ELSEVIER

Statistics & Probability Letters 33 (1997) 405–411

STATISTICS & PROBABILITY LETTERS

Probability densities from distances and discrimination[1]

C.M. Cuadras[a,*], R.A. Atkinson[b], J. Fortiana[a]

[a] *Department of Statistics, University of Barcelona, Diagonal 645, 08028 Barcelona, Spain*
[b] *School of Mathematics and Statistics, University of Birmingham, Birmingham, UK*

Received March 1996; revised July 1996

Abstract

Given a population and a random vector $\boldsymbol{X}$, by using distances between observations of $\boldsymbol{X}$, we prove that it is, in general, possible to construct probability densities for $\boldsymbol{X}$. This distance-based approach can present problems, from a multidimensional scaling point of view, for some monotonic density functions, where the construction must be made on the basis of symmetric functions instead of distances. A measure of divergence between the true density and this construction is given. The procedure aims to offer alternative methods for performing discriminant analysis.

Keywords: Constructing densities; Discriminant function; Multidimensional scaling; Shannon entropy

1. Introduction

Let $\boldsymbol{\Pi}$ be a population represented by a random vector $\boldsymbol{X}$ with support $S \subset \mathbb{R}^p$ and probability density $f(\boldsymbol{x})$, $\boldsymbol{x} \in S$, with respect to a suitable measure λ. Suppose that a distance function $\delta(.,.)$ is defined on S, i.e., $\delta(\boldsymbol{x},\boldsymbol{x}) = 0$ and $\delta(\boldsymbol{x},\boldsymbol{y}) = \delta(\boldsymbol{y},\boldsymbol{x})$, $\boldsymbol{x},\boldsymbol{y} \in S$.

The *geometric variability* of $\boldsymbol{X}$ with respect to δ, introduced in Cuadras and Fortiana (1995), is given by

$$V_\delta(\boldsymbol{X}) = \tfrac{1}{2}\int_{S\times S} \delta^2(\boldsymbol{x},\boldsymbol{y}) f(\boldsymbol{x}) f(\boldsymbol{y}) \lambda(\mathrm{d}\boldsymbol{x}) \lambda(\mathrm{d}\boldsymbol{y}). \tag{1}$$

If δ is the ordinary Euclidean distance d in S, then we have $V_d(\boldsymbol{X}) = \mathrm{tr}(\mathrm{var}(\boldsymbol{X})) = \mathrm{tr}(\Sigma)$, i.e., (1) is the total variation of $\boldsymbol{X}$. In Section 4, $V_d(\boldsymbol{X})$ is understood as a measure of inaccuracy and is compared to the Shannon entropy.

Suppose that ω is an individual of $\boldsymbol{\Pi}$ with observations $\boldsymbol{X}(\omega) = \boldsymbol{x}_0$. Cuadras (1989) and Cuadras et al. (1996) introduced the *proximity function*

$$\phi_\delta^2(\boldsymbol{x}_0, \boldsymbol{\Pi}) = \int_S \delta^2(\boldsymbol{x}_0,\boldsymbol{y}) f(\boldsymbol{y}) \lambda(\mathrm{d}\boldsymbol{y}) - V_\delta(\boldsymbol{X}), \tag{2}$$

* Corresponding author.

[1] Work supported in part by grants CGYCIT PB93-0784 and SGR95-00085.

PII 0167-7152(96)00154-X

to be used in the problem of assigning ω to two possible populations Π_1, Π_2. The so-called distance based or DB discriminant rule is:

$$\text{Assign } \omega \text{ to } \Pi_i \quad \text{if } \phi_\delta^2(x_0, \Pi_i) = \min\{\phi_\delta^2(x_0, \Pi_1),\ \phi_\delta^2(x_0, \Pi_2)\}. \tag{3}$$

Given a sample of X, it is worth noting that $\phi_\delta^2(x, \Pi)$ can be estimated without knowing $f(x)$. Therefore, choosing a suitable distance, the sample version of rule (3) can be easily obtained.

Suppose now that (S, δ) is embedded in a Euclidean (or Hilbert) space L, i.e., there exists $\psi: R^p \to L$ such that $x \to \psi(x)$ satisfies

$$\delta(x, y) = d_L(\psi(x), \psi(y)), \tag{4}$$

where d_L is the Euclidean distance defined by a scalar product $\langle \cdot, \cdot \rangle$ in L. Examples of this embedding were given in Cuadras and Fortiana (1993, 1995). Cuadras et al. (1996) proved that

$$\phi_\delta^2(x_0, \Pi) = \|\psi(x_0) - E(\psi(X))\|_L^2, \tag{5}$$

measuring the proximity between x_0 and Π, and giving support to the interpretation of (3) as a distance-based discriminant rule.

The extension of rule (3) to $g > 2$ populations is easy. It is even feasible, by using the Jensen differences (Rao, 1982)

$$\Delta^2(\Pi_i, \Pi_j) = \int \delta^2(x_i, x_j) f_i(x_i) f_j(x_j) \lambda(\mathrm{d}x_i) \lambda(\mathrm{d}x_j) - V_\delta(X_i) - V_\delta(X_j),$$

to find distances between populations from distances between individuals, and to perform a graphic representation of the g populations via metric scaling, thus obtaining a generalization of canonical variate analysis (Krzanowski, 1994a).

The objective of this paper is to construct densities from distances or from symmetric functions, to discuss a related multidimensional scaling problem (Theorem 1), to compare the geometric variability with Shannon entropy (Theorem 2) and to apply this procedure to discriminant analysis.

2. Probability densities from a distance

From now on the proximity function will be indicated by $\phi_\delta^2(x)$. When δ is the Mahalanobis distance and $\Pi = \mathrm{N}_p(\mu, \Sigma)$, the proximity function (2) yields

$$\phi_\delta^2(x) = (x - \mu)' \Sigma^{-1} (x - \mu), \tag{6}$$

and (3) is equivalent to using the linear discriminant function (LDF). Cuadras (1992a) showed, by choosing appropriate distances, the advantages of rule (3) with categorical and mixed variables. See also Cuadras (1992b).

We can obtain the multivariate normal density $\mathrm{N}_p(\mu, \Sigma)$ from (6). In general, from any distance δ satisfying (4), we can construct a probability density defining

$$f_\delta(x) = c \exp(-\phi_\delta^2(x)/2) = c \exp\left(-\tfrac{1}{2} \|\psi(x) - E(\psi(X))\|_L^2\right), \tag{7}$$

where c is a normalizing constant. Alternatively, as affine transformations of the squared distance

$$\tilde{\delta}^2 = \begin{cases} a\delta^2 + b, & x \neq y,\ a, b \geqslant 0, \\ 0, & x = y, \end{cases} \tag{8}$$

give

$$\tilde{V}_\delta = a \cdot V_\delta + b/2,$$

$$\tilde{\phi}^2_\delta(\boldsymbol{x}) = a \cdot \phi^2_\delta(\boldsymbol{x}) + b/2,$$

by suitable choices of a and b, we may consider the closed construction

$$\tilde{f}_\delta(\boldsymbol{x}) = \exp(-\tilde{\phi}^2_\delta(\boldsymbol{x})), \tag{9}$$

such that $\tilde{f}_\delta$ is a probability density. Note that f_δ and $\tilde{f}_\delta$ are monotonically related (MR, see below) and can be different from f.

3. Probability densities which may not be generated by a distance

The following question arises. Given $f(\boldsymbol{x})$, is it possible to find a distance δ which provides a monotonic function of $f(\boldsymbol{x})$ following the above procedure? It has been shown that the answer is affirmative for a multinormal distribution. (For other distributions, see Cuadras, 1989; Cuadras et al., 1996). If for a (non-normal) $f(\boldsymbol{x})$ we can find a δ such that $\phi^2_\delta(\boldsymbol{x})$ and $f(\boldsymbol{x})$ are monotonically related (MR), i.e.,

$$\phi^2_\delta(\boldsymbol{x}_1) < \phi^2_\delta(\boldsymbol{x}_2) \quad \text{iff } f(\boldsymbol{x}_1) > f(\boldsymbol{x}_2),$$

and therefore

$$f_\delta(\boldsymbol{x}_1) > f_\delta(\boldsymbol{x}_2) \quad \text{iff } f(\boldsymbol{x}_1) > f(\boldsymbol{x}_2),$$

then we can work with $f_\delta(\boldsymbol{x})$ in some multivariate problems.

For example, given two populations $\boldsymbol{\Pi}_1, \boldsymbol{\Pi}_2$, if we also have (using an obvious notation)

$$f(\boldsymbol{x}, \boldsymbol{\Pi}_1) \leqslant f(\boldsymbol{x}, \boldsymbol{\Pi}_2) \quad \text{iff } f_\delta(\boldsymbol{x}, \boldsymbol{\Pi}_1) \leqslant f_\delta(\boldsymbol{x}, \boldsymbol{\Pi}_2),$$

the DB rule (see (3)) is equivalent to the maximum likelihood (ML) rule in discrimination. In practice, for real data, DB could perform better than ML, e.g., yielding a lower estimated probability of misallocation using f_δ instead of f, indicating that f has been misspecified.

However, this construction is not always possible, and even if it is, it may not be natural. One reason is that, in general, $\phi^2_\delta(\boldsymbol{x})$ is a quadratic expression in $\boldsymbol{x}$, having a minimum, and hence it is rather problematic to relate it to $f(\boldsymbol{x})$, when $f(\boldsymbol{x})$ is monotonic in the usual Euclidean norm $\|\boldsymbol{x}\|$. Consider, for example, the exponential density $f_\alpha(x) = \alpha^{-1} \exp(-\alpha^{-1}x)$, $x > 0$. The theorem below shows what can happen when this situation occurs.

Theorem 1. *Suppose that $f(\boldsymbol{x})$ is monotonic in $\|\boldsymbol{x}\| = d(\boldsymbol{0}, \boldsymbol{x})$, $\boldsymbol{x} \in S$, where d is the ordinary Euclidean distance, and that $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n \in S$ satisfies:*

If

$$\|\boldsymbol{x}_1\| < \cdots < \|\boldsymbol{x}_n\|$$

then

$$d(\boldsymbol{x}_1, \boldsymbol{x}_2) < d(\boldsymbol{x}_1, \boldsymbol{x}_k), \quad k > 2. \tag{10}$$

If there exists a distance $\delta(.,.)$ and a ψ such that $\phi^2_\delta(\boldsymbol{x}) = \|\psi(\boldsymbol{x}) - E(\psi(\boldsymbol{X}))\|^2_L$ is also monotonic in $\|\boldsymbol{x}\|$, then ψ breaks the rank order (10).

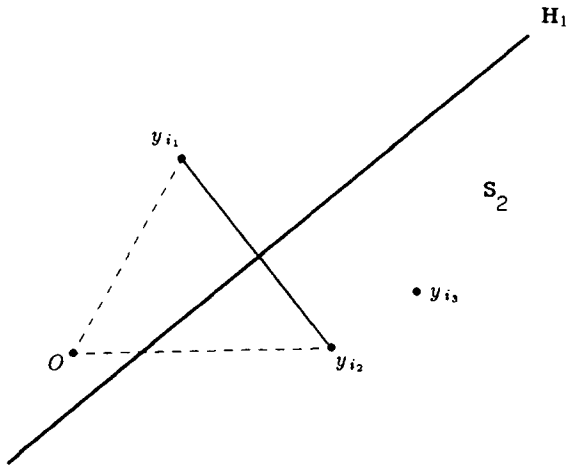

Fig. 1.

Proof. Let us consider $Y = \psi(X) - E(\psi(X))$. Then $E(Y) = \mathbf{0}$ and $\phi_\delta^2(x) = \|y\|_L^2$. If $x_1,\ldots,x_n$ are i.i.d. as X, we can arrange them such that $\|x_{i_1}\| < \cdots < \|x_{i_n}\|$. If $\phi_\delta^2(x)$ is monotonic in $\|x\|$, e.g., increasing, then $\|y_{i_1}\|_L < \cdots < \|y_{i_n}\|_L$. Suppose that the rank order of $x_{i_1},\ldots,x_{i_n}$ is preserved, i.e., the images y_i also satisfy

$$d_L(y_{i_1},y_{i_2}) < d_L(y_{i_1},y_{i_k}), \quad k > 2. \tag{11}$$

Let H_1 be the hyperplane of codimension 1 bisecting the line joining y_{i_1} and y_{i_2}. To preserve the monotonicity and the rank order (11), it is necessary that each y_{i_k}, for $k > 2$, lies in S_2, the opposite side of y_{i_1}, otherwise y_{i_k} would be closer to y_{i_1} (see Fig. 1). For very large n, it means that the center of gravity $\bar{y}$ should lie in S_2. However, by the law of large numbers, we can find a sequence such that

$$\left\| \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i \right\|_L < \varepsilon,$$

where $\varepsilon > 0$ is arbitrarily small. This is a contradiction and therefore (11) cannot hold. Hence there is a y_{i_k} such that

$$d_L(y_{i_1},y_{i_2}) = \delta(x_{i_1},x_{i_2}) > d_L(y_{i_1},y_{i_k}) = \delta(x_{i_1},x_{i_k}), \quad k > 2,$$

and thus the distance δ defined on S, with distance image d_L in L, breaks the rank order (Eq. (10)) of the ordinary distance d in S. □

A consequence of Theorem 1 is that the mapping ψ may not give a suitable representation of S from a multidimensional scaling point of view, with undesirable consequences in representing multivariate data. Linear discrimination and Mahalanobis distance are related to canonical variate analysis. Similarly, DB discrimination is related to generalized canonical variate analysis, a method for representing $g > 2$ groups when a general distance is given (Krzanowski, 1994a; Cuadras et al., 1996). If the rank order is broken, the representation of g groups with respect to δ would be inconsistent with the ordinary (finite dimensional) distance in S.

4. Probability densities from symmetric functions

We have proved that a natural distance does not always exist to provide a proximity function MR to $f(\boldsymbol{x})$. However, if we consider symmetric functions $s(.,.)$ instead of distances, the problem can be solved.

Let V_s and $\phi_s^2(\boldsymbol{x})$ be the geometric variability and the proximity function obtained using s instead of δ^2 in (1) and (2). In Theorem 2 it is supposed that a distance δ, transformed by (8) if necessary, supplies $\tilde{f}_\delta$ by means of (9).

Theorem 2. *Let f be the density of $\boldsymbol{X}$ and suppose that a distance δ generates $\tilde{f}_\delta$. Then:*

(1) *There exists a symmetric function s giving a proximity function $\phi_s^2(\boldsymbol{x})$ MR to $f(\boldsymbol{x})$ and such that $V_s = H(f)$, where $H(f)$ is the Shannon entropy.*

(2) *The following inequality holds:*

$$I(f;\tilde{f}_\delta) = V_\delta - H(f) \geqslant 0, \tag{12}$$

where $I(f;\tilde{f}_\delta)$ is the Kullback–Leibler divergence.

Proof. Let us consider the symmetric function

$$s(\boldsymbol{x},\boldsymbol{y}) = -\log f(\boldsymbol{x}) - \log f(\boldsymbol{y}), \quad \boldsymbol{x},\boldsymbol{y} \in S. \tag{13}$$

Then

$$V_s = H(f) = -\int_S f(\boldsymbol{x}) \log f(\boldsymbol{x}) \lambda(\mathrm{d}\boldsymbol{x}),$$

and

$$\phi_s^2(\boldsymbol{x}) = -\log f(\boldsymbol{x}), \quad \boldsymbol{x} \in S,$$

which is MR to $f(\boldsymbol{x})$.

Let δ be any distance generating a density $\tilde{f}_\delta$. Then

$$\begin{aligned} I(f;\tilde{f}_\delta) &= \int_S f(\boldsymbol{x})[\log f(\boldsymbol{x})/\tilde{f}_\delta(\boldsymbol{x})]\lambda(\mathrm{d}\boldsymbol{x}) \\ &= \int_S f(\boldsymbol{x})[\log f(\boldsymbol{x}) + \tilde{\phi}_\delta^2(\boldsymbol{x})]\lambda(\mathrm{d}\boldsymbol{x}) \\ &= V_\delta - H(f) \geqslant 0. \quad \square \end{aligned}$$

The Shannon entropy $H(f)$ also measures the variability of $\boldsymbol{X}$ and its geometric interpretation is worth examining, e.g., following Csizár (1975). Inequality (12) shows that $H(f)$ is a lower bound for the geometric variability of a distance δ giving a density $\tilde{f}_\delta$. In particular, $f = \tilde{f}_\delta$ (a.e.) iff $V_\delta = H(f)$. Thus:

Corollary. *A density with null or negative entropy cannot be generated by a distance using* (9).

One example is $f_\alpha(x) = \alpha^{-1}\exp(-\alpha^{-1}x)$, $x > 0$, for which $H(f) \leqslant 0$ if $0 < \alpha \leqslant \mathrm{e}^{-1}$. We cannot obtain f_α by a distance but by the symmetric function (13). Note that this construction is *not unique*, as $s_\alpha(x,y) = 1 + \alpha^{-2}xy + 2\log\alpha$ also provides $\phi_s^2(x) = -\log f_\alpha(x) = \alpha^{-1}x + \log\alpha$. However, for $\alpha > \mathrm{e}^{-1}$, (7) or (9) may supply $f = f_\delta$. But then Theorem 1 should be taken into account, as $f_\alpha(x)$ is monotonic in x.

Another example of symmetric function is

$$s_c(\boldsymbol{x},\boldsymbol{y}) = \exp(i\boldsymbol{x}'\boldsymbol{y}), \quad \boldsymbol{x},\boldsymbol{y} \in S,$$

Table 1
Number of misallocations in diagnosing cancer data

Method	Distance	Π_1	Π_2	Total
EDF	Euclidean (1)	29	37	66
LDF	Mahalanobis (2)	31	27	58
QDF	Mahalanobis (3)	13	35	48
LM	Mahalanobis (4)	21	24	45
QLM	Mahalanobis (5)	23	20	43
DB	Gower	18	21	39

(1) Covariance matrix $\Sigma = \mathbf{I}$. (2) Same Σ in each group. (3) Different Σ in each group. (4) Same Σ in each state and group. (5) Different Σ in each state and group.

where $i = \sqrt{-1}$, which gives a proximity function closely related to the characteristic function. This function is not MR to $f(\boldsymbol{x})$, but the inverse Fourier transformation gives $f(\boldsymbol{x})$.

Thus, the problem of constructing densities can always be solved using symmetric functions, although distances would be more appropriate in discriminant analysis, in order to relate discrimination with the graphical representation of the populations, by generalized canonical variate analysis.

5. A mixed data discriminant analysis example

Let us consider a real data illustration, where the use of well-known discriminant functions can be improved by DB discrimination, i.e., by using densities obtained from appropriate distances. Krzanowski (1980) described a cancer data set consisting of 11 mixed variables (7 continuous, 2 binary and 2 three-state categorical), measured on 137 individuals, 78 with benign tumours (Π_1) and 59 malignant (Π_2). Table 1 gives the results of performing a discriminant analysis, using several methods: linear discrimination (LDF), quadratic discrimination (QDF), Euclidean discrimination (EDF, Marco et al., 1987), location model (LM), quadratic location model (QLM, Krzanowski, 1994b) and DB model. Each method can be associated to a distance (Euclidean, several versions of Mahalanobis, etc.), e.g., Krzanowski (1983) found a relation between LM and Mahalanobis distance, using an extension for mixed data of Matusita affinity. We used the distance based on Gower's similarity coefficient (Gower, 1971) in the DB method. The number of misallocations is computed by the leaving-one-out procedure.

The results obtained (Table 1) for EDF, LDF, QDF, LM and QLM show that performance is improved as the number of model parameters increases. However, DB is considered as a non-parametric allocation rule (see Krzanowski and Marriott, 1995, Section 9.29) and its best performance suggests that the probability density built from Gower's distance is more appropriate in fitting the data.

6. Conclusions

Given a random vector $\boldsymbol{X}$, with probability density function $f(\boldsymbol{x})$ with support $S \subset \mathbb{R}^p$, we have shown that sometimes a distance δ can provide a proximity function monotonically related to $f(\boldsymbol{x})$. But if $f(\boldsymbol{x})$ is monotonic in $\|\boldsymbol{x}\|$, then δ can distort the ordinary distance in S. However, if we consider symmetric functions rather than distances, then we can always reproduce monotonically $f(\boldsymbol{x})$.

The dispersion of $\boldsymbol{X}$, with respect to δ, is measured by the geometric variability, which has the Shannon entropy as a lower bound when the generated density is presented in the closed form (9).

A final consequence of these results is that it is possible to perform a discriminant analysis using distances or symmetric functions between observations, instead of probability densities. The discriminant analysis is reduced to modeling such functions.

References

Csiszár, I. (1975), I-Divergence geometry of probability distributions and minimization problems, *Ann. Probab.* **3**, 146–158.

Cuadras, C.M. (1989), Distance analysis in discrimination and classification using both continuous and categorical variables, in: Y. Dodge, ed., *Statistical Data Analysis and Inference* (Elsevier, Amsterdam) pp. 459–474.

Cuadras, C.M. (1992a), Some examples of distance based discrimination, *Biomet. Lett.* **29**, 1–18.

Cuadras, C.M (1992b), Probability distributions with given multivariate marginals and given dependence structure, *J. Multivariate Anal.* **42**, 51–66.

Cuadras, C.M. and J. Fortiana (1993), Continuous metric scaling and prediction, in: C.M. Cuadras and C.R. Rao, eds., *Multivariate Analysis Future Directions 2* (Elsevier, Amsterdam) pp. 47–66.

Cuadras, C.M. and J. Fortiana (1995), A continuous metric scaling solution for a random variable, *J. Multivariate Anal.* **52**, 1–14.

Cuadras, C.M., J. Fortiana and F. Oliva (1996), The proximity of an individual to a population with applications in discriminant analysis, *J. Classification*, **14**, in press.

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

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. (1994a), Ordination in the presence of group structure, for general multivariate data, *J. Classification* **11**, 195–207.

Krzanowski, W.J. (1994b), Quadratic location discriminant functions for mixed categorical and mixed variables, *Statist. Probab. Lett.* **19**, 91–95.

Krzanowski, W.J. and F.H.C. Marriott (1995), *Multivariate Analysis – Part 2* (Arnold, London).

Marco, V.R., D.M. Young and D.W. Turner (1987), The Euclidean distance classifier: an alternative to linear discriminant analysis, *Comm. Statist. B. Simulation Comput.* **16**, 485–505.

Rao, C.R. (1982), Diversity: its measurement, decomposition, apportionment and analysis, *Sankhya* **44A**, 1–22.