DISCRIMINANT ANALYSIS

L. PAUL FATTI, *UNIVERSITY OF THE WITWATERSRAND*

DOUGLAS M. HAWKINS and E. LIEFDE RAATH, *COUNCIL FOR SCIENTIFIC AND INDUSTRIAL RESEARCH*

1. PRELIMINARIES

1.1 General Introduction

Discriminant analysis is concerned with the problem of classifying an object of unknown origin into one of two or more distinct groups or populations on the basis of observations made on it. As evidenced by the examples given below, this problem occurs frequently in various fields as diverse as medicine, anthropology and mining, and the techniques of discriminant analysis have been used successfully in many situations. Computer packages for performing the necessary calculations involved in applying some of the techniques have been readily available for some time, although there are still some serious omissions in most of these packages.

Some examples

1. Haemophilia is a sex-linked genetic disease which is transmitted only by females, but whose symptoms are manifest only in males. Under normal medical examination it is impossible to distinguish between females carrying the disease and those not. In order to try and identify female carriers, the levels of a coagulant factor and its related antigen (Factor VIII and Factor VIII RA) in the blood have been suggested as possible discriminators between carriers and non-carriers.

A pilot study was carried out by Gomperts et al (1976) to test how well Factor VIII and its related antigen discriminate between carriers and non-carriers. A sample of 26 white females, of which 11 were known, for genetic reasons, to be carriers and 15 were known to be non-carriers was selected and the Factor VIII and Factor VIII RA levels measured in each subject. Using the linear discriminant function based on the logarithms of the data and the jackknife reclassification procedure to be described in the next section, ten out of the eleven carriers and thirteen out of the fifteen non-carriers were classified correctly. In a parallel study on black females (whose

Factor VIII and Factor VIII RA levels tend to be different from the whites) all ten carriers and fourteen out of fifteen non-carriers were correctly classified. Thus it would appear that carriers of haemophilia can be identified, reasonably reliably, on the basis of their levels of Factor VIII and Factor VIII RA.

2. A common problem occurring in anthropology is that of identi= fying the tribe, race or even sex of a cranium excavated amongst the remains of an ancient civilization (for example De Villiers, 1976). By comparing various measurements (lengths and angles) made on this skull with those made on large numbers of individuals, males and fe= males, from the various tribes at present inhabiting the region, it may be classified to the tribe from which it most probably came. In this problem, we might also be interested in the possibility that it did not originate from any of these tribes, but from another unknown tribe (possibly now extinct). Another possibility is of its occupying some position intermediate between the tribes which could result from intermarriage between members of the different tribes.

A problem arising frequently in this type of application is that of choosing, amongst the large number of possible measurements that can be made on the skull, that subset giving the best discrimination between the various populations.

3. The final example comes from a stratigraphic problem in mining. In the Witwatersrand gold fields the gold bearing reef is one band (the "pay band") in a sedimentary succession, and is usually visually unrecognisable. In badly faulted areas this pay band may fault away, and the miner wishes to know the position in the sedimen= tary succession of the blank band facing him, from which he can de= duce the new position of the pay band.

The trace element geochemistry of a chip of rock from this band provides a means of classification. Given random samples of rock chips from each of the bands, and measurements of the concentrations (in log(parts per million)) of a number of trace elements taken on each chip as well as on the chip of unknown origin, the unknown band has been correctly identified in a high proportion of cases (Hawkins and Rasmussen (1973)).

An interesting feature of this problem is that the rock bands themselves may be considered to be a sample from a "super-population" of rock bands. In different situations, different sets of bands are involved, all drawn randomly from this super-population. The random

effects model, discussed later, may well be appropriate here.

There is also another use of discriminant analysis. Multiple regression is generally presented as a method of making a prediction of an unknown variable from a set of predictors, but in many cases one uses it, not to actually make the prediction, but to study the regression coefficients to see how the dependent variable is affected by the predictors. In a closely analogous way, a discriminant function is computed, not to actually classify observations to their source, but to gain a better understanding of what it is that distinguishes the populations Π_i from one another.

1.2 The basic principles of discriminant analysis

Consider the problem of classifying an observation (vector) X into one of k groups or populations $\Pi_1, \Pi_2, \ldots, \Pi_k$ where Π_i is characterizing a probability density function $f_i(X)$. Suppose further that the observation has a prior probability π_i of coming from Π_i, where $\sum_{i=1}^{k} \pi_i = 1$, and that the cost associated with classifying it into Π_i when it has actually come from Π_j is c_{ij}.

Anderson (1958) shows that the rule that minimizes the expected cost of misclassification is to assign the observation X to Π_i if

$$\sum_{\ell=1}^{k} \pi_\ell \; c_{i\ell} \; f_\ell(X) < \sum_{\ell=1}^{k} \pi_\ell \; c_{j\ell} \; f_\ell(X) \quad , \; j=1,\ldots,k; j\neq i \qquad (1.1)$$

where $c_{jj} = 0$, $j = 1,\ldots,k$.

In the situation where the costs of misclassification are all equal, this rule simplifies to: assign X to Π_i if

$$\pi_i \; f_i(X) = \max_{j=1,\ldots,k} \pi_j \; f_j(X) \qquad (1.2)$$

Assignment rules (1.1) and (1.2) have been derived considering the discriminant analysis problem from a decision-theoretic viewpoint. Viewing it from a purely probabilistic viewpoint instead, the optimal rule is to assign X to that population Π_i for which the posterior probability is the greatest. Now, using Bayes theorem, the posterior probability of Π_j given X is proportional to $\pi_j \; f_j(X)$, so that the optimal probabilistic rule is also (1.2). So, when the costs of misclassification are all equal, the optimal decision-theoretic and probabilistic classification rules are equivalent.

For the two-group (k=2) case rules (1.1) and (1.2) become,

respectively: Assign X to Π_1 if:

$$\pi_2 \, c_{12} \, f_2(X) < \pi_1 \, c_{21} \, f_1(X) \qquad (1.3)$$

and to Π_2 otherwise;

and, assign X to Π_1 if:

$$\pi_2 \, f_2(X) < \pi_1 \, f_1(X) \qquad (1.4)$$

and to Π_2 otherwise.

In practice, the probability density functions $f_i(X)$ i=1,...,k are seldom known. Usually one assumes that they have some particular parametric form (e.g. a multivariate normal distribution) which depends on some unknown parameters. Random samples, called training samples, consisting of observations known to have come from each specific one of these k populations, are then used to construct sample-based classification rules corresponding to (1.1) to (1.4) above.

There are two different methods of setting up such sample-based variants of (1.1) to (1.4). In the *estimative* approach the training samples are used to estimate the unknown parameters using such methods as maximum likelihood. These estimates are then substituted ("plugged in") for the unknown parameters in f_i, and we behave as if the estimates were the true unknown values. Provided consistent estimators are used, the estimative approach leads asymptotically, as the training samples become infinitely large, to the correct optimal classification rule. With small training samples, however, the estimative approach has but a tenuous claim to good theoretical properties, though it is within the general class of empirical Bayes procedures.

The more recent *predictive* approach is a fully Bayesian method. This means that the data are regarded as given, and the unknown parameters as random variables which in due course are integrated out of the model. The steps in this are:

(i) Set up an a priori distribution $g(\theta)$ for the unknown parameters.

(ii) Construct $f(X,T|\theta)$ - the joint distribution of the unknown to be classified, X, and the training sample, T.

(iii) The joint distribution of X, T and θ is $g(\theta)f(X,T|\theta)$. Integrate θ out of this expression, getting, in due course, the conditional distribution of X given T.

As is generally the case with Bayesian methods, criticism of this model usually centres about the realism or otherwise of the

prior distribution $g(\theta)$. This is commonly assumed uninformative, but unless the training sample is extremely small, the exact specification of $g(\theta)$ has little effect on the classification.

The estimative approach consists of setting up the generally very simple algebra for computing the ratio $f_i(X)/f_j(X)$ and then plugging in the estimates of any unknown parameters. In the predictive approach, one must set up the full model corresponding to the particular form of $f_i(X)$ and then integrate out θ. This latter may be no easy feat, and to date predictive procedures have essentially been worked out only for certain models based on underlying normal data.

In the next two sections the classical approach to discriminant analysis, being the estimative approach applied to the multivariate normal distribution, is described. Thereafter, the predictive approach is considered, followed by a section on other approaches to discriminant analysis. In the final section a number of miscellaneous topics are considered.

2. CLASSICAL DISCRIMINANT ANALYSIS

The most common assumption in discriminant analysis is that X is a p-dimensional vector of observations, and that if it comes from Π_i then it follows a multivariate normal distribution with mean vector ξ_i and covariance matrix Σ_i. We will also assume that the costs of misclassification are all equal, so that the decision-theoretic and probabilistic classification rules are the same.

Following the common notation, we will use $N(\xi,\Sigma)$ to denote the p-variate multivariate normal distribution with density

$$f(X) = \frac{1}{\{(2\pi)^p|\Sigma|\}^{\frac{1}{2}}} \exp\{-\tfrac{1}{2}(X-\xi)^T \Sigma^{-1}(X-\xi)\}$$

Later we will also use the notation

$$\text{etr}(A) = \exp\{\text{trace } (A)\}.$$

2.1 Known parameters

In the situation where all the parameters ξ_i, Σ_i and π_i, i=1,...,k are known, classification rule (1.2) becomes: assign X to Π_i if:

$$\delta_i^2(X) + \ell n\ |\Sigma_i| - 2\ \ell n\ \pi_i = \min_{j=1,\ldots,k} \{\delta_j^2(X) + \ell n|\Sigma_j| - 2\ \ell n\ \pi_j\} \quad (2.1)$$

where $\delta_j^2(X) = (X - \xi_j)^T \Sigma_j^{-1}(X - \xi_j)$ is the Mahalanobis distance from X to Π_i. Note that, for equal Σ_j and π_j, j=1,...,k, (2.1) is a minimum distance rule.

For two groups (2.1) simplifies to: assign X to Π_1 if

$$Q_{12}(X) \geq \ell n \quad (\pi_2/\pi_1) \tag{2.2}$$

and to Π_2 otherwise;

where

$$Q_{12}(X) = -\tfrac{1}{2} X^T(\Sigma_1^{-1} - \Sigma_2^{-1})X + X^T(\Sigma_1^{-1}\ \xi_1 - \Sigma_2^{-1}\ \xi_2) - \tfrac{1}{2}(\xi_1^T\ \Sigma_1^{-1}\ \xi_1 - \xi_2^T\ \Sigma_2^{-1}\ \xi_2) + \tfrac{1}{2}\ \ell n\ (|\Sigma_2|/|\Sigma_1|) \tag{2.3}$$

is called the quadratic discriminant function. The boundary between the regions in which X would be classified to Π_1 and to Π_2 may, in principle, have any quadratic shape. The more surprising ones include an annulus on which X is allocated to Π_1 while both inside and outside the annulus it is allocated to Π_2; and a degenerate case in which any X whatever is allocated to Π_1.

2.1.1 *The case of equal covariance matrices*

As elsewhere in statistics, the assumption that the covariance matrices in different groups are equal, i.e. $\Sigma_j = \Sigma$, j=1,...,k simplifies matters considerably. Under this assumption, classification rule (2.1) simplifies to: assign X to Π_i if

$$U_i(X) = \max_{j=1,\ldots,k} U_j(X) \tag{2.4}$$

where $U_j(X)$ is the linear function:

$$U_j(X) = X^T\ \Sigma^{-1}\ \xi_j - \tfrac{1}{2}\ \xi_j^T\ \Sigma^{-1}\ \xi_j + \ell n\ \pi_j \tag{2.5}$$

For two groups (2.4) simplifies further to: assign X to Π_1 if

$$U_{12}(X)=U_1(X)-U_2(X) = (X-\tfrac{1}{2}(\xi_1+\xi_2))^T\Sigma^{-1}(\xi_1-\xi_2) \geq \ell n(\pi_2/\pi_1) \text{ and to } \Pi_2 \text{ otherwise.} \tag{2.6}$$

$U_{12}(X)$ is called the linear discriminant function (LDF).

As a practical matter we might mention that it is seldom wise to compute and report only the LDF. Generally one would wish to guard against, and check for, the possibility that the unknown X does not come from any of the populations sampled. This possibility is easily checked given $\delta_i^2(X)$ (which follows a central χ_p^2 distribution if X

comes from Π_i) or $U_i(X)$ (which is normal $(\frac{1}{2}\mu+\ell n\ \pi_i, \mu)$ if X comes from Π_i; $\mu = \xi_i^T\ \Sigma^{-1}\ \xi_i$). It cannot be checked using only the LDF which may give acceptable values even when X is completely incompatible with any of the source populations.

2.1.2 *Probability of misclassification*

The probability of misclassification under any classification rule is a measure of the expected performance of that rule when classifying observations of unknown origin. In order to obtain this, we note that if u_{ij} is the linear discriminant function corresponding to groups Π_i and Π_j, then, if X is from Π_i, u_{ij} is normally distributed with mean $\frac{1}{2}\ \delta_{ij}^2$ and variance δ_{ij}^2, whereas if X is from Π_j then it has mean $-\frac{1}{2}\ \delta_{ij}^2$ and variance δ_{ij}^2, where

$$\delta_{ij}^2 = (\xi_i - \xi_j)^T\ \Sigma^{-1}(\xi_i - \xi_j) \qquad (2.7)$$

is the Mahalanobis distance between Π_i and Π_j.

For the two-group problem with equal covariance matrices this yields the following probabilities of misclassification using rule (2.4):

$$P_1 = P[\text{Misclassification}|X \text{ from } \Pi_1] = \Phi\left(\frac{\ell n(\pi_2/\pi_1) - \frac{1}{2}\ \delta_{12}^2}{\delta_{12}}\right) \qquad (2.8)$$

and

$$P_2 = P[\text{Misclassification}|X \text{ from } \Pi_2] = \Phi\left(\frac{-\ell n(\pi_2/\pi_1) - \frac{1}{2}\delta_{12}^2}{\delta_{12}}\right) \qquad (2.9)$$

where $\Phi(\cdot)$ is the standard normal distribution function.

The expected probability of misclassification for a randomly chosen observation from Π_1 or Π_2 :

$$P = \pi_1\ P_1 + \pi_2\ P_2$$

is called the error rate of the classification rule. For $\pi_1 = \pi_2$, both misclassification probabilities P_1 and P_2, given in (2.8) and (2.9), are equal, so that the error rate becomes:

$$P = \Phi(-\tfrac{1}{2}\ \delta_{12}) \qquad (2.10)$$

For k > 2 groups, the probabilities of misclassification are expressed in terms of multiple integrals over (k-1)-dimensional normal probability density functions, that can only be evaluated analytically in certain special cases. However, using Bonferroni's first

inequality, we obtain the following upper bound for the probability of misclassification:

$$p_i = P[\text{misclassification} | X \text{ from } \Pi_i] \leq \sum_{\substack{j=1 \\ j\neq i}}^{k} \Phi\left(\frac{\ell n(\pi_j/\pi_i) - \frac{1}{2}\delta_{ij}^2}{\delta_{ij}}\right) \qquad (2.11)$$

$$= \sum_{\substack{j=1 \\ j\neq i}}^{k} \Phi(-\tfrac{1}{2}\delta_{ij}) \text{ if } \pi_j = \pi_i, j=1,\ldots,k \qquad (2.12)$$

In the case of unequal covariance matrices, no simple expressions exist for the probabilities of misclassification, whether there are two or more populations.

2.1.3 *Canonical variables*

In his 1936 paper, Fisher suggested a different approach to the two-group classification problem, which turns out to give the same procedure in terms of $U_{12}(X)$ defined in 2.6: this approach is to seek some scalar linear combination $\alpha^T X$ of the p measurements yielding the largest possible Student's t statistic between Π_1 and Π_2. The approach can be generalized to k > 2 populations, but this generalization no longer corresponds to the linear discriminant functions 2.5, and confusion between the two approaches has misled a number of users of discriminant analysis.

The generalization of Fisher's method leads to defining new "canonical variables" Y_i by

$$Y_i = \alpha_j^T X$$

where α_i is the solution of the eigenvector/eigenvalue problem: $(\Sigma_B - \lambda_i \Sigma)\alpha_i = 0$ corresponding to the i-th largest eigenvalue λ_i,

$$\Sigma_B = k^{-1} \sum_{j=1}^{k} (\xi_j - \xi_.)(\xi_j - \xi_.)^T$$

is the between groups covariance matrix, and

$$\xi_. = k^{-1} \sum_{j=1}^{k} \xi_j .$$

These variables have the property that within populations they are independent normal variates with standard deviation 1; between populations they are also independent, and Y_i has variance λ_i, which is a maximum possible. Of the Y_i so defined, only s = min(k-1,p) can have a nonzero λ_i.

The proportion of the total between groups variance, relative to the within group variance, explained by Y_i is $\lambda_i / \sum_{j=1}^{s} \lambda_j$. The classification rule based on the first r canonical variables, and assuming equal prior probabilities, is: assign X to Π_i if:

$$\sum_{\ell=1}^{r} \{\alpha_\ell^T (X - \xi_i)\}^2 = \min_{j=1,\ldots,k} \sum_{\ell=1}^{r} \{\alpha_\ell^T (X - \xi_j)\}^2 \qquad (2.13)$$

This rule is sub-optimal unless r=s or $\lambda_{r+1} = \ldots = \lambda_s = 0$. For k=2, classification based on the first canonical variable is equivalent to using the linear discriminant function. For $k > 2$ groups, the most common use of the canonical variables is for graphical display using, say, the first two canonical variables.

Another use for the canonical variables when $k \leq p$ is in testing whether X has not come from any of the populations $\Pi_1,\ldots,\Pi_k$, or any combination of them. It is not difficult to show that the last p-k+1 eigenvalues λ_i, i=k,...,p will be identically zero, and that the corresponding canonical variables Y_i, i=k,...,p, will be independently normally distributed with corresponding means $\alpha_i^T \xi$. and unit variances, no matter which of the k populations X comes from. If, however, X comes from a totally different population Π_* with mean vector ξ_* and covariance Σ then the canonical variables will have means $\alpha_i^T \xi_*$. A test for this can therefore be constructed by computing the sum of squared deviations of the Y_i: $T = \Sigma_k^p (Y_i - \alpha_i^T \xi.)^2$ and comparing it with the chi-squared distribution with p-k+1 degrees of freedom.

Looking at the problem from a geometrical point of view, we see that these last canonical variables are orthogonal to the hyperplane containing all ξ_i. Thus T will have a central χ^2 distribution if the unknown comes from a normal distribution whose mean is coplanar with all ξ_i, and will have a noncentral χ^2 distribution otherwise.

2.2 Unknown parameters

When the parameters of the distributions in the k populations are unknown, the usual procedure in classical discriminant analysis is to estimate them from training samples $\{X_{ij}, j=1,\ldots,N_i\}$ from each of the populations Π_i, i=1,...,k.

Let

$$X_{i.} = N_i^{-1} \sum_{j=1}^{N_i} X_{ij}$$

and

$$S_i = n_i^{-1} \sum_{j=1}^{N_i} (X_{ij} - X_{i.})(X_{ij} - X_{i.})^T$$

where $n_i = N_i - 1$

be the sample mean and covariance matrix corresponding to the training sample from Π_i.

The usual, estimative, approach to discriminant analysis is to replace the parameters in the classification rules given above by their sample estimates (Anderson, 1958). Applying this approach to (2.1) yields the sample-based classification rule: Assign X to Π_i if

$$D_i^2(X) + \ell n|S_i| - 2 \ell n \hat{\pi}_i = \min_{j=1,\ldots,k} \{D_j^2(X) + \ell n|S_j| - 2 \ell n \hat{\pi}_j\} \tag{2.14}$$

where

$$D_j^2(X) = (X - X_j)^T S_j^{-1} (X - X_j) \tag{2.15}$$

is the sample-based Mahalanobis distance between X and Π_j, and $\hat{\pi}_j$ is some estimator of π_j.

The corresponding two-group rule is: Assign X to Π_1 if:

$$R_{12}(X) \geq (\hat{\pi}_2/\hat{\pi}_1) \text{ and to } \Pi_2 \text{ otherwise} \tag{2.16}$$

where

$$R_{12}(X) = -\tfrac{1}{2} X^T(S_1^{-1} - S_2^{-1})X + X^T(S_1^{-1} X_{1.} - S_2^{-1} X_{2.}) \tag{2.17}$$
$$-\tfrac{1}{2} (X_{1.}^T S_1^{-1} X_{1.} - X_{2.}^T S_2^{-1} X_{2.}) + \tfrac{1}{2} \ell n(|S_2|/|S_1|)$$

is the sample-based quadratic discriminant function.

Under the assumption that the population covariance matrices in the different populations are equal, the sample-based rule becomes: assign X to Π_i if

$$V_i(X) = \max_{j=1,\ldots,n} V_j(X) \tag{2.18}$$

where

$$V_j(X) = X^T S^{-1} X_{j.} - \tfrac{1}{2} X_{j.}^T S^{-1} X_{j.} - \ell n \hat{\pi}_j \tag{2.19}$$

and $S = n^{-1} \sum_{j=1}^{k} n_j S_j$, where $n = \sum_{j=1}^{k} n_j = \sum_{j=1}^{k} N_j - k$,

is the pooled sample covariance matrix.

For two groups, the sample-based linear discriminant function is