1

Multidimensional Dependencies in Classification and Ordination

Carles M. Cuadras[1]

ABSTRACT The relations between two distance matrices on the same finite set are analyzed, via metric scaling, by correlating principal axis. Some applications are given and illustrated with examples.

1.1 Introduction

Dissimilarities, similarities and distances are fundamental concepts in multidimensional scaling and related topics. Euclidean and Mahalanobis distance also play a basic role in techniques such as regression and discrimination. In many cases, distances are computed by observing variables on individuals, but in general, the dependence between variables is not taken into account. Mahalanobis, and its extension Rao's distance, is an important exception. This distance between two observations $\boldsymbol{x}, \boldsymbol{y}$, say,

$$(\boldsymbol{x}-\boldsymbol{y})'\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{y})$$

depends on the covariance matrix Σ, hence its computation is not possible when the variables are categorical, binary or mixed. In such situations, distances are obtained from similarities such as Jaccard (binary), matching (categorical) and Gower (mixed) coefficients. Other coefficients are possible, but none of them has the property of including the relationships among variables.

Motivated by this problem, [CF97b] recently introduced *related metric scaling*, a new multidimensional scaling method to represent objects when two distances are defined on them. This method is based on the construction of a joint distance that has some compatible properties, especially identifying and discarding redundant information.

[1]University of Barcelona. Department of Statistics. Diagonal, 645, 08028 Barcelona. Spain

1.2 Metric scaling

Given a set Ω of n individuals or objects labelled $\{1, 2, \ldots, n\}$, say, and an $n \times n$ Euclidean distance matrix $\Delta = (\delta_{ij})$, where $\delta_{ij} = \delta(i, j) \geq 0$ is a distance measure between i and j, let us recall that the Euclidean coordinates, in some R^p, are the rows of an $n \times p$ matrix $\boldsymbol{X}$ such that

$$\boldsymbol{B} = \boldsymbol{X}\boldsymbol{X}'. \tag{1}$$

Matrix $\boldsymbol{B}$ is obtained from Δ by $\boldsymbol{B} = \boldsymbol{H}\boldsymbol{A}\boldsymbol{H}$, where $\boldsymbol{H}$ is the centring matrix and $\boldsymbol{A} = -(\delta_{ij}^2)/2$. Let the spectral decomposition $\boldsymbol{B} = \boldsymbol{U}\Lambda\boldsymbol{U}'$, where Λ is diagonal and contains the eigenvalues of $\boldsymbol{B}$ in decreasing order. Then a suitable solution of (1) is

$$\boldsymbol{X} = \boldsymbol{U}\Lambda^{1/2}. \tag{2}$$

The rows of $\boldsymbol{X}$ are called *principal coordinates*, their Euclidean distances reproduce Δ and the columns of $\boldsymbol{X}$, understood as variables, can be interpreted as principal components: they are uncorrelated, the variances are maximum, and the first coordinates give the best fit to the initial distance, i.e.,

$$\delta_{ij}^2 \simeq d_{ij}^2(2) = (x_{i1} - x_{j1})^2 + (x_{i2} - x_{j2})^2$$

if a two-dimensional representation by means of the coordinates (x_{i1}, x_{i2}), $i = 1, \ldots, n$ is desired.

Note that $\boldsymbol{B}$ is positive semi-definite (p.s.d.), i.e., $\boldsymbol{B} \geq 0$, the columns of $\boldsymbol{X} = \boldsymbol{U}\Lambda^{1/2}$ are eigenvectors of $\boldsymbol{B}$ and only $q < p = \text{rank}(\boldsymbol{B})$ principal axes are used to represent the n objects $\{1, 2, \ldots, n\}$. The rows of $\boldsymbol{U}$ are called *standard coordinates*.

For an extensive description of metric scaling, see [CC94]. Finally, let us write

$$\Delta \sim \boldsymbol{B}$$

to indicate that $\boldsymbol{B}$ is related to Δ and recall that $\boldsymbol{B}$ is p.s.d. if and only if Δ is a Euclidean distance matrix. The relation between $\Delta = (\delta_{ij})$ and $\boldsymbol{B} = (b_{ij})$ is

$$\delta_{ij}^2 = b_{ii} + b_{jj} - 2b_{ij},$$

or, in matrix form

$$\Delta^{(2)} = \mathbf{1}\ \boldsymbol{b}' + \boldsymbol{b}'\mathbf{1} - 2\boldsymbol{B},$$

where $\mathbf{1}$ is the vector of ones, $\Delta^{(2)} = (\delta_{ij}^2)$ and $\boldsymbol{b} = (b_{11} \ldots b_{nn})'$.

The *rank order* of $\Delta = (\delta_{ij})$ is

$$\delta_{i_1 j_1} \leq \delta_{i_2 j_2} \leq \cdots \leq \delta_{i_m j_m} \qquad m = n(n-1)/2.$$

Although the rank order is a concept rather used in non-metric multidimensional scaling, we will study its invariance in related metric scaling. This invariance is a convenient property in proximity analysis. The rank order through a principal axis (a column of $\boldsymbol{X}$) can be similarly defined using the distances defined by the corresponding principal coordinates.

1.3 Joining two distances

Suppose that, from two different sources of information, we obtain two Euclidean distance matrices $\Delta_1 = (\delta_1(i,j))$, $\Delta_2 = (\delta_2(i,j))$ on the same set Ω labelled $\{1,2,\ldots,n\}$ (Example: n political parties, $\delta_1(i,j)$ measures ideology discrepancy, $\delta_2(i,j)$ the times that i and j do not vote the same in an assembly).

Let $\Delta_\alpha \sim \boldsymbol{B}_\alpha = \boldsymbol{X}_\alpha \boldsymbol{X}'_\alpha$, $\alpha = 1,2$, where $\boldsymbol{X}_1, \boldsymbol{X}_2$ are the matrices of principal coordinates. The *orthogonality* between Δ_1 and Δ_2 is defined by

$$\boldsymbol{X}'_1 \boldsymbol{X}_2 = 0$$

Orthogonality implies total incorrelation between principal axes.

Our aim is to define a joint distance δ_{12} from δ_1, δ_2. The average, or better the sum

$$\delta^{*2} = \delta_1^2 + \delta_2^2 \tag{3}$$

is a simple construction, with some advantages, but it implicity presupposes orthogonality. A more convenient definition is now given.

DEFINITION: Let $\boldsymbol{x}_i(\alpha)', \boldsymbol{x}_j(\alpha)'$ be the principal coordinates of i and j corresponding to the distance Δ_α, $\alpha = 1,2$. The joint distance matrix Δ_{12} is defined as

$$\delta_{12}^2(i,j) = \delta_1^2(i,j) + \delta_2^2(i,j) - \tau_{12}(i,j), \tag{4}$$

where

$$\tau_{12}(i,j) = (\boldsymbol{x}_i(1) - \boldsymbol{x}_j(1))' \, \Lambda_1^{-1/2} \boldsymbol{X}'_1 \boldsymbol{X}_2 \Lambda_2^{-1/2} \, (\boldsymbol{x}_i(2) - \boldsymbol{x}_j(2)) \, .$$

It can be proved that the joint distance matrix Δ_{12} is such that $\Delta_{12} \sim \boldsymbol{B}_{12}$ where

$$\boldsymbol{B}_{12} = \boldsymbol{B}_1 + \boldsymbol{B}_2 - \left(\boldsymbol{B}_1^{1/2} \boldsymbol{B}_2^{1/2} + \boldsymbol{B}_2^{1/2} \boldsymbol{B}_1^{1/2} \right) /2 \tag{5}$$

with

$$\boldsymbol{B}_\alpha^{1/2} = \boldsymbol{U}_\alpha \Lambda_\alpha^{1/2} \boldsymbol{U}'_\alpha = \boldsymbol{X}_\alpha \Lambda_\alpha^{-1/2} \boldsymbol{X}'_\alpha, \quad \alpha = 1,2.$$

The term τ_{12} encodes the dependence or redundancy between the two distances, and furnishes the joint distance with properties of compatibility with the rank order, the invariance of principal axes, the correlation between coordinates.

1.4 Properties of the joint distance

The joint distance (4) has some interesting properties. In view of these, the construction of δ_{12} has some analogies with the construction of probability distributions with given marginals [Cua92]. Here δ_1, δ_2 play the role of marginal distances.

1) If $\delta_1 = 0$ then $\delta_{12} = \delta_2$.

   If $\delta_2 = 0$ then $\delta_{12} = \delta_1$.

2) If $\delta_1 = \delta_2$ then $\delta_{12} = \delta_1 = \delta_2$.

3) If δ_1 and δ_2 are orthogonals, then

$$\delta_{12}^2 = \delta_1^2 + \delta_2^2. \tag{6}$$

Comment: Properties 1) and 2) show that the joint distance do not change when the second distance is zero or both distances are identical. Property 3) is a distance version of the Pythagoras theorem.

4) If $\delta_{ij}(2) = c$, $i \neq j$, where $c > 0$ is a constant, then δ_1 and δ_{12} have the same rank order.

5) If δ_1 and δ_2 have a common principal axis, i.e., $\boldsymbol{B}_1$ and $\boldsymbol{B}_2$ has a common eigenvector, with eigenvalues $\lambda_k(1), \lambda_k(2)$ respectively, then δ_{12} shares the same principal axis with eigenvalue

$$\lambda_k = \lambda_k(1) + \lambda_k(2) - (\lambda_k(1)\lambda_k(2))^{1/2}. \tag{7}$$

Comment: 4) says that if δ_2 is not informative δ_{12} essentially preserves δ_1. Regarding 5), note that constructing δ_{12} using (3) also shares a common axis, but the eigenvalue is $\lambda_k(1) + \lambda_k(2)$. Thus the inertia becomes inflated. Instead, (4) gives $\lambda_k = \lambda_k(1) = \lambda_k(2)$ if both eigenvalues are equal, see (7). Also note that the rank order defined by this axis is invariant.

6) If δ_1 and δ_2 are Euclidean, with related coordinates $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$, then δ_{12} is also Euclidean and does not depend on the coordinates.

Comment: The formula giving $\boldsymbol{B}_{12}$ ensures that δ_{12} is independent of the coordinates. Moreover $\boldsymbol{B}_{12} \geq 0$ as a consequence of

$$\boldsymbol{B}_{12} - \left(\boldsymbol{B}_1^{1/2} - \boldsymbol{B}_2^{1/2}\right)^2 = \left(\boldsymbol{B}_1^{1/2} + \boldsymbol{B}_2^{1/2}\right)/2 \geq 0$$

7) Let $\boldsymbol{R}_{12}$ be the $p_1 \times p_2$ matrix whose entries are the correlations between the p_1 columns of $\boldsymbol{X}_1$ and the p_2 columns of $\boldsymbol{X}_2$ and let $\boldsymbol{R}_{21} = \boldsymbol{R}_{12}'$. Then

$$\boldsymbol{B}_{12} = \boldsymbol{X}_1\boldsymbol{X}_1' + \boldsymbol{X}_2\boldsymbol{X}_2' - \frac{1}{2}\left(\boldsymbol{X}_1\boldsymbol{R}_{12}\boldsymbol{X}_2' + \boldsymbol{X}_2\boldsymbol{R}_{21}\boldsymbol{X}_1'\right).$$

Comment: This correlation matrix is given by

$$\boldsymbol{R}_{12} = \boldsymbol{U}_1'\boldsymbol{U}_2,$$

where $\boldsymbol{U}_1$ and $\boldsymbol{U}_2$ are the standard coordinates.

If $p_1 = p_2$, a global measure of association between Δ_1 and Δ_2 is

$$\eta^2 = \det(\boldsymbol{R}_{12})^2.$$

This measure reduces to the squared correlation coefficient between $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ when $p_1 = p_2 = 1$.

8) $0 \leq \eta^2 \leq 1$,
$\eta^2 = 0$ if δ_1 and δ_2 are orthogonal,
$\eta^2 = 1$ if $\delta_1 = \delta_2$.

1.5 Joining two classifications

Suppose that (C_1, γ_1) and (C_2, γ_2) are two hierarchical clustering schemes on the same set $\Omega = \{1, \ldots, n\}$.

That is, C_1 is a hierarchical structure of nested clusters and γ_1 is a level function satisfying $\gamma_1(c) \leq \gamma_1(c')$ if $c \subset c' \in C_1$. It is well-known that (C_1, γ_1) is related to an ultrametric distance u_1 [Joh67]. Similarly (C_2, γ_2) is related to u_2. Reciprocally, given an ultrametric distance u we can construct a hierarchical clustering scheme (C, γ).

Suppose that we wish to gather two hierarchies $(C_1, \gamma_1), (C_2, \gamma_2)$ in order to obtain a possibly more objective classification represented by (C_{12}, γ_{12}). One way to do this is simply to construct the joint distance u_{12}, to fit an ultrametric u_{12}^* to u_{12}, and to build (C_{12}, γ_{12}) from u_{12}^*.

The joint distance can preserve certain clusters. Suppose that u_1 defines a clustering

$$\Omega = \Omega_1 + \cdots + \Omega_k + \cdots + \Omega_g,$$

where each Ω_k is a maximal cluster and $\#(\Omega_k) = n_k > 1$, for some k, where $n_1 + \cdots + n_g = n$, and

$$h_k = u_1(i, j) \quad i \neq j \in \Omega_k$$

are the common distances. [CO87] showed that, related with Ω_k, there is a matrix $\boldsymbol{X}_k$ of principal coordinates whose columns define $n_k - 1$ principal axes with common variances

$$\lambda_k = \frac{1}{2} h_k^2.$$

If u_2 defines another clustering having a coincident cluster Ω_k, then it defines the same principal axes with common variances λ_k'. Consequently, from property 5), these axes remain with the joint distance u_{12}. Thus Ω_k is also a cluster related to u_{12}. In general, the joint classification does not break common clusters if there are several.

1.6 Joining probability densities

Improving the construction of probability densities from distances is another application of the joint distance (4). Let $\boldsymbol{X}$ be a random vector with density $f(\boldsymbol{x})$ with respect to a suitable measure, e.g., Lebesgue measure. [CF95] define the geometric variability of $\boldsymbol{X}$ with respect to a distance δ as

$$V_\delta(\boldsymbol{X}) = \frac{1}{2}\int_S \delta^2(\boldsymbol{x},\boldsymbol{y}) f(\boldsymbol{x}) f(\boldsymbol{y}) d\boldsymbol{x}\, d\boldsymbol{y},$$

where S is the support of $\boldsymbol{X}$. [CCO97] introduced and studied the proximity function to apply in discrimination

$$\phi^2_\delta(\boldsymbol{x}) = \int_S \delta^2(\boldsymbol{x},\boldsymbol{y}) f(\boldsymbol{y}) d\boldsymbol{y} - V_\delta(\boldsymbol{X}), \qquad \boldsymbol{x} \in S,$$

which can be estimated without knowing $f(\boldsymbol{x})$, and used as a discriminant function with some advantages.

By affine transformation $\delta^2 \longrightarrow a\delta^2 + b$ giving $\phi^2_\delta \longrightarrow a\phi^2_\delta + b/2$, we can consider the probability density generated by δ

$$f_\delta(\boldsymbol{x}) = \exp\left(-\phi^2_\delta(\boldsymbol{x})\right),$$

i.e., choosing a, b such that $\int_S f_\delta(\boldsymbol{x}) d\boldsymbol{x} = 1$.

To compare f to f_δ, [CCF97b] showed that

$$I(f; f_\delta) = V_\delta(\boldsymbol{X}) - H(f) \geq 0,$$

where $I(f; f_\delta)$ is the Kullback-Leibler divergence between f and f_δ and $H(f)$ is the Shannon entropy. When $V_\delta(\boldsymbol{X})$ is close to $H(f)$, this indicates the agreement between the true density f and f_δ.

Suppose we now have two dependent random vectors $\boldsymbol{X}_1, \boldsymbol{X}_2$ with related distances δ_1, δ_2, respectively. So $\delta_\alpha(\boldsymbol{x},\boldsymbol{y})$ is the distance between observations of $\boldsymbol{X}_\alpha$, $\alpha = 1, 2$. Let us consider a joint distance for the joint random vector $\boldsymbol{X} = (\boldsymbol{X}_1, \boldsymbol{X}_2)$. The use of distance (3) leads to the proximity function

$$\phi^2_{\delta^*}(x_1, x_2) = \phi^2_{\delta_1}(x_1) + \phi^2_{\delta_2}(x_2).$$

The probability density generated by δ^* is then

$$\begin{aligned} f_{\delta^*}(x_1, x_2) &= \exp\left(-\phi^2_{\delta^*}(x_1, x_2)\right) \\ &= f_{\delta_1}(x_1) f_{\delta_2}(x_2), \end{aligned}$$

i.e., δ^* implicitly assumes independence of $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$.

Alternatively, suppose that we consider the joint distance δ_{12} defined in (4). Then it can be proved that

$$\begin{aligned} V_{\delta^*}(\boldsymbol{X}) &= V_{\delta_1}(\boldsymbol{X}_1) + V_{\delta_2}(\boldsymbol{X}_2) \\ &\geq V_{\delta_{12}}(\boldsymbol{X}) \geq H(f). \end{aligned}$$

Thus the density constructed from δ_{12}

$$f_{\delta_{12}}(x_1, x_2) = \exp\left(-\phi^2_{\delta_{12}}(x_1, x_2)\right),$$

is closer to f, the true density of $(\boldsymbol{X}_1, \boldsymbol{X}_2)$. See [CF97a] for further details.

1.7 Related metric scaling

The graphical representation of Ω using principal coordinates, computed from a joint distance Δ_{12}, obtained from Δ_1, Δ_2 using (4), is the objective of the related metric scaling. The eigendecomposition of $\boldsymbol{B}_{12}$, see (5), gives the principal coordinates.

To achieve a consistent representation, both distance matrices must have the same geometric variability:

$$\frac{1}{n^2}\sum_{i<j}\delta^2_{ij}(1) = \frac{1}{n^2}\sum_{i<j}\delta^2_{ij}(2).$$

This equality is possible by multiplying one distance by a suitable constant, which amounts to a change of measurement unit, but without changing the representation.

A generalization is as follows. Suppose that $\Delta_1, \ldots, \Delta_g$ are $g > 2$ distance matrices of order $n \times n$ on the same finite set. Let $\Delta_\alpha \sim \boldsymbol{B}_\alpha$, $\alpha = 1, \ldots, g$ and consider

$$\boldsymbol{B} = \sum_{\alpha=1}^{g}\boldsymbol{B}_\alpha - \frac{1}{g}\sum_{\alpha\neq\beta=1}^{g}\boldsymbol{B}_\alpha^{1/2}\boldsymbol{B}_\beta^{1/2}. \tag{8}$$

It can be proved that $\boldsymbol{B}$ is p.s.d., and has some interesting properties. The joint distance matrix is Δ such that $\Delta \sim \boldsymbol{B}$. If $\boldsymbol{X}_1, \cdots, \boldsymbol{X}_g, \boldsymbol{X}$ are the principal coordinates obtained from $\Delta_1, \cdots, \Delta_g, \Delta$, the related metric scaling representation or principal coordinate representation of Ω from Δ using $\boldsymbol{X}$ satisfies:

1) If $\Delta_1 = \cdots = \Delta_g$ then $\boldsymbol{X}_1 = \cdots = \boldsymbol{X}_g = \boldsymbol{X}$.

2) If $\boldsymbol{X}'_\alpha\boldsymbol{X}_\beta = 0$, $\alpha \neq \beta = 1, \cdots, g$, then $\boldsymbol{X} = [\boldsymbol{X}_1; \cdots; \boldsymbol{X}_g]$.

3) In general, $\boldsymbol{X}$ provides an average representation which takes into account the redundancy of the marginal distances.

The redundancy can be measured by using $\theta = g(1-\xi)/(g-1)$, where

$$\xi = \frac{tr(\boldsymbol{B})}{\sum_{i=1}^{g} tr(\boldsymbol{B}_i)}.$$

We have $0 \leq \theta \leq 1$, with $\theta = 0$ or $\theta = 1$ if there is orthogonality or equality, respectively.

Figure 1 is a simple illustration. The faces are described by the distances between landmark points and the 3 cases above (equality, orthogonality and redundancy) are well represented. Note that we need 4 dimensions in the second case (this is expressed displaying both faces), that is, we cannot obtain a joint 2-dimensional representation of these two orthogonal faces.

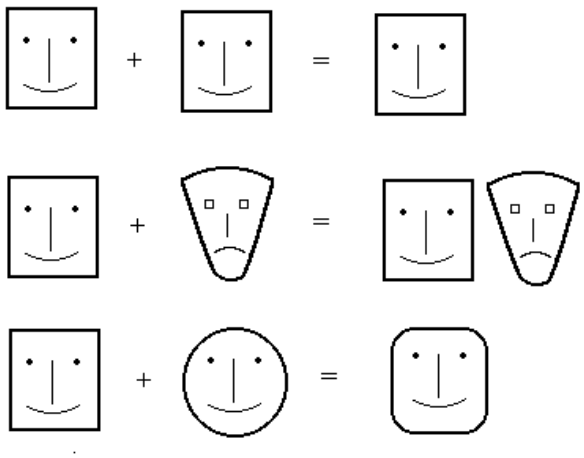

Figure 1. Related metric scaling representation of two faces showing equality, orthogonality and redundancy.

1.8 Two examples

Suppose we have five political parties labelled R, C, S, I, P. The first distance matrix Δ_1 measures diferencies according to some sociopolitical variables. $\Delta_2 = (\delta_{ij}(2))$ is such that $\delta_{ij}(2)$ is the percentage of occasions on which i and j vote differently in an assembly during one year.

Figure 2 is a multidimensional scaling (MDS) representation of the parties. The right part of the figure is a classification using an ultrametric tree. Distance Δ_1 is used in both cases. Note two maximal clusters $\{\mathrm{I},\mathrm{S}\}$, $\{\mathrm{R},\mathrm{C}\}$ and that P is an isolated object, joining the other parties at a higher level.

Figure 3 is the representation using Δ_2. A major agreement between I and P closes both and now P forms a cluster with C, R.

Figure 4 uses the joint distance matrix Δ_{12}. As for the MDS representation, it is an 'average' of the other two representations. Note that Δ_{12} preserves the clusters $\{\mathrm{I},\mathrm{S}\}$ and $\{\mathrm{R},\mathrm{C}\}$. For other examples, see [CF97b].

.

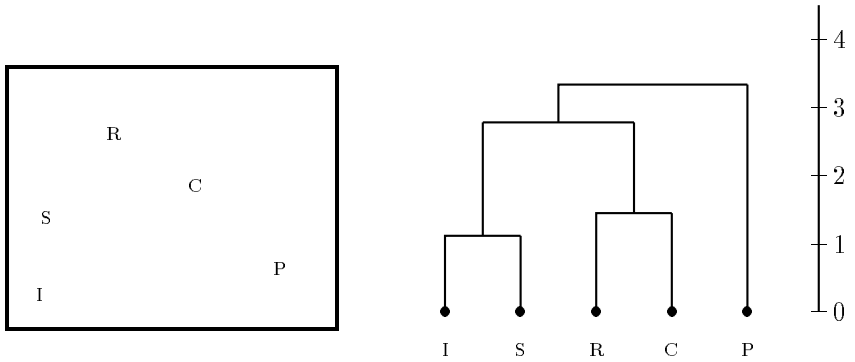

Figure 2. Representation of five political parties according to their ideology.

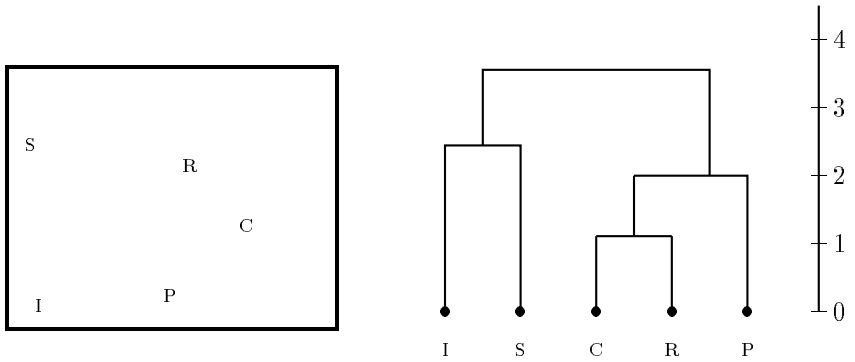

Figure 3. Representation according the agreement in voting.

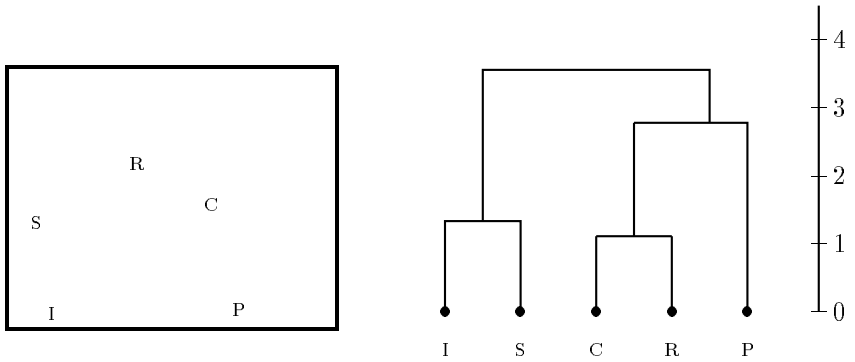

Figure 4. Related metric scaling representation and classification using a joint distance.

Finally, the joint distance related to $\boldsymbol{B}$ in (8) has been applied to obtain a joint representation of $g = 50$ pictures of metaphase planes of the human chromosomes. Figure 5 is the related metric scaling representation, giving an ideal position of the chromosome pairs. [CCF97a].

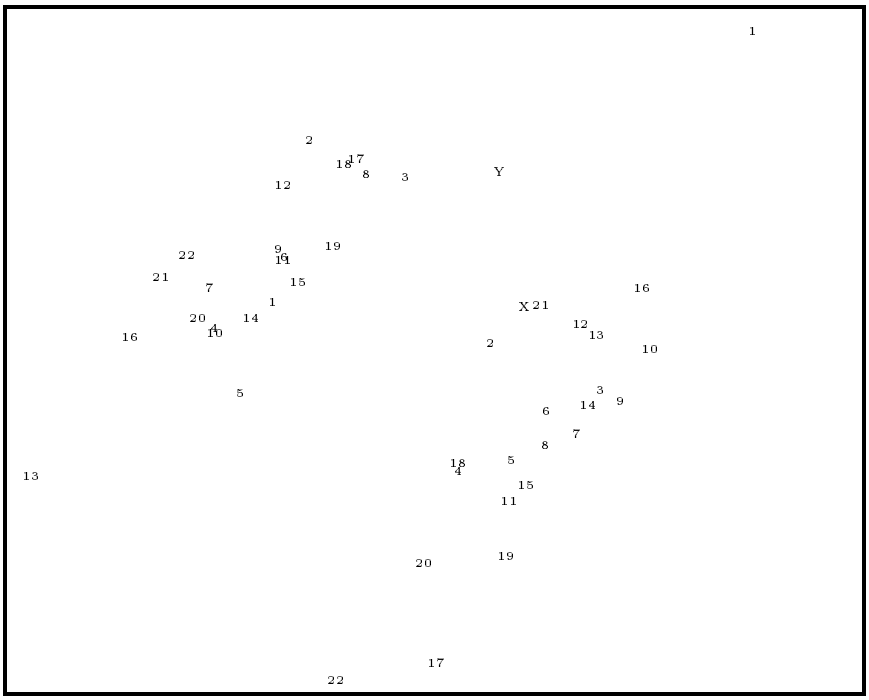

Figure 5. Ideal position of the 23 pairs of human chromosomes in the metaphase plane obtained by related metric scaling from g=50 pictures.

1.9 Conclusions

It has been shown, theoretically and with examples, that the joint distance (4) obtained from two given distances on the same finite set provides a joint representation, preserving the inertia of the principal axes and the common clusters. The construction of probability densities related to two random vectors is another application. This procedure can be generalized to $g > 2$ distances and some measures of association and redundancy can be obtained.

Acknowledgements: Work supported in part by grants CGYCIT PB96-1004-C02-01 and 1997SGR-00183.

1.10 References

[CC94] T.F. Cox and M.A. Cox. *Multidimensional Scaling.* Chapman & Hall, London, 1994.

[CCF97a] C.M. Cuadras, C. Arenas, M.D. Coll, T. Escudero and C. Fuster. MDS position of the human chromosomes in the metaphase plane. *Proceedings VI Conferencia Española de Biometria, Córdoba*, pages 211–212, 1997.

[CCF97b] C.M. Cuadras, R.A. Atkinson and J. Fortiana. Probability densities from distances and discriminant analysis. *Statistics & Probability Letters*, 32:99–103, 1997.

[CCO97] C.M. Cuadras, J. Fortiana and F. Oliva. The proximity of an individual to a population with applications to discriminant analysis. *J. of Classification*, 14:117–136, 1997.

[CF95] C.M. Cuadras and J. Fortiana. A continuous metric scaling solution for a random variable. *J. of Multivariate Analysis*, 51:1–14, 1995.

[CF97a] C.M. Cuadras and J. Fortiana. *Distributions with given marginals and moment problems*, chapter Continuous scaling on a bivariate copula, pages 137–142. V. Benes and J. Stepan, Kluwer Ac. Press, Prague, 1997.

[CF97b] C.M. Cuadras and J. Fortiana. *Visualization of Categorical Data*, chapter Visualizing categorical data with related metric scaling, pages 365–376. J. Blasius and M. Greenacre, eds. Academic Press, 1997.

[CO87] C.M. Cuadras and J.M. Oller. Eigenanalysis and metric multidimensional scaling on hierarchical structures. *Qüestiió*, 11:37–58, 1987.

[Cua92] C.M. Cuadras. Probability distributions with given multivariate marginals and given dependence structure. *J. of Multivariate Analysis*, 42:51–66, 1992.

[Joh67] S.C. Johnson. Hierarchical clustering schemes. *Psychometrika*, 32:241–254, 1967.