UNIVERSITAT DE BARCELONA

A DISTANCE BASED APPROACH TO DISCRIMINANT ANALYSIS AND ITS PROPERTIES

by

C. M. Cuadras

AMS Subject Classification: 62H30

Mathematics Preprint Series No. 90
January 1991

A DISTANCE BASED APPROACH TO DISCRIMINANT ANALYSIS AND ITS PROPERTIES

C.M. Cuadras
Departament d'Estadistica
Diagonal, 645
08028 Barcelona, Spain

ABSTRACT

A general distance based method for allocating an observation to one of several known populations, on the basis of both continuous and discrete explanatory variables, is proposed and studied. This method depends on a given statistical distance between observations, and leads to some classic discriminant rules by taking suitable distances. The error rates can be easily computed and, unlike other distance classification rules, the prior probabilities can be taken into account.

Key words and phrases: Linear and quadratic discriminant functions; Location model; Statistical distances; Given marginals.

AMS subject classification: 62H30

Work supported in part by CGYCIT grant PS88-0032.

1. INTRODUCTION

The problem of allocating an individual to one of two populations was proposed by Fisher (1936), who obtained the linear discriminant function (LDF). Optimality of LDF under the assumption of multivariate normality, with a common covariance matrix, is a well-known property. LDF gives a good performance even when the above assumptions are violated and several studies have been carried out on LDF as well as alternative methods (Krzanowski, 1977, Seber, 1984, Raveh, 1989). The discrimination problem leads to a quadratic discriminant function (QDF) when the covariance matrices are not equal.

LDF is essentially based on continuous variables (Wald, 1944, Anderson, 1958). However, most problems in applied statistics fall in the mixed case, i.e., the variables are both continuous and discrete. An interesting approach, based in the location model (LM), has been given by Krzanowski (1975, 1986, 1987). LM is applicable when the data contain both continuous and binary variables. This method computes an LDF for each pattern of the binary variables, is optimal under the assumption of conditional normality and can be extended to the multistate case for discrete variables (Lachenbruch and Goldstein, 1979). However, the LM approach needs a considerable computational effort and has not been implemented in the standard packages (Knoke, 1982). Moreover LM cannot be used when there are not enough data for many locations (Vlachonikolis and Marriott, 1982).

Logistic regression (LR) is another interesting approach, which includes quite a wide class of models and also allows discrimination with mixed variables (Anderson, 1972). LR also requires a large amount of computing.

LDF, QDF, LM and LR are all based on the ratio of probability density functions, the maximum likelihood rule (ML). The ML rule is a particular case of the Bayes discriminant rule (BR), the rule obtained after supposing that the populations have prior probabilities. BR provides a general approach to discriminant analysis which has certain optimal properties from a theoretical point of view.

Another general approach, first studied by Matusita (1956) is based on the concept of distance. Let $d(\omega, \pi_i)$ be a distance between ω to π_i, $i = 1, 2$, where ω is an individual to be allocated. The allocation rule is: allocate ω to the nearest population π_i.

LDF is closely related to the Mahalanobis distance and both approaches, ML and distance, are equivalent for multivariate normal data (Mardia *et al*, 1979). If we substitute the covariance matrix by the identity matrix, we obtain the Euclidean distance classifier (EDF), which is an alternative to LDF (Marco *et al*, 1987), and has advantages when the number of variables is large relative to the training sample size. EDF is a distance rule based on Euclidean distance. The distance rule is also equivalent to BR (including ML rule), for the LM model, as is proved by Krzanowski (1986, 1987).

A distance-based approach (DB) for regression and classification with mixed variables was proposed by Cuadras (1989) and Cuadras and Arenas (1990). Although considerable research has been done on discrimination, and the State of the Art of allocation has reached such a high level that it seems difficult to furnish original ideas, this paper has the aim of examining further aspects and applications of discrimination by the above mentioned DB approach.

Given every couple of individuals ω, ω', the DB method uses a distance $d(\omega, \omega')$ to compute discriminant functions. The assumption underlying this method is the following: in some circumnstances (mixed data, for example), it is easier to obtain a distance $d(.,.)$ than a probability density function $p(\cdot)$.

2. A DISTANCE-BASED CLASSIFICATION RULE IN TWO POPULATIONS OBTAINED FROM SAMPLE DATA

Let π_1, π_2 be two populations and suppose that ω is an individual to be allocated. Suppose that a distance function $\delta(.,.)$ is defined on the basis of several (mixed) variables. The distance-based rule (1), or DB rule, is introduced and some of its properties are studied by Cuadras (1989).

2.1 Training sets

Suppose that samples C_1 and C_2 of sizes n_1 and n_2 are availables from π_1 and π_2, respectively. Suppose that an $n_1 \times n_1$ intradistance matrix $\boldsymbol{D}_1 = (\delta_{ij}(1))$ can be computed on C_1 by using the distance $\delta(.,.)$, as well as the distance $\delta_i(1), i = 1, \cdots, n_1$, from ω to each of the set C_1.

$D_2 = (\delta_{ij}(2))$, and $\delta_i(2)$, $i = 1, \cdots, n_2$ are analogously introduced.

Let us define the discriminant functions as follows:

$$f_k(\omega) = \frac{1}{n_k} \sum_i \delta_i^2(k) - \frac{1}{2n_k^2} \sum_{ij} \delta_{ij}^2(k) \qquad k = 1,2 \tag{1}$$

A decision rule for allocation of ω is:

$$\text{allocate } \omega \text{ to } \pi_i \text{ if } f_i(\omega) = \min\{f_1(\omega), f_2(\omega)\} \tag{2}$$

This rule leads to a minimum-distance classification rule, as a consequence of theorem 1.

Let us denote $C = \{1, 2, \cdots, n\}$. $D = (\delta_{ij})$ is an $n \times n$ distance matrix, $(n+1)$ is a new individual, $\delta_1, \cdots, \delta_n$ are the distances from $(n+1)$ to each of C and $f(n+1)$ is defined according to (1). Then $(n+1)$ and C can be represented by the points $P, P_1, \cdots, P_n \in \boldsymbol{R}^r \times g\,\boldsymbol{R}^s$, where $g = \sqrt{-1}$ and $r + s = n - 1$. The coordinates are given by

$$\begin{aligned} P &= (\boldsymbol{x}', g\,\boldsymbol{y}', z) \\ P_i &= (\boldsymbol{x}'_i, g\,\boldsymbol{y}'_i, 0) \qquad i = 1, \cdots, n \end{aligned}$$

and it is satisfied that

$$\delta_{ij}^2 = d^2(P_i, P_j) = \|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2 - \|\boldsymbol{y}_i - \boldsymbol{y}_j\|^2$$

where $\|\cdot\|$ means the Euclidean norm. An explicit construction of $P, P_1, \cdots, P_n$ is given in the following:

Theorem 1

Set $\overline{P} = (\overline{\boldsymbol{x}}, \overline{\boldsymbol{y}}, 0)$, where $\overline{\boldsymbol{x}} = (\sum \boldsymbol{x}_i)/n$, $\overline{\boldsymbol{y}} = (\sum \boldsymbol{y}_i)/n$.
Then

$$f(\omega) = d^2(P, \overline{P})$$

In particular, if D is a Euclidean distance matrix and n+1, C are represented by $\boldsymbol{x}, \boldsymbol{x}_1, \cdots, \boldsymbol{x}_n \in \boldsymbol{R}^p$, then

$$f(\omega) = d^2(\boldsymbol{x}, \overline{\boldsymbol{x}}) = ||\boldsymbol{x} - \overline{\boldsymbol{x}}||^2$$

Proof. Let $\boldsymbol{B} = (b_{ij}) = \boldsymbol{HAH}$, where $\boldsymbol{H}$ is the centring matrix and $\boldsymbol{A} = (a_{ij}) = -\frac{1}{2}(\delta_{ij}^2)$. If $\boldsymbol{B} \geq 0$, i.e. $\boldsymbol{D}$ is a Euclidean distance matrix, let $\boldsymbol{\Lambda}_x$ be the diagonal matrix of positive eigenvalues and $\boldsymbol{X}$ the matrix of corresponding eigenvectors. The rows of $\boldsymbol{X}, \boldsymbol{x'}_1, \cdots, \boldsymbol{x'}_n$, verify $\delta_{ij}^2 = ||\boldsymbol{x}_i - \boldsymbol{x}_j||^2$ and constitute the classical solution of MDS (Mardia *et al.*, 1979).

In general ($\boldsymbol{D}$ need not be Euclidean), we must also consider a diagonal matrix $\boldsymbol{\Lambda}_y$ of negative eigenvalues of $\boldsymbol{B}$, whose corresponding eigenvectors are the columns of the matrix $\boldsymbol{Y}$, and $\boldsymbol{y'}_1, \cdots, \boldsymbol{y'}_n$ are the rows of $\boldsymbol{Y}$ (Lingoes, 1971). Each element i (say) of C can be represented by $P_i = (\boldsymbol{x'}_i, g\,\boldsymbol{y'}_i, 0)$ and $(n+1)$ by $(\boldsymbol{x'}, \boldsymbol{y'}, z)$. Note that, as $\boldsymbol{1}$ is an eigenvector of $\boldsymbol{B}$ of eigenvalue 0, then $\boldsymbol{X'1} = 0$ and $\boldsymbol{Y'1} = 0$, so $\overline{\boldsymbol{x}} = (\sum \boldsymbol{x}_i)/n = 0$, $\overline{\boldsymbol{y}} = (\sum \boldsymbol{y}_i)/n = 0$.

Gower (1968) proves that if $\boldsymbol{B} \geq 0$, the coordinates of $(n+1)$ are given by

$$\begin{aligned} \boldsymbol{x} &= \frac{1}{2}\boldsymbol{\Lambda}_x^{-1}\boldsymbol{X'}\boldsymbol{v} \\ z^2 &= \frac{1}{n}\left(\sum_i \delta_i^2 - \operatorname{tra}\boldsymbol{B}\right) - \boldsymbol{x'x} \end{aligned}$$

where

$$\boldsymbol{v} = \left(b_{11} - \delta_1^2, \cdots, b_{nn} - \delta_n^2\right)'$$

If $\boldsymbol{D}$ is a non–Euclidean distance matrix, it is easily proved that

$$\boldsymbol{y} = \frac{1}{2}\boldsymbol{\Lambda}_y^{-1}\boldsymbol{Y'}\boldsymbol{v}$$

where $\boldsymbol{v}$ is the same vector but z must be obtained from

$$z^2 = \frac{1}{n}\left(\sum_i \delta_i^2 - \operatorname{tra}B\right) - (\boldsymbol{x'x} - \boldsymbol{y'y})$$

On the other hand

$$\sum_{i,j} \delta_{ij}^2 = 2n \operatorname{tra} \boldsymbol{B}$$

hence

$$f(n+1) = z^2 + (\boldsymbol{x}'\boldsymbol{x} - \boldsymbol{y}'\boldsymbol{y})$$

Since $\overline{\boldsymbol{x}} = 0$ and $\overline{\boldsymbol{y}} = 0$, all coordinates of $\overline{P}$ are null. Hence the distance between $\overline{P}$ and $P = (\boldsymbol{x}', g\,\boldsymbol{y}', z)$ is $f(n+1)$. □

Now it is clear that decision rule (2) is equivalent to

$$\text{allocate } \omega \text{ to } \pi_i \text{ if } d^2(P,\overline{P}) < d^2(Q,\overline{Q})$$

otherwise to π_2, where $P, \overline{P}, Q$ and $\overline{Q}$ are suitably constructed. Note that this construction is not needed to obtain $f_1(\omega)$ and $f_2(\omega)$.

Decision rule (2) leads to standard discriminant functions for continuous data and usual distances. Let $\boldsymbol{x}_{k1}, \cdots, \boldsymbol{x}_{kn_k}$, $k = 1,2$, be the training samples and $\boldsymbol{x}$ the observation to be allocated. Then (2) is based on the LDF

$$L(\boldsymbol{x}) = \left[\boldsymbol{x} - \frac{1}{2}(\overline{\boldsymbol{x}}_1 + \overline{\boldsymbol{x}}_2)\right]' \boldsymbol{S}^{-1} (\overline{\boldsymbol{x}}_1 - \overline{\boldsymbol{x}}_2),$$

where $\boldsymbol{S}$ is the pooled sample covariance matrix, provided that the square distance is the Mahalanobis distance

$$(\overline{\boldsymbol{x}}_{ki} - \overline{\boldsymbol{x}}_{kj})' \boldsymbol{S}^{-1} (\overline{\boldsymbol{x}}_{ki} - \overline{\boldsymbol{x}}_{kj}). \qquad (3)$$

If the covariance matrices need not be equal and we replace $\boldsymbol{S}$ for $\boldsymbol{S}_k$ in (3), then rule (2) is based on a QDF.

Finally, if we choose the Euclidean distance, this decision rule is based on the EDF

$$E(\boldsymbol{x}) = \left[\boldsymbol{x} - \frac{1}{2}(\overline{\boldsymbol{x}}_1 + \overline{\boldsymbol{x}}_2)\right]' (\overline{\boldsymbol{x}}_1 - \overline{\boldsymbol{x}}_2).$$

2.2 Error rates

The leaving–one–out method to estimate the probability of misclassification (Lachenbruch, 1975) can be applied with simple additional computation. Let us denote $A = (a_{ij})$, $B = (b_{ij})$ and $C = (c_{ij})$, where

$a_{ij} = \delta^2_{ij}(1)$, $b_{ij} = \delta^2_{ij}(2)$ and $c_{ij} = \delta^2(i,j)$, where $i \in C_1$ and $j \in C_2$. A, B and C are $n_1 \times n_1, n_2 \times n_2$ and $n_1 \times n_2$ matrices respectively.

If

$$a = \sum_{i<j} a_{ij}, \quad a_i = \sum_{r=1}^{n_1} a_{ir}, \quad c_{i.} = \sum_{r=1}^{n_2} c_{ir},$$

and b, b_j and $c_{.j}$ are similarly defined, to allocate $i \in C_1$ we use the discriminant function

$$f_1(i) = (n_1 - 1)^{-1} a_i - (n_1 - 1)^{-2} (a - a_i)$$

and

$$f_2(i) = n_2^{-1} c_{i.} - n_2^{-2} b.$$

Hence

$$m_1 = \text{freq}_{i \in C_1} \left[f_1(i) - f_2(i) > 0 \right] \tag{4}$$

is the frequency of individuals misclassified in C_1. The frequency m_2 of misclassification in C_2 is analogously obtained as a function of b, b_j and $c_{.j}$. If the $n = n_1 + n_2$ observations are random samples from $\pi_1 \cup \pi_2$, the error rate is estimated by

$$\hat{e} = \frac{m_1 + m_2}{n_1 + n_2} \tag{5}$$

Thus $\hat{e}$ is directly computed from A, B and C.

Now suppose that C_1 and C_2 are nonoverlapping classes, that is,

$$d^2 \left(\boldsymbol{x}_i, \overline{\boldsymbol{x}}_{-i} \right) < d^2 \left(\boldsymbol{x}_i^*, \overline{\boldsymbol{y}} \right) \qquad \forall i \in C_1 \tag{6}$$

and

$$d^2\left(\boldsymbol{y}_j,\overline{\boldsymbol{y}}_{-j}\right) < d^2\left(\boldsymbol{y}_j^*,\overline{\boldsymbol{x}}_i\right) \qquad \forall j \in C_2 \tag{7}$$

where $\boldsymbol{x}_i, \boldsymbol{x}_i^*$ are the Euclidean representation of i as a member of C_1, C_2, respectively, and $\overline{\boldsymbol{x}}_{-i}$ is the mean on C_1 after leaving out individual i. Similarly we have $\boldsymbol{y}_j, \boldsymbol{y}_j^*$ and $\overline{\boldsymbol{y}}_{-j}$.

It is easily proved that

$$\begin{aligned} f_1(i) &= d^2\left(\boldsymbol{x}_i,\overline{\boldsymbol{x}}_{-i}\right), \\ f_2(i) &= d^2\left(\boldsymbol{x}_i^*,\overline{\boldsymbol{y}}\right). \end{aligned}$$

Thus, if both (6) and (7) apply, $m_1 = m_2 = 0$ and a perfect classification for nonoverlappling classes is obtained.

2.3 Distance between samples

Suppose that the observable variables are both quantitative and qualitative (nominal, ordinal, dichotomous) and that the distribution of the variables is unknown. Then we may use any distance function chosen among the wide repertory available. One candidate is the square distance $d_{ij}^2 = 1 - s_{ij}$, where s_{ij} is the all–purpose measure of similarity proposed by Gower (1971) and well described by Seber (1984). This distance provides a Euclidean distance matrix and gives good results in regression with mixed data, see Cuadras and Arenas (1990). When some values are missing, d_{ij} can also be obtained but it could be a non–Euclidean distance. However, rule (2) also applies. Therefore, this DB method can be used for handling missing values in discriminant analysis.

For continuous variables, letting $(x_{i1}, \cdots, x_{ip}), (x_{j1}, \cdots, x_{jp})$ be two observations for individuals i and j, we also use the square distance

$$d_{ij}^2 = |x_{i1} - x_{j1}| + \cdots + |x_{ip} - x_{jp}| \tag{8}$$

which will be confronted with the Mahalanobis and Euclidean distances in the examples given in section 6.

3. CLASSIFICATION WHEN THE DISTRIBUTIONS ARE KNOWN

Suppose that the observable variables are related to a random vector with a probability density $p_i(x)$ with respect to a suitable measure λ, if $\boldsymbol{x}$ comes from π_i, $i = 1, 2$.

Let $\boldsymbol{x}_0$ be an observation to be allocated and $\delta(.,.)$ a distance function. The discriminant function which generalizes (1) is given by

$$\begin{aligned} f_i(\boldsymbol{x}_0) &= \int \delta^2(\boldsymbol{x}_0, \boldsymbol{x}) p_i(\boldsymbol{x}) \, d\lambda(\boldsymbol{x}) - \frac{1}{2} \int \delta^2(\boldsymbol{x}, \boldsymbol{y}) p_i(\boldsymbol{x}) p_i(\boldsymbol{y}) \, d\lambda(\boldsymbol{x}) \, d\lambda(\boldsymbol{y}) \\ &= H_{i0} - \frac{1}{2} H_i. \end{aligned} \tag{9}$$

H_{i0} is the expectation of the random variable $\delta^2(\boldsymbol{x}_0, \boldsymbol{x})$ on π_i. We assume $\boldsymbol{x}, \boldsymbol{y}$ to be independent to obtain H_i, the expectation of $\delta^2(\boldsymbol{x}, \boldsymbol{y})$ on $\pi_i \times \pi_i$. The decision rule for allocating $\boldsymbol{x}_0$ is

$$\text{allocate } \boldsymbol{x}_0 \text{ to } \pi_i \text{ if } f_i(\boldsymbol{x}_0) = \min\{f_1(\boldsymbol{x}_0), f_2(\boldsymbol{x}_0)\} \tag{10}$$

Before presenting some properties of the DB rule (10), we introduce a distance between individuals based on the so-called Rao distance (Rao, 1945) and studied later by Burbea and Rao (1982), Oller and Cuadras (1985) and others.

3.1 A distance between individuals

Suppose that the random vector $\boldsymbol{X}$ is related to a statistical model $S = \{p(\boldsymbol{x}; \boldsymbol{\theta})\}$, where $\boldsymbol{\theta}$ is an n-dimensional vector parameter and $p(x, \boldsymbol{\theta})$ is a probability density. Assume the usual regularity conditions and consider the efficient score

$$Z_\theta = \frac{\partial}{\partial \boldsymbol{\theta}} \log p(X; \boldsymbol{\theta}) \tag{11}$$

Then $E(Z_\theta) = 0$ and $\boldsymbol{G} = E(Z_\theta Z_\theta')$ is the Fisher information matrix, where Z_θ is interpreted as a column vector.

Transform two observations $\boldsymbol{x}_1, \boldsymbol{x}_2$ to $\boldsymbol{z}_1, \boldsymbol{z}_2$ by using (11). The square distance between $\boldsymbol{x}_1, \boldsymbol{x}_2$ with respect to the statistical model $S = \{p(\boldsymbol{x}, \theta)\}$ is given by

$$\delta^2(\boldsymbol{x}_1, \boldsymbol{x}_2) = (\boldsymbol{Z}_1 - \boldsymbol{Z}_2)' \boldsymbol{G}^{-1} (\boldsymbol{Z}_1 - \boldsymbol{Z}_2) \tag{12}$$

This distance was introduced by Cuadras (1988, 1989) and especially Oller (1989), who provides further theoretical justification.

If $p(\boldsymbol{x}, \theta)$ is the $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ distribution ($\boldsymbol{\Sigma}$ being fixed), then $Z_\mu = \boldsymbol{\Sigma}^{-1}(X - \boldsymbol{\mu})$, $\boldsymbol{G} = \boldsymbol{\Sigma}^{-1}$ and (12) reduces to the familiar Mahalanobis distance.

If $X = (x_1, \cdots, x_m)$ has the multinomial distribution $\prod p_k^{x_k}$, where $x_k \in \{0, 1\}$, $k = 1, \cdots, m$, x_1 falls in the cell r and x_2 falls in the cell s, we obtain the square distance

$$\delta^2(\boldsymbol{x}_1, \boldsymbol{x}_2) = (1 - \delta_{rs})(p_r^{-1} + p_s^{-1}) \tag{13}$$

where δ_{rs} is the Kronecker delta.

In general, distance (12) is related to the quadratic form $Z_\theta' \boldsymbol{G} Z_\theta$, where Z_θ is given by (11), which satisfies (Oller, 1989)

$$E\left(\boldsymbol{Z'}_\theta \boldsymbol{G}^{-1} \boldsymbol{Z}_\theta\right) = E\left(\operatorname{tra} \boldsymbol{G}^{-1} \boldsymbol{Z'}_\theta \boldsymbol{Z'}_\theta\right) = \operatorname{tra}\left(\boldsymbol{G}^{-1} \boldsymbol{G}\right) = n. \tag{14}$$

Suppose next that the random variables are related to a random vector $W = (X_1, X_2)$, where X_i is a random vector with density $p_i(\boldsymbol{x}_i, \boldsymbol{\theta}_i)$ with respect to a measure λ_i, $i = 1, 2$. Suppose that there is a density $p(\boldsymbol{w}, \theta_1, \theta_2)$ with respect to a suitable measure μ, with marginals $p_i(x_i, \theta_i)$, $i = 1, 2$.

Letting

$$Z_{\theta_i} = \frac{\partial}{\partial \boldsymbol{\theta}_i} \log p_i(\boldsymbol{x}_i, \boldsymbol{\theta}_i)$$

let us define

$$\boldsymbol{G}_{ij} = E(Z_{\theta_i} Z'_{\theta_j}) \qquad i, j = 1, 2$$

and the symmetric matrix

$$G = \begin{pmatrix} G_{11} & G_{12} \\ G_{21} & G_{22} \end{pmatrix},$$

the expectation taken with respect to $p(\boldsymbol{w}, \boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$.

Since

$$\begin{aligned} G_{ii} &= \int\int Z_{\theta_i} Z'_{\theta_i}\, p(\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{\theta}_1, \boldsymbol{\theta}_2)\, d\lambda_1(\boldsymbol{x}_1)\, d\lambda_2(\boldsymbol{x}_2) \\ &= \int Z_{\theta_i} Z'_{\theta_i}\, p_i(\boldsymbol{x}_i, \boldsymbol{\theta}_i)\, d\lambda_i(\boldsymbol{x}_i) \end{aligned}$$

it is clear that $\boldsymbol{G}_{ii}$ is the Fisher information matrix for $p_i(\boldsymbol{x}_i, \boldsymbol{\theta}_i)$, $i = 1, 2$. However, $\boldsymbol{G}$ is not a Fisher information matrix for $p(\boldsymbol{w}, \boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$.

Assume that G_{12} is a suitable fixed matrix and $\boldsymbol{w}_1 = (\boldsymbol{x}_{11}, \boldsymbol{x}_{12})$, $\boldsymbol{w}_2 = (\boldsymbol{x}_{21}, \boldsymbol{x}_{22})$ are two observations transformed to $\boldsymbol{z}_1 = (\boldsymbol{z}_{11}, \boldsymbol{z}_{12})$, $\boldsymbol{z}_2 = (\boldsymbol{z}_{21}, \boldsymbol{z}_{22})$ by means of (11), i.e.,

$$\boldsymbol{z}_{ij} = \frac{\partial}{\partial \boldsymbol{\theta}_j} \log p_j(\boldsymbol{x}_{ij}, \boldsymbol{\theta}_j).$$

We define the square distance between $\boldsymbol{w}_1$ and $\boldsymbol{w}_2$

$$\delta^2(\boldsymbol{w}_1, \boldsymbol{w}_2) = (\boldsymbol{z}_1 - \boldsymbol{z}_2)' \boldsymbol{G}^{-1} (\boldsymbol{z}_1 - \boldsymbol{z}_2) \qquad (15)$$

This distance can be used when $\boldsymbol{w}$ is a mixed random vector, for example, x_1 is the continuous type and x_2 is the discrete type.

When both x_1 and x_2 have absolutely continuous distributions with respect to the Lebesgue measure, the pdf $p(\boldsymbol{w}, \boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$ exists.

Theorem 2

Suppose that (11) gives a transformation from $\boldsymbol{x}_i$ to $\boldsymbol{z}_{\theta_i}$ which is one–to–one and satisfies the condition of the change of variables for multiple integrals. For every fixed $(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$ and a given matrix $\boldsymbol{G}_{12}$ (satisfying some restrictive conditions) there exists a pdf $p(\boldsymbol{w}, \theta_1, \theta_2)$ with marginals $p(\boldsymbol{x}_i, \boldsymbol{\theta}_i)$, $i = 1, 2$, such that $\boldsymbol{G}_{12} = E(Z_{\theta_1} \cdot Z'_{\theta_2})$.

This result can be obtained as a consequence of the problem of constructing probability distributions with given multivariate marginals and a given intercorrelation matrix. However, this construction is quite complicated. A solution is given by Cuadras (1990).

Although the density $p(\boldsymbol{w}, \boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$ exists, the true function is unknown in practice. Nevertheless, distance (15) can be computed. Consequently, a DB classification rule is available for mixed variables, if only the marginal distributions are known. This, again, is the main idea of this paper.

3.2 Parameter estimation with given marginals

Parameters $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ are unknown in applications. Suppose that $\boldsymbol{w}_i = (\boldsymbol{x}_{i1}, \boldsymbol{x}_{i2})$, $i = 1, \cdots, N$, is a random sample from a population with pdf $p(w, \theta_1, \theta_2)$. However, only the marginals pdfs $p_i(x_i, \theta_i), i = 1, 2$, are known.

Let $\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{\theta}}_1, \hat{\boldsymbol{\theta}}_2)$ be the ML estimation obtained considering the product pdf $p_1(x_1, \theta_1) \cdot p_2(x_2, \theta_2)$, i.e.,

$$\sum_{i=1}^{N} \sum_{j=1}^{2} U_j(\boldsymbol{x}_{ij}, \hat{\boldsymbol{\theta}}_j) = 0$$

where

$$U_j(\cdot, \boldsymbol{\theta}_j) = \frac{\partial}{\partial \boldsymbol{\theta}_j} \log p_j(\cdot, \boldsymbol{\theta}_j).$$

The true value $\boldsymbol{\theta}_0$ is characterized by (Kent, 1982, Royall, 1986)

$$E\left\{\sum_{j=1}^{2} U_j(\boldsymbol{X}_j, \boldsymbol{\theta}_j)\right\} = 0$$

where the expectation is taken with respect to $p(\boldsymbol{w}, \boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$.

Since

$$E\{U_1(\boldsymbol{X}_1, \boldsymbol{\theta}_1)\} = \int\int U_1(\boldsymbol{x}_1, \boldsymbol{\theta}_1) p(\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{\theta}_1, \boldsymbol{\theta}_2) \, d\boldsymbol{x}_1 \, d\boldsymbol{x}_2$$

$$= \int U_1(\boldsymbol{x}_1, \boldsymbol{\theta}_1)\, p_1(\boldsymbol{x}_1, \boldsymbol{\theta}_1)\, d\boldsymbol{x}_1 = 0$$

and similarly for $U_2(\boldsymbol{x}_2, \boldsymbol{\theta}_2)$, $\hat{\boldsymbol{\theta}}$ is a consistent estimate of $\boldsymbol{\theta}_0$. (See also Huster *et al.*, 1989). Finally, a consistent estimation of G_{kj} is given by

$$\hat{G}_{kj} = \frac{1}{N} \sum_{i=1}^{N} U_k\left(x_{ik}, \hat{\theta}_k\right) U_j'\left(x_{ij}, \hat{\theta}_j\right).$$

3.3 Some properties of the classification rule

The discriminant function (9) and the decision rule (10) satisfy certain properties.

1) H_i is the diversity coefficient of π_i and H_{i0} is the average difference between π_0 and π_i, where $\pi_0 = \{\boldsymbol{x}_0\}$. Since $H_0 = 0$ we find that

$$f_i(x_0) = H_{i0} - \frac{1}{2}(H_i + H_0)$$

is the Jensen difference between π_i and π_0.

2) Suppose that $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ are independent random vectors and the distances $\delta_1^2(.,.), \delta_2^2(.,.)$ are defined by using x_1, x_2 respectively. According to Oller (1989), let us define the square distance between $(\boldsymbol{x}_{11}, \boldsymbol{x}_{12})$ and $(\boldsymbol{x}_{21}, \boldsymbol{x}_{22})$ by

$$\delta^2(\boldsymbol{x}_{11}, \boldsymbol{x}_{12}; \boldsymbol{x}_{21}, \boldsymbol{x}_{22}) = \delta_1^2(\boldsymbol{x}_{11}, \boldsymbol{x}_{21}) + \delta_2^2(\boldsymbol{x}_{21}, \boldsymbol{x}_{22}),$$

and let $(\boldsymbol{x}_{01}, \boldsymbol{x}_{02})$ be an observation to be allocated. Then, using an obvious notation, it is verified that

$$f_i(\boldsymbol{x}_{01}, \boldsymbol{x}_{02}) = f_{i1}(\boldsymbol{x}_{01}) + f_{i2}(\boldsymbol{x}_{02}). \tag{16}$$

3) If X is $N(\boldsymbol{\mu}_i, \boldsymbol{\Sigma})$ and the Mahalanobis distance is adopted, then

$$f_i(\boldsymbol{x}_0) = (\boldsymbol{x}_0 - \boldsymbol{\mu}_i)' \boldsymbol{\Sigma}^{-1} (\boldsymbol{x}_0 - \boldsymbol{\mu}_i)$$

and rule (10) leads to a minimum–distance rule. This result is even valid for nonnormal data, see theorem 3.

4) If X has the multinomial distribution $\prod p_{ik}^{x_k}$, when X come from π_i, and if we use distance (13), then

$$f_i(x_0) = (1 - p_{ik})/p_{ik} \tag{17}$$

if $\boldsymbol{x}_0$ falls in the cell k. Thus, the classification rule is:

$$\text{allocate } \boldsymbol{x}_0 \text{ to } \pi_i \text{ if } p_{ik} = \min\{p_{1k}, p_{2k}\}.$$

Let us now suppose that given $\boldsymbol{x}_1, \boldsymbol{x}_2 \in \pi_i$, there exists $\varphi : \pi_i \rightarrow \boldsymbol{R}^q, \boldsymbol{t}_1 = \varphi(\boldsymbol{x}_1), \boldsymbol{t}_2 = \varphi(\boldsymbol{x}_2)$, such that

$$\delta^2(\boldsymbol{x}_1, \boldsymbol{x}_2) = (\boldsymbol{t}_1 - \boldsymbol{t}_2)'(\boldsymbol{t}_1 - \boldsymbol{t}_2) \qquad t_1, t_2 \in \boldsymbol{R}^q$$

i.e., $\delta(.,.)$ is a q–dimensional Euclidean distance. One way to define $\boldsymbol{t}_i = \varphi(\boldsymbol{x}_i)$ is

$$\boldsymbol{t}_i = \boldsymbol{G}^{-\frac{1}{2}}\boldsymbol{Z}_i$$

where $\boldsymbol{Z}_i$ is obtained from (11) and $\boldsymbol{G}$ is the Fisher information matrix. Generally, if $\boldsymbol{x}_1, \cdots, \boldsymbol{x}_N$ are given and $\delta(.,.)$ is a Euclidean distance, $\boldsymbol{t}_1, \cdots, \boldsymbol{t}_N$ can be obtained by metric scaling.

Assume further that

$$\boldsymbol{\mu}_i = E_{\pi_i}[\varphi(\boldsymbol{x})]$$

exists and

$$E_{\pi_i}(||\boldsymbol{t} - \boldsymbol{\mu}_i||)^2 < \infty$$

where E_{π_i} means expectation with respect to $p_i(\boldsymbol{x})$.

Theorem 3

Let $\boldsymbol{x}_0$ be an observation to be allocated, $\boldsymbol{t}_0 = \varphi(\boldsymbol{x}_0)$ and $f_i(\boldsymbol{x}_0)$ given by (9). Then

$$f_i(\boldsymbol{x}_0) = ||\boldsymbol{t}_0 - \boldsymbol{\mu}_i||^2 \tag{18}$$

Proof. Let $\boldsymbol{x}_1, \cdots, \boldsymbol{x}_N$ be iid random vectors from π_i and let $\boldsymbol{t}_i = \varphi(\boldsymbol{x}_i)$, $i = 1, \cdots, N$. Then

$$\delta^2(\boldsymbol{x}_r, \boldsymbol{x}_s) = (\boldsymbol{t}_r - \boldsymbol{t}_s)'(\boldsymbol{t}_r - \boldsymbol{t}_s)$$

and, after some algebra, we find

$$\frac{1}{2N^2} \sum_{r,s} \delta^2(\boldsymbol{x}_r, \boldsymbol{x}_s) = \frac{1}{N} \sum_r (\boldsymbol{t}_r - \bar{\boldsymbol{t}})'(\boldsymbol{t}_r - \bar{\boldsymbol{t}})$$

where $\bar{\boldsymbol{t}} = (\sum_r \boldsymbol{t}_r)/N$. From the law of large numbers it follows that as $N \to \infty$ the above expression tends to

$$\frac{1}{2} E_{\pi_i^2}\{\delta^2(\boldsymbol{x}, \boldsymbol{y})\} = E_{\pi_i}\{\|\boldsymbol{t} - \boldsymbol{\mu}_i\|^2\}.$$

Similarly

$$\frac{1}{N} \sum_r \delta^2(\boldsymbol{x}_0, \boldsymbol{x}_r) = \|\boldsymbol{t}_0 - \bar{\boldsymbol{t}}\|^2 + \frac{1}{N} \sum_r (\boldsymbol{t}_r - \bar{\boldsymbol{t}})'(\boldsymbol{t}_r - \bar{\boldsymbol{t}})$$

and, as $N \to \infty$ this expression tends to

$$E_{\pi_i}\{\delta^2(\boldsymbol{x}_0, \boldsymbol{x})\} = \|\boldsymbol{t}_0 - \boldsymbol{\mu}_i\|^2 + E_{\pi_i}\{\|\boldsymbol{t} - \boldsymbol{\mu}_i\|^2\}$$

and (18) follows. □

For distance (15) an alternative proof of Theorem 3 is given. Using a suitable notation:

$$\delta^2(\boldsymbol{x}_0, \boldsymbol{x}) = (\boldsymbol{z}_0 - \boldsymbol{z})' \boldsymbol{G}^{-1} (\boldsymbol{z}_0 - \boldsymbol{z}),$$

$$E(\boldsymbol{z}'\boldsymbol{G}^{-1}\boldsymbol{z}) = E(\mathrm{tra}\, \boldsymbol{G}^{-1}\boldsymbol{z}\boldsymbol{z}') = \mathrm{tra}\,(\boldsymbol{G}^{-1}\boldsymbol{G}) = n.$$

Hence, because $E(\boldsymbol{z}) = \mathbf{0}$,

$$E\left[\delta^2(\boldsymbol{x}_0, \boldsymbol{x})\right] = \boldsymbol{z}'_0 G^{-1} \boldsymbol{z}_0 + n.$$

$$\delta^2(\boldsymbol{x}, \boldsymbol{y}) = (\boldsymbol{z}_x - \boldsymbol{z}_y)' G^{-1} (\boldsymbol{z}_x - \boldsymbol{z}_y),$$

$$\begin{aligned} E\left[\delta^2(\boldsymbol{x}, \boldsymbol{y})\right] &= E(\boldsymbol{z}'_x \boldsymbol{G}^{-1} \boldsymbol{z}_x) + E(\boldsymbol{z}'_y \boldsymbol{G}^{-1} \boldsymbol{z}_y) - 2E(\boldsymbol{z}'_x \boldsymbol{G}^{-1} \boldsymbol{z}_y) \\ &= 2n. \end{aligned}$$

We conclude that the discriminant function is given by

$$f(\boldsymbol{x}_0) = \boldsymbol{z}'_0 \boldsymbol{G}^{-1} \boldsymbol{z}_0,$$

the square distance between $\boldsymbol{z}_0 = \frac{\partial}{\partial \theta} \log p(\boldsymbol{x}_0, \boldsymbol{\theta})$ and $E(\boldsymbol{z}) = \mathbf{0}$.

4. THE BAYESIAN APPROACH

The Bayes discriminant rule (BR) allocates $\boldsymbol{x}_0$ to the population for which $q_i p_i(\boldsymbol{x})$ is greatest, where q_1 and q_2 are the prior probabilities of drawing an observation from π_1 and π_2 respectively. BR leads to ML when $q_1 = q_2 = 1/2$.

If π_i is $N(\boldsymbol{\mu}_i, \boldsymbol{\Sigma})$, $i = 1, 2$, ML is equivalent to the minimum distance rule provided that the Mahalanobis distance is adopted. For mixed data and using the location model LM, Krzanowski (1986) using a distance based on Matusita affinity, proves that LM is also equivalent to a minimum distance rule. However, neither the LM approach nor the Matusita approach takes the prior probabilities into account. In contrast, this is possible in the DB method.

4.1 Incorporating prior probabilities

Suppose that ω_0 is known to belong to π_i with probability $q_i, i = 1, 2$. Let us consider discriminant function (9) with distance (13). Given $\omega \in \pi_1$ we find

$$\begin{aligned} \delta^2(\omega_0, \omega) &= 0 \quad \text{if} \quad \omega_0 \in \pi_1, \\ &= q_1^{-1} + q_2^{-1} \quad \text{if} \quad \omega_0 \in \pi_2. \end{aligned}$$

Hence

$$H_{10} = q_1 \cdot 0 + q_2(q_1^{-1} + q_2^{-1}) = q_1^{-1}$$

Moreover $H_1 = 0$, so we obtain the function $f_1(\omega_0) = q_1^{-1}$. Similarly $f_2(\omega_0) = q_2^{-1}$. However, rule (10) remains invariant if we add the same constant to f_1 and f_2. Therefore, in order to construct a function that is a square distance, let us introduce the prior discriminant function

$$f_i(\omega_0) = q_i^{-1} - 1 \qquad i = 1, 2.$$

Note that $f_i(\omega_0)$ is consistent with the discriminant function (17).

Thus rule (10) allocates ω_0 to the population for which the prior probability is greater.

Suppose now that a vector observation X is known with density $p_i(\boldsymbol{x})$. Taking into account property (16), let us introduce the posterior discriminant functions

$$f_i^B(\boldsymbol{x}_0) = H_{i0} - \frac{1}{2}H_i + q_i^{-1} - 1 \qquad i = 1,2. \tag{19}$$

where H_{i0}, H_i are defined in (9).

A decision rule for allocating $\boldsymbol{x}_0$ is:

$$\text{allocate } \boldsymbol{x}_0 \text{ to } \pi_i \text{ if } f_i^B(\boldsymbol{x}_0) = \min\{f_1^B(\boldsymbol{x}_0), f_2^B(\boldsymbol{x}_0)\} \tag{20}$$

Discriminant function (19) is constructed by using distance (15), but it also applies for any other distance $\overline{\delta}(.,.)$.

However, note that rule (10) is invariant after multiplying $\overline{\delta}(.,.)$ by a positive constant, but the decision taken from rule (20) could be affected. In order to avoid this arbitrariness, it is necessary to impose a standardizing condition similar to (14) to the quadratic form related to $\delta(.,.)$.

4.2 Multivariate normal

The DB rule based on (18) and the BR are closely related. Suppose that π_i is $N(\boldsymbol{\mu}_i, \boldsymbol{\Sigma})$, $i = 1,2$. Then ML is based on the LDF

$$V(\boldsymbol{x}) = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)' \boldsymbol{\Sigma}^{-1} \boldsymbol{x} - \frac{1}{2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)' \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2),$$

BR is based on

$$B(\boldsymbol{x}) = V(\boldsymbol{x}) + \log(\theta/1 - \theta),$$

and DB is based on

$$D(\boldsymbol{x}) = V(\boldsymbol{x}) + \left(\theta - \frac{1}{2}\right) / [\theta(1-\theta)],$$

where θ and $1-\theta$ are the prior probabilities of drawing an observation from π_1 and π_2 respectively.

For $\theta = \frac{1}{2}$ we obtain $V(\boldsymbol{x}) = B(\boldsymbol{x}) = D(\boldsymbol{x})$, otherwise functions $B(\boldsymbol{x})$ and $D(\boldsymbol{x})$ are different. However, the difference between $\log(\theta/1-\theta)$ and $(\theta - \frac{1}{2})/[\theta(1-\theta)]$ is not too marked in the interval (0.2, 0.8). See Fig. 1.

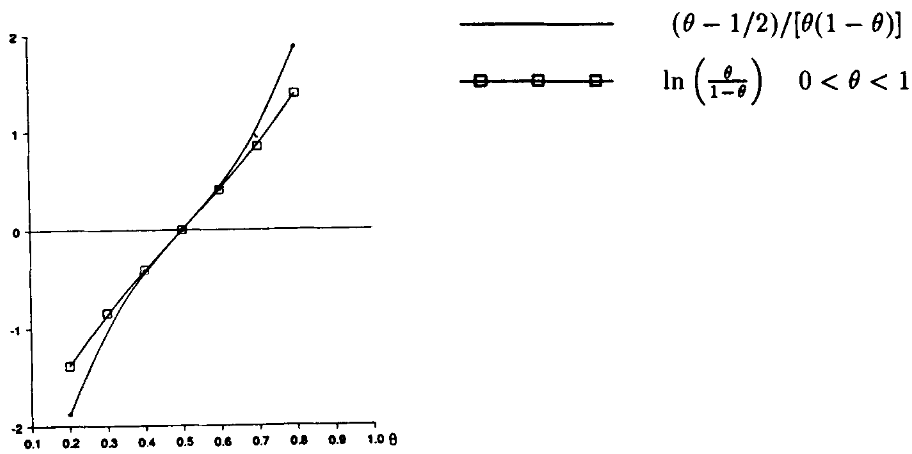

Figure 1. Graphical comparison between distance based rule (continuous curve) and Bayes rule (discrete curve), for the multivariate normal distribution. θ and $1-\theta$ are the prior probabilities.

4.3 Multinomial distribution

Suppose that a random event A has the conditional probabilities $P(A/\pi_i) = p_i$, $i = 1, 2$. If $\omega_0 \in A$, the Bayes decision rule is allocate ω_0 to π_i if $p_1\theta > p_2(1-\theta)$, otherwise to π_2. Therefore BR is based on

$$b = \log p_1 - \log p_2 + \log(\theta/1-\theta)$$

and DB is based on

$$d = (p_2^{-1} - p_1^{-1})/2 + \left(\theta - \frac{1}{2}\right)/[\theta(1-\theta)]$$

For $\theta = \frac{1}{2}$, b and d are equivalent, otherwise the decision could be different. To study this difference, let us consider the Borel set $K \subset (0,1)^3$

$$K = \{(p_1, p_2, \theta) | b \cdot d > 0\}$$

We take the same decision as long as $(p_1, p_2, \theta) \in K$. After some tedious computations, the Lebesgue measure of K is

$$\mu(K) = 0.971$$

which is very close to 1, the Lebesgue measure of $(0,1)^3$.

Thereby, the decision taken after using either BR or DB, is almost the same. This is quite interesting, because although the two decision rules are based on different criteria, their results coincide in practice.

5. CLASSIFICATION INTO SEVERAL POPULATIONS

Suppose that we have k mutually exclusive populations $\pi_1, \cdots, \pi_k$. The theory developed is easily extended to $k > 2$. Let C_i be a sample from π_i and $\boldsymbol{D}_i$ a distance matrix. If ω_0 is an individual to be allocated and $\delta_j(i)$ is the distance from ω_0 to $j \in C_i$, the discriminant function $f_i(\omega)$ is defined as in (1), and the decision rule for allocating ω is

$$\text{allocate } \omega \text{ to } \pi_i \text{ if } f_i(\omega) = \min\{f_1(\omega), \cdots, f_k(\omega)\}. \qquad (21)$$

It is obvious that theorem 1 also applies here, so (20) is a minimum–distance discriminant rule. In addition, the error rate computations are easily obtained.

When the distributions are known (up to parameters), extensions of (9) and (10), as well as therorem 3, do not present further difficulties.

Finally, if prior probabilities $q_1, \cdots, q_k$ are known, we may use the posterior discriminant functions

$$f_i^B(\boldsymbol{x}_0) = f_i(\boldsymbol{x}_0) + q_i^{-1} - 1 \qquad i = 1, \cdots, k,$$

where $\boldsymbol{x}_0$ is the observation to be allocated.

6. SOME EXAMPLES

Three real data examples are used to compare the DB approach with LDF, QDF, EDF and LM approachs.

Data set 1 is the "avanced breast cancer data" used by Krzanowski (1975, 1986, 1987) to illustrate the LM. In this data set, of 186 cases of ablative surgery for advanced breast cancer, 99 were classified as "successful or intermediate" (π_1) and 87 as "failure" (π_2). Gower's distance (section 2.3) is used on the basis of 6 continuous variables and 3 binary variables for the DB method, but we take ranks on the continuous variables instead of numerical values, as the range of most variables was too large.

Data set 2 is the well-known Fisher's Iris data (Fisher, 1936), which consider $n_1 = n_2 = n_3 = 50$ observations on three species of iris, *I. setosa* (π_1), *I. versicolor* (π_2) and *I. virginica* (π_3) and the discrimination problem on the basis of 4 continuous variables. The DB method uses the Euclidean distance (which yields the EDF) and distance (8).

Data set 3, taken from Mardia *et al.* (1979, p. 328), is concerned with the problem of discriminating between the species *Chaetocnema concinna* (π_1) and *Ch. heikertingeri* (π_2) on the basis of 2 continuous variables and samples of sizes $n_1 = n_2 = 18$. Gower's distance (version for continuous variables) is also used.

The leaving–one–out method was used, as described in section 2.2, for computing the error rates. Table 1 summarizes the results obtained.

Table 1

Numbers of misclassifications obtained with the leaving–one–out method.

1. Cancer data

	π_1	π_2	
LM	34	27	61
LDF	41	31	72
EDF	45	43	88
DB	33	32	65

2. Iris data

	π_1	π_2	π_3	
LDF	0	2	1	3
EDF	1	5	7	13
DB	0	2	3	5

3. Chaetocnema data

	π_1	π_2	
LDF	0	1	1
QDF	1	1	2
EDF	2	3	5
DB	0	0	0

7. SUMMARY AND CONCLUSIONS

A distance based approach (DB), which uses a statistical distance between observations, is described and theoretical properties are presented. This method reduces to linear discrimination (LDF), or even quadratic discrimination (QDF) when the sample Mahalanobis distance is used. Moreover the DB approach yields a general discriminant function which reduces to the Euclidean discriminant function (EDF) when the Euclidean sample distance is used.

This method offers a rather simple algebraic way to consider the discrimination problem with mixed variables as well as the missing values case, provided that a suitable distance is adopted. No restriction seems to be necessary on the numbers of binary variables, so the DB method may be an alternative to discrimination based on the location model (LM).

The leaving–one–out method for computing the probability of misclassification can be applied following an elementary matrix computation in this approach.

Finally, unlike other methods based on distances, the prior probabilities can be taken into account in the DB method, with results quite similar to those given by the Bayes decision rule.

8. REFERENCES

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

Anderson, J.A. (1972). Separate sample logistic discrimination. *Biometrika,* 59, 19-35.

Burbea, J. & Rao, C.R. (1982). Entropy differential metric, distance and divergence measures in probability spaces: A unified approach. *J. Multiv. Anal.,* 12, 575-596.

Cuadras, C.M. (1988). Distancias Estadísticas. *Estadística Española,* 30 (119), 295-378.

Cuadras, C.M. (1989). Distance analysis in discrimination and classification using both continuous and categorical variables. In: *Statistical Data Analysis and Inference* (Y. Dodge, ed.). North-Holland, Pu. Co., 459-473.

Cuadras, C.M. (1990). Multivariate distributions with given multivariate marginals and given dependence structure. *Math. Preprint Series* No. 80, Univ. of Barcelona.

Cuadras, C.M. & Arenas, A. (1990). A distance based regression model for prediction with mixed data. *Commun. Statist.-Theory Meth.,* 19 (6), 2261-2279.

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

Gower, J.C. (1968). Adding a point to vector diagrams in multivariate analysis. *Biometrika,* 27, 857-874.

Gower, J.C. (1971). A general coefficient of similarity and some of its properties. *Biometrika,* 27, 857-874.

Huster, W.J., Brookmeyer, R. & Self, S.G. (1989). Modelling paired survival data with covariates. *Biometrika,* 45, 145-156.

Kent, J.T. (1982). Robust properties of likehood ratio tests. *Biometrika,* 69, 19-27.

Knoke, J.D. (1982). Discriminant analysis with discrete and continuous variables. *Biometrics,* 38, 191-200.

Krzanowski, W.J. (1975). Discrimination and classification using both binary and continuous variables. *J. Am. Stat. Assoc.,* 70, 782-790.

Krzanowski, W.J. (1977). The performance of Fisher's linear discriminant function under non-optimal conditions. *Technometrics,* 19, 191-200.

Krzanowski, W.J. (1986). Multiple discriminant analysis in the presence of mixed continuous and categorical data. *Comp. & Maths. with Appls.,* 12A, 179-185.

Krzanowski, W.J. (1987). A comparison between two distance-based discriminant principles. *J. of Classification,* 4, 73-84.

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

Lachenbruch, P.A. & Goldstein, M. (1979). Discriminant analysis. *Biometrics,* 35, 69-85.

Lingoes, J.C. (1971). Some boundary conditions for a monotone analysis of symmetric matrices. *Psychometrika,* 36, 195-203.

Marco, V.R., Young, D.M. & Turner, D.W. (1987). The Euclidean distance classifier: an alternative to linear discriminant function. *Commun. Statist.-Simula.,* 16 (2), 485-505.

Mardia, K.V., Kent, J.T. & Bibby, J.M. (1979). *Multivariate Analysis.* Academic Press, London.

Matusita, K. (1956). Decision rule, based on the distance, for the classification problem. *Ann. Inst. Stat. Math.,* 8, 67-77.

Oller, J.M. (1989). Some geometrical aspects of data analysis and statistics. In: *Statistical Data Analysis and Inference* (Y. Dodge, ed.). North-Holland, Pu. Co., 41-58.

Oller, J.M. & Cuadras, C.M. (1985). Rao's distance for negative multinomial distributions. *Sankhyā,*47, 75-83.

Rao, C.R. (1945). Information and accuracy attainable in the estimation of statistical parameters. *Bull. Calcutta Math. Soc.,* 37, 81-91.

Raveh, A. (1989). A nonmetric approach to linear discriminant analysis. *J. Am. Stat. Assoc.,* 84, 176-183.

Royall, M.M. (1986). Model robust confidence intervals using maximum likelihood estimators. *International Statistical Review,* 54, 221-226.

Schmitz, P.I.M., Habbema, J.D.F., Hermans, J. & Raatgever, J. W. (1983). Comparative performance of four discriminant analysis methods for mixtures of continuous and discrete variables. *Commun. Statist.-Simula. Computa.,* 12 (6), 727-751.

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

Vlachonikolis, I.G. & Marriott, F.H.C. (1982). Discrimination with mixed binary and continuous data. *Appl. Statist.,* 31, 23-31.

Wald, A. (1944). On a statistical problem arising in the classification of an individual into one of two groups. *Ann. Math. Statist.*, 15, 145-163.

REMARK

A computer program for PC to perform a distance based discriminant analysis with mixed variables, is available.