Journal of Classification 14:117-136 (1997)

The Proximity of an Individual to a Population with Applications in Discriminant Analysis

C. M. Cuadras

Universitat de Barcelona

J. Fortiana

Universitat de Barcelona

F. Oliva

Universitat de Barcelona

Abstract: We develop a proximity function between an individual and a population from a distance between multivariate observations. We study some properties of this construction and apply it to a distance-based discrimination rule, which contains the classic linear discriminant function as a particular case. Additionally, this rule can be used advantageously for categorical or mixed variables, or in problems where a probabilistic model is not well determined. This approach is illustrated and compared with other classic procedures using four real data sets.

Keywords: Categorical and mixed data; Distances between observations; Multidimensional scaling; Discrimination; Classification rules.

The authors thank M. Abrahamowicz, J. C. Gower and M. Greenacre for their helpful comments, and W. J. Krzanowski for providing us with a data set and his quadratic location model program. Work supported in part by CGYCIT grant PB93-0784.

Authors' addresses: Departament d'Estadística, Universitat de Barcelona, Diagonal 645, 08028-BARCELONA, Spain. E-mail: C. M. Cuadras, carlesm@porthos.bio.ub.es; J. Fortiana, fortiana@cerber.mat.ub.es; F. Oliva, francesc@porthos.bio.ub.es.

1. Introduction

Distances between pairs of observations, between pairs of populations or between an observation and a population form the basis of many methods of multivariate analysis. The Euclidean and Mahalanobis distances, the most widely used for continuous data, exist in the three variants, but their functional appearance is the same for all of them, which can lead to confusion.

If, following Arabie (1991), we realize that other distances may be preferable in certain problems of data analysis, either due to data-induced requirements or to the desire for an intrinsic mathematical model (as opposed to an *ad hoc* formulation), it becomes apparent that the three concepts and their inter-relationships need to be reconsidered.

When populations are identified with probability distributions in a parametric statistical model, there exists an intrinsic way of obtaining distances between them, the Rao distance (see Rao 1945, Atkinson and Mitchell 1981, Burbea and Rao 1982, Oller and Cuadras 1985, Mitchell 1992). In the non-parametric case, a definition of distance, based on the concept of affinity between distributions, with applications in statistics and classification, is due to Matusita (1956) and has a close relation with the Mahalanobis distance for the multivariate normal distribution. Krzanowski (1983) extended this distance to the case of mixed variables.

The position is far from simple in the case of distances between individuals. For binary and continuous data there are many definitions (Gower and Legendre 1986) and the distance can be modelled using subjective judgements (Gordon 1990). Under the framework of a parametric model it is possible to define a distance between observations intrinsically (Cuadras 1989, Oller 1989, Miñarro and Oller 1992). For mixed data, as far as we know, the only available distance is the one based on Gower's (1971) similarity coefficient.

In contrast, there are very few distances between an observation and a population. Two strategies have been used for this purpose: Either a population is represented as an ideal point in the space of individuals (Takane et al. 1987), or an observation is treated as a degenerate population. Krzanowski (1987) extends the Matusita distance to this latter case; he then uses the Matusita rule to allocate the observation to a population in the context of the location model in discrimination with mixed variables (Krzanowski 1975), i.e., the observation is allocated to the nearest population.

The aim of this paper is to study geometrical, probabilistic and statistical aspects of an alternative proximity function, introduced by Cuadras (1989), also with the purpose of classification.

Given a dissimilarity δ defined for every pair $\mathbf{x},\mathbf{y}$ of observations of a population Π, i.e., $\delta(\mathbf{x},\mathbf{x}) = 0$ and $\delta(\mathbf{x},\mathbf{y}) = \delta(\mathbf{y},\mathbf{x}) \geq 0$, we construct from δ a proximity function $\phi^2(\mathbf{x},\Pi)$. This construction, based on a suitable dissimilarity, is quite general, since it can be used in discrimination (Cuadras 1992b), regression (Cuadras and Arenas 1990), and generalized ordination (Krzanowski 1994a). We will refer to δ as a *distance* when it satisfies the triangular inequality.

The assumptions underlying this approach lie in the belief that, in some circumstances (mixed variables, missing values, data consisting of character strings,...) it is more natural to work with dissimilarities between observations rather than postulating a given probability distribution. Anderson (1966, p.24) stated that "A classification procedure cannot be distribution-free." Here a probability distribution is assumed to exist, but only implicitly determined by dissimilarities between observations. Its relation with Mahalanobis distance and its applications to discriminant analysis are presented and compared with linear and quadratic discrimination (Anderson 1958, Lachenbruch 1975) and with the location model (Krzanowski 1975, 1986, 1987, 1993).

This approach can be seen as another avatar of the general concept of multidimensional scaling, a technique which has been shown to be useful in many areas of data analysis and statistics (see Cuadras et al. 1995).

2. Geometric Variability and the Proximity Function

The usual measure of variability of a second-order random variable X (i.e., such that $E(|X|^2) < \infty$) is the variance var(X). For a second-order random vector $\mathbf{X}$ with covariance matrix $\Sigma = \mathrm{Var}\,(\mathbf{X})$, a scalar measure of variability is the *total variation* tr (Σ) (Mardia et al. 1979, p.31). It is worth noting that both measures are related to the Euclidean distance, e.g., $2\mathrm{var}\,(X) = E[(X_1 - X_2)^2]$, where X_1, X_2 are independently distributed as X.

Let us represent a population Π by a random vector $\mathbf{X}$, defined on a sample space Ω, with values in $S \subset \mathbf{R}^p$, for some $p \geq 1$, with probability density function f with respect to a suitable measure λ.

Suppose that $\delta(\cdot,\cdot)$ is a dissimilarity on S. Cuadras and Fortiana (1995) define the *geometric variability* of $\mathbf{X}$ with respect to δ as

$$V_\delta(\mathbf{X}) = \frac{1}{2} \int_{S\times S} \delta^2(\mathbf{x},\mathbf{y}) f(\mathbf{x}) f(\mathbf{y}) \lambda(d\mathbf{x}) \lambda(d\mathbf{y}) . \qquad (1)$$

This quantity is a variant of Rao's diversity coefficient DIVC (Rao 1982a, 1982b), in which the distance is not squared and there is no factor 1/2. When δ is the Euclidean distance, we have $V_\delta(X) = \mathrm{var}\,(X)$ for $p = 1$ and in general $V_\delta(\mathbf{X}) = \mathrm{tr}\,(\mathrm{Var}\,(\mathbf{X}))$. For other dissimilarities δ, $V_\delta(\mathbf{X})$ is a generalized measure of dispersion of $\mathbf{X}$.

Assume now that we have two populations Π_1 and Π_2, represented as two independent S-valued random vectors $\mathbf{X}$ and $\mathbf{Y}$, defined on Ω, absolutely continuous with respect to a common measure λ. Let us denote the probability density functions of $\mathbf{X}$ and $\mathbf{Y}$ by f and g, respectively. The quantity

$$\Delta^2(\Pi_1,\Pi_2) = \int_{S\times S}\delta^2(\mathbf{x},\mathbf{y})f(\mathbf{x})g(\mathbf{y})\lambda(d\mathbf{x})\lambda(d\mathbf{y}) - V_\delta(\mathbf{X}) - V_\delta(\mathbf{Y}) \qquad (2)$$

is the Jensen difference (Rao 1982a, 1982b) between the distributions of Π_1 and Π_2:

$$J(f,g) = H(\frac{1}{2}f + \frac{1}{2}g) - \frac{1}{2}H(f) - \frac{1}{2}H(g)\,, \qquad (3)$$

taking as the diversity function $H(f) = 4V_\delta(\mathbf{X})$.

Given $\mathbf{x}_0 \in \mathbf{R}^p$, we define the *proximity* of $\mathbf{x}_0$ to the population Π with respect to δ as

$$\phi_\delta^2(\mathbf{x}_0,\Pi) = \int_S \delta^2(\mathbf{x}_0,\mathbf{x})f(\mathbf{x})\lambda(d\mathbf{x}) - V_\delta(\mathbf{X})\,. \qquad (4)$$

Definition (4) appears as a particular case of (2), taking as $\mathbf{Y}$ the constant $\mathbf{x}_0$, and using the fact that for every two probability measures ν_1 and ν_2 there exists a measure λ such that both ν_1 and ν_2 are absolutely continuous with respect to λ.

A natural analog of the concept of mean value in the present context is that of a δ-*center* of the population Π, which we define as an $\mathbf{x}_0 \in S$ such that $\phi_\delta^2(\mathbf{x}_0,\Pi)$ is a minimum. Some of the usual properties of a mean cannot be extended: in general it is not unique, and the fact that $\mathbf{x}_i$ is a δ-center of Π_i, $i = 1,2$ does not imply $\delta^2(\mathbf{x}_1,\mathbf{x}_2) = \Delta^2(\Pi_1,\Pi_2)$.

Remark 1. Geometric variability and proximity function are quantities that refer to a single population. In the context of discriminant analysis, (1) and (4) could be obtained using a different dissimilarity for each of several populations. However, to compute $\Delta^2(\Pi_1,\Pi_2)$, a single (global) dissimilarity is needed.

Remark 2. The dissimilarity δ appears squared in (1), (2) and (4) as a notational convenience for the metric scaling interpretation of these quantities and in order to recover classical concepts, e.g., that of variance when δ is the Euclidean distance.

For this same reason, $\Delta^2(\Pi_1,\Pi_2)$ is written conventionally as a square, yet the right hand side of (2) is not necessarily positive. It will be so whenever V_δ is a concave function on the space of probability distributions, and this property is satisfied if $\delta^2(\mathbf{x},\mathbf{y})$ is a negative definite function (see Rao, *op. cit.* and Lau 1985).

2.1. Basic Properties of the Proximity Function}

Theorem 1 below gives support to the interpretation of $\Delta^2(\Pi_1,\Pi_2)$ and $\phi_\delta^2(\mathbf{x}_0,\Pi)$ as measures of proximity. In the statement, $(L, (\cdot,\cdot)$, stands for a Euclidean (or Hilbert) space with scalar product $(\cdot,\cdot)$, and $\|u\| = (u,u)^{1/2}$ is the natural norm of $u \in L$.

Theorem 1. *Assume that there exists a representation of* (S,δ) *in an* $(L,(\cdot,\cdot))$, *i.e. a function* $\psi : S \rightarrow L$, *such that for all* $(\mathbf{x},\mathbf{y}) \in S \times S$ *the equality* $\delta^2(\mathbf{x},\mathbf{y}) = \|\psi(\mathbf{x}) - \psi(\mathbf{y})\|^2$ *holds, and that the expected values* $E(\|\psi(\mathbf{X})\|^2)$ *and* $E(\|\psi(\mathbf{Y})\|^2)$ *are finite. Then*

$$V_\delta(\mathbf{X}) = E(\|\psi(\mathbf{X})\|^2) - \|E(\psi(\mathbf{X}))\|^2 , \tag{5a}$$

$$\Delta^2(\Pi_1,\Pi_2) = \|E(\psi(\mathbf{X})) - E(\psi(\mathbf{Y}))\|^2 , \tag{5b}$$

$$\phi_\delta^2(\mathbf{x}_0,\Pi) = \|\psi(\mathbf{x}_0) - E(\psi(\mathbf{X}))\|^2 . \tag{5c}$$

Proof: Denoting $\tilde{x} = \psi(\mathbf{x})$, etc., we have

$$V_\delta = \frac{1}{2}\iint [\|\tilde{x}\|^2 + \|\tilde{y}\|^2 - 2\,(\,\tilde{x},\tilde{y})]\, f(\mathbf{x})f(\mathbf{y})\, \lambda(d\mathbf{x})\lambda(d\mathbf{y})$$
$$= E(\|\tilde{\mathbf{X}}\|^2) - (E(\tilde{\mathbf{X}}), E(\tilde{\mathbf{X}})) ,$$

i.e., (5a). Similarly,

$$\Delta^2(\Pi_1,\Pi_2) = \int[\|\tilde{x}\|^2 + \|\tilde{y}\|^2 - 2(\tilde{x},\tilde{y})]\, f(\mathbf{x})g(\mathbf{y})\, \lambda(d\mathbf{x})\lambda(d\mathbf{y})$$
$$- V_\delta(\mathbf{X}) - V_\delta(\mathbf{Y}) = E(\|\tilde{\mathbf{X}}\|^2) + E(\|\tilde{\mathbf{Y}}\|^2) - 2(E(\tilde{\mathbf{X}}), E(\tilde{\mathbf{Y}}))\,(E(\|\tilde{\mathbf{X}}\|^2)$$
$$-\|E(\tilde{\mathbf{X}})\|^2)\,(E(\|\tilde{\mathbf{Y}}\|^2) - \|E(\tilde{\mathbf{Y}})\|^2),$$

which gives (5b). Since (5c) can be considered as a particular case of (5b), as discussed above, the proof is complete. Alternatively, a direct proof *mutatis mutandis* in the preceding one could be written. ■

The following properties of the proximity function ϕ^2 associated with a squared dissimilarity δ^2 are readily verified:

- The transformation rules of ϕ^2 corresponding to affine transformations of δ^2 are:

  For $a \in \mathbf{R}^+$, if $\tilde{\delta}^2 = a\,\delta^2$, then $\tilde{\phi}^2 = a\,\phi^2$.

  For $b \in \mathbf{R}^+$, if $\tilde{\delta}^2 = \delta^2 + b$, then $\tilde{\phi}^2 = \phi^2 + b/2$.

- Assume that $\mathbf{X} = (\mathbf{X}_1,\mathbf{X}_2)$ and that δ_i^2 is related to $\mathbf{X}_i$, with associated proximity function ϕ_i^2, $i = 1,2$. If we define the squared dissimilarity

$$\delta^2 = \delta_1^2 + \delta_2^2 , \tag{6}$$

  then its associated proximity function is

$$\phi^2 = \phi_1^2 + \phi_2^2 \,. \tag{7}$$

Definition (6) is natural when $\mathbf{X}_1$ and $\mathbf{X}_2$ are stochastically independent (Oller 1989). Cuadras (1989, 1992a) proposed an extension for dependent variables.

2.2 Examples of Proximity Functions

Example 1. Let $\Pi = N_p(\mu,\Sigma)$ and $\delta^2(\mathbf{x},\mathbf{y}) = (\mathbf{x}-\mathbf{y})'\,\Sigma^{-1}\,(\mathbf{x}-\mathbf{y})$, for $\mathbf{x},\mathbf{y} \in \mathbf{R}^p$, the Mahalanobis distance. Then the transformation $\psi(\mathbf{x}) = \Sigma^{-1/2}\,\mathbf{x}$, taking values in $\mathbf{R}^p$ with its ordinary Euclidean scalar product, provides the proximity function

$$\phi^2(\mathbf{x}_0,\Pi) = (\mathbf{x}_0-\mu)'\,\Sigma^{-1}\,(\mathbf{x}_0-\mu)\,.$$

A generalization of this formula is given in Section 3.2.

Example 2. Let X be a random variable with values on an interval $I \equiv (a,b) \subseteq \mathbf{R}^* = \mathbf{R} \cup \{-\infty\} \cup \{\infty\}$. Denote its c.d.f. by F, and consider the distance $\delta(x,y) = (\,|\,x-y\,|\,)^{1/2}$. In this case the following identity holds (Cuadras and Fortiana 1995)

$$V_\delta(F) = \textstyle\int_a^b F(x)\,[1-F(x)]\,dx\,. \tag{8}$$

If Y is another random variable with values on I, with c.d.f G, and independent of X, a straightforward computation, using (3), gives

$$\Delta^2(F,G) = \textstyle\int_a^b [F(x)-G(x)]^2\,dx\,, \tag{9}$$

that is, the distance (2) between populations associated with $\delta(x,y) = (\,|\,x-y\,|\,)^{1/2}$ is the Cramér-von Mises distance.

Taking, in (9), G as the degenerate probability distribution with unit mass at $x_0 \in (a,b)$, we obtain the proximity function

$$\phi^2(x_0,F) = \textstyle\int_a^b [F(x)]^2\,dx - \int_{x_0}^b [2F(x)-1]\,dx\,, \tag{10}$$

which has a minimum for x_0 such that $2F(x_0)-1=0$, i.e., the median of X is the δ-center of F. Note that if in (10) we let F be the degenerate probability distribution with unit mass at $y_0 \in (a,b)$, we recover the squared distance $|\,x_0-y_0\,|$.

An explicit Euclidean representation is available when F is a uniform $(0,1)$ distribution. Let $L = l^2$, the Hilbert space of square summable sequences $\{x_n\}_{n\,\in\,\mathrm{N}}$, with its standard inner product $\Sigma_{i=1}^\infty x_i y_i$. The mapping $\psi : (0,1) \rightarrow L$ given by

$$\psi(x) = \frac{\sqrt{2}}{\pi}\left[\frac{1}{1}\cos(\pi x), \frac{1}{2}\cos(2\pi x), \frac{1}{3}\cos(3\pi x), \ldots\right], \quad x \in (0,1),$$

provides a Euclidean representation of a uniform $(0,1)$ random variable X in L, i.e.,

$$\delta^2(x,y) = |\,x-y\,| = \frac{2}{\pi^2}\sum_{j=1}^{\infty}\frac{(\cos j\pi x - \cos j\pi y)^2}{j^2}, \quad x,y \in (0,1).$$

The random variables $Z_j = (\sqrt{2}\cos(j\pi X))/(j\pi)$ have mean 0 and are uncorrelated, therefore the sequence $\{Z_j\}_{j\in \mathrm{N}}$ can be interpreted as a (countable) principal coordinate solution for the distance $\delta(x,y) = (\,|\,x-y\,|\,)^{1/2}$. See Cuadras and Fortiana (1993, 1995) for details and an additional example.

Computing in (10) the proximity function, we obtain $\phi^2(x_0,F) = x_0^2 - x_0 + 1/3$, $x_0 \in (0,1)$. As $E(\psi(X)) = (0,0,\ldots)$, the equality (5c) is equivalent to

$$x_0^2 - x_0 + \frac{1}{3} = \frac{2}{\pi^2}\sum_{j=1}^{\infty}\left[\frac{\cos j\pi x_0}{j}\right]^2,$$

which can be checked directly, by expanding $x_0^2 - x_0 + \dfrac{1}{3}$ in Fourier cosine series on $(0,1)$ and using the identity $\cos(2t) = 2\cos^2 t - 1$.

3. Distance-based Discrimination

The motivation underlying the construction of (4), and its most important application so far, is to provide a theoretical framework for the distance-based allocation rule for discrimination (briefly DB rule) proposed by Cuadras (1989).

Suppose that we have g populations Π_α, $\alpha = 1, \ldots, g$. Let ϕ_α^2 be the proximity function for Π_α, computed by using the squared dissimilarity between observations δ_α^2, $\alpha = 1, \ldots, g$.

The DB rule for allocating an individual for which $\mathbf{x}_0$ has been observed is:

$$\text{Allocate } \mathbf{x}_0 \text{ to } \Pi_\alpha \text{ if } \phi_\alpha^2(\mathbf{x}_0) = \min\{\phi_1^2(\mathbf{x}_0), \ldots, \phi_g^2(\mathbf{x}_0)\}. \qquad (11)$$

Theorem 1-(5c) shows that this DB rule, although computed on dissimilarities between observations, is in fact a Matusita rule, i.e., a rule based on the distance between an observation and a ''mean'' of the population, at least when a Euclidean representation exists.

One outstanding feature of the DB discriminant rule is its adaptability to different types of data. By adjusting δ, the data analyst can reflect in the model properties of the problem like scales of measurement, weights and relationships of variables.

Table 1: Proximity functions and the DB rule for some statistical models and classic equivalents.

Statistical Model (for Π_α)	Squared Distance $\delta^2_\alpha(\mathbf{x},\mathbf{y})$	Proximity function $\phi^2_\alpha(\mathbf{x}_0)$ =	Equivalent to
Multivariate Bernoulli with m states $E_1,\ldots,E_m$ $f_\alpha(\mathbf{x})=\prod_{k=1}^m q_{\alpha k}^{x_k},\ x_k\in\{0,1\}$ $\sum_{k=1}^m x_k=1$	$(1-\delta_{kl})\left(\frac{1}{q_{\alpha k}}-\frac{1}{q_{\alpha l}}\right)$ If $(\mathbf{x},\mathbf{y})$ falls in (E_k,E_l)	$\frac{1-q_{\alpha k}}{q_{\alpha k}}$ If $x_{0k}=1$	ML
Multinomial $f_\alpha(\mathbf{x})=\frac{n!}{\prod_{k=1}^m x_k!}\prod_{k=1}^m q_k^{x_k}$ $\sum_{k=1}^m x_k=n$	$\frac{1}{n}\sum_{k=1}^m\frac{(x_k-y_k)^2}{q_{\alpha k}}$	$\sum_{k=1}^m\frac{(x_{0k}-n\,q_{\alpha k})^2}{n\,q_{\alpha k}}$	Chi–Square
Normal, $\Sigma_1=\Sigma_2=\mathbf{I}$ $N_p(\mu_\alpha,\mathbf{I})$	Euclidean	$(\mathbf{x}_0-\mu_\alpha)'(\mathbf{x}_0-\mu_\alpha)$	EDF
Normal, $\Sigma_1=\Sigma_2=\Sigma$ $N_p(\mu_\alpha,\Sigma)$	Mahalanobis	$(\mathbf{x}_0-\mu_\alpha)'\Sigma^{-1}(\mathbf{x}_0-\mu_\alpha)$	LDF
Normal, $\Sigma_1\neq\Sigma_2$ $N_p(\mu_\alpha,\Sigma_\alpha)$	Mahalanobis + additive constant	$(\mathbf{x}_0-\mu_\alpha)'\Sigma_\alpha^{-1}(\mathbf{x}_0-\mu_\alpha)$ $+\log\lvert\Sigma_\alpha\rvert$	QDF
Any regular model $f_\alpha(\mathbf{x},\theta)$	Score distance	$\mathbf{Z}_0'\,\mathbf{G}^{-1}\,\mathbf{Z}_0$	

3.1 Classic Discriminant Functions

Using appropriate distances, (11) reduces to some classic and well studied rules (see Table 1).

Linear Discriminant. If Π_α is $N_p(\mu_\alpha,\Sigma_\alpha)$, $\alpha = 1,2$, with $\Sigma_1 = \Sigma_2$, and taking the Mahalanobis distance (Example 1), the halved difference

$$L(\mathbf{x}_0)=\frac{1}{2}\,[\phi_2^2(\mathbf{x}_0)-\phi_1^2(\mathbf{x}_0)] \tag{12}$$

is equal to the linear discriminant function (LDF; Lachenbruch 1975, pp. 9-11). Therefore, (11) coincides in this case with the LDF rule.

Quadratic Discriminant. If Π_α is $N_p(\mu_\alpha,\Sigma_\alpha)$, $\alpha = 1,2$, with $\Sigma_1 \neq \Sigma_2$, then we take the squared distance

$$\delta^2_\alpha(\mathbf{x},\mathbf{y}) = \begin{cases} (\mathbf{x}-\mathbf{y})' \Sigma_\alpha^{-1}(\mathbf{x}-\mathbf{y}) + \log \mid \Sigma_\alpha \mid , & \text{if } \mathbf{x} \neq \mathbf{y}\,, \\ 0\,, & \text{if } \mathbf{x} = \mathbf{y}\,, \end{cases} \tag{13}$$

to compute $\phi^2_\alpha(\mathbf{x}_0)$, $\alpha = 1,2$. i.e., the Mahalanobis distance plus a constant (without loss of generality we can suppose $\mid \Sigma_\alpha \mid \geq 1$; otherwise an arbitrary constant can be added). Now the halved difference of proximity functions

$$Q(\mathbf{x}_0) = \frac{1}{2}\,[\phi^2_2(\mathbf{x}_0) - \phi^2_1(\mathbf{x}_0)]\,, \tag{14}$$

is equal to the quadratic discriminant function (QDF; Lachenbruch 1975, p.20). Note that in (13) we have only added a constant to a squared distance, a well-known procedure in multidimensional scaling (see Lingoes 1971, Mardia 1978).

Euclidean Discriminant. A particular case is obtained when $\Sigma_1 = \Sigma_2 = \mathbf{I}$, where $\mathbf{I}$ is the identity matrix. Then δ is the usual Euclidean distance and the DB rule provides the Euclidean discriminant function (EDF)

$$E(\mathbf{x}_0) = [\mathbf{x}_0 - \frac{1}{2}\,(\mu_1 + \mu_2)]'\,(\mu_1 - \mu_2)\,. \tag{15}$$

This function has been studied by Marco et al. (1987), and has advantages when the number of variables is large in relation to the training sample size.

3.2 Generalized Discriminant Functions

More generally, if we have a regular statistical model $\{f(\mathbf{x};\theta);\ \mathbf{x} \in S \subset \mathbf{R}^p, \theta \in \Theta \subset \mathbf{R}^k\}$, and $\theta = \theta_\alpha$ for Π_α, $\alpha = 1,2$, an appropriate distance is based on Rao's *efficient scores* (Rao 1973, pp.367). Writing $\mathbf{z} = \frac{\partial}{\partial\,\theta} \log f(\mathbf{x};\theta)$, for $\mathbf{x} \in S$, and similarly $\mathbf{Z} = \frac{\partial}{\theta} \log f(\mathbf{X};\theta)$, for an S-valued random vector, this distance is defined by

$$\delta^2(\mathbf{x}_1,\mathbf{x}_2) = (\mathbf{z}_1 - \mathbf{z}_2)'\,\mathbf{G}_\theta^{-1}\,(\mathbf{z}_1 - \mathbf{z}_2)\,, \tag{16}$$

where $\mathbf{G}_\theta = E(\mathbf{Z}\,\mathbf{Z}')$ is the Fisher information matrix. Then, as $E(\mathbf{Z}'\,\mathbf{G}_\theta^{-1}\mathbf{Z}) = E(\mathrm{tr}\,(\mathbf{G}_\theta^{-1}\mathbf{Z}\,\mathbf{Z}')) = \mathrm{tr}\,(\mathbf{G}_\theta^{-1}\mathbf{G}_\theta) = k$, we easily find the discriminant function

$$\phi^2(\mathbf{x}_0) = \mathbf{z}'_0\,\mathbf{G}_\theta^{-1}\mathbf{z}_0\,. \tag{17}$$

The functions for the multinomial and multinormal ($\Sigma_1 = \Sigma_2$) models (Table 1) are particular cases of (17).

3.3 The Bayesian DB Discriminant Rule

If probability densities $\{f_\alpha(\mathbf{x})\}_{\alpha=1,\ldots,g}$ (with respect to a suitable common measure λ), and prior probabilities $\{q_\alpha\}_{\alpha=1,\ldots,g}$, for $\{\Pi_\alpha\}_{\alpha=1,\ldots,g}$ are given, the classic Bayes discriminant rule (BR) is

$$\text{Allocate } \mathbf{x}_0 \text{ to } \Pi_\beta \text{ if } q_\beta f_\beta(\mathbf{x}_0) = \max_{\alpha=1,\ldots,g} \{q_\alpha f_\alpha(\mathbf{x}_0)\}\,. \tag{18}$$

A Bayes version of the distance-based rule, which we indicate by BDB, is constructed as follows: For the multivariate Bernoulli model with parameter vector $(q_1,\ldots,q_g)$, we have the proximity functions (see Table 1)

$$\phi[P]_\alpha^2 = q_\alpha^{-1} - 1, \quad \alpha = 1,\ldots,g\,. \tag{19}$$

In the absence of additional information, the "prior" DB rule is equivalent to allocating the new individual to Π_β, where β is such that q_β is maximum. Since prior and observed information may be understood to be independent of each other, the proximity functions ϕ_α^2 related to the variables are combined with $\phi[P]_\alpha^2$ by using the additive property (7), obtaining the posterior proximity functions

$$\phi[B]_\alpha^2 = \phi_\alpha^2 + q_\alpha^{-1} - 1, \quad \alpha = 1,\ldots,g\,. \tag{20}$$

Clearly, for the multinomial and multivariate normal model, BR is equivalent to BDB if $q_1 = \ldots = q_g = 1/g$. Cuadras (1991, 1992b) evaluated the difference between the two rules in some particular cases, concluding that the discrepancy is not too marked on a broad interval of prior probabilities, i.e., excluding the most extreme values.

4. Sample Properties of Proximity Functions

In applied problems, the distance function is typically a datum, but the probability distributions for each population are unknown, or known only up to parameters. Natural estimators of V_{δ_α}, Δ^2 and ϕ_α^2, given samples $\mathbf{x}(\alpha)_1,\ldots,\mathbf{x}(\alpha)_{n_\alpha}$ of Π_α, $\alpha = 1,\ldots,g$, and an additional observation $\mathbf{x}_0$ to be classified, are:

$$\hat{V}_{\delta_\alpha}(\alpha) = \frac{1}{2n_\alpha^2} \sum_{ij} \delta_\alpha^2(\mathbf{x}(\alpha)_i, \mathbf{x}(\alpha)_j)\,, \tag{21a}$$

$$\hat{\Delta}^2(\Pi_\alpha,\Pi_\beta) = \frac{1}{n_\alpha n_\beta} \sum_{ij} \delta^2(\mathbf{x}(\alpha)_i, \mathbf{x}(\beta)_j) - \hat{V}_\delta(\alpha) - \hat{V}_\delta(\beta)\,, \tag{21b}$$

$$\hat{\phi}_\alpha^2(\mathbf{x}_0) = \frac{1}{n_\alpha} \sum_i \delta_\alpha^2(\mathbf{x}_0, \mathbf{x}(\alpha)_i) - \hat{V}_\delta(\alpha)\,. \tag{21c}$$

Theorem 2 below is a sampling analog of Theorem 1. Like the latter, it gives support to the interpretation of the DB discriminant rule as a Matusita (i.e., a minimum distance) rule. It should be emphasized that the Euclidean coordinates which appear in its statement and proof are only a theoretical scaffolding: its explicit computation is not generally needed in actual applications. In particular, DB allocations need only the dissimilarities between observations, which are used to compute the estimations (21c) of proximity functions.

Let $\mathbf{D}^{(2)}(\alpha) = (\delta_\alpha^2(\mathbf{x}(\alpha)_i, \mathbf{x}(\alpha)_j))$ be the $n_\alpha \times n_\alpha$ matrix of squared intradistances between observations in Π_α, and let $\mathbf{U}(\alpha)$ be a matrix of complex Euclidean coordinates for Π_α. That is, the n_α rows $u_i(\alpha)$ in $\mathbf{U}(\alpha)$ are the vectors representing the individuals from Π_α in some $L_\alpha = \mathbf{R}^{s_\alpha} \times i\mathbf{R}^{t_\alpha}$, where $i = \sqrt{-1}$, $s_\alpha > 0$, $t_\alpha \geq 0$, $(\alpha = 1, \ldots, g)$.

The principal coordinates solution for $\mathbf{D}^{(2)}(\alpha)$ gives one such representation, with the additional property that the centroid $\bar{\mathbf{u}}(\alpha) = (1/n_\alpha)\Sigma \mathbf{u}_i(\alpha)$ is the null vector. This assumption is too restrictive in the present context: As observed above (see Remark 1), a global distance function is required for (23b), and a global principal coordinates solution will be accordingly built in the proof. In it, the centroids $\bar{\mathbf{u}}(\alpha)$, $\alpha = 1, \ldots, g$, are not null. However, for any Euclidean representation, (Gower 1982, Eq.1), there exists an $n_\alpha \times 1$ vector $\mathbf{b}(\alpha)$ such that

$$\mathbf{U}(\alpha)\mathbf{U}(\alpha)' = -\frac{1}{2}\mathbf{D}^{(2)}(\alpha) + \mathbf{1}_{n_\alpha}\mathbf{b}(\alpha)' + \mathbf{b}(\alpha)\mathbf{1}'_{n_\alpha}, \tag{22}$$

where $\mathbf{1}_{n_\alpha}$ is the $n_\alpha \times 1$ vector of 1's.

Theorem 2. *For* $1 \leq \alpha, \beta \leq g$,

$$\hat{V}_{\delta_\alpha}(\alpha) = \frac{1}{n_\alpha}\sum_{i=1}^{n_\alpha}\|\mathbf{u}_i(\alpha)\|^2 - \|\bar{\mathbf{u}}(\alpha)\|^2, \tag{23a}$$

$$\hat{\Delta}^2(\Pi_\alpha, \Pi_\beta) = \|\bar{\mathbf{u}}(\alpha) - \bar{\mathbf{u}}(\beta)\|^2, \tag{23b}$$

$$\hat{\phi}_\alpha^2(\mathbf{x}_0) = \|\mathbf{u}_0(\alpha) - \bar{\mathbf{u}}(\alpha)\|^2, \tag{23c}$$

where $\mathbf{u}_0(\alpha)$ *is a vector representing* $\mathbf{x}_0$ *in* L_α.

Proof. From (22),

$$\hat{V}_{\delta_\alpha}(\alpha) = \frac{1}{2n_\alpha^2}\mathbf{1}'_{n_\alpha}\mathbf{D}^{(2)}(\alpha)\mathbf{1}_{n_\alpha}$$

$$= \frac{2}{n_\alpha}\mathbf{1}'_{n_\alpha}\mathbf{b}(\alpha) - \frac{1}{n_\alpha^2}\mathbf{1}'_{n_\alpha}\mathbf{U}(\alpha)\mathbf{U}(\alpha)'\mathbf{1}_{n_\alpha}$$

$$= \frac{2}{n_\alpha}\mathbf{1}'_{n_\alpha}\mathbf{b}(\alpha) - \|\bar{\mathbf{u}}(\alpha)\|^2 ,$$

and using (22) again, since tr $\mathbf{D}^{(2)}(\alpha) = 0$,

$$\frac{1}{n_\alpha}\sum_{i=1}^{n_\alpha}\|\mathbf{u}_i(\alpha)\|^2 = \frac{1}{n_\alpha}\mathrm{tr}\,[\mathbf{U}(\alpha)\mathbf{U}(\alpha)'] = \frac{2}{n_\alpha}\mathbf{1}'_{n_\alpha}\mathbf{b}(\alpha) ,$$

which gives (23a). Cuadras (1989) proved result (23c), by extending a result of Gower (1968). A more straightforward proof follows from:

$$\frac{1}{n_\alpha}\sum_{i=1}^{n_\alpha}\delta_\alpha^2(\mathbf{x}_0,\mathbf{x}(\alpha)_i) = \frac{1}{n_\alpha}\sum_{i=1}^{n_\alpha}\|\mathbf{u}_0(\alpha) - \mathbf{u}_i(\alpha)\|^2$$

$$= \|\mathbf{u}_0(\alpha)\|^2 - 2\mathbf{u}_0(\alpha)'\bar{\mathbf{u}}(\alpha) + \frac{1}{n_\alpha}\sum_{i=1}^{n_\alpha}\|\mathbf{u}_i(\alpha)\|^2 ,$$

taking into account (23a).

Result (23b) is due to Digby and Gower (1981). See also Cuadras (1991). The proof below follows that of Digby and Gower: Let $\mathbf{D}^{(2)}$ be the full $n \times n$ matrix of squared distances between the $n = \Sigma_{\alpha=1}^{g} n_\alpha$ sample observations. Consider its principal coordinate solution $\mathbf{B} = \mathbf{H}(-1/2\mathbf{D}^{(2)})\mathbf{H} = \mathbf{U}\mathbf{U}'$, where $\mathbf{H} = \mathbf{I}_n - 1/n\mathbf{1}_n\mathbf{1}'_n$ is the $n \times n$ centering matrix.

Given $1 \le \alpha \le g$, denote by $\mathbf{1}(\alpha)$ the $n \times 1$ column vector containing 1's at the n_α positions corresponding to the Π_α sample, and 0's elsewhere. For $1 \le \alpha,\beta \le g$,

$$\bar{\mathbf{u}}(\alpha)\bar{\mathbf{u}}(\beta)' = \frac{1}{n_\alpha,n_\beta}\mathbf{1}(\alpha)'\mathbf{U}\mathbf{U}'\mathbf{1}(\beta)$$

$$= \frac{1}{n_\alpha,n_\beta}\mathbf{1}(\alpha)'\mathbf{B}\mathbf{1}(\beta) - \frac{1}{2n_\alpha,n_\beta}\mathbf{1}(\alpha)'\mathbf{H}\mathbf{D}^{(2)}\mathbf{H}\mathbf{1}(\beta) .$$

Since $\mathbf{H}\mathbf{1}(\alpha) = \mathbf{1}(\alpha) - (n_\alpha/n)\mathbf{1}_n$, and similarly for $\mathbf{1}(\beta)$, we have

$$-2n_\alpha n_\beta\bar{\mathbf{u}}(\alpha)\bar{\mathbf{u}}(\beta)' = \mathbf{1}(\alpha)'\mathbf{D}^{(2)}\mathbf{1}(\beta) - \frac{n_\alpha}{n}\mathbf{1}'_n\mathbf{D}^{(2)}\mathbf{1}(\beta)$$

$$- \frac{n_\beta}{n}\mathbf{1}'_n\mathbf{D}^{(2)}\mathbf{1}(\alpha) + \frac{n_\alpha n_\beta}{n}\mathbf{1}'_n\mathbf{D}^{(2)}\mathbf{1}_n .$$

Substituting in $\|\bar{\mathbf{u}}(\alpha) - \bar{\mathbf{u}}(\beta)\|^2 = \bar{\mathbf{u}}(\alpha)\bar{\mathbf{u}}(\alpha)' + \bar{\mathbf{u}}(\beta)\bar{\mathbf{u}}(\beta)' - 2\bar{\mathbf{u}}(\alpha)\bar{\mathbf{u}}(\beta)'$ and simplifying, gives (23b). ■

4.1 Error Rates

The probability of misallocation for the DB discriminant rule can be estimated by several methods (McLachlan 1992, Chap.10). The *leave-one-out* estimator (Lachenbruch 1975) will be used in Section 5 for comparative purposes. It is worth noting that the amount of computation required is surprisingly small if the distance function does not contain parameters to be estimated from the sample, as is the case with the Minkowski distances, or with the distance considered in the second example in Section 5 (the square root of a Minkowski distance).

This statement can be verified as follows: Given a sample $\{\mathbf{x}_i\}_{i=1,\dots,n}$ of $n = \Sigma_{\alpha=1}^{g} n_\alpha$ observations, assume that $\mathbf{x}_k$ is an observation belonging to Π_α, and denote by $\hat{\phi}[k]^2_\beta$ the estimated proximity between $\mathbf{x}_k$ and Π_β, $(1 \leq \beta \leq g)$, evaluated from the sample minus the k-th observation. Then,

$$\hat{\phi}[k]^2_\beta = c_{\alpha\beta}\hat{\phi}^2_\beta(\mathbf{x}_k), \text{ where } c_{\alpha\beta} = \begin{cases} 1, & \text{if } \beta \neq \alpha, \\ (n_\alpha/(n_\alpha - 1))^2, & \text{if } \beta = \alpha. \end{cases} \tag{24}$$

This equality is clear for $\beta \neq \alpha$, and for $\beta = \alpha$, writing $\hat{\phi}^2_\alpha(\mathbf{x}_k)$ in (21c) as

$$\hat{\phi}^2_\alpha(\mathbf{x}_k) = \frac{1}{n_\alpha}\sum_{i=1}^{n_\alpha} \delta^2_\alpha(\mathbf{x}_k,\mathbf{x}_i) - \frac{1}{2n^2_\alpha}\sum_{i,j=1}^{n_\alpha} \delta^2_\alpha(\mathbf{x}_i,\mathbf{x}_j) = A/n_\alpha - B/(2n^2_\alpha),$$

removal of the k-th observation from the training sample amounts to: 1) Substituting $(n_\alpha - 1)$ for n_α in the denominators, and 2) Eliminating the k-th row and column from the matrix of squared distances. We obtain the equality

$$\hat{\phi}[k]^2_\alpha = A/(n_\alpha - 1) - (B - 2A)/(2(n_\alpha - 1)^2)$$

which, after simplification, yields (24). Then the estimation e of the probability of misallocation easily follows.

It is worth noting that the DB discriminant rule is robust to non-overlapping samples. It can easily be seen that, in fact, $e = 0$ in such situations.

4.2 DB Canonical Variate Analysis

If $g > 2$, we can construct a $g \times g$ matrix $\hat{\Delta}^{(2)} = (\hat{\Delta}^2(\Pi_\alpha,\Pi_\beta))$ of squared distances between populations and perform a metric scaling to exhibit them in a low-dimensional Euclidean space.

This useful representation originated in Digby and Gower's (1981) paper. The same construction is given in a more general form by Cuadras (1991), and is applied in Fortiana et al. (1995). Fortiana (1993) and Krzanowski (1994a) also follow this approach, making a link with classic canonical variate analysis.

More precisely, the classic representation appears as a particular case, when the distance between individuals is the Mahalanobis distance, for (23b) implies that, in this case, the distance between groups α and β is

$$\hat{\Delta}^{(2)}(\Pi_\alpha,\Pi_\beta) = (\bar{\mathbf{x}}_\alpha - \bar{\mathbf{x}}_\beta)\mathbf{S}^{-1}(\bar{\mathbf{x}}_\alpha - \bar{\mathbf{x}}_\beta)',$$

where **S** is the pooled within-groups sample covariance matrix, and the sample means $\bar{\mathbf{x}}_\alpha$ and $\bar{\mathbf{x}}_\beta$ are written as row vectors to keep notation consistent with that of (23b). The conclusion now follows from the well-known result of Gower's (1966) which states the equivalence of principal coordinates and canonical variates for the Mahalanobis distance function.

Of course, this method is especially interesting for non-Mahalanobis distances, and for mixed data. Another possibility, as observed by Krzanowski (1994a) lies in the representation of individual points in the resulting diagram, by using Gower's (1968) *adding-a-point* formula.

The possibility of this representation is a consequence of (23b) and (23c). Firstly, (23b) allows us to interpret $\hat{\Delta}^{(2)}$ as a matrix of squared distances between the population centroids, ensuring its Euclideanarity provided that $\mathbf{D}^{(2)}$ (notation as in the proof of (23b)) has this property. Second, equality (23c) allows us to interpret the vector $(\hat{\phi}_1^2(\mathbf{x}_0), \ldots, \hat{\phi}_g^2(\mathbf{x}_0))'$ of sample proximity functions from a given $\mathbf{x}_0$ to the populations as the vector $\mathbf{d}^{(2)}$ of squared Euclidean distances from $\mathbf{x}_0$ to the centroids of the populations, needed in Gower's formula

$$\mathbf{u}_0 = \frac{1}{2}(\mathbf{Y}'\mathbf{Y})^{-1}\mathbf{Y}'(\mathbf{c} - \mathbf{d}^{(2)}),$$

which gives a vector $\mathbf{u}_0$ representing $\mathbf{x}_0$ in the space L of principal coordinates of $\hat{\Delta}^{(2)}$. **Y** is the matrix having as rows the vectors representing $\Pi_1, \ldots, \Pi_g$ in L, and **c** is the $g \times 1$ vector having as entries the diagonal elements of $\mathbf{Y}\mathbf{Y}'$.

5. Four Comparative Real Data Examples

We used several real data sets to compare the DB method with linear (LDF), quadratic (QDF) and Euclidean (EDF) discriminant functions. As discussed in Section 3.1, LDF, QDF and EDF can be interpreted as particular cases of DB by using appropriate distances (namely, Mahalanobis, Mahalanobis plus an additive term and Euclidean distances, respectively). In this section we adopt the distance based on Gower's similarity coefficient to carry out the DB discrimination.

Data set 1, *DNA data,* (Fortiana et al. 1995) consists of sequences of length 360 *bp* (base pairs) taken from a given segment (called the *I*-region from the *D*-loop) of the mitochondrial DNA for a set of 120 individuals

Table 2: Total number of observations in each population and number of misallocations

DNA data

	Π_1	Π_2	Π_3	Π_4	Total
Number of observations	25	41	37	17	120
Misallocations					
DB	2	2	1	7	12
EDF	2	16	0	7	25
wDB	1	3	2	7	13

Skulls data

	Π_1	Π_2	Π_3	Total
Number of observations	30	30	30	90
Misallocations				
DB	11	16	15	42
LDF	10	18	15	43
QDF	10	24	10	44
EDF	11	19	14	44

Cancer data

	Π_1	Π_2	Total
Number of observations	78	59	137
Misallocations			
DB	18	21	39
LDF	31	27	58
QDF	13	35	48
EDF	29	37	66
LM	21	24	45
QLM	23	20	43

Students data

	Π_1	Π_2	Π_3	Total
Number of observations	25	117	26	168
Misallocations				
DB	12	69	16	97
LDF	11	85	13	109
QDF	10	98	9	117
EDF	10	82	11	103
LM	10	88	12	110
QLM	13	84	9	106

belonging in 4 human groups: !Kung (Π_1), Bantu-speaking Africans (Π_2), Pygmies (Π_3) and Hadza (Π_4). These data are extracted from a larger set, studied in Bertranpetit et al. (1995), where precise bibliographic references for the sources can be obtained.

Since all variables are four-state qualitative, with values on the set of bases

$$\{ \text{Adenine, Guanine, Cytosine, Thymine} \},$$

they should be converted into $1080 = 3 \times 360$ dummy variables (Lachenbruch 1975, p.54), in order to apply LDF or QDF, but then the sample size is insufficient to give a nonsingular covariance matrix, so neither method can be used. EDF can be applied, but results depend on the actual coding of each base as three values of the dummy indicator variables. A typical result is included for comparison in Table 2. A third method (wDB), using a weighted modification of the distance based on Gower's similarity coefficient, has been considered for this data set, where weights are inversely proportional to the probabilities of transition between each pair of bases.

Data set 2, *Skulls data*, comes from Manly (1986, Table 1.2, p.4-5), and reports four biometric measurements on male Egyptian skulls from five epochs (Early predynastic, Late predynastic, 12-th & 13-th dynasties, Ptolemaic period and Roman period). Canonical variate analysis reveals that some groups overlap, so the discrimination is made taking only the three separate groups: Early predynastic (Π_1), 12-th & 13-th dynasties (Π_2), and Roman period (Π_3), $n_1 = n_2 = n_3 = 30$.

Data set 3, *Cancer data*, consists of eleven measurements (seven continuous, two binary and two three-state categorical variables) on 137 women with breast tumours, 78 benign (Π_1), and 59 malignant (Π_2). These data are described in Krzanowski (1980) and as they involve mixed variables, we also perform a comparison with the location model for discrimination (LM) and with the quadratic location model (QLM) (Krzanowski 1994b).

Data set 4, *Students data*, taken from Mardia et al. (1979, p.294), is also used by Krzanowski (1983). This data set is concerned with the average grade (a single quantitative variable) and a qualitative variable with three states: 2, 3 or 4 A-levels, obtained by 382 students who were classified in seven groups. The number of misallocations for the complete data are high: LDF (313), QDF (319), EDF (312), DB (290), LM (310), as the 7 groups overlap somewhat. So, we select 3 more separate groups, i.e., as denoted by Mardia et al. (1979), the following groups: 'I' (Π_1), 'II(ii)' (Π_2) and '$\rightarrow$' (Π_3) respectively, the sample sizes being $n_1 = 25, n_2 = 117, n_3 = 26$.

Error rates were computed using Lachenbruch's leave-one-out procedure, as described in Section 4.1, and are given in Table 2.

Our understanding of these results is as follows: For data set 1, LDF cannot work, as explained above. For data set 2, LDF works well because Mahalanobis distance is appropriate. However, for data set 4, the better performance of DB than LDF may be interpreted as suggesting that the distance based on Gower's similarity coefficient is more appropriate than Mahalanobis distance. In other words, the probability model implicit in using the first distance fits the data better than the multivariate normal. Coincidentally, Mardia et al. (1979, p.293) stated that "the assumptions of normality would be completely unwarranted in this example." A similar reasoning could explain why the location model is not the best for data set 3.

6. Discussion

We have studied a proximity function which can be interpreted as a squared distance between an individual and a population. It is constructed using only distances $\delta(\cdot,\cdot)$ between individuals. The advantage of this approach appears for mixed variables, where it is difficult to construct a probabilistic model, while a proximity function is more accessible.

Takane et al. (1987) also use a proximity function defining a Euclidean distance between subject points and ideal points representing populations. They assume a probability distribution which depends on this proximity function. The construcion of (4) is quite different, as it can provide a non-Euclidean distance, which is not an obstacle for discrimination, and does not require an underlying probabilistic model.

For the multivariate normal distribution $N_p(\mu,\Sigma)$, the natural proximity function ϕ^2, i.e., the Mahalanobis distance between $\mathbf{x}$ and μ, is monotonically related to the probability density function $f(\mathbf{x})$. An open problem is to decide whether this relationship can be extended to other cases.

If we allow in (4) any symmetric function $s(\cdot,\cdot)$ instead of a squared dissimilarity $\delta^2(\cdot,\cdot)$, the problem has a solution for every probability density $f(\mathbf{x})$, taking

$$s(\mathbf{x},\mathbf{y}) = -\log f(\mathbf{x}) - \log f(\mathbf{y})\,.$$

Then, $\phi^2(\mathbf{x}) = -\log f(\mathbf{x})$ and the geometric variability is the Shannon entropy

$$H(f) = -\int f(\mathbf{x}) \log f(\mathbf{x})\, d\mathbf{x}\,.$$

However, it is not clear what geometric interpretation could be given to such generalized "dissimilarities."

The distance-based discrimination is the main application of the proximity function. The use of dissimilarities between individuals is useful for this method (Krzanowski 1993). The distance based on Gower's (1971) all-purpose coefficient has the additional advantage of dealing with missing data. However, like all distance functions satisfying additivity with respect to variables Gower (1992, 1993), it implicitly ignores any association (e.g., correlation) between variables (Krzanowski 1994a). Comparing the product of probability densities with the additive expression (6) in the case of independent variables, we find some connections with the construction of distributions with given marginals (Cuadras and Augé 1981, Cuadras 1992a). From this perspective, a distance between observations for dependent variables under a parametric model is proposed and studied in Cuadras (1989, 1991) and justified in Cuadras (1992a), but this is an open question when the context is general. The recommendation of Sneath and Sokal (1973, Section 4.7) concerning the use of distance functions which are as simple as possible can be taken into account, together with their performance in discrimination.

Finally, let us consider the relation between linear discrimination, Mahalanobis distance and canonical variate analysis. This relation is optimal under multivariate normality. Similarly, we can relate the location model for discrimination, Krzanowski's (1986) distance from an individual to a population and Krzanowski's (1983) representation of populations with mixed variables. This relation is also considered optimal under conditional normality (Olkin and Tate 1961). The generalization for any type of data (categorical, continuous, mixed) is the distance-based discrimination Cuadras (1989, 1991, 1992b), the proximity function (4) and the generalized ordination to represent the populations (Digby and Gower 1981, Krzanowski 1994a).

References

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

ANDERSON, T. W. (1966), ''Some Nonparametric Multivariate Procedures Based on Statistically Equivalent Blocks,'' in *Multivariate Analysis I*, Ed., P. R. Krishnaiah, New York: Academic Press, 5-27.

ARABIE, P. (1991), ''Was Euclid an Unnecessarily Sophisticated Psychologist? *Psychometrika, 56*, 567-587.

ATKINSON, C., and MITCHELL, A. F. S. (1981), ''Rao's Distance Measure,'' *Sankhya, The Indian Journal of Statistics, Series A, 43*, 345-365.

BERTRANPETIT, J., SALA, J., CALAFELL, F., UNDERHILL, P. A., MORAL, P., and COMAS, D. (1995), ''Human Mitochondrial DNA Variation and the Origin of Basques,'' *Annals of Human Genetics, 59*, 63-81.

BURBEA, J., and RAO, C. R. (1982), ''Entropy Differential Metric, Distance and Divergence Measures in Probability Spaces: a Unified Approach,'' *Journal of Multivariate Analysis, 12*, 575-596.

CUADRAS, C. M. (1989), ''Distance Analysis in Discrimination and Classification Using Both Continuous and Categorical Variables,'' in *Statistical Data Analysis and Inference*, Ed., Y. Dodge, Amsterdam: Elsevier Science Publishers B. V.

CUADRAS, C. M. (1991), ''A Distance-Based Approach to Discriminant Analysis and Its Properties,'' Math. Preprint Series 90, Univ. de Barcelona.

CUADRAS, C. M. (1992a), ''Probability Distributions With Given Multivariate Marginals and Given Dependence Structure,'' *Journal of Multivariate Analysis, 42*, 51-66.

CUADRAS, C. M. (1992b), ''Some Examples of Distance Based Discrimination,'' *Biometrical Letters, 29*, 3-20.

CUADRAS, C. M., and ARENAS, C. (1990), ''A Distance Based Regression Model for Prediction With Mixed Data,'' *Communications in Statistics A. Theory and Methods, 19*, 2261-2279.

CUADRAS, C. M., and AUGÉ, J. (1981), ''A Continuous General Multivariate Distribution and Its Properties,'' *Communications in Statistics A. Theory and Methods, 10*, 339-353.

CUADRAS, C. M., and FORTIANA, J. (1993), ''Continuous Metric Scaling and Prediction,'' in *Multivariate Analysis, Future Directions 2*, Eds., C.M. Cuadras and C.R. Rao, Amsterdam: Elsevier Science Publishers B.V., 47-66.

CUADRAS, C. M., and FORTIANA, J. (1995), ''A Continuous Metric Scaling Solution for a random variable,'' *Journal of Multivariate Analysis, 52,* 1-14.

CUADRAS, C. M., FORTIANA, J., and OLIVA, F. (1995), ''Representation of Statistical Structures, Classification and Prediction Using Multidimensional Scaling,'' in *From Data to Knowledge: Theoretical and Practical aspects of Classification, Data Analysis and Knowledge Organization,* Ed., W. Gaul, and D. Pfeifer, Berlin: Springer Verlag, 20-31.

CUADRAS, C. M., and RAO, C. R. (Eds.) (1993), *Multivariate Analysis, Future Directions 2,* Amsterdam: Elsevier Science Publishers B. V. (North-Holland).

DIGBY, P. G. N., and GOWER, J. C. (1981), ''Ordination Between- and Within-Groups applied to Soil classification,'' in *Down-to-Earth Statistics: Solutions looking for Geological Problems,* Ed., D. F. Merriam, Syracuse University Geology contributions, 63-75.

DODGE, Y. (Ed.) (1989), *Statistical Data Analysis and Inference,* Amsterdam: Elsevier Science Publishers B. V. (North-Holland).

FORTIANA, J. (1993), ''Distance-Based Approach to some Multivariate Statistical Methods (in Spanish),'' Ph.D. thesis, Universitat de Barcelona, Col · lecció de Tesis Doctorals Microfitxades núm. 1728, Publicacions de la Universitat de Barcelona, ISBN 84-475-0107-8.

FORTIANA, J., CUADRAS, C. M., and BERTRANPETIT, J. (1995), ''The Distance-based Approach in Discriminant Analysis for recognizing DNA Sequences,'' (submitted for publication).

GORDON, A. D. (1990), ''Constructing Dissimilarity Measures,'' *Journal of Classification, 7,* 257-269.

GOWER, J. C. (1966), ''A *Q*-technique for the Calculation of Canonical Variates,'' *Biometrika, 53,* 588-589.

GOWER, J. C. (1968), ''Adding a Point to Vector Diagrams in Multivariate Analysis,'' *Biometrika, 55,* 582-585.

GOWER, J. C. (1971), ''A General Coefficient of Similarity and Some of Its Properties,'' *Biometrics, 27,* 857-874.

GOWER, J. C. (1982), ''Euclidean Distance Geometry,'' *Math. Scientist, 7,* 1-14.

GOWER, J. C. (1992), ''Generalized Biplots,'' *Biometrika, 79,* 475-493.

GOWER, J. C. (1993), ''Recent Advances in Biplot Methodology,'' in *Multivariate Analysis, Future Directions 2,* Eds., C.M. Cuadras and C.R. Rao, Amsterdam: Elsevier Science Publishers, B.V., 295-325.

GOWER, J. C., and LEGENDRE, P. (1986), ''Metric and Euclidean Properties of Dissimilarity Coefficients,'' *Journal of Classification, 3,* 5-48.

KRZANOWSKI, W. J. (1975), ''Discrimination and Classification Using Both Binary and Continuous Variables,'' *Journal of the American Statistical Association, 70,* 782-790.

KRZANOWSKI, W. J. (1980), ''Mixtures of Continuous and Categorical Variables in Discriminant Analysis,'' *Biometrics, 36,* 493-499.

KRZANOWSKI, W. J. (1983), ''Distance Between Populations Using Mixed Continuous and Categorical Variables,'' *Biometrika, 70,* 235-243.

KRZANOWSKI, W. J. (1986), ''Multiple Discriminant Analysis in the Presence of Mixed Continuous and Categorical Data,'' *Computers and Mathematics with Applications, 12A,* 179-185.

KRZANOWSKI, W. J. (1987), ''A Comparison Between Two Distance-Based Discriminant Principles,'' *Journal of Classification, 4,* 73-84.

KRZANOWSKI, W. J. (1993), ''The Location Model for Mixtures of Categorical and Continuous Variables,'' *Journal of Classification, 10,* 25-49.

KRZANOWSKI, W. J. (1994a), ''Ordination in the Presence of Group Structure, for General Multivariate Data,'' *Journal of Classification, 11*, 195-207.

KRZANOWSKI, W. J. (1994b), ''Quadratic Location Discriminant Functions for Mixed Categorical and Continuous Data,'' *Statistics & Probability Letters, 19*, 91-95.

LACHENBRUCH, P. A. (1975), *Discriminant Analysis*, London: Hafner Press, Collier Macmillan Publishers.

LAU, K.-S. (1985), ''Characterization of Rao's Quadratic Entropies,'' *Sankhya, The Indian Journal of Statistics, Series A, 47*, 295-309.

LINGOES, J. C. (1971), ''Some Boundary Conditions for a Monotone Analysis of Symmetric Matrices,'' *Psychometrika, 36*, 195-203.

MANLY, B. F. J. (1986), *Multivariate Statistical Methods: A Primer*, London: Chapman and Hall.

MARCO, V. R., YOUNG, D. M., and TURNER, D. W. (1987), ''The Euclidean Distance Classifier: An Alternative to Linear Discriminant Function,'' *Communications in Statistics B. Simulation and Computation, 16*, 485-505.

MARDIA, K. V. (1978), ''Some Properties of Classical Multidimensional Scaling,'' *Communications in Statistics A. Theory and Methods, 7*, 1233-1241.

MARDIA, K. V., KENT, J. T., and BIBBY, J. M. (1979), *Multivariate Analysis*, London: Academic Press.

MATUSITA, K. (1956), ''Decision Rule, Based on the Distance, for the Classification Problem,'' *Annals of the Institute of Statistical Mathematics, 8*, 67-77.

MCLACHLAN, G. J. (1992), *Discriminant Analysis and Statistical Pattern Recognition*, John Wiley Sons, Inc., New York.

MIÑARRO, A., and OLLER, J. M. (1992), ''Some Remarks on the Individuals-Score Distance and Its Applications to Statistical Inference,'' *Qüestiió, 16*, 43-57.

MITCHELL, A. F. S. (1992), ''Estimative and Predictive Distances,'' *Test, 1*, 105-121.

OLKIN, I., and TATE, R. F. (1961), ''Multivariate Correlation Models With Mixed Discrete and Continuous Variables,'' *The Annals of Mathematical Statistics, 32*, 448-465, Correction in *Mathematical Statistics, 36*, 343-344.

OLLER, J. M. (1989), ''Some Geometrical Aspects of Data Analysis and Statistics,'' in *Statistical Data Analysis and Inference*, Ed., Y. Dodge, Amsterdam: Elsevier Science Publishers B.V., 41-58.

OLLER, J. M., and CUADRAS, C. M. (1985), ''Rao's Distance for Negative Multinomial Distributions,'' *Sankhya, The Indian Journal of Statistics, Series A, 47*, 75-83.

RAO, C. R. (1945), ''Information and the Accuracy Attainable in the Estimation of Statistical Parameters,'' *Bulletin of the Calcutta Mathematical Society, 37*, 81-91.

RAO, C. R. (1973), *Linear Statistical Inference and Its Applications, 2nd Ed.*, New York: John Wiley & Sons.

RAO, C. R. (1982a), ''Diversity and Dissimilarity Coefficients: A Unified Approach,'' *Theoretical Population Biology, 21*, 24-43.

RAO, C. R. (1982b), ''Diversity: Its Measurement, Decomposition, Apportionment and Analysis,'' *Sankhya, The Indian Journal of Statistics, Series A, 44*, 1-22.

SNEATH, P.H.A., and SOKAL, R. S. (1973), *Numerical Taxonomy. The Principles and Practice of Numerical Classification*, San Francisco, CA: W. H. Freeman and Co.

TAKANE, Y., BOZGODAN, H., and SHIBAYAMA, T. (1987), ''Ideal Point Discriminant Analysis,'' *Psychometrika, 52*, 371-392.