Linear Embedding of Binary Hierarchies and Its Applications

Boris Mirkin

Abstract. The discrete binary hierarchy (DBH) is a concept underlying many important issues in analysis of complex systems: knowledge structures, test-and-search organization, evolutionary trees, taxonomy, data handling, etc. It appears that any DBH corresponds to an orthonormal basis of the Euclidean space related to the hierarchy leaves. The properties of these bases form a mathematical framework which can be applied to such problems as clustering and multiresolution image/signal processing. Clustering applications are based on a DBH-based analogue of the singular-value-decomposition of data matrices. A theoretical support for a method in divisive clustering is provided along with some decomposition-based interpretation aids. Data processing applications appear parallel to those involving the concepts of wavelets and quadtrees. However, DBH-based techniques seem to offer some potential improvements based on relaxing "continuity and homogeneity" restrictions of classical theories.

1. Introduction

The discrete binary hierarchy is a nested set of subsets ("clusters") of a finite N-element set such that any nonsingleton cluster is split in exactly two smaller clusters. It appears that any discrete binary hierarchy (in its ordered form) one-to-one corresponds to an orthonormal basis of the $N-1$-dimensional Euclidean space. The properties of these bases form a mathematical framework which is applied here to the problems of clustering and multiresolution image and signal processing.

In clustering, a divisive clustering strategy is substantiated as a method for the fitting of an approximation clustering model. The binary hierarchy provides for decompositions of the variance, covariance and the entries themselves via clusters, which gives additional interpretation aids to those usually employed in clustering. In image analysis, the binary hierarchy framework appears closely connected with some most exploited concepts, as wavelets and quadtrees, that correspond to "homogeneous and continouous" hierarchies. It should be expected that the binary hierarchies can lead to further advances in signal and image data processing by relaxing some restrictions of the classical approaches.

The remainder of the paper is arranged as follows. In Section 2, a linear embedding theory is outlined for discrete hierarchies: the concepts of hierarchy

1991 *Mathematics Subject Classification.* 62H30, 90C27,05C50, 05C05.

The author thanks the Office of Naval Research for its support under grant number N00014-96-1-0208 to Rutgers University.

and ordered hierarchy are introduced in 2.1; three-value nest indicator functions and bases for binary and non-binary hierarchies are introduced in 2.2 and 2.3; the decompositions of the data via binary hierarchies are analyzed in 2.4; between-hierarchy transformations are considered in 2.5. Section 3 is devoted to hierarchical clustering: a binary-hierarchy based approximation clustering model is analyzed in 3.1 where a sequentially fitting approach is discussed as a method of divisive clustering. Two algorithms for splitting steps of the method are introduced in 3.2. An illustrative example is treated in 3.3, accompanied with many interpretation aids derived in Section 2. In Sections 4 and 5, potential applications for processing spatial data, both uni- and two-dimensional, are considered. In 4.1, the concepts of hierarchy layers and corresponding linear subspaces are introduced and used for data compression/decompression along the hierarchy. In 4.2, parallel concepts of wavelet-based multiresolution analysis theories are described. In 5.1, a concept of bihierarchy is introduced as a device for treating planar objects such as digitalized images. Its applications to clustering and fast compression/decompression on the plane are considered in 5.2 and 5.3, respectively. In Section 6, the main issues raised in the paper are outlined.

2. Hierarchies and Corresponding Orthonormal Bases

2.1. Hierarchies and Ordered Hierarchies. Hierarchies can be represented both in graph-theoretic and in set-theoretic terms. In this paper, only set-theoretic representation will be considered. Let I be a finite set consisting of N entities. A set of its subsets $S_W = \{S_w : S_w \subseteq I, w \in W\}$ called *clusters* is a *hierarchy* if it satisfies the following properties:

1. For any $i \in I$, $\{i\} \in S_W$;
2. $I \in S_W$;
3. The clusters S_w, $w \in W$, are nested, that is, $S_w \cap S_{w'} \in \{\emptyset,\ S_w\ S_{w'}\}$, for every $w, w' \in W$;

A hierarchy is a *binary hierarchy* if it satisfies the following additional condition:

4. For every non-singleton cluster S_w, $w \in W$, there exist two clusters $S_{w1}, S_{w2} \in S_W$ which are its proper subsets, such that $S_{w1} \cup S_{w2} = S_w$.

The definition implies that the clusters $S_{w1}, S_{w2} \in S_W$ in item 4 are defined in a unique way; sometimes they are referred to as *children* of cluster S_w which is considered their *parent*.

In graph-theoretic terms, a hierarchy is a leaf-labeled rooted tree; its nodes correspond to the clusters, and edges join the parents with their children. The root corresponds to I while the singletons to the leaves, each labeled with an entity $i \in I$. Every interior node, except for the root, is adjacent to at least three other nodes. In the binary hierarchies, every non-trivial cluster (that is, not a singleton or the root) is adjacent to exactly three nodes: its parent and children.

Obviously, the number of leaves equals N while the number of edges $N - 1$. For any binary hierarchy, $N - 1$ is also the number of its non-singleton clusters.

Three rooted trees in Fig.1 present two binary hierarchies because the clusters corresponding to the nodes of trees (a) and (c) are the same. A drawn (with no intersections) tree of a binary hierarchy is what can be called an *ordered hierarchy*: the children of every internal cluster are ordered with regard to each other so that, say, the left child "precedes" the right one, according to this order. For any binary hierarchy, S_W, there are exactly $N-1$ non-singleton clusters and thus 2^{N-1}

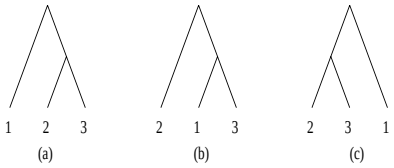

FIGURE 1. Three trees presenting two binary hierarchies on $I = \{1, 2, 3\}$.

possible ordered versions, each corresponding to a drawn (with no edge crossings) representation of the hierarchy. Obviously, an ordered hierarchy corresponds with a linear ordering, P_W of I, defined so that iP_Wj iff in the minimum cluster $S_w \in S_W$ containing both i and j, the child containing i precedes (is drawn to the left of) the child containing j. All the ordered hierarchy clusters are intervals of this unique linear ordering, P_W, of I; that is, for any $S \in S_W$, $i, j \in S$ and iP_WkP_Wj implies $k \in S$. The tree in Fig.1(a) corresponds to the natural order, 123, and that in Fig.1(b) to the order 231. Conversely, given a linear ordering, P of I, such that all clusters of S_W are its intervals, implies that a corresponding tree can be drawn with no edge crossing. This can be put as follows.

STATEMENT 1. *A hierarchy, S_W, is ordered if and only if there exists a uniquely defined linear ordering, P_W, of I such that all the hierarchy clusters are its intervals and the hierarchy order is P_W trivially extended to clusters.*

2.2. Bases for Binary Hierarchies. Let S_W be an ordered binary hierarchy. For any nonsingleton cluster $S_w = S_{w1} \cup S_{w2}$ $(w, w1, w2 \in W)$ of S_W, its three-valued *nest indicator function* ϕ_w is defined as:

$$\phi_{iw} = \begin{cases} a_w & if \quad i \in S_{w1} \\ -b_w & if \quad i \in S_{w2} \\ 0 & if \quad i \notin S_w \end{cases} \tag{2.1}$$

where the reals a_w and b_w are well defined by the following conditions: (1) vector ϕ_w is centered; that is the sum of its components is zero; (2) vector ϕ_w has its norm, that is, the square root of the sum of its components squared, equal to 1, (3) S_{w1} precedes S_{w2} in the hierarchy order. To be more precise, let us denote by n_w, n_{w1}, n_{w2} the cardinalities of clusters S_w, S_{w1} and S_{w2}, respectively. Obviously, $n_{w1} + n_{w2} = n_w$. Then, (1) means that $n_{w1}a_w - n_{w2}b_w = 0$ while (2) gives $n_{w1}a_w^2 + n_{w2}b_w^2 = 1$. These two equations lead to the following values of a_w and b_w:

$$a_w = \sqrt{\frac{n_{w2}}{n_{w1}n_w}},\ b_w = \sqrt{\frac{n_{w1}}{n_{w2}n_w}} \tag{2.2}$$

It turns out, the set of the vectors $\Phi_W = (\phi_w),\ w \in W$, is an orthonormal basis of the (N-1)-dimensional space of all the $N-$dimensional centered vectors. Since the vectors ϕ_w are centered and normed by definition, it is sufficient to prove that these vectors are mutually orthogonal.

STATEMENT 2. *Every two vectors ϕ_w and $\phi_{w'}$ from the set $\Phi_W = (\phi_w),\ w \in W$, defined for a binary hierarchy S_W by formula (2.1) are orthogonal; that is, their scalar product equals zero, $(\phi_w, \phi_{w'}) = 0$ $(w \neq w')$.*

Proof: Let us consider the scalar product $(\phi_w, \phi_{w'}) = \sum_{i \in I} \phi_{iw}\phi_{iw'}$. If $S_w \cap S_{w'} = \emptyset$ then $\phi_{iw}\phi_{iw'} = 0$ for any $i \in I$ since either $i \notin S_w$ or $i \notin S_{w'}$. Otherwise, one of the sets includes the other, say, $S_{w'} \subset S_w$, which implies that $S_{w'}$ is included in one of the children S_{w1}, S_{w2} of S_w, say, $S_{w'} \subseteq S_{w1}$. Then, $\phi_{iw} = a_w$ for any $i \in S_{w'}$, which implies that $\sum_{i \in I} \phi_{iw}\phi_{iw'} = a_w \sum_{i \in I} \phi_{iw'} = 0$, since vector $\phi_{w'}$ is centered. □

In matrix terms, the statement means that

$$\Phi^T \Phi = I_{N-1} \tag{2.3}$$

where I_n is the diagonal $n \times n$ identity matrix having all the diagonal entries equal to 1 and non-diagonal entries to 0.

It is not difficult also to prove that

$$\Phi \Phi^T = I_N - U/N \tag{2.4}$$

where U is the matrix having all its entries equal to 1 and, thus, each entry of U/N is equal to $1/N$. Equation (2.4) means that $\Phi\Phi^T$ is the orthogonal projector onto the subspace of all centered vectors.

The basis Φ_W can be considered as assigned to an unordered binary hierarchy. Since ordering subclusters, S_{w1} and S_{w2}, in this case is arbitrary, the matrix Φ corresponding to a binary hierarchy, S_W, is defined up to a right matrix factor, E, which is a diagonal matrix having its diagonal entries e_{ii} equal to 1 or -1 for any $i \in I$. The matrix ΦE corresponds to the same binary hierarchy as Φ, for any E defined above. This implies that every binary hierarchy can be ordered in 2^{N-1} ways.

2.3. Bases for Arbitrary Hierarchies. An orthonormal $(N-1)$-dimensional basis can be similarly defined for any hierarchy S_W. If $S_w \in S_W$ has $q \geq 3$ children $S_{wp} \in S_W$, $p = 1, ...q$, so that $S_w = S_{w1} \cup ... \cup S_{wq}$, a ternary nest indicator function can be defined for each of the children, S_{wp} (that is, for the edge between S_w and S_{wp}), as follows:

$$\phi_{iwp} = \begin{cases} a_{wp} & if \quad i \in S_{wp} \\ -b_{wp} & if \quad i \in S_w - S_{wp} \\ 0 & if \quad i \notin S_w \end{cases} \tag{2.5}$$

where the reals a_{wp} and b_{wp} satisfy the same conditions as above: $\sum_{i \in I} \phi_{iwp} = 0$, $\sum_{i \in I} \phi_{iwp}^2 = 1$, and S_{wp} is considered preceding $S_w - S_{wp}$. It is not difficult to prove that

$$a_{wp} = \sqrt{\frac{n_w - n_{wp}}{n_{wp} n_w}}, \; b_{wp} = \sqrt{\frac{n_{wp}}{(n_w - n_{wp}) n_w}} \tag{2.6}$$

Let us define a subset Φ of the nest indicator functions as follows. For every non-singleton $S_w \in S_W$, take in Φ all except one its nest indicator functions. It is not difficult to prove that Φ consists of exactly $N-1$ vectors.

STATEMENT 3. *The set Φ is a basis of the $(N-1)$-dimensional space of N-dimensional centered vectors. The nest indicator functions in Φ corresponding to non-siblings are mutually ortogonal.*

Proof: The same argument as in the proof of Statement 2 is applicable here except for the analysis of siblings which are absent from Φ, in the binary case. Let $S \in S_W$ consist of q children, $S_1, ..., S_q$, of which the functions $\phi_1, ..., \phi_{q-1}$

are in Φ and ϕ_q is not. Let $\sum_{p=1}^{q-1}\alpha_p\phi_p = 0$ for some reals, α_1,..., α_{q-1}. This sum, for an $i \in S_q$, equals $(-1/n)\sum_{p=1}^{q-1}\alpha_p\sqrt{\frac{n_p}{n-n_p}}$. For an $i \in S_1$, the sum differs by the first term only, which makes difference between these two values equal to $\alpha_1(\sqrt{\frac{n-n_1}{n_1}} - \sqrt{\frac{n_1}{n-n_1}})/\sqrt{n} = 0$. This implies that $\alpha_1 = 0$ if $n_1 \neq n/2$. The same is true for every coefficient α_p, $p = 1, ..., q-1$. Now let $n_1 = n/2$, which implies that equation $n_p = n/2$ is not true for any other $p < q$. Then $\alpha_2 = 0$. Considering difference of the sum values for $i \in S_1$ and $j \in S_2$ implies in this case that α_1 must be zero, too. Thus vectors $\phi_1, ..., \phi_{q-1}$ are linear independent. □

2.4. Decomposition of a Data Matrix via a Binary Hierarchy. Let us consider a $N \times n$ data matrix $Y = (y_{ik})$. Let us suppose all the columns $y_k = (y_{ik})$, $i \in I$, centered, that is, all the averages $\bar{y}_k = \sum_{i\in I} y_{ik}/N$ preliminarily subtracted from the components of corresponding column-vectors y_k, $k = 1, ..., K$.

Since every column-vector y_k, $k = 1, ..., K$ can be decomposed by the elements of basis Φ_W (for any ordered S_W), the following matrix equality holds:

$$Y = \Phi C \tag{2.7}$$

where $\Phi = (\phi_{iw})$ is the $N \times (N-1)$ matrix of values of the nest indicator functions in (2.1) and $C = (c_{wk})$ is a $(N-1) \times K$ matrix.

Since $\Phi^T\Phi$ is the identity matrix, multiplying the equality in (2.7) by Φ^T leads to

$$C = \Phi^T Y \tag{2.8}$$

which gives the value of every entry of matrix C expressed through the data as follows:

$$c_{wk} = \sum_{i\in I}\phi_{iw}y_{ik} = \sqrt{\frac{n_{w1}n_{w2}}{n_w}}(y_{w1k} - y_{w2k}) = \sqrt{\frac{n_{w1}n_w}{n_w - n_{w1}}}(y_{w1k} - y_{wk}) \tag{2.9}$$

where y_{wk}, y_{w1k} and y_{w2k} are the averages of the k-th variable in S_w, S_{w1} and S_{w2}, respectively.

It should be noted that this expression depends on the order of clusters in S_W: if Φ is changed for ΦE, then C is changed for EC. The latter expression in (2.9) is also valid for the basis corresponding to a non-binary hierarchy, as defined above.

Now consider a K-dimensional vector of the averages of the variables in a subset S_w, $w \in W$, and denote it by y_w. Then, the equality in (2.9) implies that the Euclidean norm $\sqrt{(c_w, c_w)}$ of the vector $c_w = (c_{wk})$ is equal to

$$\mu_w = \sqrt{\frac{n_{w1}n_{w2}}{n_w}}d(y_{w1}, y_{w2}) \tag{2.10}$$

where $d(x, y)$ is the Euclidean distance between vectors x, y. The value μ_w is positive if $x \neq y$, and zero if $x = y$. The norm is invariant to the between-cluster ordering and thus is well defined for nonordered binary hierarchies.

Defining M to be a diagonal $(N-1) \times (N-1)$ matrix with μ_w, $w \in W$, as its diagonal entries, and considering vectors c_w as being normed, the equation in (2.7) becomes an analogue of the singular-value decomposition (SVD) of the matrix Y (see Golub and Van Loan (1989)) since, in this case, $Y = \Phi MC$ where Φ is matrix of an orthonormal vector set and M is a diagonal matrix with nonnegative diagonal entries. The weighted distances in (2.10) are analogues to the singular values; they

will be referred to as the *cluster values* and the entries of C can be referred to as cluster loadings (by the analogy with the principal component anlysis loadings, Jolliffe, 1986).

For the sake of simplicity, the vectors c_w, $w \in W$, will not be considered normed, thus holding all the formulas above as they are.

On the other hand, the expression in (2.10) holds for any norm $\|\|$ as a function defined for the vectors $c_w, w \in W$, if the distance is accordingly defined as $d(y_{w1}, y_{w2}) = \|y_{w1} - y_{w2}\|$. Moreover, the function $\|\|$ suffices to be any monotone function thus defining d as a dissimilarity measure which might fail to satisfy some metric properties (as the triangle inequality).

Another useful property of the equation (2.7) is that

$$Y^T Y = C^T C \tag{2.11}$$

which is proved by multiplying (2.7) with its transposed version since $\Phi^T \Phi$ is the identity matrix.

Equations (2.7) and (2.11) provide us with useful decompositions of the major data characteristics via the binary hierarchy clusters. This relates to: (a) variances of the variables, (b) between-variable covariations, and (c) the entries themselves. Since the columns of Y are centered, the elements (y_k, y_l) of the matrix $Y^T Y$ have the meaning of covariance (or even correlation) coefficients between the variables k and l (multiplied by N). This allows equation (2.11) to be rewritten using formula (2.9) as follows:

$$(y_k, y_l) = \sum_{w \in W} \frac{n_{w1} n_{w2}}{n_w} (y_{w1k} - y_{w2k})(y_{w1l} - y_{w2l}). \tag{2.12}$$

When $k = l$, we have the variance of the variable k decomposed by the clusters:

$$(y_k, y_k)/N = \sum_{w \in W} \frac{p_{w1} p_{w2}}{p_w} (y_{w1k} - y_{w2k})^2. \tag{2.13}$$

Summing up equations (2.13) and employing (2.10), we arrive at an equation

$$Tr(Y^T Y)/N = \sum_{i,k} y_{ik}^2/N = \sum_{w \in W} \frac{p_{w1} p_{w2}}{p_w} d^2(y_{w1}, y_{w2}) = \sum_{w \in W} \mu_w^2 \tag{2.14}$$

decomposing the squared data scatter (the total data variance) into the sum of cluster contributions which are the cluster values squared. The last decomposition (2.7) of the entries can be expressed using (2.9), as follows:

$$y_{ik} = \sum_{\{w1 : i \in w1\}} (y_{w1k} - y_{wk}) \tag{2.15}$$

where summing is applied to all filter of proper clusters, S_{w1}, containing i (S_w is the parent of S_{w1}).

According to (2.15), it is the between-center difference, $y_{w1k} - y_{wk}$, which characterizes the contribution of a cluster, S_{w1}, to the entries of all $i \in S_{w1}$.

All the four decompositions, (2.12) – (2.15), do not depend on an ordering of S_W. The decomposition (2.14) has been employed in clustering and (2.13), (2.15) in analysis of variance (ANOVA).

2.5. Transformation Matrices. Let Φ and Φ' be basis matrices corresponding to two ordered binary hierarchies on I. According to (2.7) and (2.8),

$$\Phi' = \Phi C$$

where

$$C = C(\Phi, \Phi') = \Phi^T \Phi'.$$

The matrix $C(\Phi, \Phi')$ can be referred to as the *transformation matrix* (transforming Φ into Φ'). Exploiting the latter equation in (2.9), its entries can be expressed as:

$$c_{ww'} = \sqrt{\frac{nn_1n'_1}{n'n_2n'_2}}\left(\frac{n'_2}{n'_1}\Delta(w1') - \Delta(w2')\right) \tag{2.16}$$

where $\Delta(w')$ is the difference between the probability of $S'_{w'}$ in S_{w1} and that in S_w, and n, n_1, n_2 refer to the cardinalities of the cluster S_w and its subclusters (in the Φ-hierarchy) while n', n'_1, n'_2 relate to those of the cluster $S'_{w'}$ and its subclusters (in the Φ'-hierarchy). More precisely,

$$\Delta(w') = p(S'_{w'}/S_{w1}) - p(S'_{w'}/S_w)$$

where the conditional probability, $p(S/T)$, $S \subseteq T \subseteq I$, is defined, as usual, as $|S \cap T|/|T|$.

For any three binary hierarchies (not necessarily distinct), equation (2.4) implies

$$C(\Phi, \Phi'') = C(\Phi, \Phi')C(\Phi', \Phi'') \tag{2.17}$$

which makes the set of all the ordered binary hierarchy bases $\{\Phi\}$ a finite group since the transformation matrices are normal (that is, they "rotate" the space having their determinants equal to unity) and thus nonsingular.

Let us call two transformation matrices, C and D, as *order-equivalent* if

$$c_{ww'} = \begin{cases} d_{ww'} & \text{if} \quad (w, w') \in S_1 \times S_2 \cup (I - S_1) \times (I - S_2) \\ -d_{ww'} & \text{if} \quad (w, w') \in (I - S_1) \times S_2 \cup S_1 \times (I - S_2) \end{cases}$$

for some $S_1, S_2 \subseteq I$. Order-equivalence is obviously an equivalence relation. Its equivalence classes correspond to transformations between nonordered binary hierarchies.

STATEMENT 4. *For any two binary hierarchies, the set of all transformation matrices between their ordered versions, is an equivalence class of the order-equivalence relation.*

Statement 4 implies that the group of transformation matrices between ordered hierarchies factored with regard to the order-equivalence is not a group. Specifically, the order-equivalence classes are not closed with regard to multiplication. Let us consider, for example, hierarchies presented in Figure 1. Depending on their orderings, we may have the following transformations between them:

$$A = \begin{pmatrix} -1/2 & \sqrt{3}/2 \\ \sqrt{3}/2 & 1/2 \end{pmatrix}, \quad B = \begin{pmatrix} -1/2 & -\sqrt{3}/2 \\ -\sqrt{3}/2 & 1/2 \end{pmatrix}$$

Matrix A corresponds to the orders (123) in (a) and (213) in (b) while the orders for B are (132) and (231), respectively. Obviously, $A^TA = A^2$ and $B^TB = B^2$ are equal to the identity matrix while

$$A^TB = AB = \begin{pmatrix} -1/2 & \sqrt{3}/2 \\ -\sqrt{3}/2 & 1/2 \end{pmatrix}$$

which is not the identity matrix. Neither does it belong to the group of transformation matrices: its determinant is $1/2$, and so it is not normal.

However, the absolute values of the entries in all order-equivalent matrices are equal. This implies that transformation matrices can be used, for instance, to analyze the differences between hierarchies. It should be noted that the value (2.16) can be considered as a rather non-standard way for evaluation of between-cluster similarity. To illustrate the transformation matrices as an "analitical" device for representing geometrical differences, let us consider three hierarchies presented in Figure 2.

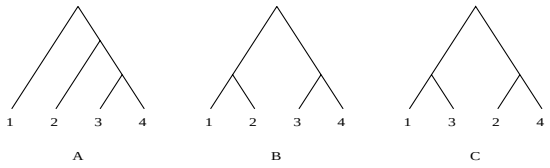

FIGURE 2. Three binary hierarchies over a four-element set.

The transformation matrices between them:

$$C(A,B) = \begin{pmatrix} \sqrt{1/3} & \sqrt{2/3} & 0 \\ \sqrt{2/3} & -\sqrt{1/3} & 0 \\ 0 & 0 & 1 \end{pmatrix}, \quad C(A,C) = \begin{pmatrix} \sqrt{1/3} & \sqrt{2/3} & 0 \\ -\sqrt{1/6} & \sqrt{1/3}/2 & \sqrt{3}/2 \\ \sqrt{1/2} & -1/2 & 1/2 \end{pmatrix},$$

and

$$C(B,C) = \begin{pmatrix} 0 & \sqrt{1/2} & \sqrt{1/2} \\ \sqrt{1/2} & 1/2 & -1/2 \\ \sqrt{1/2} & -1/2 & 1/2 \end{pmatrix}.$$

Obviously, the first matrix is somewhat simpler: hierarchies A and B in Figure 2 have a common cluster, $\{3,4\}$.

It seems quite natural to evaluate the overall between-hierarchy difference by a norm of the transformation matrix. However, the Euclidean norm, $Tr(C^TC) = \sum_{w,w'} c^2_{ww'}$ cannot do the job, because it is constantly equal to $N-1$, for any two hierarchies, as follows from (2.11) and the definition of Φ. Moreover, it can be easily proven that the matrices of squared entries of transformation matrices, $(c^2_{ww'})$, are doubly stochastic: the sum of elements in every column or row of such a matrix equal to 1.

The other norms are still available. In our example, the sums of the entries' absolute values (norm L_1) are equal to 3.79, 4.66, and 4.83, respectively, which seems in line with our intution in pair-wise comparisons of the trees in Figure 2.

3. Application to Hierarchical Clustering

3.1. Approximation Clustering Model. Traditionally, clustering is considered a discipline devoted to finding "cohesive" groups of points in a geometric space. Such a direct goal can be underlied with theoretical considerations of which we are interested in approximating. In this approach, the observed data is considered as a noisy information on the underlying discrete cluster structure. In such a setting the clustering problem is a problem of approximation of the noisy data with an adequate clustering structure (Hartigan (1972), Shepard and Arabie (1979), De Sarbo (1982), Mirkin (1990, 1994), Chaturvedi and Carroll (1994), Mirkin, Arabie, and Hubert (1995), etc.).

Let us refer to a set of subsets, $S_{W'}$, as a *hierachical cluster structure* if it satisfies requirement (3) in the definition of binary hierarchy so that clusters in $S_{W'}$ are nested, though (some) singletons or/and the root, I, may be not in $S_{W'}$. A graph corresponding to a hierachical cluster structure is a forest being only a part of a binary hierarchy tree; the leaves of the forest correspond to inclusion-minimal clusters in $S_{W'}$. The matrix of the nest indicator functions of non-leaf clusters in $S_{W'}$ will be denoted by Φ'. Obviously, any hierachical cluster structure can be completed into a binary hierarchy by further partitioning its minimal non-singleton clusters and pair-wise merging its maximal clusters.

Representing hierachical cluster structures with nest indicator functions, we arrive at the following approximation clustering model:

$$Y = \Phi' C' + E \tag{3.1}$$

where Φ' stands for a current hierachical cluster structure, C' is an unknown matrix of cluster loadings and E is the matrix of residuals to be minimized with regard to arbitrary C' and admissible Φ'.

The least-squares criterion,

$$D(\Phi', C') = Tr[(Y - \Phi' C')^T (Y - \Phi' C')] = \sum_{i=1}^{N} \sum_{k=1}^{M} (y_{ik} - \sum_{w \in W'} \phi_{iw} c_{wk})^2 \tag{3.2}$$

will be the only scalar measure of the residuals considered in this paper.

STATEMENT 5. *Given a hierachical cluster structure, Φ, the least-squares estimate for C' is determined by formula $C'(\Phi') = \Phi'^T Y$ which is analogous to (2.8). The minimum value of $D(\Phi', C')$ equals*

$$D(\Phi', C'(\Phi')) = Tr(Y^T Y) - \sum_{w \in W'} \mu_w^2 \tag{3.3}$$

where cluster values μ_w for non-leaf clusters are defined by formula (2.9) and are zeros for the leaf (minimal) clusters.

Proof: The equation $C' = \Phi'^T Y$ is derived as a necessary condition for minimality of (3.2). Putting this into (3.2), the equality $D(\Phi', C'(\Phi')) = Tr(Y^T Y - C'^T C')$ follows. Equations $Tr(C'^T C') = \sum_{w \in W'} \sum_k c_{wk}^2$ and (2.14) prove the statement. □

In fact, formula (2.14) gives decomposition of the squared data scatter, $Tr(Y^T Y)$, in two parts: explained, $\sum_{w \in W'} \mu_w^2$ and non-explained, $D(\Phi', C'(\Phi'))$,

by the cluster structure Φ'. An important feature of the formula $C'(\Phi') = \Phi'^T Y$ is that it holds only when the least-squares approximation is considered while the generic equality (2.8) holds always.

Let us define a set, A, of admissible hierachical cluster structures $S_{W'}$ by the following two conditions: (a) $I \in S_{W'}$, so that the structure is a tree, not forest, and (b) the number of clusters, $|W'|$ is fixed. When $|W'| = 2N - 1$, the set A consists of all binary hierarchies, so we consider $|W'| < 2N - 1$.

According to equation (3.3), any least-squares fit to the model (3.1) must maximize the criterion

$$\sum_{w \in W'} \mu_w^2 = \sum_{w \in W'} \frac{p_{w1} p_{w2}}{p_w} d^2(y_{w1}, y_{w2})$$

so that the problem is to find $|W'|$ consecutive divisions of I maximizing the sum of the weighted between-center distances $d^2(y_{w1}, y_{w2})$.

The author has no nontrivial suggestions on globally resolving the problem. A major issue here is that it is unknown whether the optimal structures satisfy the so-called minimal distance rule or not. The minimal distance rule requires that the distance from any point in any cluster to the cluster's center is smaller than to the center of any other cluster. This rule drastically reduces the number of potential cluster structures to check.

Thus we suggest a greedy-wise procedure of sequential extraction of clusters from the data according to the least-squares criterion. This procedure is analogous to the standard one-by-one extraction procedure of the principal component analysis and described, in a general form called the SEFIT algorithm, in Mirkin, 1990.

At each iteration of SEFIT, w, the input information includes the subtree S'_W available (initially, $S'_W = \{I\}$) and a data matrix, Y, updated. There are two steps at the iteration, according to the general procedure: (1) updating S'_W by splitting a leaf-cluster to maximize the cluster contribution, μ_w^2, added; (2) updating Y by subtracting the item found, $y_{ik} \leftarrow y_{ik} - c_{wk}\phi_{iw}$. The computation ends when w reaches a pre-specified number of clusters.

Curiously, there is no need in step (2) of updating the data matrix since the value maximized at step (1),

$$\mu_w^2 = \frac{n_{w1} n_{w2}}{n_w} d^2(y_{w1}, y_{w2}) \tag{3.4}$$

is invariant with respect to subtracting cluster values from larger clusters, because $d(x, y) = d(x - a, y - a)$ for any real a.

Thus, SEFIT in this context reduces to what has been known in the clustering discipline as the divisive clustering with splitting criterion (3.4). This criterion is well known in clustering. Ward (1963) is credited for introducing it in the agglomerative clustering context; Edwards and Cavalli-Sforza (1965) have considered the same criterion for divisive clustering. Gower (1967) provided an example demonstrating a peculiarity of the criterion reflecting the fact that factor $n_{w1} n_{w2}/n_w$ in (3.4) favors equal distribution of the entities among the clusters and, thus, the criterion may fail to immediately separate some outliers. Though for a long time treated as a shortcoming, the peculiarity does not appear to actually be so. Moreover, in many clustering studies, tendency of the cluster cardinalities to the same number

has been suggested as a good criterion of clustering (see, for example, Braverman and Muchnik, 1983).

Let us consider how the criterion (3.4) can be applied in the case when all the variables are binary descriptors of qualitative categories represented by zero-one columns (one for Yes, zero for No).

Let us compute the within-cluster average of a zero-one variable k. Do not forget, that the variable has been centered initially, which means that the entries $1 - p_k$ and $-p_k$ stand for 1 and 0, respectively, where p_k denotes the relative frequency of ones in column k.

Thus, the average is $y_{wk} = (1 - p_k)p_{wk}/p_w - p_k(1 - p_{wk}/p_w)$ where p_{wk} is the frequency of simultaneously observing descriptor k and cluster S_w and p_w is the frequency of S_w. This leads to:

$$y_{wk} = p_{wk}/p_w - p_k. \tag{3.5}$$

which implies

$$c_{wk}^2 = \frac{n_{w1}n_{w2}}{n_w}(\frac{p_{w1k}}{p_{w1}} - \frac{p_{w2k}}{p_{w2}})^2 \tag{3.6}$$

This looks quite natural: the first factor "takes care" to get the split closer to halving (which corresponds to the information concepts of the search theory) while the second follows the difference between the frequencies of ones in the subclusters. It should be noted that this measure closely resembles the so-called "twoing rule" measure used in CART techniques for conceptual clustering; see Breiman et al. (1984), p. 38, 107, 127-129.

The criterion (3.4) in this case is just the weighted distance between within-cluster probability profiles:

$$\mu_w^2 = \frac{n_{w1}n_{w2}}{n_w}d^2(p(w1), p(w2)) \tag{3.7}$$

where $p(w)$ is the vector of (conditional) probabilities of categories k in cluster S_w.

This shows that the least-squares criterion can be employed for clustering not only when all the variables are quantitative, but also when there are nominal variables present. Curiously, the formulas above fit into the problem of (multiple) alignment of biological sequences in the so-called continuous sequence space (Vingron and Sibbald, 1993). Basically, this space can be considered as a nominal data table where variables correspond to sequence positions and the categories are letters of a biomolecular alphabet.

3.2. Splitting Algorithms. Let us consider the step of splitting of a cluster S_w, in the divisive strategy, in more detail. Depending on the formula for c_{wk} in (2.9), the value of the maximized criterion μ^2 can be expressed by formula (3.4) or

$$\mu_w^2 = \frac{n_w n_{w1}}{n_{w2}}d^2(y_{w1}, y_w) \tag{3.8}$$

Computationally, criterion (3.4)/(3.8) leads to a difficult, though not NP-complete splitting task. Indeed, as is well known, the optimal splits must satisfy the mimimal distance rule, which means that the convex halls of the subclusters are linearly separated. The number of splits generated by hyperplanes is known to be less than N^K (Bock, 1974) where K is the dimensionality of the variable space, which shows the complexity of the problem. Still no further reduction of complexity of the problem has been achieved.

We describe now two local search algorithms, for each of the two formulas for μ_w^2.

Formula (3.4) implies an algorithm which is just a version of the moving-center (K-Means) technique.

Splitting by Maximizing (3.4)
Initially, the most distant points y_1 and y_2 in S_w are determined to be used as the initial centers of the clusters.
Then, sequentially, the usual next two steps are performed iteratively: (a) assigning the entities to the clusters (the nearest center wins) and (b) recomputing the centers (as the centers of gravity of the clusters obtained in (a)). The computation ends when step (a) leaves the clusters unchanged.

Evidently, this version of the K-Means technique is nothing but the alternating minimization of the square-error clustering criterion (Jain and Dubes, 1988) by two groups of the variables, those related to membership of the entities to the clusters (a) and to the cluster centers (b). Simultaneously, it is an alternating maximization algorithm for the criterion (3.4).

The second algorithm, based on formula (3.8), is a seriation algorithm.

Splitting by Maximizing (3.8)
Initially, a point y_1 is found maximizing its distance to y_w, the center of S_w, to set $S_{w1} = \{y_1\}$. On a general step, a current S_{w1} along with its center y_{w1} is considered and an entity-point y_j, closest to y_{w1} by Euclidean distance, is sought. It is added to S_{w1} if the quotient $q = d^2(y_{w1}, y_w)/d^2$, where d is the distance between y_w and the center of $S_w \cup \{y_j\}$, satisfies the inequality

$$q < \frac{n_1 n_2 + n_2}{n_1 n_2 - n_1},$$

and the process ends if not.

The inequality involved is equivalent to the fact that value of μ_w^2 (3.8) increases when y_j is added to S_w. Basically, there is a trade-off between an increase of the coefficient n_{w1}/n_{w2} and correponding decrease of the distance $d^2(y_{w1}, y_w)$. The distance may only decrease in the adding process.

Though the analogy between the one-by-one strategy of principal component analysis and the square-error divisive clustering seems rather deep, any binary hierarchy defines a different SVD-like basis while there is only one genuine SVD basis consisting of the singular vectors employed in the principal component analysis.

The algorithms described can be extended to any dissimilarity function d and, thus, amount to a family of divisive clustering algorithms which overlap but not coincide with that of Lance and Williams (1967).

A computational strategy for divisive clustering, based on the theory above, can be set as follows:

1. Standardize the entity-to-variable data by shifting the origin into the point of the variable averages and norming the variables by a chosen norm.

2. Choose a dissimilarity function (it may be different from the distance driven by the norm chosen for standardizing).
3. Choose a clustering strategy (only the divisive one has been discussed above) and create a cluster hierarchy S_W with the strategy.
4. Draw a tree hierarchy representation reflecting the cluster values μ_w by the heights of the corresponding division nodes.
5. Interpret the hierarchy designed using:

   1) the drawn pattern of clustering;

   2) contributions of the clusters and cluster–variable pairs to the square scatter of the data as reflected in values of $\sum_{k=1}^{n} c_{wk}^2$ and c_{wk}^2 (2.9), respectively ($w \in W, k = 1, ..., K$);

   3) the cluster variable-to-variable covariance/correlations, $N_{w1}N_{w2}(y_{w1k} - y_{w2k})(y_{w1l} - y_{w2l})/N_w$, as items in the additive decomposition of the overall covariance (2.12);

   4) decompositions (2.15) of the entries y_{ik} by clusters.

3.3. An Illustrative Example. Let us consider data on sorting of terms related to the human body collected by G.A. Miller (1968) and reported in Rosenberg (1982). The natural hierarchy of the body parts should be reflected in the underlying cluster structure. The four variables represent dissimilarities of 16 body terms with "Head", "Arm", "Chest", and "Leg", respectively, as presented in Table 1. The hierarchical classification found with the divisive clustering algorithm at each

TABLE 1. An extract from Miller's sorting data (1968): number of subjects (out of 50) who did not put any given row-terms into the same category with 4 column-terms presented

i	Term	Head	Arm	Chest	Leg
1	Body	45	50	37	50
2	Cheek	19	50	49	50
3	Ear	18	49	50	49
4	Elbow	49	8	50	47
5	Face	14	48	47	48
6	Hand	48	14	50	46
7	Knee	49	47	50	8
8	Lip	18	49	50	49
9	Lung	48	49	17	49
10	Mouth	19	49	50	49
11	Neck	31	45	38	45
12	Palm	50	16	49	48
13	Thigh	47	45	48	5
14	Toe	49	47	50	13
15	Trunk	42	46	19	45
16	Waist	44	45	26	46

step maximizing the contribution to the total variance is presented in Figure 3 as indexed with the corresponding cluster values (reflected in the heights of the vertical edges). The squared cluster values μ_t^2, which are equal to contributions of the cluster divisions to the total variance, are presented (per cent) for contributing the most.

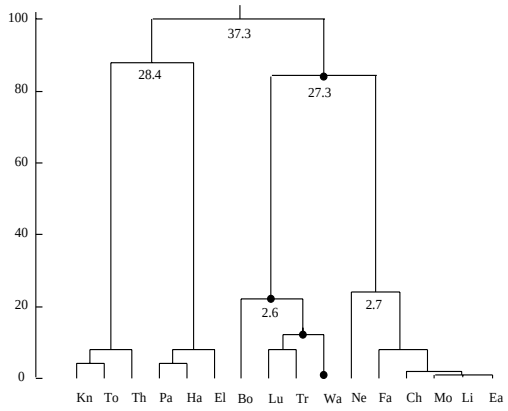

FIGURE 3. Binary hierarchy found for the data from Table 1 with the first splitting method for the least-squares criterion; the numbers show contributions of the major splits to the data variance.

The classification also leads to a decomposition of all the variances and correlations between the original variables. The general pattern of correlation is pair-wise negative as is seen in the correlation matrix:

1	1.00			
2	−0.48	1.00		
3	−0.24	−0.27	1.00	
4	−0.45	−0.15	−0.24	1.00
	1	2	3	4

.

Its decomposition by the first three separations, due to formula (2.12), is presented by the following respective matrix terms:

1	0.44			
2	−0.43	0.42		
3	0.32	−0.31	0.23	
4	−0.42	0.41	−0.30	0.40
	1	2	3	4

(first division),

1	0.00			
2	−0.01	0.56		
3	0.00	−0.01	0.00	
4	0.01	−0.57	0.01	0.58
	1	2	3	4

(second division),

1	0.43			
2	−0.02	0.00		
3	−0.54	0.02	0.60	
4	−0.01	0.00	0.02	0.00
	1	2	3	4

(third division).

These three items take into account most part of the variance and correlation. It can be seen, that all the variables are important for the first separation, although the third variable is somewhat less important (with its only 23% of the variance accounted) while the contribution of the first variable is some higher (44% of the variance). The second separation is due to the variables 2 and 4 while the third separation is made by the variables 1 and 3 (since the other variables in either case do not contribute to the variance at all).

Decomposition of the correlation coefficients confirms and details this conclusion. In particular, the negative correlations between the variables 1 and 3, as well as between 2 and 4, become positive at the first separation and sharper at the third and second separations, respectively. All the other correlations disappear in the clusters. The variance of variable 3 is not exhausted by the three first separations: this variable contributes to the separation of the smaller Head cluster.

The last interpretation aid concerns decomposition of all the standardized data entries y_{ik} by the clusters due to equation (2.15). Let us demonstrate the decomposition for the 16-th entity, Waist, belonging to the four clusters nested shown by the bold nodes in Figure 3:

1	0.52 =	−0.52+	1.09−	0.01−	0.05
2	0.28 =	0.50−	0.04−	0.06−	0.12
3	−1.46 =	−0.37−	1.20−	0.36+	0.47
4	0.36 =	0.49−	0.03−	0.05−	0.04

Every single column of the decomposition relates to its cluster reflecting the features of the cluster: the smaller values of the variables 1 and 3 in the first cluster correspond to its Head–Chest nature while the next cluster shows a split between these variables: enlarged 1 and reduced 3 correspond to the Chest membership of the entity. The last column represents individual traits of the entity.

Another tree (Figure 4) is generated with the divisive strategy when the criterion is changed for the so-called Chebyshev (uniform) metric and the second splitting algorithm has been applied. The data had been standardized as follows: the origin was shifted into the point of the average values of the variables, norming of the variables was performed by Chebyshev norm (the maximal absolute deviation from the average became one after norming was completed).

Contribution of the first split to the total variance in Figure 4 (44.9 %) is much higher than that in Figure 3 (37.3%)). This seems strange. How it could occur that the maximized contribution (in Figure 3) turned out less than the contribution achieved when another (Chebyshev) criterion was optimized (Figure 4)? To answer the question, let us consider decomposition of the variances of the variables by the

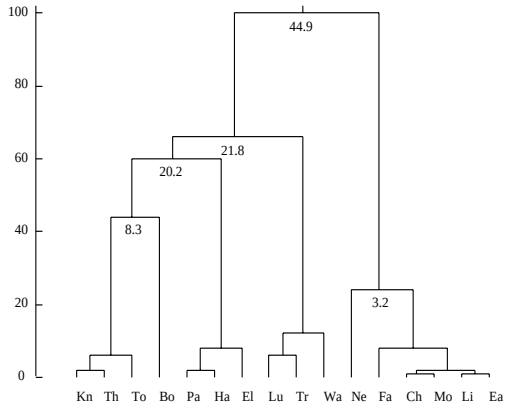

FIGURE 4. Binary hierarchy found for the data from Table 1 with the Chebyshev norm; the numbers show relative contributions of the major splits to the data variance.

clusters:

$$\begin{array}{llllll} 1 & 0.36 = & 0.33+ & 0.00+ & 0.00 + ... \\ 2 & 0.18 = & 0.03+ & 0.02+ & 0.12 + ... \\ 3 & 0.20 = & 0.02+ & 0.15+ & 0.00 + ... \\ 4 & 0.19 = & 0.03+ & 0.03+ & 0.07 + ... \end{array}$$

Again, only three major splits are represented in the decomposition. The variances (and, thus, the contributions to the square data scatter) of the variables now are different from the very beginning, which seems to determine the order they are involved in the division process: the most contributing variable 1 turns out to be the principal base of the first division; variable 3 having the second-large variance contributes mostly to the second division; the less contributing variables 3 and 4 are serving at the following divisions. Such a sequential involvement of the variables may generate a more complete account of the variance in splitting, which is reflected in the higher level of the variance extracted in the upper splits in Figure 4 as compared to those in Figure 3. This conclusion is supported by the results of the Euclidean-norm-based divisive clustering applied to the data standardized with Chebyshev norm (Figure 5). The variance contributions in the upper splits there are even greater (since the criterion is proper, in this instance); evidently, it is the left four-element cluster in Figure 4 disappearance which makes that increasing of the variances in Figure 5 possible. The contents of the clusters in the latter figure also seem quite satisfactory.

It looks that a general regularity is manifested in the example: Chebyshev norming generates a difference in the variances of the variables influencing the order of their involvement in splits (fusions) and thus increasing the contributions of the higher splits. This principle might cause the empirically observed facts that norming by range (which is quite similar to Chebyshev norming) made after centering by the average allows a best fit into Monte-Carlo generated cluster structures (Milligan and Cooper, 1988).

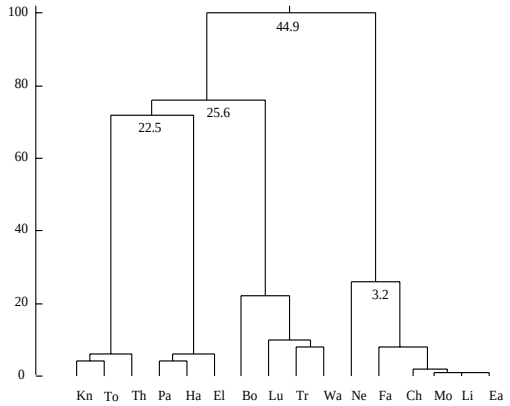

FIGURE 5. Binary hierarchy found with the least-squares criterion for the data from Table 1 normed by the Chebyshev norm

4. Application to Analysis of Spatial Data

4.1. Layered Hierarchies. The concept of ordered hierarchy fits into the so-called spatial data structures: digitalized intervals, rectangles or hyper-rectangles consisting of one-, two- or three- dimensional pixels ordered according to the coordinate axes (Samet, 1990). Let us initially consider I a unidimensional pixel set. An ordered binary hierarchy will be referred to as a spatial binary hierarchy if its order coinsides with the spatial ordering of I so that all clusters are unidimensional intervals as in the hierarchies A and B presented in Figure 6.

Any binary hierarchy can be equivalently represented by its decomposition into what will be called here layered hierarchy. A set of nested partitions of I, $\mathcal{L}$=$\{L^0, L^1, ..., L^n\}$, will be referred to as a *layered hierarchy* if (a) $L^0 = \{I\}$, (b) $L^n = \{\{i\} : i \in I\}$, (c) $L^m \subset L^{m-1}$, and (d) there exists a binary hierarchy, S_W, such that if S is a nonsingleton class of partition L^{m-1} then $S \in S_W$ and the children of S in S_W are classes of L^m $(m = 1, ..., n)$. Obviously, all classes in $\mathcal{L}$ are clusters of S_W and, moreover, there is an obvious one-to-one correspondence between the hierarchy S_W and layered hierarchy $\mathcal{L}$. The partitions $L^m \in \mathcal{L}$ will be called layers of the hierarchy.

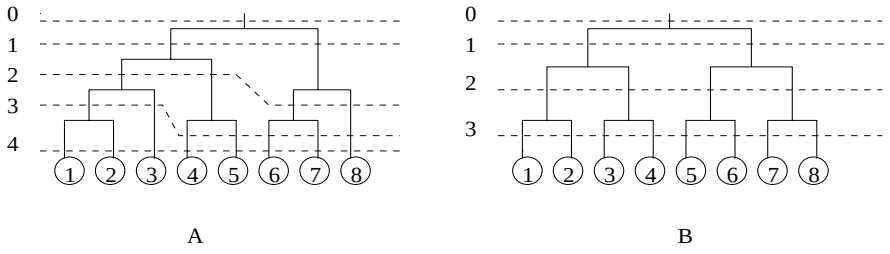

FIGURE 6. Two layered spatial binary hierarchies on an eight-element set.

The layers of hierarchies A and B in Figure 6 are presented by dashed lines. The number of layers in tree B is smaller than in A because B is a complete

hierarchy. A binary hierarchy is referred to as a *complete* binary hierarchy if every interior cluster has children splitting it into parts of equal cardinality so that the total number of entities (pixels) is a power of two, $|I| = 2^n$, and each layer L^m has exactly 2^m classes consisting of 2^{n-m} elements each ($m = 0, 1, ..., n$). Obviously, for $|I| = 2^n$, the minimum number of layers, $n + 1$, is achieved only if S_W is complete (as B in Figure 6).

In the problems of data compression, the layers of a layered hierarchy can be exploited for approximate compression of the data. More specifically, with a layer $L_m = \{L_{mt}\}$ taken, a data vector $f = (f_i)$, $i \in I$, can be substituted by the vector of within class averages $f_{mt} = \sum_{i \in L_{mt}} f_i/|L_{mt}|$, which is considered as the data at m-th level of resolution. The smaller m, the coarser the resolution; the larger m, the finer the resolution. Taking into account the spatial character of the data, a different averaging operator can be employed, giving, say, smaller weights to entities which are farther from the middle.

The layers can be used also for recalculating the averages while running along the hierarchy bottom-to-up since, obviously,

$$nf_{mt} = n_1 f_{mt1} + n_2 f_{mt2}$$

where f_{mt1} and f_{mt2} are the averages within children of L_{mt}, and n, n_1, n_2 are cardinalities of L_{mt} and its respective children. The children obviously belong in L_{m+1}. It is not difficult also to exploit the hierarchy for recalculating the averages running up-down along the hierarchy. Let us save, for every cluster S_w, in addition to f_w, the between-split difference $d_w = f_{w1} - f_{w2}$, where f_{w1} and f_{w2} are the averages of f within respective children of S_w. The formulas

$$f_{w1} = f_w + \frac{n_{w2}}{n_w} d_w, \quad f_{w2} = f_w - \frac{n_{w1}}{n_w} d_w \tag{4.1}$$

provide for calculating the average values in L_{m+1} by the averages of L_m. This allows to make decompression of the data in a fast way: to recalculate all the averages starting from any upper layer, as for instance from the grand mean $f_0 = \sum_{i \in I} f_i/|I|$. The price for that: values d_w kept along the hierarchy. The cluster cardinalities kept is a part of "hard" information about the hierarchy; they do not depend on data. Formula (4.1) becomes especially simple for a complete binary hierarchy:

$$f_{w1} = f_w + d_w/2, \quad f_{w2} = f_w - d_w/2 \tag{4.2}$$

In Figure 7, the A and B hierarchies from Figure 6 are exploited for compressing a vector f whose values are the boxed digits: F version keeps all the averages, D all the differences. It can be seen that hierarchy A provides for a safe data compression: only one average, f_0 in F, and two differences, 1.6 and 1, are needed to decompress the data entirely: the other differences are zero and thus can be dropped out of consideration. This is because hierarchy A fits into data, f, better than B does.

This methodology can be put in the linear space framework as follows.

Let us consider a layered hierarchy $\mathcal{L}$ corresponding to a binary hierarchy S_W on I. Let us define, for any $L_m \in \mathcal{L}$ and $L_{mt} \in L_m$, normed binary indicator vector χ_{mt} where $\chi_{mt}(i) = 1/\sqrt{|L_{mt}|}$ if $i \in L_{mt}$ and $= 0$ otherwise. The χ vectors corresponding to different classes of L_m are, obviously, mutually orthogonal. Let us denote by V_m the subspace in $|I|$-dimensional space generated by the normed binary indicator vectors of m-th layer, L_m. Let us denote by D_m the subspace generated by the nest indicator vectors, $\phi_{mt}(i)$, of the nonsingleton classes $L_{mt} \in L_m$.

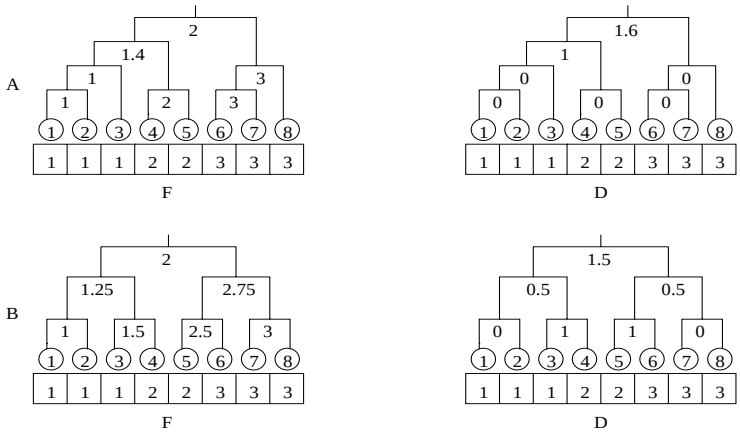

FIGURE 7. Compression and decompression of the boxed data with hierarchies A and B from Figure 6.

It is quite evident that the vectors χ_{mt} and ϕ_{mt} are pair-wise orthogonal, which implies that the spaces V_m and D_m are orthogonal too. Moreover, the following statement holds.

STATEMENT 6. *For any m ($m = 1, ..., n$), the subspace D_{m-1} is the orthogonal complement of V_{m-1} in V_m so that*

$$V_{m-1} \oplus D_{m-1} = V_m. \tag{4.3}$$

Consider a $|I|$-dimensional vector f projected into the subspace V_m:

$$f_i = \sum_t v_{mt}\chi_{mt}(i) + e_i \tag{4.4}$$

where e_i is the residual value. Due to equation (4.3), this can also be written as

$$f_i = \sum_t v_{m-1,t}\chi_{m-1,t}(i) + \sum_t c_{m-1,t}\phi_{m-1,t}(i) + e_i \tag{4.5}$$

with the same residuals.

The coefficients in (4.4) and (4.5) corresponding to a cluster $S_w \in S_W$ are: $v_w = f_w\sqrt{n_w}$ and $c_w = \sqrt{n_{w1}n_w/n_{w2}}(f_{w1} - f_w)$ where n_w, n_{w1}, n_{w2} are the cardinalities and f_w, f_{w1}, f_{w2} the within class averages for S_w and its children, S_{w1}, S_{w2}, respectively. The latter expression is the scalar product of f and ϕ_w and coincides with that in (2.9) while the former is equal to the scalar product of f and χ_w. These lead to the following formulas for fast recalculating the coefficient values along the hierarchy bottom-up:

$$v_w = v_{w1}\sqrt{n_{w1}/n_w} + v_{w2}\sqrt{n_{w2}/n_w}, \;\; c_w = v_{w1}\sqrt{n_{w2}/n_w} - v_{w2}\sqrt{n_{w1}/n_w} \tag{4.6}$$

and up-down

$$v_{w1} = c_w\sqrt{n_{w2}/n_w} + v_w\sqrt{n_{w1}/n_w}, \;\; v_{w2} = -c_w\sqrt{n_{w1}/n_w} + v_w\sqrt{n_{w2}/n_w} \tag{4.7}$$

These formulas are especially simple for complete hierarchies where all the coefficients in (4.6) and (4.7) become equal to $1/\sqrt{2}$.

4.2. Wavelets and Multiresolution Analysis. The contents of the previous section parallels some developments in data processing based on the so-called wavelet transformations. The concept of wavelet became quite popular immediately after it was introduced some ten years ago; it associates the most profound results of the theories of real-valued functions with the most urgent problems of image and other huge data compression and decompression (see, for example, reviews by Mallat (1989), Kay (1994), and Jawerth and Sweldens (1994)). The (discrete) wavelet theory involves two basic constructions: a multiresolution approximation of the space of all square-integrable real-valued functions L^2 and a dilation/translation family of functions $\chi_{mt} = 2^{m/2}\chi(2^m x - t)$ obtained from a so-called *scale* function $\chi(x)$ (which integrates to unity) with m "doubling" dilations of the space and translation of the origin by t. A basic function χ for the theory is the so-called box function $\chi(x) = \chi_{[0,1]}(x)$, that is, the indicator function of the interval $[0, 1]$ which is equal to 1 within the interval and 0 outside the interval.

The dilation/translation family may yield the functional approximation

$$f(x) = lim_{m\to\infty} \sum_t a_{mt}\chi_{mt}$$

to allow the sum $\sum_t a_{mt}\chi_{mt}$ to be considered as an approximation of any function $f \in L^2$ at resolution m for any fixed m. Here and below in this section, m and t are arbitrary integers.

A multiresolution approximation of L^2 is a sequence $\{V_m\}$ of closed subspaces of L^2 satisfying the following properties:

- M1 $V_m \subset V_{m+1}$;
- M2 The union of all V_ms is dense in L^2, and the intersection of them consists of 0 only;
- M3 $f(x) \in V_m$ if and only if $f(2x) \in V_{m+1}$;
- M4 $f(x) \in V_m \to f(x - 2^{-m}t) \in V_m$;
- M5 V_0 is isomorphic to the set of all integer sequences that are square-summable.

The meaning of the properties are as follows: V_m are approximation subspaces which are nested, thus every finer resolution $m+1$ contains all the information necessary to find the coarser resolution m (M1); the approximation can be as complete or as rough as necessary (M2); every resolution level doubles the scale (M3, M4); there is a one-to-one correspondence between the representation of f at resolution m and the coefficients a_{mt} (M5).

Let us define the subspace D_m to be the orthogonal complement of V_m in V_{m+1}. Thus, it contains all the detail lost in moving from an approximation at the finer resolution $m + 1$ to the coarser resolution m, and satisfies the equality $V_{m+1} = V_m \oplus D_m$.

It appears, given a multiresolution approximation $\{V_m\}$, there exists a unique scaling function $\chi \in V_0$ and an associated $\phi \in D_0 = V_1 - V_0$ (called a wavelet) such that $\{\chi_{mt}\}$ forms an orthonormal basis for V_m and $\{\phi_{mt}\}$ forms an orthonormal basis for D_m. Thus, for any $f \in V_m$, there are two orthonormal decompositions holding:

$$f(x) = \sum_t a_{mt}\chi_{mt}(x) \tag{4.8}$$

and

$$f(x) = \sum_t a_{m-1,t}\chi_{m-1,t}(x) + \sum_t b_{m-1,t}\phi_{m-1,t}(x) \tag{4.9}$$

Decomposition (4.9) is interpreted as the reconstruction of a finer resolution involving both the coarser resolution decomposition and the "lost detail" decomposition through wavelets. The decompositions are obvious parallels to those in (4.4) and (4.5).

The pair of the scale and wavelet functions can be taken to satisfy the following equations:

$$\chi(x) = \sum_t c_t\chi(2x-t), \quad \phi(x) = \sum_t (-1)^t c_{1-t}\chi(2x-t) \tag{4.10}$$

which implies that the wavelet function corresponding to the box function is the so-called Haar wavelet $\phi(x)$ which is equal to 1 for $0 \le x < 1/2$, -1 for $1/2 \le x < 1$, and 0 for all other x.

Graphs of the box and Haar functions are shown in Figure 8.

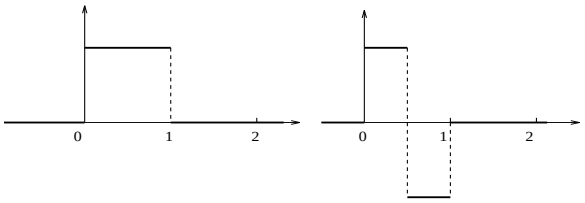

FIGURE 8. Graphs of the box and Haar functions.

The equations (4.8) to (4.10) are used to define the so-called fast wavelet transform allowing calculation of every coefficient at a finer resolution through the coefficients of a coarser resolution, and vice versa.

To apply these to image/signal processing, the following framework is employed. Let there be a pixellated unidimensional image at resolution m being a 2^m-dimensional vector v^m. This can be represented by a function $f(x) \in V_m$ defined as $f(x) = \sum_t v_t^m \chi_{mt}(x)$ where non-zero coefficients are from v^m. To calculate a coarser data sequence v^{m-1} which has half as many non-zero entries as v^m, the equations (4.8) and (4.9) are used; decompression of the data also can be done based on these equations. Moreover, the following holds.

STATEMENT 7. *The formulas (4.6) and (4.7) applied for a complete spatial binary hierarchy are a computational implementation of the fast wavelet transform based on the box scale and Haar wavelet functions.*

Sticking to the simplest box and wavelet functions restricts flexibility of the binary hierarchy approach. However, the discreteness of binary hierarchies makes up for that allowing compression and decompression of information without requiring any continuity or/and smoothness conditions which are mandatory in the classical case. Moreover, none of the "spatial" restrictions of the quantitative theories holds here: the hierarchy may be incomplete, the cluster cardinalities different, and the clusters may be spatially disconnected.

5. Extension onto Rectangle Objects

5.1. Bihierarchies and Quad-trees. The constructions above can be extended onto two-dimensional pixellated images via the following concept. A hierarchy S_W defined on $I = I' \times I''$ will be referred to as a *bihierarchy* if any of its clusters, S_w, is a Cartesian rectangle, that is, $S_w = A \times B$ for some $A \subseteq I'$ and $B \subseteq I''$, and the children of S_w are $A1 \times B1$, $A1 \times B2$, $A2 \times B1$, and $A2 \times B2$ for some partitions, $\{A1, A2\}$ and $\{B1, B2\}$, of A and B, respectively. (To allow more freedom in handling "one-dimensional" strip clusters, $\{i'\} \times B$ or $A \times \{i''\}$, we can admit some of the subsets as being empty.) The sets, A and B, can be referred to as the ranges of S_w in I' and I'', respectively. A bihierarchy will be called *spatial* if I' and I'' are ordered and the ranges of all clusters are intervals of these orders. A specific case of a bihierarchy is the Cartesian product of two binary hierarchies, $S_W = S'_{W'} \times S''_{W''}$, the clusters of which are all possible Cartesian products of clusters of $S'_{W'}$ and $S''_{W''}$.

A (divisive) bihierarchical cluster structure is an "upper" part of a bihierarchy (defined by relaxing the condition that every singleton $(i', i'') \in I' \times I''$ belongs to the bihierarchy).

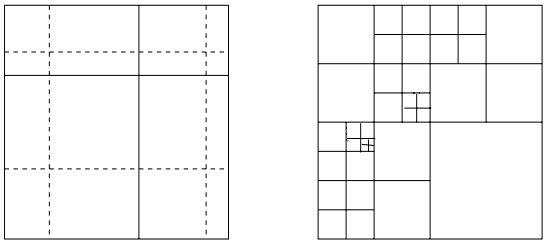

FIGURE 9. Higher splits of a Cartesian product of two spatial binary hierarchies (a) and a quad-tree (b).

A well-known structure in image data analysis, the quad-tree (see, for example, Burt and Adelson, 1983, Samet, 1990) fits into the concepts introduced. In our terms, a quadtree is a bihierarchical cluster structure for a complete spatial bihierarchy (see Figure 9, (b)).

For a cluster S_w in a bihierarchy, S_W, with its ranges A and B subdivided in $A1, A2$ and $B1, B2$, respectively, three nest indicator functions are needed according to the general description in section 2.3. A natural way of defining the indicators would be by considering the four children as produced via double dichotomy. In such a double dichotomy cluster $S_w = A \times B$ can be divided, firstly, in two strips, say, $A1 \times B$ and $A2 \times B$, and secondly, each of the strips is further split into the final children $Ak \times Bj$, $k, j = 1, 2$. The three splits can be assigned with corresponding nest indicator functions. The bihierarchy can be regarded as a contracted version of the binary hierarchy involving the double dichotomy described.

However, we'll consider here another triple of nest indicator functions (also different from those defined in section 2.3). Each of the ranges implies its nest indicator function, $\phi_A(i')$ and $\phi_B(i'')$, defined with correspondingly modified formulas (2.1) and (2.2). The three cluster nest indicator functions, ϕ_A, ϕ_B, and ϕ_{AB}, then, can be defined for all $(i', i'') \in S_w = A \times B$ as (1) $\phi_A(i', i'') = \phi_A(i')\chi_B(i'')$,

(2) $\phi_B(i',i'') = \chi_A(i')\phi_B(i'')$, and (3) $\phi_{AB}(i',i'') = \phi_A(i')\phi_B(i'')$ where $\chi_S(i) = 1/\sqrt{(|S|)}$ when $i \in S$ and $= 0$ when $i \notin S$. (When A or B is a singleton, only one of these three functions remains valid.) These functions, obviously, are centered and normed (with regard to all $(i',i'') \in I' \times I''$) and, moreover, are mutually orthogonal. Thus, the nest indicator functions of all interior clusters $S_w \in S_W$ form an orthonormal basis, Φ, of the space of $|I' \times I''|$-dimensional centered matrices (considered as vectors). The coefficients of decomposition of a matrix vector $y(i',i'')$ defined on $I' \times I''$ by the fragment of Φ related to a cluster $S_w = S_{AB}$ are scalar products of $y(i',i'')$ and corresponding nest indicator functions that can be shown to have the following format:

$$c_A = \sqrt{\frac{n_{A1}n_{A2}}{n_A}}\sqrt{n_B}(y_{1.} - y_{2.}),$$

$$c_B = \sqrt{n_A}\sqrt{\frac{n_{B1}n_{B2}}{n_B}}(y_{.1} - y_{.2}) \tag{5.1}$$

$$c_{AB} = \sqrt{\frac{n_{A1}n_{A2}}{n_A}}\sqrt{\frac{n_{B1}n_{B2}}{n_B}}(y_{11} - y_{12} - y_{21} + y_{22})$$

where y_{kj}, $y_{k.}$, or $y_{.j}$ is the average of $y(i',i'')$ on $Ak \times Bj$, $Ak \times B$ or $A \times Bj$, respectively $(k,j = 1,2)$.

These expressions can be easily extended to the situation of three-way data $Y = (y(i',i'',k))$ by adding an index k where necessary.

Applications to analysis of rectangle data can be done by extending the developments above to bihierarchies.

5.2. Bihierarchical Clustering. Following the sequential extraction strategy SEFIT discussed in section 3, we arrive at the problem of splitting the ranges of a given rectangle $A \times B \subseteq I' \times I''$ to maximize $\mu^2_{AB} = c^2_A + c^2_B + c^2_{AB}$ where the items are defined in (5.1):

$$\mu^2_{AB} = \frac{n_{A1}n_{A2}}{n_A}\frac{n_{B1}n_{B2}}{n_B}(y_{11} - y_{12} - y_{21} + y_{22})^2 + \frac{n_{A1}n_{A2}}{n_A}n_B(y_{1.} - y_{2.})^2 + n_A\frac{n_{B1}n_{B2}}{n_B}(y_{.1} - y_{.2})^2 \tag{5.2}$$

This can be done with a local search algorithm. For instance, to find an initial partition, let us split A to maximize c^2_A and, in parallel, B to maximize c^2_B. This can be done with an algorithm for splitting a cluster described in section 3.2. Then, the partition found can be iteratively updated by exchanging rows between $A1$ and $A2$ or columns between $B1$ and $B2$ (one item in a time) until μ^2_{AB} cannot be increased anymore.

5.3. Up-to-Bottom Decompression. A bihierarchy can be employed for data compression and decompression in the same fashion as it was described above for hierarchies. We will not maintain here the linear subspace terminology since it does not much differ from that described above. Let us just show how the data compressed as within cluster averages can be decompressed up-down employing the

three differences involved in (5.1) and kept as coefficients of the "wavelet" bases consisting of those parts of Φ that correspond to layers of a bihierarchy S_W:

$$d_{AB} = y_{11} - y_{12} - y_{21} + y_{22},\ d_A = y_{1.} - y_{2.},\ d_B = y_{.1} - y_{.2}.$$

STATEMENT 8. *In a bihierarchy, the children's averages can be expressed through the within cluster S_w average, y_w, and the d-coefficients above as follows:*

$$y_{11} = y_w + \frac{n_{A2}}{n_A}\frac{n_{B2}}{n_B}d_{AB} + \frac{n_{A2}}{n_A}d_A + \frac{n_{B2}}{n_B}d_B,$$

$$y_{12} = y_w - \frac{n_{A2}}{n_A}\frac{n_{B1}}{n_B}d_{AB} + \frac{n_{A2}}{n_A}d_A - \frac{n_{B1}}{n_B}d_B,$$

$$y_{21} = y_w - \frac{n_{A1}}{n_A}\frac{n_{B2}}{n_B}d_{AB} - \frac{n_{A1}}{n_A}d_A + \frac{n_{B2}}{n_B}d_B,$$

$$y_{22} = y_w + \frac{n_{A1}}{n_A}\frac{n_{B1}}{n_B}d_{AB} - \frac{n_{A1}}{n_A}d_A - \frac{n_{B1}}{n_B}d_B.$$

Proof: The proof follows with a little arithmetic from the basic equations connecting y_w, $y_{k.}$ and $y_{.j}$ with y_{kj}, $k, j = 1, 2$, as, for instance $n_A n_B y_w = n_{A1}n_{B1}y_{11} + n_{A1}n_{B2}y_{12} + n_{A2}n_{B1}y_{21} + n_{A2}n_{B2}y_{22}$, and definitions of d_{AB}, d_A, d_B. □

These formulas can be converted into the language of V_m and D_m spaces as it was done in the case of hierarchies.

6. Conclusion

The following issues discussed in the paper seem of an interest:

1. Every binary cluster hierarchy is associated with an orthonormal basis of the centered variable space providing a SVD-like decomposition of the data matrix by the elements of the cluster structure.
2. The set of interpretation aids based on the SVD-like decomposition adds the decompositions of the single variable variances, variable-to-variable covariances/correlations, and entity-to-variable entries by the clusters to the known decomposition of the overall variance.
3. An existing divisive clustering strategy can be explained as a "greedy" one-by-one fitting strategy for a clustering approximation model in terms of the SVD-like decomposition.
4. Norming of the data with norms which are different from the Euclidean one (like Chebyshev's norm related to the range of a variable rather than to its density) might lead to better clustering results because of a natural ordering of the variables emerging.
5. The binary hierarchies, applied to spatial data processing, are very closely related to wavelets and quadtrees which correspond to the so-called complete (spatially and numerically) hierarchies and bihierarchies, respectively.
6. Bottom-up and up-down computations along the binary hierarchies are parallel to the so-called fast wavelet transforms, and can be used in data compression/decompression problems.
7. The discrete character of binary hierarchies allows relaxing many restrictions of the wavelet-based techniques since the hierarchy clusters may be split into parts which are neither of equal sizes nor spatially continuous. Still, fast recalculation formulas hold for such general hierarchies and bihierarchies, which should be exploited in data processing.

8. Combining hierarchy-based clustering with the follow-up data processing may be an adequate tool for processing sets of data that have a steady structure (as documents of a kind or images of a body organ).

7. Acknowledgement

The author thanks F.R. McMorris for his numerous revising suggestions.

References

[1] Braverman, E.M., and Muchnik, I.B. (1983) *Structural Methods for Processing Empirical Data*, Moscow: Nauka (in Russian).

[2] Breiman, L., Friedman, J.H., Olshen, R.A., and Stone, C.J. (1984) *Classification and Regression Trees*, Belmont, Ca: Wadswarth International Group.

[3] Burt, P.J., and Adelson, E.H. (1983) "The Laplacian pyramid as a compact image code", *IEEE Transactions on Communications V COM-31*, 532-540.

[4] Chaturvedi, A., and Carroll, J.D. (1994) "An alternating optimization approach to fitting INDCLUS and generalized INDCLUS models", *Journal of Classification, 11*, 155-170.

[5] Diday, E. (1986) "Orders and overlapping clusters by pyramids", In: J. de Leeuw, W. Heiser, J. Meulman, and F. Critchley (Eds.) *Multidimensional Data Analysis*, Leiden: DSWO Press, 201-234.

[6] Edwards, A.W.F. and Cavalli-Sforza, L.L. (1965) "A method for cluster analysis", *Biometrics, 21*, 362-375.

[7] Golub, G.H. and Van Loan, C.F. (1989) *Matrix Computations*, Baltimore: J. Hopkins University Press.

[8] Gower, J.C. (1967) "A comparison of some methods of cluster analysis", *Biometrics, 23*, 623-637.

[9] Hansen, P., Jaumard, B., and Da Silva, E. (1993) "Average-linkage divisive hierarchical clustering", *Les Cahiers du GERAD, G-91-55*, Montréal.

[10] Hartigan, J.A. (1972) "Direct Clustering of a Data Matrix", *Journal of American Statistical Association, 67*, 123–129.

[11] Jain, A.K. and Dubes, R.C. (1988) *Algorithms for Clustering Data*, Englewood Cliffs, NJ: Prentice Hall.

[12] Jolliffe, I. T. (1986) *Principal Component Analysis*. New York: Springer-Verlag.

[13] Kay, J. (1994) "Wavelets", In *Advances in Applied Statistics, 2*, 209-224.

[14] Lance, G.N., and Williams, W.T. (1967) "A general theory of classificatory sorting strategies: 1. Hierarchical Systems", *Comp. Journal, 9*, 373-380.

[15] Mallat, S.G. (1989) "Multiresolution approximations and wavelet orthonormal bases on $L^2(R)$", *Transactions of the American Mathematical Society, 315*, 69-87.

[16] Milligan, G.W., and Cooper, M.C. (1988) "A study of standardization of the variables in cluster analysis", *Journal of Classification, 5*, 181-204.

[17] Mirkin, B.G. (1990) "A sequential fitting procedure for linear data analysis models", *Journal of Classification*, 7, 167-195.

[18] Mirkin, B.G. (1994) "Approximation of association data by structures and clusters". In: P. Pardalos, H. Wolkowicz (Eds.) *Quadratic Assignement and Related Problems*, DIMACS Series v. 16, American Mathematical Society, 293-316.

[19] Mirkin, B.G., Arabie, P., and Hubert, L. (1995) "Additive two-mode clustering: The error-variance approach revisited", *Journal of Classification, 12*, 243-263.

[20] Rosenberg, S. (1982) "The method of sorting in multivariate research with applications selected from cognitive psychology and person perception", In N. Hirchberg and L.G. Humphreys (Eds.) *Multivariate Applications in the Social Sciences*, University of Illinois at Urbana–Champaign: L. Erlbaum Assoc., 117 - 142.

[21] Samet, H. (1990) *The Design and Analysis of Spatial Data Structures*. Addison-Wesley Series on Computer Science and Information Processing. Addison-Wesley Publishing Company.

[22] Shepard, R. N., and Arabie, P. (1979) "Additive clustering: representation of similarities as combinations of discrete overlapping properties", *Psychological Review*, *86*, 87-123.

[23] Vingron, M., and Sibbald, P.R. (1993) "Weighting in sequence space: a comparison of methods in terms of generalized sequences", *Proc. Natl. Acad. Sci. USA, 90*, 8777-8781.

[24] Ward, J.H., Jr (1963) "Hierarchical grouping to optimize an objective function", *Journal of American Statist. Assoc., 58*, 236-244.

DIMACS, Rutgers University, P.O.Box 1179, Piscataway, NJ 08855-1179 USA.
E-mail address: mirkin@dimacs.rutgers.edu