Analysis of Distance Matrices for Studying Data Structures and Separating, Classes

ARTICLE *in* QUANTITATIVE STRUCTURE-ACTIVITY RELATIONSHIPS · JANUARY 1993

DOI: 10.1002/qsar.19930120408

2 AUTHORS, INCLUDING:

Alessandro Giuliani
Istituto Superiore di Sanità
362 PUBLICATIONS **4,462** CITATIONS

Analysis of Distance Matrices for Studying Data Structures and Separating Classes

R. Benigni*

Istituto Superiore di Sanitá-Laboratory of Comparative Toxicology and Ecotoxicology – Via Regina Elena 00161 Rome, Italy

A. Giuliani

Institute for Research on Senescence – Pomezia, Italy

Abstract

This paper demonstrates that the computation of Euclidian distances between objects generates an intrinsic geometry, which, besides containing all the information explicitly present in the original space, also provides new information about the object interrelationships. In other words, by computing distances the data are mapped onto a new, and more flexible, reference frame. Such properties of the distance function also permit the derivation of non linear descriptors of the data space, and can be usefully exploited for pattern recognition purposes. In particular, it is demonstrated that this approach is able to separate embedded classes, which have been repeatedly reported in QSAR research.

Key words: Class separation, Euclidian distance, Principal component analysis, Distance matrix, Pattern recognition

Introduction

Pattern recognition problems, i.e. the study of data in which the objects are divided into groups characterized by different patterns of variable values, are encountered in many different scientific areas. In some cases, these patterns are linearly separable – and so are easily analyzable with well-founded analytical tools, such as linear discriminant analysis –; but, in other cases, discrimination is a more demanding task. In the study of structure-activity and activity-activity relationships, the occurrence of non-linearly separable classes has been reported by several authors, and different *ad hoc* approaches have been devised to overcome this problem [1, 2, 3].

In this paper we present a new approach to the problem of separating two classes of points. This approach originated from our observation that an "asymmetric" data structure can often be solved by calculating the Euclidian distances between the objects, and then incorporating one or more of these distance variables into a linear discriminant equation. An "asymmetric" structure is a cluster of objects that belong to one class, embedded in a cloud of other objects belonging to another class [1]. The rationalization of our result is intuitive in the case of an embedded cluster with spherical symmetry: the distance from the central point of the cluster will be sufficient to discriminate between the two classes. When the embedded cluster has a more irregular shape, the discrimination can be attained by using a combination of distances from a few objects (our unpublished results). Based on these observations, we have analyzed the information provided by the calculation of distances among objects, and studied how this information can be used for practical purposes, such as the discrimination between two non-linearly separable groups.

Results and Discussion

The information provided by calculating distances between objects will be demonstrated with a series of simulated examples. First, let us consider a 2-dimensional uniform distribution of points (N = 100), which have been generated by N random extractions of two variables X and Y, with values ranging 0 to 10 (Fig. 1'). Since X and Y are uncorrelated, the Principal Component Analysis (PCA) of the data gives two PCs, each explaining about 50% of the variance: the two PCs correspond to a rotation of the original variables. On the other hand, the computation of the Euclidian distances among objects produces a symmetrical N × N matrix, which can be considered as a multivariate data matrix with N objects defined by N variables. The N distance variables can be analyzed by PCA. Accepting eigenvalues > 1, the PCA produces 4 PCs, explaining 98% of variance. Table 1 shows that PC1 and PC2 are correlated with X and Y, thus being homologous to the two PCs obtained from the original X and Y variables. PC3 and PC4 (explaining together 20% of variance) represent new information, since they are uncorrelated with the original variables.

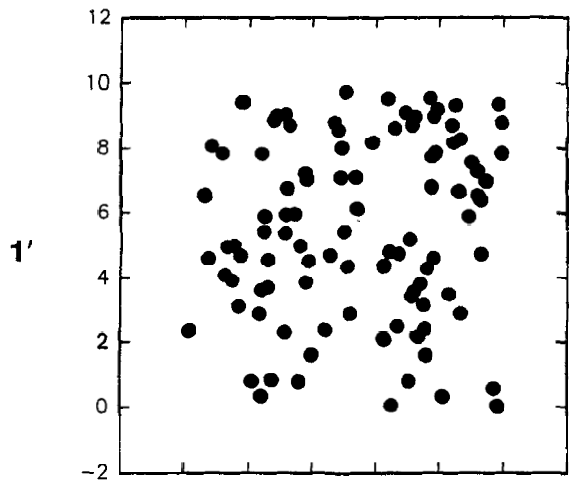
To give an idea of the geometrical properties of the data description provided by these four PCs, we have projected the scores of each of them on the original variables X and Y (Fig. 1''). PC1 and PC2 clearly appear as rotations of X and Y,

Table 1. Uniform distribution: Correlations among original variables and PCs obtained from the distance matrix.

	PC1	PC2	PC3	PC4	X	Y
PC1	1.000					
PC2	0.000	1.000				
PC3	0.000	0.000	1.000			
PC4	0.000	0.000	0.000	1.000		
X	−0.823	0.547	−0.044	0.000	1.000	
Y	−0.647	−0.751	−0.023	0.012	0.118	1.000

* to receive all correspondence

0931-8771/93/0412-0397$5.00+.25/0



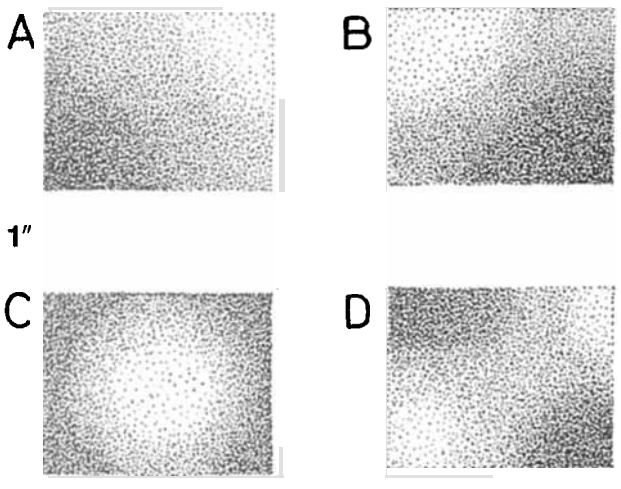
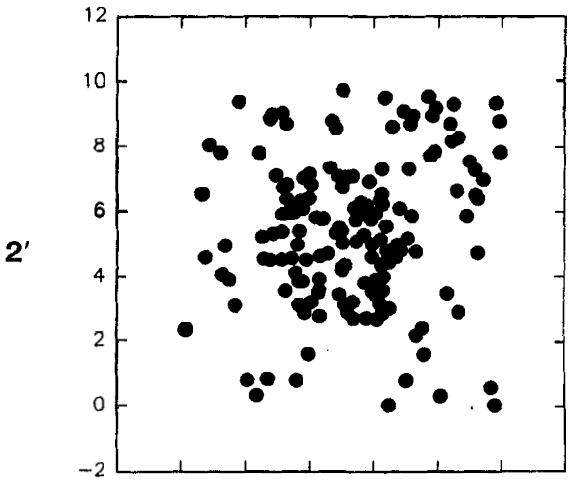


Figure 1. 1': Uniform distribution of random points. 1": Figs. 1"A, 1"B, 1"C, and 1"D display the PC1, PC2, PC3, and PC4 scores, which are plotted against the original x,y coordinates of the points. PC1, PC2, PC3, and PC4 are the Principal Components of the matrix of Euclidian distances between the points of Fig. 1'. Heavily shaded areas represent high scores, and *vice versa*. Proportions of explained variability: PC1: 0.44; PC2: 0.33; PC3: 0.19; PC4: 0.02.



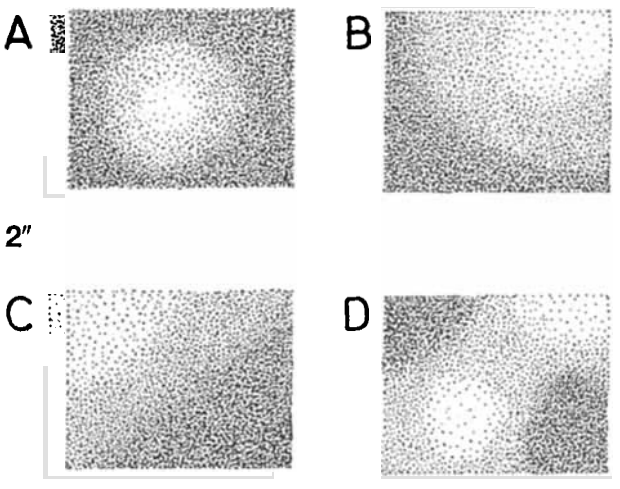


Figure 2. 2': Uniform distribution modified by adding further points to the central area. 2": The scores of the PCs, obtained by PCA of the distance matrix of Fig. 2' points, are displayed in Figs. 2"A, 2"B, 2"C, and 2"D. Proportions of explained variability: PC1: 0.43; PC2: 0.30; PC3: 0.24; PC4: 0.01.

except for a curvature due to the intrinsic quadratic character of the Euclidian distance function. On the contrary, PC3 provides a data description in terms of central symmetry (it is actually highly correlated with the distance of the points from the center of the data field, with $r = 0.927$). PC4 describes the data in terms of symmetry from the corners of the data field. This new information can be interpreted as information on the relationships among objects.

Since the matrices with N objects and N variables are degenerate, the robustness of the above results was controlled by repeating the PCA on subsets of distance variables. The original $N \times 2$ data matrix was analyzed with the Km clustering algorithm, and 5-, 10-, and 20-clusters repartitions of the objects were constructed. Then three distance matrices, in which the N objects were defined in terms of distances from the 5, 10, and 20 centers of the clusters respectively, were calculated. In these matrices (100×5, 100×10, 100×20), the number of the objects was much higher than that of the variables, thus ruling out the problem of matrix degeneration. The PCA of these 3 distance matrices produced PC patterns substantially identical to that obtained with the $N \times N$ distance matrix (results not shown).

The PC pattern derived from the analysis of distances proved to be considerably stable in relation to small perturbations of the data structure. We eliminated 20 points from the uniform distribution of Fig. 1', and created empty strips with different shapes (rectilinear, or circular, or irregular), thus separating the data into two areas of different shapes. The distance matrix PC pattern always resulted to be very similar to that of the original uniform distribution (results not shown).

In the context of this paper, an important point is that of the effectiveness of the PCs, which are derived from the distance matrices, in being able to discriminate between different classes of objects. If the Fig. 1′ population is divided into two linearly separable classes, a combination of two PCs is sufficient to separate the classes: obviously, there is no advantage over using the original X and Y variables. On the contrary, the advantage is clear-cut when non linearly separable classes are dealt with. Let us consider a series of different asymmetric cases, where the Class 1 objects have $a < x < b$, and $c < Y < d$ (with different combinations of a, b, c, d), and are interspersed among the Class 2 objects. In each of these cases, a linear discriminant analysis applied to X and Y variables is unable to separate the classes, whereas linear combinations of the four PCs obtained from the distance matrix discriminate between the two classes with fairly good accuracies (Table 2). This result can be explained in geometrical terms, by recalling that combinations of straight and conical lines (which is what the 4 PCs actually are) permit the approximation of a variety of shapes. Thus, the information gained with PC3 and PC4 helps solve non linearly separable class structures. Moreover, it should be noted that the linear discriminant analysis can also be applied directly to the distance variables (Table 2): usually, the class separation is better than with the PCs, but the geometrical interpretation is less clear, and there is the possibility of chance correlations since the number of variables is equal to the number of objects.

As pointed out above, small perturbations of the uniform distribution do not substantially affect the distance PC pattern. A remarkable perturbation has been obtained by adding to the uniform distribution a further 60 objects, which were randomly sorted with the constraint of being localized in an area of 2.5 radius around the center of the data field (Fig. 2′). The PCA of X and Y variables again gives two PCs, each explaining about 50% of variance; in fact, the addition of the new points around the center of the data space keeps X and Y uncorrelated ($r = 0.109$). On the contrary, the PCA of the distance matrix of Fig. 2′ data is able to reflect the strong departure of the new data field from the previous uniform distribution. Four PCs with eigenvalue > 1 are obtained (Fig. 2″) their spatial features are similar to those of the PCs derived for the uniform distribution, but their rank of importance is changed. PC1 now has central symmetry, and reflects the heavy density of points in the center of the data space.

Fig. 3′ shows another kind of perturbation of the uniform distribution, with a heavy density of points in an eccentric position in the data field. In this case, the PCA of X and Y variables produces a first component (variance explained = 0.69), which is more important than the second

Table 2. Uniform distribution: Discrimination of an embedded class.

Class 1 boundaries	F-ratios of PCs	Performance (a)	Number of distance variables	Performance (b)
a 3	PC1 3.30*	Cl1 100.	3	Cl1 93.3
b 7	PC2 0.38	Cl2 78.		Cl2 95.3
c 3	PC3 45.76*			
d 7	PC4 0.12	Acc 82.		Acc 95.
a 5	PC1 27.74*	Cl1 81.3	4	Cl1 87.5
b 9	PC2 1.01	Cl2 83.3		Cl2 88.1
c 5	PC3 1.64			
d 9	PC4 5.49*	Acc 83.		Acc 88.
a 6	PC1 22.76*	Cl1 100.	3	Cl1 89.5
b 10	PC2 6.35*	Cl2 72.8		Cl2 88.9
c 4	PC3 6.69*			
d 8	PC4 4.32*	Acc 78.		Acc 89.
a 6	PC1 156.3 *	Cl1 100.	4	Cl1 100.
b 10	PC2 0.70*	Cl2 98.7		Cl2 98.7
c 6	PC3 5.08*			
d 10	PC4 25.39*	Acc 99.		Acc 99.

Discriminating non-linearly separable classes. Classes 1 and 2 are partitions of an uniform distribution. Class 1 is the embedded class, with $a < X < b$, $c < Y < d$.

F-ratio statistics of the PCs are at Step 0 of Stepwise Linear Discriminant Analysis (BMDP7M Program from the BMDP Statistical Software); the PCs are those generated by PCA of the distance matrix; the asterisks indicate the PCs entered into the final discriminant equation.

The Number of Distance Variables refers to the Distance Variables, which were entered into the final discriminant equation after Stepwise Linear Discriminant Analysis of the distance matrix.

Performance:

a) refers to the Discriminating Analysis based on the PCs, and b) to the Analysis based on the distance variables.

Cl1: percentage of Class 1 objects correctly classified; Cl2: percentage of Class 2 objects correctly classified; Acc: percentage of objects of both classes correctly classified

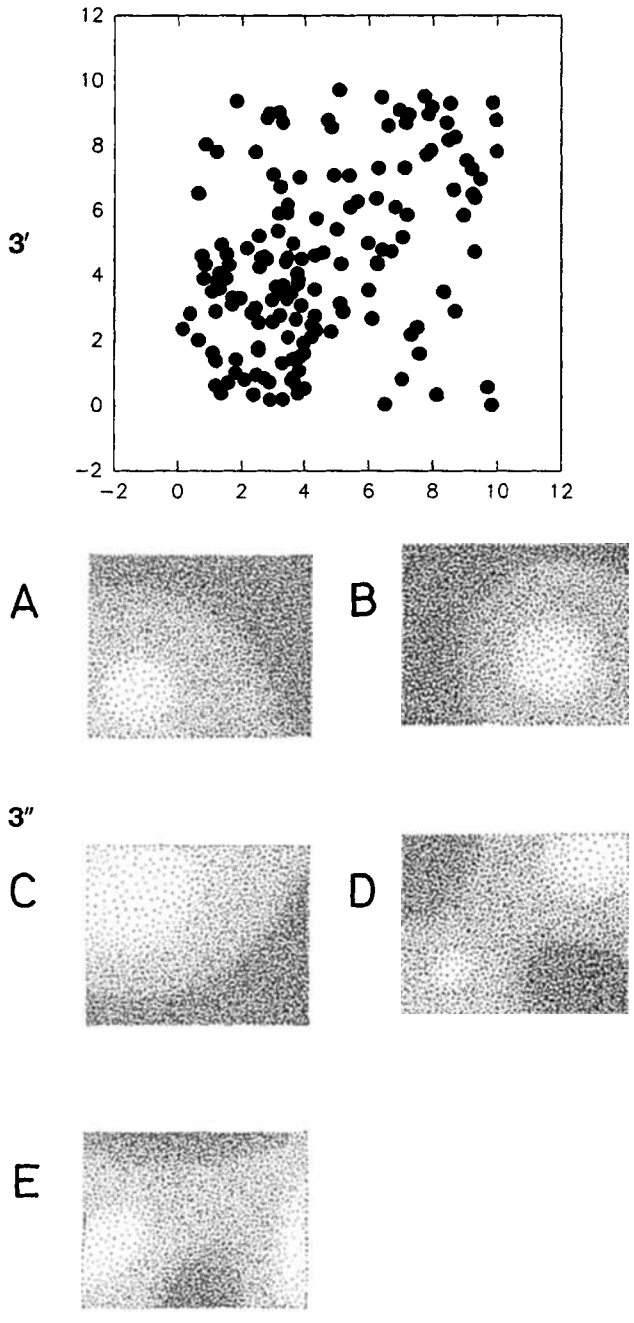


Figure 3. 3′: Uniform distribution modified by adding further points in an eccentric area (low x,y values). 3″: The scores of the PCs, obtained by PCA of the distance matrix of Fig. 3′ points, are displayed in Figs. 3″A, 3″B, 3″C, 3″D, and 3″E. Proportions of explained variability: PC1: 0.60; PC2: 0.21; PC3: 0.16; PC4: 0.01, PC5: 0.01.

component, since a correlation between X and Y has taken shape (r = 0.307). The PCA of the distance matrix generates 5 PCs with eigenvalues > 1 (Fig. 3″). It should be noted that PC1 is again centered on the area with the highest density of points, and the entire PC pattern is remarkably different than the patterns seen before: this again indicates that the PCA of distance matrices accurately reflects the features of the data structures.

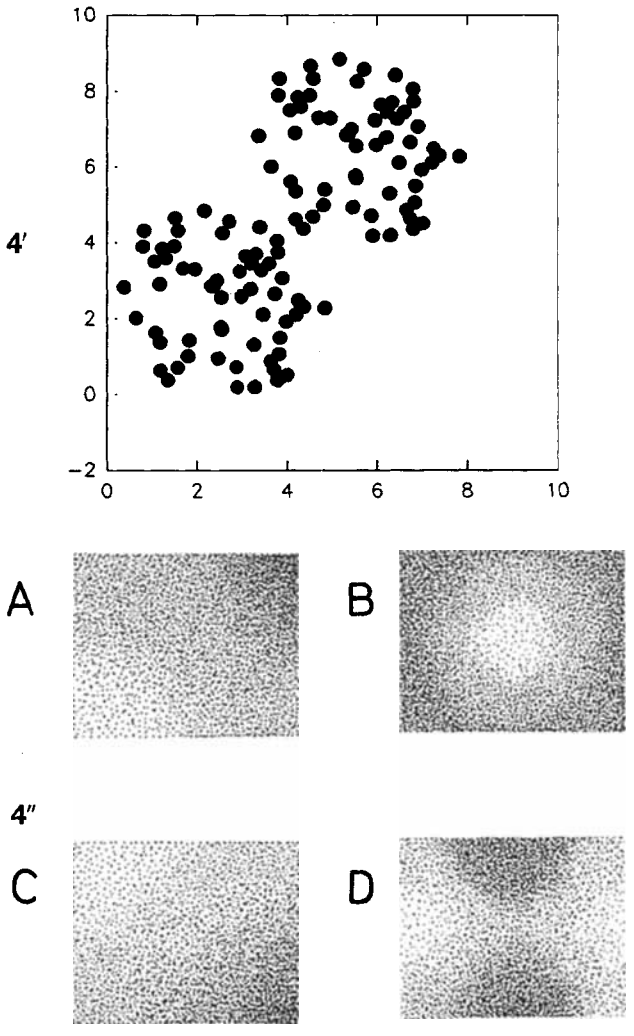


Figure 4. 4′: Linearly separable clusters of points. 4″: The scores of the PCs, obtained by PCA of the distance matrix of Fig. 4′ points, are displayed in Figs. 4″A, 4″B, 4″C, and 4″D. Proportions of explained variability: PC1: 0.71; PC2: 0.18; PC3: 0.08; PC4: 0.02.

In terms of class discrimination, Fig. 2′ and Fig. 3′ cases can be considered at the same time. In the non trivial case in which the area dense with points is not an artifact due to the sampling procedure, but coincides with the embedded class – thus having a physical meaning –, PC1 alone provides a good discrimination. If the embedded class is in any other area, the situation is essentially similar to the uniform distribution, and so adequate linear combinations of PCs can be found to solve the problem (results not shown).

As a final case, let us consider a situation in which two non embedded clusters exist (Fig. 4′). The PCA of X and Y variables produces an overwhelming first component, which goes along the straight line connecting the two clusters (variance explained = 0.71). The PCA of the distance matrix generates 4 PCs (Fig. 4″). Since the class structure is linearly discriminable, the use of the distance matrix PCs (in particular PC1, which alone separates with 100% accuracy) does not

improve the discrimination ability. However, it should be stressed once again that the distance matrix PC pattern is sensitive to the data structure, and it changes according to it.

Conclusions

Distances play an important role in the analysis of data. For example, they are at the basis of Clustering algorithms, of Correspondence Analysis [4], and of Multidimensional Scaling [5]. Moreover, the computation of distance matrices permits the transformation of qualitative data into quantitative data (e.g. through the computation of Hamming distances), on which algorithms requiring at least continuous variables can be applied [6]. In this case, the distance matrix is directly used as a multivariate unit/variable matrix, on which to apply the algorithms adequate to the problem under study [7]. Another important application is the use of distance matrices to overcome the problem of missing data [8]).

In this paper, we have demonstrated that the computation of Euclidian distances between objects generates an intrinsic geometry, which, besides containing all the information explicitly present in the original space, also provides new information, which is based on the symmetry properties of the distance function.

Our simulation on a data field devoid of structure (i.e. 2-dimensional uniform point distribution) has shown how the distance computation has added two new non linear descriptors (PC3 and PC4) to the simple reconstruction of the data field (PC1 and PC2). These four dimensions basically maintain the same shape when new structural features – which take away the data field from the uniform distribution – are inserted (e.g. addition of points in specific areas): the new structures are described by changing the relative importance of the PCs. In other words, by computing distances the data are mapped onto a new, and more flexible, reference frame. Such properties of the distance function also permit the derivation of non linear descriptors of the data space, and can be usefully exploited for pattern recognition purposes – e.g. the solution of the asymmetric case. We are presently implementing our approach as a computer program; it is also possible to find – *via* an optimization technique – an optimal pattern of weights, to be given to the original variables before the computation of the distances. This implementation has been succesfully applied to a number of real cases from the literature (manuscript in preparation).

It should be stressed that the computation of distances and the derivation of PCs is an unsupervised process. The treatment of the data precedes the pattern recognition step: this allows the emerging structure not to be influenced by the specific problem, but to be a detailed description of the data field with new variables. These variables, which consist of a series of linear and non linear views on the data space, will be used for the pattern recognition only in a second phase. Since these are "natural" views not driven by the problem to be solved, they are also important for the exploration of the scientific aspect of the problem, and not only for obtaining the formally best discrimination.

Acknowledgements

Dr. Marta Menghini is gratefully acknowledged for critically reading the manuscript. Mrs. Eve Silvester is acknowledged for the patient editing of the text. Mr. G. Briancesco is acknowledged for his assistance.

References

[1] Dunn, W. J. and Wold, S., *Med. Chem.*, 25, 595–599 (1980).
[2] McFarland, J. W. and Gans, D. J., *J. Med. Chem.*, 30, 46–49 (1987).
[3] Rose, V. S., Wood, J. and McFie, H. J. H., *QSAR*, 10, 359–368 (1991).
[4] Lebart, L., Morineau, A. and Warwick, K. M., *Multivariate descriptive statistical analysis*, Wiley, New York (1984).
[5] Gower, J. C., in E. Lloyd (Ed) *Handbook of Applicable Mathematics*, vol. VI: Statistics, part B, Wiley, New York, 727–781 (1984).
[6] Benigni, R. and Giuliani, A., *Mutat. Res.*, 147, 139–151 (1985).
[7] Kruskal J., in J. Van Ryzin (Ed) *Classification and clustering*, Academic Press, New York, 17–44 (1977).
[8] Sneath, P. H. A., *J. Gen. Microbiol.*, 129, 1045–1073 (1983).