Approximation Clustering: a Mine of Semidefinite Programming Problems

Boris Mirkin

DIMACS, Rutgers University, P.O.Box 1179, Piscataway, NJ 08855-1179 USA
`mirkin@dimacs.rutgers.edu`

Abstract. Clustering is a discipline devoted to finding homogeneous groups of data entities. In contrast to conventional clustering which involves data processing in terms of either entities or variables, approximation clustering is aimed at processing of the data matrices as they are. Currently, approximation clustering is a set of clustering models and methods based on approximate decomposition of the data table into scalar product matrices representing weighted subsets, partitions or hierarchies as the sought clustering structures. Some of the problems involved are of semidefinite programming, the others seem quite similar.

1 Introduction

Clustering models may differ depending on the nature of data. We distinguish here among three types of data: column-conditional, similarity and aggregable ones. The first two are those usually considered in clustering: a column-conditional data set is represented by an entity-to-variable matrix so that the entries within any column (variable) can be compared and/or averaged, and a similarity data table admits comparing and/or averaging the entries through all the table. The concept of aggregable data relates to a matrix whose entries can be summed up to their total value.

Each of the forthcoming sections is devoted to approximation clustering problems emerging for a single type of data. In section 2, the column-conditional data are considered. The concepts of conventional and approximation clustering are discussed. A bilinear clustering model, which is an extension of a nonstandard form of the principal component analysis model, is introduced. Both crisp and fuzzy cluster approximation models are considered and a doubly-greedy strategy, SEFIT, for fitting them is described. The strategy involves a major step (1), extracting clusters one-by-one, and a minor step within each of the clusters (2), completing a cluster with a local search procedure. It features a standard decomposition of the data into

"explained" and "unexplained" parts, and, in the case of crisp clustering, is closely connected with the conventional approaches. Bilinear fuzzy clustering imitates an "ideal type" concept in classification studies. Analogous analysis is provided for the case of hierarchic clustering which is presented with 3-valued indicators rather than with the traditional binary ones. All the optimization problems considered in this section are those of semidefinite programming.

In section 3, analogous constructions are applied to the similarity data. Two approximation clustering problems are discussed: additive clustering and uniform partitioning.

In section 4, approximation clustering is considered for the aggregable data. It extends the so-called correspondence analysis approach rather than that of the principal component analysis, which allows us to find more adequate formulations to approximation clustering with this type of data. The problems here extend those of SDPs to the situations when the low-rank matrices sought are rectangular rather than square ones.

The contents is briefly discussed in the conclusion (section 5).

2 Conventional and Approximation Clustering

2.1 Representation of Data and Clustering. All the column-conditional data will be considered as presented in a quantitative table format. A quantitative data table is a matrix $X = (x_{ik})$, $i \in I, k \in K$, where I is the set of entities, K is the set of variables, and x_{ik} is the value of variable $k \in K$ at entity $i \in I$. The number of entities will be denoted by N and number of the variables by n, which means that $N = |I|$ and $n = |K|$.

Based on the matrix X, the data can be considered in either of the following three geometric frameworks: (1) Space of the Entities, (2) Space of the Variables, and (3) Matrix Space. Depending on the framework chosen, the subject of clustering is considered differently. Geometric clustering, conceptual clustering and approximation clustering concepts will be considered below as those respectively related to the frameworks listed.

(1) Geometric Clustering

This is the most conventional approach, frequently referred to as the "classical clustering". The data table is considered as a set of the entities, $i \in I$, presented with corresponding row-vectors $x_i = (x_{ik})$, $k \in K$, as the elements of a space, usually Euclidean space R^n. Sometimes this n-dimensional space is referred to as the *variable* space since its dimensions correspond to the variables.

Table 1 Six two–dimensional points as given originally (variables x_1 and x_2), traditionally standardized (columns y_1, y_2), and scale-changed (columns z_1, z_2).

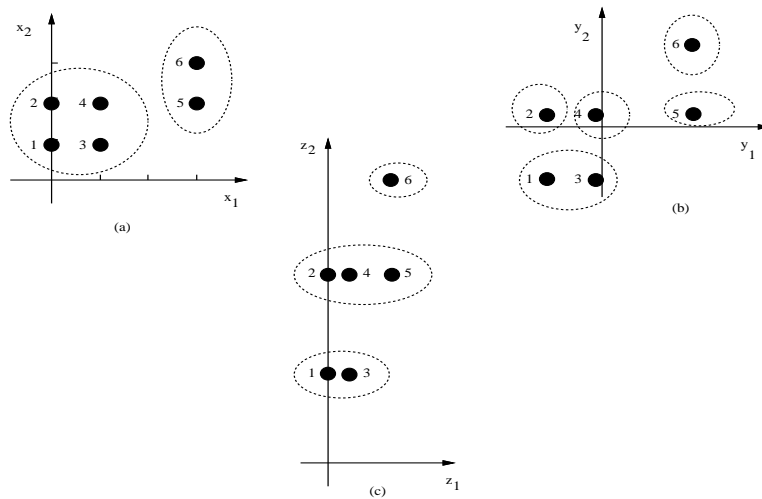| Point	x_1	x_2	y_1	y_2	z_1	z_2
1	0	1	-1.07	-1.21	0	5
2	0	2	-1.07	0.24	0	10
3	1	1	-0.27	-1.21	1	5
4	1	2	-0.27	0.24	1	10
5	3	2	1.34	0.24	3	10
6	3	3	1.34	1.70	3	15

Figure 1 The row entities presented at a visual display; tentative clusters are shown with dotted ovals.

In Fig. 1 (a), the rows from a 6 by 2 data table presented in the first two columns of Table 1 are shown as 6 points of 2-dimensional Euclidean plane of the variables x_1, x_2. Tentative clusters are encircled. The other two pictures, (b) and (c), show how the cluster pattern changes when the variables are scaled differently (both of the variables have been standardized in (b) and x_2 has been multiplied by 5 in (c)).

In the box framed, some distinctive features of the geometric approach are listed.

Geometric Clustering

Clustering: Finding cohesive isolated groups of points (Methods: K-Means, Agglomerative Clustering, ...)

Advantages:
- Geometric format
- Can treat large number of the variables

Drawbacks:
- No inferences about parameters (dissimilarities, number of clusters,...)
- No means for interpreting

---

(2) Conceptual Clustering

This approach is supposed to overcome some of the drawbacks of the geometric clustering. The data matrix is considered as a set of its columns, that is, the variables. Some features of the approach are highlighted in the framed box below.
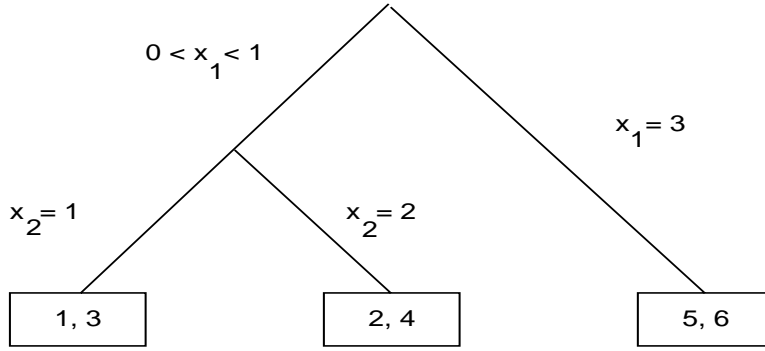
Figure 2 A conceptual clustering tree; each node corresponds to conjunction of the statements along the path from root to the node.

Conceptual Clustering

Clustering: Finding a qualitative variable(s) maximally correlated with given ones (Methods: Classification Decision Trees, Conceptual Clustering, Sequential Partitioning,...)

Advantages:
- Easy-to-interpret solution
- Large number of the entities admitted

Drawbacks:
- No inference for criterion
- Small number of variables in cluster structure
- Rigid format of cluster structure

In Fig. 2, a conceptual cluster structure based on the data in Table 1 is presented.

(3) Approximation Clustering

In this setting, it is the $N \times n$ matrix $X = (x_{ik})$ itself considered as a vector in an $(N \times n)$-dimensional space. Each cluster is represented with two items: (a) a list of the entities, S, or corresponding indicator vector $z = (z_i)$ where $z_i = 1$ if $i \in S$ and $z_i = 0$ otherwise, and (b) a standard element (called also center or centroid or prototype), $c_S \in R^n$. Let, in our example, the first cluster be $S_1 = \{1, 2, 3, 4\}$, the second, $S_2 = \{5, 6\}$, and their prototypes just the vectors of the means, $c_1 = (0.5, 1.5)$ and $c_2 = (3, 2.5)$.

This cluster structure can be related to the data via the following matrix equation:

$$
\begin{pmatrix} 0 & 1 \\ 0 & 2 \\ 1 & 1 \\ 1 & 2 \\ 3 & 2 \\ 3 & 3 \end{pmatrix} = \begin{pmatrix} 0.5 & 1.5 \\ 0.5 & 1.5 \\ 0.5 & 1.5 \\ 0.5 & 1.5 \\ \hline 3.0 & 2.5 \\ 3.0 & 2.5 \end{pmatrix} + \begin{pmatrix} -0.5 & -0.5 \\ -0.5 & 0.5 \\ 0.5 & -0.5 \\ 0.5 & 0.5 \\ 0 & -0.5 \\ 0 & 0.5 \end{pmatrix}
$$

$$
= \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 0 \\ 0 \end{pmatrix} (\ 0.5 \quad 1.5 \ ) + \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 1 \\ 1 \end{pmatrix} (\ 3.0 \quad 2.5 \ ) + \begin{pmatrix} -0.5 & -0.5 \\ -0.5 & 0.5 \\ 0.5 & -0.5 \\ 0.5 & 0.5 \\ 0 & -0.5 \\ 0 & 0.5 \end{pmatrix}
$$

This equation means that $X = ZC + E$ where Z is the cluster indicator matrix and E is the residual matrix which should be minimized with respect to the cluster structure sought. Matrix C has centroid vectors as its rows.

Some features of the approximation clustering approach are listed in the box. More detail will be provided in the remainder.

Approximation Clustering

Clustering is approximation of a given data matrix with a "cluster structure" matrix (Methods: Additive Clustering, Principal Cluster Analysis,...)

Advantages:
- Format: Data = Cluster structure + Residual
- Various cluster structures can be treated
- Can be translated in terms of the conventional approaches
- The computation parameters are interrelated
- Interpretation aids follow from the model

Drawbacks:
- No model for the residuals
- Complexity of approximation problems

Let us consider a most important geometric characteristic of the data

$$
L_p(X) = \sum_{i \in I} \sum_{k \in K} |x_{ik}|^p
$$

which is referred to as the p-scatter of the data and is the p-th power of Minkowski p-norm of the data matrix, $L_p(X) = l_p(X)^p$. This can be used as a measure of approximation of the data by the cluster structure. Obviously, for any finite p, the following decompositions hold:

$$
L_p(X) = \sum_{i \in I} d_p^p(0, i)^p = \sum_{k \in K} l_p(x_k)^p \tag{2.1}
$$

where $d_p(i, 0)$ is Minkowski p-norm (distance from zero) of row-vector $i \in I$, and x_k is the column-vector, $k \in K$. These equations may be employed in reinterpreting the matrix-based contributions to the scatter in terms of either the variables or entities or both (Mirkin [1996]).

2.2 Approximation Bilinear Model and SEFIT. Let us consider the following bilinear model corresponding to the matrix equation above:

$$y_{iv} = \sum_{t=1}^{m} c_{tv} z_{it} + \epsilon_{iv} \tag{2.2}$$

where $i \in I$ are entities,

$v \in K$ are variables or (binary) categories,

y_{iv} are observed data (standardized),

$t = 1, ..., m$ are factors or clusters, usually all unknown,

$z_t = (z_{it})$ are factor scores (when no constraints assumed) or hard cluster membership functions (1/0 values required) or fuzzy cluster membership values (when they are to be between 0 and 1),

$c_t = (c_{tv})$ are factor loadings (or, cluster standard points) with no constraints assumed,

ϵ_{iv} are residuals, to be minimized with regard to z_t and/or c_t.

The model, in matrix terms, is

$$Y = ZC + E$$

where $Y = (y_{iv})$, $Z = (z_{it})$, $C = (c_{tv})$, and $E = (\epsilon_{iv})$. The least-squares criterion, $(Y - ZC, Y - ZC)$, where scalar product is applied to the matrices considered as $N \times n$ vectors, leads to the optimal load matrix

$$C = (Z^T Z)^{-1} Z X,$$

for any given Z, since there are no constraints on values c_{tv}. Substituting this in the least-squares criterion, the problem becomes of maximizing

$$(YY^T, P_Z) \tag{2.3}$$

with regard to admissible matrices Z, where $P_Z = Z(Z^T Z)^{-1} Z^T$ is the projection matrix on the linear subspace spanning the columns of Z. When the columns of Z are mutually orthogonal, this criterion simplifies into a direct SDP format of finding of an admissible Z maximizing

$$F(Z) = (YY^T, ZZ^T) \tag{2.4}$$

when columns of Z are normed.

As it is known, the least-squares solution to the equation (2.2) can be presented by the first m elements of the singular value decomposition of matrix Y (called, in data analysis, the principal components).

In approximation clustering, hard clusters are represented by binary columns, and criterion (2.4) is valid for the partitioning (nonoverlapping clusters) problem for which it can be rewritten as

$$F(Z) = \sum_{t=1}^{m} z_t^T B z_t / z_t^T z_t = \sum_{t=1}^{m} \sum_{i,j \in S_t} b_{ij} / N_t \tag{2.5}$$

where $B = (b_{ij}) = YY^T$, and z_t is a 1/0 dummy variable corresponding to cluster $S_t = \{i : z_{it} = 1\}$ so that $N_t = (z_t, z_t) = \sum_i z_{it}$ is the cardinality of S_t.

To fit bilinear clustering model (2.2), the author has developed a doubly local search procedure which is applicable to a wide class of additive data models. In particular, it works when the clusters are supposed to be crisp non-overlapping or overlapping; fuzzy membership functions are also permitted (Mirkin [1990]). The procedure, SEFIT (SEquential FITting), finds clusters one-by-one, reiterating the

following two steps:

Sequential Fitting

(1) Find z_i^*, c_v^* minimizing
$$\sum_{i,v}(y_{iv} - c_v z_i)^2$$
with regard to arbitrary c_v and admissible z_i; take them as the solution at given iteration t;

(2) Take residual data
$$y_{iv} \Leftarrow y_{iv} - z_i^* c_v^*$$
and go to (1) until Stop-Condition is satisfied.

When the clusters are to be crisp non-overlapping, each next cluster is sought among the entities remaining unclustered yet; the computation ends when no unclustered entities remain. When clusters can be overlapping or even fuzzy, the Stop-Condition above is based on the following decomposition which holds for any sequence of vectors z_t found at Step (1) (no mutual orthogonality is required) when c_t is optimal for given z_t ($t = 1, ..., m$):

$$\sum_{i,v} y_{iv}^2 = \sum_{t=1}^{m} \sum_v c_{tv}^2 \sum_i z_{it}^2 + \sum_{i,v} \epsilon_{iv}^2 \qquad (2.6)$$

The left part is the data scatter, the right term is the least-squares criterion minimized, and the mid-term expresses that part of the data scatter which is "explained" by the clusters found; the computations should be stopped when this term is large enough or when the single cluster contribution, $\sum_v c_{tv}^2 \sum_i z_{it}^2$, becomes small.

This method is referred to as the principal cluster analysis when crisp clusters are sought or as the ideal type additive fuzzy clustering when the clusters are to be fuzzy (Mirkin [1996]).

Its relation to conceptual clustering is based on the fact that the maximized mid-term in (2.6) is equal to the sum of correlation coefficients between the partition constructed and the variables given. The correlation coefficient is the so-called correlation ratio squared when the variable is quantitative, and it is a contingency coefficient (as Pearson's chi-square) when the variable is nominal (Mirkin [1990]).

The minimization problem at Step (1) is a rank-one SDP: maximize $f(z) = z^T B z / z^T z$ with regard to all admissible N-dimensional z where $B = YY^T$, for a data matrix Y. Vector z is restricted to be Boolean, in crisp clustering, or to satisfy inequality $0 \le z \le \alpha$, in fuzzy clustering, where α is a vector whose components are unities or their complements to the cumulative membership values for the previously found clusters. The inequality $0 \le z \le \alpha$ applies when there is a prior restriction that Z must characterize a fuzzy partition so that its columns sum up to the vector having all its components equal to unity.

However, the bilinear format of the problem encourages us to apply a convenient alternating minimization strategy at Step (1), which leads to geometrically explicit results in either of the cases, crisp or fuzzy clustering. In the case of crisp clustering, Step (1) yields $c = (c_v)$ as a point in an "extreme part" of the data set while the corresponding membership is presented by those entity points that are around c rather than around 0. In the case of fuzzy clustering, the membership value, for

every point i, is proportional to the length of projection of the point onto interval between 0 and c (the membership equals zero when the projection falls out the interval) (see Mirkin [1990] for detail). Thus, explicitly involving the coefficient matrix C in the problem can be considered as a feature leading to a reasonable solution to the underlying semidefinite programming problem by exploiting the alternating minimization method.

For $m = 1$, criterion (2.5) can be related to the problem of finding a maximum density subgraph in a graph whose vertex set is I and B is the edge weight matrix. As it is known, the maximum density subgraph problem can be resolved with a polynomial-time algorithm when B is non-negative, which is usually not the case here. However, in the geometric context, the problem can be proven polynomial: the principal cluster can be separated from the other part by a linear hyperplane, which implies that the number of possible separation variants is not larger than N^n (see Bock [1974], p. 175).

The problem of finding a partition maximizing total density, (2.5), of the within-class subgraphs has been shown to be NP-complete in the case of arbitrary B. However, to the author's knowledge, it has never been explored for the case when $B = YY^T$ which is a semidefinite matrix, too. The same argument as in Bock [1974] seems work to prove that the problem with a pre-defined number of clusters is polynomial, also.

2.3 Bilinear Hierarchic Clustering. To discuss hierarchic clustering, let us remind that a set of subsets $S_W = \{S_w : S_w \subseteq I, w \in W\}$ called clusters is referred to as a binary hierarchy if it satisfies the following properties:

1. All singletons belong to S_W; that is, $\{i\} \in S_W$, for any $i \in I$;
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 $w, w' \in W$;
4. For every non-singleton cluster S_w, $w \in W$, there exist two clusters $S_{w1}, S_{w2} \in S_W$ 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 well defined; sometimes they are referred to as the children of cluster S_w which is considered their parent.

Let us define an orthonormal basis of the space of all N-dimensional centered vectors, corresponding to a binary hierarchy S_W. 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 follows:

$$\phi_{iw} = \begin{cases} a_w & if & i \in S_{w1} \\ -b_w & if & i \in S_{w2} \\ 0 & if & i \notin S_w \end{cases} \qquad (2.7)$$

where values a_w and b_w are to satisfy the following two conditions: (1) vector ϕ_w is centered, $\sum_i \phi_{iw} = 0$; (2) vector ϕ_w has its norm equal to 1. It is easy to see that

$$a_w = \sqrt{\frac{n_{w2}}{n_{w1} n_w}}, \text{ and } b_w = \sqrt{\frac{n_{w1}}{n_{w2} n_w}} \qquad (2.8)$$

where n_w, n_{w1}, and n_{w2} are cardinalities of S_w and its two children, S_{w1} and S_{w2}, respectively.

It turns out, vectors ϕ_w are mutually orthogonal, $(\phi_w, \phi_{w'}) = 0$, which is trivial when $S_w \cap S_{w'} = \emptyset$ and also true when $S_w \cap S_{w'} \neq \emptyset$ since in the latter case one

of the clusters is a part of the other and, thus, its components are non-zero when the other vector's components are constant. Therefore, the set $\{\phi_w : w \in W\}$ is an ortho-normal basis of the $(N-1)$-dimensional space of all N-dimensional centered vectors, and any column-centered data matrix Y can be decomposed as

$$Y = \Phi C \tag{2.9}$$

where $\Phi = (\phi_{iw})$ is the $N \times (N-1)$ matrix of the values of the nest indicator functions in (2.7) and $C = (c_{wk})$ is an $(N-1) \times n$ matrix.

Since $\Phi^T \Phi$ is the identity matrix, multiplying equality in (2.9) by Φ^T leads to

$$C = \Phi^T Y \tag{2.10}$$

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

$$c_{wk} = \sum_{i \in I} \phi_{iw} y_{ik} = \sqrt{\frac{n_{w1} n_{w2}}{n_w}} (\bar{y}_{w1k} - \bar{y}_{w2k}), \tag{2.11}$$

where $\bar{y}_{w1k}$ and $\bar{y}_{w2k}$ are the averages of k-th variable in S_{w1} and S_{w2}, respectively. By analogy with the factor loadings in the principal component analysis, the entries of C can be referred to as the cluster loadings.

Let us denote by y_w the n-dimensional vector of the averages of the variables in a subset S_w, $w \in W$. The equality in (2.11) implies that the Euclidean norm $\sqrt{(c_w, c_w)}$ of vector $c_w = (c_{wk})$ is equal to

$$\mu_w = \sqrt{\frac{n_{w1} n_{w2}}{n_w}} d(y_{w1}, y_{w2}) \tag{2.12}$$

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 $x = y$. It is an analogue of the singular value concept in decomposition (2.9) considered as an analogue of the singular-value decomposition.

Another useful property of decomposition (2.9) is that

$$Y^T Y = C^T C \tag{2.13}$$

which is a decomposition of the between-variable covariance (or correlation) coefficients by clusters of the hierarchy S_W.

When the hierarchy is partly unknown so that only higher clusters are given, the exact equality in (2.9) must be changed for approximate equation $Y = \Phi C + E$ where residuals in E are to be minimized. It is not difficult to prove that, when Φ is given (as a part of a basis), the least-squares estimator for C still satisfies equation (2.10). Moreover, the data scatter is decomposed as follows:

$$(Y, Y) = \sum_{t=1}^{m} \mu_t^2 + (E, E) \tag{2.14}$$

so that to find an optimal Φ requires maximizing $\sum_{t=1}^{m} \mu_t^2$.

In the framework of the SEFIT strategy, splitting are to be done sequentially, starting with the all set I, each time maximizing corresponding μ_t^2. It is exactly the criterion

$$\mu_w^2 = \frac{n_{w1} n_{w2}}{n_w} d^2(y_{w1}, y_{w2}), \tag{2.15}$$

Ward [1963] is credited for, though Ward used it for agglomerative clustering while the algorithm here is of divisive clustering.

As it is currently well-known, the criterion (2.15) added to the sum of squared Euclidean distances from the entities to the cluster centers (called "within group of squared errors" [WGSS] criterion) is equal to the overall variance of the variables in the set S_w to be partitioned. Thus, maximizing (minimizing) of (2.15) is equivalent to minimizing (maximizing) of WGSS over two-class partitions, which is a polynomial complexity task (when the space dimensionality is fixed) (see Bock [1974]). Alternating minimization strategy for this criterion is just a version of the well-known moving center (k-means) technique. 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 following 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)).

It seems doubtful that the SEFIT strategy leads to maximum $\sum_{t=1}^m \mu_t^2$ over all possible m binary splits. Moreover, it is unknown how hard the m-split problem is.

3 Clustering of Similarity Data

3.1 Additive Clustering. Let $A = (a_{ij})$, $i, j \in I$, be a given similarity or association matrix and $\lambda \mathbf{s} = (\lambda s_i s_j)$ a weighted set indicator matrix, which means that $s = (s_i)$ is the indicator of a subset $S \subseteq I$ along with its intensity weight λ somehow defined. When A can be considered as a noisy information on a set of weighted "additive clusters" $\lambda_t s_t s_t^T$, $t = 1, ...m$, the following model is assumed (Shepard and Arabie [1979], Mirkin [1987]):

$$a_{ij} = \sum_{t=1}^m \lambda_t s_{it} s_{jt} + \epsilon_{ij} \tag{3.1}$$

where ϵ_{ij} are the residuals to be minimized.

In matrix terms, the model is $A = S\Lambda S^T + E$ where S is the $N \times n$ matrix of cluster indicator functions, Λ is the $m \times m$ diagonal matrix of $\lambda_1, ..., \lambda_m$, and $E = (\epsilon_{ij})$.

When the clusters are assumed mutually nonoverlapping (that is, the indicator functions s_t are mutually orthogonal) or/and when fitting of the model is made with SEFIT strategy, the data scatter decomposition holds as follows:

$$(A, A) = \sum_{t=1}^m [s_t A s_t^T / s_t^T s_t]^2 + (E, E) \tag{3.2}$$

where the least-squares optimal λ_ts are put as the within cluster averages of the (residual) similarities.

It can be seen, from (3.2), that the least-squares fitting of the additive clustering model (in the SEFIT framework or under the nonoverlapping assumption) requires maximizing of the intermediate term in (3.2), which differs from (2.5) only in that the terms are squared here. No mathematical results on maximizing the criterion are known (beyond some properties of local search algorithms based on adding/removing/switching the entities or agglomerating the clusters).

3.2 Uniform Partitioning. A special case of model (3.1) relates to the situation when clusters are nonoverlapping and all the intensity weights λ_t are equal to the same value λ which may be pre-specified or least-squares fitted.

This clustering model can be expressed with equation $A = \lambda SS^T + E$ where λ is a real and $S = (s_{it})$ is the indicator matrix of a partition. Minimizing of

the least-squares criterion, $(A - \lambda SS^T, A - \lambda SS^T)$, is a semidefinite programming problem. This problem will be referred to as the uniform partitioning model.

With λ fixed (for the sake of brevity, assume $\lambda > 0$), the uniform partitioning criterion is equivalent to criterion

$$F(S, \lambda) = \sum_{t=1}^{m} \sum_{i,j \in S_t} (a_{ij} - \lambda/2) \tag{3.3}$$

to be maximized. Value $\lambda/2$ is a "soft" similarity threshold requiring that, in general, the larger similarities fall within the clusters, and the smaller similarities, between the clusters. The least-squares optimal λ is the within-cluster average of the similarities in A.

The rationales for considering the uniform partitioning problem include the following (Mirkin [1996]):

(1) In an optimal partition, the average within class similarity is not larger than $\lambda/2$, and the average between class similarity is not smaller than $\lambda/2$. This gives an exact meaning of "clusterness" to the uniform partition classes.

(2) In a thorough experimental study, G. Milligan [1981] has demonstrated that the usual correlation coefficient between A and SS^T belongs to the best goodness-of-fit indices of clustering results. On the other hand, the correlation coefficient characterizes quality of the matrix (bi)linear regression model, $A = \lambda SS^T + \mu U + E$ (where U is the matrix with all its entries equal to unity). Therefore, the experimental results may be considered as justifying use of the latter model as a clustering model; the uniform partitioning problem is just a shortened version of it.

(3) Criterion (3.3) appears to be equivalent to that of the index-driven consensus problem in various settings. (A partition S is called an index-driven consensus partition if it maximizes $\sum_{k=1}^{n} \xi(S^k, S)$ where $S^1, ..., S^n$ are some given partitions on I and $\xi(S^k, S)$ is a between-partition correlation index.) In particular, the problem of approximation of a graph by a graph consisting of cliques fits within this one.

(4) In the context of the so-called Lance-Williams agglomerative clustering, the uniform partitioning criterion appears to be the only one leading to the flexible Lance-Williams algorithms (with constant coefficients) as the optimization ones. We refer to an agglomerative clustering algorithm as an optimization one if its every agglomeration step, merging R_u and R_v into $R_u \cup R_v$, maximizes increment, $\delta_F(u, v) = F(R_u \cup R_v) - F(R_u) - F(R_v)$, of a cluster criterion, F.

When λ is to be least-squares adjusted, the least-squares criterion is

$$(A, SS^T)^2 / (SS^T, SS^T)$$

to be maximized by S. This problem can also be addressed in the framework of alternating optimization: (1) reiteration of the steps of optimization of criterion (3.3), with λ fixed, and (2) calculation of the within-class-average λ, for the partition found.

4 Clustering of Aggregable Data

4.1 Box Clustering. We refer to a data matrix $P = (p_{ij})$, $i \in I$, $j \in J$, as an aggregable one if it makes sense to add the entries up to their total, $p_{++} = \sum_{i \in I} \sum_{j \in J} p_{ij}$, as it takes place for contingency, flow or mobility data. The partial sums, $p_{i+} = \sum_{j \in J} p_{ij}$ and $p_{+j} = \sum_{i \in I} p_{ij}$, are called usually marginals.

There can be two different goals for the aggregable data analysis: 1) analysis within row or column set similarities, 2) analysis between row and column set interrelations.

To analyze row/column interrelations, a cluster structure called box clustering should be utilized (Mirkin, Arabie and Hubert [1995]). Two subsets, $V \subseteq I$ and $W \subseteq J$, and a real, μ, form a box cluster as presented with $|I| \times |J|$ matrix having its entries equal to $\lambda v_i w_j$ where v and w are Boolean indicators of the subsets V and W, respectively.

For the aggregable data, a specific approximation clustering strategy emerges based on the following two features: (1) it is transformed data entries, $q_{ij} = p_{ij}/p_{i+}p_{+j} - 1$, are to be approximated rather than the original data p_{ij} (q_{ij} can be considered a reasonable measure of association between i and j); (2) it is a weighted least-squares criterion employed rather than the common unweighted one (see Mirkin [1996]). To be more specific, we consider what we call box clustering model as an equation,

$$q_{ij} = \sum_{t=1}^{m} \mu_t v_{it} w_{jt} + \epsilon_{ij} \qquad (4.1)$$

to be fitted by minimizing

$$L^2 = \sum_{i \in I} \sum_{j \in J} p_{i+}p_{+j}(q_{ij} - \sum_{t=1}^{m} \mu_t v_{it} w_{jt})^2 \qquad (4.2)$$

with regard to real μ_t and Boolean v_{it}, w_{jt}, $t = 1, ..., m$.

The following rationales can be suggested to support the box clustering model:

(1) It is a clustering extension of the correspondence analysis method (widely acknowledged to be a genuine method in analysis and visualization of contingency data, see, for instance, Greenacre [1993]), as was shown in Lebart and Mirkin [1993].

(2) When a box $\mu_t v_t w_t^T$ is orthogonal to the other boxes, that is, $(v_t w_t^T, v_u w_u^T) = 0$ for any $u \neq t$, the optimal μ_t is also a q-measure applied to subsets V_t and W_t:

$$\mu_t = q_{V_t W_t} = (p_{V_t W_t} - p_{V_t+}p_{+W_t})/p_{V_t+}p_{+W_t}) \qquad (4.3)$$

where $p_{V_t W_t} = \sum_{i \in V_t} \sum_{j \in W_t} p_{ij}$.

(3) The box clusters found with a doubly-greedy SEFIT-based strategy (box clusters are extracted one-by-one, and each box cluster is formed with sequential adding/removing a row/column entity) represent quite deviant fragments of the data table (Mirkin [1996]).

The problem of finding of an optimal box, at a single SEFIT step, by maximizing

$$\frac{(\sum_{i \in V} \sum_{j \in W} p_{i+}p_{+j}q_{ij})^2}{\sum_{i \in V} p_{i+} \sum_{j \in J} p_{+j}} \qquad (4.4)$$

over $V \subseteq I$ and $W \subseteq J$, seems to be an unknown combinatorial problem deserving further investigation; yet, it is not known whether the problem can be resolved by a polynomial-time algorithm or not.

4.2 Bipartitioning. We refer to a box clustering problem as that of bipartitioning when the boxes are generated by partitions on each of the sets, I and J. Let $S = \{V_t\}$ be a partition of I, and $T = \{W_u\}$, of J, so that every pair (t, u) labels corresponding box (V_t, W_u) and its weight μ_{tu}. In corresponding specification of the model (4.1)-(4.2), the optimal values μ_{tu} are $q_{V_t W_u}$ in (4.3).

Due to mutual orthogonality of the boxes (V_t, W_u), a decomposition of the weighted squared scatter of the data, q_{ij}, onto the minimized criterion L^2 (4.2) and the bipartition part which is just the sum of terms having format of (4.4), can be made analogously to those in (2.6) and (3.2). The optimization problem here is an analogue to that in (3.2). An equivalent reformulation of the problem involves aggregation of the data based on the so-called Pearson chi-square coefficient. Let us aggregate the $|I| \times |J|$ table $P = (p_{ij})$ into $|S| \times |T|$ table $P(S, T) = (p_{tu})$ where $p_{tu} = \sum_{i \in V_t} \sum_{j \in W_u} p_{ij}$. In this notation, the original table is just $P = P(I, J)$. Then, the chi-square coefficient is defined as

$$X^2(S, T) = \sum_{t,u} \frac{(p_{tu} - p_{t+}p_{+u})^2}{p_{t+}p_{+u}}.$$

It is not difficult to see, that the data scatter decomposition, due to the bilinear model under consideration, is nothing but

$$X^2(I, J) = X^2(S, T) + L^2 \qquad (4.5)$$

which means that the bipartitioning problem is equivalent to that of finding such an aggregate table $P(S, T)$ that maximizes $X^2(S, T)$.

Alternating and agglomerating optimization clustering procedures can be easily extended to this case (Mirkin [1996]). Reformulated in geometric clustering terms, they involve the so-called chi-squared distance, an important concept in the theory of correspondence analysis (Benzécri [1973], Greenacre [1993]).

4.3 Aggregation of Flow Tables. The flow table is an aggregable data table $P = (p_{ij})$ where $I = J$ as, for instance, in brand switching or intergenerational mobility or input-output tables. Aggregation problem for such a table can be stated as that of bipartitioning with coinciding partitions, $S = T$, or equivalently, of finding an aggregate table $P(S, S)$ maximizing corresponding Pearson contingency coefficient $X^2(S, S)$. Another formulation involves finding such a partition, $S = \{V_1, ..., V_m\}$, that the aggregate q-values, q_{tu}, satisfy equations

$$q_{ij} = q_{tu} + \epsilon_{ij}, \ i \in V_t, \ j \in V_u \qquad (4.6)$$

and minimize the weighted least-squares criterion, $\sum_{t,u} \sum_{i \in V_t} \sum_{j \in V_u} p_{i+}p_{+j}(q_{ij} - q_{tu})^2$.

The rationales for this criterion include not only those listed in this section above, but also some mathematically rigorous connections between the clustering model and problems of aggregating corresponding Markov chains.

5 Conclusion

In the approximation cluster analysis framework, the following research subjects seem of particular interest: (1) developing clustering models that are adequate statistical models for substantive problems; (2) analyzing relations between approximation clustering and other data analysis approaches (first of all, the conventional cluster analysis); (3) analyzing and solving corresponding optimization problems.

As we have seen, the approximation clustering problems belong to or are closely related to semidefinite programming problems. Some of the problems are equivalent to those quite known, as multi-min-cut (which is a special case of the uniform partitioning) and maximum-density-subgraph, the others are unstudied at all (as those of box clustering). However, it seems that the bilinear format of the original

problem allows for exploiting such straightforward techniques as alternating optimization (based on both of the subspaces, Z and C, involved), which may give a computational advantage in comparison to the general semidefinite programming approaches.

More strange and unconventional are the problems arisen when the least-squares approximation criterion is changed for the less traditional least-moduli or least-maximum criteria (with the same equations): these seem have nothing to do with the semidefinite programming, though the alternating minimization approach remains a workable tool for those extended problems.

References

Benzécri, J.-P. [1973] *L'Analyse des Données*, Dunod, Paris.

Bock, H.H. [1974] *Automatische Klassifikation*, Vandenhoeck & Ruprecht, Goettingen.

Greenacre, M.J. [1993] *Correspondence Analysis in Practice*, Academic Press, San Diego, Ca.

Lebart, L., and Mirkin, B. [1993] *Correspondence analysis and classification*, In: C.M. Quadras and C.R. Rao (Eds.) Multivariate Analysis: Future Directions 2, North-Holland, Amsterdam, pp. 341-357.

Milligan, G.W. [1981] *A Monte Carlo study of thirty internal criterion measures for cluster analysis* Psychometrika, **46**, 187-199.

Mirkin, B. [1987] *Additive clustering and qualitative factor analysis methods for similarity matrices*, Journal of Classification, **4**, 7-31.

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

Mirkin, B. [1996] *Mathematical Classification and Clustering*, Kluwer Academic Press, Dordrecht.

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

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

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