Listy Biometryczne – Biometrical Letters
Vol. 29 (1992), No. 1, 3-20

Some examples of distance based discrimination

C.M. Cuadras

Universitat de Barcelona, Departament d'Estadistica ,
Diagonal 645, 08028 Barcelona,
Spain

Summary

In the problem of allocating an observation into one of several populations, the linear, quadratic and other discriminant functions are obtained from a distance based point of view. This approach depends on a defined distance between observations, seems to be useful for mixed data and can be used for handling missing values. The error rates can easily be computed. A comparison with the location model for discriminant analysis is performed. The method is applied to previously published data and also checked with simulated data.

1. Introduction

The most usually applied discriminant function for assigning an individual to one of two populations Π_1 and Π_2 , is Fisher's linear discriminant function (LDF)

$$L(\mathbf{x}) = [\mathbf{x} - \frac{1}{2}(\bar{\mathbf{x}}_1 + \bar{\mathbf{x}}_2)]'\mathbf{S}^{-1}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2) , \qquad (1)$$

where $\mathbf{x}$ is a vector of observations obtained by taking measurements on p continuous variables, $\bar{\mathbf{x}}_1$ and $\bar{\mathbf{x}}_2$ are the means and $\mathbf{S}$ is the pooled sample covariance matrix, all three of which are computed from training samples of sizes n_1 , n_2 obtained from Π_1 , Π_2 respectively. The discriminant rule is:

allocate $\mathbf{x}$ to Π_1 if $L(\mathbf{x})>0$

and otherwise to Π_2 .

If it is supposed that population covariance matrices are not equal, the assignment is based on the quadratic discriminant function (QDF)

Key words: linear discriminant function; location model; statistical distances; mixed variables; Fisher information matrix.

$$Q(\mathbf{x}) = \frac{1}{2}\log\frac{|S_2|}{|S_1|} - \frac{1}{2}(\bar{\mathbf{x}}_1'S_1^{-1}\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2'S_2^{-1}\bar{\mathbf{x}}_2) + \mathbf{x}'(S_1^{-1}\bar{\mathbf{x}}_1 - S_2^{-1}\bar{\mathbf{x}}_2) - \frac{1}{2}\mathbf{x}'(S_1^{-1} - S_2^{-1})\mathbf{x}\,, \qquad (2)$$

where S_1 and S_2 are the sample covariance matrices. The QDF reduces to the LDF when S_1 and S_2 are replaced by S.

Discrimination based on LDF is optimal under the assumption of multivariate normality and equal covariance matrices (Anderson, 1958). However, LDF is frequently used when both assumptions are violated, with good results. This is so because linear discrimination is robust to non-normality and several studies and comparisons have been carried out on the behaviour of the LDF under non-optimal conditions (Gilbert,1969; Efron,1975; Krzanowski,1977; Lachenbruch and Goldstein, 1979; Raveh, 1989).

Most applications of discriminant analysis fall, however, within the mixed case, that is, where the variables are both continuous and discrete. An appealing approach to mixed discrimination is the location model (LM). This model assumes a normal multivariate distribution for each pattern of discrete variables, e.g. 2^k different states for k binary variables. Krzanowski(1975) uses this model for discrimination using both continuous and binary variables.

Let x be the continuous part of the vector of observations. The LM approach is based on the discriminant functions

$$(\boldsymbol{\mu}_1^{(m)} - \boldsymbol{\mu}_1^{(m)})'\,\boldsymbol{\Sigma}^{-1}[\mathbf{x} - \frac{1}{2}(\boldsymbol{\mu}_1^{(m)} + \boldsymbol{\mu}_2^{(m)})] - \log(p_{2m}/p_{1m}) \qquad (3)$$

where, for each state m of the binary variables, $\boldsymbol{\mu}_i^{(m)}$ is the mean of x and p_{im} is the probability of obtaining and observation in Π_i, i=1,2, respectively. $\boldsymbol{\Sigma}$ is the common covariance matrix. As the number of states could be very large, both a log-linear model and a regression model are used to estimate p_{im} and $\boldsymbol{\mu}_i^{(m)}$ respectively. LM has the following advantages:

a) This approach is optimal under the assumption of conditional normality and it can be extended to the multistate case (Lachenbruch and Goldstein,1979).

b) Using real mixed data sets, LM gives comparable or better results than LDF (Krzanowski,1975; Vlachonikolis and Marriott, 1982).

c) LM is equivalent to the so-called minimum distance rule (Krzanowski, 1986).

In contrast, LM needs a considerable computational effort, it has not been implemented in standard statistical packages (Knoke, 1982) and the number of discrete variables should be limited to six (Krzanowski,1983; a method for deleting mixed variables in LM is given by Krusińska,1989). Finally, simulation studies for LM would be rather complicated (Schmitz *et al*, 1983) and the prior probabilities q_1, q_2 of Π_1, Π_2 respectively, are not taken into account.

Logistic discrimination, as well as other discriminant methods, also allows discrete variables, but the efficiency with respect to LDF is quite similar (Efron,1975; Vlachonikolis and Marriott,1982; Schmitz *et al*,1983; Seber,1984), so it is not considered here.

2. The maximum likelihood, the Bayes and the minimum distance rules

Let $\Pi_1,...,\Pi_g$ be $g \geq 2$ mutually exclusive populations. On the basis of a random vector $\mathbf{X}$, let $\mathbf{x}_o$ be an observation to be allocated. If $p_i(\mathbf{x})$ is the density of $\mathbf{x}$ in Π_i, $i=1,...,g$, with respect to a suitable measure λ, the most general allocation rule is the maximum likelihood rule (ML). This rule is based on the discriminant functions

$$V_{ij}(\mathbf{x}_o) = \log p_i(\mathbf{x}_o) - \log p_j(\mathbf{x}_o). \tag{4}$$

When the probabilities of drawing an observation of Π_i, $i=1,...,g$, are known, i.e. $q_i = \Pr(\Pi_i)$, $i=1,...,g$, then the Bayes discriminant rule (BR) is based on

$$B_{ij}(\mathbf{x}_o) = V_{ij}(\mathbf{x}_o) + \log(q_i) - \log(q_j). \tag{5}$$

An alternative approach is based on the concept of distance and has been studied by K.Matusita in several papers (1956,1964,1973). The M rule is: allocate $\mathbf{x}_o$ to the nearest population, i.e., if $d(\mathbf{x}_o, \Pi_i)$ is a suitable distance, the rule is:

$$\text{allocate } \mathbf{x}_o \text{ to } \Pi_i \quad \text{if} \quad d(\mathbf{x}_o,\Pi_i) = \min\{d(\mathbf{x}_o,\Pi_1),...,d(\mathbf{x}_o,\Pi_g)\} \tag{6}$$

The M rule is quite general. For multivariate normal data and taking the Mahalanobis distance, M leads to a ML rule (Mardia *et al*, 1979). Using an extension of the Matusita affinity to mixed variables, Krzanowski (1986,1987) proves that the M rule is also equivalent to the LM discrimination.

3. The distance based classification rule

The distance based approach (DB) for regression and discriminant analysis was introduced by Cuadras (1989). DB uses a distance $\delta(\mathbf{x}_i,\mathbf{x}_j)$ between observations instead of the distance $d(\mathbf{x},\Pi_k)$, and it reduces to LDF and QDF in special cases.

In short, and assuming the notation of the previous section, the DB approach to allocate an observation $\mathbf{x}_o$ by means of a distance function $\delta(.,.)$, uses the discriminant functions

$$f_i(\mathbf{x}_o) = H_{io} - \frac{1}{2} H_i \qquad i = 1,...,g, \tag{7}$$

where

$$H_{io} = \int \delta^2(\mathbf{x}_o,\mathbf{x})\, p_i(\mathbf{x})\, d\lambda(\mathbf{x})$$

is the expectation of $\delta^2(\mathbf{x}_o,\mathbf{x})$ in Π_i, and

$$H_i = \int \delta^2(\mathbf{x},\mathbf{y})\, p_i(\mathbf{x})\, p_i(\mathbf{y})\, d\lambda(\mathbf{x})\, d\lambda(\mathbf{y})$$

is the expectation of $\delta^2(\mathbf{x},\mathbf{y})$ in $\Pi_i \times \Pi_i$ when $\mathbf{x}$ and $\mathbf{y}$ are assumed to be independent of each other.

The DB decision rule for allocating $\mathbf{x}_o$ is

$$\text{allocate } \mathbf{x}_o \text{ to } \Pi_i \quad \text{if} \quad f_i(\mathbf{x}_o) = \min\{f_1(\mathbf{x}_o),\ldots,f_g(\mathbf{x}_o)\}. \tag{8}$$

This classification rule depends on the density $p_i(\mathbf{x})$ and the distance $\delta(.,.)$, and it has interesting properties. If $\mathbf{x}$ has the multinomial distribution, with n pairwise exclusive states,

$$p_i(\mathbf{x}) = \prod_{k=1}^{n} p_{ik}^{x_k} \qquad x_k \in \{0,1\},$$

where $\Sigma p_{ik} = \Sigma x_k = 1$, and we use the distance

$$\delta^2(\mathbf{x}_1,\mathbf{x}_2) = (1-\delta_{rs})(p_{ir}^{-1} + p_{is}^{-1}) \tag{9}$$

if, inside Π_i, $\mathbf{x}_1$ falls in state r and $\mathbf{x}_2$ falls in state s, $1 \le r, s \le n$, where δ_{rs} is the Kronecker delta, then the discriminant function is given by

$$f_i(\mathbf{x}_o) = (1 - p_{ik})/p_{ik} \tag{10}$$

if $\mathbf{x}_o$ falls in state k. Thus rule (8) leads to the allocation of $\mathbf{x}_o$ to Π_i if $p_{ik} = \max\{p_{1k},\ldots,p_{gk}\}$.

If it is supposed that Π_i is the $N(\boldsymbol{\mu}_i, \boldsymbol{\Sigma})$ population and we use the square Mahalanobis distance $(\mathbf{x}_i - \mathbf{x}_j)' \boldsymbol{\Sigma}^{-1}(\mathbf{x}_i - \mathbf{x}_j)$, then we obtain

$$f_i(\mathbf{x}_o) = (\mathbf{x}_o - \boldsymbol{\mu}_i)' \boldsymbol{\Sigma}^{-1} (\mathbf{x}_o - \boldsymbol{\mu}_i). \tag{11}$$

Hence the DB rule is based on the LDF. If Π_i is $N(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$, $i=1,\ldots,g$, where $\boldsymbol{\Sigma}_1, \ldots, \boldsymbol{\Sigma}_g$ need not be equal, then DB is based on a QDF. Note that here the DB rule yields a minimum-distance rule. Under general conditions, it can be proved that the DB rule is essentially an M rule for any distribution of the variables.

Suppose that $\mathbf{x}$ and $\mathbf{y}$ are independent random vectors. Let $\delta_1(.,.)$, $\delta_2(.,.)$ be distance functions related to $\mathbf{x}$ and $\mathbf{y}$ respectively. Let us define the square distance

$$\delta^2(.,.) = \delta_1^2(.,.) + \delta_2^2(.,.)$$

for $\mathbf{z} = (\mathbf{x},\mathbf{y})$. Then (7) yields

$$f_i(\mathbf{z}) = f_i(\mathbf{x}) + f_i(\mathbf{y}). \tag{12}$$

Thus the discriminant functions used in the DB approach are additive.

As an application of (12), suppose that $\mathbf{x}$ is multinomial and $\mathbf{y}$ is independently multivariate normal. Then a classification rule for the mixed case is based on

$$f_i(\mathbf{x}, \mathbf{y}) = (\mathbf{y} - \boldsymbol{\mu}_i)' \boldsymbol{\Sigma}^{-1} (\mathbf{y} - \boldsymbol{\mu}_i) + (1 - p_{ik}) / p_{ik}. \quad (13)$$

4. Discrimination with mixed variables

From the point of view of distance, both the M and the LM rules (Krzanowski,1986) are based on the square distance $\delta^2 = 2(1-\rho)$, where ρ is the affinity between two density functions (Matusita,1956).

The DB approach can be used for any distance function, but it is preferable to adopt the distance between observations introduced by Cuadras (1988) and Oller (1989), based on the so-called "the Rao-distance". Let $p(\mathbf{x}, \theta)$ be a density function parameterized by θ and assume the usual regularity conditions. The square distance between the observations $\mathbf{x}_1, \mathbf{x}_2$ is given by

$$\delta^2(\mathbf{x}_1, \mathbf{x}_2) = (\mathbf{z}_1 - \mathbf{z}_2)' \mathbf{G}_\theta^{-1} (\mathbf{z}_1 - \mathbf{z}_2) \quad (14)$$

where $\mathbf{z} = \dfrac{\partial}{\partial \theta} \log p(\mathbf{x}, \theta)$ is the efficient score, interpreted as a column vector, and

$$\mathbf{G}_\theta = \mathrm{E}(\mathbf{z}.\mathbf{z}') = -E\left(\frac{\partial^2 \log p(\mathbf{x},\theta)}{\partial \theta . \partial \theta'}\right)$$

is the Fisher information matrix. If the distribution is multivariate normal (14) yields the Mahalanobis distance and we obtain distance (9) for the multinomial distribution.

Suppose now that the observable random vector $\mathbf{w}$ is partitioned into $(\mathbf{x},\mathbf{y})$, where $\mathbf{x}$ is (possibly) discrete and $\mathbf{y}$ is continuous. If $\mathbf{x}$ is multinomial and $\mathbf{y}$ is multivariate normal, and $\mathbf{x}$, $\mathbf{y}$ are independent each other, we may use the results of section 3 to obtain the discriminant function (13). However, in practice $\mathbf{x}$ and $\mathbf{y}$ are correlated. An extension of square distance (14) is given by Cuadras (1989).

Let $p_1(\mathbf{x},\theta_1)$ and $p_2(\mathbf{y},\theta_2)$ be density functions and let us write $\mathbf{w} = (\mathbf{x},\mathbf{y})$, $\mathbf{u} = \dfrac{\partial}{\partial \theta} \log p_1(\mathbf{x},\theta_1)$, $\mathbf{v} = \dfrac{\partial}{\partial \theta} \log p_2(\mathbf{y},\theta_2)$, $\mathbf{z} = (\mathbf{u}',\mathbf{v}')'$. The square distance between $\mathbf{w}_1$ and $\mathbf{w}_2$ is given by

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

where

$$\mathbf{G} = \begin{pmatrix} \mathbf{G}_{uu} & \mathbf{G}_{uv} \\ \mathbf{G}_{vu} & \mathbf{G}_{vv} \end{pmatrix}$$

and $\mathbf{G}_{uu} = E(\mathbf{u}\mathbf{u}')$, $\mathbf{G}_{uv} = \mathbf{G}'_{vu} = E(\mathbf{u}\mathbf{v}')$, $\mathbf{G}_{vv} = E(\mathbf{v}\mathbf{v}')$, the expectations taken with respect to a density $p(\mathbf{w}, \theta_1, \theta_2)$ with marginals $p_1(\mathbf{x},\theta_1)$, $p_2(\mathbf{y},\theta_2)$. The existence of the density $p(\mathbf{w}, \theta_1, \theta_2)$ is proved by Cuadras (1991a).

In practice parameters θ_1, θ_2 are unknown. From a sample of size N of w, we can obtain the maximum likelihood estimate of θ_1 taking into account density $p_1(x,\theta_1)$ and similarly for θ_2. The $\hat{\theta}_1$, $\hat{\theta}_2$ obtained are consistent estimates of the true value of the parameters with respect to $p(w, \theta_1,\theta_2)$. So we can easily obtain a consistent estimate of G. However, also in practice, only the marginal densities are known. The DB approach provides a discriminant decision rule, as it is based on distance (15) which can be estimated and consequently a sample version of discriminant rule (7) is available.

Example: Let x be distributed as Poisson with mean λ and y distributed as $N(\mu,\sigma^2)$, with σ^2 fixed. The efficient scores are

$$u = \frac{x-\lambda}{\lambda}, \quad v = \frac{y-\mu}{\sigma^2}.$$

$$E(u) = 0,\ E(v) = 0,\ E(u^2) = 1/\lambda,\ E(v^2) = 1/\sigma^2$$

and

$$E(uv) = \frac{1}{\lambda\sigma^2} E\{(x-\lambda)(y-\mu)\} = \sigma_{xy} / (\lambda\sigma^2),$$

where σ_{xy} is the covariance between x and y. The maximum likelihood estimate of λ and μ obtained from a sample $(x_1, y_1),...,(x_N, y_N)$ are the sample means $\bar{x}$ and $\bar{y}$ respectively. So the estimate of G is given by

$$\hat{\mathbf{G}} = \begin{pmatrix} s_x^2 / \bar{x}^2 & s_{xy} / (\bar{x}.s_y^2) \\ s_{xy} / (\bar{x}.s_y^2) & 1/s_y^2 \end{pmatrix}$$

where s_x^2 and s_y^2 are the sample variances and s_{xy} is the sample covariance. Note that

$$s_x^2/(\bar{x}^2 s_y^2) - s_{xy}^2/(\bar{x}^2 s_y^4) = (\bar{x}^2 s_y^4)^{-1}(s_x^2 s_y^2 - s_{xy}^2) > 0,$$

hence $\hat{\mathbf{G}}$ is a positive definite matrix. Distance (15) yields

$$((x_1 - x_2)/\bar{x}, (y_1 - y_2)/s_y^2)'\hat{\mathbf{G}}^{-1}((x_1 - x_2)/\bar{x}, (y_1 - y_2)/s_y^2) =$$

$$= (x_1 - x_2, y_1 - y_2)'\mathbf{S}^{-1}(x_1 - x_2, y_1 - y_2),$$

where S is the sample covariance matrix. Therefore the DB rule is based on the LDF. However, if x is distributed as negative binomial, i.e., with probability density function

$$p(x) = \frac{\Gamma(x+r)}{\Gamma(r)x!} p^r(1-p)^x, \quad x = 0,1,2,...$$

where both r and p are unknown parameters, then the DB rule is not linear.

5. The effect of prior probabilities

If the prior probability q_i that $\mathbf{x}_o$ belongs to Π_i is known and we take distance (9), we find $\delta^2(\mathbf{x}_o,\mathbf{x}) = 0$ if $\mathbf{x}_o \in \Pi_i$ and $\delta^2(\mathbf{x}_o,\mathbf{x}) = q_i^{-1} + q_j^{-1}$ if $\mathbf{x}_o \in \Pi_j$, hence $H_{io} = q_i^{-1} + (g-2)$ and, as $H_1 = 0$, we obtain the prior discriminant function $f_i(\mathbf{x}_o) = q_i^{-1} + (g-2)$. As the prior and the posterior information may be interpreted as independent of each other, taking into account (12), we find that $q_i^{-1} + (g-2)$, or better $q_i^{-1} - 1$ (in order to consider a distance, as adding the constant 1-g does not affect the classification rule) should be added to (7) to obtain the posterior discriminant functions

$$f_i(\mathbf{x}_o) = H_{io} - \frac{1}{2}H_i + q_i^{-1} - 1, \quad i = 1,...,g.$$

Suppose now that Π_i is $N(\boldsymbol{\mu}_i,\boldsymbol{\Sigma})$. The ML decision rule is based on the discriminant function

$$V_{ij}(\mathbf{x}_o) = (\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)'\boldsymbol{\Sigma}^{-1}\mathbf{x}_o - \frac{1}{2}(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_i + \boldsymbol{\mu}_j),$$

while the Bayes decision rule (BR) is based on

$$B_{ij}(\mathbf{x}_o) = V_{ij}(\mathbf{x}_o) + \log q_i - \log q_j.$$

The DB rule is based on

$$D_{ij}(\mathbf{x}_o) = V_{ij}(\mathbf{x}_o) + \frac{1}{2}(q_j^{-1} - q_i^{-1}).$$

The decision rule using ML is

allocate $\mathbf{x}_o$ to Π_i if $V_{ij}(\mathbf{x}_o) > 0$ for every $j \neq i$,

and similarly for BR and DB.

For $q_1 = \ldots = q_g = 1/g$, we find $V_{ij} = B_{ij} = D_{ij}$. In general, there is a certain difference between B_{ij} and D_{ij}, but this difference is not too marked. For example, for g=2, this difference as a function of $\gamma = q_1$ can be appreciated below.

$\gamma =$	0.2	0.3	0.4	0.5	0.6	0.7	0.8
$\log\gamma - \log(1-\gamma) =$	-1.38	-0.84	-0.4	0	0.4	0.84	1.38
$[(1-\gamma)^{-1} - \gamma^{-1}]\frac{1}{2} =$	-1.87	-0.95	-0.41	0	0.41	0.95	1.87

Suppose next that the random events $A_1,...,A_m$ have the conditional probabilities

$$\Pr(A_k / \Pi_i) = p_{ik}, \quad k=1,...,m,\ i=1,...,g.$$

If A_k occurs, the BR rule is

allocate to Π_i if $q_i p_{ik} > q_j p_{jk}$ for any $j \neq i$.

The Bayes decision rule is based on the discriminant function

$$b_{ij} = \log p_{ik} + \log q_i - \log p_{jk} - \log q_j,$$

while the DB rule is based on

$$d_{ij} = (p_{jk}^{-1} - p_{ik}^{-1}) + \frac{1}{2}(q_j^{-1} - q_i^{-1}).$$

Again suppose g=2. Let us write $\alpha = p_{ik}$, $\beta = p_{jk}$, $\gamma = q_1$ and consider the functions

$$b(\alpha,\beta,\gamma) = \log(\alpha) - \log(\beta) + \log[\gamma/(1-\gamma)],$$

$$d(\alpha,\beta,\gamma) = (\beta^{-1} - \alpha^{-1})/2 + (\gamma - 0.5)/\gamma(1-\gamma).$$

For $\gamma = 0.5$ we take the same decision because $b(\alpha,\beta,\gamma) > 0$ iff $d(\alpha,\beta,\gamma) > 0$. In general, let us consider the unit cube $U \subset R^3$ and the measurable set

$$C = \{(\alpha,\beta,\gamma) \mid b(\alpha,\beta,\gamma).\ d(\alpha,\beta,\gamma) > 0\} \subset U.$$

We take the same decision as long as $(\alpha,\beta,\gamma) \in C$. The volume of U is 1 and, after some tedious calculations, the volume of C is found to be 0.968. In other words, the decision taken is practically the same using either BR or DB.

6. Distance based discrimination under estimation

The sample discriminant rule for the DB approach can be stated from a data analysis point of view using only distances, i.e., without knowing the probability density function. Suppose that samples $C_1,...,C_g$ of sizes $n_1,...,n_g$ are obtained from $\Pi_1,...,\Pi_g$, respectively. Suppose that a distance matrix $\mathbf{D}_k = (\delta_{ij}(k))$ can be computed on the observations of C_k by means of a distance function $\delta(.,.)$, which can be computed on the basis of the observable variables. Let $\mathbf{x}_0$ be an observation to be allocated and let $\delta(\mathbf{x}_0,i)$ be the distance from $\mathbf{x}_o$ to each element of C_k, where $i \in C_k$.

The sample counterpart of (7) is the discriminant function

$$f_k(\mathbf{x}_o) = \frac{1}{n_k}\sum_i \delta^2(\mathbf{x}_o,i) - \frac{1}{2n_k^2}\sum_{ij}\delta_{ij}^2(k) \quad k = 1,...,g$$

and the DB rule is also given by (8).

Actually this rule is based on several distances between means. If D_k is a Euclidean distance matrix, $\mathbf{x}_o$ and C_k can be related to a Euclidean configuration P, $P_1,...,P_{nk}$ in such a way that $f_k(\mathbf{x}_o) = d^2(\bar{P},P)$, where d(.,.) stands for the Euclidean distance and $\bar{P}$ is the centroid of P_i, $i = 1,...,n_k$. This property is essentially valid for a non-Euclidean distance (Cuadras,1989). Therefore the DB rule leads to an M rule.

For mixed data Gower (1971) proposes the Euclidean square distance $d_{ij}^2 = 1 - s_{ij}$, where s_{ij} is given by

$$s_{ij} = [\sum_{k=1}^{p} (1 - | x_{ik} - x_{jk} | / R_k) + a + \alpha] / [p + (q - d) + r] \qquad (17)$$

where p is the number of continuous variables, a and d are the number of positive and negative matches, respectively, for the q dichotomous variables, and α is the number of matches for the r multistate variables. R_k is the range of the continuous variable k. It is necessary to introduce appropriate corrections on s_{ij} when some values are missing, but then $\mathbf{D}_k$ could be a non-Euclidean distance matrix. However it is not an obstacle for the DB approach which, consequently, can be used in the case of missing values. Note that if we take ranks on the continuous variables, we obtain a nonparametric approach.

Finally, the estimate of the probability of missclassification by using the leaving-one-out method, can be computed easily from the symmetric supermatrix

$$\mathbf{D} = \begin{pmatrix} \mathbf{D}_{11} \ldots \mathbf{D}_{1g} \\ \ldots\ldots\ldots\ldots \\ \mathbf{D}_{g1} \ldots \mathbf{D}_{gg} \end{pmatrix}$$

where $\mathbf{D}_{kk} = \mathbf{D}_k$ is defined as above and $\mathbf{D}_{rs} = (\delta_{ij})$ contains the $n_r \cdot n_s$ distances from each of C_r to each of C_s.

Further theoretical and practical aspects of this method have been studied and reported, not in a final form, in a mimeographed lecture notes of the author (Cuadras,1991b).

7. Real examples

The aim of this section and section 8 is not to give an exhaustive comparison of the performances of the ML, BR and DB methods but to show that the DB approach is useful and may be taken into account for discriminant analysis. Five real data examples were studied and the leaving-one-out procedure for obtaining the individuals misclassified was used by computing the distance matrices described in the previous section.

7.1. Multinomial data

The first example from linguistics was studied by F. Rosés and it is unpublished. In Catalan language, a word containing any of the following strings

ia ià ie ié iè io ió iò

after a consonant letter (i.e. ciència) may be pronounced by using either one (e.g. "ciència") or two syllables (e.g. "ci-ència"). For many catalan words the pronunciation has been established as monosyllable (Π_1) or bisyllable (Π_2), but there is no valid general rule for any word containing these strings.

To predict whether a word containing these strings should be pronounced monosyllable (Π_1) or bisyllable (Π_2), two samples of 146 and 43 words, belonging to Π_1 and Π_2 respectively, were obtained at random from a dictionary and recorded (see Table 1) according to five categorical variables (see Table 2).

Table 1
Words obtained at random and coded as multinomial data

First group (136 words)															Second group (43 words)				
1	4	3	9	1	2	3	3	11	2	6	1	2	2	0	1	1	2	17	2
1	3	2	3	1	2	3	3	3	1	6	1	2	2	1	6	3	2	12	2
3	1	2	9	1	2	3	3	11	1	6	3	2	5	1	6	1	2	17	2
3	3	1	6	1	2	3	3	2	1	6	1	2	16	1	6	1	2	14	2
3	3	2	12	1	2	3	3	11	1	6	1	2	15	0	6	3	2	14	2
3	4	2	11	2	2	3	3	10	1	6	3	2	3	1	6	1	2	9	2
3	1	2	16	1	2	3	3	15	1	2	3	3	1	1	8	3	2	16	2
5	3	2	2	1	6	1	2	15	1	2	3	3	15	1	8	3	2	3	2
5	3	2	2	1	6	1	2	2	1	2	3	3	3	1	8	3	2	1	2
5	3	2	14	1	6	1	2	9	1	2	3	3	15	1	8	3	2	14	0
5	3	2	11	1	6	1	2	2	1	2	3	3	3	1	8	3	2	14	2
5	3	2	2	1	6	1	2	3	1	2	3	3	11	1	8	3	2	11	0
5	3	2	2	1	6	1	2	15	1	2	3	3	14	1	1	1	2	2	2
5	3	2	12	1	6	3	2	14	0	2	3	3	14	1	1	1	2	2	2
5	3	2	14	1	6	3	2	14	1	2	3	3	9	1	1	3	2	14	0
5	3	2	2	1	6	1	2	1	1	2	3	2	3	1	1	1	2	16	0
5	3	2	5	1	7	3	2	2	1	5	3	2	2	1	6	1	2	17	2
6	1	2	2	1	7	3	3	2	1	5	3	2	3	1	6	1	2	10	2
6	1	1	6	1	7	3	3	2	1	5	3	2	2	1	8	3	2	1	2
6	1	2	2	1	7	3	3	2	1	5	3	2	11	1	8	3	2	16	2
6	1	2	3	1	7	3	3	2	1	5	3	2	5	1	8	3	2	10	2
7	3	3	2	1	7	3	3	2	1	5	3	1	6	1	8	3	2	14	0
7	3	3	2	1	7	3	2	15	1	5	3	2	11	1	8	3	2	1	2
7	3	3	2	1	7	3	3	2	1	5	3	2	3	1	8	3	2	1	2
7	3	3	2	1	7	3	2	2	1	2	3	3	21	1	1	1	2	14	0
7	3	3	2	1	8	3	2	16	1	2	3	2	11	1	1	1	2	3	2
7	3	3	2	1	8	3	2	15	1	2	3	2	12	1	1	3	2	13	2
7	3	3	2	1	8	3	2	13	1	2	3	2	14	1	6	1	2	17	2
7	3	3	2	1	8	3	2	15	1	8	3	2	11	1	6	1	2	10	0
7	3	3	2	1	8	3	2	16	1	8	3	2	5	1	6	3	2	17	0
8	3	2	14	1	8	3	2	3	1	8	3	2	14	1	2	3	2	3	2
8	3	2	10	1	8	3	2	13	1	8	3	2	9	1	5	3	2	9	1
8	3	2	2	1	1	3	2	9	1	6	1	2	2	1	5	3	2	3	2
8	3	2	14	1	1	3	2	2	1	6	1	2	14	0	5	3	2	14	0
8	3	2	14	1	1	3	2	12	1	6	1	2	9	1	2	3	2	15	0
8	3	2	17	1	1	4	3	2	1	6	1	2	9	1	2	3	2	14	2
8	3	2	3	1	1	3	2	16	1	1	3	2	3	1	2	3	2	3	2
1	1	2	3	1	1	4	3	13	1	1	1	2	2	1	2	3	2	16	2
1	4	2	10	1	1	4	3	2	1	1	3	2	2	1	8	3	2	1	0
1	3	2	2	1	1	4	3	4	1	2	3	3	3	1	8	3	2	9	0
1	1	2	2	1	1	3	2	11	1	2	3	2	9	1	6	1	2	16	0
1	1	2	2	1	1	4	3	2	1	1	4	3	14	1	6	1	2	14	0
1	3	2	9	1	1	1	2	14	1	6	3	2	1	1	7	3	3	17	2
1	4	3	5	1	1	4	3	2	1	6	3	2	2	0					
1	3	2	14	1	1	3	2	3	1	6	1	2	15	1	1	4	3	14	1

Table 2

The variables and codes used in Example 7.1.

Variable	Name	States and code (in parenthesis)
1	String	ia (1) ià (2) ie (3) iÍ (4) iè (5) io (6) ió (7) iò (8)
2	Location with respect to the syllable	Pre-tonic (1) Tonic with accent (2) Tonic with accent on the second vocal (3) Post-tonic (4)
3	Location with respect to the word	Initial (1) Medial (2) Final (3)
4	Consonant before i	B(1) C(2) D(3) ... V(17)
5	Spanish pronuntiation	Monosyllable (1) Bisyllable (2) Nonexistent in Spanish (0)

(For example, "embòlia" is coded as 1 4 3 9 1 and belongs to Π_1).

As the information is qualitative with many states coded conventionally, LDF and QDF need not be applied here (see Lachenbruch,1975, p.54). The ML rule could be used but, because the multinomial variables are not independent, this rule is rather complicated. Using a log-linear (LL) model (see Krzanowski,1988, p.352) is also problematic because there are many empty cells and parametric estimation fails when first-order interactions are considered. So, only main effects can be included. The model is then equivalent to the assumption of independence, but the data does not fit this model, as the chi-square statistic is very significant. By contrast, the DB approach using Gower's distance (which reduces to the matching coefficient in this case) can be applied. However, for comparison purposes, the four rules LDF, QDF, LL and DB are used with the following misclassifications showed in Table 3.

Table 3

Multinomial data. Comparisons among the distance based approach (DB), the linear discrimination (LDF), the quadratic discrimination (QDF) and the log-linear model (LL) for the multinomial data on Table 1. The misclassifications are obtained by using the leaving-one-out procedure taking variables 1 to 3 (a), 1 to 4 (b) and all variables (c).

	a			b			c		
	Π_1	Π_2	Total	Π_1	Π_2	Total	Π_1	Π_2	Total
LDF	73	8	81	54	14	68	39	15	54
QDF	89	1	90	89	1	90	36	2	38
DB	50	11	61	56	8	64	7	1	8
LL	60	9	69	46	12	58	-	-	-

Remarks:

1) LDF and QDF are used only for comparison purposes.

2) LL cannot classify two words in b because we find null likelihood in both groups.

3) LL cannot classify many words in c by the above reason.

Note that, as variable 5 has a high influence on the classification, we first consider variables 1 to 3 and 1 to 4, but only for 3 variables the leaving-one-out procedure for LL is not conflictive.

7.2. Continuous data

The second real data set is Smith's data consisting of 25 normal (Π_1) and 25 psychotic (Π_2) individuals, who were classified on the basis of 2 continuous variables. This set was studied by Kendall (1957) and Mardia *et al.* (1979) to illustrate the QDF. The third set is Lubischew's data consisting of 2 measurements on samples of three species of Flea-Beetles *Chetocnema concinna* (Π_1), *C. heikertingeri* (Π_2) and *C. heptapotamica* (Π_3), the samples sizes being 21, 31 and 22 respectively, which illustrate the LDF as shown by Seber (1984) and Krzanowski (1988). From Table 4 it is seen that the results obtained are quite similar to those given by the classic methods.

Table 4

Continuous data. Comparisons among the distance based approach (DB), the linear discrimination (LDF) and the quadratic discrimination (QDF). The misclassifications are obtained by using the leaving-one-out procedure.

2.Smith's data (Kendall, 1957, p.154)				3.Flea-Beetles species (Seber, 1984, p.333)				
	Π_1	Π_2	Total		Π_1	Π_2	Π_3	Total
LDF	0	4	4	LDF	1	0	0	1
QDF	2	2	4	QDF	1	0	0	1
DB	1	3	4	DB	1	0	0	1

7.3. Mixed data

The fourth set is "the advanced breast cancer data" used by Krzanowski (1975). From 186 cases of ablative surgery for advanced breast cancer, 99 were classified as "successful or intermediate"(Π_1) and 87 as "failure"(Π_2). The study includes 6 continuous variables and 3 binary variables. Gower's distance (17) was used in the DB method. For the continuous variables in the cancer data, we take ranks instead of numerical values, as the ranges were too large. The fifth set, taken from Mardia et al. (1979, p.294), is also used by Krzanowski (1982). This data is concerned with the average grade (a single quantitative variable) and a qualitative variable with three states: 2, 3 or 4 A-levels. Gower's distance is used for the DB method. The frequency of misclassifications for the complete data are high: LDF(313), QDF(319), DB(280), LM(310), as the 7 groups were quite overlapped. So, we select 3 separate groups, i.e., as denoted by Mardia *et al.*, I, II(i) and '→ ', which we denote by Π_1 , Π_2 and Π_3 , respectively, the sample sizes being $n_1 = 25$, $n_2 = 67$, $n_3 = 26$. See Table 5.

Table 5
Mixed data. Comparisons among the distance based approach (DB), the location model (LM) and the linear discrimination (LDF) The misclassifications are obtained by using the leaving-one-out procedure.

2.Cancer data (Krzanowski, 1975)				3.Students data (Mardia *et. al.*, 1979, p.294)				
	Π_1	Π_2	Total		Π_1	Π_2	Π_3	Total
LM	34	27	61	LM	13	38	7	58
LDF	41	31	72	LDF	11	43	7	61
DB	33	32	65	DB	12	23	11	46

8. Simulations

As LDF and QDF provide optimal solutions for continuous normal data, no simulation is studied for these data. So, we focus our attention on mixed data, i.e., the variables are both continuous and discrete (binary or multistate). In addition, if the conditional distribution for each state of the discrete variables is normal, the LM approach is also considered optimal (Lachenbruch and Goldstein,1979) and no simulation seems to be necessary. Consequently, this study is concentrated in other different situations. For normal data, the methodology is based on Schmitz *et al.*(1983).

8.1. Two populations and four normal variables

Let us consider the mean vectors

$$\boldsymbol{\mu}_0 = (0\,0\,0\,0), \qquad \boldsymbol{\mu}_1 = (1\,1\,1\,1),$$

and the covariance matrices

$$\Sigma_o = \begin{pmatrix} 1.0 & 0.5 & 0.5 & 0.5 \\ 0.5 & 1.0 & 0.5 & 0.5 \\ 0.5 & 0.5 & 1.0 & 0.5 \\ 0.5 & 0.5 & 0.5 & 1.0 \end{pmatrix}, \quad \Sigma_1 = \begin{pmatrix} 1.0 & -0.5 & 0.5 & -0.5 \\ -0.5 & 1.0 & -0.5 & 0.5 \\ 0.5 & -0.5 & 1.0 & -0.5 \\ -0.5 & 0.5 & -0.5 & 1.0 \end{pmatrix}.$$

The distributions $N(\boldsymbol{\mu}_o, \boldsymbol{\Sigma}_o), N(\boldsymbol{\mu}_o, \boldsymbol{\Sigma}_1), N(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_o)$ and $N(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$ are denoted as (0,0),(0,1),(1,0) and (1,1) respectively.

Samples of size 100 of a random vector (x_1, x_2, x_3, x_4) with distribution (0,0),(0,1),(1,0) and (1,1), respectively, were generated. Variables x_3 and x_4 were discretized into two and three categories respectively:

$$x'_3 = 0 \text{ if } x_3 \le 0, \quad x'_3 = 1 \text{ if } x_3 > 0,$$

$$x'_4 = 1 \text{ if } x_4 \le 1, \quad x'_4 = 2 \text{ if } -1 < x_4 \le 1, \quad x'_4 = 3 \text{ if } x_4 > 1.$$

Twelve different discriminant analyses were performed on these populations on the basis of (x_1, x_2, x_3', x_4'). The results are given in Table 6. Giving rank number 1 for the best and rank 4 for the worst method, the average ranges obtained are:

LDF	QDF	DB	LM
3.04	1.66	2.16	3.12

Table 6

Comparisons among LDF, QDF, DB and LM for several combinations of normal populations (four variables), where two variables are continuous and two variables are discretized. The misclassifications are obtained by using the leaving-one-out procedure.

	(0,0)	(1,0)		(0,0)	(1,0)		(0,0)	(1,0)		(0,1)	(1,0)	
LDF	34	25	59	31	27	58	31	27	58	15	22	37
QDF	31	24	55	32	27	59	33	24	57	12	21	33
DB	31	27	58	29	26	55	28	26	54	19	21	40
LM	33	27	60	32	26	58	30	28	58	14	27	41
	(0,1)	(1,0)		(0,1)	(1,0)		(0,0)	(1,1)		(0,0)	(1,1)	
LDF	29	11	40	12	28	40	31	15	46	29	14	43
QDF	27	7	34	7	25	32	30	8	38	28	10	38
DB	24	12	36	11	26	37	31	10	41	26	11	37
LM	23	10	33	12	27	39	28	16	44	26	16	42
	(0,0)	(1,1)		(0,1)	(1,1)		(0,1)	(1,1)		(0,1)	(1,1)	
LDF	28	10	38	12	12	24	9	12	21	11	7	18
QDF	28	6	34	9	11	20	9	8	17	11	8	19
DB	27	13	40	11	11	22	9	10	19	10	10	20
LM	28	15	43	12	14	26	10	16	26	9	9	18

8.2. Two populations and four exponential variables

Let (x_1, x_2, x_3, x_4) be a random vector where each x_i follows the negative exponential distribution with pdf

$$f(x, \theta) = \exp(-(x-\theta)) \quad \text{if } x \geq \theta,$$
$$= 0 \quad \text{otherwise.}$$

Cuadras(1991a) proposes a method of constructing probability distributions with given marginals and given covariance matrix. The covariance matrix used in this example is Σ_o (see section 8.1). Samples of size 100 of (x_1, x_2, x_3, x_4) following this multivariate distribution for $\theta = 0$ (so the mean vector is (1,1,1,1)) and also for $\theta = 1$ (so the mean vector is (0,0,0,0))

were generated. Variables x_3 and x_4 were discretized into two and three categories respectively:

$$x'_3 = 0 \text{ if } x_3 \le 0, \quad x'_3 = 1 \text{ if } x_3 > 0,$$

$$x'_4 = 1 \text{ if } x_4 \le 0.5, \quad x'_4 = 2 \text{ if } 0.5 < x_4 \le 1, \quad x'_4 = 3 \text{ if } x_4 > 1.$$

The misclassifications obtained for discriminating the population with $\theta = 0$ and $\theta = 1$ are given in Table 7. The average ranges obtained are:

LDF	QDF	DB	LM
3.16	2.21	1	3.62

Table 7

Comparisons among LDF, QDF, DB and LM for twelve pairs of populations. The marginal distribution of the four variables is exponential and the covariance matrix is given, except that two variables are discretized. The misclassifications are obtained by using the leaving-one-out procedure.

	1	2	3	4	5	6	7	8	9	10	11	12
LDF	58	59	52	59	56	56	62	58	49	48	55	52
QDF	53	52	46	58	49	49	55	55	44	51	58	52
DB	49	48	43	52	45	44	53	50	39	38	44	44
LM	59	53	48	60	55	53	66	60	49	53	59	55

8.3 Three populations and eight variables

Let us denote by $\mathbf{u}=(1,...,1)$ the vector of ones and consider the 1x8 mean vectors

$$\boldsymbol{\mu}_o = 0\mathbf{u}, \quad \boldsymbol{\mu}_1 = \mathbf{u}, \quad \boldsymbol{\mu}_2 = 2\mathbf{u},$$

and the 8×8 covariance matrices

$$\boldsymbol{\Sigma}_o = \begin{pmatrix} 1.0 & 0.5 & \dots & 0.5 \\ 0.5 & 1.0 & \dots & 0.5 \\ \dots & \dots & \dots & \dots \\ 0.5 & 0.5 & \dots & 1.0 \end{pmatrix} \quad \boldsymbol{\Sigma}_1 = \begin{pmatrix} 1.0 & 0.7 & \dots & 0.7 \\ 0.7 & 1.0 & \dots & 0.7 \\ \dots & \dots & \dots & \dots \\ 0.7 & 0.7 & \dots & 1.0 \end{pmatrix}$$

$$\boldsymbol{\Sigma}_2 = \begin{pmatrix} 1.0 & -0.5 & \dots & -0.5 \\ 0.5 & 1.0 & \dots & 0.5 \\ \dots & \dots & \dots & \dots \\ -0.5 & 0.5 & \dots & 1.0 \end{pmatrix}$$

The normal distribution $N(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_j)$ is denoted by (i,j), i=0,1,2, j=0,1,2. Again samples of size 100 of a random vector $(x_1,...,x_8)$ with distributions (i,j), i=0,1,2, j=0,1,2, were generated. Variables $x_3,...,x_8$ were dichotomized:

$$x'_j = 0 \text{ if } x_j \le 0, \; x'_j = 1 \text{ if } x_j > 0, \; j=3,...,8.$$

Several discriminant analysis were performed on the basis of $(x_1, x_2, x'_3, ..., x'_8)$, where x_1, x_2 are continuous and $x'_3, ..., x'_8$ are binary variables. The misclassifications obtained are presented in Table 8. As the LM method finds zero cell estimates for 2 continuous and 6 binary variables, the last binary variable was discarded and the simulations repeated with 2 continuous and 5 binary variables. The average ranges obtained for eight variables are

LDF	QDF	DB
2.33	2.54	1.12

Omitting a binary variable, the LM works, and the average ranges are

LDF	QDF	DB	LM
3.125	3.625	1.375	1.875

Table 8

Comparisons among LDF, QDF, DB and LM for several combinations of three normal populations (eight variables), where two variables are continuous and six variables are dichotomized. The misclassifications (left column) are obtained by using the leaving-one-out procedure. This computation is repeated omitting a binary variable (right column).

	(0,0)(1,1)(2,2)		(0,1)(1,2)(2,0)		(0,2)(1,0)(2,1)		(0,0)(1,1)(2,2)	
LDF	109	109	105	95	94	97	101	99
QDF	134	124	93	96	96	94	92	95
DB	100	104	90	86	87	92	92	95
LM	-	109	-	87	-	91	-	83
	(0,1)(1,2)(2,2)		(0,2)(1,0)(2,1)		(0,2)(1,1)(2,0)		(0,2)(1,1)(2,1)	
LDF	85	81	94	97	98	98	100	105
QDF	103	95	96	94	111	118	110	118
DB	74	72	87	92	90	89	96	98
LM	-	75	-	91	-	105	-	106
	(0,0)(1,2)(2,0)		(0,1)(1,2)(2,2)		(0,0)(1,0)(2,1)		(0,1)(1,0)(2,0)	
LDF	106	101	87	83	132	130	142	147
QDF	89	95	104	98	141	145	137	140
DB	93	93	75	71	121	122	119	118
LM	-	88	-	75	-	128	-	121

9. Conclusions

The advantages of this distance-based method can be summarized:

a) For continuous variables and adopting the Mahalanobis distance, the method is based on the linear (or quadratic) discriminant function.

b) It is possible to take into account the prior probabilities with similar results to the Bayes decision rule.

c) The mixed variable case can be tackled in a simple algebraic way, provided that a suitable distance is defined. No restrictions are necessary on the number of binary variables and a solution for handling missing values is available.

d) The leaving-one-out procedure for estimating the probability of misclassification can be easily applied.

e) Parametric estimation may fail in both log-linear (for multinomial data) and location model (for mixed data) when there are many discrete variables or many states. Except for the capacity of the computer, no limitations exist for the distance based method.

10. Computer programs and data sets

The linguistic data (section 7.1) was provided to us by F. Rosés and both BMDP and SPSS were used for performing a log-linear discrimination. The cancer data (section 7.3) was obtained while the author visited the Institute of Computer Sciences, Wroclaw, and is due to W.J. Krzanowski, who also provided us with a general location model program for discriminant analysis. A multivariate package, called MULTICUA, created by C. Arenas, C.M. Cuadras and J. Fortiana, was used for performing the linear, quadratic and distance based discriminant analyses. Also, a program from A. Miñarro was employed to generate multivariate normal data. Finally, the generation of correlated data with given marginals, is obtained by using a program written by the author.

Acknowledgements

This work was partially supported by CGYCIT grant PS88-0032. I am grateful to J. Fortiana, W.J. Krzanowski, A. Miñarro, F. Oliva and F. Rosés for providing me with the data and computational assistance. Thanks are also due to the referees for their valuable suggestions.

REFERENCES

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

Cuadras, C.M. (1988). Distancias Estadisticas. *Estadistica Española* **30**, 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, Amsterdam, 459-473.

Cuadras, C.M. (1991a). Probability distributions with given multivariate marginals and given dependence structure. To appear in *Journal of Multivariate Analysis*, 41.

Cuadras, C.M. (1991b). A distance based approach to discriminant analysis and its properties. *Mathematics Preprint Series*, No. 90, Second version, Univ. of Barcelona.

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

Efron, B. (1975). The efficency of logistic regression compared to normal discriminant analysis. *J. Am. Stat. Assoc.* **70**, 892- 898.

Gilbert, E.S. (1969). The effect of unequal variance-covariance matrices on Fisher's linear discriminant function. *Biometrics* **25**, 505-515.

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

Kendall, M.G. (1957). *A course in Multivariate Analysis*. Charles Griffin, London.

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

Krusińska, E. (1989). New procedure for selection of variables in location model for mixed discrimination. *Biometrical Journal* **31**, 511-523.

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. (1983). Stepwise location model choice in mixed-variable discrimination. *Applied Statistics* **32**, 260-266.

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

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

Krzanowski, W.J. (1988). *Principles of Multivariate Analysis: a user's perspective*. Clarendon Press, Oxford.

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

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

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.

Matusita, K. (1964). Distance and decision rule. *Ann. Inst. Stat. Math.* **16**, 305-315.

Matusita, K. (1973). Discrimination and the affinity of distributions. In: *Discriminant Analysis and Applications*. (T. Cacoullos, ed.). Academic Press, New York, 213-223.

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

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

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. Comp.* **12**, 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. *Applied Statistics* **31**, 23-31.

Received 25 February 1991; revised 10 August1991