SOME UNSOLVED PRACTICAL PROBLEMS IN DISCRIMINANT ANALYSIS

by

Peter A. Lachenbruch

Department of Biostatistics
University of North Carolina at Chapel Hill

Institute of Statistics Mimeo Series No. 1050

December 1975

SOME UNSOLVED PRACTICAL PROBLEMS IN DISCRIMINANT ANALYSIS

by

Peter A. Lachenbruch

University of North Carolina

I. Introduction

A number of procedures have been proposed for assigning an individual to one of two or more groups on the basis of a multivariate observation. A review of these appears in Lachenbruch (12). The procedures that will be considered in this article are:

(1) The linear discriminant function. If the parameters are known, this is the optimum rule for assigning to one of two normal populations with the same covariance matrix. In practice, an observation is assigned to the first population if

$$D_s(\underset{\sim}{x})=(\underset{\sim}{x}-\tfrac{1}{2}(\bar{\underset{\sim}{x}}_1+\bar{\underset{\sim}{x}}_2))' \underset{\sim}{S}^{-1}(\bar{\underset{\sim}{x}}_1-\bar{\underset{\sim}{x}}_2)>C$$

where $\bar{\underset{\sim}{x}}_i$ is the sample mean vector in the $i^{\underline{th}}$ population, S is the sample covariance matrix, and C is a cutoff point chosen by the investigator.

(2) The quadratic discriminant function. If the parameters are known, this is the optimum rule for assigning to one of two normal populations with different covariance matrices. In practice, an observation is assigned to the first population if

$$Q_s(\underset{\sim}{x})=(\underset{\sim}{x}-\bar{\underset{\sim}{x}}_2)' \underset{\sim}{S}_2^{-1}\ (\underset{\sim}{x}-\bar{\underset{\sim}{x}}_2)-(\underset{\sim}{x}-\bar{\underset{\sim}{x}}_1)' \underset{\sim}{S}_1^{-1}\ (\underset{\sim}{x}-\bar{\underset{\sim}{x}}_1)>C$$

(3) The multiple group discriminant function. This rule is theoretically optimal when there are g multivariate normal populations with the same covariance matrix. In practice, an observation is assigned to the $i^{\underline{th}}$

population if

$$M_j(\underset{\sim}{x})=(\underset{\sim}{x}-\tfrac{1}{2}\bar{\underset{\sim}{x}}_j)'\underset{\sim}{S}^{-1}\bar{\underset{\sim}{x}}_j+C_j$$

is maximized for j=i, and C_j is a constant depending on the a priori probability of the $j^{\underline{th}}$ population.

(4) The eigenvector, or canonical vector, approach to the multiple group problem. This is an extension of Fisher's original approach to the discriminant problem. The variables are transformed to eigenvectors, and assignment may be made using any number of the transformed variables as in the multiple discriminant method.

(5) Density estimation techniques. These have become more feasible with the availability of computers. One estimates the densities of the distributions in each population, and assign $\underset{\sim}{x}$ to the $i^{\underline{th}}$ population if

$$\hat{f}_i(\underset{\sim}{x}) = \max_j \hat{f}_j\ (\underset{\sim}{x})$$

In assignment problems in biomedical research, one or more of these techniques is often used. The assumptions underlying these techniques are not always evident to the user, nor are the consequences of their violation. The assumptions include multivariate normality, common covariance matrices, and correct assignment of the initial groups. While a good deal is known in the two group situation, the robustness of these procedures in the multiple group situation is essentially unknown. The purpose of this paper is to delineate these problems systematically and to suggest useful areas of research.

II. Types of Problems

In the two group setting, the following areas have been explored in recent years.

1. Unequal covariance matrices: Gilbert (6) considered the case in which all parameters are known. The quadratic discriminant function is then optimal. She noted that if the

parameters were such that the covariance matrices were not too dissimilar the linear discriminant function performed almost as well as the quadratic discriminant function. This assumed that (using the well known diagonalizing transformation) all elements of the non-identity matrix were equal. The worst case, of course, was the one in which the means in the two populations were equal. Marks and Dunn (13) considered the same problem when the parameters were estimated from initial samples. Their results corresponded closely to those of Gilbert. Kronmal and Wahl (7) indicated the amount of improvement the quadratic discriminant gives over the linear. They recommend its use when unequal covariances are present.

2. Dichotomous variables: In many medical problems the variables consist of a symptom checklist (presence or absence). The effects of using the linear discriminant function in this situation have been studied by Gilbert (5), Moore (15) and Brown (4). They found that the linear function generally performs as well as the estimated optimal rule. Reasons for this are not hard to find. If there are k variables, this is equivalent to a contingency table with $2x2^k$ possible cells, and each cell probability must be estimated. If the linear discriminant function is used, there are $2k+k(k+1)/2=k(k+5)/2$ parameters to estimate. (2k for the means and $k(k+1)/2$ for the covariances) For example, for k=8, the estimated optimal rule must estimate 512 parameters, while the linear discriminant function must estimate 52. Some other procedures have

been considered such as treating all variables as independent, or using log-linear models for the cell probabilities. These seem to work fairly well, but their use has been limited because of the attractiveness of using an already packaged program.

3. Discrete Distributions: Revo (16) considered the case of ordered discrete distributions exemplified by the bivariate Poisson, bivariate negative multinomial, and a discretized normal distribution. In these cases, he found that the linear function performed very well. Estimating parameters in the full model did fairly well, but this is valid only if the model is adequate. The multinomial procedure estimates cell probabilities poorly, because of the necessity for extremely large samples.

4. Continuous Non-Normal Distributions: Lachenbruch, Sneeringer, and Revo (10) studied this problem and found that the individual probabilities of misclassification were drastically affected by this sort of distribution. It was recommended that the data be transformed prior to analysis, to ensure closer approximations to normality.

5. Contamination of distributions: Ahmed (1) and Ahmed and Lachenbruch (2) showed that contamination could affect the performance of the linear discriminant function considerably. A number of alternative estimation procedures were evaluated, and it was shown that some of the simpler M-estimates (see Andrews et al) (3) improved the overall behavior in the scale contamination problem. The location contamination problem is more

difficult; a study is currently in progress.

6. Initial Misclassification of Samples: Lachenbruch (8, 11), and McLachlan (14) have shown that if the rate of misclassification is not too high and is the same in both populations, there is minimal effect. However, if the initial misclassification rates are unequal, the error rates are affected. If there is non-random initial misclassification (e.g. those more similar to the other group are more likely to be misclassified) there is relatively little effect on the true error rates, but the apparent error rates (those found by resubstituting the initial samples into the computed discriminant function) give far too optimistic a picture.

III. The Unsolved Problems

The available information on the robustness of multiple group procedures is based on extrapolation from results from the two group problems which may or may not hold in the multiple group setting. There are two main reasons for this lack of solid information. First, statisticians have recently begun to study the robustness of discriminant procedures. Second, in multiple group problems, very difficult decisions must be made as to "realistic" violations of assumptions. For example, in studying robustness to unequal covariance matrices, in a three group context, should two of the three groups have equal matrices and one different, or should all three have different covariances? Is there a convenient canonical form as there is in the two group problem?

In both the two group and multiple group problem, there are a variety of packaged programs to perform the analysis in a stepwise manner. Only limited work has been done on the variable selection problem in the two group problem, and none has appeared on the multiple group problem. The problems involved here include: a) What are appropriate significance levels to use? b) What is the probability of failing to include an "important" variable? (this may not merely mean one whose coefficient is significantly different from zero, but possibly may mean one which improves the error rate) c) How robust are the methods to non-normality? d) What are appropriate tests for variables in the quadratic discriminant problem?

In the case of unequal covariance matrices, the quadratic discriminant function is the optimal method when all parameters are known. Its robustness to failure of assumptions has not been extensively studied. There have been some indications that it is quite sensitive to non-normality when the non-normality is long-tailed. This seems to be because the coefficients, which depend on the covariance matrices are seriously affected by the extreme observations. However, no systematic study of this has been made, and therefore no general statements can be given. No investigation has been undertaken of the effects of initial misclassification on the quadratic function. We may speculate that the effects of initial misclassification will be to make the covariance matrices more alike, and that any advantages of using the quadratic function over the linear function will tend to disappear. The violation of the common covariance assumption is not applicable to the quadratic discriminant function as that is what it is designed to handle.

If more than two groups are present, say g of them, two methods are presently used. The first method is based on the assumption of multivariate normality and common covariances. It yields a set of g linear functions. An observation is assigned to the group corresponding to the largest function. The second method is based on canonical vectors which maximize the between group variance relative to the within-group variance. These vectors are related to those involved in multivariate analysis of variance. They are particularly useful in describing the space in which the observations are located. Sampling studies are difficult to design for the multiple group problem because there is no standard form as in the two group case. If common covariances hold, one may assume the identity matrix as the covariance matrix but the means will lie in a g-1 (or less) dimensional subspace. The placement of the means is arbitrary. One way around this difficulty might be to assign direction cosines randomly, and choose an appropriate multiplier. An alternative method is to find least favorable and most favorable configurations for the two methods. As shown in Lachenbruch (9), the most favorable configuration for the canonical vector method is when the means lie on a straight line, for then a single canonical vector can represent all information to be used in classification. The least favorable configuration for the canonical vectors procedure occurs when the means are placed on the vertices of a regular simplex. One may ask what configuration of means leads to the smallest error rate (or highest correct classification rate) for a given average pairwise distance? I suspect that the answer is the regular simplex, but I have not been able to prove it as yet. In the collinear case, the groups that are on the extremes of the line have very little overlap compared to the adjacent ones. Robustness properties of multiple group procedures have not been studied. One might expect non-normality to affect the linear functions more than the canonical vectors since the latter is based on maximizing a ratio of quadratic forms, while the former seems to depend heavily on the distributional assumption. However, it is difficult to assess the extent of the effect. Unequal covariances matrices would change the optimal rule to a quadratic one. The effect on the canonical vector procedure would depend on the extent of inequality, the mixing proportions, and the configuration of means. Initial misclassification tends to reduce the distance between means, and to reduce the size of the variances, so one might expect that the multiple group rule will not be too seriously affected, but no speculations can be made regarding the canonical vector rule. Studies of contamination are just beginning for the two group case, and we are learning which modifications help there. It is hoped that many of the techniques which work in the two group case will also work in the multiple group case. It appears that location contamination will be the more serious problem.

Density estimate methods would seem to suffer from no problems due to unequal covariances or continuous non-normal distributions. However, these methods do require fairly large samples, and are not easily computed, particularly for many variables. It would appear that initial misclassification has the effect of making the distributions more similar and would tend to reduce the discriminatory power of the procedures. However, the loss of discriminatory power was expected using the linear discriminant function and that expectation was incorrect. If scale or location contamination occurs primarily in the tails and is not too heavy the density estimates should be unaffected and the classification procedures remain fairly good.

Table 1 indicates the unsolved problems in the area of robustness and variable selection. Currently, work is going on on the location contamination problem for the linear function in the two group case. My own priorities and interest are to study the robustness of the multiple linear discriminant function with respect to the effects of non-normality. I feel that one should include the canonical vectors approach in any such study as it is often suggested as an alternative to the multiple linear function. A slightly lower priority project is to look at the effects of non-normality on the quadratic function.

TABLE 1

UNSOLVED PRACTICAL PROBLEMS IN DISCRIMINANT ANALYSIS

Problem	Linear	Quadratic	Multiple		Density
			Linear	Canonical	Estimates
Initial Misclassification	1	2	2	2	2
Unequal Covariances	1	3	2	2	3
Non-Normality					
a) Dichotomous	1	2	2	2	2
b) Continouos	1	2	2	2	3
Contamination					
a) Scale	1	2	2	2	2
b) Location	2	2	2	2	2
Variable Selection	1,2	2,1	2,1	4	2

Notes:

1. Problem has been studied, and solution that seems to work satisfactorily in practice has been given.

2. This problem has not been studied, or the results given to date are not conclusive.

3. This technique is an appropriate solution for the problem indicated.

4. If the canonical vectors are ordered by the magnitude of the eigenvalues, the variable selection problem is essentially solved.

1. Ahmed S. (1975) Discriminant analysis when the initial samples are contaminated. Dissertation University of North Carolina

2. Ahmed S. and Lachenbruch P. (1975) Discriminant analysis when one or both of the initial samples is contaminated: Large Sample Results. EDV in Medizin und Biologie, 6, 35-42

3. Andrews D., Bickel, P., Hampel, F., Huber, P., Rogers W., and Tukey J. (1972) Robust Estimates of Location Princeton: Princeton University Press

4. Brown A.M. (1971) Classification using dichotomous responses. D.Sc. Dissertation University of Pittsburgh

5. Gilbert E. (1968) On discrimination using qualitative variables. JASA, 63, 1399

6. Gilbert E. (1969) The effect of unequal variance-covariance matrices on Fisher's linear discriminant function. Biometrics, 25, 505-516

7. Kronmal R. and Wahl P. (1975) Personal Communication

8. Lachenbruch P.A. (1966) Discriminant analysis when the initial samples are misclassified. Technometrics, 8, 657

9. Lachenbruch P.A. (1973) Some results on the multiple group discriminant problem. Discriminant Analysis and Applications T. Cacoullos ed. New York: Academic Press

10. Lachenbruch, P.A., Sneeringer C., and Revo L.T. (1973) Robustness of the linear and quadratic discriminant function to certain types of nonnomality. Commun Stat, 1, 39-57

11. Lachenbruch P.A. (1974) Discriminant analysis when the initial samples are misclassified II: Non-random misclassification models. Technometrics, 16

12. Lachenbruch P.A. (1975) Discriminant Analysis. New York: Hafner Press

13. Marks S. Dunn O.J. (1974) Discriminant functions when covariance matrices are unequal. JASA, 69, 555-559

14. McLachlan G.J. (1972) Asymptotic results for discriminant analysis when the initial samples are misclassified. Technometrics, 14, 415-422

15. Moore D.H. (1973) Evaluation of five discrimination procedures for binary variables. JASA, 68, 399-404

16. Revo L.T. (1970) On classifying with certain types of ordered variates: An evaluation of several procedures. North Carolina Institute of Statistics Mimeo Series No. 708