and $\mu^{(4)}$ and covariance matrices $\Sigma^{(1)}$ and $\Sigma^{(4)}$. The probability density functions for π_2 and π_3 were selected as bivariate normal with means $\mu^{(2)}$ and $\mu^{(3)}$. The covariance matrices are $\Sigma^{(2)}$ and $\Sigma^{(3)}$. This data for the four cases studied are listed in Table 3.

CASE	i	N_i	MEAN VECTOR	THE ELEMENTS OF THE COVARIANCE MATRIX
1	1	20(20)100	(0,0)	$\sigma_{11} = 1, \sigma_{22} = 1, \sigma_{12} = 0$
	2	20(20)100	(.5,1)	$\sigma_{11} = .6, \sigma_{22} = .9, \sigma_{12} = 0$
	3	20(20)100	(1.6,.2)	$\sigma_{11} = 1.5, \sigma_{22} = .9, \sigma_{12} = .4$
	4	20(20)100	(1,2)	$\sigma_{11} = 1, \sigma_{22} = 1, \sigma_{12} = 0$
2	1	20(20)100	(1.5,-1)	$\sigma_{11} = 1, \sigma_{22} = 1.5, \sigma_{12} = .4$
	2	20(20)100	(0,0)	$\sigma_{11} = 1, \sigma_{22} = 1, \sigma_{12} = 0$
	3	20(20)100	(.3,1)	$\sigma_{11} = .6, \sigma_{22} = 1, \sigma_{12} = .2$
	4	20(20)100	(.5,0)	$\sigma_{11} = 1, \sigma_{22} = 1, \sigma_{12} = 0$
3	1	20(20)100	(-2,0)	$\sigma_{11} = 1, \sigma_{22} = 1, \sigma_{12} = 0$
	2	20(20)100	(.5,1)	$\sigma_{11} = 1.5, \sigma_{22} = 1, \sigma_{12} = .2$
	3	20(20)100	(-1,-.2)	$\sigma_{11} = 1, \sigma_{22} = 1, \sigma_{12} = 0$
	4	20(20)100	(0,.1)	$\sigma_{11} = .9, \sigma_{22} = 1, \sigma_{12} = .1$
4	1	20(20)100	(1.3,2)	$\sigma_{11} = 1, \sigma_{22} = 1, \sigma_{12} = .4$
	2	20(20)100	(.2,.5)	$\sigma_{11} = 1, \sigma_{22} = 1, \sigma_{12} = 0$
	3	20(20)100	(3,1)	$\sigma_{11} = 1.9, \sigma_{22} = 1, \sigma_{12} = .4$
	4	20(20)100	(0,0)	$\sigma_{11} = 1.8, \sigma_{22} = 2.4, \sigma_{12} = .6$

Table 3. Input Data for the Second Simulation

In this part of the study we include the technique in which it is assumed each $p_i(x)$ $i = 1,2,3$ is normal and from the samples estimate the parameters given in turn estimates of the density $p_i(x;\mu,\Sigma)$ by $p_i(x;\hat{\mu},\hat{\Sigma})$ where $\hat{\mu}$ and $\hat{\Sigma}$ are the usual maximum likelihood estimators [4, Chapter 3] for μ and Σ. Samples of size 20, 40, 60, 80 and 100 were generated to estimate the probability densities $p_i(x)$, $i = 1,2,3$. An additional 100 observations were

generated from the first population (the bimodal population) and the classification performed with the four techniques. The results are listed in Table 4a, 4b, 4c and 4d.

SAMPLE SIZE	TECHNIQUE	π_1	π_2	π_3	NO DECISION
20	Kendall	15	2	6	77
	Optimal	40	47	13	0
	Non-Parametric	46	43	11	0
	Assumed Norm	30	58	12	0
40	Kendall	12	0	0	88
	Optimal	40	56	17	0
	Non-Parametric	27	56	17	0
	Assumed Norm	16	70	14	0
60	Kendall	19	0	2	79
	Optimal	40	47	13	0
	Non-Parametric	45	42	13	0
	Assumed Norm	28	64	8	0
80	Kendall	16	0	1	83
	Optimal	40	47	13	0
	Non-Parametric	37	47	16	0
	Assumed Norm	31	56	13	0
100	Kendall	16	0	0	84
	Optimal	40	47	13	0
	Non-Parametric	32	55	13	0
	Assumed Norm	20	65	15	0

Table 4a. The Comparison, Case I.

SAMPLE SIZE	TECHNIQUE	R_1	R_2	R_3	NO DECISION
20	Kendall	22	8	0	70
	Optimal	51	25	24	0
	Non-Parametric	47	30	23	0
	Assumed Norm	47	19	34	0
40	Kendall	9	0	0	91
	Optimal	51	25	24	0
	Non-Parametric	49	37	14	0
	Assumed Norm	25	51	24	0
60	Kendall	15	0	1	84
	Optimal	51	25	24	0
	Non-Parametric	51	19	30	0
	Assumed Norm	40	30	30	0
80	Kendall	5	3	0	92
	Optimal	51	25	24	0
	Non-Parametric	46	33	21	0
	Assumed Norm	42	29	29	0
100	Kendall	12	2	0	86
	Optimal	51	25	24	0
	Non-Parametric	50	31	19	0
	Assumed Norm	46	26	28	0

Table 4b. The Comparison, CaseII.

SAMPLE SIZE	TECHNIQUE	R_1	R_2	R_3	NO DECISION
20	Kendall	16	3	0	81
	Optimal	37	20	43	0
	Non-Parametric	20	26	54	0
	Assumed Norm	12	13	75	0
40	Kendall	3	2	0	95
	Optimal	37	20	43	0
	Non-Parametric	34	24	42	0
	Assumed Norm	16	32	52	0
60	Kendall	5	1	1	93
	Optimal	37	20	43	0
	Non-Parametric	40	25	35	0
	Assumed Norm	24	26	50	0
80	Kendall	5	4	0	91
	Optimal	37	20	43	0
	Non-Parametric	26	21	53	0
	Assumed Norm	18	24	58	0
100	Kendall	5	3	0	92
	Optimal	37	20	43	0
	Non-Parametric	34	26	40	0
	Assumed Norm	17	28	55	0

Table 4c. The Comparison, Case III

SAMPLE SIZE	TECHNIQUE	R_1	R_2	R_3	NO DECISION
20	Kendall	13	1	0	86
	Optimal	44	52	4	0
	Non-Parametric	39	49	12	0
	Assumed Norm	28	59	13	0
40	Kendall	19	0	0	81
	Optimal	44	52	4	0
	Non-Parametric	32	63	5	0
	Assumed Norm	23	69	8	0
60	Kendall	14	0	0	86
	Optimal	44	52	4	0
	Non-Parametric	47	45	8	0
	Assumed Norm	35	55	10	0
80	Kendall	10	0	0	90
	Optimal	44	52	4	0
	Non-Parametric	40	48	12	0
	Assumed Norm	23	64	13	0
100	Kendall	17	0	0	83
	Optimal	44	52	4	0
	Non-Parametric	38	57	5	0
	Assumed Norm	27	65	8	0

Table 4d. The Comparison, Case IV.

We note that the comparison remains relatively invariant under changing sample sizes. The non-parametric technique remains nearly as effective as the optimal technique. Kendall's technique still gives large numbers of no-decision results.

5. CONCLUDING REMARKS

We have considered here only those measurements that are quantitative. Discriminating techniques which include quantal or qualitative data have been ignored. However, the concepts and techniques remain valid. For a discussion of problems associated with such data the interested reader is referred to [13], [14], and [15].

A valid way to evaluate the techniques discussed in this paper would be to compute probabilities of misclassification under perhaps normality assumptions. This is indeed a difficult problem since integration over the regions $R_1,\ldots,R_m$ can be very complicated [16][17]. However, techniques employed by Lachenbruch [18], Lachenbruch and Mickey [19], and Dunn and Vardy [20] can be mimicked to obtain estimates of probabilities of misclassification.

6. REFERENCES

[1] Kendall, M.G. and Stuart A., The Advanced Theory of Statistics, Hafner, (1966), Vol. 3, pp. 314-341.

[2] Kendall, M.G., Discrimination and Classification, Proc. of WAFB Symposium on Multivariate Analysis, Academic Press, New York, (1966), pp. 165-185.

[3] Fix, E and Hodges, J.L., Discriminatory Analysis, Project Report 21-49-004, No. 4 and No. 11, USAF School of Aviation Medicine, Randolph Field, San Antonio, Texas.

[4] Anderson, T.W., Introduction to Multivariate Statistical Analysis, Wiley and Sons. (1958), pp. 126-152.

[5] Kendall, M.G., A Course in Multivariate Analysis, Griffith and Co., (1957), pp. 144-179.

[6] Wilks, S.S., Mathematical Statistics, Wiley and Sons, (1962), pp. 573-587.

[7] Kabe, D.G., Some Results on the Distribution of Two Random Matrices Used in Classification Procedures, Annals of Math. Statis., Vol. 36, (1965), pp. 181-185.

[8] Parzen, E., On Estimation of a Probability Density Function and Mode, Annals of Math. Statis., Vol. 33, (1965), pp. 1065-1976.

[9] Cacoullos, T., Estimation of a Multivariate Density, Annals of the Inst. of Stat. Math., Vol. 18, (1966), pp. 174-183.

[10] Martz, H.F., Emperical Bayes Estimation in Multiple Linear Regression, Doctorate Dissertation, Virginia Polytechnic Institute, (1967), pp. 28-42.

[11] Hammersley, J.M., and Handscomb, Monte Carlo Methods, Methuen, (1964).

[12] Spanier, J. and Gelbard, E.M., Monte Carlo Principles and Neutron Transport Problems, Addison-Wesley, (1969), pp. 1-36.

[13] Solomon, H., Classification Procedures Based on Dichotomous Response Vector, Contributions to Probability and Statistics, Edited by Olkin and others, Stanford Univ. Press, (1960), pp. 414-423.

[14] Linhart, H., Techniques for Discriminant Analysis with Discreet Variables, Metrika, Vol. 2, (1959), pp. 138-149.

[15] Cochran, W.G., and Hopkins, C.E., Some Classification Problems with Multivariate Qualitative Data, Biometrics, Vol. 17, (1961), pp. 10-32.

[16] Owen, D.B., Tables for Computing Bivariate Normal Probabilities, Annals of Math. Stat., (1956), 27, No. 4, pp. 1075-1090.

[17] Owen, D.B., Bibliography on the Multivariate Normal Integrals and Related Topics, Annals of Math. Statis., Vol. 34, No. 3, (1963), pp. 829-838.

[18] Lachenbruch, P.A., An Almost Unviased Method of Obtaining Confidence Intervals for the Probability of Misclassification in Discriminant Analysis, Biometrics, Vol. 23, (1967), pp. 639-645.

[19] Lachenbruch, P.A., and Mickey, M.R., Estimation of Error Rates in Discriminant Analysis, Technometrics, Vol. 10, (1968), pp. 1-11.

[20] Dunn, O.J., and Vardy, P.D., Probabilities of Correct Classification in Discriminant Analysis, Biometrics, Vol. 22, (1966), pp. 908-924.

[21] Bradley, J.V., Distribution-Free Statistical Tests, Prentice-Hall, Inc., (1968).

Lecture II.

ON COMPUTING MINIMAX PROCEDURES IN DISCRIMINATION ANALYSIS[1]

P. L. Odell[2] and T. G. Newman[2]

1. Introduction

In his text [1] Anderson outlines the theory of statistical discriminate analysis based on a Bayes and on a minimax criteria of evaluation. For completeness we state the problems.

Problem I. (The Bayesian Discriminate Problem) Let $\pi_1, \pi_2, \ldots, \pi_m$ denote m distinct p-variate populations whose multivariate probability density functions $p_1(x), p_2(x), \ldots, p_m(x)$ are known. Let $q_1, q_2, \ldots, q_m$ be the known a priori probabilities that a sample is selected from population $\pi_1, \pi_2, \ldots, \pi_m$, respectively. Let $c(i|j)$ be the cost of misclassifying an individual from population π_j as being from population π_i such that

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

Given a sample X selected at random from one of the populations give a decision rule R which minimized the expected cost of misclassification for assigning individuals to i^{th} population π_i $i = 1,2,\ldots,m$. Such a rule R is called a Bayesian procedure.

A decision rule R for assigning an individual to one of the populations π_i $i = 1,2,\ldots,m$ which minimizes the expected cost of

---

1 This research was supported in part by NASA-Manned Spacecraft Center under Contract NAS-9-6963.

2 Texas Tech University, Lubbock, Texas 79409

misclassification c, where

$$c = \sum_{i=1}^{m} q_i \left\{ \sum_{\substack{j=1 \\ j\neq i}}^{m} c(j|i)P(j|i,R) \right\} \tag{1}$$

where

$$P(j|i,R) = \int_{R_j} p_i(x)dx \tag{2}$$

is the probability that X belongs to R_j given it is actually from π_i, where the procedure R is a partition of the sample space into m mutually exclusive and exhaustive regions $R_1, R_2, \ldots, R_m$. The following theorem proved in [1] summarizes an approach to compute the optimal procedure, R.

Theorem 1. The procedure R, or equivalently, the regions of classification $R_1, R_2, \ldots, R_m$, that minimizes the expected cost of misclassification (1) are defined by assigning X to R_k if

$$\sum_{\substack{i=1 \\ i\neq k}}^{m} q_i p_i(x) c(k|i) \quad < \quad \sum_{\substack{i=1 \\ i\neq j}}^{m} q_i p_i(x) c(j|i) \qquad j = 1,2,\ldots,m. \tag{3}$$

We use the notation $R = (R_1, \ldots, R_m)$ to be a set of disjoint but not necessarily connected regions whose union is the total Euclidean p-space, the sample space. The partition R constructed as indicated in Theorem 1 is called a Bayes procedure.

It is important to note that one must know a great amount in order to apply Theorem 1. Unfortunately in practice, there are cases in which the a priori probabilities q_i are unknown. If $c(j|i)$ are unknown or not assumed equal, then the problem is not

well-posed. If the set $q = \{q_1,\ldots,q_m\}$ are unknown, then a strategy for selecting a decision procedure can be based on the minimax criteria. One can define this problem as

Problem II. (The Minimax Discriminate Problem) Let the conditions in Problem I remain true, except that the a priori probabilities are not known. Given a sample X selected at random from one of the populations $\pi_1,\ldots,\pi_m$ give a decision rule that will minimize the maximum expected loss, where

$$r(i,R) = \sum_{\substack{j=1 \\ j\neq i}}^{m} c(j|i)P(j|i,R) \qquad i,j = 1,2,\ldots,m \tag{4}$$

is the expected loss if observation is from π_i and assigned to π_j.

Von Mises [2] considered Problem II and observed that the partition $R = \{R_1,\ldots,R_m\}$ that solves Problem II is characterized by two properties, (i) the probabilities of correct classification

$$P(i|i,R) = \int_{R_i} p_i(x)dx. \tag{5}$$

are equal for all $i = 1,2,\ldots,m$, and (ii), on the border of R_i and R_j, the ratio $p_i(x)/p_j(x)$ is constant. The value of the constant is simply

$$q_j c(i|j)/q_i c(i|j). \tag{6}$$

In [1] this problem is solved when $m = 2$, $p_i(x)$ $i = 1,2$ are normal with identical covariance matrices. Also, Problem II is discussed in general terms for $m = 3$. However, there exist no general algorithms available for computing the regions R for $m > 2$ and for densities other than normal.

It is the purpose of this paper to discuss the problems associated with the actual computing of the procedure $R = (R_1,\ldots,R_m)$ which solves Problem II. Theorem 1 gives the solution to Problem I and little computational difficulties exist.

2. Fundamental Concepts [1]

Suppose that we are confronted with Problem II, that is we do not have a priori probabilities. Hence, c in (1) cannot be defined. One can define expected cost of misclassification on the condition that the observation comes from a given population. The expected cost of misclassifying an observation given that the sample actually came from π_i with respect to a procedure R is given by (4).

A procedure R is at least as good as a competing procedure R^* if

$$r(i,R) \leq R(i,R^*) \qquad i = 1,2,\ldots,m \ , \tag{7}$$

and R is better if at least one inequality is strict. R is said to be admissible, if there is no procedure R^* that is better. A class of procedures is complete if for every procedure R^* outside the class, there is a procedure R in the class that is better. A minimal complete class is a complete class such that no proper subset is a complete class.

The following Theorems summarize well-known [1] facts.

Theorem 2. If $q_i > 0$ $(i = 1,2,\ldots,m)$, then every Bayes procedure is admissible.

Theorem 3. If $\Pr\{p_i(x) \neq 0 | \pi_j\} = 0$, then a Bayes procedure is admissible.

Theorem 4. If $\Pr\{P_i(x)/p_j(x) = k|\pi_h\} = 0 \quad i \neq j \quad 0 \leq k < \infty$, then every admissible procedure is a Bayes procedure.

Theorem 5. If $\Pr\{p_i(x)/p_j(x) = k|\pi_k\} = 0 \quad i \neq j \quad 0 \leq k < \infty$, the class of Bayes procedures is minimal complete.

Let

$$\alpha_i(R) = 1 - P(i|i,R) \tag{8}$$

be the probability of making a wrong decision using procedure R and sampling from π_i. When R is a Bayes procedure $\alpha_i(R)$ is a function of $q_1, q_2, \ldots, q_m$ in the following manner.

$$\alpha_i(q_1,\ldots,q_m;R) = 1 - \Pr\left[\frac{P_1(x)}{P_i(x)} \leq \frac{q_i C(1|i)}{q_1 C(i|1)}\right.$$

$$\frac{P_2(x)}{P_i(x)} \leq \frac{q_i C(2|i)}{q_2 C(i|2)}; \ldots; \frac{P_{i-1}(x)}{P_i(x)} \leq \frac{q_i C(i-1|i)}{q_{i-1} C(i|i-1)};$$

$$\left.\frac{P_{i+1}(x)}{P_i(x)} \leq \frac{q_i C(i+1|i)}{q_{i+1} C(i|i+1)}; \ldots; \frac{P_m(x)}{P_i(x)} \leq \frac{q_i C(m|i)}{q_m C(i|m)}\right] \tag{9}$$

We now consider the minimax problem. There is a Bayes solution for which

$$\alpha_1 = \alpha_2 = \ldots = \alpha_m \tag{10}$$

for the totality of points for which $\alpha_1 = \alpha_2 = \ldots = \alpha_{m-1}$ is connected and includes point $\alpha_m = 1$ and for which $\alpha_m = 0$. By continuity there is a point for which (10) holds. Since this procedure is admissible, there is no other procedure which has smaller maximum probability of error. Thus (10) gives the minimax procedure. The quantities (9) and (10) give Von Mises conclusion (ii) and (i), respectively.

3. An Algorithm

If $p_1(x), p_2(x), \ldots, p_m(x)$ are known, and the joint probability density functions of the ratios

$$p_i(x)/p_j(x) \qquad i \neq j = 1,2,\ldots,m \tag{11}$$

can be obtained analytically, one can write in integral form the quantities $\alpha_1, \alpha_2, \ldots, \alpha_m$ as functions of $q_1, q_2, \ldots, q_{m-1}$ since

$$q_m = 1 - \sum_{L=1}^{m-1} q_i \ . \tag{12}$$

Define

$$\begin{aligned} f_1(q_1 \ldots q_{m-1}) &= \alpha_1 - \alpha_2 \\ f_2(q_{11} \ldots q_{m-1}) &= \alpha_2 - \alpha_3 \\ &\vdots \\ f_{m-1}(q_1 \ldots q_{m-1}) &= \alpha_{m-2} - \alpha_{m-1} \ . \end{aligned} \tag{13}$$

In matrix notation then (13) can be written as

$$f = P\alpha$$

where

$$f = \begin{bmatrix} f_1 \\ \vdots \\ f_{m-1} \end{bmatrix} ; \ P = \begin{bmatrix} 1 & -1 & 0 & 0 & 0 \\ 0 & 1 & -1 & 0 & 0 \\ \vdots & & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & -1 & 0 \\ 0 & 0 & 0 \ldots & 1 & -1 \end{bmatrix} \quad \alpha = \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_m \end{bmatrix}$$

We note that P is (m-1) x m matrix whose rank is m-1. Also we note that if $q^* = (q_1^*, \ldots, q_{m-1}^*)$ such that $f \equiv \phi$, then $\alpha_1 = \alpha_2 = \ldots = \alpha_m$, and the procedure $R = (R_1, \ldots, R_m)$ defined by (3) in Theorem 1, with each q_i replaced by q_i^* $i = 1,2,\ldots,m-1$ and q_m^* is given by

(12). Hence, we have reduced Problem II to solving $f = \phi$ for q^*. Unfortunately, it cannot be assured in general that q^* is unique in the sense that (10) holds. Suppose that q and q^* are such that

$$\alpha_1(q) = \alpha_2(q) = \ldots = \alpha_m(q) = \alpha(q)$$

and

$$\alpha_1(q^*) = \alpha_2(q^*) = \ldots \alpha_m(q^*) = \alpha(q^*)$$

where

$$\alpha(q) > \alpha(q^*)$$

Then q^* would lead to the minimax procedure. Except in pathelogical cases one would expect a finite number of solutions for $f = P\alpha = \phi$. One must remember the solution $q_1^*, q_2^*, \ldots, q_{m-1}^*$ are such that

$$0 < q_i^* < 1 \qquad i = 1,\ldots,m-1 \; . \tag{15}$$

That is, we solve for all solutions q^* such that

$$P\alpha = 0 \tag{16}$$

and (15) is true.

One can use the following iteration formula for solving for q^*

$$q^N = q^{N-1} - [f'(q^{N-1})]^{-1} f(q^{N-1}) \tag{17}$$

where the (m-1) x (m-1) matrix

$$\left| f'(q^{N-1}) \right| = \left\{ \frac{\partial f_i}{\partial q_j} \right\}_{q=q^{N-1}} \tag{18}$$

and the elements $\partial f_i / \partial q_j$ for some cases must be approximated using one of the standard formulas for numerical differentiation [2]. One

such formula is

$$\frac{\partial f_i}{\partial q_j} = \frac{1}{2\Delta} \left| f_i(q_j^{N-1} + \Delta) - f_i(q_j^{N-1} - \Delta) \right| \tag{19}$$

where Δ is an incremental value of q_j and

$$f_i(q_j^{N-1} + \Delta) = f_i(q_1^{N-1}, q_2^{N-1}, \ldots, q_j^{N-1} + \Delta, \ldots, q_{m-1}^{N-1})$$

$$f_i(q_j^{N-1} - \Delta) = f_i(q_1^{N-1}, q_2^{N-1}, \ldots, q_j^{N-1} - \Delta, \ldots, q_{m-1}^{N-1})$$

During the iteration if for some k, $q_i^k > 1$ one replaces that element with unity and continues to iterate. If for some k $q_i^k < 0$, one replaces that element with zero. These rules will assure that condition (15) holds. If $q^N = q^{N+1}$, one assumes that

$$q^N = q^*$$

the desired solution.

We will summarize the process in the following.

Algorithm. Let

$$q^{N-1} = (q_1^{N-1}, q_2^{N-1}, \ldots, q_{m-1}^{N-1})$$

be an approximate solution of (16) such that (15) holds, then

a. Compute the Bayes procedure $R^{N-1} = \{R_i^{N-1};\ i = 1,2,\ldots,m\}$ using Theorem 1.

b. Using the probability densities $p_i(x)$ $i = 1,2,\ldots,m$, compute the sets

$$P(i|i,R^{N-1}) = \int_{R_i^{N-1}} P_i(x)\,dx$$

and $f_i(q_1^{N-1},\ldots,q_{m-1}^{N-1})$ $\quad i = 1,2,\ldots,m-1.$

Using the sets

$$Q_1 = \{q_1^{N-1} \neq \Delta_1, q_2^{N-1},\ldots,q_{m-1}^{N-1}\}$$

$$Q_2 = \{q_1^{N-1}, q_2 \neq \Delta_2,\ldots,q_{m-1}^{N-1}\}$$

$$\vdots$$

$$Q_{m-1} = \{q_1^{N-1}, q_2,\ldots,q_{m-1}^{N-1} \neq \Delta_{m-1}\}$$

compute the set

$$f_i(q_j \pm \Delta_j) \quad i,j = 1,2,\ldots,m-1$$

d. Compute the matrix $[\partial f_i/\partial q_j]_{q=q^{N-1}}$ using (19) or some other appropriate formula for numerical differentiation if necessary.

e. Compute q_N^N using (17).

f. Check if $0 < q_i^N < 1$ for all i and replace those elements less than zero by zero and those greater than unity by unity.

g. Replace q^{N-1} by q^N and repeat the process.

h. When $q^N = q^{N-1}$, assume that $q^* = q_N$ and compute the minimax decision rule R^* by computing the Bayesian procedure using Theorem 1.

A reasonable choice of $q^{(1)}$ would be

$$q_1^1 = q_2^1 = \ldots = q_m^1 = 1/m.$$

Also one should check

$$\alpha_1(q^*) = \alpha_2(q^*) = \ldots = \alpha_m(q^*)$$

or equivalently

$$P(1|1,R^*) = P(2|2,R^*) = \ldots = P(m|m,R^*).$$

4. An Example

Let $m = 3$, that is, we are to classify an observation x as being from one of the populations π_1, π_2, or π_3. Let the observation be bivariate, that is $p_1(x)$, $p_2(x)$ and $p_3(x)$ are bivariate normals. Let the costs $C(i|j)$ be unity for all i and j.

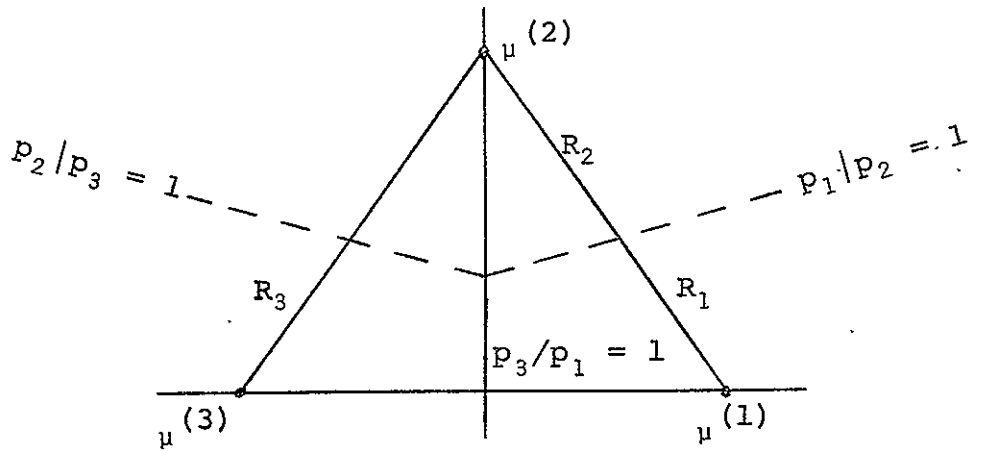


Fig. 1. The First Approximation to the Minimax Solution

Let $q_1, q_2, q_3 = 1 - q_1 - q_2$ be unknown probabilities that an observation comes from π_1, π_2 and π_3, respectivelv. Let $p_i(x)$ be bivariate normal with the following parameters

$$\mu^{(1)} = (1,0)^T \qquad \Sigma^{(1)} = \delta^2 I$$

$$\mu^{(2)} = (0,2)^T \qquad \Sigma^{(2)} = \delta^2 I$$

and

$$\mu^{(3)} = (-1,0) \qquad \Sigma^{(3)} = \delta^2 I$$

where δ^2 is a known scalar. Then

$$R_1 = \{(x_1,x_2);\quad x_2 \le \tfrac{1}{2}x_1 + (\tfrac{1}{2}\delta^2 \ln q_2/q_1 + 3/4)$$

$$\text{and} \qquad x_1 \ge \tfrac{1}{2}\delta^2 \ln[(1 - q_2 - q_1)/q_1]$$

$$R_2 = \{(x_1,x_2);\quad x_2 \ge \tfrac{1}{2}x_1 + [\tfrac{1}{2}\delta^2 \ln q_2/q_1 + 3/4]$$

$$\text{and} \qquad x_2 \ge -\tfrac{1}{2}x_1 + [\tfrac{1}{2}\delta^2 \ln(1 - q_1 - q_2)/q_1 + 3/4]$$

$$R_3 = \{(x_1,x_2);\quad x_2 \le -\tfrac{1}{2}x_1 + [\tfrac{1}{2}\delta^2 \ln q_2/q_1 + 3/4]$$

$$\text{and} \qquad x_1 \le \tfrac{1}{2}\delta^2 \ln[(1 - q_1 - q_2)/q_1]$$

We note that from the geometry of our example that if $q_1=q_2=q_3=1/3$ that the regions R_1, R_2, R_3 are not minimax. Yet these values seem to be reasonable first estimates. R_1, R_2, and R_3 are computed by letting

$\ln q_2/q_1 = 0$, $\ln(1 - q_1 - q_2)/q_1 = 0$, and $\ln(1 - q_1 - q_2)/q_2 = 0$.

We can write in general

$$P(1|1;R) = \int_{\frac{1}{2}\delta^2 \ln q_3/q_1}^{\infty} \int_{-\infty}^{\frac{1}{2}x_1 + [\frac{1}{2}\delta^2 \ln q_2/q_1 + 3/4]} p_1(x_1,x_2)\,dx_1 dx_2$$

$$P(2|2,R) = \int_{\frac{1}{2}\delta^2 \ln q_2/q_1 + 3/4}^{\infty} \int_{-2x_2 - [\frac{1}{2}\delta^2 \ln q_3/q_1 + 3/4]}^{2x_2 - [\frac{1}{2}\delta \ln q_2/q_1 + 3/4]} p_2(x_1,x_2)\,dx_1 dx_2$$

and

$$P(3|3,R) = \int_{-\infty}^{\frac{1}{2}\delta^2 \ln q_3/q_1} \int_{-\infty}^{-\frac{1}{2}x_1 + [\frac{1}{2}\delta^2 \ln q_2/q_1 + 3/4]} p_3(x_1,x_2)dx_1dx_2 \ .$$

Then

$$f_1(q_1,q_2) = P(1|1,R) - P(2|2,R) = 0$$

$$f_2(q_1,q_2) = P(2|2,R) - P(3|3,R) = 0$$

where $q_3 = 1 - q_1 - q_2$.

Using Leibnitz's rule [3] for differentiation of an integral, one can obtain the matrix $\{\partial f_i/\partial q_j\}$ as a function of q_1 and q_2. Using our initial guess for q_1, q_2 one can proceed to compute iteratively the solution, say q_1^* and q_2^*. These values q_1^* and q_2^* using Theorem 1 give the minimax regions R_1^*, R_2^*, and R_3^*. These regions are R_1, R_2, and R_3 of this section with q_1, q_2, q_3 replaced by q_1^*, q_2^* and $1 - q_1^* - q_2^*$.

5. References

[1] Anderson, T.W., An Introduction to Multivariate Statistical Analysis, Wiley and Sons (1958).

[2] Korn, G.A., DIGITAL COMPUTER USES'S HANDBOOK, McGraw-Hill, (1967), pp. 2-111 to 2-116.

[3] Taylor, A.E., ADVANCED CALCULUS,Ginn and Company, (1955), p. 523.

Lecture III.

WILK'S SCATTER TECHNIQUES AND STATISTICAL DISCRIMINATION[1]

P. L. Odell[2] and T. G. Newman[2]

1. Introduction

In his text [1] Wilks discusses a concept of multidimensional scatter [2] and implies that the concept can lead to ways for reducing the dimension of some statistical discriminate problems. Sebestyan [3], in developing a technique for recognizing patterns, has developed a theory similar to the theory of scatter developed by Wilks and apparently has had some affect on engineering applications [4], [5].

The discriminate problem of interest in this paper can be described as follows:

The Discriminate Problem. Given the distinct populations $\pi_1, \pi_2, \ldots, \pi_m$ $m \geq 2$ and m sets of p-dimensional observations

$$\{X_j^{(i)}, \quad j = 1,2,\ldots,N_i\}, \qquad i = 1,2,\ldots,m \tag{1}$$

each from one of the populations, formulate a "good" decision rule to assign an individual from one of the populations by using the data contained in an observation X made on that individual.

Several techniques [6], [7], [8] are available to solve this

---

1 This research was supported in part by NASA - Manned Spacecraft Center under contract NAS 9-6963.

2 Texas Tech University, Lubbock, Texas 79409.

problem. These formulations either minimize expected cost of misclassification or minimize maximum expected cost of misclassification. So called non-parametric techniques are discussed in [8], [9], and [10]. These techniques are all applicable to the data after one has reduced the dimensions by techniques discussed in this paper.

2. The Problem

We define the i^{th} sample scatter matrix defined by

$$S_i = \sum_{j=1}^{N_i} (X_j^{(i)} - \bar{X}^{(i)})(X_j^{(i)} - \bar{X}^{(i)})^T \tag{2}$$

where $$\bar{X}^{(i)} = \sum_{j=1}^{N_i} X_j^{(i)}/N_i$$

The determinate of the matrix

$$S_w = \sum_{i=1}^{m} S_i \tag{3}$$

is called the <u>within</u> scatter of the m samples. The determinate of the matrix

$$S_p = \sum_{i=1}^{m} \sum_{j=1}^{N_i} (X_j^{(i)} - \bar{X})(X_j^{(i)} - \bar{X})^T \tag{4}$$

where $$\bar{X} = \sum_{i=1}^{m} \sum_{j=1}^{N_i} X_j^{(i)}/(N + N + \ldots + N_m) \tag{5}$$

is called the <u>pooled</u> scatter. It is easy to show that

$$S_p = S_w + S_B \tag{6}$$

where the determinate of the matrix

$$S_B = \sum_{i=1}^{m} N_i (\bar{X}^{(i)} - \bar{X})(\bar{X}^{(i)} - \bar{X})^T \tag{7}$$

is called the between scatter. It is important to note that the rank of the matrix S_B is the maximal value of p and m (almost surely).

In order to reduce the size we introduce a r x p matrix C of rank r and such that if the vector

$$Z_j^{(i)} = CX_j^{(i)} \tag{8}$$

is defined for $j = 1,2,\ldots,N_i$ and $i = 1,2,\ldots,m$. We select the rows of C in such a manner so that $S_B(Z)$ is minimal and $S_w(Z)$ remains constant in such a way that

$$\frac{|S_w|}{|S_p|} = \frac{|S_w|}{|S_w+S_B|} = \frac{|S_w(Z)|}{|S_w(Z)+S_B(Z)|} = \frac{|S_w(Z)|}{|S_p(Z)|} \tag{9}$$

where $S_{(\cdot)}(Z)$ denotes $S_{(\cdot)}$ with the X's replaced by the Z's as defined by (8). That is, one selects C in such a manner that

$$Q = |S_w(Z)| + [K - |S_B(Z)|]\lambda \tag{10}$$

is minimum where K is a constant. Let r = 1, then C is a 1 x p matrix. Then (10) reduces to

$$Q = CS_wC^T + [K - CS_BC]\lambda$$

since

$$S_B(Z) = CS_BC^T \tag{11}$$

$$S_w(Z) = CS_wC^T \tag{12}$$

Taking the derivative of Q with respect to the elements of C^T and equating with the p x 1 zero vector $\emptyset$ yields necessary conditions for a maximum

$$(S_W - \lambda S_B)C^T = 0 \tag{13}$$

and

$$K - CS_W C^T = 0 \tag{14}$$

It is well-known [11] that (13) has a non-zero solution only if

$$|S_W - \lambda S_B| = 0 \tag{15}$$

Let P be a non-singular matrix (P exists [11]) such that

$$P^T S_W P = I, \tag{16}$$

the p x p identity matrix, and

$$P^T S_B P = \begin{bmatrix} \lambda_1 & & & & \Phi \\ & \lambda_2 & & & \\ & & \ddots & & \\ & & & r_0 & \\ & & & & \ddots \\ \Phi & & & & 0 \end{bmatrix} \tag{17}$$

where

$$\lambda_1 > \lambda_2 > \ldots > \lambda_r > 0 \tag{18}$$

the characteristic roots in (15). Note that S_B is almost surely positive semi-definite and S_W is almost surely positive definite. One should note that

$$|P^T(S_B - \lambda S_W)P| = |P^T S_B P - \lambda P^T S_W P|$$

$$= |L - \lambda I|$$

where

$$L = \begin{bmatrix} \lambda_1 & & & & \Phi \\ & \lambda_2 & & & \\ & & \ddots & & \\ & & & \lambda_{r_0} & \\ & & & & \ddots \\ \Phi & & & & 0 \end{bmatrix}$$

or

$$|P^T||S_B - \lambda S_w||P| = \lambda^{P-S}(\lambda_1 - \lambda) \ \ldots \ (\lambda_r - \lambda),$$

and finally

$$|S_B - \lambda S_w| = \frac{\lambda^{P-S}(\lambda_1 - \lambda) \ \ldots \ (\lambda_r - \lambda)}{|P|^2} \tag{19}$$

Let C_j^T be the j^{th} characteristic vector associated with the characteristic root λ_j, that is

$$(S_B - \lambda_j S_w)C_j = \phi \qquad j = 1,2,\ldots,r$$

and

$$C_j S_w C_j^T = K,$$

the constant.

Let $C_1^T, C_2^T, \ldots, C_r^T$ be the characteristic vectors associated with the characteristic roots $\lambda_1, \ldots, \lambda_r$. The matrix

$$C = \begin{bmatrix} C_1 \\ C_2 \\ \vdots \\ C_j \end{bmatrix}$$

is the matrix that reduces the dimensions from a p-dimensional to one of only $j \leq r \leq p$ dimensions.

From (13) and (14) it follows that

$$C_j S_B C_j^T = \lambda_j C_j S_w C_j^T$$

$$= \lambda_j K. \tag{20}$$

Also if $C_i \neq C_j$, then from (13)

$$C_i S_B C_j^T - \lambda_j C_i S_w C_j^T = 0$$

and

$$C_j S_B C_i^T - \lambda_i C_j S_w C_i^T = 0.$$

From this we subtract and note that $C_i S_w C_j^T = C_j S_w C_i^T$. We find that

$$(\lambda_j - \lambda_i) C_i S_w C_j^T = 0$$

which implies that

$$C_i S_w C_j^T = \emptyset \tag{21}$$

since $\lambda_j \neq \lambda_i$, which in turn implies

$$C_i S_B C_j^T = 0. \tag{22}$$

Also

$$S_w(Z) = C S_w C^T \tag{23}$$

$$= \begin{bmatrix} K & 0 & 0 \\ 0 & K & 0 \\ 0 & 0 & K \end{bmatrix}$$

$$= \text{Dia}\{K\}$$

is a r x r diagonal matrix with the constant K as the common diagonal element. The statements (20), (21), (22) and (23) lead to the result that

$$S(Z) = S_w(Z) + S_B(Z)$$

$$= \text{Dia}\{K\} + \text{Dia}\{\lambda_j K\}$$

or

$$|S(Z)| = K^r(1 + \lambda_1)(1 + \lambda_2)\ldots(1 + \lambda_r)$$

$$|S_w(Z)| = K^r$$

Hence

$$\frac{|S_w(Z)|}{|S(Z)|} = \frac{1}{(1 + \lambda_1)(1 + \lambda_2) \ldots (1 + \lambda_r)}$$

Finally we note that (16) implies that

$$|S_w| = 1/|P|^2$$

and

$$|P^T S_w P + P^T S_B P| = |I + L|$$

$$= (1 + \lambda_1)(1 + \lambda_2)\ldots(1 + \lambda_r)$$

implies

$$|S_w + S_B| = \frac{(1 + \lambda_1)(1 + \lambda_2)\ldots(1 + \lambda_r)}{|P|^2}$$

which results in

$$\frac{|S_w|}{|S|} = \frac{1}{(1 + \lambda_1)(1 + \lambda_2)\ldots(1 + \lambda_r)} = \frac{|S_w(Z)|}{|S(Z)|} . \qquad (24)$$

In summary, we selected C so that the between scatter for Z's are maximum yet (24) holds. Note that we can force the dimension from p to 1 by selecting λ_1 the largest characteristic root.

3. A Proposed Numerical Simulation

In order to clarify what we have described in section 2, consider the following simulation plan: Let π_1, π_2, π_3 be three populations with associated observations $X^{(1)}$, $X^{(2)}$, and $X^{(3)}$ such that

$$X^{(i)} \sim N[e_i, I] \qquad i = 1,2,3 \qquad p = 1,2,3,4$$

where

$$e_1 = (a,0,0,0)^T$$

$$e_2 = (0,a,0,0)^T$$

and

$$e_3 = (0,0,a,0)^T$$

One then generates three samples, one from each of the populations by the usual methods [12] for generating normal variates, computes the matrices S_W and S_B, determines C_1, C_2, C_3. Determine the set of 3 x 1 vectors defined by

$$Z_j^\gamma = \begin{bmatrix} C_1 \\ C_2 \\ C_3 \end{bmatrix} X_j^\gamma \qquad \gamma = 1,2,3 \qquad j = 1,2,3,\ldots,N_\gamma.$$

Determine the set of 2 x 1 vectors defined by

$$Z_j^\gamma = \begin{bmatrix} C_1 \\ C_2 \end{bmatrix} X_j^\gamma \qquad \gamma = 1,2,3 \qquad j = 1,2,3,\ldots,N_\gamma.$$

Determine the set of scalars defined by

$$Z_j^\gamma = C\, X_j^\gamma \qquad \gamma = 1,2,3 \qquad j = 1,2,3,\ldots,N_\gamma.$$

Generate an additional 100 samples of X's from one of the density, say

$$X^{(1)} \sim N(e_1, I).$$

Classify the 100 samples according to the two techniques described below. Repeat the experiment 100 times and compare the average percentages of correct classification for each level of dimension reduction.

In order to study the effect of closeness of the populations, let the value of the constant, a, take on a sequence of positive values that tend to zero. In the cases proposed for study the sequence selected should be {2,1,3/4,1/2}. The results then may be summarized in table form.

Two techniques which are suggested to be used to assign the observations to π_1, π_2, or π_3. The first technique assumes the probability density function of X's are normal and estimates the mean μ_j $j = 1,2,3$ and variances Σ_j $j = 1,2,3$, then assigns X to π_i if

$$p_1(x;\hat{\mu}_i,\hat{\Sigma}_i) > p_2(x;\hat{\mu}_j,\hat{\Sigma}_j)$$

and

$$> p_3(x;\hat{\mu}_k,\hat{\Sigma}_k);$$

where

$$i,j,k = 1,2,3 \text{ and } i \neq j,\ i \neq k, \text{ and } j \neq k.$$

The second technique is to estimate p_i by using the non-parametric estimator p_i formulated in Lecture I and then assign X to π_i if

$$\hat{p}_i > \hat{p}_j$$

and

$$> \hat{p}_k$$

$i,j,k = 1,2,3$ and $i \neq j$, $i \neq k$, and $j \neq k$.

Details of these techniques are discussed in a previous lecture and hence will not be discussed here. The reader is referred to the first lecture.

4. Concluding Remarks

Let the set A_{ij} be defined as follows.

$$A_{ij}(x) = \{x; p_i(x)/p_j(x) \geq 1\}$$

and

$$A_{ij}(z) = \{z; p_i(z)/p_j(z) \geq 1\}$$

where $p_i(\cdot)$ is the probability density of $(\cdot)$ given that $(\cdot)$ from the i^{th} population.

The question that arises is as follows: If x belongs to A_{ij}, does $z = cx$ belong to $A_{ij}(z)$?

Let $p_i(x)$ be normal, such that $E(X) = \mu^{(i)}$ and $Cov(X) = \Sigma^{(i)}$ $i = 1,2$. Then $Z = CX$ is distributed normally with mean $C\mu^{(x)}$ and covariance matrix $C\Sigma^{(i)}C^T$. Now suppose that $p_i(x)/p_j(x) = 1$, then

$$p_i(z)/p_j(z) = \frac{|C\Sigma^{(j)}C^T|}{|C\Sigma^{(i)}C^T|} e^{-\frac{1}{2}\left[(x - \mu^{(i)})^T C^T (C\Sigma^{(i)}C^T)^{-1} C(X - \mu^{(i)})\right.}$$

$$\left. - (x - \mu^{j})^T C^T (C\Sigma^{j}C^T)^{-1} C(x - \mu^{(j)})\right]$$

But we know that

$$\alpha^T C^T (C\Sigma^{(i)}C^T)^{-1} C\alpha \leq \alpha^T \Sigma^{(i)-1} \alpha$$

for every vector α.

Let $\Sigma^{(1)} = \Sigma^{(2)} = I$, then

$$\frac{p_i(z)}{p_j(z)} = e^{-\frac{1}{2}[(x - \mu^{(i)})C^T(CC^T)^{-1}C(x - \mu^{(i)})}$$

$$- (x - \mu^{(j)})C^T(CC^T)^{-1}C(x - \mu^{(j)})]$$

which implies that in general one cannot be sure that if x is such that

$$\frac{p_i(x)}{p_j(x)} = 1$$

then it is not true that

$$\frac{p_i(z)}{p_j(z)} = 1.$$

Clearly if $C^T(CC^T)^{-1}C = I$ when $\Sigma^{(1)} = \Sigma^{(2)} = I$, then these conditions would hold. But rank considerations force $C^T(CC^T)^{-1}C$ to be of rank r although $C^T(CC^T)^{-1}C$ is a $p \times p$ matrix.

Further study is necessary to define those conditions such that no loss of discrimination power results from the suggested reduction of dimension. The proposed simulation discussed in Part III of this lecture should imply facts that may lead to some analytical statements concerning the effects of reduction of dimension.

5. References

[1] Wilks, S.S., Mathematical Statistics, Wiley and Sons, (1962), pp. 573-587.

[2] Wilks, S.S., Collected Papers: Contributions to Mathematical Statistics, John Wiley and Sons, Inc., New York, (1967) pp. 597-614.

[3] Sebestyen, G.S., Decision-Making Processes in Pattern Recognition, The Macmillan Company, New York, 1962.

[4] Marill, T. and D.M. Green, On the Effectiveness of Receptors in Recognition Systems, IEEE Transactions on Information Theory, Vol IT-9, pp. 11-17, 1963.

[5] Cooper, D.B. and Paul W. Cooper, Nonsupervised Adaptive Signal Detection and Pattern Recognition, Information and Control, Vol. 7, pp. 416-444 (1964).

[6] Anderson, T.W., Introduction to Multivariate Statistical Analysis, Wiley and Sons. (1958), pp. 126-152.

[7] Kendall, M.G. and Stuart A., The Advanced Theory of Statistics, Hafner, (1966), Vol. 3, pp. 314-341.

[8] Kendall, M.G., A Course in Multivariate Analysis, (1957), Charles Griffin and Company Limited, London, pp. 144-170.

[9] Kendall, M.G., Discrimination and Classification, Proc. of WAFB Symposium on Multivariate Analysis, Academic Press, New York, (1966), pp. 165-185.

[10] Odell, P.L., Non-Parametric Discrimination Techniques, Lecture I of this report, pp. 4-25.

[11] Birkhoff, G., and MacLane, S., Survey of Modern Algebra, rev. ed., Macmillan, New York, 1953.

BIBLIOGRAPHY

Bartlett, M. S. and Please, N. W., (1963). Discrimination in the case of zero mean differences. Biometrika 50, 17-21.

Geisser, S. (1964a). Posterior odds for multivariate normal classifications. J.R. Statist. Soc. B26, 69-76.

Geisser, S. (1964b). Estimation in the uniform covariance case. J.R. Statist. Soc. B26, 477-83.

Okamoto, M. (1961). Discrimination for variance matrices. Osaka Math. J. 13, 1-39.

Penrose, L. S. (1946-47). Some notes on discrimination. Ann. Eugen. 13, 228-37.

Richter, D. and Geisser, S. (1960). A statistical model for diagnosing zygosis by ridge-count. Biometrics 16, 110-14.

Stocks, P. (1933). A biometric investigation of twins. Part II. Ann. Eugen. 5, 1-55.

Maynard Smith, S. and Penrose, L. S. (1955). Monozygotic and dizygotic twin diagnosis. Ann. Hum. Genet., Lond., 19, 273-89.

Smith, C. A. B. (1946-47). Some examples of discrimination. Ann. Eugen., Lond., 13, 272-82.

Stocks, P. (1931). A biometric investigation of twins. Part I. Ann. Eugen., Lond., 4, 49-108.

Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. Ann. Eugenics, 7, 179-188.

Fix, E. and Hodges, J. L. (1951). Discriminatory Analysis. Project Report 21-49-004, Numbers 4 and 11. U.S. Air Force School of Aviation Medicine Randolph Field, San Antonio, Texas.

Kendall, M. G. (1958). Rank Correlation Methods. Charles Griffin, London.

Kendall, M. G. (1963). A Course in Multivariate Analysis. Charles Griffin, London.

Kendall, M. G. and Sundrum, R. M. (1953). Distribution-free methods and order properties. Rev. Internat. Statist. Inst. 3, 124-134.

Sokal, R. R. and Sneath, P. H. A. (1963). Principles of Numerical Taxonomy. Freeman, San Francisco and London.

Welch, B. L. (1939). Note on discriminant functions. Biometrika, 31, 218-220.

Sitgreaves, Rosedith.,(1952). On the distribution of two random matrices used in classification procedures. Ann. Math. Stat., 23, p. 263.

Wald, A., (1944). On a statistical problem in the classification of an individual into one of two groups. Ann. Math. Stat., 15, p. 145.

Cornfield, J. (1962). Joint dependence of risk of coronary heart disease on serum cholesterol and systolic blood pressure: a discriminant function analysis. Federation Proc. 21 (II), 53-61.

Cramer, E. (1967). Equivalence of two methods of computing discriminant coefficients. Biometrics 23, p. 153.

Lachenbruch, P. A. and Mickey, M. R. (1965). Estimation of the multiple correlation coefficient and mean square prediction error. Paper presented at Biometric Society meetings, WNAR, Riverside, California.

Lachenbruch, P. A. and Mickey, M. R. (1968). Estimation of error rates in discriminate analysis. Technometrics 10, 1-10.

Olkin, I. and Pratt, J. W. (1958). Unbiased estimation of certain correlation coefficients. Ann. Math. Statist. 29, 201-11.

Sitgreaves, Rosedith (1961). Some results on the distribution of the W-Classification statistic. pp. 241-61 in Studies in Item Analysis and Prediction. Stanford University Press, Stanford, California.

Wherry, R. J. (1931). A new formula for predicting the shrinkage of the multiple correlation coefficient. Ann. Math. Statist. 2, 440-57.

Russell, J. S., Moore, A. W., and Coaldrake, J. E. (1976). Relationships between subtropical semiarid forest of Acacia Harpophylla (brigalow), microrelief and chemical properties of associated gilgai soil. Aust. J. Bot. 15, 481-98.

Seal, H. L. (1964). Multivariate Statistical Analysis for Biologists. Methuen, London.

Kronmal, Richard (1964). The Evaluation of a Pseudorandom Normal Number Generator. Jour. Assn. for Computing Machinery 11, 357-63.

Cornfield, J., Gordon, T. and Smith, W. (1961). Quantal Response Curves for Experimentally Uncontrolled Variates. Bull. Inter. Stat. Inst. 38, 97-115.

Fisher, R. A. (1938). The Statistical Utilisation of Multiple Measurements. Ann. Eug., Lond., 7, 179-88.

von Neumann, J. (1961). Various Techniques used in Connection with Random Digits. Paper No. 13 in Monte Carlo Methods. National Bureau of Standards, Applied Mathematics Series No. 12. Washington: U.S. Govt. Printing Office.

Warner, H., Toronto, A., Veesey, L. and Stephenson, R. (1961). A Mathematical Approach to Medical Diagnosis. J. Amer. Med. Assn. 177, 177-83.

Lazarsfeld, P. F. Some observations on dichotomous systems. Sociology Department, Columbia University, New York. Duplicated report (second draft), October 1956.

Bahadur, R. R. A representation of the joint distribution of responses to n dichotomous items. USAF SAM Series in Statistics, Report No. 59-42, Randolph AFB Texas: USAF School of Aviation Medicine, January 1959.

Wishart, J. The generalized product moment distribution in samples from a normal multivariate population. Biometrika, 1928, 20A, 32-52.

Hotelling, H. The generalization of Student's ratio. Ann. Math. Stat., 1931, 2, 360-78.

Wilks, S. S. Certain generalizations in the analysis of variance. Biometrika, 1932, 24, 471-94.

Hotelling, H. Analysis of a complex of statistical variables into proncipal components. J. Educ. Psychol., 1933, 24, 417-41, 498-520.

Fisher, R. A. The sampling distribution of some statistics obtained from nonlinear equation. Ann. Eugen., 1939, 9. 238-49.

Hsu, P. C. On the distribution of the roots of certain determinantal equations. Ann. Eugen., 1939, 9, 250-58.

Mood, A. M. On the distribution of normal second-moment matrices. Ann. Math. Stat., 1951, 22, 266-73.

James, A. T. Normal multivariate analysis and the orthogonal group. Ann. Math. Stat., 1954, 25, 40-75.

Olkin, I. On distribution problems in multivariate analysis. Institute of Statistics Mimeograph Series Report No. 43, Chapel Hill, N.C.: University of North Carolina, 1951.

Olkin, I., and Roy, S. N. On multivariate distribution theory. Ann. Math. Stat., 1954, 25, 329-39.

Roy, S. N. p-statistics or some generalizations in analysis of variance appropriate to multivariate problems. Sankhya, 1939, 4, 381-96.

Parzen, Emanual "On uniform convergence of families of sequences of random variables," Univ. Calif. Publ. Statist., Vol. 2, No. 2 (1954), pp. 23-54.

Ward, Joe "Hierarchical grouping to optimize an objective function, J. Amer. Statist. Assoc., Vol. 58 (1963), pp. 236-244.

Anderson, T. W. (1958). An Introduction to Multivariate Statistical Analysis. J. Wiley and Sons, Inc., New York.

Dixon, W. J., and Massey, F. J., Jr. (1957). Introduction to Statistical Analysis. 2nd Ed., McGraw-Hill, New York.

Hollingsworth, T. H. (1959). Using an electronic computer in a problem of medical diagnosis. J. Roy. Stat. Soc. 122, 221-31.

Lubischew, A. A. (1962). On the use of discriminate functions is taxonomy. Biometrics 18, 455-77.

Myers, J. H., and Forgy, E. W. (1963). The development of numerical credit evaluation systems. J. Amer. Stat. Assn. 58, 799-806.

Weiner, J. M., Allen, M., and Marmorston, J. (1965). Manual of Computer Programs for Preliminary and Multivariate Analysis. University of Southern California School of Medicine. Los Angeles.

Bartlett, M. S. (1951). An inverse matrix adjustment arising in discriminant analysis. Ann. Math. Statist. 22, 107-11.

Hills, M. (1966). Allocation rules and their error rates. J. Roy. Statist. Soc. B28, 1-31.

Hoel, P. G. and Peterson, R. P. (1949). A solution to the problem of optimum classification. Ann. Math. Statist. 20, 433-8.

Lachenbruch, P. A. (1965). Estimation of error rates in discriminant analysis. Ph.D. Dissertation (unpublished), University of California, Los Angeles.

Mahalanobis, P. C. (1936). On the generalized distance in statistics. Proc. Nat. Inst. Sci. India 2, 49-55.

Anderson, T. W., (1951), Classification by multivariate analysis. Psychometrika, 16, p. 31.

Kabe, D. G., (1963). Some results on the distribution of two random matrices used in classification procedures. Ann. Math. Stat., 34, p. 181 (a correction to this article appeared in Ann. Math. Stat., 1964, 34, p. 924).

Okamoto, M., (1963). An asymptotic expansion for the distribution of the linear discriminant function. Ann. Math. Stat., 34, p. 1286.

Hotelling, H. Relations between two sets of variates. Biometrika, 1936, 28, 321-77.

Dempster, A. P. (1964). On the difficulties inherent in the fiducial argument. J. Amer. Statist. Assn, 59, 56-66.

Geisser, S. (1965). Bayesian estimation in multivariate analysis. Ann. Math. Statist. 36, 150-159.

Geisser, S. and Cornfield, J. (1963). Posterior distributions for multivariate normal parameters. J. Roy. Statist. Soc. Ser. B 25 368-376.

Bartlett, M. S., "Tests of significance in factor analysis," Britis J. Psych. (Statist. Sec.), Vol. 3 (1950), pp. 77-85.

Duncan, D. B., "On some theory of mixed linear models and its application in the estimation of missile trajectory," paper No. 332 from the Department of Biostatistics, The Johns Hopkins Univ.

Durbin, J. Estimation of parameters in time-series regression models, J. Roy. Statist. Soc., Vol. 22 (1960), pp. 139-153.

Forsyth, G. E. and Golub, G., On the stationary values of a second degree polynomial on the unit sphere, J. SIAM, to appear.

Henderson, C. R., Dempthorne, Oscar, Searle, S. R., and Drosigk, C. The estimation of environmental and genetic trends from records subject to culling, Biometrics, Vol. 15 (1959), pp. 192-218.

Mann, H. B. and Wald, A., On the statistical treatment of linear stochastic difference equations, Econometrics, Vol. 11 (1943), pp. 173-220.

Rao, C. R., Tests with discriminant functions in multivariate analysis, Sankhya, Vol. 7 (1946), pp. 407-414.

_____, A note on the distribution of $D^2_{p+q} - D^2$ and some computational aspects of D^2 statistic and discriminant function, Sankhya, Vol 10 (1950), pp. 257-268.

_____, On some problems arising out of discrimination with multiple characters, Sankhya, Vol. 9 (1949), pp. 343-364.

_____, Estimation and tests of significance in factor analysis, Psychometrika, Vol. 20 (1955), pp. 93-111.

_____, Some problems involving linear hypotheses in multivariate analysis, Biometrika, Vol. 46 (1959), pp. 49-58.

_____, The use and interpretation of principal component analysis in applied research, Sankhya, Vol. 26 (1964), pp. 329-358.

Sebestyen, G. S., Decision-Making Processes in Pattern Recognition, The Macmillan Company, New York, 1962.

Nilsson, Nils J., Learning Machines, McGraw-Hill, New York, 1965.

Cooper, D. B. and Paul W. Cooper, Nonsupervised Adaptive Signal Detection and Pattern Recognition, Information and Control, Vol. 7, pp. 416-444 (1964).

Marill, T. and C. M. Green, On the Effectiveness of Receptors in Recognition Systems, IEEE Transactions on Information Theory, Vol. IT-9, pp. 11-17, 1963.

Bochner, S. (1955). Harmonic Analysis and the Theory of Probability. Univ. of California Press.

Ling, C. and Wu, H. (1965). Tables of values of three infinite integrals. Math. of Computation 19 487-494.

Loeve, M. (1963). Probability Theory (3rd ed.). Van Nostrand, Princeton.

Parzen, E. (1962). On estimation of a probability density function and mode. Ann. Math. Statist. 33 1065-1076.

Rutherford, J. (1966). Some parametric empirical Bayes techniques. Doctoral theses, Virginia Polytechnic Institute.

Schlafli, L. Theorie der vielfachen Kontinuitat (Berne, 1852), Ges. Math. Abh. Vol. I, p. 209 (Basel, 1950).

Carver, W. B., The polygonal regions into which a plane is divided by n straight lines. Amer. Math. Monthly, 48(1941), 667-675.

Blumenthal, L. M. and Gillam, B. E. Distribution of points in n-space. Amer. Math. Monthly, 50 (1943), 181-185.

Buck, R. C. Partition of space, Amer. Math. Monthly, 50 (1943), 541-544.

Robinson, H. A. A problem of regions. Amer. Math. Monthly 52 (1945), 33-34.

Edwards, A. W. F. and Cavalli-Sforza, L. L. A method for Cluster Analysis, Biometrics, 21 (2)(1965), 362-375.

Winder, R.O. Partitions of N-space by Hyperplanes, SIAM J. Appl. Math. 14 (4)(1966), 811-818.

Aitken, A. C., On least squares and linear combinations of observations. Proceedings of the Royal Society of Edinburgh, A, Vol. 55 (1934), 42-47.

Goldberger, A. S., Econometric Theory, New York: John Wiley & Sons, (1964).

Halperin, Max, Normal regression theory in the presence of intra-class correlation, Annals of Mathematical Statistics, Vol. 22 (1951), 575-80.

Johnston, J., Econometric Methods, New York: McGraw-Hill, (1963).

Plackett, R. L., Principles of Regression Analysis, Oxford: Oxford University Press, 1960, problem 8 on p. 50.

Watson, G. S., Serial correlation in regression analysis I, Biometrika, Vol. 42, 327-41.

Blackwell, David Comparison of experiments, Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability, Berkeley and Los Angeles, University of California Press, 1951, pp. 93-102.

Cox, D. R. Note on grouping, J. Amer. Statist. Assn., Vol. 52 (1957), pp. 543-547.

Doob, J. L. Stochastic Processes, Nes York, Wiley,1953.

Dubins, L. E. and Savage, L. J., A Tchebycheff-like inequality for stochastic processes, Proc. Nat. Acad. Sci. U.S.A., Vol. 53 (1965), pp. 274-275.

Fisher, W. D., On grouping for maximum homogeneity, J. Amer. Statist. Assn., Vol. 53 (1958), pp. 789-798.

Forgy, Edward Cluster analysis of multivariate data: efficiency vs. interpretability of classifications, abstract, Biometrics, Vol. 21 (1965), p. 768.

Halmos, Paul R. Measure Theory, New York, Van Nostrand, 1950.

MacQueen, J. The classification problem, Western Management Science Institute Working Paper No. 5, 1962.

_____, On convergence of K-means and partitions with minimum average variance, abstract, Ann. Math. Statist., Vol. 36, (19650, p. 1084.

Marschak, Jacob Towards an economic theory of organization and information, Decision Processes, edited by R. M. Thrall, C.H. Coombs, and R.C. Davis, New York, Wiley, 1954.

_____, Remarks on the economics of information, Proceedings of the scientific program following the dedication of the Western Data Processing Center, University of California, Los Angeles, January 29-30, 1959.

_____, Linear Statistical Inference and Its Applications, New York, Wiley, 1965.

_____, Generalized inverses for matrices and its applications in mathematical statistics, Research Papers in Statistics (Festschrift volume for J. Neyman), pp. 263-280, New York, Wiley, 1966.

Cox, D. R. and Brandwood, L. (1959). On a discriminatory problem connected with the works of Plato. Jour. Roy. Stat. Soc. B. 21, 195-200.

Linhart, H. (1959). Techniques for discriminant analysis with discrete variables. Metrika 2, 138-49.

Ogawa, J. (1951). Contributions to the theory of systematic statistics. Osaka Math. Jour. 4, 175-213.

Rao, C. R. (1952). Advanced Statistical Methods in Biometric Research. John Wiley and Sons, New York, Chapter 8.

Stokes, D. E., Campbell, A. and Miller, M. E. (1958). Components of electoral decision. Amer. Pol. Sci. Rev. 52, 367-87.