Statistical Science
1989, Vol. 4, No. 1, 34–69

Discriminant Analysis and Clustering

Panel on Discriminant Analysis, Classification, and Clustering

Abstract. The general objectives of this report are to provide a summary of the state-of-the-art in discriminant analysis and clustering and to identify key research and unsolved problems that need to be addressed in these two areas. It was prepared under the auspices of the Committee on Applied and Theoretical Statistics of the Board on Mathematical Sciences, National Research Council by its Panel on Discriminant Analysis, Classification, and Clustering. Both methodological and theoretical aspects are reviewed, and a survey of available software and algorithms is provided.

Key words and phrases: Agglomerative methods, algorithms, classification, evolutionary distances, high density clusters, kernel methods, logistic regression, pattern recognition, minimum spanning tree, mixtures, nearest neighbor methods, single linkage, complete linkage, ultrametric distances, variable selection, software, displays, and diagnostics.

1. INTRODUCTION

An interest in "classification" permeates many scientific studies and also arises in the contexts of many applications. From speech and speaker recognition problems in acoustics, to problems of numerical taxonomy in biology, and problems of classifying diseases by symptoms in health sciences, as well as problems of classifying artifacts in archaeology, or identifying market segments in market research, the central interest is in classifying "objects," "subjects" or entities of some kind. When the classification is based on measurements of a set of characteristics or variables, statistical techniques are available to aid the systematic process. The major concern of this report is with such statistical methods.

Classification is an inherently multivariate problem. Whether the interest is in deciding admissions to college, diagnosing a patient's illness for treatment purposes or pattern recognition in specific applications, the most likely scenario is one in which the data on hand pertain to many variables measured on each entity and not one involving just a single variable. This high-dimensional nature of classification provides an opportunity but also presents some difficulties to the developer of appropriate statistical methodology.

One can distinguish two broad categories of classification problems. In the first, one has data from known or prespecifiable groups as well as observations from entities whose group membership, in terms of the known groups, is unknown initially and has to be determined through the analysis of the data. For instance, one may have several repeated utterances of a specific word by different persons, and acoustic parameters extracted from each utterance labeled by the particular speaker would constitute the known replicate representations (also called training samples). In such a situation, if some additional utterances of the same word become available but one does not know from which person these utterances arose, one may need to make such an assignment statistically (i.e., the so-called speaker recognition problem) where the classification is with respect to the known speakers (groups). In the pattern recognition literature (see, e.g., Duda and Hart, 1973) this type of classification problem is referred to as *supervised pattern recognition* or *learning with a teacher*. In statistical terminology it falls under the heading of *discriminant analysis*.

On the other hand there are classification problems where the groups are themselves unknown *a priori* and the primary purpose of the data analysis is to

The members of the Panel on Discriminant Analysis, Classification, and Clustering were Ramanathan Gnanadesikan, Bellcore, Chairman; Roger K. Blashfield, University of Florida; Leo Breiman, University of California, Berkeley; Olive J. Dunn, University of California, Los Angeles; Jerome H. Friedman, Stanford University; King-Sun Fu, Purdue University (deceased); John A. Hartigan, Yale University; Jon R. Kettenring, Bellcore; Peter A. Lachenbruch, University of California, Los Angeles; Richard A. Olshen, University of California, San Diego; and F. James Rohlf, SUNY, Stony Brook. The corresponding authors for this article are R. Gnanadesikan and J. R. Kettenring, Bellcore, 445 South Street, Morristown, New Jersey 07960-1910.

determine the groupings from the data themselves so that the entities within the same group are in some sense more similar or homogeneous than those that belong to different groups. Many problems of numerical taxonomy, as well as market segments that are determined on the basis of demographics and psychographic profiles of people, provide examples of this second type of classification problem where the groups are data-dependent and not prespecified. This type of classification problem is referred to as *unsupervised pattern recognition* or *learning without a teacher*, and in statistical terminology falls under the heading of *cluster analysis*.

Although discriminant analysis and cluster analysis constitute a useful dichotomy of classification problems, there are of course many real-life problems that combine the features of both situations. One might have some preliminary or imprecise idea of the groups from which the data arise but wish some verification of the meaningfulness of the prespecified groups in certain problems. Some combination of the tools from the two types, or perhaps entirely different and as yet unavailable tools, may be appropriate for these situations.

The earlier-mentioned widespread prevalence of the classification problem (in all of its guises) in many fields, stimulated by the easy access to both numerical and graphical computing facilities, has seen the development of a plethora of new approaches and algorithms for discriminant analysis and cluster analysis in the last two decades. If one were to consider classification problems in three stages, viz. input, algorithms and output, it would be fair to say that the vast majority of the work has focused on the second of these. It is clear, however, that careful thought about what variables to use and how to characterize and/or summarize them as inputs to methods of classification are very important issues that would involve both statistical and subject matter considerations in applications. Similarly, the most challenging aspect of most analyses of data tends not to be the choice of a particular method but interpretation of the output and results of algorithms.

The three stages clearly interact with each other and statistical issues and methods play central roles in all three of them. To illustrate this point, the importance of choosing the variables and/or features to use initially for classification purposes has been mentioned. Nevertheless, despite the care with which this is done by a user, there may be a tendency to include "too many" rather than "too few" variables from the point of view of informativeness of the variables. (The opposite problem of using too few variables sometimes occurs, too, giving rise to poor results.) Sorting out the resultant redundancy among the variables and identifying those that have incremental statistically useful information for classification purposes are problems that can benefit from statistical methods for variable selection. It is usual to consider algorithms for variable selection as part of the process of understanding and interpreting the results of an initial application of a discriminant or cluster analysis procedure.

The discriminant analysis situation has been a more integral part of the historical development of multivariate statistics, although the cluster analysis case received most of its impetus from fields such as psychology and biology until relatively recently. In part, the lack of statistical emphasis in cluster analysis may be due to the greater inherent difficulty of the technical problems associated with it. Even a precise and generally agreed upon definition of a cluster is hard to come by. The data-dependent (presumably "random") nature of the clusters, the number of them and their composition appear to cause fundamental difficulties for formal statistical inference and distribution theory. Except for *ad hoc* algorithms for carrying out cluster analyses themselves, counterparts of many other statistical methods that exist for the discriminant analysis case are by and large unavailable for the cluster analysis situation.

The main dual purposes of this report are to take stock of the current state of the art in both discriminant and cluster analysis and to identify important problems that still need to be addressed in both domains. In Section 2, the focus is on methodology while in Section 3 theoretical aspects of the subject are reported. The fourth section provides a survey of available software and algorithms for both discriminant and cluster analysis. The final section contains a brief summary of the current state of the art and lists some problems that need more attention from researchers.

Although some editorial efforts were expended at putting in references to material across sections written by different people, inevitably, there remain some duplication of coverage and inconsistency of notation, which we hope are not too distracting.

2. METHODS

2.1 Introduction

In this section, brief descriptions of the methods of discriminant analysis and of cluster analysis are provided. The intention is not to provide details or derivations, because those are available in a number of books, but merely skeletal descriptions of the essential steps in the statistical procedures and algorithms.

Section 2.2 is concerned with methods of discriminant analysis and includes classical two-group linear discriminant analysis, classification into one of several populations, heterogeneous covariance matrices, classification by logistic regression, kernel and nearest neighbor methods and classification trees. Section 2.3 pertains to methods of clustering and includes a general discussion of algorithmic approaches.

2.2 Methods of Discriminant Analysis

2.2.1 General Remarks

The discriminant analysis situation is characterized by the following: one has two types of multivariate observations—the first, called *training samples*, are those whose group identity (i.e., membership in a specific one of say G given groups is known *a priori*), and the second type, referred to as *test samples*, consists of observations for which such *a priori* information is not available and which have to be assigned to one of the G groups.

The variables constituting the multivariate observations and the "groups" involved will depend on the particular application. For instance, in anthropometry, the variables might be different measurements on fossils and the groups might be a known taxonomy of the fossils (e.g., different races or different stages of evolution). In a medical application, the variables could be the results of various clinical tests and the groups could be collections of patients known to have different diseases. In an acoustical application, the variables might be the a set of acoustical parameters extracted from the utterance of a specific word by an individual whereas the groups are repeated utterances of the same word by different individuals. In each of these cases, there are observations whose group identity is known (the *training samples*) but there will also be some observations whose classification is unkown (e.g., a fossil whose race group is unknown, a patient whose disease category is unknown or an utterance whose source speaker is unknown).

Before discussion of the major numerically oriented methods of discriminant analysis, mention should be made of a number of developments in computer graphics for representing multivariate data that are useful informal aids for classification. Schematic graphical displays of multivariate observations proposed by Anderson (1957), Andrews (1972), Chernoff (1973b) and Kleiner and Hartigan (1981) can be and have been used for informal classification of objects. The essential idea is to represent either the individual training samples or some typical value (e.g., the mean of a group) via a schematic display, do the same for the test samples and then by inspection of these displays decide to assign a test sample to the group whose training sample displays (or typical value display) look "visually closest" to the display of the test case. In practice, large numbers of observations or variables, as well as poorly understood visual perception biases, can limit the usefulness of these graphical techniques.

In thinking of the more numerically oriented methods of discriminant analysis, it is useful to distinguish two stages of the analysis, although not all of the available statistical methods either make such a distinction or are equally useful for the two stages. The first stage, concerned solely with the training samples, is to find a representation of these observations so as to, in some sense, clearly separate the G groups. The resulting representation, usually a spatial one, is often called the *discriminant space*. Such a representation when presented graphically has major descriptive and diagnostic value in analyzing data.

The second stage of a discriminant analysis is concerned with assigning the test samples (i.e., those observations whose group identity is initially unknown) to one of the G specified groups. At this stage, the focus in on *correct classification*. Some measure of correct classification, using the training samples and not the test samples, is often used to evaluate the performance of discriminant analysis methods (see discussion below on evaluation). An important scientific consideration, that is sometimes not emphasized adequately in the statistics literature on discriminant analysis, is that in the real world it may turn out that an item whose classification is unknown may not belong to any of the prespecified groups but indeed be a member of an entirely different or hitherto unknown group (see Rao, 1960, 1962; Andrews, 1972).

Statistical considerations in discriminant analysis have to do with distributional assumptions concerning the observations, measures of separation among the groups, algorithms for carrying out both stages of the discriminant analysis and the study of the properties of proposed algorithms. Historically, Fisher (1936) was the first to propose a procedure for the two-group ($G = 2$) case based on maximizing the separation between the groups in the spirit of analysis of variance. This procedure is equivalent to the likelihood ratio procedure that arises if one assumes multivariate normality (with a common covariance matrix) for the observations from both groups. The initial extensions of this were concerned with multiple groups and with heterogeneous covariance matrices across groups, but still retained the multivariate normal assumption. These normality-based methods are the ones most widely used in practice. Provided the measured variables are not constrained to take on only a few distinct values, as in the case of binary variables, transformations of them might enhance their normality and enable the more sensible use of the normality-based procedures (see further discussion of transformations below).

There are real situations involving variables, such as binary or categorical ones, that are not sensibly transformed. Distribution-free and nonparametric methods, which move away from the normality assumption, have been developed relatively recently to handle such data. See, e.g., Hand (1981, Chapter 5) and Lachenbruch (1975, Chapter 4).

After developing a classification rule, the natural next step is to evaluate its performance. When information on the cost of misclassification is available, then one might look at the expected (average) cost of misclassification. However, such information is not usually available, and an oft used criterion is just the error rate itself (i.e., the proportion of items that are misclassified).

A number of possible rates may be considered:

1. The optimum error rate—the rate which would hold if all parameters are known.
2. The actual error rate—the rate which holds for a classification rule under consideration when it is used to classify all possible future samples.
3. The apparent error rate—the rate we obtain by resubstituting the training sample and determining the misclassifications.

It is possible to evaluate the overall error rate or the individual group rates. Both are of interest. The likelihood ratio procedure (e.g., see Sections 2.2.2 and 2.2.3) determines the rule so as to minimize the overall rate for specified parametric distributions. However, this may lead to a rule which has a high error rate in one of the groups, and this may be unacceptable to the user. In such cases, the "cutoff point" involved in the rule can be altered to give a more balanced set of error rates. This usually does not increase the overall error rate greatly.

Many procedures that depend heavily upon the assumption of normality have been proposed to estimate the error rates. Consideration is given here to estimators that may be used in any context. First, the apparent error rate (or *resubstitution estimator*) simply classifies the training sample using the rule calculated from it. The estimator is typically over optimistic and can badly mislead the user if the sample size is not much larger than the number of variables in the rule. It is also hazardous if there is initial misclassification in the training samples. However, for those cases in which the number of initially correctly classified observations is sufficiently large, the bias will be small. The second method of estimation is called *leave-one-out* and is similar in spirit to the jackknife. This procedure omits an observation, recalculates the classification rule from the remaining observations, classifies the deleted observation, and repeats these steps for each observation in turn. Counting the errors of misclassification yields an almost unbiased estimate of the error rate. Unfortunately, the variables indicating misclassification are correlated so that this estimate has a large variance. In many cases, the mean square error of the leave-one-out method is larger than that of the resubstitution estimator. The third procedure is the *bootstrap* method (Efron, 1982). This seems to combine the best features of the previous two estimators: it is almost unbiased and it has a small variance. The major drawbacks of the bootstrap are its expense and its inability, even asymptotically, to deal with sufficiently large biases. One must compute as many classification rules as there are replicates. If the classification rule is based on density estimation, this could become prohibitively expensive. A fourth possibility, closely analogous to the leave-one-out method, is *cross-validation*. One splits the training sample into k parts, uses all but one to develop the classification rule and classifies the left out part. This process is repeated k times, and error rates are averaged. Popular choices of k are the sample size (the jackknife case) and two. Provided enough data are available to carry it out, this has the advantages of being nearly unbiased. However, as for the jackknife, the mean square error may be large.

In summary, the apparent error rate is optimistically biased and should be used with caution when the sample sizes are small relative to the number of variables. The other methods mentioned can be useful alternatives in this case. Otherwise, the apparent error rate should be a satisfactory estimator. For a bibliography on error rates, see Toussaint (1974).

In the sections that follow, specific methods of discriminant analysis are outlined and for many of them some discussion is provided of their absolute/relative performances, including error rate behaviors.

2.2.2 Classical Two-group Linear Discriminant Analysis

The most widely used rule for classifying an observation $\mathbf{x}$ into one of two populations, Π_1 or Π_2, is that which classifies $\mathbf{x}$ into Π_1 if

$$(1)\quad v = (\bar{\mathbf{x}}^{(1)} - \bar{\mathbf{x}}^{(2)})'\mathbf{S}^{-1}(\mathbf{x} - (1/2)(\bar{\mathbf{x}}^{(1)} + \bar{\mathbf{x}}^{(2)})) \geq c$$

or into Π_2 otherwise. Here $\bar{\mathbf{x}}^{(1)}$ and $\bar{\mathbf{x}}^{(2)}$ denote the vector means of two independent samples (the training samples) of sizes n_1 and n_2, respectively, and $\mathbf{S}$ denotes the pooled sample covariance matrix, $((s_{ii'}))$, where

$$s_{ii'} = \sum_{g=1}^{2} \sum_{\alpha=1}^{n_g} \frac{(x_{i\alpha}^{(g)} - \bar{x}_i^{(g)})(x_{i'\alpha}^{(g)} - \bar{x}_{i'}^{(g)})}{n_1 + n_2 - 2};$$

$\mathbf{x}$ is a p component vector. The linear discriminant function (LDF) $(\bar{\mathbf{x}}^{(1)} - \bar{\mathbf{x}}^{(2)})'\mathbf{S}^{-1}\mathbf{x}$ was suggested by Fisher (1936) who introduced it as that linear combination of the p variables which separates the two (training) samples as much as possible. Specifically, for any linear combination, say $\mathbf{d}'\mathbf{x}$, the squared difference between the two sample means, divided by the pooled estimate of the variance of that difference is maximized by $\mathbf{d} = \mathbf{S}^{-1}(\bar{\mathbf{x}}^{(1)} - \bar{\mathbf{x}}^{(2)})$. This property of the LDF is a strong argument in favor of its use for classification purposes for populations with a common covariance matrix.

The cutoff point c in (1) can be chosen in various ways. Sometimes it is chosen so that the number misclassified from the two training samples is as small as possible. If the p variables used in the discrimination are normally distributed, and if their covariance matrices are the same in the two populations, then a frequently used cutoff point is

$$c = \ln(\hat{\pi}(2)/\hat{\pi}(1)). \tag{2}$$

Here $\hat{\pi}(g)$ is some estimate of $\pi(g)$, the *a priori* probability that an individual to be classified comes from Π_g. With this value of c, the classification rule is a sample estimate of the rule that classifies into Π_1 if

$$\begin{aligned} u &= (\boldsymbol{\mu}^{(1)} - \boldsymbol{\mu}^{(2)})'\boldsymbol{\Sigma}^{-1}(\mathbf{x} - (1/2)(\boldsymbol{\mu}^{(1)} + \boldsymbol{\mu}^{(2)})) \\ &\geq \ln \frac{\pi(2)}{\pi(1)} \end{aligned} \tag{3}$$

and into Π_2 otherwise; here $\boldsymbol{\mu}^{(1)}$, $\boldsymbol{\mu}^{(2)}$, $\boldsymbol{\Sigma}$ are the population counterparts of $\bar{\mathbf{x}}^{(1)}$, $\bar{\mathbf{x}}^{(2)}$, $\mathbf{S}$.

If the two populations have normal distributions with equal covariance matrices, then (3) is the best possible classification rule in the sense that the expected probability of misclassification is as small as possible. That is, $P = \pi(1)P(2 \mid 1) + \pi(2)P(1 \mid 2)$ is minimized, where $P(2 \mid 1)$ is the probability of misclassifying an individual from Π_1 and $P(1 \mid 2)$ is the probability of misclassifying an individual from Π_2.

Occasionally the $\pi(g)$ are known; e.g., in developing a function to discriminate between carriers and noncarriers of a genetically based disease, the prior probability that an individual is a carrier might be known. Sometimes $\pi(g)$ might be approximated well from knowledge of the relative sizes of the two populations. When little is known about the relative population sizes, it is usual to set $\hat{\pi}(1) = \hat{\pi}(2) = 1/2$ so that $\ln(\hat{\pi}(2) \mid \hat{\pi}(1)) = c = 0$.

Another method determines the cutoff c so that $P(2 \mid 1) = P(1 \mid 2)$; then an observation from Π_1 is just as likely to be misclassified as an observation from Π_2. This method has the advantage that no knowledge of the *a priori* probabilities is necessary. To accomplish this the cutoff c is determined so that

$$\Phi\left(\frac{c - \Delta^2/2}{\Delta}\right) = \Phi\left(-\frac{c + \Delta^2/2}{\Delta}\right), \tag{4}$$

whose solution is $c = 0$. Here $\Phi(\cdot)$ is the distribution function of the univariate standard normal distribution and

$$\Delta^2 = (\boldsymbol{\mu}^{(1)} - \boldsymbol{\mu}^{(2)})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}^{(1)} - \boldsymbol{\mu}^{(2)})$$

is the *Mahalanobis squared distance* between the two population means. This approach again suggests use of (1) with $c = 0$ in practice.

It was suggested by Wald (1944) that misclassification cost rather than misclassification probability should be used as a criterion in discrimination. If

$$C = \pi(1)P(2 \mid 1)C(2 \mid 1) + \pi(2)P(1 \mid 2)C(1 \mid 2)$$

is the expected cost, with $C(2 \mid 1)$ the cost of misclassifying an individual from Π_1 into Π_2 and similarly for $C(1 \mid 2)$, the rule that minimizes expected cost is to assign $\mathbf{x}$ to Π_1 if

$$\begin{aligned} u &= (\boldsymbol{\mu}^{(1)} - \boldsymbol{\mu}^{(2)})'\boldsymbol{\Sigma}^{-1}(\mathbf{x} - (1/2)(\boldsymbol{\mu}^{(1)} + \boldsymbol{\mu}^{(2)})) \\ &\geq \ln \frac{C(1 \mid 2)\pi(2)}{C(2 \mid 1)\pi(1)}. \end{aligned} \tag{5}$$

In practice it is usually difficult to estimate the relative costs. In some situations, however, when the ratio $\pi(2)/\pi(1)$ is known to be small, the cost ratio $C(1 \mid 2)/C(2 \mid 1)$ is clearly large, so that setting $c = 0$ is not unreasonable.

If one wishes to select a cutoff point so that the expected costs of misclassifying observations from each of the two populations are approximately equal, then in (1) c is chosen so that

$$C(2 \mid 1)\Phi\left(\frac{c - D^2/2}{D}\right) = C(1 \mid 2)\Phi\left(-\frac{c + D^2/2}{D}\right), \tag{6}$$

where $D^2 = (\bar{\mathbf{x}}^{(1)} - \bar{\mathbf{x}}^{(2)})'\mathbf{S}^{-1}(\bar{\mathbf{x}}^{(1)} - \bar{\mathbf{x}}^{(2)})$, the Mahalanobis squared distance between the two sample means.

Relation to Regression Analysis. It is possible to obtain the coefficients of the LDF by using a regression program. A dummy variable y is introduced that takes on the value $n_2/(n_1 + n_2)$ for observations from Π_1 and $-n_1/(n_1 + n_2)$ for observations from Π_2. If the two data sets are then treated as a single sample of size $n_1 + n_2$, the coefficients in the regression of y on $\mathbf{x}$ are proportional to $\mathbf{S}^{-1}(\bar{\mathbf{x}}^{(1)} - \bar{\mathbf{x}}^{(2)})$. (See Anderson, 1958.)

Tests of Hypotheses. Under the normality assumption and with equal covariance matrices, the hypothesis that the p variates have no discriminatory power can be stated as either H_0: $\Delta^2 = 0$ or H_0: $\boldsymbol{\mu}^{(1)} = \boldsymbol{\mu}^{(2)}$. This can be tested (Rao, 1965) with the statistic $(n_1n_2/(n_1 + n_2))D^2$; this statistic is known as Hotelling's T^2. When multiplied by $(n_1 + n_2 - p - 1)/p(n_1 + n_2 - 2)$, it has an F distribution with degrees of freedom p, $n_1 + n_2 - p - 1$, and with noncentrality parameter $(n_1n_2/(n_1 + n_2))\Delta^2$. This F test is the same as the F test that would be made in the regression analysis with the dummy variable y to test whether all the regression coefficients are zero. The F test of whether a subset of the regression coefficients are zero may be used to test whether that subset of variables adds anything to the discrimination. See Rao (1952) for details.

For deciding between the hypotheses that $\mathbf{x}$ is from Π_1 or from Π_2, a Bayesian approach can be taken. This approach compares *a posteriori* probabilities. In particular, under the normality and equal covariance matrix assumptions, one can easily estimate the needed *a posteriori* probabilities $P(\Pi_g \mid \mathbf{x})$. These probabilities are

$$P(\Pi_1 \mid \mathbf{x}) = \frac{\pi(1)\exp(u)}{\pi(2) + \pi(1)\exp(u)} \tag{7}$$

and

$$P(\Pi_2 \mid \mathbf{x}) = \frac{\pi(2)}{\pi(2) + \pi(1)\exp(u)},$$

and they can be estimated by

$$\hat{P}(\Pi_1 \mid \mathbf{x}) = \frac{\pi(1)\exp(v)}{\pi(2) + \pi(1)\exp(v)} \tag{8}$$

and

$$\hat{P}(\Pi_2 \mid \mathbf{x}) = \frac{\pi(2)}{\pi(2) + \pi(1)\exp(v)}.$$

Advantages of the LDF. Clear advantages of this discrimination method are simplicity and the availability of package programs. Further, the idea of replacing p variates—if they are in the same units—by a linear index is sometimes easily accepted by the statistical layman. Also, if the researcher's aim is to estimate *a posteriori* probabilities rather than to classify, these are particularly simple to obtain. If, as is frequently the case, his purpose is to understand the difference between Π_1 and Π_2 rather than to classify, the sizes of the standardized coefficients in the LDF may give him some clue. Also, projections of the training samples onto the LDF can be studied graphically. Indeed, Fisher's (1936) original paper shows histograms of such projections. The histograms are not only visually useful for looking at the separation between the two groups but also have diagnostic value in checking the reasonableness of the assumptions of normality and homoscedasticity.

These advantages have led to the widespread use of the LDF. Without the assumptions of normality and equal covariance matrices, the main justification for its use is that it spreads the two sample means apart as far as possible, scaled in a particular way, using a linear combination of variables. Because the LDF is often used with all types of nonnormality and with unequal covariance matrices, its performance under these departures becomes important.

To evaluate LDF performance, one might use as a criterion the expected value of any of the following:

$$\begin{aligned} &P = \pi(1)P(2 \mid 1) + \pi(2)P(1 \mid 2),\\ &\max[P(2 \mid 1), P(1 \mid 2)],\\ &C = \pi(1)P(2 \mid 1)C(2 \mid 1) + \pi(2)P(1 \mid 2)C(1 \mid 2),\\ &\max[P(2 \mid 1)C(2 \mid 1), P(1 \mid 2)C(1 \mid 2)]. \end{aligned} \tag{9}$$

These are, respectively, the total error rate, the maximum group-specific error rate, the total cost and the maximum of group-specific costs.

An estimate of any one of the four quantities in (9) for any particular discriminant function might be selected by a researcher for evaluating that particular function. To evaluate the LDF method, however, one must estimate the expected value over all possible LDF's (that is, over the distribution of $\bar{\mathbf{x}}^{(1)}$, $\bar{\mathbf{x}}^{(2)}$ and $\mathbf{S}$).

Considerable work has been done on the robustness of the LDF, much of it being in comparison with other specific alternative methods. In general, the LDF is thought to perform relatively well for moderate sample sizes in comparison with other more complicated methods. Its performance is often improved by the use of transformations of the variables.

Variable Selection. Some of the strengths of the LDF can also be a source of weakness. It has become dangerously easy for the researcher to toss a large number of variables into the computer and then on the basis of coefficient size to make extremely doubtful statements concerning the relative importance of the different variables in discrimination.

It has been shown that when many variables are included, the LDF may do extremely well in classifying the observations in the two training samples, but perform worse in classifying new observations than an LDF based on fewer variables. Good practice therefore necessitates selecting a small number of variables relative to the sizes of the two training samples. There are as many possible ways of doing this as there are in the corresponding regression problem. (In discrimination there may be a greater tendency to have large numbers of variables than in regression.)

Often several variables known by the researcher to be highly correlated can be replaced by just one variable. Sometimes the variables included are simply those whose scaled between group squared distances, $D_i^2 = (\bar{x}_i^{(1)} - \bar{x}_i^{(2)})^2/s_{ii}$, are the largest. This often works quite well, but it is not a foolproof method. Indeed, even if the Mahalanobis squared distance between the two populations based on only the ith variable, $(\mu_i^{(1)} - \mu_i^{(2)})^2/\sigma_{ii}$, equals zero, it is possible that the ith variable may increase the Mahalanobis squared distance considerably when it is used with some other variables.

Often a stepwise discriminant program is used for variable selection. In a forward selection program, variables are included one at a time; at each step the next variable included is the one that increases the sample Mahalanobis squared distance the most. In a backward stepwise procedure one begins with the entire set and then at each step drops the variable that decreases the Mahalanobis squared distance the least.

With a large number of variables, the stepwise procedures seem a sensible way to select, say 3 to 5 variables. One must always bear in mind, however, that the set of variables selected may not be the best possible set, even for the purpose of classifying the original observations. For the purpose of identifying which variables are important in discriminating between Π_1 and Π_2, a single run of a stepwise program is particularly inadequate. If one believes that several important variables have been found, they should be dropped, and the stepwise procedure done without them in order to see how well the other variables discriminate.

Some simulation studies have indicated that with many variables a combination of first using the Mahalanobis squared distance based on each single variable to reduce the number of variables and then a stepwise program is a reasonable plan (Farver and Dunn, 1979).

As mentioned earlier, an F test can be used to test whether a subset of the variables adds significantly to the separation of the two groups (Rao, 1952). It is a routine matter, in many problems, to compute significance levels for all possible subsets. These can be plotted and studied informally as a guide to selecting subsets for the discriminant analysis (McKay, 1978).

For further discussion of variable selection in discriminant analysis see, e.g., Hand (1981, Chapter 6), McKay and Campbell (1982a, b) and Seber (1984, Section 6.10).

2.2.3 Classification into One of Several Populations

When an observation $\mathbf{x}$ is to be classified into one of G populations where $G > 2$, then the procedure that minimizes the expected value of the probability of misclassification is to classify $\mathbf{x}$ into Π_k if

$$\pi(g)p_g(\mathbf{x}) \le \pi(k)p_k(\mathbf{x}) \qquad \text{for } g = 1, \cdots, G, \quad g \neq k, \tag{10}$$

where $\pi(g)$ is the *a priori* probability that an observation belongs to the gth group and $p_g(\mathbf{x})$ is the probability density function for the gth group. This is the Bayes procedure.

For multivariate normal populations, this Bayes rule becomes

for each $g = 1, \cdots, G$, classify into Π_k if

$$\begin{aligned} u_{gk}(\mathbf{x}) &= (\boldsymbol{\mu}^{(k)} - \boldsymbol{\mu}^{(g)})'\boldsymbol{\Sigma}^{-1} \\ &\quad \cdot (\mathbf{x} - (1/2)(\boldsymbol{\mu}^{(k)} + \boldsymbol{\mu}^{(g)})) \\ &\ge \ln\frac{\pi(g)}{\pi(k)}, \quad g = 1, \cdots, G, \quad g \neq k. \end{aligned} \tag{11}$$

The sample-based estimate of the Bayes rule is to classify $\mathbf{x}$ into Π_k if

$$\begin{aligned} v_{gk}(\mathbf{x}) &= (\bar{\mathbf{x}}^{(k)} - \bar{\mathbf{x}}^{(g)})'\mathbf{S}^{-1}(\mathbf{x} - (1/2)(\bar{\mathbf{x}}^{(g)} + \bar{\mathbf{x}}^{(k)})) \\ &\ge \ln\frac{\pi(g)}{\pi(k)}, \quad g = 1, \cdots, G, \quad g \neq k. \end{aligned} \tag{12}$$

Each of the v_{gk} provides the usual LDF for discriminating between two groups, and thus to classify $\mathbf{x}$ one may first decide between pairs of groups in the usual way and finally decide among all G groups. The situation is somewhat analogous to a baseball league in which each team plays every other team once to determine the winner; the analogy breaks down, however, for in classification, one population always emerges as winner.

In using (12) to approximate the Bayes solution, it is necessary to know or estimate the $\pi(g)$'s, the *a priori* probabilities. If one does not know the $\pi(g)$, one may seek the minimax solution and choose the cutoff points so that the expected probabilities (or costs) of misclassification are all equal, no matter from which population an observation is drawn. It has been shown that the minimax solution is the same as the Bayes solution for some set of $\pi(1), \cdots, \pi(G)$. Therefore, one may use (12) and find $c_1, \cdots, c_G$ to replace $\ln \pi(1), \cdots, \ln \pi(G)$ such that the estimated expected costs are approximately equal based on the rule: classify into Π_k if

$$v_{gk}(\mathbf{x}) \ge c_g - c_k, \quad g = 1, \cdots, G, \quad g \neq k.$$

These constants can be determined by trial and error.

An alternate approach (Rao, 1948, 1952) to classifying into one of several populations is a generalization of Fisher's original idea of choosing a linear function with maximum squared distance between means as compared with the variances.

The within-population covariance matrix is $\boldsymbol{\Sigma}$; the covariance matrix of the G population means is the "between population" covariance matrix $\mathbf{B} = \sum_{g=1}^{G} (\boldsymbol{\mu}^{(g)} - \bar{\boldsymbol{\mu}})(\boldsymbol{\mu}^{(g)} - \bar{\boldsymbol{\mu}})'/(G - 1)$, where $\bar{\boldsymbol{\mu}} = \sum_{g=1}^{G} \boldsymbol{\mu}^{(g)}/G$. One seeks $\boldsymbol{\alpha}$ such that $\gamma = \boldsymbol{\alpha}'\mathbf{B}\boldsymbol{\alpha}/\boldsymbol{\alpha}'\boldsymbol{\Sigma}\boldsymbol{\alpha}$

is a maximum. When $G = 2$, the solution is $\boldsymbol{\alpha} = \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}^{(1)} - \boldsymbol{\mu}^{(2)})$. For general G, the extreme values of γ are obtained by using the eigenvectors, $\boldsymbol{\alpha}_1, \cdots, \boldsymbol{\alpha}_s$ of the matrix $\Sigma^{-1}\mathbf{B}$.

There are no more than $s = \min(G - 1, p)$ nonzero eigenvalues, $\gamma_1 \geq \gamma_2 \geq \cdots \geq \gamma_s$, of $\Sigma^{-1}\mathbf{B}$. These have eigenvectors $\boldsymbol{\alpha}_1, \cdots, \boldsymbol{\alpha}_s$ which are linearly independent. If one uses $u_1 = \boldsymbol{\alpha}_1'\mathbf{x}, \cdots, u_s = \boldsymbol{\alpha}_s'\mathbf{x}$ as the discriminant functions, then the rule is to classify into Π_i if

$$(13) \quad \sum_{j=1}^{s} [\boldsymbol{\alpha}_j'(\mathbf{x} - \boldsymbol{\mu}^{(i)})]^2 = \min_{g} \sum_{j=1}^{s} [\boldsymbol{\alpha}_j'(\mathbf{x} - \boldsymbol{\mu}^{(g)})]^2.$$

Rule (13) involves knowledge of the population parameters. This rule is equivalent to (11) if the prior probabilities are equal. The corresponding sample-based rule substitutes the pooled within-groups estimate of the covariance matrix, $\mathbf{W}$, for $\boldsymbol{\Sigma}$, and

$$\hat{\mathbf{B}} = \sum_{g=1}^{G} n_g(\bar{\mathbf{x}}^{(g)} - \bar{\mathbf{x}})(\bar{\mathbf{x}}^{(g)} - \bar{\mathbf{x}})'/(G - 1)$$

for $\mathbf{B}$, where $\bar{\mathbf{x}} = \sum_{g=1}^{G} n_g\bar{\mathbf{x}}^{(g)}/n$ and $n = \sum_{g=1}^{G} n_g$. Then the sample discriminant functions $v_j = \hat{\boldsymbol{\alpha}}_j'\mathbf{x}$ are used in place of the u_j where $\hat{\boldsymbol{\alpha}}_j$ is an eigenvector of $\mathbf{W}^{-1}\hat{\mathbf{B}}$, $j = 1, \cdots, s$.

This method has certain advantages in reducing the use of a large number of variates to a small number of canonical variables. Although the sample v_j are not uncorrelated, as are the u_j's, the sample estimates of their covariances are zero, so that calculations are greatly simplified. When not all the canonical variables are used in classification, the procedure using (13) cannot be expected to be optimal, but its sample-based counterpart may be better because the additional canonical variables may be mostly reflecting noise. Indeed, in practice, only the first few canonical variables are often used. When all the canonical variables are used, the procedure gives the same results as the one using all the original variables.

Projections of the training samples onto the canonical variables, especially the first few of them, can be useful in much the same way as projections onto the LDF in the two-groups case. Scatter plots of such projections (see, e.g., Rao, 1952; Gnanadesikan, 1977) can be studied for separations among the groups and for evaluating the reasonableness of assumptions such as the homogeneity of the group covariance matrices.

2.2.4 Heterogeneous Covariance Matrices Case

The Quadratic Discriminant Function. Given two populations with mean vectors and covariances matrices, $\boldsymbol{\mu}_g$, $\boldsymbol{\Sigma}_g$, $g = 1, 2$, and *a priori* probabilities $\pi(1)$ and $\pi(2)$ that an observation belongs to each of them, the quadratic discriminant rule is to assign $\mathbf{x}$ to Π_1 if

$$(14) \quad \begin{aligned} &\mathbf{x}'(\boldsymbol{\Sigma}_2^{-1} - \boldsymbol{\Sigma}_1^{-1})\mathbf{x} - 2\mathbf{x}'(\boldsymbol{\Sigma}_2^{-1}\boldsymbol{\mu}_2 - \boldsymbol{\Sigma}_1^{-1}\boldsymbol{\mu}_1) \\ &\qquad + (\boldsymbol{\mu}_2'\boldsymbol{\Sigma}_2^{-1}\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1'\boldsymbol{\Sigma}_1^{-1}\boldsymbol{\mu}_1) \\ &\geq \ln(|\boldsymbol{\Sigma}_2|/|\boldsymbol{\Sigma}_1|) + 2\ln(\pi(2)/\pi(1)), \end{aligned}$$

and to Π_2 otherwise. If the two populations are normally distributed, the quadratic discriminant rule is the best discriminant rule, in the sense of minimizing the expected probabilities of misclassification. It reduces to the LDF if $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2$.

The sample-based rule corresponding to (14) is that $\mathbf{x}$ is assinged to Π_1 if

$$(15) \quad \begin{aligned} &\mathbf{x}'(\mathbf{S}_2^{-1} - \mathbf{S}_1^{-1})\mathbf{x} - 2\mathbf{x}'(\mathbf{S}_2^{-1}\bar{\mathbf{x}}_2 - \mathbf{S}_1^{-1}\bar{\mathbf{x}}_1) \\ &\qquad + (\bar{\mathbf{x}}_2'\mathbf{S}_2^{-1}\bar{\mathbf{x}}_2 - \bar{\mathbf{x}}_1'\mathbf{S}_1^{-1}\bar{\mathbf{x}}_1) \\ &\geq \ln(|\mathbf{S}_2|/|\mathbf{S}_1|) + 2\ln(\pi(2)/\pi(1)). \end{aligned}$$

Best Linear Discriminant Function. An attractive simplicity of Fisher's LDF is that it is a linear function in the original variables. The preceding discussion, however, established that when the covariance matrices across groups are not the same, even under normality assumptions, the optimal discriminant function is no longer linear in the variables. Nevertheless, as an approximation, one may limit consideration to linear functions and seek a "best LDF" for two normally distributed populations whose covariance matrices are unequal. The "best LDF" procedure was developed independently by Riffenburgh and Clunies-Ross (1960), Clunies-Ross and Riffenburgh (1960), Anderson and Bahadur (1962) and Jennrich (1962). It is the linear combination of measurements that discriminates best between the two populations.

The sample-based best linear rule is that an observation $\mathbf{x}$ is classified into Π_1 if

$$\mathbf{x}'\mathbf{b} \geq \bar{\mathbf{x}}_1\mathbf{b} - t_1\mathbf{b}'\mathbf{S}_1\mathbf{b} = \bar{\mathbf{x}}_2\mathbf{b} + t_2\mathbf{b}'\mathbf{S}_2\mathbf{b},$$

or otherwise into Π_2. Here $\bar{\mathbf{x}}_g$, $\mathbf{S}_g$, $g = 1, 2$ are the sample mean vector and covariance matrix in group g and

$$\mathbf{b} = (t_1\mathbf{S}_1 + t_2\mathbf{S}_2)^{-1}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2),$$

and t_1 and t_2 are chosen to minimize the estimated expected probabilities (or costs) of misclassification. The quantities, t_1 and t_2, can be normalized so that the minimization can be carried out with respect to a single variable; see Anderson and Bahadur (1962).

Asymptotically, the best LDF must perform better than Fisher's LDF if $\boldsymbol{\Sigma}_1 \neq \boldsymbol{\Sigma}_2$ and reduce to Fisher's if $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2$. However, under normality the quadratic discriminant rule performs better asymptotically than either the best LDF or the Fisher LDF rule when $\boldsymbol{\Sigma}_1 \neq \boldsymbol{\Sigma}_2$; under equality all three methods are asymptotically the same. For small samples, however, the quadratic discriminant function can behave appreciably worse than the linear functions as has been shown in simulation studies, see Marks and Dunn, 1974. This tendency increases as the two populations are moved farther apart and is more pronounced when more variables are used; as expected, it decreases as the sample sizes increase and as departures from equality increase.

When gross inequality is present, the best LDF has a certain advantage over the quadratic discriminant function from the standpoint of ease in interpretation. However, compared with either Fisher's LDF or the quadratic, it takes more computation time.

In Fisher's LDF the coefficients remain the same if one changes the *a priori* probabilities $\pi(g)$; in the best LDF these coefficients vary as one varies the $\pi(g)$. Thus the coefficients seem even less meaningful for the best LDF.

Marks and Dunn (1974) find that with small departures from equality of covariance matrices the best LDF performs quite well. For extremely large departures, it performs appreciably better than the usual LDF, but usually in such cases the quadratic discriminant function performs still better.

The quadratic discriminant function appears to perform poorly under non-normality. This is not surprising because the difference between the linear and quadratic discriminant functions is most marked in the tails of the distributions. When population means coincide, the quadratic discriminant function comes into its own; in this situation, the LDF becomes useless.

2.2.5 Two-group Classification by Logistic Regression

The logistic regression discriminant procedure involves a (linear) discriminant function for use with certain non-normal populations. Suggested by Cornfield (1962), it was used by Truett, Cornfield and Kannel (1967) in the Framingham study. Hand (1981, Section 5.3.1) and Lachenbruch (1975, Chapter 6) provide more detail and references than are provided here.

In the logistic regression procedure, the data set is considered to consist of a single sample of size $n = n_1 + n_2$ from the combined population. For each observation $\mathbf{x}_j$, $j = 1, \cdots, n = n_1 + n_2$, a y variable is introduced. For the n_1 observations that are from Π_1, $y = 1$; for the n_2 observations from Π_2, $y = 0$.

The variable y is a binary variable and the *a priori* probability that y equals 1 is $\pi(1)$. If the $\mathbf{x}$ variables are normally distributed with equal covariance matrix $\mathbf{\Sigma}$ in both groups, the *a posteriori* probability that y equals 1 is of the form (see (7))

$$P(y = 1 \mid \mathbf{x}) = \frac{\pi(1)\exp u}{\pi(2) + \pi(1)\exp u} \tag{16}$$
$$= \frac{\exp(\alpha + \mathbf{x}'\boldsymbol{\beta})}{1 + \exp(\alpha + \mathbf{x}'\boldsymbol{\beta})}.$$

Equation (16) holds for a wider class of distributions than the normal, and for any such distribution, one may obtain a (linear) discriminant function, $v = \hat{\alpha} + \mathbf{x}'\hat{\boldsymbol{\beta}}$ by estimating the parameters α and $\boldsymbol{\beta}$. One method of estimating α and $\boldsymbol{\beta}$ is the maximum likelihood method. With a sample of size $n = n_1 + n_2$ and binomial parameter $\exp(\alpha + \mathbf{x}'\boldsymbol{\beta})/(1 + \exp(\alpha + \mathbf{x}'\boldsymbol{\beta}))$ the likelihood function is

$$\prod_{i=1}^{n_1} \left(\frac{\exp(\alpha + \mathbf{x}_i'\boldsymbol{\beta})}{1 + \exp(\alpha + x_i'\boldsymbol{\beta})} \right) \tag{17}$$
$$\cdot \prod_{j=n_1+1}^{n} \left(\frac{1}{1 + \exp(\alpha + \mathbf{x}_j'\boldsymbol{\beta})} \right),$$

assuming that the first n_1 observations are from Π_1 and the remaining are from Π_2. The estimates $\hat{\alpha}$ and $\hat{\boldsymbol{\beta}}$ are chosen to maximize (17).

In practice, it is usual to have two samples rather than a single sample from the combined population. Then n_1 and n_2 are chosen by the researcher rather than being random variables. In this case, the same procedure is nevertheless used. After obtaining $\hat{\alpha}$ and $\hat{\boldsymbol{\beta}}$, the coefficients $\hat{\boldsymbol{\beta}}$ are retained, but a constant to replace $\hat{\alpha}$ is chosen by considering the number of the training samples misclassified by the discrimination rule for various possible choices of the value of the constant. The final choice of constant may be the one that yields the lowest total number misclassified or the one that minimizes the maximum proportion of misclassified observations.

Selection of variates is a problem in this method as in others. Package programs for stepwise logistic regression are available. See Section 4.2.3 for details.

Strengths and Weaknesses. A disadvantage of the logistic regression approach is that it involves more extensive computations, a factor that becomes important when using stepwise procedures.

It is clear that under normality the logistic procedure cannot be expected to classify as well as does the LDF. Efron (1975), in comparing the asymptotic relative efficiency of the two procedures under normality, found that when $\pi(1) = \pi(2) = 1/2$ (the case most favorable to logistic regression) the asymptotic relative efficiency decreased from one at $\Delta = 0$ to about .3 at $\Delta = 3.5$, where Δ is the Mahalanobis distance between the two populations.

2.2.6 Kernel and Nearest Neighbor Methods

These methods are based on nonparametric density estimation algorithms that have been developed since the early 1950s.

The kernel methods estimate densities based on the sum of a set of functions

$$f(\mathbf{x}) = \sum_{j=1}^{n} k(\mathbf{x}, \mathbf{x}_j)/n,$$

where k must satisfy certain regularity conditions (see, e.g., Cacoullos, 1966).

The direct use of these functions is not needed (see, e.g., Silverman, 1986, Section 3.5). The Fast Fourier Transform speeds the calculations greatly. An additional advantage is the apparent resistance to the effect of outliers. Because a kernel must become small far from a point, a single outlying point will not contribute greatly to the estimate of the density of points in the middle of the distribution. Thus, even if the training sample is contaminated with outliers, the resulting allocation rule should perform well. Various types of kernels have been proposed, e.g., a multivariate normal density with a diagonal covariance matrix. Some of the most important kernels can be negative for some values of their arguments. If the kernel function is zero for distant points, outliers generally have no influence on estimates of the density at the majority of points. The nearest neighbor rule allocates points on the basis of a "majority vote." For equal prior probabilities, the k closest points to the point to be allocated are found and the unknown point is classified in the group to which the majority of these neighbors belong. This rule may be modified easily to account for unequal prior probabilities. It has been applied to continuous and discrete distributions. The density estimates are consistent and the error rates tend to the optimal ones. Another class of rules is based on Fourier series estimates of densities (see, e.g., Tarter and Kronmal, 1970).

The behavior of kernel estimates does not depend on the form of the kernel as much as it does on the smoothing parameter. This parameter determines the weight given to a point and is related to the smoothness of the estimated density function. Several solutions have been proposed but none seems to be generally accepted. Breiman, Meisel and Purcell (1977) have considered variable kernel estimates. The problems they address include selection of the smoothing parameter and methods of smoothing the estimates in regions of low density. They proposed a smoothing parameter of the form

$$\Sigma_{jk} = (\alpha_k d_{j,k}),$$

where $d_{j,k}$ is the distance from x_j to its kth nearest neighbor and α_k is a constant specific to the value of k.

An alternative approach to discrimination might be to estimate the ratio of the densities nonparametrically rather than the densities themselves. This is a problem because with the kernels involved, one would have the ratio of two sums of functions that may not be smooth. Simple approximations to these sums may merely lead us back to parametric densities.

2.2.7 Classification Trees

A rather different method of discriminant analysis is to portray the problem in terms of a binary tree. The tree provides a hierarchical-type of representation of the data space that can be readily used as a basis for classification by tracing down the appropriate branches of the tree.

This line of development was started by Morgan and Sonquist (1963) and Morgan and Messenger (1973). It has been vigorously pursued and refined by several people. Recent work is described in depth in the book by Breiman, Friedman, Olshen and Stone (1984). The earlier work is often referred to as AID (for automatic interaction dectection) whereas the contributions by Breiman, Friedman, Olshen and Stone are known by the acronym, CART (for classification and regression trees). The primary differences between AID and CART are in details of how the binary trees are formed.

In its simplest form, the CART method produces a tree that is based on individual variables. For example, the split at the top of the tree might be determined by the question, "Is $x_5 \leq 6.2$?". This will determine a left and right branch. The left branch corresponding to $x_5 \leq 6.2$ might then be divided according to the question, "Is $x_3 \geq 1.4$?" and the right branch, for which $x_5 > 6.2$, might be split according to the question, "Is $x_1 \geq 0$?". The methodology has three components to it: the set of questions, rules for selecting the best splits and a criterion for choosing the extent of the tree. With the tree in place, each terminal node of the tree can then be associated with one of the G groups.

More sophisticated questions can also be handled by this approach, such as, "Is $\sum a_j x_j \leq c$?" or "Is $x \in A$?". The variables themselves can be categorical, continuous or a mixture of both.

Many of the issues that arise in classical discriminant analysis show up in this procedure as well. These include selection of variables, use of misclassification costs and prior distributions, construction of classification rules using training samples, estimation of error rates, etc.

Generally speaking, CART is a flexible procedure that can result in very intuitive and easy-to-use classification rules. At the same time, there has not been enough widespread use of these methods to know how generally effective they are. Arriving at the best tree structure is a nontrivial matter and the tree itself may not be reliably determined. The descriptive value of the LDF is lost in the sense that the tree is a higher-level summary that is further removed from the raw data. Moreover, the discriminant function approach focuses more directly on spatial separations among groups, as revealed in scatter plots of the discriminant variables.

The current state of CART is perhaps best summarized by its developers (Breiman, Friedman, Olshen and Stone, 1984, page viii):

> Binary trees give an interesting and often illuminating way of looking at data in classification or regression problems. They should not be used to the exclusion of other methods. We do not claim they are always better. They do add a flexible nonparametric tool to the data analyst's arsenal.

2.3 Methods of Cluster Analysis

2.3.1 General Remarks

Cluster analysis involves the search through data for observations that are similar enough to each other to be usefully identified as part of a common cluster. This is a very intuitive and natural objective and one that is easy to think about. For example, the galaxies of stars in the universe can be described as clusters in a three-dimensional setting.

However, to be even a bit more precise about what is meant by a cluster can quickly get one bogged down in controversy and details. In fact, there is no generally accepted precise definition. Some would claim that clusters correspond to *real* underlying groups or populations and the challenge is to discover them. Others tend to think of clusters in a much weaker structural sense but still find the data-determined groups to be useful. For now, it will suffice to take the rather ambiguous attitude that clusters consist of observations that are close together and that the clusters themselves are clearly separated. If each observation is associated with one and only one cluster, then the clusters constitute a partition of the data that can be very useful for statistical purposes.

For instance, it is often possible to summarize a large multivariate data set in terms of a "typical" member of each cluster. This would be more meaningful than only looking at a single "typical" member of the entire data and much more concise than individual descriptions of each observation.

Another use occurs when one is attempting to model data in the presence of cluster structure. Better results may be achieved by taking this structure into account before attempting to estimate any of the relationships that may be present.

Finding the partition into clusters is not as easy as it may sound. Except in small problems, to "do it right," i.e., to consider all possible partitions of the data into clusters, is computationally out of the question. Consequently, numerous different algorithms have evolved as compromise procedures for finding clusters in a reasonably efficient way. Some authors prefer to start with a model, e.g., a mixture model (see Section 3.3.5), of clusters and then to find a practical algorithm for extracting the clusters in the context of that model. In the following discussion, such models are not discussed at all.

The development of algorithms has, for the most part, come out of applications-oriented disciplines such as biology and psychology rather than statistics. The explanation would appear to be that experts in these fields have developed tailored methods to solve their own problems because a general body of adequate clustering methodology was lacking.

Mentioning a few examples of applications of clustering methods may help to convey the types of problems they can contribute to:

> *Taxonomy*. Clustering species of bees into higher-level taxonomic groups (Michener and Sokal, 1957).
>
> *Genetics*. Studying genetic diversity within and between populations of an endangered fish species (Vrijenhoek, Douglas and Meffe, 1985).
>
> *Medicine*. Developing clusters of patients based on physiological variables (Siegel, Goldwyn and Friedman, 1971).
>
> *Speech processing*. Constructing a speaker-independent word recognition system (Rabiner, Levinson, Rosenberg and Wilpon, 1979).
>
> *Glaciology*. Mapping the Antarctic and Arctic regions in terms of clusters of types of sea ice and fern (Rotman, Fisher and Staelin, 1981).
>
> *Archaeology*. Grouping broaches from an Iron Age site in Switzerland based on their attributes (Hodson, Sneath and Doran, 1966).
>
> *Education*. Dividing up a class of workers in the telephone industry based on their common training needs (Kettenring, Rogers, Smith and Warner, 1976).
>
> *Business*. Clustering corporations according to their financial characteristics (Chen, Gnanadesikan and Kettenring, 1974).

These examples are typical of many in the literature in that the clustering was done with the aid of familiar numerical algorithms. These algorithms will be discussed in more detail in Section 2.3.2, but it is worth pointing out here that they are the products of the type of research on clustering methodology that was going on in the late 1950s and 1960s. The algorithms are pretty straightforward and easy to describe. More recently there has been a very pronounced trend toward more complex algorithms that attempt to achieve better results through their sophistication and exploitation of currently available computing power. Another trend has been the development of dynamic graphic display devices that can be very effective at revealing characteristics of the data including clusters.

Somewhat ironically, given their early lack of involvement, statisticians have recently been using cluster analysis as a building block for other procedures, especially in the area of regression diagnostics (Landwehr, Pregibon and Shoemaker, 1984; Gray and Ling, 1984).

2.3.2 Algorithms

Because detailed discussions of specific clustering algorithms are readily available (see, e.g., Anderberg, 1973; Cormack, 1971; Everitt, 1980; Hartigan, 1975; Seber, 1984, Chapter 7; Sneath and Sokal, 1973), the focus here will be more on general approaches.

Clustering data is often convincingly useful even if an unambiguously "correct" solution is lacking. The same can be said about attempts to classify the existing clustering algorithms: they are not as clean-cut as one might like but they do help to summarize the types that are available.

Among the numerical algorithms whose primary function is to reveal clusters, three general types can be distinguished: hierarchical, partitioning and overlapping. Only the second of these is strictly compatible with the loose definition of clustering used in the previous section.

The hierarchical algorithms result in a tree-like representation of the data, often called a dendrogram. At the top of the tree each observation is represented as a separate "cluster." At intermediate levels observations are grouped into fewer "clusters" than at the higher levels. At the bottom, all of the observations are merged into one "cluster." In some problems, the entire tree structure may be of interest. In others, the tree is just a convenient tool for obtaining a partition. This is usually done by cutting the tree at a suitable level which forces a particular partition.

Some hierarchical algorithms form the tree from the bottom up in a divisive fashion, but most work agglomeratively from the top down. Hartigan (1975, page 12) attributes this to the difficulty in finding effective splitting rules as well as the possible expense involved in executing them. Nevertheless, aside from their pragmatic advantage, the current emphasis on the agglomerative approach may be overdone because it may be possible to build more sophisticated algorithms that are less sensitive to local idiosyncracies in the data by working in the other direction.

A further distinction among the hierarchical algorithms is in the type of data they require. Some operate directly on pairwise measures of similarity or dissimilarity between every pair of observations. This is appealing from at least two points of view: first, the initial data, which commonly take the form of n observations on p variables, are not used by the algorithm once the interpoint distances have been determined; and, second, sometimes the raw form of the data is a set of pairwise dissimilarities or "distances" between points and it is convenient to be able to cluster points directly with these as input.

Perhaps the best known and most widely used of the hierarchical algorithms are the single linkage (nearest neighbor), complete linkage (farthest neighbor) and average linkage methods. In the single linkage approach, successive mergings are made according to the rule that the two clusters to be joined are the ones with the smallest interpoint distance between them. The complete linkage procedure focuses on the largest pairwise distances and joins those clusters that have the smallest of these values. The average linkage method operates similarly but on the average distances between members of pairs of clusters.

These three hierarchical methods have been singled out not only because of their fairly widespread use but also because they illustrate some of the trade-offs among the algorithms. The single and complete linkage methods have the attractive feature that the topologies of the dendrograms are invariant under monotone transformations of the distances. However, the single linkage method is frequently shunned by practitioners because of its propensity to produce long, stringy clusters that are of little interest (see, e.g., Sneath and Sokal, 1973, page 223). The complete linkage method has the opposite problem of being "biased" in the direction of small compact clusters (see, e.g., Sneath and Sokal, 1973, pages 222 and 223). Other criticisms of complete linkage have been raised by Hartigan (1981); see also Section 3.3.3. For more discussion of the pros and cons of the single and complete linkage methods, see Shepard and Arabie (1979). The average linkage method is a compromise between the extremes of the other two, but it does not have their invariance feature.

The broad-based popularity of the hierarchical approach to clustering is illustrated by the fact that all but two of the practical applications mentioned earlier were based on some method of this type. Simplicity and availability are probably the primary reasons for their frequent use rather than performance or optimality.

The partitioning methods offer a class of alternatives that are generally more flexible on the one hand and more difficult to use on the other. In a typical algorithm of this type, an initial specification of cluster "centers" is made. Then observations are assigned to the clusters according to their nearest cluster centers. Cluster "centers" are refined and observations are reallocated. The procedure continues until some type of stability is achieved. Among the details that vary across the algorithms are the starting points, the frequency in which updating of the cluster centers occurs, the flexibility to change the number of clusters and the manner in which clusters are added or deleted. Perhaps the best known of the partitioning procedures is the *k*-means algorithm (see, e.g., Hartigan, 1975, Chapter 4).

One can imagine situations where a standard hierarchical or partitioning algorithm would be inappropriate for the data because of the need to allow for overlapping clusters. Although easy to contemplate, there has been relatively little work in this area. Perhaps this is due more to the shortage of satisfactory algorithms than to the potential for applications. Several methods are mentioned by Seber (1984, pages 387 and 388). See also Arabie (1977), Arabie and Carroll (1980) and Shepard and Arabie (1979).

This cursory discussion of clustering methods has, to this point, concentrated on numerical algorithms for identifying clusters. However, many practitioners seem to rely on methods whose primary objective is something else.

A common example is the use of principal components analysis: the data are projected down into the space of the first two or three principal components and clusters are then identified by eye. Reliance on the eyes may seem unscientific, but they do offer great flexibility and efficiency in processing what they can see. The more important issue is the appropriateness of the projection. It offers, in some sense, the two- or three-dimensional space of maximum variance in the data, or it can be thought of as the two- or three-dimensional plane of closest fit to the data configuration. However, neither objective equates to cluster seeking. In fact, it is easy to conjure up data for which such a projection would be useless for "seeing" clusters.

This illustrates the risks involved in relying on such methods for extracting clusters. If cluster analysis is a serious objective, then one is probably better off using clustering methods—in spite of their limitations and imperfections.

The static graphic displays mentioned in Section 2.2.1 could also be used for cluster detection by subjective visual grouping of the pictorial representations of the data. Without any clues as to the cluster structure, this can be hard when either the number of variables or observations is large.

Sophisticated dynamic graphic systems that allow one to see data from many different perspectives are perhaps the best current hope for a genuine methodology breakthrough in multivariate data analysis generally and cluster analysis particularly. An easy-to-use and accessible system that will systematically traverse the data space along directions most likely to reveal clustering is a realistic objective for the near future. The directional guidance will come from numerical intelligence gleaned from the data, the cluster identification will come from human intelligence and what is seen by eye, and the implementation will be eased by the hardware and software tools now emerging for artificial intelligence. Many of the parts for such a system are already in place or under vigorous development (see, e.g., Asimov, 1985; Buja, Hurley and McDonald, 1986; Donoho, Donoho and Gasko, 1985; Fisherkeller, Friedman and Tukey, 1974; Friedman and Tukey, 1974; Huber, 1985).

2.3.3 Perspective

To place clustering methodology in perspective, it may be helpful to dissect the main steps in the process of using these methods and to comment on some of the stumbling blocks. Three stages can be identified: (i) the *input stage* where the data are adjusted as needed into a form suitable for clustering, (ii) the *algorithm* where a clustering method is applied to the adjusted input data and (iii) the *output stage* where the results of applying the algorithm are studied for statistical sensibleness. Although the choice of algorithm or algorithms is surely important, the other two stages are at least as crucial for achieving sound results.

The input stage involves the choice, transformation and scaling of variables plus—for many algorithms—commitment to a distance metric. It is obvious that the analysis depends upon the selection of useful variables in the first place. Coming up with an effective list is not always easy. For example, cultural and personal biases may enter (Sokal, 1974). To be safe, there is a temptation to throw in everything that comes to mind, but that is also a trap. Extra variables that do not reveal anything about the cluster structure tend to dilute the analysis and cause the standard algorithms to go astray. There are few statistical procedures to assist in variable selection for clustering; see Fowlkes, Gnanadesikan and Kettenring (1987) for one method.

One can also expect that the clustering results will be very sensitive to transformations of the input variables. In archaeology, analyses are often based on trace elements and it is commonly argued that logarithms of such variables should be employed. Such transformations can, in effect, create, accentuate, diminish or destroy clusters.

The scaling or weighting of variables needs careful thought. In its simplest form, this may involve a conscious scaling up of a variable in order to magnify its impact relative to other variables. If these variables are measured in the same units, then this rescaling is relatively easy to rationalize.

Of more concern is how to equalize the roles of the variables, especially when their measurement units are not comparable, or, going further, to make the results invariant to nonsingular linear transformations of the data. These are tricky problems that are circular in the sense that one really needs to know the cluster structure to begin to grapple with them correctly.

A common solution to equalizing the roles is to divide each variable by the square root of its total variance. However, this form of equalization is artificial and can, e.g., inappropriately downplay a variable that exhibits strong cluster structure. The only defense for this approach is that it may be better than doing nothing.

A more effective way to rescale the individual variables would be to utilize estimates of the within cluster variability in place of the total variance. Statistics based on the smallest absolute pairwise differences of the data on a particular variable are natural to consider for this purpose.

This line of thinking presumes that the within-cluster variability is roughly comparable across clusters. If this is true in a multivariate sense as well, then pairwise differences in the vector observations can be used in an iterative fashion to develop an estimate $\mathbf{W}^*$, of the within-cluster variability without knowing the clusters in advance (Art, Gnanadesikan and Kettenring, 1982). Scaling the data by $\mathbf{W}^{*-1/2}$ should then render the clusters roughly spherical in shape and hence amenable to detection by algorithms, like the k-means one, that are particularly effective at detecting such clusters.

More work is needed on effective ways of scaling the data when the assumption of within-cluster homogeneity is inappropriate either in the univariate or multivariate sense. In such cases, it may be necessary to consider several possible scalings.

For algorithms taking distances or dissimilarities as their inputs, one must consider, in addition to the previously mentioned issues, the type of distance metric that will most effectively reflect the kinds of differences between observations that are important for a particular problem. Two very popular types of distances are Manhattan, which is the sum of absolute differences across variables for two observations, and Euclidean, which is the square root of the sum of squared differences. Many other types are discussed in the standard cluster analysis books; see, e.g., Sneath and Sokal (1973, Chapter 4).

An overview of algorithms has already been provided in Section 2.3.2. Each has limitations, but their overall performance can be ameliorated by careful choice of the inputs to them.

A temptation worth resisting is to take the output of any clustering algorithm and to accept it without scrutiny. Issues worth investigating include cluster location, dispersion, orientation, separation, tightness and stability. Elementary data analytic displays and summary statistics can help address many of these. Resampling and perturbation techniques are potentially of use for checking on stability, but exactly what should be done is not so clear.

Several ideas in this vein are mentioned in Gnanadesikan, Kettenring and Landwehr (1977). Some examples include:

> *Distances.* Plot the distance of each object to all the cluster centroids to check on the strength of its association with a particular cluster.
>
> *Summary statistics.* For any two clusters, measure their separation on each variable according to the p-value of the usual t statistic to find out which ones provide relatively more discrimination.
>
> *Projections.* Treating the clusters as fixed groups, display them in the space of the first few discriminant variables to assess separation, tightness, orientation and dispersion; see also Gnanadesikan, Kettenring and Landwehr (1982).
>
> *Sensitivity analysis.* Check stability by adding noise to the original data and comparing clusters from the original and perturbed data sets.

There is a need for more ideas and more experimentation on effective ways of analyzing the output of clustering algorithms. This would include further development of practical inferential tools for assessing cluster validity. For further reading on statistical inference in clustering, see, e.g., Bock (1985), Fowlkes and Mallows (1983), Sneath and Sokal (1973, pages 284–287), as well as the discussion in Section 3.3 of this report.

3. THEORY

3.1 Introduction

This section emphasizes certain theoretical statistical aspects of the techniques of discriminant and cluster analyses discussed in Chapter 2. Section 3.2

pertains to discriminant analysis and includes a discussion of the performance of explicit discrimination rules, the estimation of misclassification costs and nonparametric techniques. Section 3.3 is concerned with the theory of clustering algorithms and includes a discussion of high density clusters, complete linkage, single linkage, minimum spanning trees, mixture models, inference about the number of clusters and ultrametric and evolutionary distances.

3.2 Theoretical Issues in Discriminant Analysis

3.2.1 Introduction

The questions and techniques which are addressed in this chapter are quite simple to state, but are rich in areas of application. Problems of computer-aided diagnosis in medicine, military surveillance and speech recognition, for example, can sometimes be formulated in a common way. Observations are available from a source that belongs to a unique population among a set of populations. For example, the observations might be gathered by radar as a plane flies over a ship at sea. The ship is the source of the data. It is assumed that the ship is one (the unique population) of five ship types (the set of populations), and the classification question is which one. Such data are often termed "test" data or "test samples." The set of candidate populations is assumed to be finite. Indeed, two is perhaps the most popular number. The values of a set of features—that is, covariates, or "independent" variables such as the radar measurements on the ship—are available for each unit to be classified, and it is on the basis of these data that assignments are made. In problems which are our focus here, something is assumed known about the conditional distributions of features given population (or perhaps more commonly "class") membership. These distributions are generally known or assumed to be known from some previous experience—in which case the problem of class assignment can be rather simple—or learned from other data, a "learning" or "training" sample. Prior probabilities of class membership and costs of misclassifying candidate observations are implicit to most schemes for "classification" or "discrimination." They will be made explicit in what follows. But before that beginning of our more technical discussion, we draw the principal distinction between the topic of this section and the notion of "classification" which is usually associated with clustering. Namely, here it is assumed without question that the classes are well-defined. Thus, what is discussed here applies to the assignment of a cancer patient to one of several recognized stages of illness on the basis of some data, and decidedly not to the question of how many stages it makes sense to ascribe to the cancer itself. Solution to the latter question, however fundamental to science and technology, is a requisite preprocessing to the tasks we confront here. Also, we will have little to say regarding the fundamental question of feature selection, which has been so prominent a part of recent literature on regression (see, e.g., Shibata, 1981). On the other hand, the probability distribution of the data within a given class will not be assumed to be known in most of our discussion. In fact, it will be evident that to do effective discrimination one need not know these distributions—only one aspect of the rank ordering of certain linear combinations of their (generalized) densities.

The formulation of "discrimination" which follows is adapted from the recent book by Breiman, Friedman, Olshen and Stone (1984). Suppose that the variable Y can assume any integer value between, say, 1 and $J < \infty$ and that the value of Y is unknown; Y represents "class." [Note: In Section 2 the total number of classes or groups was denoted as G instead of J.] But suppose that features $\mathbf{X}$ of Y are observed or are otherwise available. On the basis of $\mathbf{X}$ we wish to infer Y. Assume further that there are J densities $f(\cdot \mid Y = j)$ with respect to some dominating measure μ on a space $\mathbf{X}$ which is the range of $\mathbf{X}$. We use a dot to indicate the argument of the conditional density f and a vertical bar to denote "given" or "given that." Generally speaking, $\mathbf{X}$ can be taken to be Euclidean, but it is the decided exception in practice for all the f's to be absolutely continuous (with respect to Lebesgue measure) because discrete features are common in applications. We denote by $\pi(j) = P(Y = j)$ the "prior" probability that an observation whose class membership is unknown is of class j. Although the use of prior probabilities can be controversial in other settings, it seems difficult to formulate discrimination in a satisfactory way without them. Also, in the present context they often have a compelling frequentistic basis. Applications of the technologies under discussion are often in the context of new data like those of an existent and at least moderately well-understood data base.

A more controversial aspect of our formulation is the set of numbers $C(i \mid j)$, the cost of classifying an observation to class i given it is of class j. It is typically convenient to take $C(i \mid i) = 0$ and $C(i \mid j) > 0$ for $i \neq j$. Consider the problem of classifying a patient who enters an emergency room with a complaint of chest pain as to whether he has suffered a heart attack. The patient who desires that the most extensive medical resources be made available to him just in case he has actually had a heart attack may have C's that are very different from those of the director of a coronary care unit, who must make careful and responsible allocations of scarce, expensive resources. (See Breiman, Friedman, Olshen and Stone, 1984, pages 176 and 177.)

A (nonrandomized, measurable) decision rule d—such rules are all that we need consider in discrimination—arises from a partition of $\mathbf{X}$ into disjoint subsets. If $\mathbf{X}$ belongs to the jth of these, we decide that $Y = j$, that is, $d(\mathbf{X}) = j$.

The expected cost of the rule d is

$$\begin{aligned}&\sum_{j=1}^{J} \pi(j)\left[\sum_{i=1}^{J} C(i\mid j)P(d(\mathbf{X})=i\mid Y=j)\right]\\ &= \sum_{j=1}^{J} \pi(j)\left[\sum_{i=1}^{J} C(i\mid j)\int_{\{d(\mathbf{X})=i\}} f(\mathbf{x}\mid Y=j)\mu(d\mathbf{x})\right].\end{aligned} \tag{18}$$

A rule d_B is called a "Bayes rule" if its cost is as small as possible. All Bayes rules have the same expected cost. One concludes that if $d(\mathbf{x}) = i$ implies

$$\sum_{j=1}^{J} \pi(j)C(i\mid j)f(\mathbf{x}\mid Y=j) \le \sum_{j=1}^{J} \pi(j)C(i'\mid j)f(\mathbf{x}\mid Y=j)$$

for all i', then d is a Bayes rule. Specialize to the case $J = 2$ to see that for μ almost all $\mathbf{x}$, $d_B(\mathbf{x}) = 1$ implies

$$\frac{f(\mathbf{x}\mid Y=1)}{f(\mathbf{x}\mid Y=2)} \ge \frac{\pi(2)C(1\mid 2)}{\pi(1)C(2\mid 1)}. \tag{19}$$

So obviously we "know" a Bayes rule if we know the "densities" $f(\cdot\mid Y=1)$ and $f(\cdot\mid Y=2)$. This observation is precisely why discrimination or classification as pursued here is a subject of serious inquiry. Imagine, as is the case in many applications, that $\mathbf{X}$ is, say, twenty-dimensional, and that we have available learning samples of, say, hundreds or even thousands of observations from $f(\cdot\mid Y=1)$ and $f(\cdot\mid Y=2)$, which are not assumed to be of any particular functional form. Think of the crudest partitioning of the axes of $\mathbf{X}$ into two parts each; the resulting product partition has a total of 2^{20} bins, or more than a million. Most will have no members of the learning sample at all. The estimation of the densities is thus hopeless. But from (19) it is clear that we need not know the densities exactly. We only need to know when their ratio exceeds a specified constant. Fortunately, it is far easier to know when that occurs than it is to know the densities themselves.

One needs a benchmark from which to gauge the performance of any classifier d; an obvious candidate is the no data Bayes rule. Such a rule assigns every observation to a class i for which

$$\sum_{j=1}^{J} \pi(j)C(i\mid j)$$

is minimized. In case of ties, the usual convention is to take the smallest minimizing index.

We have discussed learning samples in casual terms, but by so doing we obscure what can be a troubling distinction in theoretical work and an important difference between prospective and retrospective studies. In the case of the former, one generally assumes that the learning sample is of the form $(\mathbf{X}_1, Y_1), \cdots, (\mathbf{X}_N, Y_N)$, where the pairs are independent and independent of $(\mathbf{X}, Y)$, and each $(\mathbf{X}_i, Y_i)$ is distributed as $(\mathbf{X}, Y)$; N is assumed to be a nonrandom constant. In retrospective studies it often happens that an existing data base is searched for preassigned numbers of pairs in each of the J classes. Then the assumption of unconditional independence is not reasonable, and the basis for choosing the prior probabilities may be somewhat unclear. Yet it can be plausible to assume that Y_n is $1, \cdots, J$ valued for each $n = 1, 2, \cdots$ and that conditioned on $Y_n = j_n$ for $n \ge 1$, the random variables $\mathbf{X}_n$, $n \ge 1$ are independent. As Efron (1975, page 898) indicates, in most practice the distinction between unconditional and conditional independence is ignored. He details how in normal linear discrimination and in logistic regression—both to be discussed—one can deal theoretically with the more difficult case of conditional independence. In other contexts, theoretical problems regarding the subtle issue at hand are raised by Olshen in his discussion of Stone's (1977) landmark paper and by Stone in his reply (page 641) to the discussants. See also Gordon and Olshen (1978), and especially Chapter 12 of Breiman, Friedman, Olshen and Stone (1984). For ease of exposition, in what follows the point of view is that the observations that make up the learning sample are independent and identically distributed.

3.2.2 The Fisher Linear Discriminant and Some of Its Children

As with much of what is worthwhile in statistics, the starting point for a discussion of explicit rules for discrimination is a result of Fisher (1936). The Fisher linear discriminant and procedures to which it has led have been remarkably useful in practice.

Suppose that $J = 2$ and that given $Y_i = j$, $\mathbf{X}_i \sim N_p(\boldsymbol{\mu}_j, \boldsymbol{\Sigma})$. That is, $\mathbf{X}$ has a p-dimensional normal distribution with mean $\boldsymbol{\mu}_j$ and covariance $\boldsymbol{\Sigma}$. Then, if $\boldsymbol{\Sigma}$ is of full rank and prime denotes transpose, one calculates that a Bayes rule

$$d(\mathbf{x}) = 1 \quad \text{if } \beta_0 + \boldsymbol{\beta}'\mathbf{x} > 0,$$

where

$$\begin{aligned}\beta_0 &= \log\left(\frac{\pi(1)C(2\mid 1)}{\pi(2)C(1\mid 2)}\right)\\ &\quad - \tfrac{1}{2}(\boldsymbol{\mu}_1'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_2),\\ \boldsymbol{\beta}' &= (\boldsymbol{\mu}_1' - \boldsymbol{\mu}_2')\boldsymbol{\Sigma}^{-1}.\end{aligned} \tag{20}$$

Otherwise $d(\mathbf{x}) = 2$.

The coefficients $\boldsymbol{\beta}'$ in (20) have been described by Truett, Cornfield and Kannel (1967) as "the amount by which the logit of risk increases for unit increase in the risk factor." The logit of risk is the log odds (given its features) that an observation of unknown class actually is of class 1. There is no loss in assuming that $\mathbf{\Sigma}$ is nonsingular, because singular cases can always be made nonsingular by an appropriate reduction of dimension. In practice, $\boldsymbol{\mu}_1$, $\boldsymbol{\mu}_2$ and $\mathbf{\Sigma}$ are seldom assumed known and must be estimated from the learning sample. If $N_1 = \#\{i \le N: Y_i = 1\}$ and $N_2 = \#\{i \le N: Y_i = 2\}$, then the method of maximum likelihood gives

$$\hat{\boldsymbol{\mu}}_1 = \frac{1}{N_1} \sum_{i \le N: Y_i = 1} \mathbf{X}_i,$$

$$\hat{\boldsymbol{\mu}}_2 = \frac{1}{N_2} \sum_{i \le N: Y_i = 2} \mathbf{X}_i, \tag{21}$$

$$\hat{\mathbf{\Sigma}} = \frac{1}{N} \left\{ \sum_{i \le N: Y_i = 1} (\mathbf{X}_i - \hat{\boldsymbol{\mu}}_1)(\mathbf{X}_i - \hat{\boldsymbol{\mu}}_1)' + \sum_{i \le N: Y_i = 2} (\mathbf{X}_i - \hat{\boldsymbol{\mu}}_2)(\mathbf{X}_i - \hat{\boldsymbol{\mu}}_2)' \right\};$$

if, in addition, $\pi(1)$ and $\pi(2)$ are unknown, then

$$\hat{\pi}(1) = N_1/N \quad \text{and} \quad \hat{\pi}(2) = N_2/N.$$

The maximum likelihood estimates can be substituted for their corresponding parameters in (20). There results one version of an estimated linear discriminant. This estimate of d and various stepwise versions of it are widely used in practice, even when the conditional distributions of $\mathbf{X}$ given Y not only do not share a common covariance, but also when the distributions are not normal at all. Robustness of the Fisher linear discriminant has been noticed by many users and studied by some. One reference on the subject is Lachenbruch (1982).

Even though the most important issues involve errors of classification, one may ask from a decision theoretic point of view how the estimators defined by (20) and (21) perform when the model is correct. Thus, for some positive definite matrix $\mathbf{Q}$, one might study the "risk"

$$E\{(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})'\mathbf{Q}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \mid \boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \mathbf{\Sigma}\}, \tag{22}$$

and ask whether any other estimator does at least as well riskwise as $\hat{\boldsymbol{\beta}}$ for all $\boldsymbol{\mu}_1$, $\boldsymbol{\mu}_2$ and $\mathbf{\Sigma}$, and better for some values. The phenomenon, which leads to the coordinates of a vector of estimated parameters being pulled toward a prechosen value, with salutary decision-theoretic implications, surfaces here provided $N_1 \ge 2$, $N_2 \ge 2$ and $N > p + 3$. This idea figures in what has come to be known as James-Stein estimation. See Stein (1956) and James and Stein (1961). Haff (1986) has shown that in this case the estimated Fisher linear discriminant coefficients $\hat{\boldsymbol{\beta}}$ can be improved by improving the estimate $\hat{\mathbf{\Sigma}}^{-1}$ of $\mathbf{\Sigma}^{-1}$. Basically, if one adds a multiple of $\mathbf{Q}$ which depends on the trace of $\hat{\mathbf{\Sigma}}^{-1}\mathbf{Q}$ to $\hat{\mathbf{\Sigma}}$ before inverting and substituting into (20), then the resulting estimate of $\boldsymbol{\beta}$ improves upon $\hat{\boldsymbol{\beta}}$, at least in the sense given by (22).

Another departure from the substitution of maximum likelihood estimates into (20) is given by "logistic regression." (See Section 2.2.5.) Its starting point is the observation that if the stated normal model is correct, then

$$P(Y = 1 \mid \mathbf{X}) = \frac{\exp(\beta_0 + \boldsymbol{\beta}_1'\mathbf{X})}{[1 + \exp(\beta_0 + \boldsymbol{\beta}_1'\mathbf{X})]}, \tag{23}$$

$$P(Y = 2 \mid \mathbf{X}) = 1 - P(Y = 1 \mid \mathbf{X}).$$

To estimate $(\beta_0, \boldsymbol{\beta}_1')$, one can maximize the conditional (binomial) likelihood based on (23). (This maximization must be done on a computer by numerical maximization techniques (see Section 4.2.3), but is relatively easy as such problems go because the likelihood is unimodal and logarithmically concave.) These estimates can be substituted into (20). Of course, if the original normal model is correct, this estimated d must somehow be less efficient than that based on (21), which utilizes the full likelihood function. However, the "logistic" likelihood function is valid under more general assumptions of the exponential family than is the likelihood function which leads to (21). Efron (1975) has investigated the efficiency of logistic regression relative to estimated Fisher linear discrimination when the normal model is correct and found it to be "between one half and two thirds as effective as normal discrimination for statistically interesting values of the parameters."

Yet another departure from the model with which this section begins is this. Suppose that given $Y_i = j$, $\mathbf{X}_i \sim N_p(\boldsymbol{\mu}_j, \mathbf{\Sigma}_j)$, where $\mathbf{\Sigma}_1$ may not equal $\mathbf{\Sigma}_2$. (See also Section 2.2.4.) It is straightforward to compute a Bayes rule, but several individuals have taken a different point of view. Suppose first that $p = 1$. Then in an obvious notation and as Becker (1968) indicates,

$$\frac{|\mu_1 - \mu_2|}{\sigma_1 + \sigma_2}$$

is a useful measure of the separation of the two conditional distributions. If $p > 1$, then because linear functions of normal vectors are normal, one might consider $\mathbf{b}'(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$ for various nonrandom vectors $\mathbf{b}$. If $Y = j$, $\mathbf{b}'\mathbf{X}_j \sim N(\mathbf{b}'\boldsymbol{\mu}_j, \mathbf{b}'\mathbf{\Sigma}_j\mathbf{b})$, and thus Becker's measure extends to

$$S(\mathbf{b}) = \frac{|\mathbf{b}'(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)|}{(\mathbf{b}'\mathbf{\Sigma}_1\mathbf{b})^{1/2} + (\mathbf{b}'\mathbf{\Sigma}_2\mathbf{b})^{1/2}},$$

a criterion actually studied earlier by Anderson and Bahadur (1962). Suppose now that one chooses to compute a linear function of $\mathbf{X}$, say $\mathbf{b}_0'\mathbf{X}$, so that

$d(\mathbf{X}) = 1$ if $\mathbf{b}_0'\mathbf{X} \le c$, and otherwise $d(\mathbf{X}) = 2$. If the goal is to minimize $P(d(\mathbf{X}) \ne Y \mid Y)$ (that is, the probability of a mistake), then pick $\mathbf{b}_0$ to maximize $S(\mathbf{b})$ and

$$c = \frac{(\mathbf{b}_0'\boldsymbol{\Sigma}_2\mathbf{b}_0)^{1/2}\mathbf{b}_0'\boldsymbol{\mu}_1 + (\mathbf{b}_0'\boldsymbol{\Sigma}_1\mathbf{b}_0)^{1/2}\mathbf{b}_0'\boldsymbol{\mu}_2}{(\mathbf{b}' \sum_1 \mathbf{b}_0)^{1/2} + (\mathbf{b}_0' \sum_2 \mathbf{b}_0)^{1/2}}.$$

This fact follows from work of Anderson and Bahadur (1962). Chernoff (1972, 1973a) indicates how to compute $\mathbf{b}_0$ and notes that the common probability of a mistake is $\Phi(-S(\mathbf{b}_0))$. Of course, in order to implement the Anderson-Bahadur-Chernoff ideas in practice, one must estimate those parameters which are not known.

In case $J > 2$, the slick appearance of the Fisher linear discriminant Bayes rule disappears. Generalizations of logistic regression to this case have been studied by many authors. Chapters 4 through 6 of the book by McCullagh and Nelder (1983) are an excellent recent summary. We choose not to provide details here because the issues which arise pertain to modeling conditional distributions of $\mathbf{X}$ given Y, and the main aspects are not discussed in the explicit context of discrimination.

3.2.3 Estimating Misclassification Costs

When a decision rule d is estimated from a learning sample, then it follows from Fubini's theorem that the expression (18) for the expected cost of d is really the conditional expected cost given the learning sample. The unconditional expected cost is the expectation of (18) with respect to the distribution of the learning sample: n independent copies of the joint distribution of $(\mathbf{X}, Y)$. Of course, that distribution is assumed here to be unknown, and we are left to estimate the unconditional expected cost from data. (See earlier discussion in Section 2.2.1.) In the best of all worlds, we have available a genuinely independent test sample, taken to be independent of the learning sample, distributed as $(\mathbf{X}, Y)$, and large enough to permit reasonable inferences. This happy situation occurred, for example, with the work of Goldman, Weinberg, Weisberg, Olshen, Cook, Sargent, Lamas, Dennis, Deckelbam, Fineberg, Stiratelli and the Medical Housestaffs at Yale-New Haven Hospital and Brigham and Women's Hospital (1982), but unfortunately seems to be the exception rather than the rule in practice. In the absence of such a test sample we are left to estimate the misclassification cost associated with the prospective use of d from the learning sample.

The starting point for the estimation of misclassification costs from the learning sample is the so-called "resubstitution" estimate. If the prior probabilities are assumed to be known and the learning sample is of cardinality N, then the resubstitution estimate is

$$\text{(24)} \qquad \sum_{j=1}^{J} \frac{\pi(j)}{\#\{i \le N : Y_i = j\}} \sum_{i \le N : Y_i = j} C(d(\mathbf{X}_i) \mid j).$$

If the π's are not assumed to be known, then the analogue to (24) is

$$\text{(25)} \qquad \frac{1}{N} \sum_{k=1}^{N} C(d(\mathbf{X}_k) \mid Y_k).$$

These resubstitution estimates may occasionally be of practical value with some parametric procedures such as the Fisher linear discriminant. However, they are subject to enormous over optimistic biases for nonparametric techniques such as those discussed in the next section. One approach to overcoming the biases is that of cross-validation. By m-fold cross-validation (see Breiman, Friedman, Olshen and Stone, 1984) is meant this. The learning sample is divided at random into m disjoint subsets of approximately equal size. Call them $\mathbf{L}_1, \cdots, \mathbf{L}_m$. Successively, the data of $\mathbf{L}_v$ are deleted to yield $\mathbf{L}^{(v)}$. The classifier $d^{(v)}$ is computed from those data in $\mathbf{L}^{(v)}$ according to the same algorithm by which d was calculated from all of the data. Then $\mathbf{L}_v$ is used as a test sample. The process is repeated for $v = 1, \cdots, m$ and the results averaged. In order to simplify the exposition we suppose in what follows that (25) applies. Thus, the estimated cost of misclassification is

$$\frac{1}{m} \sum_{v=1}^{m} \frac{m}{mN - N} \sum_{i:(\mathbf{X}_i, Y_i) \in L_v} C(d^{(v)}(\mathbf{X}_i) \mid Y_i).$$

The most widely advertised choice of m is N, the "leave-one-out" method. However, N-fold repetitions of computationally expensive procedures is not practical and seems not to be best for theoretical reasons in some cases (see Efron, 1983, page 327).

The discussion thus far has been vague as to how the $\mathbf{L}_v$ are chosen. For example, one might think of stratifying the sampling by class even in the present context in which the π's are not assumed known. With cross-validation stratified, each class is as nearly as possible equally represented in each of the m $\mathbf{L}_v$. The results on the effects of enforced stratification are skimpy and specialized—see Breiman, Friedman, Olshen and Stone (1984, pages 80, 179, 245–247) and Olshen, Gilpin, Henning, Lewinter, Collins and Ross (1985, Section 5)—but thus far what theory there is suggests that stratification does not hurt, and typically helps.

One popular approach to the estimation of misclassification costs was termed the "bootstrap" by Efron when he introduced it in 1979. His starting point is the resubstitution estimate (25). To this he adds a bias adjustment, which is arrived at as follows. Generate a random sample, a "bootstrap" sample, with replacement from the learning sample. Compute a classifier d^B from the bootstrap sample by the same algorithm which produced d. Classify the learning sample and the bootstrap sample by d^B, and compute their misclassification costs. Because d^B is tailored to the bootstrap sample, it typically will do better for the bootstrap sample than for the learning sample. The difference in these two misclassification costs is the estimated bias adjustment. If a superscripted asterisk denotes membership in the bootstrap sample, then the bias adjustment can be written

$$(26)\quad \frac{1}{N}\sum_{k=1}^{N} C(d^B(\mathbf{X}_k) \mid Y_k) - \frac{1}{N}\sum_{k=1}^{N} C(d^B(\mathbf{X}_k^*) \mid Y_k^*).$$

Expression (26) can and should be computed from independent bootstrap samples and averaged to obtain an overall bias adjustment, which then is added to (25) to give the final estimated misclassification cost for d.

By way of comparison, the bootstrap technique tends to be less variable in its estimation of misclassification costs than is cross-validation. But it can be more biased. For parametric procedures in which resubstitution estimates are not so severely biased, the bootstrap typically outperforms cross-validation, as Efron's work indicates (see Efron, 1983, and its references). However, in utterly nonparametric situations, the bootstrap can badly underestimate misclassification costs (Breiman, Friedman, Olshen and Stone, 1984, Section 11.7) and even be inconsistent. The cited example—which amounts to a single nearest neighbor rule applied to a problem where $\mathbf{X}$ and Y are actually independent—and one way out of the severe biases possible with the bootstrap rest on the same simple observation for a starting point. That is, in any given bootstrap sample, the expected number of observations (from among the N in the learning sample) which actually appear is

$$N\left(\frac{N-1}{N}\right)^N = N(1 - e^{-1}),$$

or approximately 63.2% of the learning sample.

Efron (1983) has several approaches to correcting bootstrap biases, but none more intriguing or more successful in his simulations than the ".632 estimator." The .632 estimate of misclassification cost is a weighted average of the resubstitution estimate (25) and the bootstrap estimate computed for those members of the learning sample not occurring in the bootstrap sample. The respective weights are .368 and .632. Neither Efron's (1983) work on this new estimate of misclassification cost nor the results of Gong (1982) are definitive; and, moreover, the .632 rule only partly meets the criticism of Breiman, Friedman, Olshen and Stone. Yet it still seems to merit further study.

3.2.4 Nonparametric Techniques

It is clear from Section 3.2.1 that any procedure by which probability densities are estimated carries with it a technique for discrimination. So, for example, kernel and series expansion procedures for density estimation imply corresponding procedures for the problem at hand. At the same time, it is clear, too, from Section 3.2.1 that for most problems of practical interest the relevant densities cannot be estimated at all well. Nonetheless, there are available a variety of procedures which confront the discrimination problem directly and are not tied to parametric assumptions. In this section two approaches, nearest neighbors—of which mention has been made—and recursive partitioning, are discussed. These techniques are applicable to the general regression problem as well. Not only are discrimination and density estimation closely connected, but also discrimination is closely related to regression, and regression can be viewed as a special instance of generalized density estimation. It is to these connections that we now turn.

Consider a general $(\mathbf{X}, Y)$ pair with Y real-valued and $E(\mid Y \mid) < \infty$; if $\mathbf{X}$ is as before, then $h(\mathbf{x}) = E(Y \mid \mathbf{X} = \mathbf{x})$ is the regression of Y on $\mathbf{X}$. It is a particular case of this situation that Y has finite range and assumes only the values $1, \cdots, J$. Then, with an obvious indicator function notation, one may write

$$(27)\qquad Y = \sum_{j=1}^{J} jI_{[Y=j]}.$$

The problem of discrimination can be viewed thus as a special regression problem in which Y can be written as in (27) and the estimate of $h(\mathbf{x})$ is of the same form, say

$$(28)\qquad \sum_{j=1}^{J} jI_{[f(\mathbf{X},\text{learningsample})=j]}.$$

If we return to the general regression problem, then we may think of μ defined for measurable subsets B of $\mathbf{X}$ by $\mu(B) = E(YI_{[\mathbf{X}\in B]})$. The measure μ is absolutely continuous with respect to the distribution of $\mathbf{X}$, and in fact the regression function $h(\mathbf{x})$ is the density (Radon-Nikodym derivative) of μ with respect to that distribution.

There could hardly be a simpler approach to discrimination than that of the single nearest neighbor rule. Its motivation is that of the physician who tentatively diagnoses his present patient as having the same disease as what was known to be the correct diagnosis for that past patient whose symptoms and history most closely match those at hand. With this motivation it may seem reasonable to the physician to compare the present patient with a number of previous patients who are also quite similar, although not necessarily the "most" similar. Thus we are led to more general nearest neighbor type procedures. In subsequent discussion, for convenience we take $C(i \mid j) = 1$ if $i \neq j$, 0 otherwise. Much of the presentation is based on the important paper of Stone (1977).

Assume no particular functional form for the joint distribution of $\mathbf{X}$ and Y. Then it follows from our assumptions regarding C and the formulation of Section 3.2.1 that a Bayes rule d_B satisfies

$$(29) \qquad d_B(\mathbf{x})\ i \quad \text{if } P(Y = i \mid \mathbf{X}) \geq P(Y = i' \mid \mathbf{X}) \text{ for all } i'.$$

Denote its cost of misclassification by R. Motivated by the informal discussion of the previous paragraph, imagine estimating $P(Y = i \mid \mathbf{X})$ from the learning sample by

$$(30) \qquad \sum_{k=1}^{N} W_{Nk}(\mathbf{X}; \mathbf{X}_1, \cdots, \mathbf{X}_N) I_i(Y_k),$$

where the W's are weights and $I_i(Y_k) = 1$ if $Y_k = i$, and 0 if not. The estimated conditional probabilities can then be plugged into (29) to estimate a Bayes rule. The cited simple nearest neighbor rule is clearly of this form, where

$$W_{Nk} = W_{Nk}(\mathbf{X}) = W_{Nk}(\mathbf{X}; \mathbf{X}_1, \cdots, \mathbf{X}_N) = 1$$

for $\mathbf{X}_i$ the set of covariates in the learning sample which is "closest" to $\mathbf{X}$. We can ask what happens as N grows without bound to the expected cost of a Stone type rule, i.e., a rule of the form (29), (30). In our formulation, we suppose (with slight loss of generality) that $W_{Nk} \geq 0$, all N, k, and that $\sum_{k=1}^{N} W_{Nk} = 1$. Stone indicates that the limiting expected cost of any such classifier is at most

$$(31) \qquad R\left(2 - \frac{J}{J-1} R\right) \leq 2R$$

no matter what be the joint distribution of $(\mathbf{X}, Y)$ provided only that two conditions are satisfied:

$$(32) \qquad \text{for each} \quad a > 0 \quad \sum_{k=1}^{N} W_{Nk}(\mathbf{X}) I_{[\|\mathbf{X}_k - \mathbf{X}\| > a]} \rightarrow 0$$

in probability; and there is a $D \geq 1$ such that, for every (measurable) nonnegative function f on $\mathbf{X}$

$$(33) \qquad E\left(\sum_{k=1}^{N} W_{Nk}(\mathbf{X}) f(\mathbf{X}_k)\right) \leq DE(f(\mathbf{X})).$$

If we take $\mathbf{X}$ formally to be Euclidean, then any positive definite norm will do in (32), which is simply a requirement that the classifier be asymptotically "local." Although (32) permits rules to do well in large samples for general distributions of $(\mathbf{X}, Y)$, any classifier which satisfies it cannot be asymptotically efficient in any reasonable sense relative to a Bayes rule for a parametric problem.

Note that $W_{Nk}(\mathbf{X}) f(\mathbf{X}_k)$ has the same distribution as $U_{Nk}(\mathbf{X}_k) f(\mathbf{X})$, where

$$U_{Nk}(\mathbf{X}_k) = W_{Nk}(\mathbf{X}_k; \mathbf{X}_1, \cdots, \mathbf{X}, \cdots, \mathbf{X}_N).$$

So the left-hand expectation in (33) can be written

$$E\left(f(\mathbf{X}) \sum_{k=1}^{N} U_{Nk}(\mathbf{X}_k)\right).$$

Therefore, (33) says that the random transformation which takes $f(\mathbf{X})$ to $f(\mathbf{X}) \sum_{k=1}^{N} U_{Nk}(\mathbf{X})$ must be bounded from the linear space of random variables with finite expectation to itself. This kind of condition resembles other necessary and sufficient conditions for convergence in ergodic theory. Stone's Theorem 1 puts it as both sufficient and in a certain sense necessary for his consistency results. Indeed, for any classifier constructed so that (32), (33) and

$$(34) \qquad \max_k W_{Nk}(\mathbf{X}) \rightarrow 0 \quad \text{in probability as } N \rightarrow \infty$$

are satisfied, the limiting expected misclassification cost is R itself. Of course, (34) entails that with arbitrarily large probability, no finite number of observations determine the rule.

An instance of the result which precedes and includes (31) was discovered by Cover and Hart (1967) in the case of single nearest neighbor classifiers. Their arguments apply to the situation where, with probability 1, nearest neighbors are uniquely defined. In Stone's work, any weight attached to kth nearest neighbors is divided equally when there are ties. Also, his notion of distance can involve a random scaling so that certain coordinatewise affinely invariant rules are covered. The first important theoretical work on the consistency of nearest neighbor classifiers in the presence of some regularity was by Fix and Hodges (1951). They gave the $K = K(N)$ nearest neighbors equal weights, where $K \rightarrow \infty$ and $K/N \rightarrow 0$.

In practice, nearest neighbor rules can be victimized by missing data, and by noise coordinates which should have been eliminated in the selection of covariates. Also, a criticism which might be made of Stone's (1977) work is that the weights are insufficiently adaptive because they ignore the Y's of the learning sample. An extension to the case where the W's depend on $Y_1, \cdots, Y_N$ as well as $\mathbf{X}_1, \cdots, \mathbf{X}_N$ was made by Gordon and Olshen (1980). They were studying tree-structured recursive partitioning rules, to which we turn next.

Tree-structured rules are the subject of the recent book by Breiman, Friedman, Olshen and Stone (1984). (We assume that the reader is familiar with the basic notion of a binary tree. A formal definition and details are given in Section 10.1 while an informal approach that relates trees, partitions and classifiers can be found in Sections 2.2 and 2.3 of this paper.) These techniques deal in a salutary fashion with the three mentioned shortcomings of nearest neighbor procedures, and they have proven successful in practice. The basic theme of recursive partitioning is based on the learning sample, **X**. The range of **X** is successively partitioned, or split, into "boxes" by a sequence of linear inequalities. (Unordered discrete covariates can be handled, too.) The partitioning amounts to choosing a sequence of yes-no questions that can be answered by knowing values of the features. A binary decision tree is associated with the process of partitioning, and the associated classifier d is constant on the terminal nodes, which correspond to terminal subsets of the partitioning; **X** corresponds to the root node of any tree **T**. The partitioning of a node $t \in \mathbf{T}$ is according to some criterion which is designed to produce daughter nodes more homogeneous as to class content than their parents. More precisely, if for any node s, $\hat{P}(s) = \{\#i \le N : \mathbf{X}_i \in s\}/N$, and $i(s)$ is an index of the "impurity" of the node s, then one may state the rule for partitioning t:

Form left daughter node t_L and right daughter node t_R so as to maximize

$$\hat{P}(t)i(t) - [\hat{P}(t_L)i(t_L) + \hat{P}(t_R)i(t_R)].$$

A popular index of impurity is the so-called Gini index

$$\sum_{i,j} C(i \mid j)p(i \mid t)p(j \mid t),$$

where

$$p(l \mid s) = \frac{\#\{i \le N : \mathbf{X}_i \in s \text{ and } Y_i = l\}}{\#\{i \le N : \mathbf{X}_i \in s\}},$$

and usually $C(i \mid j)$ is taken to be 0 or 1 according to whether $i = j$ or not. A large tree is grown initially, according to the cited splitting criterion or some other. (There are built-in constraints to this initial tree development which restrict the tendency of tree-structured methods to "sliver" nodes.) At each terminal node a Bayes rule is estimated from the members of the learning sample which belong to that node. Because the terminal nodes partition **X**, the process completely specifies a classifier d.

For any tree **T** and any $\alpha \ge 0$, a measure of the merit of **T** is

$$R_\alpha(\mathbf{T}) = \text{Resub}(\mathbf{T}) + \alpha \mid \tilde{\mathbf{T}} \mid,$$

where Resub(**T**) is the resubstitution estimate of d's misclassification cost, and $\mid \tilde{\mathbf{T}} \mid$ is the number of terminal nodes of **T**. Clearly, $R_\alpha(\mathbf{T})$ involves a trade-off of "bias" in its first term and "variance" in its second. For each α the large initial tree mentioned in the last paragraph has a subtree which is optimal in the sense of being the smallest subtree which minimizes R_α. As α increases, there arises a finite, *nested* sequence of optimally pruned subtrees. (To prune a tree at node t is to delete the branch of the tree that has t as its "root node.") The estimation of how well each will perform if used prospectively is accomplished by cross-validation; for some reason 10-fold cross-validation has been used in a variety of applications. The optimally pruned subtree for which the cross-validated misclassification cost is smallest is an obvious candidate for prospective use. Arguments have been advanced and techniques developed for some further pruning of this initially grown tree—see Breiman, Friedman, Olshen and Stone (1984).

If the partitions of **X** are nested as N grows, then the martingale convergence theorem bears upon the consistency of recursive partitioning decision rules. Regardless, available arguments all lean heavily on the uniform convergence of empirical probabilities of certain sets to their true probabilities. Thus, the pioneering work of Vapnik and Chervonenkis (1971, 1974) bears upon the asymptotic properties of tree-structured methods. Consistency in the sense described for nearest neighbor rules has been established for many recursive partitioning rules, and consistency with probability 1 has been too. Always the diameter of p is required to tend to 0, but also $\log N/N = o(\hat{P}(t))$ is required. Finally, the sizes of nodes must become "small" asymptotically. For details, see Chapter 12 of Breiman, Friedman, Olshen and Stone (1984) and the papers of Gordon and Olshen (1978, 1980, 1984).

There are new procedures which may prove to be competitive with those that have been discussed. Both projection pursuit classification and additive logistic regression seem particularly promising. But, at this writing, substantial track records are lacking and their story will wait for another day.

3.3 Statistical Theory in Clustering

3.3.1 Introduction

Classification, placing sets of objects in similar classes, is necessary for language and thought and is the foundation of statistical data collection and of probability judgements. You believe this toss of a coin gives heads with probability ½ because you classify it with other remembered coin tosses, half giving heads and half giving tails. You predict rain after thunder because you classify the thunder with other thunders followed by rain.

The statistician is pleased to inform the biologist that his fossil shellfish divide distinctly into three clusters evidenced by a trimodal distribution of the measurements of number of whorls and relative diameter of the innermost and outermost chambers of the shell. The biologist is not surprised because they looked like three different species and he made those measurements that he thought would best distinguish them. Classification precedes measurement.

Statistical theory cannot provide a complete theory of classification. We cannot say how similarities should be judged, although we can give technical assistance in constructing distances. See Section 2.3.3. Different classifications are right for different purposes, so we cannot say any one classification is best. Statistical theory in clustering provides a testing ground for various clustering methods—we discover how well the methods work for various idealized forms of data, and reject those methods that fail, at least for application to similar types of real data.

One general model is that the data form a random sample $X_1, X_2, \cdots, X_n$ from some population with a probability distribution P. A technique produces some clusters in the sample. A theoretical model generates some clusters in the population with the distribution P. We evaluate the technique by asking how well the sample clusters agree with the population clusters.

Perhaps you wish to classify the 50 United States by their agricultural products, where the population is the sample. Nevertheless, you might not want to use a clustering technique, such as complete linkage (maximum diameter) clustering that produces clusters that do not depend asymptotically on the distribution P. Frequently, you have available a sample that cannot be regarded as random. You collect all the specimens available at the site, but you wish to form a taxonomy of general utility. Some species will be represented much more highly than others; if you treat the whole as a random sample, the overrepresented species will receive too much attention. In survey sampling theory, much attention is paid to probability sampling, where the probability that each individual in the population enters the sample is known; perhaps this theory can be adapted to clustering problems. We do not usually know selection probabilities, but we might be able to progress by assuming the selection probabilities are the same within each cluster. For example, in the normal mixture model we could estimate the normal parameters for each component in the population, but not the mixing proportions, because these would be confounded with unknown selection probabilities.

3.3.2 High Density Clusters

A model suggested by clusters of stars is that a cluster corresponds to a high density region in p-dimensional space (Hartigan, 1975).

Let P be the population distribution, let $\mathbf{x}$ be a typical point in p-dimensional Euclidean space, let f be the density of P with respect to Lebesgue measure. The population clusters are the maximal connected subsets of the high density region $\{\mathbf{x} \mid f(\mathbf{x}) \geq c\}$ for each c. The family of population clusters forms a tree, in that two clusters are disjoint, or one includes the other. This model is thus suitable for examining hierarchical techniques. Taking the density of P with respect to Lebesgue measure rather than some other measure ensures that the population clusters are the same if a nonsingular linear transformation is performed on the space.

For discrete data, we might assume that P is supported by the vertices of a cube in p-dimensional space, and take f to be the density with respect to the uniform distribution on the vertices. A set of vertices is connected if any two vertices in the set may be connected by a chain of cube edges between vertices in the set. The same definition of high density clustering may then be used.

Methods of density estimation produce clusters in the sample, namely the high density clusters corresponding to the estimate. This is to be expected because density estimates at a point depend on nearby sample points, and the definition of "nearby" corresponds to the similarity assessments in clustering methods. (Probability rests on the similarity between what we know and what we are guessing!) Conversely, hierarchical clustering methods may be interpreted as estimates of density contours, although the density itself is only specified by the clustering up to a monotone transformation.

3.3.3 Agglomerative Methods for High Density Clusters

Agglomerative methods define a distance between any two possible clusters, and the clusters are constructed by beginning with n singleton clusters, one for each point, and successively joining the closest pairs of clusters to form new clusters. These methods are poor estimates of high density clusters.

Complete linkage, in which the distance between clusters is the maximum distance between points in the two clusters, is the worst of all standard methods for high density clustering. If the distribution P is carried by a compact set C on which the minimum density is positive, I conjecture that the asymptotic behavior of the complete linkage clustering depends only on C, not on P. To be more precise, fix three points $\mathbf{x}$, $\mathbf{y}$, $\mathbf{z}$ and, for a given sample size n, let the closest sample points to them be $\mathbf{x}_n$, $\mathbf{y}_n$, $\mathbf{z}_n$. Then the conjecture is that the probability that $\mathbf{x}_n$ and $\mathbf{y}_n$ are clustered together, before $\mathbf{x}_n$ and $\mathbf{z}_n$ are, depends only on C as $n \longrightarrow \infty$.

Complete linkage remains a popular method because it gives nice evenly bifurcating trees for almost all data sets—the real world, not so nice, does not show through. If P is supported by disconnected sets, then complete linkage will discover those sets, which depend only on C, the supporting set. What upsets complete linkage is the little fuzz of observations between the high density regions.

Why does complete linkage fail? After we have joined the small clusters together, all clusters have roughly the same diameter (if the maximum diameter of the clusters is d, no neighboring pair of clusters can amalgamate to a cluster of diameter less than d, so at least one of the pair must have diameter $d/2$ or larger, assuming that some pair of points in the two clusters are negligibly close). Later decisions are made entirely on the pairwise distances between clusters, which do not depend on the number of points in the clusters; thus at this stage information about the distribution of points is already lost.

Average linkage defines distance between clusters as the average distance between pairs of points in the two clusters. It is known in the numerical taxonomy literature as the unweighted pair group method. It behaves somewhat better than complete linkage in sensitivity to the population distribution because the distance measure is affected by the number of points in the clusters. If two neighboring clusters are formed that cut across a high density region, the distance between clusters will be smaller than usual because of the many close pairs of points, and so these neighboring clusters will be quickly joined identifying the high density region. See Figure 1, where the clusters (2, 3) and 4 are joined before (2, 3) and 1 so that the high density cluster (3, 4) is separated from the high density cluster 1. The weighted pair group method, in which distance between two clusters is just the average distance between component *clusters* (rather than points), should be no better than complete linkage, because after a small amount of joining the numbers of points in the various clusters becomes irrelevant.

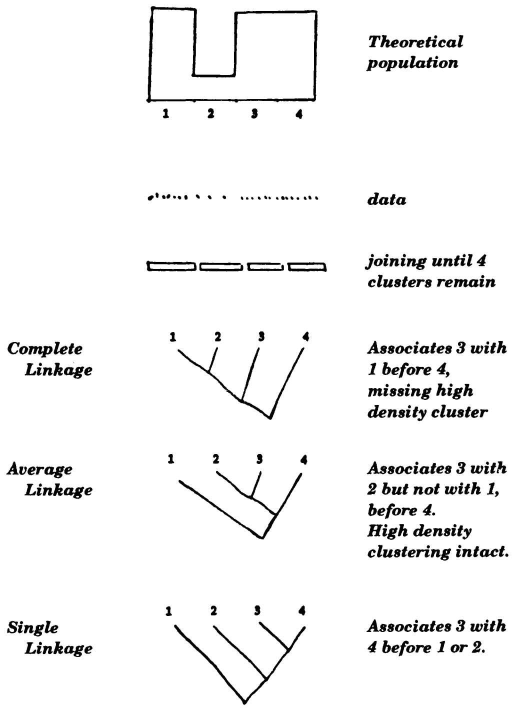

Theoretical population
1 2 3 4
data
joining until 4 clusters remain
Complete Linkage
1 2 3 4
Associates 3 with 1 before 4, missing high density cluster
Average Linkage
1 2 3 4
Associates 3 with 2 but not with 1, before 4. High density clustering intact.
Single Linkage
1 2 3 4
Associates 3 with 4 before 1 or 2.

FIG. 1. *Comparative behavior of complete, average and single linkage.*

The centroid method measures distance between clusters 1 and 2 by $n_1 n_2 \rho^2(\bar{\mathbf{x}}_1, \bar{\mathbf{x}}_2)/(n_1 + n_2)$, where n_i is the number of points and $\bar{\mathbf{x}}_i$ the mean point in the ith cluster and ρ is Euclidean distance. This ensures that the two clusters are joined to least increase the within cluster sum of squares; the method is the hierarchical analogue of the k-means algorithm. The resulting clusters are sensitive to the population distribution; the intermediate clusters (those obtained by a moderate amount of joining) are smaller in diameter in high density regions. Nevertheless these clusters are not consistent for high density clusters—it is easy to have the edge of a large cluster join with a neighboring small cluster rather than with the other parts of the large cluster.

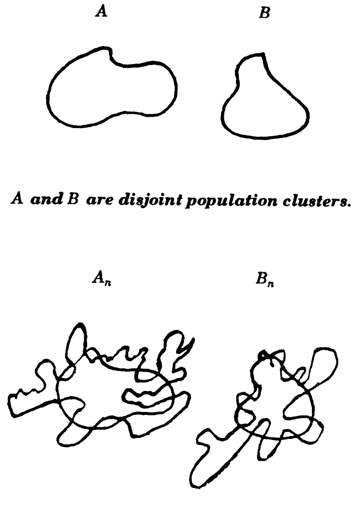
3.3.4 Single Linkage, the Minimum Spanning Tree and Percolation

Single linkage clustering measures the distance between clusters as the minimum distance between pairs of points in the two clusters. Single linkage clustering is consistent for high density clusters in one dimension in the sense that two fixed disjoint population clusters will eventually lie within some two disjoint sample clusters with probability one. Only approximate consistency holds in more than one dimension: let A and B be two disjoint population clusters, and define the distance between two sets C and D,

$$\rho(C, D) = \sup_{x \in C} \inf_{y \in D} \rho(x, y).$$

If $\rho(C, D)$ is small, every point of C has some point of D close to it, so that D approximately includes C. As $n \to \infty$, with probability one, there exist disjoint single linkage clusters A_n and B_n such that $\rho(A, A_n) \to 0$, $\rho(B, B_n) \to 0$ (Hartigan, 1981). The single linkage clusters are straggly affairs whose contours by no means approximate the population density contours, but each of the two single linkage clusters has a point near each point in the two population clusters (Figure 2).

Single linkage clusters have a number of equivalent characterizations that make single linkage attractive for theoretical study. For example, divide the points into two clusters so that the minimum distance between the two clusters is as large as possible, and continue dividing the clusters obtained in the same way. This produces single linkage clusters; the other agglomerative methods have no simple divisive



A and B are disjoint population clusters.

A_n and B_n are disjoint sample clusters approximating A and B.

FIG. 2. *Single linkage's approximate consistency.*

characterization, and so it is to be expected that the large clusters they produce have no known desirable properties.

Replace each point by a sphere of radius d and consider the maximal connected subsets of the union of spheres, for all d. These are the single linkage clusters. Clusters of this type are studied in percolation theory (Broadbent and Hammersley, 1957; Smythe and Wierman, 1978) and so asymptotic results about single linkage clusters may be obtained from percolation asymptotics.

The nearest neighbor density estimate is, in p dimensions,

$$f_n(\mathbf{x}) = C \Big/ \inf_i \rho^p(\mathbf{x}, \mathbf{x}_i).$$

The high density clusters of f_n are the maximal connected subsets of unions of spheres of the previous paragraph, the single linkage clusters. The nearest neighbor density estimate is a poor estimate, not consistent for the true density; it is remarkable that single linkage clusters retain approximate consistency. Following Wishart (1969) and Ling (1973) we should use high density clusters corresponding to some form of kth nearest neighbor density estimation, where for consistency $k \to \infty$ as $n \to \infty$ but $k/n \to 0$. For example, we define distance between clusters in a joining algorithm by the kth smallest among distances between pairs of points in the two clusters. This clustering method is the analogue of kth nearest neighbor discriminant procedures, in which a new point is allocated to the class that appears most frequently in its k nearest neighbors.

Another characterization of single linkage clusters is through the *ultrametric*

$$\rho^*(\mathbf{x}, \mathbf{y}) = \inf_{\mathbf{x}=\mathbf{x}_1,\mathbf{x}_2\cdots\mathbf{x}_k=\mathbf{y}} \sup_i \rho(\mathbf{x}_i, \mathbf{x}_{i+1}).$$

The ultrametric satisfies $\rho^*(\mathbf{x}, \mathbf{y}) \le \sup[\rho^*(\mathbf{x}, \mathbf{z}), \rho(\mathbf{y}, \mathbf{z})]$ and determines a family of clusters $\{\mathbf{x} \mid \rho^*(\mathbf{x}, \mathbf{y}) \le C\}$ for various C, $\mathbf{y}$, that turn out to be the single linkage clusters.

Finally, the minimum spanning tree is the graph of minimum total length connecting the sample points. Gower and Ross (1969) show that single linkage clusters are the connected sets obtained by successively deleting, largest to smallest, the links of the minimum spanning tree. Thus single linkage computation and asymptotics are intimately related to minimum spanning tree computation and asymptotics.

3.3.5 Mixtures

The population density

$$f = \sum_{i=1}^{k} p_i f_i$$

is a mixture of *components* f_i in proportions p_i. This may be viewed as a model for k clusters. The f_i and p_i are unidentified without some further constraints. The usual assumption is that each f_i is a member of the same parametric family, the multivariate normal (e.g., Wolfe, 1970; Day, 1969; Dick and Bowden, 1973); however, the mixture model may also be applied to general sample spaces, not only to points in p-dimensional Euclidean space. In discriminant analysis, a random observation $\mathbf{X}$ is associated with a classification I into one of k classes. Suppose that I takes the value i with probability p_i, and that $\mathbf{X}$ given $I = i$ has density f_i. Then the marginal density of $\mathbf{X}$ is just $f = \sum p_i f_i$. In discriminant analysis we know $\mathbf{X}$ and I; in clustering we know only $\mathbf{X}$; thus the mixture model for clustering corresponds to the marginal probability model for discriminant analysis. Lack of knowledge of the classification variable makes the general mixture model unidentifiable however; so further constraints are needed for clustering.

Let $p(i \mid \mathbf{x}) = p_i f_i(\mathbf{x}) / \sum p_i f_i(\mathbf{x})$ denote the posterior probability that the observation $\mathbf{x}$ belongs to class i. These posterior probabilities are useful in maximum likelihood estimation for the model

$$f(\mathbf{x}) = \sum p_i f_i(\mathbf{x}, \boldsymbol{\theta}_i)$$

where the f_i are known up to the parameter $\boldsymbol{\theta}_i$ taking values in r-dimensional Euclidean space. Assume the f_i are differentiable with respect to $\boldsymbol{\theta}_i$. Then the maximum likelihood estimates for p_i, $\boldsymbol{\theta}_i$ based on observations $\mathbf{X}_1, \mathbf{X}_2, \cdots, \mathbf{X}_n$ satisfy

$$\text{(i)} \quad \sum_{j=1}^{n} p(i \mid \mathbf{X}_j) \frac{d}{d\boldsymbol{\theta}_i} \log f_i(\mathbf{X}_j, \boldsymbol{\theta}_i) = 0,$$

$$\text{(ii)} \quad p_i = \sum_{j=1}^{n} \frac{p(i \mid \mathbf{X}_j)}{n}.$$

The estimation proceeds in alternating steps: given $p(i \mid \mathbf{X}_j)$, estimate $\boldsymbol{\theta}_i$ weighting the observations $\mathbf{X}_j$ by their probability of belonging to the ith component; then given these estimates $\boldsymbol{\theta}_j$, estimate p_i and $p(i \mid \mathbf{X}_j)$; then repeat the first step.

Rao (1948) appears to have been the first to use maximum likelihood for normal mixtures. See also Day (1969), Wolfe (1970), Hosmer (1973) and Everitt and Hand (1981). We can't show that the alternating procedure leads to the true maximum likelihood estimates, or even that the likelihood increases after each step, even in the simple case of normal mixtures in one dimension. The standard maximum likelihood regularity conditions do not hold in the normal case, and the usual asymptotic consistency and distribution results do not always hold.

For well-separated clusters, the posterior probabilities $p(i \mid \mathbf{X}_j)$ are all near 0 or 1, and the maximum likelihood solution is approximated by dividing the observations into k clusters, estimating $\boldsymbol{\theta}_i$ by maximum likelihood separately within the clusters, and finding that division into k clusters that maximizes the product of the likelihoods. This procedure is the strict maximum likelihood estimate for the model in which the observations $\{\mathbf{X}_j\}$ are supposed drawn from components $\{I_j\}$, and the components are regarded as unknown *parameters*. The likelihood is

$$L[\mathbf{X}_1, \cdots, \mathbf{X}_n, I_1, \cdots, I_n, \boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \cdots, \boldsymbol{\theta}_k] = \prod f_{I_j}(\mathbf{X}_j, \boldsymbol{\theta}_{I_j}).$$

In practice we can rarely afford to search all partitions, but we use an alternating step algorithm:

$$\text{(i)}' \quad \text{Given } I_j, \quad \text{select } \boldsymbol{\theta}_i \text{ to maximize } \prod_{I_j = i} f_i(\mathbf{X}_j, \boldsymbol{\theta}_i).$$

$$\text{(ii)}' \quad \text{Given } \boldsymbol{\theta}_i, \quad \text{select } I_j \text{ to maximize } f_{I_j}(\mathbf{X}_j, \boldsymbol{\theta}_{I_j}).$$

Thus given the clusters we estimate $\boldsymbol{\theta}_i$ by maximum likelihood, and given the $\boldsymbol{\theta}_i$ we specify cluster membership to make $\mathbf{X}_j$ most likely. This is the alternating method for mixture maximum likelihood when the $p(i \mid \mathbf{X}_j)$ are all zero or one.

In the particular case when

$$f_i(\mathbf{X}) = (2\pi)^{-p/2} \exp[-\tfrac{1}{2}(\mathbf{X} - \boldsymbol{\mu}_i)'(\mathbf{X} - \boldsymbol{\mu}_i)],$$

the above algorithm has a simplified form known in the clustering literature as k-means. (See, e.g., MacQueen, 1967; Hartigan, 1975). We select $\boldsymbol{\mu}_i$ to be the mean of the observations in the ith cluster, and we allocate the observation $\mathbf{X}_j$ to that cluster i that minimizes the distance between $\mathbf{X}_j$ and $\boldsymbol{\mu}_i$.

3.3.6 The Number of Clusters: Modes

In the high density clustering model, we associate a family of clusters with each mode of the density f: f has a mode at $\mathbf{m}$ if there is a neighborhood $\mathbf{M}$ of $\mathbf{m}$ such that $f(\mathbf{x}) \le f(\mathbf{m})$ for $\mathbf{x} \in \mathbf{M}$, and $f(\mathbf{x}) < f(\mathbf{m})$ for $\mathbf{x}$ in the boundary of $\mathbf{M}$. There are disjoint high density clusters only if f is multimodal. Thus we can test for the presence of clusters by testing for multimodality.

For X one-dimensional, unimodal and bimodal densities may be fit by maximum likelihood giving a likelihood ratio test for bimodality, but it is difficult to handle the large contributions to the likelihood made by small intervals between neighboring observations. A better test is the *dip* test, which measures the maximum difference between the empirical distribution function, and the unimodal distribution function chosen to minimize that maximum difference. The dip approaches zero for unimodal distributions, and some non-zero value for multimodal distributions, as the sample size increases. It is therefore consistent for distinguishing unimodal from multimodal distributions. It is argued in Hartigan and Hartigan (1985) that the uniform is the appropriate null unimodal distribution, because the dip is asymptotically stochastically larger for the uniform than for other unimodal distributions; the asymptotic distribution of the dip and some empirically determined distributions for finite sample sizes are given in that paper.

The dip does not generalize simply to many dimensions. The minimum spanning tree provides a kind of ordering of the n sample points that may be used to generate an analogue of the dip statistic: select a particular sample point $\mathbf{x}_0$ to be the mode or root, and consider probability distributions P supported by the links of the minimum spanning tree. Define P to be unimodal if P has a density, with respect to the uniform distribution on the tree, that is a nondecreasing function of $\mathbf{x}$ as $\mathbf{x}$ moves toward the root. At each point $\mathbf{x}$ on the tree define a distribution function value $F(\mathbf{x})$ to be the probability that a random point $\mathbf{X}$ is such that $\mathbf{x}$ lies between $\mathbf{X}$ and $\mathbf{x}_0$. Let F_n be the empirical distribution function corresponding to the empirical distribution which gives each sample point probability $1/n$. Define

$$d(F, F_n) = \sup_{\mathbf{x}} | F_n(\mathbf{x}) - F(\mathbf{x}) |,$$

$$D(F_n, \mathbf{x}_0) = \inf_{F} d(F, F_n),$$

where F is unimodal with mode $\mathbf{x}_0$,

$$DIP(F_n) = \inf_{\mathbf{x}_0} D(F_n, \mathbf{x}_0).$$

This procedure locates an optimal mode $\mathbf{x}_0$ and states how well the data fit the unimodal hypothesis, $DIP(F_n)$. In the one-dimensional case the usual definition of dip gives the same value. The asymptotic behavior of the multivariate version is unknown.

3.3.7 The Number of Clusters: Components

If the components of a multivariate normal mixture are sufficiently well-separated, there will be one mode for each component. In this case the number of clusters is the number of components or the number of modes, but in general the number of modes is fewer than the number of components, so testing for the presence of more than one component is less conservative than testing for the presence of more than one mode.

Wolfe (1971) considers the likelihood ratio test for say one component against two, but notes that the regularity conditions which are usually required for the log likelihood ratio to be proportional to a chi-square are no fulfilled. See also Binder (1978) and Hartigan (1977).

Consider the simplest case: $\mathbf{X}_1, \cdots, \mathbf{X}_n$ sample from $N(0, 1)$ under the null hypothesis, and from $(1 - p)N(0, 1) + pN(\mu, 1)$ for some $0 \le p \le 1$, $-\infty < \mu < \infty$ under the alternative hypothesis. Let $Z_i = \exp(\mathbf{X}_i\mu - \frac{1}{2}\mu^2) - 1$. Then the likelihood is proportional to $L(p, \mu) = \prod_{i=1}^{n} (1 + pZ_i)$. Note that Z_i has mean 0 and variance $e^{\mu^2} - 1$; if $Z_i(\mu)$ and $Z_i(\mu')$ denote the Z-values computed for μ and μ', then $\operatorname{cov}(Z_i(\mu), Z_i(\mu')) = e^{\mu\mu'} - 1$.

For each fixed μ, $\log L(p, \mu)$ is a concave function of p that has maximum value 0 if $\sum Z_i < 0$ but maximum value approximately $(\sum Z_i)^2/2(\sum Z_i^2)$ otherwise. Thus asymptotically, $L(\mu) = \sup_p \log L(p, \mu)$ is equal to zero with probability ½, and to $(\chi_1^2)/2$ with probability ½. The $(\chi_1^2)/2$ would be expected from usual likelihood asymptotics.

If μ and μ' are widely separated, $Z_i(\mu)$ and $Z_i(\mu')$ are nearly uncorrelated, and so asymptotically $L(\mu)$ and $L(\mu')$ are nearly independent. Thus $\sup_\mu L(\mu)$ is greater than the maximum of k nearly independent $L(\mu)$ for each k. Thus $\sup_\mu L(\mu)$ is asymptotically infinite.

The likelihood ratio test does not therefore follow the usual asymptotics, and is not conservative: the usual significance test will (with probability 1 asymptotically) reject the hypothesis of a single component when only a single component is present. For each μ, $\sup_p L(p, \mu)$ has asymptotically the same distribution, and these distributions are nearly independent for well-separated μ; maximum likelihood computations will therefore be difficult; we can expect to see local maxima of $\sup_p L(p, \mu)$ near every value of μ.

If μ has prior density normal with mean 0 and variance 1, large values of μ are inhibited and the maximum posterior density will occur only with μ moderate. The corresponding ratio test may have better asymptotic behavior than the likelihood ratio test. More generally, if the mixture model has components with means $\mu_1, \mu_2, \cdots, \mu_k$ we might assume the μ_k to be *a priori* a sample from a normal; this prevents the artificially large separation of μ's that occurs in likelihood estimation and testing.

The behavior of the likelihood ratio statistic in the k-means case has been examined in one dimension by Hartigan (1978) and in higher dimensions by Pollard (1982).

3.3.8 Ultrametric and Evolutionary Distances

Assume that there are N objects, and $N(N - 1)/2$ distances between pairs of objects. From these distances we wish to form clusters of close objects. One way to go about constructing the clusters is to require that the distances satisfy certain properties in the final clustering. For example, all distances within two disjoint clusters must be smaller than all distances between the clusters. Or, each pair of points in the same cluster must be connected by a chain of points such that neighboring points in the chain are closer than some neighboring points in a chain connecting points in different clusters. (This definition leads to single linkage.) Another way is to suppose that the clusters correspond to some ideal distance matrix, and to attempt to approximate the given distance matrix d with a best fitting cluster distance D. For example, hierarchical clustering might correspond to an *ultrametric* D, a distance satisfying $D(i, j) \le \sup[D(i, k), D(j, k)]$ and we would find the ultrametric D closest to d. See Hartigan (1967), Johnson (1967) and Jardine, Jardine and Sibson (1967). Another plausible definition, the *evolutionary* model from Fitch and Margoliash (1967) is based on an evolutionary tree generating the objects. The distance between any pair of objects is the sum of links on the unique path connecting them in the tree. If there exists an ancestor in the tree such that all points are equidistant (in sums of links) from the ancestor, then this evolutionary distance reduces to an ultrametric. Given the tree, the best fitting evolutionary distance or ultrametric can be fitted by regression methods; the hard part is searching for the best tree.

Baker (1974) has considered probability models in which an observed distance matrix d varies by some amount from an ultrametric D, and has investigated empirically how well the various hierarchical techniques recover the true ultrametric D. The results are opposite to those obtained using the high density model: complete linkage does well and single linkage poorly.

Euclidean distances in p-dimensional space will form an ultrametric distance matrix on at most $p + 1$ points. For a density f, we can construct an ultrametric by

$$D(\mathbf{x}, \mathbf{y}) = \min_{C} \max_{\mathbf{u} \in C} 1/f(\mathbf{u})$$

where C is any path connecting $\mathbf{x}$ and $\mathbf{y}$. Thus $\mathbf{x}$ and $\mathbf{y}$ are close if they can be connected by a path of high density or equivalently if they lie together in a high density cluster. In fitting such an ultrametric to objects in p-dimensional space we would use only the small distances between objects to obtain an estimate of the density f. Single linkage works only with the small distances, whereas complete linkage depends on the large distances. This may be the explanation for Baker's results favoring complete linkage, in that he requires the fitted ultrametric to be close to the true ultrametric when averaged over *all* distances, and the large distances are neglected by single linkage. In practice, the large distances deviate most from the fitted ultrametric (however fitted) and it seems correct to downweight their contribution. Theoretically, if we wish to allow clusters of arbitrary shape and size, it also seems impossible to give large distances much weight. Perhaps we should fit an ultrametric D to minimize

$$\sum w(D)[d(i, j) - D(i, j)]^2 / \sum w(D)$$

where $w(D)$ is small or zero for D large. This moves single linkage a little way toward average linkage. More weight should be given to the large distances in high dimensional spaces.

Let the objects $1, \cdots, n$ be generated by an evolutionary tree, beginning at some ancestor, O. For a particular measurement X taking values X_i on the objects, assume that X changes in time t, $t + \Delta t$ on a particular link of the tree, by an amount that has mean 0, variance $\sigma^2 \Delta t$, and is uncorrelated with changes in different intervals or links.

Then, letting EY denote the expected value of the random variable Y,

$$E(X_i - X_j)^2 = 2\sigma^2 t_{ij}$$

where t_{ij} is the time since i and j evolved from their most recent ancestor, so $E(X_i - X_j)^2$ is an ultrametric!

If we had used different rates of evolution in the different links of the tree, so that the changes in X had variance $\sigma_i^2 \Delta t$ for link i, then $E(X_i - X_j)^2$ would be an evolutionary distance.

Suppose that X is normal, and there are p independent samples of X, namely, $X^1, X^2, \cdots, X^p$. (Here the number of objects is fixed, and the measurements are assumed sampled from an infinite population of possible measurements; it will require careful standardization to achieve something like this in practice.) Then

$$\sum_{r=1}^{p} (X_i^r - X_j^r)^2 \sim 2\sigma^2 t_{ij} \chi_p^2,$$

$$d(i, j) \sim \sqrt{2\sigma^2 t_{ij} \chi_p^2}.$$

Let $D(i, j) = \sqrt{2\sigma^2 t_{ij} p}$, an ultrametric. For large p,

$$d(i, j) \sim D(i, j)[1 + N(0, 2/p)].$$

This suggests fitting the ultrametric D by minimizing

$$\sum [D(i, j) - d(i, j)]^2 / D^2(i, j)$$

which downweights the large distances nicely, but probably not enough. We have to take note also of the high correlation between the large distances, arising from the high fraction of their paths through the tree that they have in common. We can compute these, but the criterion to be minimized is then a complex quadratic in $D - d$.

4. SOFTWARE AND ALGORITHM IMPLEMENTATION

4.1 Introduction

This chapter provides a summary of available software and algorithms for discriminant and cluster analyses. Although it is intended to be up to date, the current pace of statistical software evolution is such that some of the more recent developments may be inadvertently excluded. For instance, the new S system (Becker, Chambers and Wilks, 1988) has facilities for linear discriminant analysis and a variety of hierarchical clustering methods, but is not discussed here.

The strengths and shortcomings of programs and packages are described. Section 4.2 focuses on discriminant analysis and Section 4.3 on cluster analysis. The final section, 4.4, considers software needs.

4.2 Discriminant Analysis

4.2.1 Linear and Quadratic Discriminant Functions

Many packages are available for performing linear discriminant analyses. Fewer are available for quadratic discriminant analyses and only one (to our knowledge) is available for performing density estimate discriminant analysis. These are reviewed in the section on packages. We have not attempted to cover programs which are not widely available in the United States. BMDP, GLIM, IMSL, Minitab, P-STAT, SAS, SPSS-X and Statgraphics are now available in versions for MS-DOS compatible microcomputers. In addition, several statistical systems developed specifically for microcomputers have appeared on the market: SYSTAT, CRISP, GAUSS and STATA are examples.

The linear discriminant function is well implemented for most applications. The numerical techniques are standard, the equations and algorithms (inversion, solution of a set of linear equations) have been tested thoroughly and accuracy is not a major concern. Estimating errors in discriminant analysis is generally done by reclassifying the training sample. If the sample sizes are sufficiently large (say 3 to 5 times the number of variables in each group), this method is satisfactory and has an approximately binomial distribution. One package offers the jackknife (or leaving-one-out) method. This has a smaller bias than the resubstitution method, but because of the correlation among the pseudo-observations has a larger variance. This method should only be used for small samples when the danger of the optimistic bias of the resubstitution method is substantial. Plots of the linear discriminant variables are available in most packages. Weighting of cases is possible in SAS and SPSS, and it is not clear from the manuals whether it is possible in BMDP and P-STAT. None of the packages offer the option of proportional covariance matrices. This intermediate step between the full quadratic function and the linear function involves estimation of only one additional parameter for each additional group, rather than the full covariance matrix for the added group, and may be a satisfactory compromise in many cases.

4.2.2 Review of Packages

P-STAT. The P-STAT discriminant analysis procedure is similar to BMDP7M program. It is a backward stepwise procedure and allows from 10 to 40 groups depending on the P-STAT size. No warning on the sample size requirements for the many groups case is given. This may lead some naive users astray. Also, the assumption of common covariance is not discussed. The resubstitution method is available for estimating error rates. There are three types of runs:

1. Analyze and classify a data set.
2. Classify a known data set by using previously generated functions. This can be used as a validation method by holding out a fraction of the data.
3. Classify an unknown data set.

Output data sets contain the original group, the assigned group, the posterior probability the observation belongs to its original group and the posterior probability it belongs to its highest probability group.

The program does not automatically step in the batch mode but is stepwise when run interactively.

SPSS-X. The DISCRIMINANT procedure in SPSS-X allows one to use a fixed set of variables or to select variables in a forward manner. Removal of variables is possible, but backward stepping does not appear to be possible. Five criteria are possible to select variables. Inclusion levels are available to force variables into the discriminant function.

For the multiple groups problem, the canonical discriminant functions are computed rather than the likelihood ratio functions which minimize the total (weighted) error rate.

Cases with missing values are excluded. It is possible to select a subset of cases to analyze and then test the performance of the rule on the remainder of the cases.

Plots may be obtained which map the two-dimensional space of canonical functions and show the classification boundary, an all-groups plot which plots each case or a separate-groups plot. A variety of matrix operations are possible on the discriminant coefficients.

BMDP7M. The BMDP7M stepwise discriminant analysis program is the descendant of the oldest discriminant analysis packaged program. It offers forward and backward stepping, forcing levels for inclusion or exclusion of variables, and two criteria for variable entry. It is possible to specify prior probabilities. For estimating error rates, the resubstitution method and the jackknife method are available. Plots of canonical variables are given either by group or for any subset of groups. The error rates may be printed at any set of steps in the variable selection process. The size of the problem is a function of the number of variables, groups and cases. It is not clear if there is an upper limit on groups or variables. Quadratic discrimination does not appear to be available.

SAS Procedures. SAS offers four discriminant procedures: DISCRIM, NEIGHBOR, CANDISC and STEPDISC. CANDISC performs a canonical discriminant analysis and provides output for other SAS procedures for plotting or printing. A number of statistics are available. The DISCRIM procedure computes a linear or a quadratic discriminant function on a fixed set of variables. Prior probabilities may be specified. Classification may be done on the training sample or on a test sample. Stratified analyses may be performed by using a BY statement. The NEIGHBOR procedure performs a nearest neighbor discriminant analysis. Either the single nearest neighbor or the *k*-nearest neighbor rule may be used. Prior and posterior probabilities are printed and an error matrix is given. The STEPDISC procedure performs a stepwise discriminant analysis. It is s.milar to the BMDP7M program. The Wilks' lambda criterion is used to determine which variable enters or is removed.

Other Packages. MINITAB (Ryan, Joiner and Ryan, 1982) has no discriminant analysis procedure, although it is possible to use a linear regression program to obtain the discriminant coefficients. After using the regression procedure, one could calculate the resubstitution estimator of error rates by the MULTiply and ADD commands. Other packages would be preferred for discriminant analysis.

IMSL has two subroutines for linear discriminant analysis, ODFISH and ODNORM. In ODFISH, the canonical discriminant functions are calculated. In ODNORM, the multivariate normal discriminant functions are computed. These subroutines do not print output; this becomes the user's responsibility. There are also routines which will estimate a density function using the kernel method. The user must supply a kernel function. The subroutine computes density estimates at a set of points requested by the user. Printing is the user's responsibility. These routines are NDKER and NMPLE which estimate the density for a one-dimensional problem.

ALLOC (Hermans, Habbema and Schaefer, 1982) is a program which computes allocation rules based on density estimation. It uses multivariate normal kernels with a diagonal covariance matrix. The smoothing parameter is estimated by the program. A subsequent modification allows the program to use variable kernels to obtain better estimates of densities.

4.2.3 Logistic Regression

The major statistical packages all offer some form of logistic regression analysis. Additionally, there are a number of other programs available to perform these computations. The method was originally suggested by Cornfield (1962) in connection with the Framingham studies. Walker and Duncan (1967) suggested a weighted least squares method of estimating the parameters which has been widely used. Day and Kerridge (1967) discussed several properties of the method. Nelder and Wedderburn (1972) derived the theory of generalized linear models which has been the basis for additional important work. The program, GLIM (Generalized Linear Interactive Modeling), is an outgrowth of this work and is easily used for fitting these models which include the logistic regression model.

BMDP offers a logistic regression program based on a method developed by Jennrich and Moore (1975). This is a stepwise program and uses iteratively reweighted least squares. Conditional logistic regression is possible for matched pairs analyses.

SAS includes a procedure, LOGIST, in their supplementary programs that performs logistic regression by computing maximum likelihood estimates of the parameters. Stepwise variable selection is possible.

SPSS does not have a separate logistic regression procedure. One can get estimates of the regression coefficients if the observations can be analyzed using a categorical analysis program. Thus, continuous variables cannot be handled by SPSS.

GLIM was developed by the Numerical Algorithms Group (NAG), in conjunction with the Royal Statistical Society, to estimate parameters from the Nelder-Wedderburn models. Special cases of this model include logistic regression, log-linear categorical models, analysis of variance and multiple regression. This program fits these models using maximum likelihood. A new program, PRISM, has recently been issued by NAG which includes all facilities of GLIM.

A general criticism of these packages is that they offer little in the way of diagnostic computations for the detection of influential observations. Work by Pregibon (1981) is now available and new revisions of these programs should include these results.

4.2.4 Classification Trees

Recent work on classification trees was summarized briefly in Section 2.2.7. Batch and interactive versions of the CART methodology are available through California Statistical Software, Lafayette, California.

4.3 Cluster Analysis

The amount and diversity of cluster analysis software has been surprisingly large for a statistical method with effectively only a twenty year history. New methods are produced continually, and there appears to be no end in sight to the process of innovation. Probably hundreds of software packages and programs are available to perform cluster analysis, and it is likely that many researchers have written their own "home-grown" versions of popular algorithms. That so much clustering software has been written can be explained by two factors: (1) unlike many statistical procedures, clustering algorithms, which are often no more than heuristics, are relatively easy to program on a computer; and (2) because most sciences have different goals, analytical needs and methodological requirements, many different clustering methods have been developed to exploit these needs.

Clustering software can be placed into four major categories: (1) collections of subroutines and algorithms, (2) general statistical packages which contain clustering methods, (3) cluster analysis packages and (4) simple programs which perform one type of clustering (Blashfield, Aldenderfer and Morey, 1982). Because a comprehensive review of clustering software is beyond the scope of this report, the focus shall be only upon those programs and packages which are widely available.

4.3.1 Collections of Subroutines and Algorithms

Three major collections of software are available today in this category; books by Anderberg (1973), Hartigan (1975) and Jambu and Lebeaux (1983) plus the International Mathematical and Statistical Library (1977). Although much of this software is fairly sophisticated, it requires the user to supply all job control language of the computing system to link and subsequently run the routines. As a result, these programs are not very "user-friendly." The user must be familiar with the local control language as well as FORTRAN in order to be able to get the programs running. In general the level of user support for these routines is low. Hartigan's algorithms are described in a separate user's manual (Dallal, 1975), whereas Anderberg's algorithms are only documented in his book. Although the IMSL clustering algorithms are embedded within the documentation of the entire collection of IMSL subroutines, this does not necessarily make them any easier to use. The FORTRAN programs in the Jambu and Lebeaux book (1983) are quite extensive and represent a considerable effort by these two French writers. Like the routines in Hartigan (1975), the algorithms in Jambu and Lebeaux are unique. Despite the breadth of methods available, algorithms in this category are not recommended for use by the novice unless extensive guidance is available.

Statistical Packages Containing Clustering Software. The most convenient cluster analysis available for general use is that contained within popular packages of statistical programs such as BMDP (Dixon, 1981), SAS (SAS Institute, 1985) and SPSS-X (SPSS, 1986). The philosophy of these packages is well-known; they provide nonprogrammers with relatively easy access to sophisticated statistical methods for a wide variety of research problems. The packages provide an "umbrella" of support for the user in that they use a consistent control language that communicates the needs of the user to the computing system with a minimum of effort. These packages also contain a full range of data screening and manipulation methods which help to make complex analyses simple and feasible. If the package contains the method of interest to the user, the advantages of using existing statistical packages are substantial.

Until recently, the range of clustering options contained in most statistical packages has been severely limited. For instance, before 1980, SAS contained only one clustering method and SPSS had no clustering methods. However, this state of affairs has changed dramatically. Since 1979, BMDP has developed four procedures devoted to cluster analysis: (1) a collection of single, complete and average linkage to cluster *variables*; (2) an average linkage (centroid sorting) method to cluster *cases*; (3) a block clustering method (Hartigan, 1975) to simultaneously cluster cases and variables; and (4) an iterative *k*-means method which forms partitions among the cases. The BMDP procedures are well-annotated, have clear output and are relatively easy to use. The most serious limitations of this package are the limited range of hierarchical agglomerative methods, especially for clustering cases, and the choice of only a single similarity measure, Euclidean distance.

Earlier versions of the second statistical package, SAS, contained one method of cluster analysis—complete linkage. However, a recent release of this package (SAS, 1985) includes substantial additions. This version of SAS contains Ward's single linkage, complete linkage and average linkage plus seven other hierarchical agglomerative methods. Euclidean distance is still the only similarity measure offered. In procedure, FASTCLUS, a *k*-means method (Anderberg's centroid sorting method) has been added, and finally, a factor analysis-type variable clustering method has been included (procedure VARCLUS). The diagnostics of the package has been expanded. In addition, the output provides a great deal of information about the solutions. A major limitation, however, is that SAS continues to use "sky line" plots to represent hierarchical trees; these plots are difficult to use with large data sets. Of considerable interest in SAS is the inclusion of a new statistic, *cubic clustering criterion*, for the determination of the number of clusters.

The 1986 version of SPSS-X contains two major clustering procedures: CLUSTER and QUICK CLUSTER. The former emphasizes hierarchical agglomerative methods including seven of the most commonly used techniques (single linkage, complete linkage, average linkage, Ward's method, etc.). There are six distance measures and three types of plots available. The second procedure, QUICK CLUSTER, uses a *k*-means method with limited options for starting partitions. Interesting aspects of this procedure are provisions for missing data and the ability to handle extremely large data sets.

4.3.2 Cluster Analysis Packages

For the sophisticated and serious user of cluster analysis, cluster analysis packages represent the ultimate in flexibility and user convenience. These packages combine the advantages of general statistical packages, such as an integrated control language and data screening and manipulation procedures, with features of interest to users of cluster analysis, such as a diversity of clustering methods, special diagnostic features and enhanced graphics. Of the greatest importance is that many of the packages contain hard to find clustering methods or analytical procedures which are appropriate for special problems.

The first of these packages is NT-SYS which is and has been important because it adopts the terminology and methodology inherent in the most frequently cited book on cluster analysis, Sneath and Sokal (1973).

This package has undergone numerous revisions and updates in its 15-year existence. Moreover, there now exists a microcomputer version, called NTSYS-pc, which contains the standard hierarchical agglomerative methods, graph theoretic methods and an eigenvector routine. This version can handle similarity matrices up to 400 × 400 plus it contains three graphics programs.

The most versatile of the clustering packages is CLUSTAN. Like BMDP, SAS and SPSS-X, CLUSTAN contains procedures for hierarchical agglomerative and iterative partitioning methods. However, CLUSTAN also contains a number of other procedures including NORMIX to decompose multivariate normal mixtures (Wolfe, 1971); INVARIANT, which uses partitioning methods to optimize MANOVA statistics; DENDRITE, which is a minimum spanning tree method, plus others. In addition, CLUSTAN has cluster diagnostic and validation aids, including the procedures called RULES and COMPARE, which implement the stopping rules of Mojena (1977) and the cophenetic correlation coefficient of Mojena and Wishart (1980). A total of 38 similarity measures are contained in procedure CORREL, and the package contains a utility procedure which permits the user to define any type of similarity coefficient (DEFINE). Other important utilities are those which prepare a number of cluster diagnostics or which produce a wide variety of graphic output. The novice user of CLUSTAN should be aware that although this package contains a large number of methods, the package contains little guidance on which methods may be most appropriate for what types of data sets.

There are three other packages which are devoted to cluster analysis. CLUS (Friedman and Rubin, 1967) is an old program which used a powerful set of iterative partitioning methods. A more modern version of CLUS is the procedure INVARIANT in CLUSTAN. Another large package is BC-TRY. Like CLUS, this program was written in the 1960s and contained the innovative ideas of Tryon who was one of the earliest writers about cluster analysis (Tryon, 1939). Currently this program is being revised for redistribution. Finally, a recent clustering package for use on microcomputers has been developed called MICRO-CLUSTER (Edmonston, 1985). This package contains seven hierarchical agglomerative methods and an iterative partitioning method.

4.3.3 Simple Cluster Analysis Programs

Simple cluster analysis programs are just that: simple. These are programs written primarily in FORTRAN, and they implement one or two cluster analysis methods. In some ways, they strongly resemble the subroutine of the first category defined above, in that they require the user to be fully competent in the job control language of the computing system as well as the language in which the program is written. In general, these programs have no or few aids for checking programming errors, are poorly documented and provide limited output information. These programs are important, however, because they have often been used within particular scientific areas, or they have been used for the basis for the algorithms presented in major packages such as SAS, IMSL and OSIRIS. Some of the more popular of these simple programs are HGROUP, a method which implements Ward's minimum variance method (Veldman, 1967) JCLUST, which implements single and complete linkage as discussed in the influential article by Johnson (1967); and ISODATA, a flexible iterative partitioning method which has been used extensively in engineering (Hall and Khanna, 1977).

Another category of cluster analysis programs consists of those that handle large data sets (N is greater than 500). Unfortunately most clustering routines in statistical packages are somewhat limited in the number of cases which can be analyzed at one time. Typically, most have a practical upper limit of 200 cases. In response to this problem, a number of authors have extended the capabilities of popular hierarchical agglomerative and iterative partitioning methods to deal with very large data sets. Among the most important of these is Sibson's (1973) single linkage algorithm (SLINK). Note: SLINK is now incorporated within CLUSTAN 2.1, CLUSTER (Levisohn and Funk, 1974) and QUICLUSTER (Bell and Korey, 1975) which implement Ward's methods and programs by Defays (1977) and Rohlf (1977) for complete linkage clustering. Rohlf (1982) presented a number of different algorithms for single linkage that could be useful for large data sets. Lennington and Rossbach (1978) have developed CLASSY, an iterative partitioning method based upon the logic of ISODATA, for use with the very large data sets obtained in LANDSAT satellite research.

4.4 Needs

For parametric (multivariate normal) discriminant problems, relatively little is needed. A variety of programs are available which offer flexibility of use, adequate error rate estimation and many variable selection options. A general shortcoming is advice on usage. For many users, the only place they will learn about discriminant analysis is in the user's manual of a computer package. Some discussion of the limitations (e.g., if you have many groups, you need many observations) and robustness (e.g., transform your data if a variable is badly skewed) is needed. Some of the programs seem to have the attitude, if it can be programmed, include it. For example, in one program up to 40 groups can be included in discriminant analysis. A user with that many groups has probably not thought sufficiently about the problem. (Nevertheless, there are some problems, such as speaker recognition, where the only interesting situation is having many groups.) The quadratic discriminant function, available only in SAS, has some serious robustness problems. These should be noted.

The ability to enter previous coefficients or a set of means and covariances is useful. It is valuable to enter a simplified set of coefficients (say integers) and compare the performance of the rule to the optimal rule.

Other than the IMSL routines for unidimensional problems and ALLOC, no package has any programs for density estimation. This is a useful procedure, especially when distributions are rather far from normal. Nearest neighbor procedures, which are related to nonparametric density estimates, are available in SAS in the neighbor procedure.

Concerning cluster analysis programs, the inclusion of k-means and hierarchical agglomerative methods in the SPSS and SAS packages have helped standardize the clustering methods that are used in applied research. The SAS manual is particularly helpful concerning the use of these techniques because it provides a skeptical perspective and references some of the best articles in the field. Nonetheless, none of the packages is successful in providing sufficient cautions and indications of the practical problems that are of serious concern to new users of these procedures (e.g., guidelines on the choice of methods, the number of clusters problem, the issue of outliers and the choice of the similarity measure).

The preceding discussion has focused largely on mainframe and minicomputers because most users of these procedures have access to such computing equipment at the present time. Several of the programs available on microcomputers offer discriminant analysis routines. BMDP, SAS, P-STAT and SPSS-PC offer discriminant analysis through the usual programs. SYSTAT provides a discriminant analysis facility by using the module MGLH. Other programs offer regression capabilities which give an equivalent analysis, although not tailored exactly to the purposes of allocation.

Development of graphic methods for allocation and their computer implementation is needed. The mainframe packages usually offer plots of the sample discriminant variables which often is adequate to determine differences among groups. However, these variables are linear combinations of the observed variables and are not always easily interpretable. A "simple" exploratory graphical program would be welcome. Such a program would be interactive, with very good graphics (i.e., much better than the usual transcription of a page of text to graphic symbols). Graphic clustering procedures have also been neglected in generally available programs.

There are no interactive programs for allocation or clustering that are generally available. Such a program would allow the user to specify the variables to be included in the allocation rule, to specify the form of the rule (e.g., linear, quadratic, tree structure for discrimination), to enter new variables or delete old ones and to detect influential observations.

Regression diagnostics have become increasingly important in statistical practice, but little in the way of diagnostics is available for allocation rules. In a sense, the regression diagnostics suffice for classical linear discriminant theory, and Pregibon's work has large application in logistic regression. See also Landwehr, Pregibon and Shoemaker (1984) and Fowlkes (1987). Diagnostic procedures are generally lacking in cluster analysis. However, this lack is primarily due to the problems in developing an adequate statistical theory for clustering rather than reflecting a programming deficiency. Nevertheless, a few procedures have been developed and appear to be useful (see Section 2.3.3).

5. CLOSING PERSPECTIVE

Discriminant analysis and cluster analysis must be classified as among the most useful statistical techniques for society's problems (Gnanadesikan and Kettenring, 1984). This report has attempted to summarize and assess their status in terms of methodology, theory and software. The material, it is hoped, will be informative both to users of these techniques, who may want an update on the state-of-the-art, and to professional statisticians, who may be more interested in current research and what remains to be done.

Even casual readers of this report will have noticed tremendous differences between the conditions of discriminant analysis and cluster analysis. The former is a well-developed subject with a variety of effective methods and supporting theory. The latter is lacking in firm foundations and agreed upon methodology. Perhaps it is only along the software coordinate that there is approximate parity. Indeed, there may be more software for cluster analysis simply because of the proliferation of ad hoc methods for this purpose.

In spite of its relatively advanced state, there are still many interesting problems to be worked on in discriminant analysis. Specific mention was made in Sections 2 and 3 of the promising new areas of projection pursuit classification and additive logistic regression analysis; the need for more study of biases associated with the use of the bootstrap in estimating error rates and of the trade-offs between bias and mean square error performance of different estimates of error rates; and, generally, the opportunities for additional development and experimentation with the variety of nonparametric and semiparametric methods that are now available.

However, the greatest needs appear to be in the cluster analysis arena. Unless significant breakthroughs in theory and insights into the behavior of procedures are produced, cluster analysis is likely to remain a largely descriptive technique whose results are too dependent upon the vagaries of particular methods. A list of research problems, some of which were discussed in Sections 2 and 3, includes:

- Developing a stronger base of inferential and diagnostic tools (a high priority should be placed on the development of sample reuse techniques that will work in the clustering context).
- Closing the gap between theory and practice (the need is illustrated by the attractiveness of single linkage clustering from a theoretical point of view in spite of its frequent poor performance in practice).
- Compensating for the lack of adequate theory with empirical development of new insights about existing algorithms (clever and extensive simulation studies may be the only way around the mathematical and theoretical difficulties in this field).
- Finding tools for selection, scaling and transformation of variables that are effective at bringing out cluster structure (iterative schemes may be required because the clusters are unknown in advance).
- Learning how to make clustering algorithms robust to data idiosyncrasies (the payoff may prove to be in the "local" application of the robustness concept rather than a crude global attack that is insensitve to fine cluster structure).

Another area that is ripe for more research concerns problems that fall between discriminant analysis and cluster analysis. A fair amount of work has been done near the discriminant end of the spectrum, e.g., dealing with the situation where errors are present in the group labels of the training sample. Little is known about how to do cluster analysis in the presence of limited prior information on the composition of clusters.

ACKNOWLEDGMENT

Section 2.2 is an amalgam of written material supplied by O. J. Dunn and P. A. Lachenbruch, edited and modified by R. Gnanadesikan and J. R. Kettenring. Section 2.3 was written by Gnanadesikan and Kettenring. Section 3.2 was written by R. A. Olshen, and Section 3.3 by J. A. Hartigan. Section 4.2 was written by Lachenbruch and Section 4.3 by R. A. Blashfield. Section 4.4 combines materials written by both Lachenbruch and Blashfield.

A slightly different version of Section 3.3 has appeared in J. A. Hartigan, "Statistical theory in clustering," *J. Classification* **2** (1985) 63–76, Springer-Verlag.

We are grateful for the helpful comments that were received from current and past members of the Committee on Applied and Theoretical Statistics (CATS), and particularly indebted to R. A. Bradley and R. Pyke, former chairmen of CATS. It is a pleasure to thank Nancy K. Davidson and Hester A. Glynn for their major help in the word processing effort for this report. Thanks are also due to Bellcore for their facilities and support in this process.

Support for this project was provided by the Office of Naval Research under Contract N00014-82-0652 and by core funds from the Board on Mathematical Sciences, National Research Council, provided by the Air Force Office of Scientific Research, the Army Research Office, the Department of Energy, the National Science Foundation, the National Security Agency and the Office of Naval Research.

This report is reprinted from *Discriminant Analysis and Clustering*, 1988, with the permission of the National Academy Press, Washington, D. C.

REFERENCES

ANDERBERG, M. R. (1973). *Cluster Analysis for Applications.* Academic, New York.

ANDERSON, E. (1957). A semi-graphical method for the analysis of complex problems. *Proc. Nat. Acad. Sci. U.S.A.* **43** 923–927.

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

ANDERSON, T. W. and BAHADUR, R. R. (1962). Classification into two multivariate normal distributions with different covariance matrices. *Ann. Math. Statist.* **33** 420–431.

ANDREWS, D. F. (1972). Plots of high-dimensional data. *Biometrics* **28** 125–136.

ARABIE, P. (1977). Clustering representations of group overlap. *J. Math. Sociol. Japan* **5** 112–128.

ARABIE, P. and CARROLL, J. D. (1980). MAPCLUS: A mathematical programming approach to fitting to ADCLUS model. *Psychometrika* **45** 211–235.

ART, D., GNANADESIKAN, R. and KETTENRING, J. R. (1982). Data-based metrics for cluster analysis. *Utilitas Math.* **31A** 75–99.

ASIMOV, D. (1985). The grand tour. *SIAM J. Sci. Statist. Comput.* **6** 128–143.

BAKER, F. B. (1974). Stability of two hierarchical grouping techniques. Case I: Sensitivity to data errors. *J. Amer. Statist. Assoc.* **69** 440–445.

BECKER, P. (1968). *Recognitions of Patterns.* Polyteknisk, Copenhagen.

BECKER, R. A., CHAMBERS, J. M. and WILKS, A. R. (1988). *The New S Language.* Wadsworth and Brooks/Cole, Pacific Grove, Calif.

BELL, P. A. and KOREY, J. J. (1975). QUICLSTR: A FORTRAN program for hierarchical cluster analysis with a large number of subjects. *Behavioral Res. Methods Instrumentation* **7** 575.

Binder, D. A. (1978). Comment on "Estimating mixtures of normal distributions and switching regressions," by R. E. Quandt and J. B. Ramsey. *J. Amer. Statist. Assoc.* **73** 746–747.

Blashfield, R. K., Aldenderfer, M. S. and Morey, L. C. (1982). Cluster analysis literature on validation. In *Classifying Social Data* (H. Hudson, ed.) 167–176. Jossey-Bass, San Francisco.

Bock, H. H. (1985). On significance tests in cluster analysis. *J. Classification* **2** 77–108.

Breiman, L., Friedman, J. H., Olshen, R. A. and Stone, C. J. (1984). *Classification and Regression Trees.* Wadsworth, Belmont, Calif.

Breiman, L., Meisel, W. S. and Purcell, E. (1977). Variable kernel estimates of multivariate densities and their calibration. *Technometrics* **19** 135–144.

Broadbent, S. R. and Hammersley, J. M. (1957). Percolation processes. I: Crystals and mazes. *Proc. Cambridge Philos. Soc.* **53** 629–641.

Buja, A., Hurley, C. and McDonald, J. A. (1986). A data viewer for multivariate data. In *Computer Science and Statistics: Proceedings of the 18th Symposium on the Interface* (T. J. Boardman, ed.) 171–174. Amer. Statist. Assoc., Washington.

Cacoullos, T. (1966). Estimation of a multivariate density. *Ann. Inst. Statist. Math.* **18** 179–189.

Chen, H., Gnanadesikan, R. and Kettenring, J. R. (1974). Statistical methods for grouping corporations. *Sankhyā Ser. B* **36** 1–28.

Chernoff, H. (1972). The selection of effective attributes for deciding between hypotheses using linear discriminant functions. In *Frontiers of Pattern Recognition* (S. Watanabe, ed.) 55–60. Academic, New York.

Chernoff, H. (1973a). Some measures for discriminating between normal multivariate distributions with unequal covariance matrices. In *Multivariate Analysis III* (P. R. Krishnaiah, ed.) 337–344. Academic, New York.

Chernoff, H. (1973b). The use of faces to represent points in k-dimensional space graphically. *J. Amer. Statist. Assoc.* **68** 361–368.

Clunies-Ross, C. W. and Riffenburgh, R. H. (1960). Geometry and linear discrimination. *Biometrika* **47** 185–189.

Cormack, R. M. (1971). A review of classification (with discussion). *J. Roy. Statist. Soc. Ser. A* **134** 321–367.

Cornfield, J. (1962). Joint dependence of risk of coronary heart disease on serum cholesterol and systolic blood pressure: A discriminant function analysis. *Fed. Proc.* **21** 58–61.

Cover, T. M. and Hart, P. E. (1967). Nearest neighbor pattern classification. *IEEE Trans. Inform. Theory* **IT-13** 21–27.

Dallal, G. E. (1975). A user's guide to J. A. Hartigan's clustering algorithms. Unpublished manuscript.

Day, N. E. (1969). Estimating the components of a mixture of normal distributions. *Biometrika* **56** 463–474.

Day, N. E. and Kerridge, D. F. (1967). A general maximum likelihood discriminant. *Biometrics* **23** 313–323.

Defays, D. (1977). An efficient algorithm for a complete link method. *Comput. J.* **20** 364–366.

Dick, N. P. and Bowden, D. C. (1973). Maximum likelihood estimation for mixtures of two normal distributions. *Biometrics* **29** 781–790.

Dixon, W. J., ed. (1981). *BMDP Statistical Software.* Univ. California Press, Berkeley.

Donoho, A. W., Donoho, D. L. and Gasko, M. (1985). MacSpin graphical data analysis software. D^2 Software, Austin, Tex.

Duda, R. O. and Hart, P. E. (1973). *Pattern Classification and Scene Analysis.* Wiley, New York.

Edmonston, B. (1985). MICRO-CLUSTER: Cluster analysis software for microcomputers. *J. Classification* **2** 127–130.

Efron, B. (1975). The efficiency of logistic regression compared to normal discriminant analysis. *J. Amer. Statist. Assoc.* **70** 892–898.

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *Ann. Statist.* **7** 1–26.

Efron, B. (1982). *The Jackknife, the Bootstrap, and Other Resampling Plans.* SIAM, Philadelphia.

Efron, B. (1983). Estimating the error rate of a prediction rule: Improvements on cross-validation. *J. Amer. Statist. Assoc.* **78** 316–331.

Everitt, B. (1980). *Cluster Analysis*, 2nd ed. Halsted, New York.

Everitt, B. S. and Hand, D. J. (1981). *Finite Mixture Distributions.* Chapman and Hall, London.

Farver, T. B. and Dunn, O. J. (1979). Stepwise variable selection in classification problems. *Biometrical J.* **21** 145–153.

Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Ann. Eugenics* **7** (part 2) 179–188.

Fisherkeller, M. A., Friedman, J. H. and Tukey, J. W. (1974). Prim-9: An interactive multidimensional data display and analysis system. SLAC-Pub. 1408, Stanford Linear Accelerator Center, Stanford, Calif.

Fitch, W. M. and Margoliash, E. (1967). Construction of phylogenetic trees. *Science* **155** 279–284.

Fix, E. and Hodges, J. (1951). Discriminatory analysis, nonparametric discrimination: consistency properties. Technical Report, Randolph Field, Texas, USAF School of Aviation Medicine.

Fowlkes, E. B. (1987). Some diagnostics for binary logistic regression via smoothing. *Biometrika* **74** 503–515.

Fowlkes, E. B., Gnanadesikan, R. and Kettenring, J. R. (1987). Variable selection in clustering and other contexts. In *Design, Data, and Analysis, by Some Friends of Cuthbert Daniel* (C. L. Mallows, ed.). Wiley, New York. To appear.

Fowlkes, E. B. and Mallows, C. L. (1983). A method for comparing two hierarchical clusterings (with discussion). *J. Amer. Statist. Assoc.* **78** 553–583.

Friedman, H. P. and Rubin, J. (1967). On some invariant criteria for grouping data. *J. Amer. Statist. Assoc.* **62** 1159–1178.

Friedman, J. H. and Tukey, J. W. (1974). A projection pursuit algorithm for exploratory data analysis. *IEEE Trans. Comput.* **C-23** 881–889.

Gnanadesikan, R. (1977). *Methods for Statistical Data Analysis of Multivariate Observations.* Wiley, New York.

Gnanadesikan, R. and Kettenring, J. R. (1984). A pragmatic review of multivariate methods in applications. In *Statistics: An Appraisal* (H. A. David and H. T. David, eds.) 309–337. Iowa State Univ. Press, Ames, Ia.

Gnanadesikan, R., Kettenring, J. R. and Landwehr, J. M. (1977). Interpreting and assessing the results of cluster analyses. *Bull. Inst. Internat. Statist.* **47** 451–463.

Gnanadesikan, R., Kettenring, J. R. and Landwehr, J. M. (1982). Projection plots for displaying clusters. In *Statistics and Probability: Essays in Honor of C. R. Rao* (G. Kallianpur, P. R. Krishnaiah and J. K. Ghosh, eds.) 281–294. North-Holland, Amsterdam.

Goldman, L., Weinberg, M., Weisberg, M., Olshen, R., Cook, F., Sargent, R. K., Lamas, G. A., Dennis, C., Deckelbam, L., Fineberg, H., Stiratelli, R. and The Medical Housestaffs at Yale-New Haven Hospital and Brigham and Women's Hospital (1982). A computer-derived protocol to aid in the diagnosis of emergency room patients with acute chest pain. *New England J. Med.* **307** 588–596.

Gong, G. (1982). Cross-validation, the jackknife, and the bootstrap: Excess error estimation in forward logistic regression. Technical Report 80, Dept. Statistics, Stanford Univ.

Gordon, L. and Olshen, R. A. (1978). Asymptotically efficient solutions to the classification problem. *Ann. Statist.* **6** 515–533.

Gordon, L. and Olshen, R. A. (1980). Consistent nonparametric regression from recursive partitioning schemes. *J. Multivariate Anal.* **10** 611–627.

Gordon, L. and Olshen, R. A. (1984). Almost surely consistent nonparametric regression from recursive partitioning schemes. *J. Multivariate Anal.* **15** 147–163.

Gower, J. C. and Ross, G. J. S. (1969). Minimum spanning trees and single linkage cluster analysis. *Appl. Statist.* **18** 54–65.

Gray, J. B. and Ling, R. F. (1984). *K*-clustering as a detection tool for influential subsets regression (with discussion). *Technometrics* **26** 305–330.

Haff, L. R. (1986). On linear log-odds and estimation of discriminant coefficients. *Comm. Statist. A—Theory Methods* **15** 2131–2144.

Hall, D. J. and Khanna, D. (1977). The ISODATA method of computation for relative perception of similarities and differences in complex and real data. In *Statistical Methods for Digital Computers* (K. Enslein, A. Ralston and H. S. Wilf, eds.) **3** 340–373. Wiley, New York.

Hand, D. J. (1981). *Discrimination and Classification.* Wiley, New York.

Hartigan, J. A. (1967). Representation of similarity matrices by trees. *J. Amer. Statist. Assoc.* **62** 1140–1158.

Hartigan, J. A. (1975). *Clustering Algorithms.* Wiley, New York.

Hartigan, J. A. (1977). Distribution problems in clustering. In *Classification and Clustering* (J. Van Ryzin, ed.) 45–71. Academic, New York,.

Hartigan, J. A. (1978). Asymptotic distributions for clustering criteria. *Ann. Statist.* **6** 117–131.

Hartigan, J. A. (1981). Consistency of single linkage for high density clusters. *J. Amer. Statist. Assoc.* **76** 388–394.

Hartigan, J. A. and Hartigan, P. M. (1985). The dip test of multimodality. *Ann. Statist.* **13** 70–84.

Hermans, J., Habbema, J. and Schaefer, R. (1982). The ALLOC80 package for discriminant analysis. *Stat. Software Newletter* **8** 15–20.

Hodson, F. R., Sneath, P. H. A. and Doran, J. E. (1966). Some experiments in the numerical analysis of archaeological data. *Biometrika* **53** 311–324.

Hosmer, D. W. (1973). A comparison of iterative maximum likelihood estimates of the parameters of a mixture of two normal distributions under three different types of sample. *Biometrics* **29** 761–770.

Huber, P. J. (1985). Projection pursuit (with discussion). *Ann. Statist.* **13** 435–525.

International Mathematical and Statistical Library (1977). Reference manual library **1**, 6th ed. Houston.

Jambu, M. and Lebeaux, M. O. (1983). *Cluster Analysis and Data Analysis.* North-Holland, Amsterdam.

James, W. and Stein, C. (1961). Estimation with quadratic loss. *Proc. Fourth Berkeley Symp. Math. Statist. Probab.* **1** 311–319. Univ. California Press.

Jardine, C. J., Jardine, N. and Sibson, R. (1967). The structure and construction of taxonomic hierarchies. *Math. Biosci.* **1** 173–179.

Jennrich, R. I. (1962). Linear discrimination in the case of unequal covariance matrices. Unpublished manuscript.

Jennrich, R. and Moore, R. H. (1975). Maximum likelihood estimation by means of nonlinear least squares. *Proc. Statist. Comput. Sect. Amer. Statist. Assoc.* 57–65.

Johnson, S. C. (1967). Hierarchical clustering schemes. *Psychometrika* **32** 241–254.

Kettenring, J. R., Rogers, W. H., Smith, M. E. and Warner, J. L. (1976). Cluster analysis applied to the validation of course objectives. *J. Educ. Statist.* **1** 39–57.

Kleiner, B. and Hartigan, J. A. (1981). Representing points in many dimensions by trees and castles (with discussion). *J. Amer. Statist. Assoc.* **76** 260–276.

Lachenbruch, P. A. (1975). *Discriminant Analysis.* Hafner, New York.

Lachenbruch, P. A. (1982). Robustness of discriminant functions. *SUGI-SAS Group Proc.* **7** 626–632.

Landwehr, J. M., Pregibon, D. and Shoemaker, A. C. (1984). Graphical methods for assessing logistic regression models (with discussion). *J. Amer. Statist. Assoc.* **79** 61–83.

Lennington, R. K. and Rossbach, M. E. (1978). CLASSY: An adaptive maximum likelihood clustering algorithm. Paper presented at 1978 meeting of the Classification Society.

Levisohn, J. R. and Funk, S. G. (1974). CLUSTER: A hierarchical clustering program for large data sets ($n \geq 100$). Research Memo 40, Thurstone Psychometric Lab., Univ. North Carolina.

Ling, R. F. (1973). A probability theory of cluster analysis. *J. Amer. Statist. Assoc.* **68** 159–169.

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. *Proc. Fifth Berkeley Symp. Math. Statist. Probab.* **1** 281–297. Univ. California Press.

Marks, S. and Dunn, O. J. (1974). Discriminant functions when covariance matrices are unequal. *J. Amer. Statist. Assoc.* **69** 555–559.

McCullagh, P. and Nelder, J. A. (1983). *Generalized Linear Models.* Chapman and Hall, London.

McKay, R. J. (1978). A graphical aid to selection of variables in two-group discriminant analysis. *Appl. Statist.* **27** 259–263.

McKay, R. J. and Campbell, N. A. (1982a). Variable selection techniques in discriminant analysis. I. Description. *British J. Math. Statist. Psych.* **35** 1–29.

McKay, R. J. and Campbell, N. A. (1982b). Variable selection techniques in discriminant analysis. II. Allocation. *British J. Math. Statist. Psych.* **35** 30–41.

Michener, C. D. and Sokal, R. R. (1957). A quantitative approach to a problem in classification. *Evolution* **11** 130–162.

Mojena, R. (1977). Hierarchical grouping methods and stopping rules—An evaluation. *Comput. J.* **20** 359–363.

Mojena, R. and Wishart, D. (1980). Stopping rules for Ward's clustering method. *COMPSTAT 1980, Proc. in Computational Statistics* (M. M. Barritt and D. Wishart, eds.) **4** 426–432. Physica, Vienna.

Morgan, J. N. and Messenger, R. C. (1973). THAID: A sequential search program for the analysis of nominal scale dependent variables. Institute for Social Research, Univ. Michigan.

Morgan, J. N. and Sonquist, J. A. (1963). Problems in the analysis of survey data and a proposal. *J. Amer. Statist. Assoc.* **58** 415–435.

Nelder, J. A. and Wedderburn, R. W. M. (1972). Generalized linear models. *J. Roy. Statist. Soc. Ser. A* **135** 370–384.

Olshen, R. A., Gilpin, E., Henning, H., Lewinter, M., Collins, D. and Ross, J., Jr. (1985). Twelve month prognosis following myocardial infarction: Classification trees, logistic regression, and stepwise linear discrimination. In *Proc. of the Berkeley Conference in Honor of Jerzy Neyman and Jack Kiefer* (L. M. Le Cam and R. A. Olshen, eds.) **1** 245–267. Wadsworth, Monterey, Calif.

Pollard, D. (1982). A central limit theorem for *k*-means clustering. *Ann. Probab.* **10** 919–926.

Pregibon, D. (1981). Logistic regression diagnostics. *Ann. Statist.* **9** 705–724.

Rabiner, L. R., Levinson, S. E., Rosenberg, A. E. and Wilpon, J. G. (1979). Speaker independent recognition of isolated words using clustering techniques. *IEEE Trans. Acoust. Speech Signal Process.* **27** 336–349.

Rao, C. R. (1948). The utilization of multiple measurements in problems of biological classification. *J. Roy. Statist. Soc. Ser. B* **10** 159–203.

Rao, C. R. (1952). *Advanced Statistical Methods in Biometric Research.* Wiley, New York.

RAO, C. R. (1960). Multivariate analysis: An indispensable statistical aid in applied research. *Sankhyā* **22** 317–338.

RAO, C. R. (1962). Use of discriminant and allied functions in multivariate analysis. *Sankhyā* **A24** 149–154.

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

RIFFENBURGH, R. H. and CLUNIES-ROSS, C. W. (1960). Linear discriminant analysis. *Pacific Sci.* **14** 251–256.

ROHLF, F. J. (1977). Computational efficacy of agglomerative clustering algorithms. Technical Report RC-6831, IBM Watson Research Center.

ROHLF, F. J. (1982). Single-link clustering algorithms. In *Handbook of Statistics* (P. R. Krishnaiah and L. N. Kanal, eds.) **2** 267–284. North-Holland, Amsterdam.

ROTMAN, S. R., FISHER, A. D. and STAELIN, D. H. (1981). Analysis of multiple-angle microwave observations of snow and ice using cluster analysis techniques. *J. Glaciology* **27** 89–97.

RYAN, T., JOINER, B. and RYAN, B. (1982). *Minitab Reference Manual.* Duxbury, Boston.

SAS INSTITUTE, INC. (1985). *SAS User's Guide: Statistics, Version 5 Edition.* SAS Institute, Inc., Cary, N. C.

SEBER, G. A. F. (1984). *Multivariate Observations.* Wiley, New York.

SHEPARD, R. N. and ARABIE, P. (1979). Additive clustering: Representation of similarities as combinations of discrete overlapping properties. *Psycholog. Rev.* **86** 87–123.

SHIBATA, R. (1981). An optimal selection of regression variables. *Biometrika* **68** 45–54.

SIBSON, R. (1973). SLINK: An optimally efficient algorithm for single-link cluster methods. *Comput. J.* **16** 30–34.

SIEGEL, J. H., GOLDWYN, R. M. and FRIEDMAN, H. P. (1971). Pattern and process in the evolution of human septic shock. *Surgery* **70** 232–245.

SILVERMAN, B. W. (1986). *Density Estimation for Statistics and Data Analysis.* Chapman and Hall, London.

SMYTHE, R. T. and WIERMAN, J. C. (1978). *First Passage Percolation on the Square Lattice. Lecture Notes in Math.* **671**. Springer, New York.

SNEATH, P. H. A. and SOKAL, R. R. (1973). *Numerical Taxonomy.* Freeman, San Francisco.

SOKAL, R. R. (1974). Classification: Purposes, principles, progress, prospects. *Science* **185** 1115–1123.

SPSS, INC. (1986). *SPSSX (a Computer Program).* McGraw-Hill, New York.

STEIN, C. (1956). Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. *Proc. Third Berkeley Symp. Math. Statist. Probab.* **1** 197–206. Univ. California Press.

STONE, C. J. (1977). Consistent nonparametric regression (with discussion). *Ann. Statist.* **5** 595–645.

STONE, M. (1977). Cross-validation: A review. *Math. Operationforsch. Statist. Ser. Statist.* **9** 127–139.

TARTER, M. and KRONMAL, R. (1970). On multivariate density estimates based on orthogonal expansions. *Ann. Math. Statist.* **41** 718–722.

TOUSSAINT, G. T. (1974). Bibliography on estimation of misclassification. *IEEE Trans. Inform. Theory* **IT-20** 472–479.

TRUETT, J., CORNFIELD, J. and KANNEL, W. (1967). A multivariate analysis of the risk of coronary heart disease in Framingham. *J. Chronic Dis.* **20** 511–524.

TRYON, R. C. (1939). *Cluster Analysis.* Edwards Brothers, Ann Arbor, Mich.

VAPNIK, V. N. and CHERVONENKIS, A. YA. (1971). On the uniform convergence of relative frequencies of events to their probabilities. *Theory Probab. Appl.* **16** 264–280.

VAPNIK, V. N. and CHERVONENKIS, A. YA. (1974). *Theory of Pattern Recognition* (in Russian). Nauka, Moscow.

VELDMAN, D. J. (1967). *FORTRAN Programming for the Behavioral Sciences.* Holt, Rinehart and Winston, New York.

VRIJENHOEK, R. C., DOUGLAS, M. E. and MEFFE, G. K. (1985). Conservation genetics of endangered fish populations in Arizona. *Science* **229** 400–402.

WALD, A. (1944). On a statistical problem arising in the classification of an individual into one of two groups. *Ann. Math. Statist.* **15** 145–162.

WALKER, S. B. and DUNCAN, D. B. (1967). Estimation of the probability of an event as a function of several independent variables. *Biometrika* **54** 167–179.

WISHART, D. (1969). Mode analysis: A generalization of nearest neighbor which reduces chaining effects. In *Numerical Taxonomy* (A. J. Cole, ed.). Academic, New York.

WOLFE, J. H. (1970). Pattern clustering by multivariate mixture analysis. *Multivariate Behavioral Res.* **5** 329–350.

WOLFE, J. H. (1971). A Monte-Carlo study of the sampling distribution of the likelihood ratio for mixtures of multinomial distributions. Research Memorandum 72-2, Naval Personnel and Research Training Laboratory, San Diego, Calif.