CONTRIBUTIONS to SCIENCE, **2** (2): 183-191 (2002)
Institut d'Estudis Catalans, Barcelona

Recent statistical methods based on distances

C. Arenas* and C. M. Cuadras

Departament d'Estadística. Facultat de Biologia. Universitat de Barcelona

Asbtract

The distance concept has been applied in different fields and is fundamental in recent statistical methods, valid for non-numerical explanatory variables as well as a mixture of variables. This paper discusses and illustrates recent methods based on distances: distance-based regression, distance-based discrimination, related metric scaling and continuous scaling. These methods form part of the present research of the multivariate group, conducted by C.M. Cuadras of the Departament of Statistics at the University of Barcelona. This group, doing research at the Faculties of Biology and Mathematics, consists of seven researchers: C. Arenas, C. M. Cuadras, D. Cuadras, A. Esteve, J. Fortiana, A. Grané and F. Oliva.

Keywords: categorical and mixted data, distances between observations, regression model, discriminant function, missing data, principal co-ordinate analysis, multidimensional scaling, related metric scaling.

Resum

El concepte de distància s'ha utilitzat en diferents camps i és bàsic en alguns mètodes estadístics recents, vàlids per a variables no numèriques així com per a una barreja de diferents tipus de variables. A aquest article expliquem i il·lustrem mètodes recents basats en distàncies: regressió basada en distàncies; discriminant basat en distàncies, *related metric scaling* i *continuous scaling*. Aquests mètodes formen part de la recerca actual del grup d'anàlisi multivariant, dirigit per C. M. Cuadras, del Departament d' Estadística de la Universitat de Barcelona. Aquest grup, que treballa a les Facultats de Biologia i Matemàtiques, està format per set investigadors: C. Arenas, C. M. Cuadras, D. Cuadras, A. Esteve, J. Fortiana, A. Grané i F. Oliva.

1. Introduction

In statistics and data analysis, the geometrical concept of distance between individuals or populations have been applied in fields such as anthropology, biology, genetics, psychology, linguistics and others. The distance concept is a useful tool in hypothesis testing and parameter estimation between other applications. Also, in some statistical techniques, such as correspondence analysis or multidimensional scaling, the concept of distance is a basic tool. Distance functions are also fundamental in recent methods such as the distance-based regression analysis, the distance-based discrimination analysis, and the related metric scaling. C. M. Cuadras presented a survey about distances, its properties and applications in [8].

These methods are valid for non-numerical explanatory variables as well as mixed variables, which frequently arise in applications (medicine, biometry, psychology, etc.), but few models, have been used to overcome this situation. The purpose of the distance-based (DB) methods, regression and discrimination, is to properly handle problems with non-real value predictors, including categorical or a mixture of real-valued and categorical explanatory variables. Distance-based methods (DB) use a metric $d(\cdot,\cdot)$ defined on the set of predictors and all computations take the resulting distances between observations as the departure point. The start-up ideas can be found in a paper of C. M. Cuadras [9] and, as these methods are available for a mixture of continuous and categorical variables, they are quite useful for applications with real data. Several articles present data for applications in the botanical and anthropological fields [2], [3], [6]. As these methods are based on a metric $d(\cdot,\cdot)$ it is obvious that the results depend on the selected metric. The first part of the study for this approach considers the selection of the metric and proves that when a suitable metric is taken, these methods reduce to classic regression or discrimination methods.

* Author for correspondence: C. Arenas, Departament d'Estadística, Facultat de Biologia, Universitat de Barcelona. Diagonal 645, 08028 Barcelona, Catalonia (Spain). Tel. 34 934021561. Fax: 34 934111733. Email: carenas@ub.edu

For the representation of several groups, the *related metric scaling* is also a useful technique. This method has been applied to analyse chromosomal position ([5], [37]). This technique obtains a joint representation of n objects when two distances are available [26] and is an extension of classic metric scaling [7]. Another extensions is the *continuous scaling*. This method obtains the principal dimensions of a random variable [21].

We present a brief developmental history of these methods, the corresponding mathematical model with the main properties, a few examples and comments about their incidence.

2. Distance-Based Regression

2.1. A brief history

The first reference related to the DB-regression method was a paper of C.M. Cuadras entitled *Statistical Methods applied to the prehistoric reconstruction* [9]. This paper considered the problem of the prediction of a continuous variable from independent qualitative variables. For this situation, the limitations of the classical linear regression model are known. The usual way of proceeding would be to subject the qualitative variables to some scoring system (optimal scaling, for example) and consider all the variables as quantitative. Other options are possible [46] but an optimum solution does not exist. A methodology based on Principal Co-ordinate Analysis was introduced in [9, 10]. These works introduce the idea of constructing a similarity or a distance matrix from the original data, to apply a principal Co-ordinate Analysis and to consider a new model where the principal co-ordinates play the role of explanatory variables. Then, a formula for the prediction of a new observation was given. In order to illustrate this possible regression model, it was applied to the classical data of students given in [50]. From this initial idea, Cuadras and Arenas [19] formally defined the DB-linear regression model. This paper considers the case of a mixture of continuous and qualitative variables compared to the classic linear regression. Selected properties were studied and real data were used to illustrate the model's utility. As it is based on a Principal Co-ordinate Analysis, the number of new explanatory variables used may be too large, therefore, a possible criteria to select only some of these variables was proposed. In fact, the optimal selection of variables ([41], [49]) is still an open question and a coherent criterion for the dimension reduction does not exist in the classical formulation of principal component regression ([43], [57]). How to compute the coefficient of determination and the prediction of a new observation was also developed in [19]. Moreover, when the Euclidean distance is used the relationship and compatibility of this method to the classical linear regression model was proved. Additional properties and examples can be found in [20]. In particular, it was demonstrated that the DB model with the distance

$$d(x,y)=\sqrt{|x-y|}\,,$$

is equivalent to the regression on orthogonal polynomials. For dimension p>2, there are no theoretical results yet, but the performance of the DB method with the distance

$$d(x,y)=\sqrt{\sum_{i=1}^{p}|x_i-y_i|}\,,$$

was shown with real examples.

These ideas were extended to the non-linear regression case [29]. Namely, they introduced a coefficient in order to choose the most predictive dimensions, providing a solution to the problem of small variances and very large number of observations. They also proposed a solution to the problem of missing data and showed that the DB method can be regarded as a kind of ridge regression when the usual Euclidean distance is used. Another solution for the missing data case is proposed and justified with real data in [3]. In the application of the DB-model special matrices arise, e.g., the $n\times n$ matrices A = (a_{ij}) where

$$a_{ij} = a_{ji} = min\{i, j\},\ i= 1,..., n.$$

In Cuadras [11], the eigen-structure of these matrices was conjectured:

> *«Given an eigenvector* $\mathbf{v}$ *of* $\mathbf{A}$ *the remaining eigenvectors are obtained by permuting up to sign the components of* $\mathbf{v}$*».*

But, at the moment, only empirical results confirm this conjecture, which is still an open problem. Fortiana and Cuadras [38] proposed a parametric family of matrices, which includes the previous one, proved some theoretical results and traced the way to solve this conjecture.

A generalised DB-regression model for a predictor and response matrix respectively is described in [24]. Other interesting properties and applications related to the regression problem can be found in [13], [14] and [38].

Ad hoc software was prepared to compute the DB-method (linear or non-linear case). These programs formed part of a Multivariate Package of non standard multivariate methods [4].

2.2. The model

The DB-regression model, as it was defined in [19] for the linear case, and in [29] for the non-linear case, is discussed below.

First, let us consider the linear case and suppose that we wish to relate a continuous variable Y to a variable vector **W**, where **W** is a mixture of continuous, binary and categorical variables. Consider a set of n individuals $S = \{1,2,..., n\}$, and a distance function $d(\cdot,\cdot)$, which depend on **W**, and gives a $n\times n$ distance matrix $\mathbf{D} = (d_{ij})$. We suppose that **D** is an Euclidean distance matrix. Let $\mathbf{A} = (a_{ij})$ the matrix with elements $a_{ij} = -(d_{ij}^2)/2$ and set **B=HAH** where $\mathbf{H} = \mathbf{I}_n - \mathbf{11}'/n$ is the centring matrix. It is well-known ([50]) that **B** is positive semi-definite and assuming rank (**B**) = m, the spectral decomposition of matrix **B** is $\mathbf{B} = \mathbf{U\Lambda U}' = \mathbf{XX}'$, where $\mathbf{\Lambda}$ is diagonal and **X**

$= \mathbf{U}\Lambda^{1/2}$ is an $n \times n$ matrix of rank m. As the distances between the rows of $\mathbf{X}$ are the same as d_{ij}, the following full DB-model is proposed:

$$\mathbf{y} = \gamma_0 \mathbf{1} + \mathbf{X}\gamma + \mathbf{e},$$

where $\mathbf{1}$ is the vector of 1's, γ_0 is an unknown scalar parameter and γ is an unknown parameter vector of dimension p, and $\mathbf{y}$ is the vector of observations of Y.

As the number of columns $\mathbf{X}$ could be too large, a suitable subset should be selected. Setting $\mathbf{X} = (\mathbf{X}_{(k)}, \mathbf{Z})$, the k-dimensional general linear model is suggested

$$\mathbf{y} = \beta_0\mathbf{1} + \mathbf{X}_{(k)}\beta_{(k)} + \mathbf{e}_{(k)}.$$

Note that $\mathbf{1}, \mathbf{X}_1, \ldots \mathbf{X}_k$ are eigenvectors of $\mathbf{B}$ with eigenvalues $0, \lambda_1, \ldots \lambda_k$ respectively. Additional criteria for selecting or deleting predictive columns of $\mathbf{X}$ can be found in [19].

The ordinary least squares estimates of β_0 and $\beta_{(k)}$ are given by, $\hat{\beta}_0 = \bar{y}$ and $\hat{\beta}_{(k)} = \Lambda_{(k)}^{-1}\mathbf{X}'_{(k)}\mathbf{y}$, where $\Lambda_{(k)}$ = diag $(\lambda_1, \ldots, \lambda_k)$. For computing the coefficient of determination a useful formula is

$$R_k^2 = \sum_{i=1}^{k} r^2(Y, \mathbf{X}_i),$$

where $r(Y, \mathbf{X}_i)$ is the simple correlation coefficient between Y and the predictor variable $\mathbf{X}_i$. Also, for a new individual ω, the prediction $Y(\omega) = y_{n+1}$, can be computed by

$$y_{n+1} = \bar{y} + \mathbf{x}'_{(k)}\Lambda_{(k)}^{-1}\mathbf{X}'_{(k)}\mathbf{y} + \mathbf{z}'\Lambda_{m-k}^{-1}\mathbf{Z}'\mathbf{y},$$

with

$$\mathbf{x} = \begin{bmatrix} \mathbf{x}_{(k)} \\ \mathbf{z} \end{bmatrix}, \qquad \mathbf{X} = \begin{bmatrix} \mathbf{X_{(k)}} \\ \mathbf{Z} \end{bmatrix}, \qquad \Lambda = \begin{bmatrix} \Lambda_{\mathbf{(k)}} & \\ & \Lambda_{\mathbf{(m-k)}} \end{bmatrix}$$

Details and proofs are presented in [19].

The DB-model is compatible with the classical regression model, when the predictor variables are continuous and the Euclidean distance is used. The equivalence also holds for qualitative variables when a distance based on the matching coefficient is used.

In the non-linear case, this DB model can be applied by taking the distance

$$d(x,y) = \sqrt{\sum_{i=1}^{p} |x_i - y_i|}$$

This choice provides principal co-ordinates which behave as linear, quadratic, cubic, ..., dimensions. For p=1 and the equidistant case, this method is equivalent to an ordinary regression on k suitable Chebychev polynomials of the first kind [20]. The non-equidistant case is also related to a set of orthogonal polynomials defined by a recurrence formula. The case p>1 is unsolved, but several examples show a good performance using this model.

2.3. Examples

Two examples are given in order to illustrate the utility of the method covering the linear and non-linear case.

Example 1. Linear case

The data relates the automobile accident rate, in accidents per million vehicle miles to 13 potential independent variables: 3 binary, 3 qualitative and 7 continuous [61]. The data include 39 sections of major highways in Minnesota (USA) in 1973. In this case, in order to compare the classical regression method with the DB-method, we computed the coefficient of determination R^2 and the value of the coefficient

$$C = \sum_i (y_i - \hat{y}_i)^2 / n,$$

where $\hat{y}_i$ is the prediction obtained by leaving out the individual i from the original data (cross-validation method). We use Gower's distance [39], which is a suitable distance measure for mixed data.

The values obtained are reported in Table 1. Note that the DB-method improves the classical one by using Gower's distance.

Example 2. Non-linear case

Next we considered the data which reports a set of 38 measures on a chemical reaction ([35]). Y is the fraction of original material remaining after $\boldsymbol{x}_1$ minutes of reaction at $\boldsymbol{x}_2$ degrees Kelvin. The non-linear regression model is

$$Y = \exp\left\{-\theta_1 x_1 \exp\left[-\theta_2\left(\frac{1}{x_2} - \frac{1}{620}\right)\right]\right\} + \varepsilon$$

where θ_j, j = 1,2 are the parameters. In this case the results are quite similar, although a better fit for the non-linear model shows that this model may be better. However, the DB-method has been performed *without knowing* the function in the non-linear regression model.

More examples are presented in [11], [19], [20] and [29].

2.4. The incidence of the method

The DB-method has been referenced in several works. For instance:

- [1] in relation with a new methodology to construct a tuned QSAR model.
- [40] in relation with MANOVA models, which are not consonant with the MANOVA assumptions and in applications for economic data.
- [42] where the DB-method is used for short-term solar-flare predictions.

Table 1. Results of the classic regression model and DB-regression method for the Example 1 Section 2.3, where R^2 is the coefficient of determination and C is the cross-validation coefficient.

	R^2	C
Classic method	0.755	2.501
DB-method	0.875	1.564

- [53] in relation with predictive models based on tuned molecular quantum similarity measurements and their application to obtain quantitative structure-activity relationships.
- [54] in relation with molecular quantum similarity measurements.

3. Distance-Based Discrimination

3.1. A brief history

The first paper about the DB discrimination method [10] gives a solution to discrimination and classification using both continuous and categorical data, overcoming the classical and arbitrary procedures of codification of non-continuous variables. It is well known that if all the variables are continuous with normal distribution, the linear (LDF) and quadratic (QDF) discrimination rules are the best. The former can be applied if the hypothesis of equality of covariance matrices can be accepted. If the Mahalanobis distance is used, LDF and DB rules give the same results. Moreover, if there are mixed variables, the location model [45] is a good rule, if the normality assumption is verified for the continuous variables, however, it requires a codification of the qualitative variables. As the DB method does not suppose any probability distribution and does not need a codification of the variables, it is very useful for real data, e.g., in DNA sequences and assignation of manuscripts or voice ([59]). The utility of this method is shown in [12] with some good examples. Results and properties about the method can be found in [33] and [34]. Contributions to the typicality in discrimination were developed in [25]. Examples with real data can be found in [2], [3], [6] with applications to botany and anthropology.

Ad hoc software was prepared computing the DB-discrimination rule ([4]).

3.2. The model

Let $\mathbf{X} = (X_1,..., X_p)$ be a random vector with values on some space $E \subset R^p$ and probability density f with respect to a suitable measure λ. Suppose that $\delta(\cdot,\cdot)$ is a distance function on E, i.e., such that $\delta(\mathbf{x},\mathbf{y}) = \delta(\mathbf{y},\mathbf{x}) \geq \delta(\mathbf{x},\mathbf{x}) = 0$, $\forall \mathbf{x},\mathbf{y} \in E$. Suppose that

$$\mathbf{V}_\delta(\mathbf{X}) = \frac{1}{2}\int_{ExE} \delta^2(\mathbf{x},\mathbf{y}) f(\mathbf{x}) f(\mathbf{y}) d\lambda(\mathbf{x}) d\lambda(\mathbf{y})$$

is finite. $\mathbf{V}_\delta(\mathbf{X})$, called *geometric variability* in [21], is the measure of dispersion of $\mathbf{X}$ with respect to δ, which reduces to the total variation trace(Σ) when δ is the ordinary Euclidean distance, Σ being the covariance matrix of $\mathbf{X}$.

Given $\omega_0 \in \mathbf{\Omega}$, the *proximity function* of the observation $\mathbf{x}_0 = \mathbf{X}(\boldsymbol{\omega_0})$ to the population represented by $\mathbf{X}$ is defined as

$$\Phi(\mathbf{x}_0) = \int_E \delta^2(\mathbf{x_0},\mathbf{x}) f(\mathbf{x}) d\lambda(\mathbf{x}) - \mathbf{V}_\delta(\mathbf{X}),$$

i.e., $\mathbf{\Phi}(\mathbf{x}_0)$ is the average of the squared distance from $\mathbf{x}_0$ to the population minus the geometric variability.

If $\mathbf{x}_1,..., \mathbf{x}_n$ is a sample from $\mathbf{X}$, the sampling version of a proximity function is

$$\hat{\Phi}(\mathbf{x}_0) = \frac{1}{n}\sum_{i=1}^{n} \delta^2(\mathbf{x}_0,\mathbf{x}_i) - \frac{1}{2n^2}\sum_{i,j=1}^{n} \delta^2(\mathbf{x}_i,\mathbf{x}_j).$$

Thus $\mathbf{\Phi}(\mathbf{x}_0)$ can be estimated without knowing the density f.

For theoretical and practical aspects see [10], [12], [27], [32], [33] and [34].

Suppose that we have samples of sizes $n_1,..., n_g$ drawn from g populations or groups $\Omega_1,..., \Omega_g$ and a distance function δ between observations. We can obtain the proximity functions

$$\hat{\Phi}_k(\mathbf{x}) = \frac{1}{n_k}\sum_{i=1}^{n_k} \delta^2_{i(k)} - \frac{1}{2n_k^2}\sum_{i,j=1}^{n_k} \delta^2_{ij(k)} \ , \ \mathrm{k} = 1,...,\mathrm{g} \ ,$$

where:

$\mathbf{x}$ is the observation of $\mathbf{X}$ on one individual $\omega \in \Omega_1 \cup \cdots \cup \Omega_g$,
$\delta_{i(k)}$ is the distance from $\mathbf{x}$ to the i-*th* observation of Ω_k,
$\delta_{ij(k)}$ is the distance between two observations *i, j* of Ω_k.

Now suppose that ω is an individual to be allocated such that $\mathbf{x} = \mathbf{X}(\boldsymbol{\omega})$. The distance-based discriminant rule, or DB-rule, is:

$$\text{Allocate } \omega \text{ to } \Omega_i \text{ if } \hat{\Phi}_i(\mathbf{x}) = \min \{\hat{\Phi}_1(\mathbf{x}),..., \hat{\Phi}_g(\mathbf{x})\}.$$

In [33] it was proved that each $\hat{\Phi}_i(\mathbf{x})$ could be interpreted as a squared distance from $\mathbf{x}$ to Ω_i. Thus the DB-rule assigns an individual to the nearest group [34]. Further, it can be shown that it is equivalent to the linear discriminant or the quadratic discriminant rule when a distance like Mahalanobis is considered. Furthermore, as it is based on a distance, it can be applied to binary, qualitative, or mixed variables by using a suitable distance function ([10], [12]). This DB-rule is understood as a non-parametric discriminant rule in [47].

The results of the distance-based discriminant analysis depend on the choice of distance δ. In [51] it was proved that Gower's distance ([39]) is a suitable distance for the treatment of data with missing values. A complete discussion about the use and advantages of this distance-based method when dealing with missing values is discussed in [3].

3.3. Examples

An application to a problem in linguistics [55] was reported in [12]: to decide whether a diphthong whose first vowel is an atonic i, appearing after a consonant in Catalan, should be pronounced as monosyllabic (π_1) or bisyllabic (π_2). Random samples of 136 and 43 words whose pronunciation is known, were selected and each one was coded in five categorical variables. The leaving-one-out procedure yields 58 misclassifications for LDF, 38 for QDF and only 8 for the DB method whereas the log-linear discrimination does not work for this data.

3.4. The incidence of the method

The DB-method has been referenced in various different works. See for example [47] and [60], where a new algorithm for allocation of an individual to one of several populations is proposed. In [52] this method is mentioned in relation to the pattern recognition of the boundary shape of closed figures.

4. Related Metric Scaling

4.1. A brief history

This method arises when two or more distances are available and a joint representation of data is necessary [15], [26]. A provoking problem is resolving the position of human chromosomes. The first results were obtained in [37] and the CACROMOS program, presented in [5], allows the computations. This technique obtains a joint representation of *n* objects when two distances are available [26], and has been extended to more than two distances [15]. Recently a real application of the method to human evolution was presented in [6]. Finally a probabilistic extension of the method is proposed in [22].

4.2. The model

Consider *n* individuals and two *n* × *n* distance matrices $\mathbf{D}_k = (d_{ij(k)})$, $k = 1,2$, and its corresponding *n* × *n* inner product matrices $\mathbf{B}_k = (\mathbf{b}_{ij(k)})$, $k = 1,2$, related to $\mathbf{D}_k$ by

$$d^2_{ij(k)} = b_{ii(k)} + b_{jj(k)} - 2b_{ij(k)}$$

Let $\mathbf{B}_k = \mathbf{U}_k \mathbf{\Lambda}_k \mathbf{U}'_k$ be the spectral decomposition of $\mathbf{B}_k$. That is, $\mathbf{U}_k$ contains unitary eigenvectors and Λ_k is diagonal with the eigenvalues of $\mathbf{B}_k$. Then the matrix with the principal co-ordinates is $\mathbf{X}_k = \mathbf{U}_k \Lambda_k^{1/2}$, which satisfies $\mathbf{B}_k = \mathbf{X}_k \mathbf{X}'_k$. The problem is now to find an *average* matrix **B** summarising the information contained in the matrices $\mathbf{B}_k$, and then to find **X** such that **B** = **XX**´. The average matrix **B** can be used to obtain a final representation of the groups, which summarises all the initial information. The proposed average matrix **B** is:

$$\mathbf{B} = \mathbf{B}_1 + \mathbf{B}_2 - \frac{1}{2}(\mathbf{B}_1^{1/2}\mathbf{B}_2^{1/2} + \mathbf{B}_2^{1/2}\mathbf{B}_1^{1/2}),$$

where $\mathbf{B}_k^{1/2} = \mathbf{U}_k \Lambda_k^{1/2} \mathbf{U}'_k$. The final representation of n individuals can be obtained by using the co-ordinates of the matrix **X** such that **B** = **XX**´. This definition can be justified as follows.

Let **D** = (d_{ij}) the joint distance matrix related to **B**, i.e., $d^2_{ij} = b_{ii} + b_{jj} - 2b_{ij}$ or **B** = **HAH** (see above). It can easily be proved that:

1) If $\mathbf{D}_1 = \mathbf{D}_2$ then also $\mathbf{D} = \mathbf{D}_1 = \mathbf{D}_2$.

2) If $\mathbf{X}'_1 \mathbf{X}_2 = 0$ then $\mathbf{D}^{(2)} = \mathbf{D}_1^{(2)} + \mathbf{D}_2^{(2)}$, where

$$\mathbf{D}^{(2)} = (d^2_{ij}), \mathbf{D}_k^{(2)} = (d^2_{ij(k)}), \quad k = 1,2.$$

Thus, the definition of **D** is consistent with equality and orthogonality. In general, we can have an intermediate situation between 1) and 2), so that **D** keeps the redundant information between $\mathbf{D}_1$ and $\mathbf{D}_2$.

For two-dimensional representation, we take the matrix $\mathbf{X}_{(2)}$ with n rows and 2 columns which best fits **X** in the least square sense. This matrix is $\mathbf{X}_{(2)} = \mathbf{U}_{(2)} \Lambda_{(2)}^{1/2}$, where $\mathbf{U}_{(2)}$ and $\mathbf{\Lambda}_{(2)}$ contain the first two eigenvectors and eigenvalues of **B** respectively. Definition of matrix **B** can also be justified by some theoretical properties [15], [26].

4.3. Examples

Example 1. Anthropological data

A joint representation of ten ethnic groups was found in [6]. Working with 860 crania measurements from ten ethnic groups: Yamana (Y), Alakaluf (Al), Ona (O), Eskimo (E), Arikara (Ar), Santa Cruz (S.C.), Peruvians (P), Australians (Au), Tasmanians (T) and Melanesians (M). With this data we compared the ethnic groups Yamana, Ona, and Alakaluf with the other Amerindian races to ascertain whether there is a strong relation among them. This would be an indicator of colonisation from North to South along the American continent. As there is also the possibility that these ethnic groups come from immigrations along the Pacific, we have compared them with other groups from the Austral continent and from the South Pacific. For the Yamana, Alakaluf, and Ona samples, 65 variables were measured, however, there were a great number of missing values. For the other 7 populations, 45 biometrical traits were measured with no missing values. The first group presents two difficulties:

(1) a significant number of missing values and,

(2) the poor identification of ethnic origin and of the skulls.

Previously these skulls were completely identified using the DB discrimination method [3].

The results of the related metric scaling (Figure 1) show the real geographical situation of the groups and a clear differentiation between the American and the Pacific groups. Also, the Yamana and Alakaluf groups are closer to the other American groups than the Ona group. On the other hand, the Ona group does not seem to be related to the Australians or Amerindians, as some theories suggest.

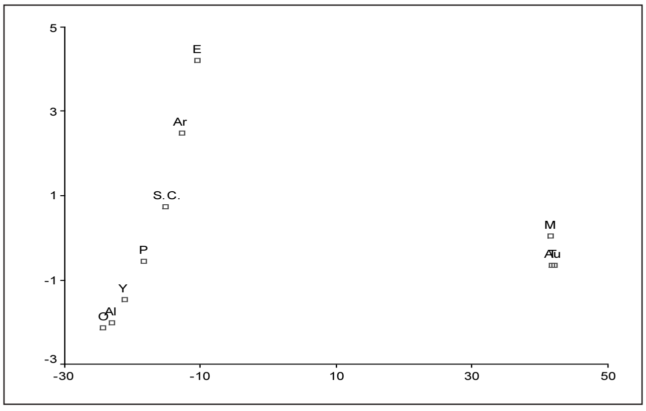

Figure 1. Graphical representation for different ethnic groups using the related metric scaling method. Yamana (Y), Alakaluf (Al), Ona (O), Eskimo (E), Arikara (Ar), Santa Cruz (S.C.), Peruvians (P), Australians (Au), Tasmanians (T) and Melanesians (M).

Example 2. Statistical research data

The related metric scaling can be used [26] to represent some aspects of statistical research in Spain. The data were collected from The Extended CIS Database [58]. They considered two sources of data: the number of papers published by 11 representative authors on 11 subjects (information on individuals) and the number of authors that had written joint papers (information on pairs of individuals). Two distance matrices were defined from the data and the related metric scaling provides a way of mixing these two types of information taking into account possible redundancies. Data and results are presented in [26].

5. Continuous Scaling

5.1. Introduction

Multidimensional Scaling is a multivariate analysis method to obtain, for a given distance matrix $\Delta = (\delta_{ij})$, $i, j \in I$, points $P_i \in R^p$, such that the distances between points give Δ,i.e.,$d(P_i,P_j) = \delta_{ij}, i,j, \in I$, where $d(\cdot,\cdot)$ is the Euclidean distance ([7]). In ordinary applications, I is a finite set (nations, stimulus, cars, etc.).

Suppose that $I \subset R$ is a continuous set, e.g., an interval. Suppose that there exists an embedding $x \to \varphi(x) \in E$, where E is a real separable Hilbert space with quadratic norm $\|\cdot\|$ such that $\delta(x, x') = \|\varphi(x) - \varphi(x')\|$, $x,x' \in I$. We may identify $\varphi(x)$ with $Q(x)$, where for $x \in I$, $Q(x) = (Q_1(x),Q_2(x),...)$ are the Euclidean coordinates such that

$$\delta^2(x,x') = \sum_{n=1}^{\infty}\left(Q_n(x) - Q_n(x')\right)^2, \qquad x,x' \in I.$$

To find φ and an optimal countable representation $Q(x)$ of $\varphi(x)$ for a given probability distribution, is the aim of continuous scaling. The Euclidean embedding or method to finding Euclidean coordinates from distances was first given by [56].

5.2. Continuous Scaling on a random variable

This approach was used [20], [21] in studying the principal dimension of a random variable X with range I=[a,b]. Also it was proved that considering the symmetric covariance kernel $K(s,t) = min\{F(s),F(t)\} - F(s)F(t)$, the eigen- decomposition

$$K(s,t) = \sum_{n=1}^{\infty}\lambda_n\psi_n(s)\psi_n(t),$$

where (λ_n,ψ_n) are eigenvalues, eigen-functions of K, then

$$\delta^2(x,x') = \sum_{n=1}^{\infty}(h_n(x) - h_n(x'))^2,$$

where $\delta(x,x') = \sqrt{|x - x'|}$ and $h_n(x) = \int_a^x \psi_n(t)dt.$

If X_t is the indicator of [X > t], i.e., if x is an observation of X, then X_t = o for $x \le$ t and X_t = 1 for $x > t$. Then

$$\delta^2(x,x') = \int_I (x_t - x'_t)^2\,dt$$

where X'_t is the indicator of X' and x_t, x'_t, are realisations of X_t, X'_t. Thus X_t, $t \in I = [a,b]$ is a continuous configuration to represent the distance $\delta(\cdot,\cdot)$ and $H(x) = (h_1(x),h_2(x),...)$ is an optimal discrete configuration to represent the same distance.

On the other hand, $h_1(x),h_2(x),...$ can be interpreted as principal components of X_t, as well as principal coordinates of distance

$$\delta(x,x') = \sqrt{|x - x'|}.$$

Thus:

var($h_n(X)$) = λ_n, cov($h_m(X),h_n(X)$) = 0 for $m \neq n$,

$$trace(K) = \sum_{n=1}^{\infty}\lambda_n$$

and the following expansion holds

$$|X - X'| = \sum_{n=1}^{\infty}\left(h_n(X) - h_n(X')\right)^2.$$

5.3. Continuous Scaling expansions

The above expansion can be generalised. Let $G(x,x')$ be the centralised inner product function for a distance $\delta(x,x')$, i.e.,

$$G(x,x') = -\frac{1}{2}\left[\delta(x,x')^2 - E_X\delta(x,X')^2 - E_{X'}\delta(X,x')^2 + \right.$$
$$\left. + E_{XX'}\delta(X,X')^2\right]$$

where X,X' are independent and identically distributed. Let us consider the eigen decomposition

$$f(x)^{1/2}G(x,x')f(x')^{1/2} = \sum_{n=1}^{\infty}\lambda_n u_n(x)u_n(x'),$$

where (λ_n,u_n) are eigenvalues, eigenfunctions of $f^{1/2}Gf^{1/2}$. Define $c_n(x) = f(x)^{-1/2}\sqrt{\lambda_n}u_n(x)$. Then

$$G(x,x') = \sum_{n=1}^{\infty}c_n(x)c_n(x')$$

and $c_n(x)$, $n \ge 1$, are uncorrelated and centered principal coordinates for the distance $\delta(x,x')$. Thus we can obtain orthogonal expansions by writing

$$G(X,x') = \sum_{n=1}^{\infty}\lambda_n c_n(X)c_n(x')$$

In particular, when

$$\delta(x,x') = \sqrt{|x - x'|},$$

we obtain the above continuous scaling solution by means of $h_n(x) = c_n(x) - c_n(a)$.

As a consequence, the random variable X itself can be expanded, e.g., as

$$X = a + \sum_{n=1}^{\infty} h_n(b) h_n(X) \qquad \text{(if } a \text{ is finite)},$$

a discrete version of the continuous expansion

$$X = a + \int_a^b X_t dt.$$

In general, the following expansions hold:

$$X = x_0 + \sum_{n=1}^{\infty} h_n(b)(h_n(X) - h_n(x_0)), \qquad x_0 \in (a,b).$$

$$X = x_0 + \sum_{n=1}^{\infty} (h_n(X)^2 - h_n(x_0) h_n(b)).$$

This general approach was proposed in [21], [22], [27] and it is proved that the geometric variability for δ (see Section 3.2) satisfies

$$V_\delta(X) = \frac{1}{2} E_{XX'} \delta^2(X, X') = \sum_{n=1}^{\infty} \lambda_n,$$

and the above expansions exist provided that $V_\delta(X)$ is finite.

5.4. Some expansions

For the uniform, exponential, and logistic distributions some expansions were found ([21], [28]) with principal dimensions:

1) $h_n(X) = \frac{\sqrt{2}}{n\pi}(1 - \cos n\pi X), \; \lambda_n = \frac{1}{(n\pi)^2}$,

   where X is uniform on [0,1] .

2) $h_n(X) = \frac{2J_0(\xi_n \exp(-X/2)) - 2J_0(\xi_n)}{\xi_n J_0(\xi_n)}, \quad \lambda_n = \frac{4}{\xi_n^2}$

   where X is exponential with mean 1. Here ξ_n is the n-*th* positive roof of J_1 and J_0, J_1 are the Bessel functions of the first order.

3) $h_n(X) = \sqrt{\frac{1}{n(n+1)}} \left[L_n(F(X)) + (-1)^{n+1}\sqrt{2n+1}\right],$

   $\lambda_n = \frac{1}{n(n+1)},$

   where X is standard logistic with $F(x) = (1+e^{-x})^{-1}$ and $L_n(x)$ is the Legendre polynomial on [0,1].

Further expansions were found for the Pareto [31], Laplace and normal distributions (unpublished manuscripts).

5.4. The usefulness of the method

As it has been noted [20], the expansion of the Cramer-von Mises statistics [36]

$$W^2 = \sum_{n=1}^{\infty} \frac{Y_n^2}{n^2 \pi^2},$$

where $Y_1, Y_2, \ldots$ are independent *N(0,1)*, is formally analogous to the expansion

$$X = \sum_{n=1}^{\infty} \frac{U_n^2}{n^2 \pi^2} \qquad X \text{ uniform on } [0,1],$$

where

$$U_n = \frac{\sqrt{2}}{n\pi}(1 - \cos nX), n \geq 1,$$

is a countable set of uncorrelated and identically distributed random variables. This suggests that these expansions may be used in goodness-of-fit assessment (notice that W^2 is the limit distribution of Cramér-von Mises statistics, used in deciding whether a sample comes from a specified distribution).

These expansions have been used [28] , [30], to distinguish the normal from the logistic distribution. Given a sample $x_1, x_2, \cdots, x_n$, they compared $h_k(x_i), i = 1, \cdots, n$, for $1 \leq k \leq 4$, to the principal dimensions $h_k(X)$, where X is logistic, and to $h_k(Y)$, where Y is normal. The relative position of the sample curve with respect to the theoretical one may help the user to distinguish both distributions.

To test stochastic dependence between two random variables X, Y, ([44]) the functions $(L_m(F(X)), L_n(G(Y)))$ were correlated, where $L_m(x)$ is the Legendre polynomial on [0,1], and F,G are the probability distributions functions of X and Y. However ([18]), this test is appropriate for marginal logistic distributions, but for other distributions (e.g., exponential), this test can be improved by using the principal directions of the marginal variables. Finally, a formula for the covariance between functions is given,

$$\operatorname{cov}(\alpha(X), \beta(Y)) = \int_{R^2} (H(x,y) - F(x)G(y)) d\alpha(x) d\beta(y)$$

where H is a bivariate distribution with marginals F, G [17] and these expansions can also be used in extending the probability plot, in constructing distributions with given marginals [16], [18], and in studying the asymptotic distribution of Rao's quadratic entropy ([48]).

Acknowledgements

This work has been supported in part by grants from Spanish DGCYT (BFM2000-0801) and the Generalitat de Catalunya (2001SGR00067).

References

[1] Amat L, Robert D, Besalu E and Carbodorca R (1998) Molecular Quantum Similarity Measures Tuned 3D QSAR - An Antitumoral Family Validation-Study. *Jour-*

nal of Chemical Information and Computer Sciences, 38, 624-631.

[2] Arenas C and Bernal M (1994) Multivariate approach to the classification of *genus Dianthus L. (Caryophyllaceae). In:* R. Gutierrez and M. J. Valderrama (Ed.), *Selected topics on stochastic modelling,* pp.215-222 (World Scientific, Singapore).

[3] Arenas C and Turbón D (1998) The usefulness of discrimination based on distances on human evolution. *Qüestiió,* 22, pp. 529-538.

[4] Arenas C, Cuadras CM and Fortiana J (1998) *Multicua: Non standard package of multivariate analysis, version 0.77.* Publicacions del Departament d'Estadística (nueva colección), nº 1, Barcelona, Spain (in Spanish).

[5] Arenas C, Escudero T, Mestres F, Coll MD and Cuadras CM (2000) CACROMOS: A computer program to reconstruct the position of chromosomes on the metaphase plate. *Hereditas,* 132*,* 157-159.

[6] Arenas C, Cuadras CM and Turbón D (2001) Joint representation of multivariate data with applications to human evolution (unpublished work).

[7] Cox TF and Cox MA (1994) *Multidimensional scaling.* Chapman and Hall, London.

[8] Cuadras CM (1988) Statistical Distances. *Estadística Española*, 30, 295-378.

[9] Cuadras CM (1988) Statistical Methods applied to the prehistoric reconstruction. *Munibe (Antropologia Arqueologia)* 6, 25-33.

[10] Cuadras CM (1989) Distance analysis in discrimination and classification using both continuous and categorical variables. *In:* Y. Dodge (Ed.), *Statistical Data Analysis and Inference,* pp. 459-473 (Elsevier Amsterdam, North Holland).

[11] Cuadras CM (1990) An eigenvector pattern arising in nonlinear regression. *Qüestiió* 14, 89-95.

[12] Cuadras CM (1992) Some examples of distance based discrimination. *Biometrical Letters*, 29, 1-18.

[13] Cuadras CM (1993) Interpreting an inequality in multiple regression. *The American Statistician,* 47(4), 256-258.

[14] Cuadras CM (1995) Increasing the correlations with the response variable may not increase the coefficient of determination: a PCA interpretation. *In*: E. Tiit, T. Kollo and H. Niemi (Ed.), *New Trends in probability and Statistics. Vol. 3. Multivariate Statistics and matrices in Statistics,* pp.75-83, (TEV, The Netherlands).

[15] Cuadras CM (1998) Multidimensional dependencies in classification and ordination. *In*: K. Fernández and A. Morineau (Ed.), *Analyses Multidimensionelles des Données,* pp. 15-25 (CISIA-CERESTA, Saint-Mandé France).

[16] Cuadras CM (2002) Correspondence analysis and diagonal expansions in terms of distribution functions. *Journal of Statistical Planning and Inference.* 103, 137-150.

[17] Cuadras CM (2002) On the covariance between functions. *Journal of Multivariate Analysis.* 81, 19-27.

[18] Cuadras CM (2002) Diagonal distribution via orthogonal expansions and test of independence. *In*: C.M. Cuadras, J. Fortiana and J. A. Rodriguez-Lallena (Ed.), *Distributions with given marginals and statistical modelling* pp. 35-42, (Kluwer A.P., Dordrecht).

[19] Cuadras CM and Arenas C (1990) A distance based regression model for prediction with mixed data. *Commun. Stat. A. Theory and Methods* 19, 2261-2279.

[20] Cuadras CM and Fortiana J (1993) Continuous metric scaling and prediction. *In*: C.M. Cuadras and C.R. Rao (Ed.), *Multivariate Analysis, Future Directions 2,* 47-66 (Elsevier Science Publishers B.V. Amsterdam, North-Holland).

[21] Cuadras CM and Fortiana J (1995) A continuous metric scaling solution for a random variable, *Journal of Multivariate Analysis,* 52, 1-14.

[22] Cuadras CM and Fortiana J (1996) Weighted Continuous metric scaling. *In*: A.K. Gupta and V.L. Girko, (Eds.), *Multidimensional Statistical Analysis and Theory of Random Matrices,* pp. 27-40, (The Netherlands).

[23] Cuadras CM and Fortiana J (1997) Continuous scaling on a bivariate copula. *In*: V. Benes and J. Stepan, (Eds). *Distributions with given marginals and moment problems*, pp. 137-142, Kluwer, Ac. Press.

[24] Cuadras CM and Fortiana J (1998) Generalized distance-based regression. *Communications to the Joint Meeting of the International Psychometric Society and the Classification Society of North America, Urbana-Champaign, Illinois, USA.*

[25] Cuadras CM and Fortiana J (1998) Typicality in discriminant analysis with mixed variables. *Data Science, Classification and Related Topics*, IFCS-98, Roma, 82-85.

[26] Cuadras CM and Fortiana J (1998) Visualizing categorical data with related metric scaling. *In*: J. Blasius and M. Greenacre (Ed.), *Visualisation of Categorical Data*, pp. 365-376 (Academic Press, London).

[27] Cuadras CM and Fortiana J (2000) The importance of geometry in Multivariate Analysis and some applications. *In*: C.R. Rao and G. Szekely (Ed.), *Statistics for the 21st Century*, pp. 93-108 (Marcel Dekker, New York).

[28] Cuadras CM and Lahlou Y (2000) Some orthogonal expansions for the logistic distribution. *Comm. Stat.-Theor. Meth.*, 29 (12), 2643-2663.

[29] Cuadras CM, Arenas C and Fortiana J (1996) Some computational aspects of a distance-based model for prediction. *Commun. Stat. Simulation and Computation* 25(3), 1-18.

[30] Cuadras CM and Cuadras D (2002) Orthogonal expansions and distintion between logistic and normal. *In*: C. Huber-Carol, N. Balakrishnan, M.S. Nikulin and M. Mesbah (Ed.), *Goodness-of-Fit tests and Model Validity*, pp. 325-338 (Birkäuser, Boston).

[31] Cuadras CM and Lahlou Y (2002) Principal components of the Pareto distribution. *In*: C.M. Cuadras, J. Fortiana and J. A. Rodriguez-Lallena (Ed.), *Distributions with given marginals and statistical modelling pp. 43-50*, (Kluwer A.P., Dordrecht).

[32] Cuadras CM, Fortiana J and Oliva F (1996) Representation of statistical structures, classification and prediction using multidimensional scaling. *In*: W. Gaul and D. Pfeifer (Ed.), *From Data to Knowledge*, pp. 20-31 (Springer-Verlag, Berlin).

[33] Cuadras CM, Fortiana J and Oliva F (1997) The proximity of an individual to a population with applications to discriminant analysis. *J. of Classification*, 14, 117-136.

[34] Cuadras CM, Atkinson RA and Fortiana J (1997) Probability densities from distances and discriminant analysis. *Statistics and Probability Letters,* 33, 405-411.

[35] Draper NR and Smith H (1981) *Applied Regression Analysis (second edition)*, J. Wiley, New York.

[36] Durbin J and Knott M (1972) Components of Cramér-von Mises Statistics. *I. Journal of the Royal Statistical Society*, B, 34, 290-307.

[37] Escudero T, Arenas C, Fuster C, Coll MD, Cuadras CM and Egozcue J (1998) Distribution of human chromosomes into two haploid sets in lymphocyte metaphases treated with colcemid: a multidimensional scaling approach, *Cytogenetic and Cell Genetics,* 81, p.116.

[38] Fortiana J and Cuadras CM (1997) A family of matrices, the discretized Brownian Bridge and distance-based regression. *Linear Algebra and its Applications,* 264, 173-188.

[39] Gower JC (1971) A general coefficient of similarity and some of its properties, *Biometrics,* 27, 857-874.

[40] Gower JC and Krzanowski WJ (1999) Analysis of Distance for Structured Multivariate Data and Extensions to Multivariate-Analysis of Variance. *Journal of the Royal Statistical Society series C-Applied Statistics*, 48, 505-519.

[41] Hill RC, Fomby TB and Johnson SR (1977) Component selection norms for principal component regression. *Commun. Stat. Theor. methods A*, 6, 309-334.

[42] Jakimiec M and Bartkowiak A (1994) Short-Term Solar-Flare Predictions by Distance-Based Regression. Bearalert Regions in 1988 and 1989 - Continuous Predictors. *Acta Astronomica*, 44, 115-140.

[43] Jollife IT (1986) *Principal Component Analysis*. Springer-Verlag, New York.

[44] Kallenberg W CM and Ledwina T (1999) Data-driven tests for independence. *J. Amer. Stat. Assoc.,* 94, 285-301.

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

[46] Krzanowski WJ (1988) *Principles of Multivariate Analysis: a User's Perspective.* Clarendon Press, Oxford.

[47] Krzanowski K and Marriott M (1995) *Multivariate Analysis*, Arnold, London.

[48] Liu Z and Rao CR (1995) Asymptotic distribution of statistics based on quadratic entropy and bootstrapping. *J. Statistical Planning and Inference*, 43, 1-18.

[49] Lott WF (1973) The optimal set of principal component restrictions on a least-squares regression. *Commun. Stat. Theory and Methods* 2(5), 449-464.

[50] Mardia KV, Kent JT and Bibby JM (1979) *Multivariate Analysis*. London. Academic Press.

[51] Montanari, A. and Mignani, S. (1994) Notes on the bias of dissimilarity indices for incomplete data sets: the case of archaeological classifications, *Qüestiió,* 18, pp. 39-49.

[52] Rao CR (1998) Geometry of circular vectors and Pattern recognition of shape of a boundary. *Proceedings of the national Academy of Sciences of The United States of America*, 95, 12783-12786.

[53] Robert D, Amat L and Carbodorca R (1999) 3-Dimensional Quantitative Structure-Activity-Relationships from Tuned Molecular Quantum Similarity Measures - Prediction of the Corticosteroid-Binding Globulin Binding-Affinity for a Steroid Family. *Journal of Chemical Information and Computer Sciences*, 39, 333-344.

[54] Robert D, Girones X and Carbodorca R (2000) Quantification of the Influence of Single-Point Mutations on Haloalkane Dehalogenase Activity - A Molecular Quantum Similarity Study. *Journal of Chemical Information and Computer Sciences* 40, 839-846.

[55] Rosés F (1990) Estudi de la "i" àtona en posició post-consonàntica. Univ. de Barcelona. Unpublished manuscript.

[56] Schoenberg IJ (1935) Remarks to Maurice Fréchet's article "Sur la definition axiomatique d'une classe d'espaces vectorielles distanciés applicables vectoriellement sur l'espace de Hilbert". *Annals of Mathematics*, 36, 724-732.

[57] Soofi ES (1988) Principal component regression under exchangeability. *Commun. Stat. Theor. Methods A*, 17, 1717-1733.

[58] Thisted RA (Ed.) (1994) *CIS Extended database.* American Statistical Association (Alexandria, VA), and the Institute of Mathematical Statistics (Hayward, CA).

[59] Valdés R (1992) Finding string distances. *Dr. Dobb's Journal* 17(4), 56-62.

[60] Villarroya A, Rios M and Oller JM (1995) Discriminant analysis algorithm- based on a distance function and on a Bayesian decision. *Biometrics*, 51, 908-919.

[61] Weisberg, S (1985) *Applied linear regression.* J. Wiley, New York

About the authors

Dr. Carles M. Cuadras was born in Figueres (Catalonia) in 1945. He was awarded a Master's degree in Mathematics by the University of Barcelona in 1968 and received a PhD in Mathematics in 1973. From 1974 to 1979 he performed scientific research at the CSIC. Since 1979 he has been Full Professor of Statistics at the University of Barcelona, where he leads a consolidated group of researchers in theoretical and applied statistics. His main research areas are multivariate analysis, distributions with given marginals, continuous and related scaling, biostatistics and distance-based models in regression and classifi-

cation. He has supervised 14 PhD theses on statistics and is author or co-author of more than 100 papers, most of them in international journals, and more than 40 invited and contributed papers to international meetings. He has edited two books (published by Elsevier and Kluwer) and is the author of six text books. He is a member of several Statistical Societies (International Biometric Society; International Statistical Institute; American Statistical Association, etc.). He has been president of the Spanish Region and Council Member of the International Biometric Society. He organized a national meeting on biostatistics (1984) and two international meetings on multivariate analysis (1992) and probability distributions with given marginals (2000).

Dra. Conchita Arenas was born in Barcelona in 1959. She was awarded a Master's degree in Mathematics by the University of Barcelona in 1982 and received a PhD in Mathematics (Statistics) in 1987. In 1989 she became Associate Professor at the University of Barcelona. Since 1982 she has worked at the Department of Statistics at the University of Barcelona and since 1988 she has been a member of the team led by Dr. C. M. Cuadras, working on multivariate analysis and distance-based models in regression. She is author or co-author of 17 articles in international journals, 19 contributed papers to international meetings and two university publications. She is a member of the International Biometric Society and other statistical societies. She is co-author of multivariate statistical software for non-standard methods.