Recent statistical methods based on distances

C. Arenas* and C. M. Cuadras
Departament d’Estadística. Facultat de Biologia. Universitat de Barcelona

Abstract

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 Department 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 mixed data, distances between observations, regression model, discriminant function, missing data, principal co-ordinate analysis, multidimensional scaling, related metric scaling.

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.
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 extension 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 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

\[
d(x, y) = \sqrt{\sum_{i=1}^{n} |x_i - y_i|}
\]

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 \( A \) 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 \( \mathbf{W} \), where \( \mathbf{W} \) 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 \( \mathbf{W} \), and gives a \( n \times n \) distance matrix \( \mathbf{D} = (d_{ij}) \). We suppose that \( \mathbf{D} \) is an Euclidean distance matrix. Let \( \mathbf{A} = (a_{ij}) \) the matrix with elements

\[
a_{ij} = -(d_{ij}^2) / 2
\]

and set \( \mathbf{B} = \mathbf{H} \mathbf{A} \mathbf{H} \) where \( \mathbf{H} = \mathbf{I} - 1/n \mathbf{1} \mathbf{1}' \) is the centring matrix. It is well-known ([50]) that \( \mathbf{B} \) is positive semi-definite and assuming rank \( \mathbf{B} = m \), the spectral decomposition of matrix \( \mathbf{B} \) is

\[
\mathbf{B} = \mathbf{U} \Lambda \mathbf{U}' = \mathbf{X} \mathbf{X}', \quad \text{where } \Lambda \text{ is diagonal and } \mathbf{X}
\]
A matrix $\mathbf{U} \Lambda^{1/2}$ is an $n \times n$ matrix of rank $m$. As the distances between the rows of $\mathbf{X}$ are the same as $d_p$, the following full DB-model is proposed:

$$y = \gamma_0 + 1 + X \gamma + e,$$

where $1$ is the vector of $1$’s, $\gamma_0$ is an unknown scalar parameter and $y$ is an unknown parameter vector of dimension $\rho$, and $y$ 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

$$y = \beta_0 + X_{(k)} \beta_{(k)} + e_{(k)}.$$

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 $\beta_0$ and $\beta_{(k)}$ are given by $\hat{\beta}_0 = \mathbf{y}$ and $\hat{\beta}_{(k)} = \mathbf{A}_{(k)}^{-1} \mathbf{x}_{(k)} \mathbf{y}$, where $\mathbf{A}_{(k)} = \text{diag} (\lambda_1, \ldots, \lambda_k)$. For computing the coefficient of determination a useful formula is

$$R^2 = \sum_{i=1}^{d} r^2 (Y, X_i),$$

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

$$y_{n+1} = \hat{\mathbf{z}} + \mathbf{x}_i^T \mathbf{A}_{(k)}^{-1} \mathbf{x}_{(k)} \mathbf{y} + \mathbf{z}^T \mathbf{A}_{(m-k)}^{-1} \mathbf{z} \mathbf{y},$$

with

$$\mathbf{x} = \begin{bmatrix} \mathbf{x}_{(k)} \\ \mathbf{z} \end{bmatrix}, \quad \mathbf{X} = \begin{bmatrix} \mathbf{X}_{(k)} \\ \mathbf{Z} \end{bmatrix}, \quad \Lambda = \begin{bmatrix} \mathbf{A}_{(k)} \\ \mathbf{A}_{(m-k)} \end{bmatrix}.$$

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, the DB model can be applied by taking the distance

$$d(x, y) = \sqrt{\sum_{j=1}^{p} |x_j - y_j|}$$

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 $x_1$, minutes of reaction at $x_2$ degrees Kelvin. The non-linear regression model is

$$Y = \exp \left\{- \theta_j x_1 \exp \left\{- \theta_j \left( \frac{1}{x_2} - \frac{1}{620} \right) \right\} + g \right\},$$

where $\theta_j, j = 1,2$ 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.**

<table>
<thead>
<tr>
<th></th>
<th>$R^2$</th>
<th>$C$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Classic method</td>
<td>0.755</td>
<td>2.501</td>
</tr>
<tr>
<td>DB-method</td>
<td>0.875</td>
<td>1.564</td>
</tr>
</tbody>
</table>
• [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 does not need a codification of the variables, it is very useful for real data, e.g., in DNA sequences and assignation of man-locuses. However, methods that do not suppose any probability distribution and that do not need a codification of the variables, it is very useful for real data, e.g., in DNA sequences and assignment of manuscripts or voice ([59]) 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 assignment 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 practical application of the method 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 \( X = (X_1, ..., X_p) \) be a random vector with values on some space \( E \subseteq \mathbb{R}^p \) and probability density \( f \) with respect to a suitable measure \( \lambda \). Suppose that \( \delta(\cdot) \) is a distance function on \( E \), i.e., such that \( \delta(x, y) = \delta(y, x) = \delta(x, x) = 0, \forall x, y \in E \). Suppose that

\[
V_\delta(X) = \frac{1}{2} \int \int_\mathbb{R} \delta^2(x, y) f(x) f(y) d\lambda(x) d\lambda(y)
\]

is finite. \( V_\delta(X) \), called geometric variability in [21], is the measure of dispersion of \( X \) with respect to \( \delta \), which reduces to the total variation \( \lambda(X) \) when \( \delta \) is the ordinary Euclidian distance, \( \Sigma \) being the covariance matrix of \( X \).

Given \( \omega_j \in \Omega \), the proximity function of the observation \( x_j = X(\omega_j) \) to the population represented by \( X \) is defined as

\[
\Phi(x_j) = \frac{1}{2} \int \int \delta^2(x, y) f(x) f(y) d\lambda(x) d\lambda(y) - V_\delta(X),
\]

i.e., \( \Phi(x_j) \) is the average of the squared distance from \( x_j \) to the population minus the geometric variability.

If \( x_1, ..., x_n \) is a sample from \( X \), the sampling version of a proximity function is

\[
\hat{\Phi}(x_i) = \frac{1}{n} \sum_{j=1}^{n} \delta^2(x_i, x_j) - \frac{1}{2n^2} \sum_{j=1}^{n} \sum_{k=1}^{n} \delta^2(x_i, x_k),
\]

Thus \( \Phi(x_j) \) 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 \( \delta \) between observations. We can obtain the proximity functions

\[
\hat{\Phi}_k(x) = \frac{1}{n_k} \sum_{j=1}^{n_k} \delta^2(x_j, x) - \frac{1}{2n_k} \sum_{j=1}^{n_k} \sum_{k=1}^{n_k} \delta^2(x_j, x_k), \quad k = 1, ..., g,
\]

where:

\( x \) is the observation of \( X \) on one individual \( \omega \in \Omega \), \( \cdot \) a distance function within \( \Omega \) or \( \cdot \) a distance function between \( \Omega \) or \( \cdot \) a distance function between two observations \( \omega \) and \( \omega' \).

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

Allocate \( \omega \) to \( \Omega_k \) if \( \hat{\Phi}_k(x) = \min \{ \hat{\Phi}_1(x), ..., \hat{\Phi}_g(x) \} \).

In [33] it was proved that each \( \hat{\Phi}_k(x) \) could be interpreted as a squared distance from \( x \) to \( \Omega_k \). 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 \( \delta \). 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 a- or an i-, appearing after a consonant in Catalan, should be pronounced as monosyllabic (\( \omega_i \)) or bisyllabic (\( \omega_a \)). 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 algo-
rithm 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 \( n \times n \) distance matrices \( D_i = (d_{ij}) \), \( k = 1,2 \), and its corresponding \( n \times n \) inner product matrices \( B_k = (b_{ij}) \), \( k = 1,2 \), related to \( D_k \) by

\[
d_{ij}^2 = b_{ii(k)} + b_{jj(k)} - 2b_{ij(k)}
\]

Let \( B_k = U_k \Lambda_k U_k^T \) be the spectral decomposition of \( B_k \). That is, \( U_k \) contains unitary eigenvectors and \( \Lambda_k \) is diagonal with the eigenvalues of \( B_k \). Then the matrix with the principal co-ordinates is \( X_k = U_k \Lambda_k^{1/2} \), which satisfies \( B_k = X_k X_k^T \). The problem is now to find an average matrix \( B \) summarising the information contained in the matrices \( B_k \), and then to find \( X \) such that \( B = XX^T \). 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:

\[
B = B_1 + B_2 - \frac{1}{2} (B_1^{1/2} B_2^{1/2} + B_2^{1/2} B_1^{1/2}),
\]

where \( B_k^{1/2} = U_k \Lambda_k^{1/2} U_k^T \). The final representation of \( n \) individuals can be obtained by using the co-ordinates of the matrix \( X \) such that \( B = XX^T \). This definition can be justified as follows.

Let \( D = (d_{ij}) \) the joint distance matrix related to \( B \), i.e.,

\[
d_{ij}^2 = b_{ij(k)} + b_{jj(k)} - 2b_{ij(k)} \text{ or } B = HAH (\text{see above}). \text{ It can easily be proved that:}
\]

1) If \( D_1 = D_2 \) then also \( D = D_1 = D_2 \).
2) If \( X_i^T X_i = 0 \) then \( D^{(3)} = D^{(1)} + D^{(2)} \), where

\[
D^{(2)} = (d_{ij}^2), D^{(1)} = (d_{ij(k)}^2), \ k = 1,2.
\]

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 \( D_1 \) and \( D_2 \).

For two-dimensional representation, we take the matrix \( X_{2n} \) with \( n \) rows and 2 columns which best fits \( X \) in the least square sense. This matrix is \( X_{2n} = U_1 \Lambda_{2n}^{1/2} \), where \( U_{2n} \) and \( \Lambda_{2n} \) 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 Ya- mana, 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).](image-url)
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 \( D = (d_{ij}) \), \( i, j \in I \), points \( P \in \mathbb{R}^d \), such that the distances between points give \( D \), i.e., \( d(P_i, P_j) = d_{ij} \), 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 \( x \rightarrow q(x) \in E \), where \( E \) is a real separable Hilbert space with quadratic norm \( || \cdot || \) such that \( \delta(x, x') = ||q(x) - q(x')|| \). \( x, x' \in I \). We may identify \( q(x) \) with \( Q(x) \), where for \( x \in I \), \( Q(x) = (Q_1(x), Q_2(x), \ldots) \) are the Euclidean coordinates such that

\[
\delta^2(x, x') = \sum_{i=1}^{\infty} (Q_i(x) - Q_i(x'))^2, \quad x, x' \in I.
\]

To find \( q \) and an optimal countable representation \( Q(x) \) of \( q(x) \) 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 \( \lambda_n, \psi_n \) 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 \( h_n(x) = \frac{1}{a} \int_a^x \psi_n(t) dt \).

If \( X \) is the indicator of \( [X > t] \), i.e., \( x \) is an observation of \( X \), then \( X_t = 0 \) for \( x \leq t \) and \( X_t = 1 \) for \( x > t \). Then

\[
\delta^2(x, x') = \int (x - x')^2 dt
\]

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

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

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

Thus:

\[
\var(h_m(X)) = \lambda_m, \quad \text{cov}(h_m(X), h_n(X)) = 0 \quad \text{for} \quad m \neq n,
\]

\[
\text{trace}(K) = \sum_{n=1}^{\infty} \lambda_n
\]

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.,

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

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 \( (\lambda_n, u_n) \) are eigenvalues, eigenfunctions of \( G^{1/2} F^{1/2} \). 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 \( c_n(x), n \geq 1, \) 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
\( h_n(x) = c_n(x) - c_0(x) \).

As a consequence, the random variable \( X \) itself can be expanded, e.g., as
\[
X = a + \sum_{n=1}^{\infty} h_n(b) h_n(X) \quad \text{(if } a \text{ is finite)},
\]
a discrete version of the continuous expansion
\[
X = a + \int_a^b X \, dt.
\]
In general, the following expansions hold:
\[
X = x_0 + \sum_{n=1}^{\infty} h_n(b)(h_n(X) - h_n(x_0)), \quad x_0 \in (a, b),
\]
\[
X = x_0 + \sum_{n=1}^{\infty} (h_n(X))^2 - h_n(x_0) h_n(b).
\]
This general approach was proposed in [21], [22], [27] and
it is proved that the geometric variability for \( \delta \) (see Section 3.2) satisfies
\[
V_\delta(X) = \frac{1}{2} E_{X',X''}(X | X'), \quad \sum_{n=1}^{\infty} \lambda_n,
\]
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) \( h_n(X) = \sqrt{\frac{2}{n\pi}} (1 - \cos n\pi X), \quad \lambda_n = \frac{1}{(n\pi)^2}, \)
   where \( X \) is uniform on \([0,1]\).

2) \( h_n(X) = \frac{2 J_0(\xi_n) \exp(-X/2)}{\xi_n J_0(\xi_n)} - 2 J_0(\xi_n), \quad \lambda_n = \frac{4}{\xi_n^2}, \)
   where \( X \) is exponential with mean 1. Here \( \xi_n \) is the \( n \)-th positive roof of \( J_0 \) and \( J_0, J_1 \) are the Bessel functions of the first order.

3) \( h_n(X) = \frac{1}{n(n+1)} L_n(F(X)) + (-1)^n I \sqrt{2n+1}, \quad \lambda_n = \frac{1}{n(n+1)}, \)
   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
\[
X = \sum_{n=1}^{\infty} \frac{U_n^2}{n^2 \pi^2} \quad \text{X uniform on } [0,1],
\]
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 \( x_1, x_2, \ldots, x_n \) they compared \( h_n(x), \lambda = 1, \ldots, n \), for \( 1 \leq k \leq 4 \), to the principal dimensions \( h_n(X) \), where \( X \) is logistic, and to \( h_n(Y) \), 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,
\[
cov(\alpha(X), \beta(Y)) = \int \frac{H(x,y) - F(x)G(y)}{\xi^2} da(x)db(y)
\]
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 DGICYT (BFM2000-0801) and the Generalitat de Catalunya (2001SGR00067).

References

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


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-
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 textbooks. 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.