and show completely different properties for cation complexation and polar interactions (Mikes, Milat, Pugin & Blein, 1994; Mikes, Lavernet, Milat, Collange, Pâris & Blein, 1994).

We express our thanks to the Interface Chimie-Biologie du CNRS and Conseil Régional de Bourgogne for financial support, Professor J. C. Tabet and Drs J. Einhorn and C. Descoins for their interest and useful discussions.

References

ARNONE, A., NASINI, G., MERLINI, L., RAGG, E. & ASSANTE, G. (1993). *J. Chem. Soc. Perkin Trans* I, 145–151.
ASSANTE, G., LOCCI, R., CAMARDA, L., MERLINI, L. & NASINI, G. (1977). *Phytochemistry*, **16**, 243–247.
BLEIN, J. P., BOURDIL, I., ROSSIGNOL, M. & SCALLA, R. (1988). *Plant Physiol.* **88**, 429–434.
DUCROT, P. H., BLEIN, J. P., MILAT, M. L. & LALLEMAND, J. Y. (1994). *J. Chem. Soc. Chem. Commun.* pp. 2215–2216.
JALAL, M. A. F., HOSSAIN, M. B., ROBESON, D. J. & VAN HER HELM, D. (1992). *J. Am. Chem. Soc.* **114**, 5967–5971.
JOHNSON, C. K. (1976). *ORTEP*II. Report ONRL-5138. Oak Ridge National Laboratory, Tennessee, USA.
KUYAMA, S. & TAMURA, T. (1957). *J. Am. Chem. Soc.* **79**, 5726–5729.
LIEBICH, B. W. (1979). *Acta Cryst.* B**35**, 1186–1190.
LOUSBERG, R. J., WEISS, U., SALEMINK, C. A., ARNONE, A., MERLINI, L. & NASINI, G. (1971). *J. Chem. Soc. Chem Commun.* pp. 1463–1464.
MIKES, V., LAVERNET, S., MILAT, M. L., COLLANGE, E., PÂRIS, M. & BLEIN, J. P. (1994). *Biophys. Chem.* **52**, 259–265.
MIKES, V., MILAT, M. L., PUGIN, A. & BLEIN, J. P. (1994). *Biophys. Biochem. Acta*, **1195**, 124–130.
MILAT, M. L. & BLEIN, J. P. (1994). *Plant Physiol. Biochem.* In preparation.
MILAT, M. L., BLEIN, J. P., EINHORN, J., TABET, J. C., DUCROT, P. H. & LALLEMAND, J. Y. (1993). *Tetrahedron Lett.* **34**(9), 1483–1486.
MILAT, M. L., FRAICHARD, A., BLEIN, J. P. & PUGIN, A. (1992). Int. Workshop on Plant Membrane Biology, Monterey, July 19–24.
MILAT, M. L., PRANGÉ, T., DUCROT, P. H., TABET, J. C., EINHORN, J., BLEIN, J. P. & LALLEMAND, J. Y. (1992). *J. Am. Chem Soc.* **114**, 1478–1479.
MOTHERWELL, W. D. S. & CLEGG, W. (1978). *PLUTO. Program for Plotting Molecular and Crystal Structures*. Univ. of Cambridge, England.
NEUMAN, A., BECQUART, J., GILLIER, H., LEROUX, Y., QUEVAL, P. & MORETTI, J. L. (1989). *Acta Cryst.* C**45**, 1966–1970.
SCHLÖSSER, E. (1962). *Phytopathol. Z.* pp. 295–312.
SCHLÖSSER, E. (1971). *Phytopathol. Mediterr.* **10**, 154–158.
SHELDRICK, G. M. (1985). *SHELXS*86. *Program for the Solution of Crystal Structures*. Univ. of Göttingen, Germany.
TABATA, N., TOMODA, H., MATSUZAKI, K. & OMURA, S. (1993). *J. Am. Chem. Soc.* **115**, 8558–8564.

Acta Cryst. (1995). B**51**, 314–328

Use of the Estimated Errors of the Data in Structure-Correlation Studies

BY OLIVIERO CARUGO

Dipartimento di Chimica Generale, Università di Pavia, Italy

(*Received* 9 *June* 1994; *accepted* 3 *October* 1994)

Abstract

Novel statistical and numerical methods of data analysis, which make extensive use of the estimated errors (e.s.d.'s) of the data are presented and applied to structure-correlation problems. The novel procedures concern both univariate (histogram representation, HR) and multivariate (cluster analysis, CA, and principal-component analysis, PCA) statistical techniques. In the case of HR, the problem of optimally selecting the dimensions of the spaces is bypassed by convoluting a series of normal functions. In the case of CA, a probability significance is given to the similarity between two (or more than two) objects. In the case of PCA, a cross-validation technique, which takes into account the e.s.d.'s of the row data, allows the determination of the dimensionality of the principal-component space, easy detection of outliers with respect to any principal component, and evaluation of a more comprehensive percentage of the variance described by the principal components.

1. Introduction

An impressive growth of interest in structure-correlation studies appeared in the last two decades (Bürgi & Dunitz, 1994; Auf der Heyde, 1994; Orpen, 1993; Bürgi, 1992; Ferretti, Dubler-Steudle & Bürgi, 1992; Domenicano, 1992). This can be considered a consequence of two main factors: on one hand, structural determinations became very fast, and consequently the number of new structures increased; on the other hand, structural data have been organized in computer-readable databases (Allen, Bergerhoff & Sievers, 1987).

Many papers have been published on data-treatment strategies (Taylor & Allen, 1994) and, in particular, certain interest has been focused on the importance and use of estimated standard deviations (e.s.d.'s) of the structural data [bond distances and angles, torsions, *etc.* (Taylor & Kennard, 1983, 1985, 1986; Mackenzie, 1974; Hamilton & Abrahams, 1970, 1972; Abrahams & Keve, 1971; Abrahams, Hamilton & Mathieson, 1970)]. The importance of the e.s.d.'s is, in fact, crucial for

many aspects of structural analysis such as, for example, comparisons of different data sets, mean-value estimations, *etc.* Moreover, more extensive use of e.s.d.'s of the data in structure-correlation studies is advisable, since positional e.s.d.'s are included more and more in databases such as the Cambridge Structural Database (1991), although they have been disregarded in the past.

In the present communication, the use of e.s.d.'s in some statistical techniques is described: histogram representation of the data (HR), cluster analysis (CA) and principal-component analysis (PCA) (Taylor & Allen, 1994; Comincioli, 1992; Malinowski, 1991; Auf der Heyde, 1990; Robert, 1989; Everitt, 1980). The main tool of the present work is the design of mathematical procedures which are easily applicable to structural problems. Each topic, HR, CA and PCA, will be discussed separately. Particular attention will be devoted to PCA, a more and more widely used statistical technique. All the calculations have been performed with locally written programs (*Fortran*77) on a MicroVAX 3100 computer.

2. Histogram representations

2.1. *Discrete histograms*

Histograms are often used to describe the distribution of the values of a given variable (Taylor & Allen, 1994). One of the most important advantages of the histogram representation of the distribution is the ability to visualize its features (unimodal, bimodal, symmetric *etc.*). When a histogram is built, there are two fundamental problems. The first is how to select a 'grid division'. Fig. 1 shows how the descriptive ability of a histogram depends on the dimension of the spaces. A very small grid division [case (*a*), 0.002 Å] implies a serrate profile; a very large grid division [case (*f*), 0.04 Å] shows a unimodal trend; a grid division similar to the mean value of the e.s.d.'s, which in the examined case is 0.009 Å [case (*d*), 0.010 Å], shows a bimodal trend. The second problem is the definition of space. A space ranging between two values $d1$ and $d2$ can be defined as $[d1,d2]$; $]d1,d2[$; $[d1,d2[$ or $]d1,d2]$. An observation of value $d1$ would be classified in first and third space, but not in the other two.

2.2. *Integrated histograms*

In the case of structural parameters (distances, angles *etc.*), the two problems mentioned above can be solved rigorously, exploiting the fact that an e.s.d. is associated with each observation.

An observed bond distance D of 1.50(1) Å can be described by a continuous function, such as, for example, the normal distribution

$$P(x) = [1/\sigma(2\pi)^{1/2}] \exp[-(x-d)^2/2\sigma^2], \qquad (1)$$

where d = 1.50 Å and σ = 0.01 Å. In every space ranging between $d1$ and $d2$, the probability of finding the observation D will be

$$\int_{d2}^{d1} P(x)\,dx. \qquad (2)$$

Therefore, the observation D will be classified in every space $d1$–$d2$ in a measure proportional to (2). In other words, the occupancies in all the spaces of the histogram are considered. Fig. 2 shows another example. Note that this procedure recalls the theory of the fuzzy clusters (Comincioli, 1992; Miyamoto, 1990). There are essentially two advantages of the proposed method: (i) there is no longer the problem of selecting a grid division, but the spaces can be chosen as small as one wants, without the drawback of diminishing too greatly their mean population (this would produce very serrate discrete histograms; Figs. 3 and 4 show some examples); (ii) there is no longer the problem of defining space (open, closed *etc.*).

2.3. *Comparison between discrete and integrated histograms*

It is interesting to compare the discrete histograms with the integrated ones obtained with (1) and (2). Both discrete and integrated histograms have been built, with various grid divisions, for the examples reported in Fig. 3. In the case of the quinones, the grid division ranged between 0.0005 and 0.050 Å, in the case of the 1,2-diaminobenzenes between 0.001 and 0.050 Å, and in the case of the dihydrazinophthalazines between 0.1 and 4.0°. For each couple of histograms (discrete and integrated) Pearson's correlation coefficient (R) of the occupancies and their interdependence (a) have been calculated as

$$a = o_{i(\text{integrated})}/o_{i(\text{discrete})}, \qquad (3)$$

$$\begin{aligned} R = {} & (\Sigma\{[o_{i(\text{discrete})} - \langle o_{i(\text{discrete})}\rangle] \\ & \times [o_{i(\text{integrated})} - \langle o_{i(\text{integrated})}\rangle]\}) \\ & \div (\{\Sigma[o_{i(\text{discrete})} - \langle o_{i(\text{discrete})}\rangle]^2 \\ & \times \Sigma[o_{i(\text{integrated})} - \langle o_{i(\text{integrated})}\rangle]^2\}^{1/2}), \end{aligned} \qquad (4)$$

where $o_{i(\text{discrete})}$ and $o_{i(\text{integrated})}$ are the populations of the ith space in the discrete and integrated histograms, and $\langle o_{i(\text{discrete})}\rangle$ and $\langle o_{i(\text{integrated})}\rangle$ are their mean values. Plot of a or R *versus* the normalized grid division (see Fig. 5) indicates that discrete and integrated histograms become almost identical only for normalized grid division higher than 1.0. This suggests that discrete histograms can be built with grid divisions not smaller than the mean e.s.d.'s of the data

3. Cluster analysis

3.1. *Description of the method*

Cluster analysis (CA) is one of the principal techniques of data analysis. The problem one is expecting to solve by CA is the classification of objects in different

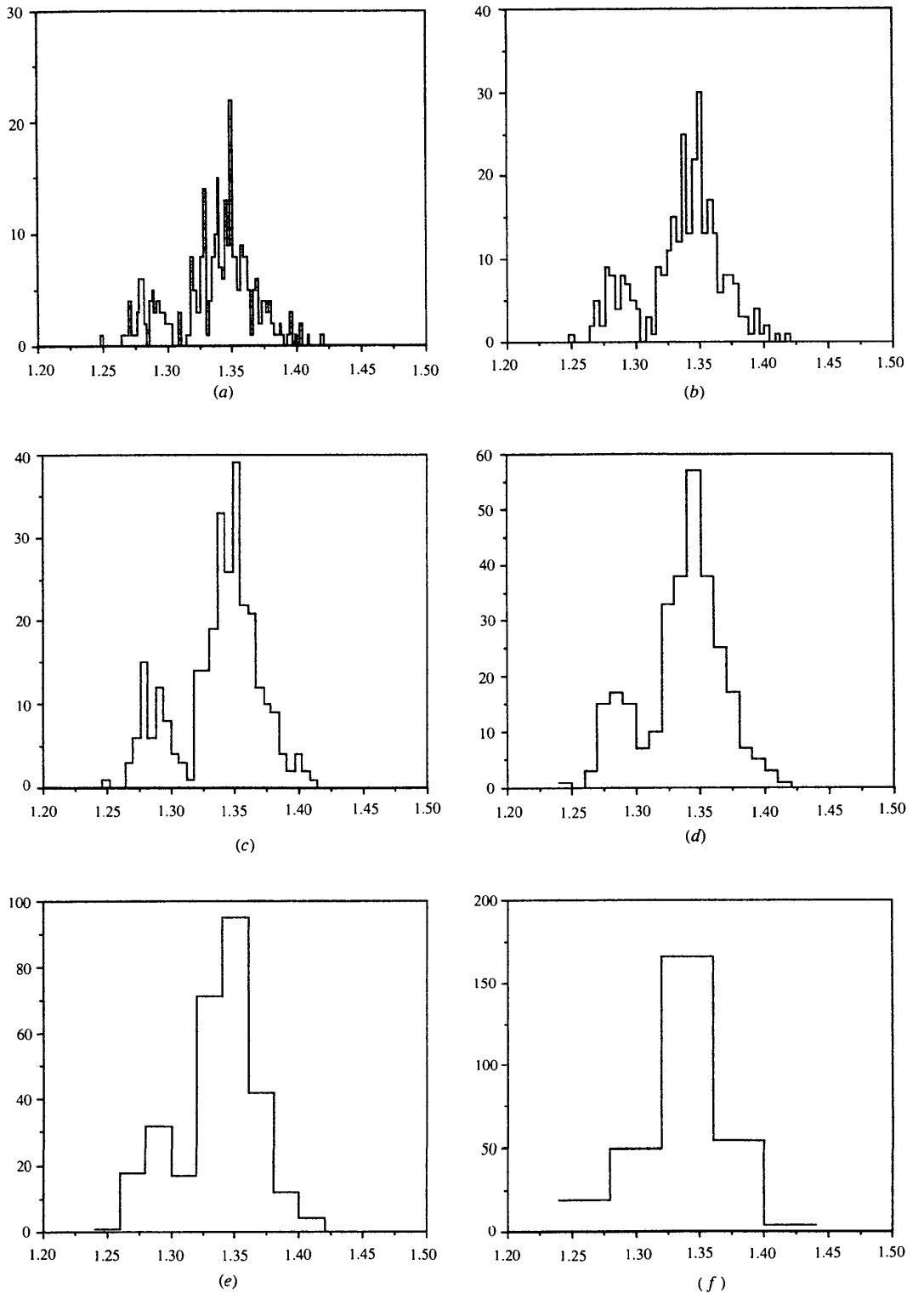


Fig. 1. Distribution of the values of the C—O bond distances in the case of *o*-benzoquinone derivatives (data from Carugo, Bisi Castellani, Djinovic & Rizzi, 1992); grid division (*a*) = 0.002; (*b*) 0.004; (*c*) 0.006; (*d*) 0.010; (*e*) 0.020 and (*f*) 0.040 Å. Increase of grid division produces the loss of some detail, the distribution is changed from bimodal to unimodal. *Vice versa* a decrease of grid division makes the distribution more and more serrated, to such a point that the distribution does not appear clearly bimodal.

clusters on the basis of their similarity (Comincioli, 1992; Everitt, 1980). The utility of CA in structure-correlation studies has been widely investigated (see for example, Taylor & Allen, 1994; Allen, Doyle & Taylor, 1991*a*,*b*; Auf der Heyde, 1990; Norskov-Lauritsen & Bürgi, 1985).

From a formal point of view, given an ensemble $I = \{I1, I2, \ldots, Im\}$ of m objects, each one identified by n variables, it is possible to define an $m \times n$ pattern matrix (m objects described by n variables), which is transformed into an $m \times m$ proximity matrix, **T**, where each element t_{ij} indicates the similarity between the two objects i and j. The proximity matrix **T** is then analyzed, in order to find which objects are similar, and which objects are different. The results of CA can be simply summarized as follows: similar objects will fall within the same cluster, while different objects will fall within different clusters. Two main problems arise: (i) how to define the proximity matrix **T**, that is, the definition of the similarity between the objects i and j, and (ii) how to discriminate between one (or two or more) cluster(s) of objects. Both problems have been extensively studied, but they seem intrinsically ill-defined. Consequently, the CA method of data analysis is inherently a rather ill-defined process (Taylor & Allen, 1994).

The problem of defining a criterion of similarity between various objects is generally solved by the Minkowski metric, as described later (see next paragraph), although some attention has been devoted also to the Mahalanobis criterion of similarity.

The problem of the clustering process can be solved by two main classes of algorithms: partitional and hi-

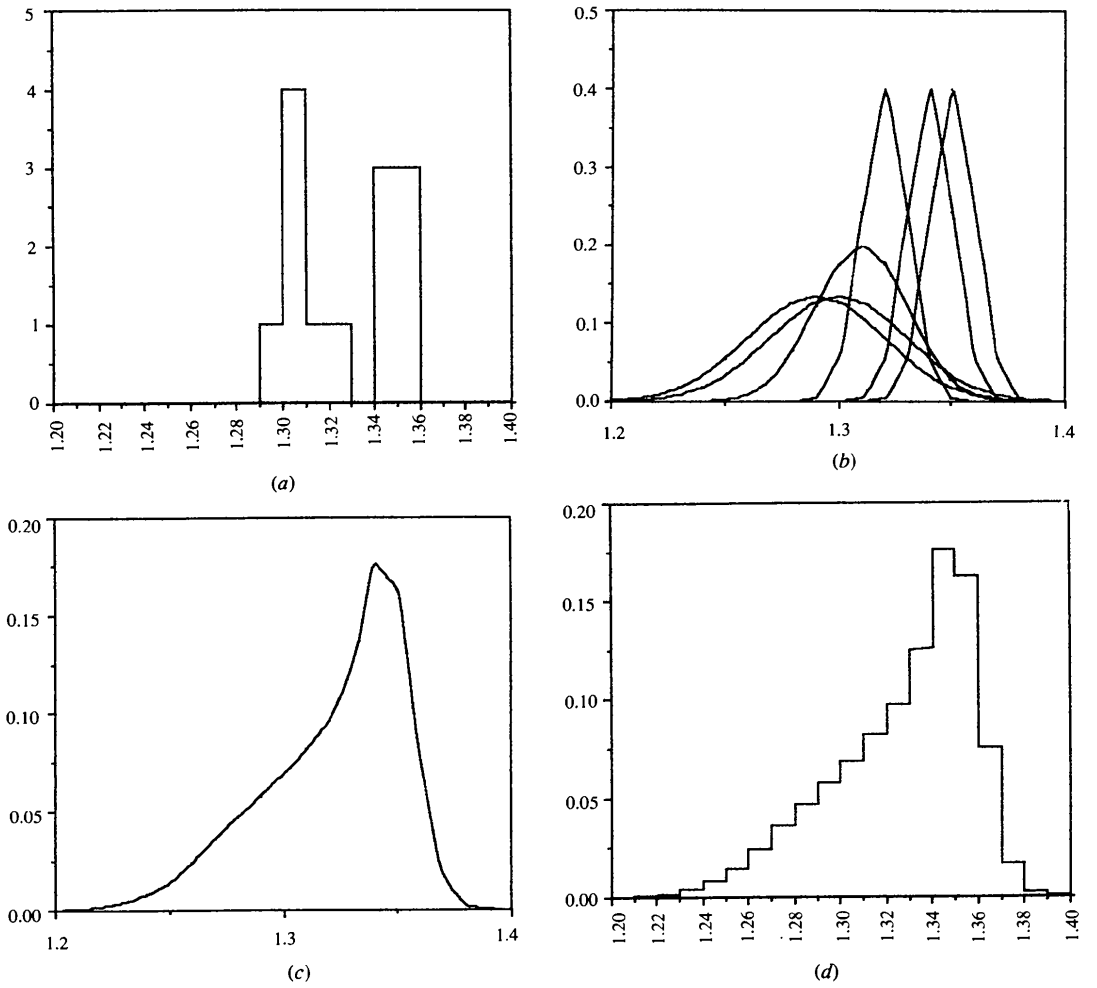


Fig. 2. The following values can be represented by a discrete histogram (*a*), (1) 1.29 (3); (2) 1.30 (3); (3) 1.30 (3); (4) 1.30 (3); (5) 1.30 (3); (6) 1.31 (2); (7) 1.32 (1); (8) 1.34 (1); (9) 1.34 (1); (10) 1.34 (1); (11) 1.35 (1); (12) 1.35 (1); (13) 1.35 (1). The integrated histogram is built as follows: (*b*) normal functions [equation (1)] for cases 1.29 (3) (observation number 1), 1.30 (3) (observations 2, 3, 4 and 5), 1.31 (2) (observation 6), 1.32 (1) (observation 7), 1.34 (1) (observations 8, 9 and 10), 1.35 (1) (observations 11, 12 and 13); (*c*) normalized sum of the curves reported in (*b*); (*d*) integrated histogram corresponding to the curve reported in (*c*). It appears that a bimodal distribution (*a*) becomes nearly unimodal (*d*): this means that the grid division used in (*a*) is too small to show a statistically significant trend.

erarchical. In the first case, the number of clusters to be found is assumed *a priori* and, therefore, partitional clusterings are of little utility for structure-correlation studies, where in general little is known about the structure of the data (if there are reasonable hypotheses on the data distribution or chemically reasonable preconceptions, the CA method becomes useless). Therefore, hierarchical clustering algorithms are preferred; these can be classified into two families: in the first (agglomerative), the m objects are initially distributed in m clusters, each of occupancy one, and larger clusters are then formed, on the basis of the similarity between the objects; in the second, the m objects are initially grouped within one cluster, which is then divided into smaller subsets.

Nevertheless, a hierarchical clustering process cannot be univocally defined. It is generally performed in steps; in each step it is necessary to decide if two clusters are similar; if m objects are analyzed, the maximum number of steps is m. There are various criteria to estimate the similarity of two clusters. If the two clusters have occupancy one, the problem is simply the comparison of two objects. If the two clusters are bigger, the evaluation of their similarity is quite ambiguous. The similarity can be taken as that of their nearest members (single-linkage or nearest-neighbor method), of that of their furthest members (furthest-neighbor method), or that of their centroids (Ward's method). More complex procedures, such as the Jarvis–Patrick method for example, have also been proposed. It is apparent, however, that the clustering procedure is strictly dependent on the criterion selected for comparing two clusters.

Another rather ambiguous point of a hierarchical clustering process is the determination of the optimum number of clusters. In general, the data can be structured in one of the following three ways: (i) all the m objects are similar; (ii) there are two or more distinct groups of objects; (iii) the m objects are quite dissimilar, but do not cluster well. In an agglomerative clustering of m

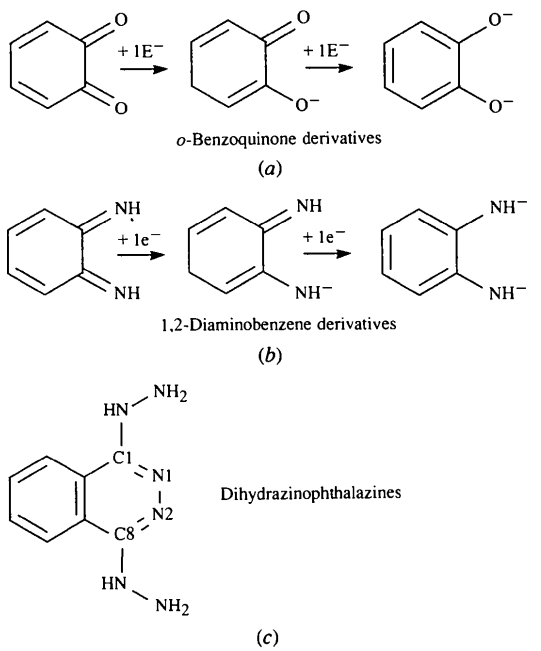


Fig. 3. Summary of the data analyzed. (*a*) 146 crystal structures of *o*-benzoquinone derivatives (data from Carugo, Bisi Castellani, Djinovic & Rizzi, 1992); variable examined: C—O bond distance. (*b*) 28 crystal structures of 1,2-diaminobenzene derivatives (data from Carugo, Djinovic, Rizzi & Bisi Castellani, 1991*b*); variable examined: C—N bond distance; (*c*) six crystal structures dihydrazinophthalazines (data from REFC = HPTNIC, HZPCBX, HPCXNI, DHYZAS10, DIGLIO, DUTCIR10); variable examined: N1—N2—C8 and N2—N1—C1 bond angles.

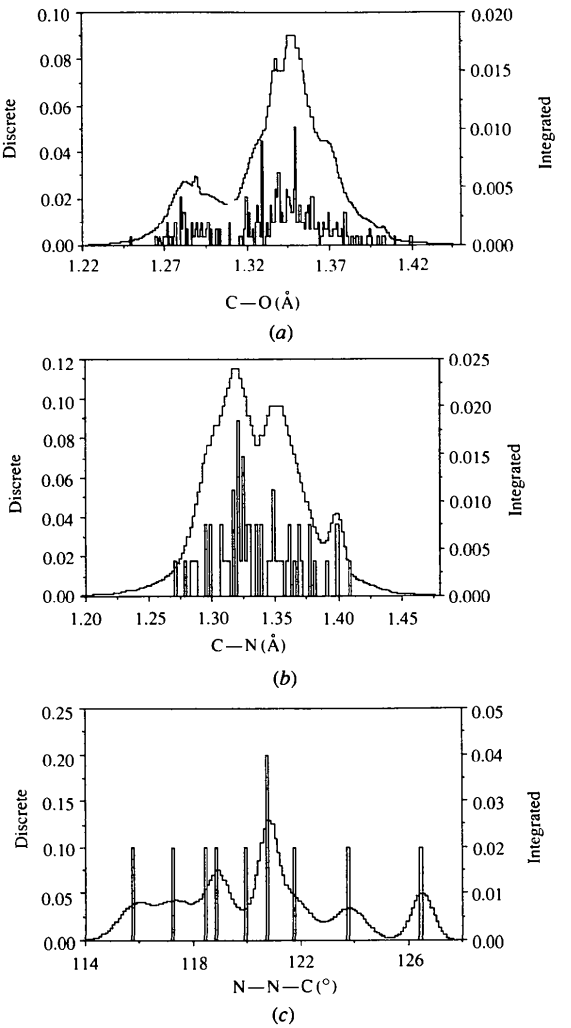


Fig. 4. Discrete and integrated histograms for the variables summarized in Fig. 3: (*a*) C—O bond distance in *o*-benzoquinone derivatives (grid division = 0.001 Å); (*b*) C—N bond distance in 1,2-diaminobenzene derivatives (grid division = 0.002 Å); (*c*) N1—N2—C8 and N2—N1—C1 bond angles of dihydrazinophthalazines (grid division = 0.1°). In all the plots, the grid division is too small for the discrete histograms, which show statistically insignificant trends; on the contrary, the same grid division does not prejudice the significance of the integrated histograms.

objects, the starting point is an m number of clusters, and the ending point is one cluster. Therefore, the agglomerative clustering cannot, by itself, distinguish which of the above three possibilities is the true one. Different procedures have been proposed in order to estimate the best number of partitions (such as, for example, the dependence of the dissimilarity of the clusters on the clustering step, or its derivative), but none of them does insure sound results.

3.2. *Proximity matrix versus probability*

A possible solution to the ambiguities of CA described above can be proposed by giving a statistical probabilistic meaning to the proximity matrix **T**. In this way, it would become possible to estimate the probability that two objects are identical and that a certain number of clusters is the optimum one. Of course, the results of clustering would still depend on the criteria used to estimate the similarity between two clusters. In the present paper, for the sake of simplicity, only the single-link criterion is presented, but the principles underlying the calculation of the proximity matrix keep their validity also if other clustering criteria are adopted.

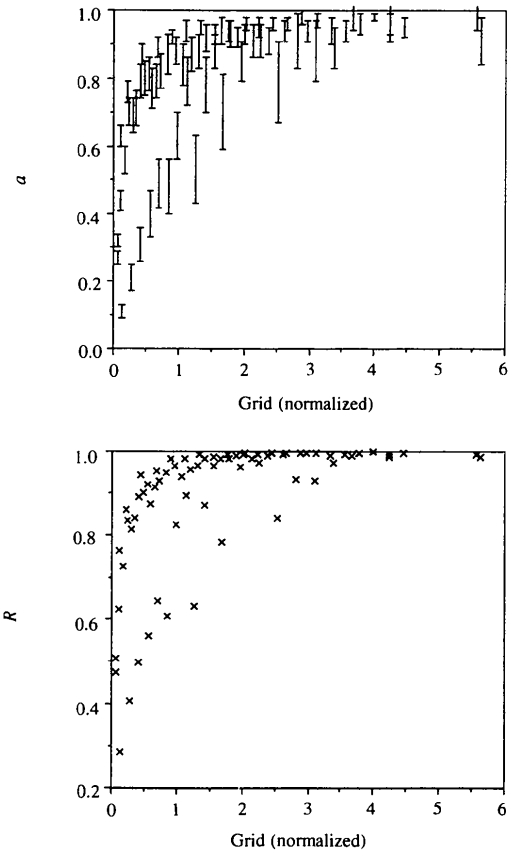


Fig. 5. Comparison between discrete and integrated histograms; dependence of the a factor [equation (3)] and of the Pearson's correlation coefficient [equation (4)] on the normalized grid division, obtained by dividing the actual one by the mean value of the e.s.d.'s; mean e.s.d.'s are 0.009 Å for the C—O bond distance of o-benzoquinone derivatives, 0.017 Å for the C—N bond distance of 1,2-diaminobenzene derivatives and 0.71° for the N1—N2—C8 and N2—N1—C1 bond angles of the dihydrazinophthalazines. It appears that both a and R approach 1.0 (that is, discrete and integrated histograms are identical) if the normalized grid division approaches 1.0 (that is a grid division near the mean e.s.d.) is chosen.

The t_{ij} elements of the proximity matrix **T** are generally termed as similarity indices, and they describe the similarity between the objects i and j. The dimensions and the meaning of the t_{ij} values depend of course on the procedure adopted for calculating them. The elements of the proximity matrix **T** can be defined as

$$t_{ij} = \Sigma[(x_{ik} - x_{jk})/(\sigma_{ik}^2 + \sigma_{jk}^2)^{1/2}]/n, \quad (5)$$

according to the fundamental paper by Cruickshank & Robertson (1953), where x_{ik} and x_{jk} are the values of the variable k in the case of the two objects i and j, and σ_{ik} and σ_{jk} are their e.s.d.'s. The meaning of (5) can be assumed as follows: if t_{ij} is lower than 1.960, the two objects i and j are identical, if t_{ij} is bigger than 2.576 they are significantly different, and if $1.960 < t_{ij} < 2.576$ their difference is possibly significant. The advantage of (5) is, therefore, the possibility to translate the proximity index in terms of probability. It is important to remember that 1.960 and 2.576 are just reference values generally accepted: they correspond to 95 and 99% significance, respectively. Although generally accepted, these values are completely arbitrary and different ones could be selected if desired.

Alternatively, since a simple and general way to evaluate the statistical significance of the t_{ij} values is to divide them by their estimated errors $\sigma(t_{ij})$, the elements of **T** can be redefined as

$$t_{ij}/\sigma(t_{ij}). \quad (6)$$

The t_{ij} values are often calculated on the basis of the Minkowski metric

$$t_{ij} = (\Sigma|x_{ik} - x_{jk}|^r)^{1/r}, \quad (7)$$

where r is a real number ≥ 1.0. According to the rules of error propagation (Taylor, 1982), the error, $\sigma(t_{ij})$, on t_{ij} is defined as

$$\sigma(t_{ij}) = (\Sigma S_k)^{[(1-r^2)/r]} \Sigma[S_k^{r-1} (\sigma_{ik}^2 + \sigma_{jk}^2)^{1/2}], \quad (8)$$

where S_k is defined as the absolute value of the difference between x_{ik} and x_{jk}. The interpretation of (6) in probability terms is analogous to that of the t-test of (5).

3.3. *An example*

24 independent structures of o-benzoquinone monooximes where analyzed (data from Carugo, Djinovic, Rizzi & Bisi Castellani, 1991*a*; Djinovic, Carugo & Bisi Castellani, 1992). Nine bond lengths were considered. A 24 × 9 pattern matrix was obtained,

and a 24 × 24 proximity matrix was treated. An agglomerative hierarchical algorithm was applied, together with the single-link criterion of similarity between two clusters. The clustering was not performed step by step (as described above: that is, by finding the lowest similarity index t_{ij} at each step), but by increasing the threshold values (that is, by considering the ith and jth objects within the same cluster if t_{ij} < threshold). The adopted procedure is presented in the *Appendix* I. Its advantage consists of the fact that two successive cluster fusions are discriminated by a difference in the threshold value, which can be interpreted in probability terms, and not just by a step, which is a simple serial number. The published values of the e.s.d.'s were multiplied by 1.5, according to Taylor & Kennard (1985).

By analyzing the data with the Minkowski metric, and by ranging the exponent r between 1 and 20, it appears (see Fig. 6) that different clustering paths are followed, depending on the definition of the matrix **T**,

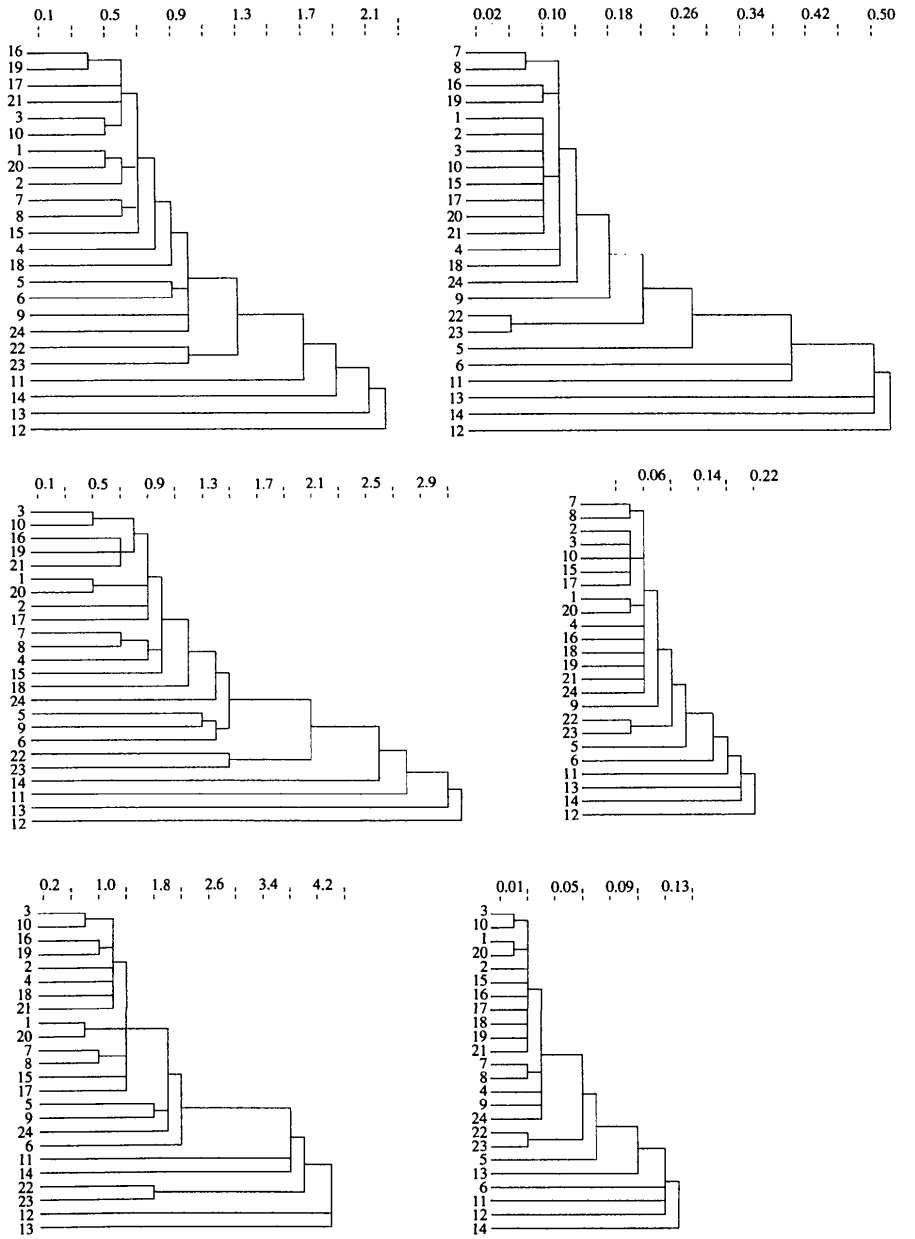


Fig. 6. Graph description of the clustering paths as a function of the similarity index. Left: proximity matrix built with equations (6)–(8). Right: proximity matrix built with equation (7). From the top: Minkowski metric with exponent r = 1, 2 and 9. The clustering path clearly changes if the errors on the row data are considered.

i.e. by calculating its t_{ij} elements with (7) alone, or with (6)–(8). Moreover, the t_{ij} values calculated with (6)–(8) are generally smaller than those calculated with (7) alone (this comparison was performed by normalizing the t_{ij} values between 0.0 and 1.0; see Fig. 7 and Table 1).

The probability ranking is not completely independent of the exponent r of the Minkowski metric. However, some common features, grossly independent of r, do appear. Fig. 6 shows, for example, that at a threshold level of 1.960 (95% probability level) the clusters in Table 2 are formed. Nevertheless, by increasing r, it is apparent that the probability that all the objects fall into the same cluster decreases.

By analyzing the data with the t-test of (5), the clustering path of Fig. 8 is observed. It is closely similar to that obtained by Minkowski metric with exponent $r = 1$. For example, at a threshold level of 1.960, clusters (12) (13) (all the others) are formed. This result suggests that low exponents r should be preferred when the Minkowski metric is used in building the proximity matrix **T**. This is actually done in structure-correlation studies, where exponents 1 ('city-block' metric) or 2 (Euclidean metric) are generally reported.

In conclusion, the main advantage of the procedure proposed above is that the clustering assumes a probability meaning. In the reported example, independently of the algorithms used to build the proximity matrix **T** [(5), or (6)–(8)], there is very little probability (less than 5% at least) that the two quinone monooxime fragments 12 and 13 fall within the same cluster of the other fragments. The same result can be obtained also by CA ignoring the e.s.d.'s of the data, but in that case it would only have

Table 1. *Data obtained from least-squares fit*

exp	r	a	R	Mean absolute error	exp	r	a	R	Mean absolute error
1		0.73 (1)	0.7788	0.0967	9		0.65 (1)	0.7400	0.1105
2		0.64 (1)	0.7414	0.1227	10		0.66 (1)	0.7419	0.1091
3		0.58 (1)	0.7249	0.1358	11		0.66 (1)	0.7433	0.1079
4		0.59 (1)	0.7228	0.1294	12		0.66 (1)	0.7444	0.1070
5		0.61 (1)	0.7262	0.1234	13		0.67 (1)	0.7452	0.1063
6		0.62 (1)	0.7305	0.1187	14		0.67 (1)	0.7458	0.1057
7		0.63 (1)	0.7344	0.1151	15		0.67 (1)	0.7463	0.1052
8		0.64 (1)	0.7376	0.1125	20		0.68 (1)	0.7477	0.1037

Table 2. *Clusters formed at a threshold level of* 1.960 *(95% probability level)*

Exponent r	Clusters
1	(12) (13) (all the others)
2	(12) (13) (11) (14) (all the others)
9	(12) (13) (11) (14) (22,23) (6) (all the others)

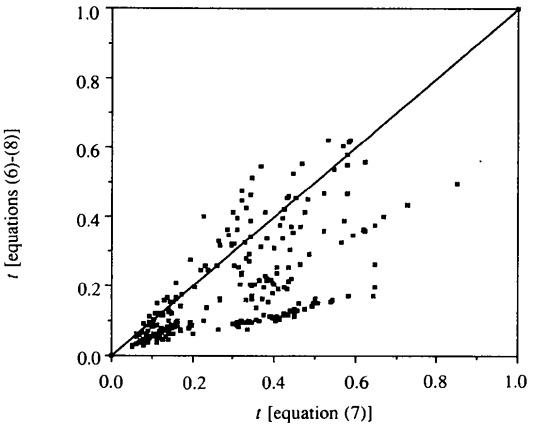


Fig. 7. Dependence of the similarity indices calculated with equations (6)–(8) on those calculated with equation (7). The plot refers to the case with exponent $r = 1$. If the data are fitted by least-squares as $t_{\text{equations (6)–(8)}} = at_{\text{equation (7)}}$, the results in Table 1 are obtained, indicating that the elements of the proximity matrix calculated with equations (6)–(8) tend to be smaller than those calculated with equation (7).

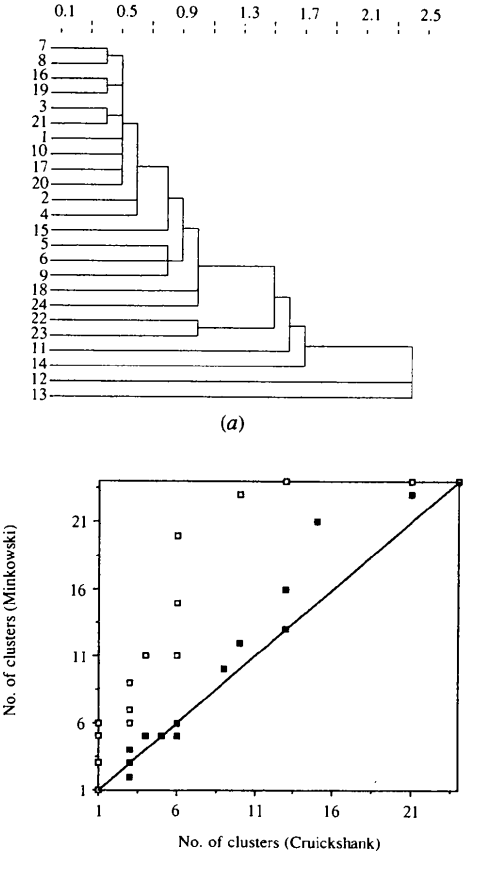


Fig. 8. (*a*) Graph description of the clustering path as a function of the similarity index [calculated with equation (5)]. Note that this graph closely resembles that obtained by the Minkowski metric with exponent = 1 (see Fig. 8, top left). (*b*) Relation between the number of clusters obtained by the Cruickshank t-test [equation (5)] and that obtained by the Minkowski metric [equations (6)–(8)] with exponent = 1 (■) and = 15 (□). Note that the clustering of the Minkowski metric with exponent = 1 closely approaches that of the Cruickshank t-test. If the exponent is set to 15, the clustering is faster.

been possible to conclude that fragments 12 and 13 are the most 'heterogeneous' fragments: it would have been impossible to judge the statistical significance of their 'heterogeneity'.

4. Principal-component analysis

4.1. *Description of the method*

The utility of principal-component analysis (PCA) in structure-correlation analysis has been widely illustrated (see for example, Morton & Orpen, 1992; Auf der Heyde & Bürgi, 1989*a*,*b*,*c*; Murray-Rust & Bland, 1978). PCA can be summarized in a simple way as follows (Taylor & Allen, 1994; Malinowski, 1991; Auf der Heyde, 1990; Robert, 1989): m objects (for example m crystal structures) characterized by n variables (for example n structural parameters like interatomic distances, angles, *etc.*) can be represented by m points in an n-dimensional space, where each axis corresponds to one of the variables; PCA transforms this n-dimensional space into a new one, where each axis describes the greatest part of the variance of the sample (the new axes are determined in the following order: the first represents the direction corresponding to the maximum variance, the second represents the direction of the maximum variance not described by the first, and so on). At the end of a PCA, the problem of determining the number of new axes sufficient to describe all the variance of the original sample arises. Of course, the n new axes can describe the totality of the variance, but it often appears that only a subset of $k < n$ new axes is sufficient to describe the totality of the variance.

The mathematical lexicon of the above procedure is as follows: the n new axes are the eigenvectors and the percentage of the total variance they describe is related to their corresponding eigenvalues. The couple consisting of an eigenvector and an eigenvalue is generally referred to as a component. The $k < n$ components which are sufficient to describe the totality of the overall variance are the primary components, while the remaining ones are the secondary components.

Such terminology derives from the mathematical apparatus performing principal-component analyses. An $m \times n$ data matrix, **D** (m objects characterized by n variables), is standardized by means of (9)–(11), in order to force all the variables to be distributed with zero mean and unit variance

$$dn_{ij} = (d_{ij} - \langle d_j \rangle)w_j, \tag{9}$$

$$w_j = [\Sigma(\langle d_j \rangle - d_{ij})^2/(m-1)]^{-1/2}, \tag{10}$$

$$\langle d_j \rangle = \Sigma d_{ij}/m, \tag{11}$$

where d_{ij} are the elements of **D** and dn_{ij} are the elements of the standardized data matrix **DN**. The $n \times n$ covariance matrix of **DN**, **Z**, which is the correlation matrix of **D**, is then computed with (12)

$$\mathbf{Z} = (\mathbf{DN}^{\mathrm{T}}\mathbf{DN})/(n-1), \tag{12}$$

and the eigenanalysis of **Z** [(13)],

$$\mathbf{Ze}_i = \lambda_i \mathbf{Ie}_i, \tag{13}$$

where **I** is the $n \times n$ identity matrix, produces n unique pairs of vectors $\mathbf{e}_i$ (eigenvectors) and scalars λ_i (eigenvalues). The $m \times n$ scores matrix, **S**, which contains the coordinates of the m objects within the n-dimensional space spanned by the n eigenvectors, is then obtained by (14) and (15),

$$\mathbf{S} = \mathbf{DN}\,\mathbf{F}, \tag{14}$$

$$\mathbf{F} = \mathbf{C}\boldsymbol{\Lambda}^{1/2}, \tag{15}$$

where **F** is the $n \times n$ factor matrix, **C** is the $n \times n$ eigenvectors matrix (where each column is one of the eigenvectors, $\mathbf{e}_i$), and $\boldsymbol{\Lambda}$ is the $n \times n$ diagonal eigenvalues matrix (where each element within the diagonal is one of the eigenvalues λ_i, while the other elements are 0.0). The cumulative variance percentage, V_k, described by the first k principal components is calculated with (16).

$$V_k = 100(\sum_{i=1}^{k} \lambda_i)/n. \tag{16}$$

4.2. *Primary and secondary components*

There are many procedures to discriminate between primary and secondary components (Malinowski, 1991). They can be grouped into three types: those disregarding the estimated errors on the data, those considering a mean error, and those taking into account the variability of the errors from one data point to the next.

Among the procedures of the first type, there are many alternatives. Attention is limited here to the eigenvalue-one criterion, originally proposed by Kaiser (1960), and adopted for example by Domenicano, Murray-Rust & Vaciago (1983) to PCA of structural data. It is based on accepting all the components with eigenvalues above unity.

Among the procedures considering only the mean error, crystallographers often adopt the method proposed by Murray-Rust & Bland (1978), defined by (17)

$$(\sum_{i=1}^{k} \lambda_i)/n - 1 + (\langle \text{e.s.d.} \rangle / \langle \text{s.d.} \rangle)^2 = 0, \tag{17}$$

where $\langle$e.s.d.$\rangle$ is the mean value of the errors of the variables, $\langle$s.d.$\rangle$ is the mean value of the standard deviation of the distribution of the variables, and λ_i are the eigenvalues. The number of primary components is k, where k satisfies (17). A major drawback of (17) is

that it often occurs that the n variables of the m objects have very variable estimated errors. Moreover, in cases when among the variables there are both 'hard' structural parameters, like for example interatomic distances, and 'soft' structural parameters, like bond angles, it would be impossible to compare the absolute values of their estimated errors as they are measured in different units. In these cases it can often appear that the number k of primary components is quite ambiguous, because of the high difference of $\langle$e.s.d.$\rangle$ among the variables. Other criteria for deducing the number of primary components which are closely related to that of (17), like for example the residual standard deviation method of the average error method (Malinowski, 1991), suffer from the same drawback of considering only the mean value of the estimated errors.

Eventually, in order to understand the procedures for discriminating primary and secondary components which take into account the variability of the errors from one data point to the next, it is necessary to stress the background of PCA more.

PCA has been developed mainly for social sciences, where the general problem is to decompose a series of data, the data matrix **D** into two subsets: one, the abstract row matrix **R**, indicates the 'true values' of the variables in each compound, and the other, the eigenvectors matrix **C**, indicates the 'true location' of the objects. Formally, the result of this decomposition can be resumed with (18)

$$\mathbf{DN} = \mathbf{R}_k\,\mathbf{C}_k, \tag{18}$$

where **DN** is the $m \times n$ standardized data matrix, $\mathbf{R}_k$ is the $n \times k$ abstract row matrix, $\mathbf{C}_k$ is the $k \times n$ eigenvectors matrix, and k is the number of primary components. The matrix $\mathbf{R}_k$ is obtained with (19)

$$\mathbf{R}_k = \mathbf{DN}\,\mathbf{C}_k^T, \tag{19}$$

and it is possible, therefore, for each value k with $1 < k < n$ to obtain a calculated (and standardized) data matrix, $\mathbf{DNC}_k$, by means of (20)

$$\mathbf{DNC}_k = \mathbf{R}_k\mathbf{C}_k. \tag{20}$$

The matrix $\mathbf{DNC}_k$ can finally be transformed into the calculated data matrix, $\mathbf{DC}_k$, by means of (9)–(11). Therefore, an approach to the discrimination between primary and secondary components can be based on the comparison between the original data matrix **D** and the calculated data matrix $\mathbf{DC}_k$. It is of course easy to make this comparison dependent on the errors of the data.

The most popular procedure for this is the χ^2 criterion, originally proposed by Bartlett (1950). It is defined as follows. Given an $m \times n$ data matrix **D**, χ_k^2 is calculated with (21)

$$\chi_k^2(\text{calculated}) = \Sigma\Sigma[(d_{ij} - dc_{ij})^2/\sigma_{ij}^2], \tag{21}$$

where σ_{ij} is the estimated error of d_{ij}, and dc_{ij} is the element of the calculated data matrix $\mathbf{DC}_k$. The value of χ_k^2(calculated) is then compared with its corresponding expected value, given by (22)

$$\chi_k^2(\text{calculated}) = (m - k)(n - k). \tag{22}$$

The number of primary components is the smallest k for which (23) is satisfied.

$$\chi_k^2(\text{expected})/\chi_k^2(\text{calculated}) > 1. \tag{23}$$

Nevertheless, the χ^2 criterion has never been applied in studies concerning crystallographic data. This is probably justified by its abstractness with regard to structural chemistry, which makes it difficult to interpret its results. For this reason we designed a new procedure of discrimination between primary and secondary components, which makes extensive use of the estimated errors, being easily interpretable in terms of crystallographic information. This new procedure can be summarized as follows.

4.3. *Fitting percentage*

We can consider as primary components those relative to the k eigenvectors needed to calculate an abstract row matrix, $\mathbf{R}_k$, that allows calculation of a calculated data matrix $\mathbf{DC}_k$ which is not statistically different from the original data matrix **D**. The problem of comparing matrices **D** and $\mathbf{DC}_k$ can be solved as reported by Cruickshank & Robertson (1953), by means of the t test [(24)]

$$t = |d_{ij} - dc_{ij}|/\sigma_{ij}, \tag{24}$$

where d_{ij} are the elements of **D**, σ_{ij} are their estimated errors, and dc_{ij} are the elements of $\mathbf{DC}_k$. The meaning of (24) can be assumed as follows: the difference between d_{ij} and d_{cij} is not significant if $t < 1.960$. The percentage of the overall original variance that can be explained by the first k components is then expressed as the percentage of the elements of $\mathbf{DC}_k$ that have $t < 1.960$ (this percentage will be referred to as FP, fitting percentage). As in the case of CA, it should be remembered that the threshold value 1.960 (95% of significance), although widely accepted, is completely arbitrary.

It is also worth noting that: (i) the FP method allows evaluation of which variables and which objects require an expansion of dimensionality of the principal-component space, in a simple and rapid way; (ii) it seems reasonable to substitute the cumulative variance percentage values V_k [(16)] with the FP ones, which offer the possibility of inclusion of the e.s.d.'s in the estimation of the percentage of the overall variance described by the principal components.

An example of application of the FP method is given in the *Appendix* II. Moreover, PCA's have been performed on the real cases summarized in Fig. 9.

4.4. *Comparison between the FP and other methods discriminating primary components*

12 applications of a new procedure for discriminating between primary and secondary components have been described, and it is thus possible to compare this new procedure with those generally adopted, which have been described above. Results of the procedures for discriminating between primary and secondary components are summarized in Table S1 (Supplementary Material).* Table 3 shows the number of primary components, highlighted in each of the 12 examples, by the fitting-percentage method, the criterion defined by (17) (Murray-Rust & Bland, 1978), the χ^2 criterion (Bartlett, 1950), and the eigenvalue-one criterion (Kaiser, 1960).

* Procedures for discriminating between primary and secondary components have been deposited with the IUCr (Reference NA0062). Copies may be obtained through The Managing Editor, International Union of Crystallography, 5 Abbey Square, Chester CH1 2HU, England.

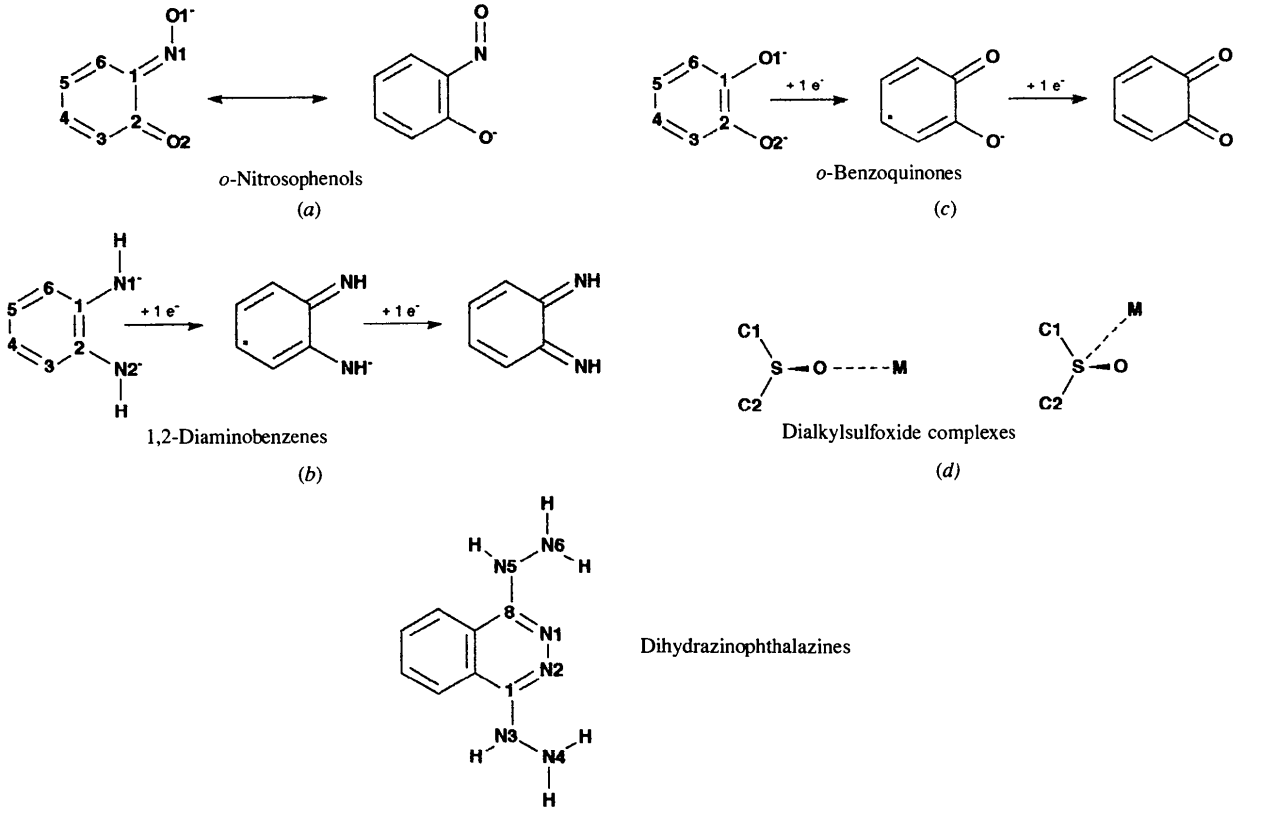


Fig. 9. Summary of the data submitted to PCA with FP procedure. All the published values of the e.s.d.'s were multiplied by 1.5, according to Taylor & Kennard (1985). A structure was considered as outliers: (i) if one of its structural parameters has a value deviating more than $3X$ from the mean value of this parameter, where X is the standard deviation of the distribution of this parameter (OUTX criterion) or: (ii) if one of the estimated errors of one of its structural parameters deviates more than $3Y$ from the mean value of the estimated error of this parameter, where Y is the standard deviation of the distribution of the estimated errors of this variable (OUTY criterion). (*a*) *o*-Nitrosophenols. 24 crystallographically independent structures, characterized by the following variables: N1—O1, C1—N1, C2—O2, C1—C2, C3—C4, C4—C5, C5—C6 and C1—C6. Data from Djinovic *et al.* (1992) and Carugo *et al.* (1991*a*). Six outliers were detected by OUTX and OUTY methods. PCA performed both with and without outliers. (*b*) 1,2-Diaminobenzenes. 28 crystallographically independent structures, characterized by the following variables: C1—N1, C2—N2, C1—C2, C3—C4, C4—C5, C5—C6 and C1—C6. Data from Carugo (1994) and Carugo, Djinovic, Rizzi & Bisi Castellani (1991*b*). No outliers were detected. In order to take into account the intrinsic C_{2v} symmetry of these molecules, the labels N1 and N2, C1 and C2, C3 and C6, and C4 and C5 were permuted (Murray-Rust, 1982). (*c*) *o*-Benzoquinones. 146 crystallographically independent structures, characterized by the following variables: C1—O1, C2—O2, C1—C2, C3—C4, C4—C5, C5—C6 and C1—C6. Data from Carugo, Bisti-Castellani, Djinovic & Rizzi (1992). Nine outliers were detected. In order to take into account the intrinsic C_{2v} symmetry of these molecules, the labels O1 and O2, C1 and C2, C3 and C6, and C4 and C5 were permuted (Murray-Rust, 1982). (*d*) Dialkylsulfoxide complexes. 175 crystallographically independent fragments, characterized by the following variables: S—O, S—C1, S—C2, O—S—C1, O—S—C2 and C1—S—C2. Data from Calligaris, Faleschini & Carugo (1995). 22 outliers were detected. The labels C1 and C2 were permuted in order to take into account the intrinsic C_s symmetry of the fragment. (*e*) Dihydrazinophthalazines. Six crystallographically independent fragments, characterized by the following variables: N1—N2, N1—C1, N2—C8, N3—C1, N3—N4, N5—C8 and N5—N6. Data from REFC = HPTNIC, HZPCBX, HPCXNI, DHYZAS10, DIGLIO and DUTCIE10. No outliers were detected. The labels N1 and N2, C1 and C8, N3 and N5, and N4 and N6 have been permuted in order to account for the C_{2v} symmetry of the molecule (Murray-Rust, 1982). (*f*) Cu–Co-SOD (not shown in the figure). The backbone bond distances (N—Cα, Cα—C, C—O and C—N^+) of the two monomers have been analyzed, for a total of 149 residues (the first and the last have been excluded). Data from Djinovic *et al.* (1992). No outliers were detected. Mean e.s.d.'s of 0.2 Å (Luzzati plot) or of 0.013 Å (Carugo, 1995) for all the variables were considered.

Table 3. *Number of primary components pointed out by the various methods considered in the present article*

Fitting percent (FP), χ^2, equation (17) (eq-176), and eigenvalue-one (λ-one).

	FP	χ^2	eq-17	λ-one
o-Nitrosophenols*	6	5	1	4
o-Nitrosophenols†	9	5	2	3
1,2-Diaminobenzenes	5	1	1	3
o-Quinones*	8	8	2	1
o-Quinones†	8	8	2	2
Dialkylsulfoxides*	6	6	4	2
Dialkylsulfoxides†	6	6	4	2
Dihydrazinophthalazines	4	4	2	3
Cu–Co–SOD monomer 1‡	1	1	1	1
Cu–Co–SOD monomer 1§	2	1	1	1
Cu–Co–SOD monomer 2‡	1	1	1	1
Cu–Co–SOD monomer 2§	3	1	1	1

* With data rejection.
† Without data rejection.
‡ With mean e.s.d. = 0.2 Å.
§ With mean e.s.d. = 0.013 Å.

It appears that the results of the fitting-percentage method and of the χ^2 criterion are quite similar: in seven cases out of 12 these two methods indicate the same number of primary components. It can be observed that they do not fit completely, but it should be remembered that they are integer numbers, and not floating-point real numbers, with the consequence that a perfect linear dependence is unlikely. Moreover, it can be observed that the discrepancies between the results obtained with the FP and the χ^2 methods are overestimated. For example, in the case of the 1,2-diaminobenzenes, the χ^2 method locates only one primary component, while the FP criterion locates five; however, in going from $k = 1$ to $k = 5$, the fitting percentage increases only very little, from 96.9 to 100.0. The same goes for other cases, where the FP method expands the dimensionality of the component space with respect to the χ^2 criterion, only because of a small percentage gain.

On the contrary, the results of the FP and χ^2 methods are considerably different from those obtained by the other two methods [the criterion defined by (17) and the eigenvalue-one criterion], which do not consider extensively the estimated errors on the data. In particular, it appears that the number of primary components indicated by the FP method is never smaller than those indicated by the other methods. Moreover, the method of (17) and the eigenvalue-one criterion give quite different results to each other, which is likely because of the fact that while the first one considers an average value of the estimated errors of the data, the latter completely disregards this kind of information.

Therefore, it is possible to conclude that the extensive use of the information given by the estimated errors of the data significantly affects the determination of the number of primary components. The methods which consider the variability of the errors from one data point to the next can be considered as better, not only because they exploit a larger amount of information, but also because they generally increase the number of primary components and, therefore, decrease the possibility of disregarding significant components of low eigenvalue.

4.5. *Comparison between the FP and V_k values*

Although the FP and the χ^2 methods give closely similar results, the first has considerable advantages other than its simplicity. It gives a quantitative estimation of the data variance percentage that a given number of eigenvectors can describe; in other words, the fitting percentage is more informative than the cumulative variance percentage V_k defined in (16).

Table 4 shows the fitting percentage and the cumulative variance percentage values for all the cases presented above. Two major features appear. On the one hand, the FP values always tend to be higher than V_k. On the other hand, as V_k increases, the difference between FP and V_k decreases. Both trends could indicate that the V_k values tend to underestimate the 'importance' of the first components with respect to the latter ones. However, since the cases reported above cover quite a wide range of structural data, it seems more reasonable to suppose that the errors on the data cause a sort of noise, hiding a part of the overall variance. In other words, the inclusion of the errors on the data results in a lower overall variance. An obvious consequence would be FP values higher than the V_k ones, especially for small V_k.

4.6. *Outlier detection with the FP method*

Another advantage of the FP method over the χ^2 approach is that it allows the easy detection of which data are responsible for the expansion of dimensionality of the component space. This can be seen from the following example, taken from the PCA of the first monomer of the Cu–Co–superoxide dismutase (SOD) (with e.s.d.'s = 0.013 Å). This molecule consists of 151 amino acids; 149 of them (the first and the last excluded) are characterized by the four peptidic bond distances N—Cα, Cα—C, C—O and C—N$^+$. The fitting percentage values of the first two components are 99.7 and 100.0. Therefore, the number of principal components is two. Out of the 596 data, only two cannot be fitted by the first principal component: the N—Cα of the residue Cys6, and the Cα—C of the residue Arg126. It is apparent immediately, and easily, which variables and which object require expansion of the number of primary components from one to two, that is, require an expansion of dimensionality of the space spanned by the principal components. It is beyond the scope of the present article to discuss the chemical or biological relevance of the anomalies of these two data. However, it is important to note that these anomalies are statistically significant. It could also be observed, that the FP method could be useful in checking the correctness of crystallographic refinement.

Table 4. *Comparison between the fitting percentage (FP) and the cumulative variance percentage [V_k see equation (16)] values*

	Component	FP	V_k
o-Nitrosophenols*	1	93.2	39.2
	2	93.8	58.6
	3	96.9	73.2
	4	98.1	84.5
	5	98.8	90.2
	6	100.0	95.3
	7	100.0	97.4
	8	100.0	98.8
	9	100.0	100.0
o-Nitrosophenols†	1	80.6	51.0
	2	88.0	73.3
	3	94.4	88.3
	4	97.7	92.2
	5	99.5	95.8
	6	99.5	97.4
	7	99.5	98.6
	8	99.5	99.5
	9	100.0	100.0
1,2-Diaminobenzenes	1	96.9	37.9
	2	96.4	58.9
	3	98.7	73.0
	4	99.6	80.8
	5	100.0	87.9
	6	100.0	93.1
	7	100.0	96.8
	8	100.0	100.0
o-Quinones*	1	87.9	57.6
	2	89.1	69.9
	3	91.2	78.8
	4	93.6	85.2
	5	97.9	90.7
	6	98.1	95.4
	7	99.5	98.0
	8	100.0	100.0
o-Quinones†	1	88.6	53.2
	2	91.0	66.2
	3	93.2	87.0
	4	96.5	85.7
	5	97.9	91.1
	6	98.3	95.1
	7	99.5	97.8
	8	100.0	100.0
Dialkylsulfoxides*	1	72.5	41.8
	2	79.6	64.7
	3	88.7	80.5
	4	89.0	89.9
	5	94.9	95.6
	6	100.0	100.0
Dialkylsulfoxides†	1	62.5	39.5
	2	81.3	63.7
	3	89.2	78.4
	4	90.9	87.3
	5	95.6	94.7
	6	100.0	100.0
Dihydrazino-phthalazines	1	80.0	35.9
	2	87.1	71.2
	3	97.1	88.3
	4	100.0	93.4
	5	100.0	97.0
	6	100.0	99.4
	7	100.0	100.0
Cu–Co–SOD monomer 1‡	1	100.0	39.5
	2	100.0	62.4
	3	100.0	82.3
	4	100.0	100.0
Cu–Co–SOD monomer 1§	1	99.7	39.5
	2	100.0	62.4
	3	100.0	82.3
	4	100.0	100.0
Cu–Co–SOD monomer 2‡	1	100.0	39.2
	2	100.0	60.5
	3	100.0	81.1
	4	100.0	100.0
Cu–Co–SOD monomer 2§	1	99.2	39.2
	2	99.7	60.5
	3	100.0	81.1
	4	100.0	100.0

* With data rejection.
† Without data rejection.
‡ With mean e.s.d. = 0.2 Å.
§ With mean e.s.d. = 0.013 Å.

5. Concluding remarks

A series of novel statistical techniques of data analysis, suitable for structure-correlation studies, have been presented. They are justified since e.s.d.'s of the data contain an important part of the information given by the data themselves, and since most of the available statistical and numerical methods of data analysis (especially multivariate) were developed for applications in fields other than crystallography (for example, social sciences), in which no e.s.d.'s of the data are known or considered. The main results produced by the novel approaches to histogram representation, cluster analysis and principal-component analysis, can be summarized as follows.

(i) Histogram representation: a novel method has been designed which solves the problem of selection of the correct grid division. It has been verified that the smallest grid division which can be selected in discrete histograms is the mean value of the e.s.d.'s of the data. Smaller grid divisions can be adopted only if integrated histograms are built.

(ii) Cluster analysis: original guidelines for the calculation of the proximity matrix have been reported, in order to give a probabilistic meaning to the similarity indexes.

(iii) Principal-component analysis: the proposed fitting-percentage method allows: (i) determination of the number of primary components, that is, the dimensionality of the principal-component space; (ii) detection of outliers with respect to a given number of principal components; (iii) evaluation of the percentage of the overall variance which is described, including the e.s.d.'s of the data.

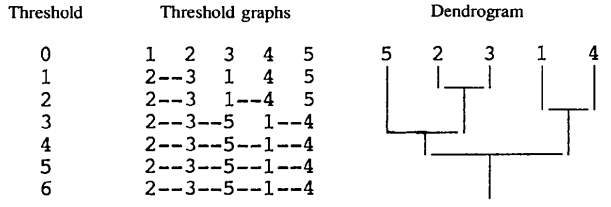
Acknowledgment is given to K. Djinovic, for helpful discussions, and to MURST (Rome) for financial support. Also a referee is acknowledged for her/his contribution in making the text clearer.

APPENDIX I
Example of hierarchical agglomerative clustering

Given a proximity matrix **T**

$$\mathbf{T} = \begin{vmatrix} 0 & 6 & 8 & 2 & 7 \\ 6 & 0 & 1 & 5 & 3 \\ 8 & 1 & 0 & 10 & 9 \\ 2 & 5 & 10 & 0 & 4 \\ 7 & 3 & 9 & 4 & 0 \end{vmatrix},$$

the subsequent agglomeration of the objects can proceed through different threshold values. For example, the first threshold can be 0, the second 1, the third 2, and so on. A series of threshold graphs is then obtained. For example, if a threshold of 1 is considered it appears that objects number 2 and 3 are similar and, therefore, belong to the same subset (and constitute nodes of a threshold graph), while all other objects are in different subsets. If a further threshold of 2 is considered, objects 2 and 3 continue to be in the same subset, a new subset is formed by objects 1 and 4, while object 5 is alone. Three threshold graphs can be traced, one with two nodes, one with only one node. The overall procedure can be summarized by a dendrogram. It is up to the analyst to select a threshold value to stop the agglomerative clustering process. The number of clusters which will be formed depends on the threshold.



(I)

APPENDIX II
Example of the application of the FP method

Given the following data matrix **D**

$$\mathbf{D} = \begin{vmatrix} 1 & 1 & 1 \\ 3 & 3 & 1 \\ 2 & 3 & 1 \\ 3 & 5 & 3 \\ 4 & 3 & 2 \end{vmatrix},$$

it is possible to calculate the eigenvalues matrix and the eigenvectors matrix **C**, by means of (9)–(13)

$$\Lambda = \begin{vmatrix} 2.040 & 0 & 0 \\ 0 & 0.683 & 0 \\ 0 & 0 & 0.277 \end{vmatrix},$$

$$\mathbf{C} = \begin{vmatrix} 0.612 & -0.393 & 0.687 \\ 0.485 & 0.872 & 0.067 \\ 0.625 & -0.292 & -0.724 \end{vmatrix}.$$

By means of (19) and (20) it is then possible to obtain the three calculated data matrices $\mathbf{DC}_k$ (with $1 < k < n$),

$$\mathbf{DC}_1 = \begin{vmatrix} 1.27 & 1.42 & 0.49 \\ 2.48 & 2.67 & 1.70 \\ 2.10 & 2.28 & 1.31 \\ 3.73 & 3.96 & 2.91 \\ 3.42 & 3.65 & 2.60 \end{vmatrix},$$

$$\mathbf{DC}_2 = \begin{vmatrix} 1.41 & 1.04 & 0.58 \\ 2.40 & 3.00 & 1.00 \\ 1.87 & 2.92 & 1.61 \\ 3.33 & 5.04 & 2.64 \\ 3.99 & 1.99 & 3.01 \end{vmatrix},$$

$$\mathbf{DC}_2 = \begin{vmatrix} 1.00 & 1.00 & 1.00 \\ 3.00 & 3.00 & 1.00 \\ 2.00 & 3.00 & 1.00 \\ 3.00 & 5.00 & 3.00 \\ 4.00 & 2.00 & 3.00 \end{vmatrix}.$$

If the estimated errors on the elements of **D** are

$$\begin{vmatrix} 0.4 & 0.3 & 0.4 \\ 0.5 & 0.2 & 0.4 \\ 0.6 & 0.4 & 0.2 \\ 0.5 & 0.6 & 0.2 \\ 0.4 & 0.3 & 0.3 \end{vmatrix}.$$

It appears that for $k = 1$ only the element d_{ij}, with $i = 5$ and $j = 2$, has a calculated value 3.65, which is statistically different from the original, 2.00; in fact

$$t = |3.65 - 2.00|/0.3 = 5.500 > 1.960.$$

Therefore, since just one element out of 15 is not reproducible with the first component, FP = 93.3%. For $k = 2$, it appears that all 15 elements of the data matrix **D** are statistically equivalent to the corresponding elements of the calculated data matrix $\mathbf{DC}_2$. Therefore, FP = 100.0% and the number of primary components is two. The cumulative variance percentage V_k [calculated with (16)] described by the first principal component is 68.0%, while the FP value for the same component is 93.3%. This indicates how underestimated the V_k values can be. It is, moreover, very easy to detect the only outlier with respect to the first principal component; the element d_{ij} with $i = 5$ and $j = 2$.

References

ABRAHAMS, S. C., HAMILTON, W. C. & MATHIESON, A. M. (1970). *Acta Cryst.* A**26**, 1–18.

ABRAHAMS, S. C. & KEVE, E. T. (1971). *Acta Cryst.* A**27**, 157–165.

ALLEN, F. H., BERGERHOFF, G. & SIEVERS, R. (1987). Editors. *Crystallographic Databases*. Chester: IUCr.

ALLEN, F. H., DOYLE, M. J. & TAYLOR, R. (1991*a*). *Acta Cryst.* B**47**, 29–40.

ALLEN, F. H., DOYLE, M. J. & TAYLOR, R. (1991*b*). *Acta Cryst.* B**47**, 41–49.

AUF DER HEYDE, T. P. E. (1990). *J. Chem. Educ.* **67**, 461–469.

AUF DER HEYDE, T. (1994). *Angew. Chem. Int. Ed. Engl.* **33**, 823–829.

AUF DER HEYDE, T. P. E. & BÜRGI, H.-B. (1989*a*). *Inorg. Chem.* **28**, 3960–3969.
AUF DER HEYDE, T. P. E. & BÜRGI, H.-B. (1989*b*). *Inorg. Chem.* **28**, 3970–3981.
AUF DER HEYDE, T. P. E. & BÜRGI, H.-B. (1989*c*). *Inorg. Chem.* **28**, 3982–3989.
BARTLETT, M. S. (1950). *Br. J. Psychol. Stat. Sci.* **3**, 77–83.
BÜRGI, H.-B. (1992). *Perspectives in Coordination Chemistry*, edited by A. F. WILLIAMS, C. FLORIANI & A. E. MERBACH, pp. 1–29. Basel: Verlag Helvetica Chimica Acta.
BÜRGI, H.-B. & DUNITZ, J. D. (1994). Editors. *Structure Correlation*. Weinheim: VCH.
CALLIGARIS, M., FALESCHINI, P. & CARUGO, O. (1995). In preparation.
Cambridge Structural Database (1991). Version 4.5. Cambridge Crystallographic Centre, Cambridge, England.
CARUGO, O. (1994). *Inorg. Chim. Acta*, **215**, 219–223.
CARUGO, O. (1995). In preparation.
CARUGO, O., BISI CASTELLANI, C., DJINOVIC, K. & RIZZI, M. (1992). *J. Chem. Soc. Dalton Trans.* pp. 837–841.
CARUGO, O., DJINOVIC, K. RIZZI, M. & BISI CASTELLANI, C. (1991*a*). *J. Chem. Soc. Dalton Trans.* pp. 1255–1258.
CARUGO, O., DJINOVIC, K. RIZZI, M. & BISI CASTELLANI, C. (1991*b*). *J. Chem. Soc. Dalton Trans.* pp. 1551–1555.
COMINCIOLI, V. (1992). *Metodi Numerici e Statistici per le Scienze Applicate*. Milan: CEA.
CRUICKSHANK, D. W. J. & ROBERTSON, A. P. (1953). *Acta Cryst.* **6**, 698–705.
DJINOVIC, K., CARUGO, O. & BISI CASTELLANI, C. (1992). *Inorg. Chim. Acta*, **202**, 59–65.
DJINOVIC, K., CODA, A., ANTOLINI, L., PELOSI, G., DESIDERI, A., FALCONI, M., ROTILIO, G. & BOLOGNESI, M. (1992). *J. Mol. Biol.* **226**, 227–238.
DOMENICANO, A. (1992). *Accurate Molecular Structures*, edited by A. DOMENICANO & I. HARGITTAI, pp. 437–468. Oxford Univ. Press.
DOMENICANO, A., MURRAY-RUST, P. & VACIAGO, A. (1983). *Acta Cryst.* B**39**, 457–468.
EVERITT, B. (1980). *Cluster Analysis*. New York: John Wiley.
FERRETTI, V., DUBLER-STEUDLE, K. C. & BÜRGI, H.-B. (1992). *Accurate Molecular Structures*, edited by A. DOMENICANO & I. HARGITTAI, pp. 412–436. Oxford Univ. Press.
HAMILTON, W. C. & ABRAHAMS, S. C. (1970). *Acta Cryst.* A**26**, 18–24.
HAMILTON, W. C. & ABRAHAMS, S. C. (1972). *Acta Cryst.* A**28**, 215–218.
KAISER, H. F. (1960). *Educ. Psychol. Meas.* **20**, 141–150.
MACKENZIE, J. K. (1974). *Acta Cryst.* A**30**, 607–616.
MALINOWSKI, E. R. (1991). *Factor Analysis in Chemistry*. New York: John Wiley.
MIYAMOTO, S. (1990). *Fuzzy Sets in Information Retrivial and Cluster Analysis*. London: Kluwer Academic Publishers.
MORTON, D. A. V. & ORPEN, A. G. (1992). *J. Chem. Soc. Dalton Trans.* pp. 641–653.
MURRAY-RUST, P. (1982). *Acta Cryst.* B**38**, 2765–2771.
MURRAY-RUST, P. & BLAND, R. (1978). *Acta Cryst.* B**34**, 2527–2533.
NORSKOV-LAURITSEN, L. & BÜRGI, H.-B. (1985). *J. Comput. Chem.* **6**, 216–228.
ORPEN, A. G. (1993). *Chem. Soc. Rev.* **22**, 191–196.
ROBERT, C. (1989). *Analyse Descriptive Multivarie*. Paris: Flammarion.
TAYLOR, J. R. (1982). *An Introduction to Error Analysis*. New York: Univ. Science Books.
TAYLOR, R. & ALLEN, F. H. (1994). *Structure Correlation*, edited by H.-B. BÜRGI & J. D. DUNITZ, pp. 111–161. Weinheim: VCH.
TAYLOR, R. & KENNARD, O. (1983). *Acta Cryst.* B**39**, 517–522.
TAYLOR, R. & KENNARD, O. (1985). *Acta Cryst.* A**41**, 85–89.
TAYLOR, R. & KENNARD, O. (1986). *Acta Cryst.* B**42**, 112–120.

Acta Cryst. (1995). B**51**, 328–337

Experimental and Theoretical Determination of Electronic Properties in L-Dopa

BY S. T. HOWARD, M. B. HURSTHOUSE AND C. W. LEHMANN*

School of Chemistry and Applied Chemistry, University of Wales College of Cardiff, Cardiff CF1 3TB, Wales

AND E. A. POYNER

Pharmaceutical Science Institute, Aston University, Aston Triangle, Birmingham B4 7ET, England

(*Received* 15 *September* 1994; *accepted* 7 *October* 1994)

Abstract

(2*S*)-3-(3′,4′-Dihydroxyphenyl)alanine (L-dopa), $C_9H_{11}NO_4$, $M_r = 197.19$, monoclinic, $P2_1$, $a = 13.619(6)$, $b = 5.232(2)$, $c = 6.062(3)$ Å, $\beta = 97.56(4)°$, $V = 428.191$ Å^3, $Z = 2$, $D_x = 1.529$ g cm^{-3}, $D_m = 1.515$ g cm^{-3} ($T = 293$ K), λ(Mo $K\alpha$) = 0.71069 Å, $\mu = 1.2$ cm^{-1}, $F(000) = 208$, $T = 173$ K, $R(F) = 0.017$ for 4208 reflections with $\sin\theta/\lambda < 1.078$ Å^{-1}. The electron distribution has been determined by multipole refinement with the Hansen/Coppens aspherical scattering factor expansion, including multipole terms up to octopoles for C, N and O and up to dipoles for H. The molecular dipole moment was determined as 12(2) D, within an e.s.d. of the *ab initio* value reported here of 11 D. The bond critical-point properties of the total electron density were determined, giving negative values for $\nabla^2\rho_c$ consistent with covalent bonds, and are in fair agreement with the *ab initio* results. An analysis of the hydrogen-bond critical points gave small positive $\nabla^2\rho$ values, consistent with ionic, closed-shell interactions between the participating atoms. A set of theoretical structure factors was generated from the *ab initio* charge distribution and subjected to multipole refinement, to enable a more detailed comparison with experiment.

*Permanent address: Department of Chemistry, University of Durham, Durham DH1 3LE, England.