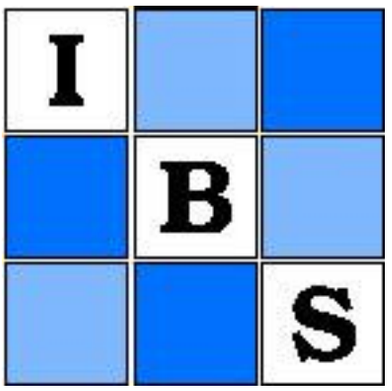

BIOMETRICS 35, 69–85
March, 1979

Discriminant Analysis

P. A. LACHENBRUCH

Department of Preventive Medicine and Environmental Health, University of Iowa, Iowa City, Iowa 52242, U.S.A.

M. GOLDSTEIN

Department of Statistics, Baruch College, City University of New York, New York, New York 10017, U.S.A.

Summary

This paper summarizes work in discriminant analysis. Normal theory and discrete results are discussed. Estimation of error rates and variable selection problems are indicated. Current research problems are considered: robustness, nonparametric rules, contamination, density estimation, mixtures of variables.

1. Theoretical Basis

The basic problem in discriminant analysis is to assign an unknown subject to one of two or more groups on the basis of a multivariate observation. It is important to consider the costs of assignment, the *a priori* probabilities of belonging to one of the groups, and the number of groups involved. The allocation rule is selected to optimize some function of the costs of making an error and the *a priori* probabilities of belonging to one of the groups. Denote by p_i the *a priori* probability of belonging to π_i the *ith* group; by c_{ji} the cost of assigning an observation to the *jth* group when the individual belongs to the *ith* group; and by D_i the region for which the assignment is made to population i. $P(D_j|\pi_i)$ is the probability that an observation from π_i falls in D_j. In the following pages we will be considering two group problems primarily, although extensions to more than two groups are usually straightforward.

Various criteria have been proposed to determine "good" allocation rules. One criterion is to minimize the cost of assignment to the groups. Then the problem of minimizing the cost of assignment is to minimize the following equation,

$$\underset{\{D_j\}}{\text{Min}} \sum_i \sum_{j \neq 1} P(D_j/\pi_i) p_i c_{ji}. \tag{1}$$

Thus we would choose the assignment regions D_j in such a way that equation (1) is minimized. Usually c_{ii} is assumed to be 0 and c_{ji} is supplied by the user (which, in practice, usually means the statistician). Because it is very difficult to determine relative costs of misclassification, it is often assumed that the c_{ji} are equal to 1 if $i \neq j$, and to 0 otherwise. In that case, the minimum cost criteria minimizes the overall error rate. At times this may lead

Key Words: Discriminant Analysis; Discrete Discriminant Analysis; Robustness; Nonparametric Rules.

to highly unbalanced error rates. For example, if one is attempting to diagnose a rare disease (say its prevalence is less than .01), the region would assign almost all cases to the non-disease group, whatever the observation. (It is always possible to obtain a rule which correctly classifies at the rate of the maximum of the p_i, simply by assigning every observation to the group which corresponds to that of the maximum of the p_i's.) For this reason it is sometimes desired to optimize a criterion other than the minimum error rate or the minimum cost criterion. For example, one may minimize the maxima of the probabilities of misclassification,

$$\underset{\{D_j\}}{\text{Min}} \ \underset{i}{\text{Max}} \ \underset{j \neq i}{\text{Max}} \ P(D_j/\pi_i). \tag{2}$$

Still another criterion that can be used is to fix the error rate in π_1 and accept the error rate that one gets in the second group. In epidemiologic terms, one has fixed the sensitivity and determined the specificity. This method may be necessary when the allocation is a first-line screen, as for example might occur in a tuberculosis screening project. In the majority of the remainder of this article the minimum average error rate will be used, and when it is not, the exception will be noted.

In developing an allocation rule distributional assumptions may be:

1. that the probability distributions of the random variable are completely known,
2. that the functional form of the distribution is known but the parameters are unknown, or
3. that nothing whatever is known about the distributions.

If the distributions are completely known, obtaining an optimal discriminant is very simple. Equation 3 gives the quantity to be minimized for this problem.

$$T(f_1, f_2, D_1, D_2) = p_1 \int_{D_2} f_1(\mathbf{x})\, d\mathbf{x} + p_2 \int_{D_1} f_2(\mathbf{x})\, d\mathbf{x}. \tag{3}$$

By an argument identical to that used in proving the Neyman-Pearson Lemma (Welch 1939) it is easy to show that the optimal rule is given by Equation 4.

$$D_1 = \left\{ \mathbf{x}: \frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} \geq \frac{p_2}{p_1} \right\} \quad \text{and} \quad D_2 = \left\{ \mathbf{x}: \frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} < \frac{p_2}{p_1} \right\}, \tag{4}$$

If the form of the distribution is assumed known but the parameters are unknown, one may obtain maximum likelihood (ML) estimates of the parameters and obtain the ML rule. For sufficiently large n, this will be an acceptable rule. The question has arisen, "What is sufficiently large?" and generally answers have been difficult to come by. As a rule of thumb, about three times as many observations as there are parameters to estimate in each group is satisfactory. This number will decrease for populations which are well separated, and will increase for populations that are close together. In practice, knowing the form of the distribution has meant that a multivariate normal distribution has been assumed.

If nothing is assumed about the underlying multivariate probability distributions, a nonparametric form of discrimination may be used. Relatively little is known about the performance of these rules in practice compared to rules based on multivariate normal distributions. In principle they should provide superior discriminating ability if the samples are quite large and the distributions are not normal. Some sampling experiments have suggested that their behavior is quite good even for small samples.

Frequently made assumptions are:

1. The underlying distributions are multivariate normal. The packaged computer programs such as the BMD, SPSS or SAS make this assumption. A hidden assumption is that the linear or quadratic discriminant function is robust to non-normality. Another distribution that is occasionally assumed is a

multinomial. When multivariate normal distributions are assumed it is usually also assumed that the covariance matrices are the same.

2. The distributions are uncontaminated; that is, the underlying distribution is purely normal. If, in fact, the distributions have longer tails than normal, or are contaminated, there can be a problem with the procedure.

3. There is an assumption that the initial samples that are used in developing the rule are correctly classified. This may be of considerable importance in some discrimination procedures.

General surveys of this field are available in Goldstein and Dillon (1978) and Lachenbruch (1975).

2. *Normal Theory*

If the underlying distributions are multivariate normal with means $\underset{\sim}{\mu}_1$ and $\underset{\sim}{\mu}_2$ and covariance matrices $\mathbf{\Sigma}_1$ and $\mathbf{\Sigma}_2$ then under the assumption that $\mathbf{\Sigma}_1 = \mathbf{\Sigma}_2$ the allocation rule is assign to π_1 if

$$f_1/f_2 = \exp\left(-\tfrac{1}{2}(\mathbf{x} - \underset{\sim}{\mu}_1)'\mathbf{\Sigma}^{-1}(\mathbf{x} - \underset{\sim}{\mu}_1) + \tfrac{1}{2}(\mathbf{x} - \underset{\sim}{\mu}_2)'\mathbf{\Sigma}^{-1}(\mathbf{x} - \underset{\sim}{\mu}_2)\right) > 1 - p/p \tag{5}$$

where $p = p_1$, $1 - p = p_2$, and $1 - p/p$ is read as $(1 - p)/p$. This is equivalent to allocating to π_1 if

$$D_T(\mathbf{x}) = [\mathbf{x} - \tfrac{1}{2}(\underset{\sim}{\mu}_1 + \underset{\sim}{\mu}_2)]'\mathbf{\Sigma}^{-1}(\underset{\sim}{\mu}_1 - \underset{\sim}{\mu}_2) > \ell n(1 - p/p). \tag{6}$$

It is easy to show that the error rates for this rule are

$$P_1 = \Pr(D_T(\mathbf{x}) < \ell n \frac{1-p}{p} \mid \pi_1) = \Phi\{[\ell n(1 - p/p) - \delta^2/2]/\delta\}$$

and

$$P_2 = \Pr(D_T(\mathbf{x}) > \ell n \frac{1-p}{p} \mid \pi_2) = \Phi\{[-\ell n(1 - p/p) - \delta^2/2]/\delta\}. \tag{7}$$

where $\delta^2 = (\underset{\sim}{\mu}_1 - \underset{\sim}{\mu}_2)'\mathbf{\Sigma}^{-1}(\underset{\sim}{\mu}_1 - \underset{\sim}{\mu}_2)$ and $\Phi(\cdot)$ is the cumulative normal distribution function.

This of course assumes that the parameters $\underset{\sim}{\mu}_1$, $\underset{\sim}{\mu}_2$ and $\mathbf{\Sigma}$ are known. In practice this is never so. Therefore a sample-based rule is usually employed in which $\underset{\sim}{\mu}_1$, $\underset{\sim}{\mu}_2$ and $\mathbf{\Sigma}$ are estimated by their ML estimators. Then one allocates to π_1 if

$$D_S(\mathbf{x}) = [\mathbf{x} - \tfrac{1}{2}(\bar{\mathbf{x}}_1 + \bar{\mathbf{x}}_2)]'\,\mathbf{S}^{-1}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2) > \ell n(1 - p/p). \tag{8}$$

The error rates for the sample linear discriminant function are

$$\begin{aligned} P_1 &= \Phi\left\{\frac{ln\ (1 - p/p) - [\underset{\sim}{\mu}_1 - \tfrac{1}{2}(\bar{\mathbf{x}}_1 + \bar{\mathbf{x}}_2)]'\mathbf{S}^{-1}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)}{SQRT(V_S)}\right\} \\ P_2 &= \Phi\left\{\frac{-ln\ (1 - p/p) + [\underset{\sim}{\mu}_2 - \tfrac{1}{2}(\bar{\mathbf{x}}_1 + \bar{\mathbf{x}}_2)]'\mathbf{S}^{-1}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)}{SQRT(V_S)}\right\} \end{aligned} \tag{9}$$

where SQRT(V_S) is the square root of $V_S = (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)'\mathbf{S}^{-1}\,\mathbf{\Sigma}\,\mathbf{S}^{-1}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)$.

The coefficients $\mathbf{S}^{-1}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)$ are referred to as the discriminant coefficients. These coefficients were derived by Fisher (1936) from a different point of view. He was searching for a linear combination of the variables which had maximum between-group difference relative to its within-group standard deviation. A relatively straightforward derivation shows that the coefficients are determined up to a multiplicative constant and that those coefficients are indeed the ones cited above.

In practice it is usually necessary to estimate error rates. We shall briefly discuss three methods here. First, one may resubstitute the training samples into the sample discriminant function and count the number of observations that have been misclassified. This method is biased, and should not be used for small samples. It is also poor if the training samples are misclassified. A second method is the leaving-one-out or jackknife method. This sequentially removes one observation from the training sample, calculates the discriminant on the remaining observations, and classifies the deleted observation. The estimate is the proportion misclassified. This removes much of the bias of the resubstitution method. For large samples, these methods are essentially the same. The third method is the "plug-in" method. Estimates of μ_i and Σ are substituted in equation 9 which lead to

$$P_1 = \Phi\{[\ell n(1 - p/p) - D^2]/D\} \qquad \text{and} \qquad P_2 = \Phi\{[-\ell n(1 - p/p) - D^2]/D\}.$$

where $D^2 = (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)\, \mathbb{S}^{-1}\, (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)$ assuming $\hat{\mu}_i = \bar{\mathbf{x}}_i$ and $\hat{\Sigma} = \mathbf{S}$). A more complete discussion is available in Lachenbruch (1975). These results can also be applied to discrete variable problems.

If the underlying populations are normal but the covariances are not the same in the two groups, then the rule for assigning to population 1 is given by Equation 10.

$$\frac{f_1}{f_2} = \frac{|\Sigma_2|^{1/2}}{|\Sigma_1|^{1/2}} \exp\left[-\tfrac{1}{2}(\mathbf{x} - \mu_1)'\Sigma_1^{-1}(\mathbf{x} - \mu_1) + \tfrac{1}{2}(\mathbf{x} - \mu_2)'\Sigma_2^{-1}(\mathbf{x} - \mu_2)\right] > \frac{1-p}{p} \tag{10}$$

This is equivalent to the formulation in Equation 11, which will be referred to as the Quadratic Discriminant Function.

$$Q_T(\mathbf{x}) = \tfrac{1}{2}\left[\ell n \frac{|\Sigma_2|}{|\Sigma_1|} + (\mathbf{x} - \mu_2)'\Sigma_2^{-1}(\mathbf{x} - \mu_2) - (\mathbf{x} - \mu_1)'\Sigma_1^{-1}(x - \mu_1)\right] > \ell n \frac{1-p}{p} \tag{11}$$

Similarly, if the parameters are unknown we will estimate them by their maximum likelihood estimates and obtain a sample QDF. It should be noted here that if the observations $\mathbf{x}$ are transformed to $\mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{b}$ where $\mathbf{A}$ is some non-singular matrix and $\mathbf{b}$ is a vector, then the assignment rule is invariant under this transformation. It is also easy to show that if Σ_1 and Σ_2 are equal, then $Q_T(\mathbf{x}) = D_T(\mathbf{x})$, the Linear Discriminant Function. A special case in which the means are equal (and without loss of generality they may be taken equal to $\mathbf{0}$), arises in twin studies where one is desiring to determine if a pair is homozygotic or heterozygotic. The assignment rule then is given in Equation 12, and all of the distinction between the two populations is based on differences in spread.

$$Q_T(\mathbf{x}) = \tfrac{1}{2}\left\{\ell n \frac{|\Sigma_2|}{|\Sigma_1|} + \mathbf{x}'(\Sigma_2^{-1} - \Sigma_1^{-1})\,\mathbf{x}\right\}. \tag{12}$$

In constructing a discriminant function, one is immediately faced with problem of variable selection. This is still an open problem with much research currently being done. A reasonable criterion for selecting a subset of variables is that it be based on minimizing a function of the error rates (or, maximizing a function of the non-error rates). Stepwise selection algorithms are available in the BMD package based on an F-test criterion. No results are currently available for quadratic rules.

3. *Classification with Discrete Variables*

3.1. *Theoretical Basis*

When multivariate data contain categorical or qualitative variables, the assumptions underlying the use of the linear or quadratic discriminant functions no longer apply. Data of

this type has led to an interest in the development and properties of classification rules which for the most part assume underlying multinomial distributions.

Aside from the realization that linear discriminants might perform poorly especially for discrete data yielding nonmonotone likelihood ratios (see Moore 1973) there is the added incentive of favorable error rates. Glick (1973) showed for the two group multinomial problem that both the mean apparent and mean actual error converge at least exponentially to the Bayes error provided that no state has the same discriminant score under both populations; no comparable rate under normal theory is known.

We suppose that data are generated by discrete random variables $X_1, X_2, \ldots, X_p$ each assuming at most a finite number of distinct values $s_1, s_2, \ldots s_p$ which then generate a sample space consisting of $s = \Pi_j s_j$ states. If the class conditional multinomial densities in π_i are h_i and the priors are p_i, $i = 1, 2$ then allocation of a future observation $\mathbf{z}$ to π_1 occurs if $g_1(\mathbf{z}) = p_1h_1(\mathbf{z}) > p_2h_2(\mathbf{z}) = g_2(\mathbf{z})$. The $g_i(\mathbf{z})$ are known as discriminant scores. If $g_1(\mathbf{z}) = g_2(\mathbf{z})$ assignment is done randomly, otherwise $\mathbf{z}$ is assigned to π_2.

When n observations are sampled from the mixed population, the number $N_i(\mathbf{x})$ from π_i with $\mathbf{X} = \mathbf{x}$ and the total number from π_i, $N_i = \sum N_i(\mathbf{x})$ are binomial random variables. Frequency estimates employed for prior probabilities and conditional densities generate estimates of discriminant scores

$$\hat{g}_i(\mathbf{x}) = (N_i/n)[N_i(\mathbf{x})/N_i] = N_i(\mathbf{x})/n \tag{13}$$

Although a sample-based rule in terms of $\hat{g}_i(\mathbf{x})$ has good large sample properties its implementation with small or moderate sample sizes is problematic. A few variables each containing only two or three levels results in a proliferation of states many of which contain no observations and hence no assignment outside of random allocation using $\hat{g}_i(\mathbf{x})$ is possible.

3.2, Models for Dichotomous Vectors

An approach that has been advocated by a number of writers to deal specifically with multivariate dichotomous observations relates to model representatives of the joint density. Bahadur (1961) showed that the joint density of p dichotomous random variables (X_1, $X_2, \ldots, X_p$) can be represented by

$$f(x_1, x_2, \ldots, x_p) = \prod_{j=1}^{p} \theta_j^{x_j}(1-\theta_j)^{1-x_j}\left\{1 + \sum_{j<k}\rho_{jk}z_jz_k + \ldots + \rho_{12\ldots p}z_1z_2\ldots z_p\right\}$$

where $E(X_j) = \theta_j$, $Z_j = (X_j - \theta_j)/SQRT[\theta_j(1-\theta_j)]$ and

$$\rho_{jk} = E(Z_jZ_k), \ldots, \rho_{12\ldots p} = E(Z_1Z_2\ldots Z_p). \tag{14}$$

Martin and Bradley (1972) and Ott and Kronmal (1976) represent the joint density in terms of linear combination of orthogonal polynomials. The Martin and Bradley model utilizes a class of orthogonal polynomials defined by

$$\psi(\mathbf{x}) = 1; \quad \psi_j(\mathbf{x}) = 2x_j - 1, j = 1, 2, \ldots p$$

$$\psi_{\mathbf{v}}(\mathbf{x}) = \prod_{j=1}^{k} \psi_{v_j}(\mathbf{x}), \mathbf{v} = (v_1, v_2, \ldots v_k), v_1 < v_2 < \ldots < v_k, k = 2, 3, \ldots p;$$

$$v_j \in \{1, 2, \ldots, k\}. \tag{15}$$

The Ott-Kronmal representation is in terms of polynomials

$$\psi_{\mathbf{r}}(\mathbf{x}) = (-1)^{\mathbf{x}'\mathbf{r}} \tag{16}$$

where $\mathbf{r}$ is an indexing vector assuming values in the state space generated by $\mathbf{X}$.

For classification purposes all three models are basically used in the same fashion. The idea is to achieve a degree of parsimony in modeling with the hope that the induced classification yields reasonable results. Preliminary evidence indicates that the Martin and Bradley model better deals with the sparseness problem than the other two. The Bahadur model on the other hand is defined in terms of more familiar parameters and hence would probably be better understood by the practitioner. Of the three however, the Ott and Kronmal model has a simple decision rule which allows for an objective determination as to whether a given parameter should be included in the model. Although all three, in varying degrees, allows some control in choosing a workable reduced model, there is no satisfactory distributional result which assists in the determination. What we usually find is the somewhat arbitrary elimination of higher order parameters as a method of approximating the density before defining the rule.

3.3 The Loglinear Representation

None of the above approaches allow for a satisfactory degree of control in the construction of models for state probabilities. However, borrowing results from the established literature on multidimensional contingency table analysis leads through the use of loglinear models to a powerful discrete classification methodology. Such models allow us to use goodness of fit statistics as a tool of model building, provide a potential handle on the sparseness issue and permit the incorporation of information pertaining to orderings in the variables.

As an illustration of these ideas suppose we consider a three variable problem (X_1, X_2, X_3) which under either Poisson, multinomial or product multinomial sampling generates a table of dimension $2 \times J \times K$. We view the first variable as one identifying the two groups of interest, that is $X_1 = 1 \Rightarrow \pi_1$. Suppose we represent the logarithm of the theoretical probability in cell (i, j, k) of the table by

$$\ln p_{ijk} = u + u_1(i) + u_2(j) + u_3(k) + u_{12}(ij) + u_{13}(ik) + u_{23}(jk) + u_{123}(ijk) \qquad (17)$$

where the u-terms satisfy similar constraints as in the analysis of variance and in addition satisfy the hierarchical principle (Bishop, Fienberg, Holland 1975).

An unsaturated model is determined by a set of sufficient configurations. Specification of a set of sufficient configurations determines which u-terms are hypothesized to be zero. Under the specified model, maximum likelihood estimators are readily generated either directly or through some successive approximation algorithm. The fitted table values can then be compared to the observed values by some fit statistic like the likelihood ratio or Pearson statistic to determine if the model appears to be consistent with the data. Once an acceptable model is decided upon classification of any future observations into π_1 is made if the log odds in favor of π_1 for a given state exceeds some threshold.

The loglinear approach has a number of obvious advantages which single it out over previous methods discussed in the literature. State sparseness is a particularly distressing problem in discrete classification. Since most rules ultimately depend upon comparison of state probability estimates to effect a classification, and since zeros in general may mean different things, it becomes important to be able to derive meaningful nonzero estimates for those states containing zero frequencies. If sufficient configurations associated with zero marginal totals are fitted, then because of a result due to Birch (1963) relating to the generation of maximum likelihood estimators, all cell estimates summing to these totals will be zero. However, if only configurations are fitted which correspond to positive totals are used then we might be successful in both reducing the number of random zeros and providing

reasonable estimates on which to base our discriminant rule. This approach has been used on a number of different sparse data sets and preliminary results indicate that the method leads to successful rules both in providing for classifications which otherwise would be unsatisfactory, and in maintaining error rates at acceptable levels.

Frequently categorical variables represent points in an ordinal scale; attitudinal or preference variables are good examples. It is important that we utilize this information on the ordered structure of the categories for otherwise we are throwing away relevant aspects of the data. During the process of model building we might determine that certain u-terms associated with independence statements between the group defining variable and some other variables do not hold. Rather than fitting a more saturated model we might be more inclined to include terms which reflect an interaction structure between the group defining variable and the orderings of the remaining variables. Toward this end suppose that X_2 and X_3 have a natural ordering which can be represented by specified scores $v_1^{(1)}, v_2^{(2)}, \ldots, v_J^{(2)}$; $v_1^{(3)}, v_2^{(3)}, \ldots, v_K^{(3)}$ respectively. A model which reflects the association between X_1 and the ordered structure of X_2 and X_3 is given by

$$\begin{aligned} \ln p_{ijk} = u + u_{1(i)} + u_{2(j)} + u_{3(k)} + u_{23(jk)} + (v_j^{(2)} - \bar{v}^{(2)})u_{1(i)}^{(2)} \\ + (v_K^{(3)} - \bar{v}^{(3)})u_{1(i)}^{(3)} + (v_j^{(2)} - \bar{v}^{(2)})(v_K^{(3)} - \bar{v}^{(3)})u_{1(i)}^{(23)}, \end{aligned} \tag{18}$$

where $\bar{v}^{(i)}$ is the average score for variable i and the terms $u_{1(1)}^{(2)}$, $u_{1(i)}^{(3)}$ and $u_{1(i)}^{(23)}$ reflect the association between X_1 and the ordered nature of X_2 and X_3. Maximum likelihood estimates can readily be found using versions of the iterative proportional fitting algorithm like Haberman's Newton-Raphson algorithm (1974). This approach has been used by a number of people, usually with good success. Although it has not been discussed in terms of the classification problem per se it certainly should be applied in situations where the ordered nature can be captured.

3.4. Variable Selection

Selecting a "good" set of variables for discrimination using discrete variables is of considerable interest. Whereas in contingency table analysis attempts are made to find a parsimonious model so that one can better understand relationships between variables, efforts to reduce the dimensionality for discrimination are motivated by conditions of sparseness and economy. The literature on variable selection procedures in the context of discrete classification is of recent vintage and only a few papers have been published which specifically utilize the distributional properties of the data.

Lachin (1973) studied a stepwise selection procedure which chooses variables on the basis of the independence of the group defining variable and a new candidate for inclusion given that a set of variables have already been chosen. In essence the method measures the change in the degree of association between the grouping variable and the remaining variables when an additional one is added to the predictive set.

Goldstein and Rabinowitz (1975) utilized the error rate bounds derived by Glick (1973) to suggest a decision rule for defining an optimal subset of variables. The procedure is not in a stepwise mode but selects variables for inclusion on the basis of maximizing a pseudo distance between discriminant scores.

Recently Goldstein and Dillon (1977) presented a stepwise selection procedure which uses distributional properties of a Kullback minimum discrimination divergence statistic. The method allows for additions of new variables given the levels of those previously included. The idea behind the method is that for certain levels of previously included

variables, new candidates will contribute to discriminating between the two groups while for other levels the same variable will contribute little. In general, this procedure will dramatically reduce the number of patterns that need to be considered, and on the basis of a number of data sets to which it has been applied, impacts little on increasing estimated error rates.

In Goldstein and Dillon (1978), Chapter Four deals specifically with the problem of variable selection and illustrates the above mentioned procedures. Further, Chapter Six presents computer programs with full documentation which operationalizes all the variable selection algorithm in addition to other procedures in discrete classification.

4. Current Research Areas

This section will discuss recent developments in discriminant analysis research. We shall consider work on robustness of linear and quadratic discriminants, nonparametric and density estimate discriminants, and recent work on problems involving mixtures of categorical and continuous variables. We specifically are not reviewing Bayesian methods since there is little difference between them and conventional methods if samples are of moderate size. For a recent review of Bayesian methods, see Giesser's article in Van Ryzin (1977).

4.1. Robustness Studies

Robustness of an allocation rule may be broadly defined as a small effect on error rates when the underlying assumptions fail to hold. This definition can be made more precise (see, for example, Huber 1977) but for our purposes this will suffice. We will talk specifically about four types of robustness studies that have been in progress and indicate some possible future areas of research.

The simplest robustness property is the robustness to misclassification of the initial or training samples. It was shown by Lachenbruch (1966) that if the initial samples are randomly misclassified at the same rate in both populations, then there is no effect on the linear discriminant function. In fact the effect on the linear discriminant function is related to the difference in rates of initial misclassification and so if there is a small amount of misclassification, there is very little effect. More precisely, if α_1 is the rate of misclassification in the first sample and α_2 is the rate of misclassification in the second sample and equal prior probabilities hold, then the probabilities of misclassification of a future observation are given by Equation 19.

$$P_{1M} = \Phi\left[\frac{-\delta(1 + \alpha_1 - \alpha_2)}{2}\right] \quad \text{and} \quad P_{2M} = \Phi\left[\frac{-\delta(1 - \alpha_1 + \alpha_2)}{2}\right], \tag{19}$$

where P_{iM} is the error rate in the i^{th} population when initial misclassification is present. This model is presumably not a realistic model for initial misclassification as it is more likely that subjects near the other population will be misclassified than individuals which are very clearly one of the populations. This problem was studied and it was found that the actual error rates were relatively unaffected by non-random initial misclassification (Lachenbruch 1974). However, the apparent error rates were grossly distorted and could not be relied upon for any sample sizes. Recently a study has been completed on the effect of initial misclassification on the quadratic discriminant function (Lachenbruch 1979). It was found that with a random misclassification model the QDF does not display any of the insensitivity to initial misclassification that the linear function appears to. Both error rates are adversely affected. The effects are proportional to the amount of contamination and the relative

fraction of initial misclassification that appears. No studies have been performed to date on the effect of initial misclassification on categorical discriminant functions. However, such effects should not be dissimilar to those noted in studies of misclassification in multi-way contingency tables.

4.2. Contamination Studies

Contamination may arise in one of two ways. One or both of the populations may have a small fraction of the population which comes from a distribution with the same mean as the major part of the population but a variance somewhat greater than the true population. This may occur when a measuring instrument varies to a greater extent than usual. Alternatively, there may be different measuring instruments which have different variances. A second kind of contamination could arise when an instrument slips in calibration and gives biased readings on a fraction of the points. The first kind of contamination is called scale contamination and the second kind is called location contamination. This is sometimes referred to as the outlier problem. Various models may be postulated for these two kinds of contaminations. The simplest scale contamination model is that the contaminated fraction, say α, of the initial sample has its covariance matrix multiplied by a scalar (> 1). Thus each variable in the population is contaminated by the same proportion and the same degree of contamination in terms of the scalar multiplier. Another model of scale contamination would be that different fractions of a given subset of variables are contaminated by different scalar multipliers. This situation may be exemplified for the linear discriminant function as in Display 1.

Display 1

Possible Contamination Models in Linear Case

$$\Pi_1\colon : N(\underline{\mu}_1, I) \qquad \Pi_2\colon N(\underline{\mu}_2, I)$$

$$\Pi_1{}^c\colon \sum_{i=1}^{k} \beta_{1i}\, N(\underline{\mu}_1, I + d_{1i}J_i) \qquad \Pi_2{}^c\colon \sum_{i=1}^{k} \beta_{2i}\, N(\underline{\mu}_2, I + d_{2i}J_i)$$

where J_i is a diagonal matrix with a 1 in the i^{th} diagonal element.

Under the pure scale contamination model in which all variables are contaminated the same fraction and the same amount, Ahmed and Lachenbruch (1975) have provided asymptotic formulas for the error rates in these two populations. With mild contamination, relatively small effects are observed. If the contaminating matrices differ substantially from the uncontaminated covariance matrix, a clearly harmful effect on contamination can be observed. In a small sample study Ahmed and Lachenbruch (in Van Ryzin 1977) considered possible remedial measures in the scale contamination problem. They considered a number of possible robust discriminants based upon variable-wise robust procedures. These included trims, hubers, hampels, tri-means and so forth. What was found indicated there was practically no loss when any of the robust procedures were applied to the data as compared to the LDF. For mild single sample contamination the LDF performed as well as the robust procedure. In general, however, the LDF does not perform as well when the initial samples have been scale contaminated. For moderate and heavy contamination the LDF's based on variate-wise hampel, huber and sin procedures tend to be superior to the usual LDF. The procedures based on robust regressions did not perform well.

The effects of location contamination can be more harmful than scale contamination depending on the direction of contamination and fraction of contaminating observations. The worst effects occur when the mean of the contaminating distribution is on the opposite side of the mean of the uncontaminated population. The asymptotic model given by Ahmed and Lachenbruch for location contamination is rather a specific one in that all components are assumed to be contaminated to the same extent and by the same contaminating distribution. A more realistic model would include situations in which only a single variable was contaminated. In particular this could include models as given from Display 1. In some current work Broffitt, Lachenbruch and Clarke (1978) have noted that for a model which allows different individual variates to be contaminated at different rates and with different contaminating distributions, the effect on the linear discriminant is fairly small and probably does not require any corrective action if the contamination fraction is less than 10%. No studies are available at present regarding the effects of scale or location contamination on quadratic discriminants.

4.3. Non-normality

Several studies have been conducted on the effects of various types of non-normality on the quadratic and linear discriminant functions. The earliest studies were concerned with the performance of the linear discriminant function on sets of dichotomous variables. It was found by several investigators that if the distribution of the dichotomous variables did not exhibit high level interaction, reflected through the non-monotonicity of the likelihood ratio, the LDF performed satisfactorily for allocating future observations (Gilbert 1968, Moore 1973). Other forms of non-normality were studied using the Johnson system of transformations which includes the log normal, the inverse hyperbolic sine normal, and the logit normal distribution. The first such study indicated a considerable decline in performance of the linear discriminant function but these results are somewhat clouded because the log-normal distribution used had extremely large skewness and kurtosis for each marginal distribution (Lachenbruch, Sneeringer and Reno 1973). Perhaps with less extreme skewness and kurtosis the linear discriminant might have been shown to be not so badly affected. This non-normality caused the probability of misclassification in one group to be increased substantially over the optimal while the probability of misclassification in the other group was decreased. Overall there was an increase in the average probability of misclassification. The average probability of misclassification tends to be insensitive to changes in distributional form because the error rates in the two groups have a large negative correlation. The problems with the large skewness and kurtosis in the log-normal distribution appears also in work by other authors [e.g., Koffler (1976)). The effects of non-normality on the quadratic discriminant have been studied recently by Clarke, Lachenbruch and Broffitt (1978). They avoided the large kurtosis and skewness problem. They found:

a) the between-sample variability of the individual error rates in the QDF on normal or non-normal distributions was quite large;

b) the actual error rates were considerably larger than the optimal rates;

c) the QDF applied to non-normal samples generally did not do substantially worse than when the QDF was applied to the normal samples which would be obtained after transformation;

d) the individual error rates were affected much more than the average error rates;

e) attempting to obtain a robust estimate (e.g. hubers, trims) of means and covariances was of little help unless the distribution was substantially skewed;

f) heavy kurtosis did not affect the performance of the QDF.

4.4. Unequal Covariance Matrices

If one attempts to use the linear discriminant function when in fact the covariance matrices are unequal, the performance of the linear discriminant may be substantially affected. Gilbert (1969) and Marks and Dunn (1974) studied the performance of the linear discriminant function under this violation. They both found that the linear discriminant function is satisfactory if the covariance matrices are not too different. It should also be pointed out that if the means are widely separated the linear function should generally do well. Marks and Dunn found that if the sample sizes are small the QDF did quite poorly. This is an agreement with the findings of Clarke, Lachenbruch and Broffit, who found that the between-sample variability of the QDF was quite large.

4.5. Classification Rules Derived Through Density Estimates

Welch's result shows that if we want to classify an object that comes from one of two populations having associated densities f_1 and f_2, then a classification rule should be based upon the likelihood ratio f_1/f_2. The induced rule, under the assumption of sampling from normal populations with equal covariance matrices results in determining on which side of a certain hyperplane of constant likelihood the measurement vector lies. When the covariance matrices are different the rule makes the determination on the basis of a hyperquadratic. Although studies have indicated that these procedures perform reasonably well for some non-normal populations, there is interest in approaching the problem in a less constrained way.

Perhaps the most direct way to proceed with the more general problem is to estimate the likelihood ratio f_1/f_2 by estimating the individual densities f_1 and f_2. It would make better sense to estimate the likelihood ratio, however, the state of research is such that more is known of density estimation than likelihood ratio estimation.

The literature on nonparametric density estimation is extensive (see e.g., Wegman 1972). A recent paper principally dealing with optimal convergence properties of various classes of nonparametric density estimates, is Wahba (1975).

Of what advantage is nonparametric density estimation in the classification problem? Recall that if f_1 and f_2 are the two underlying continuous densities from subpopulations π_1 and π_2, and p_1 and p_2 are the two prior probabilities, then if $D = (D_1, D_2)$ is a partition of the sample space χ such that $x \in D_j \Rightarrow \pi_j$ for any $x \in \chi$, the probability of correct classification for the partition or rule D is given by

$$r(D) = \sum_{j=1}^{2} \int_{D_j} p_j f_j(x)\, dx \,. \tag{20}$$

If $\mathfrak{D}$ is the collection of possible classification rules, the D^* achieves the optimal probability of correct classification if

$$r(D^*) = \sup_{D \in \mathfrak{D}} r(D) = r^* \,. \tag{21}$$

As we have seen, Welch showed that D^* is defined by

$$D_j^* = \left\{ x \,|\, j \text{ is the smallest integer such that } p_j f_j(x) = \max_{1 \le j \le 2} p_j f_j(x) \right\} . \tag{22}$$

Suppose now that a random sample of n observations is available from π (the mixture of π_1 and π_2) and that through sufficient effort an identification from the correct subpopulation for each sample point can be made. If it is determined that N_j of the total sample n are from

π_j, then we can estimate p_j by $\hat{p}_j = N_j/n$. Further, denote by $\hat{f}_j(x)$, an estimate based upon the sample of the density f_j at the point x. If we denote by $\hat{D}_j$ the partition given in (22) with p_j and $f_j(x)$ replaced by $\hat{p}_j$ and $\hat{f}_j(x)$, then with

$$\hat{r}(D) = \sum_{j=1}^{2} \int_{D_j} \hat{p}_j \hat{f}_j(x)\, dx \tag{23}$$

it can be shown that

$$\hat{r}(\hat{D}) = \sup_{D \in \mathfrak{D}} \hat{r}(D) \tag{24}$$

In a fundamental paper dealing with classification rules derived through density estimates, Glick (1972) showed that $\sup|\hat{r}(D) - r(D)| \to 0$, $r(\hat{D}) \to r^*$ and $\hat{r}(\hat{D}) \to r^*$ provided the density estimates are consistent and that $\int \sum_j \hat{p}_j \hat{f}_j(x)$ converges to unity or is bounded by a finite constant.

Thus, it is clear that under reasonable conditions on the density estimators, such induced rules have very desirable error-rate convergence properties. Note in particular that these results do not assume any restrictive conditions relating to the underlying family of distributions generating the data.

Why then have practitioners shied away from utilizing nonparametric density estimators as a methodology in discriminant analysis? The answer to this question is best served by briefly discussing just two of the many estimators that have been proposed.

(I) Kernel Estimates:

Based upon a sequence of i.i.d. random variables $X_1, X_2, \ldots, X_n$ a univariate kernel estimate assumes the form

$$\hat{f}_n(x) = \frac{1}{n} \sum_{j=1}^{n} K_n(X_j, x) = \frac{1}{nh(n)} \sum_{j=1}^{n} K\left[\frac{X_i - x}{h(n)}\right]. \tag{25}$$

$K(\cdot)$ is a function that satisfies a number of regularity conditions (see Parzen (1962)). For $\hat{f}_n(x)$ to be a mean square error consistent estimator of $f(x)$, $\lim h(n) \to 0$ and $\lim nh(n) \to \infty$ as $n \to \infty$ is required. Kernel estimates have received much study in the direction of improved convergence properties, general representations for $K_n(x, y)$ and extension to the multivariate problem.

(II) Orthogonal Series Estimates:

Schwartz (1967) considers estimates of the form

$$\hat{f}_n(x) = \sum_{j=0}^{q(n)} \hat{a}_j g_j(x) \tag{26}$$

where $g_j(x)$ is the sequence of normalized Hermite functions

$$g_j(x) = (2^j j! \pi^{1/2})^{-1/2} e^{-x^2/2} H_j(x), j = 0, 1, \ldots \quad \text{and} \quad H_j(x) = (-1)^j e^{x^2} (d^j/dx^j) e^{-x^2} \tag{27}$$

where $\hat{a}_j = \sum_i g_j(X_i)/n$.

The parameter $q(n)$ controls the tradeoff between the square bias and the variance. The smaller $q(n)$ is taken the smaller is the variance; large $q(n)$ is associated with small bias but large variance. If $q(n) \to \infty$ and $q(n)/n \to 0$ then $\hat{f}_n(x)$ is consistent in quadratic mean. Schwartz extends his results to the multivariate problem.

In the two classes of estimators discussed above a crucial decision regarding the choice of the parameters $h(n)$ and $q(n)$ need to be made. For the estimators to work reasonably well "proper" assignments are required. Unfortunately the state of research is such that little guidance is available on how to do this especially for estimation in more than one dimension.

For the kernel and orthogonal series estimators there has been some work on how to choose these smoothing parameters; unfortunately they are of little practical value. Recently, however Habbema *et al.* (1974) suggested a leaving-one-out modification of maximum likelihood estimation as an approach to choosing a sample based smoothing parameter. The procedure makes intuitive sense and probably will be a useful mechanism to assist the practitioner in formulating a reasonable density estimator. Recently Wahba (1977) has discussed a new class of estimators with a viable algorithm for estimating the optimal (in the sense of integrated mean square error) smoothing parameter from the data. Although the methods are extendable to more than one dimension they are not operational and probably quite difficult to use. Until practitioners are more secure in how to arrive at smoothing parameters so that the induced classification procedures have fair chance at competing with existing methodology, the area will probably still stay dormant.

Nonetheless, however, we believe this is an area worthy of much additional study and use even if no analytic results become available to help in the smoothing problem. One limited study by Goldstein (1975) showed for some bivariate normal populations that two classification procedures evolved from nonparametric density estimators can be made to work well.

4.6. Nonparametric Rules

The phrase nonparametric classification to most means rules for classification triggered by likelihood ratios formed from nonparametric density estimators. Having already discussed this in a previous section, we wish to point to another part of the literature that deals with rules not motivated in the usual sense through consideration of estimated likelihood ratios. As a general statement all such procedures rely on "distances" between points or distributions to effect a classification.

Fix and Hodges (1951) is probably the first paper to address the nonparametric classification problem. Their paper motivated much research activity on nonparametric density estimation, and showed for the first time how a consistency property for induced classification rules can be derived. They were the first to discuss the k nearest neighbor rules and showed them to be Bayes risk consistent provided k and the sample size increase to infinity in a prescribed way. The rule assumes a sample from a mixture of two populations with continuous distributions and the ability to identify which subpopulation generated each point. For a new observation z to be classified, a determination is made as to which population generated the k (a preassigned integer) nearest sample points from z. If the majority of these sample points are from π_j, then z is assigned to π_j. Since the initial work by Fix and Hodges these rules have received much study, the latest by Rogers and Wagner (1978).

Procedures based upon distributional distances account for a small literature and hence it is not surprising that their impact in applied work has been rather limited. Gupta (1963) considered a minimum distance rule based upon the Kolmogorov-Distance. The procedure assumes independent samples selected from k populations $\pi_1, \pi_2, \ldots, \pi_k$, and requires an evaluation of the sample c.d.f.'s $S_1, S_2, \ldots, S_k$. A new set of observations become available and they are known to have been generated from one and only one π_i, $i = 1, 2, \ldots, k$. Classification of the set to π_j is made provided

$$\min_{1 \le i \le k} D(S_0, S_i) = D(S_0, S_j) \tag{28}$$

where S_0 is the sample c.d.f. of the new sample and $D(.,.)$ represents the Kolmogorov-Distance. Gupta shows that the rule given in (28) leads to a consistent procedure.

Earlier work by Matusita (1955) is similar in spirit to Gupta's minimum distance methods. However, Matusita uses a different distributional distance and applies his methodology to Gaussian and multinomial populations. More recently Dillon and Goldstein (1978) used this distance measure in defining a new procedure for the two group multinomial problem. Monte Carlo studies reported in their paper indicate that the procedure can be most effective when the two pilot samples are disproportionate in size.

Perhaps the most interesting and potentially powerful nonparametric classification methodology is attributable to Gorden and Olshen (1978). Although their attack to classification is through estimating a pair of densities, their densities are perfectly general in that they can be discrete, continuous, mixed or singular. Their approach involves a successive partitioning algorithm of the sample space into boxes. Each partition of interest is a result of successive refinements of previous partitions with ultimately a classification made by majority vote within each box of the final partition. In a sense therefore the procedure is a very general extension of the k nearest neighbor rules and those rules which utilize nonparametric density estimators but assume the underlying distribution to be absolutely continuous with respect to Lebesgue measure. Their principal contribution is that this partitioning scheme along with the majority rule will asymptotically yield error rates equal to that of the Bayes procedure.

4.7. Mixtures of Variables

In the social, behavioral and biological sciences it is common for multivariate data to contain both continuous and discrete components. It is of considerable interest to arrive at reasonable classification rules which consider jointly all the important vector components without making the erroneous assumption of multivariate normality. As discussed in the previous section the work of Gordon and Olshen can potentially be very helpful in dealing with the mixed variable problem.

Recently Krzanowski (1975, 1976, 1977) considered the following two group problem. Let $\mathbf{X}$ be a q-dimensional binary vector generating $k = 2^q$ states. Conditional on $\mathbf{X} = \mathbf{x}$ being in state m and the observations from π_i, let $\mathbf{Y}$ be a p-dimensional normal random vector with mean $\mathbf{u}_i^{(m)}$ and covariance matrix $\mathbf{\Sigma}$, $i = 1, 2$. Letting $\mathbf{W} = (\mathbf{X}, \mathbf{Y})$ and denoting the density of $\mathbf{W}$ *in* π_i *by* $h_i(\mathbf{w}) = f_i(\mathbf{y}|\mathbf{x})f_i(\mathbf{x})$, *then assuming equal priors the optimal partition in the sense* of Welch is

$$D_1 = \{\mathbf{w} | (\mathbf{u}_1^{(m)} - \mathbf{u}_2^{(m)})' \mathbf{\Sigma}^{-1}[\mathbf{y} - \tfrac{1}{2}(\mathbf{u}_1^{(m)} + \mathbf{u}_2^{(m)})] \geq \log [f_2(\mathbf{x})/f_1(\mathbf{x})] \quad \text{and} \quad D_2 = D_1^c \tag{29}$$

The optimal rule reduces to k linear discriminants; one for each of the $k = 2^q$ states. It is straightforward to show that if $P(i|j)$ is the probability of misclassifying an observation from π_j then

$$P(2|1) = \sum_{m=1}^{k} P_{1m}\Phi\{[\ell n(P_{2m}/P_{1m}) - D_m^2/2]/D_m\}$$

and

$$P(1|2) = \sum_{m=1}^{k} P_{2m}\Phi\{[\ell n(P_{1m}/P_{2m}) - D_m^2/2]/D_m\} \tag{30}$$

where for $\mathbf{x}$ belonging to state m, $P_{im} = f_i(\mathbf{x})$ and

$$D_m^2 = (\mathbf{u}_1^{(m)} - \mathbf{u}_2^{(m)})' \mathbf{\Sigma}^{-1}(\mathbf{u}_1^{(m)} - \mathbf{u}_2^{(m)}).$$

For the sample-based version of (29), Krzanowski used a log-linear model to estimate $\{P_{lm}\}$ and a linear additive model to estimate $\mathbf{u}_i^{(m)}$ and $\boldsymbol{\Sigma}$, $i = 1, 2$; $m = 1, 2, \ldots, k$.

Using this methodology, extensions are immediate to the case where the discrete components take on more than two values. In general if X_j assumes s_j distinct values then the number of linear discriminants needed will be equal to $\Pi_j s_j$. In most instances, unless large amounts of data are available the number of sample points in many states will be too small for any reasonable estimations of mean values $\mathbf{u}_i^{(m)}$ or cut-off ratios P_{1m}/P_{2m}. Further, one need not assume the same covariance matrix under π_1 and π_2; if this is dropped then in general the sample space would be partitioned into $\Pi_j s_j$ hyperquadratics.

General extensions of these ideas occur when the assumption of conditional normality is no longer postulated. In theory this should cause no real difficulty since we assume that conditional on $\mathbf{X}$, $\mathbf{Y}$ has some nonspecified multivariate distribution with continuous components. Sample-based rules from such an assumption can evolve through utilization of appropriate nonparametric density estimators for the joint conditional distribution of $\mathbf{Y}|\mathbf{X}$. Again the two problematic areas in this approach relate to specification of smoothing parameters and the limited availability of data in given states. Since the sparseness issue is undoubtedly more of a problem in the general model than in the normal model, an alternative approach might be to utilize some similarity measure to collapse states so that more data is available on which to estimate the densities.

The mixed variable problem will, because of its practical importance, receive more notice by both theoreticians and practitioners. It is an area worthy of additional study.

Acknowledgment

This work was partially supported by Grant GM-23496 from the National Institutes of Health.

References

Ahmed, S. and Lachenbruch, P. A. (1975). Discriminant analysis when one or both of the initial samples is contaminated: large sample results. *EDV in Medizin und Biologie 6,* 35–42.

Bahadur, R. R. (1961). A representation of the joint distribution of responses to *n* dichotomous items. *Studies in Item Analysis and Prediction.* H. Solomon, ed., Stanford, California: Stanford University Press, 158–168.

Birch, M. W. (1965). Maximum likelihood in three-way contingency tables. *Journal of the Royal Statistical* Society *B25,* 220–233.

Bishop, Y., Fienberg, S., and Holland, P. (1975). *Discrete Multivariate Analysis,* MIT Press, Cambridge.

Broffitt, B., Lachenbruch, P. A. and Clarke, W. R. (1978). On the effects of location contamination on the linear discriminant function. In preparation.

Clarke, W. R., Lachenbruch, P. A. and Broffitt, B. (1978). Robustness of the quadratic discriminant function to non-normality. Submitted to *Communications in Statistics.*

Dillon, W. R. and Goldstein, M. (1978). On the performance of some multinomial classification rules. *Journal American Statistical Association 73,* 305–313.

Fisher, R. A. (1936). The utilization of multiple measurements in taxonomic problems. *Annals of Eugenics 7,* 179–188.

Fix, E. and Hodges, J. L. (1951). Discriminatory analysis, nonparametric discrimination: consistency properties. Report No. 4, Project No. 21-49-004, USAF School of Aviation Medicine, Brooks Air Force Base, Texas.

Gilbert, E. S. (1968). On discrimination using qualitative variables. *Journal of the American Statistical Association 63,* 1399.

Gilbert, E. S. (1969). The effects of unequal variance covariance matrices on Fisher's linear discriminant function. *Biometrics 25,* 505–516.

Glick, N. (1972). Sample-based classification procedures derived from density estimators. *Journal of the American Statistical Association 67,* 116–122.
Glick, N. (1973). Sample-based multinomial classification. *Biometrics 29,* 241–256.
Goldstein, M. (1975). Comparison of some density estimate classification procedures. *Journal of the American Statistical Association 70,* 666–669.
Goldstein, M. and Dillon, W. R. (1977). A stepwise discrete variable selection procedure. *Communications in Statistics 6,* 1423–1436.
Goldstein, M. and Dillon, W. R. (1978). *Discrete Discriminant Analysis,* John Wiley and Sons, New York.
Goldstein, M. and Rabinowitz, M. (1975). Selection of variates for the two-group multinomial classification problem. *Journal of the American Statistical Association 70,* 776–781.
Gordon, L. and Olshen, R. A. (1978). Asymptotically efficient solutions to the classification problem. *Annals of Statistics 6,* 515–544.
Gupta, S. D. (1963). Nonparametric classification rules. *Sanhkya Series A,* 25–30.
Haberman, S. J. (1974). Loglinear models for frequency tables with ordered classifications. *Biometrics 30,* 589–600.
Habbema, J. D. F., Hermans, J. and van den Broek, K. (1974). A stepwise discriminant analysis program using density estimation. Compstat, 1974, Proceedings in Computational Statistics, Wien, Physica Verlag, 101–110.
Huber, P. (1977). *Robust Statistical Procedures.* Philadelphia: Society for Industrial and Applied Mathematics.
Koffler, S. (1976). An evaluation and comparison of discrimination procedures for certain types of non-normal distributions. Paper presented at American Education Research Association Meetings, 1976.
Krzanowski, W. J. (1975). Discrimination and classification using both binary and continuous variables. *Journal of the American Statistical Association 70,* 782–790.
Krzanowski, W. J. (1976). Canonical representation of the location model for discrimination or classification. *Journal of the American Statistical Association 71,* 845–848.
Krzanowski, W. J. (1977). The performance of Fisher's linear discriminant function under non-optimal conditions. *Technometrics 19,* 191–200.
Lachenbruch, P. A. (1966). Discriminant analysis when the initial samples are misclassified. *Technometrics 8,* 657–662.
Lachenbruch, P. A. (1974). Discriminant analysis when the initial samples are misclassified II: Non-random misclassification models. *Technometrics 16,* 419–424.
Lachenbruch, P. A. (1975). *Discriminant Analysis.* Hafner, New York.
Lachenbruch, P. A. (1979). Note on initial misclassification effects on the quadratic discriminant function to non-normality. *Technometrics,*
Lachenbruch, P. A., Sneeringer, C. and Revo, L. (1973). Robustness of the linear and quadratic discriminant function to certain types of non-normality. *Communications in Statistics 1,* 39–57.
Lachin, J. M. (1973). On a stepwise procedure for two populations: Bayes decision rules using discrete variables. *Biometrics 29,* 551–564.
Marks, S. and Dunn, O. J. (1974). Discriminant functions when covariance matrices are unequal. *Journal of the American Statistical Association 69,* 555–559.
Martin, D. C. and Bradley, R. A. (1972). Probability models, estimation, and classification for multivariate dichotomous populations. *Biometrics 28,* 203–222.
Matusita, K. (1955). On estimation by the minimum distance methods. *Annals of the Institute of Statistical Mathematics 7,* 67–77.
Moore, D. H. II (1973). Evaluation of five discrimination procedures for binary variables. *Journal of the American Statistical Association 71,* 339–404.
Ott, J. and Kronmal, R. A. (1976). Some classification procedures for binary data using orthogonal functions. *Journal of the American Statistical Association 71,* 391–399.
Parzen, E. (1962). On the estimation of a probability density function and the mode. *Annals of Mathematical Statistics 33,* 1065–1076.
Rogers, W. H. and Wagner, T. J. (1978). A finite sample distribution free performance bound for local discrimination rules. *Annals of Statistics 6,* 506–514.
Schwartz, S. C. (1967). Estimation of a probability density by an orthogonal series. *Annals of Mathematical Statistics 38,* 1261–1265.
Van Ryzin, J. (1977). *Classification and Clustering.* New York: Academic Press.

Wahba, G. (1977). Optimal smoothing of density estimates. In *Classification and Clustering*. Academic Press Inc., John Van Ryzin, Ed.
Wahba, G. (1975). Optimal convergence properties of variable knot, kernel, and orthogonal series methods for density estimation. *Annals of Statistics 3,* 15–29.
Wegman, E. J. (1972). Nonparametric probability density estimation I: A summary of available methods. *Technometrics 14,* No. 3, 533–546.
Welch, B. L. (1939). Note on discriminant functions. *Biometrika 31,* 218–220.