L_1-*Statistical Procedures and Related Topics*
IMS Lecture Notes – Monograph Series (1997) Volume 31

L_1 and L_2 approximation clustering for mixed data: Scatter decompositions and algorithms

Boris Mirkin

Rutgers University, Piscataway, USA and
Central Economics-Mathematics Institute, Moscow, Russia

Abstract: Clustering is considered usually an art rather than a science because of lacking comprehensive mathematical theories in the discipline. The major issue raised in this paper is that L_2 and L_1 approximation bilinear clustering can provide a theoretical framework for an extensive part of partitioning and hierarchic clustering concerning its algorithmical and interpretational aspects, which is supported with a theoretical evidence.

Key words: Partitioning, hierarchy, mixed data, approximation, contingency coefficients.

AMS subject classification: 62H30, 90C27, 05C50, 05C05.

1 Introduction

Clustering is considered usually an art rather than a science because of lacking comprehensive mathematical theories in the discipline. The major issue raised in this paper is that approximation bilinear clustering can provide a theoretical framework for a part of partitioning and hierarchic clustering concerning its algorithmical and interpretational aspects. Two approximation norms, L_1 and L_2, are considered and compared.

The remainder consists of two parts devoted respectively to partitioning (Sections 2 and 3) and hierarchic clustering (Section 4), and a conclusion (Section 5). In Section 2, a bilinear model relating data to a partition is considered. The model is introduced in Section 2.1 where two model-based principles for data standardization are suggested. In Section 2.2., an L_2 decomposition of the data scatter into explained and unexplained parts is

discussed, especially in its relation to the nominal data case. It appears, some most known contingency measures, as Pearson chi-square, can be interpreted as contributions to the data scatter. Similar work is done for L_1 in Section 2.3.

In Section 3, clustering algorithms are discussed for both, L_1 and L_2, criteria. In Section 3.1, K-Means and principal cluster analysis methods are considered as locally optimal approximation techniques (the latter can be applied also for finding overlapping clusters). In Section 3.2, a prerequisite for this is outlined: interrelation between six otherwise independent parameters of cluster structure, emerging in the context of the bilinear model.

In Section 4, hierarchic clustering is put in the bilinear modeling framework. In Section 4.1, 3-valued nest indicator functions are introduced to provide for exact embedding of binary hierarchies into linear subspaces. The case of L_2 hierarchic clustering is considered in Section 4.2, which is proved similar to the case of L_2 partitioning except for that here decomposition concerns not only the data scatter but also the data entries and between-variable correlations. The case of L_1 hierarchic clustering is treated in Section 4.3. Due to the fact that the split cluster centers must be interrelated here, the alternating minimization technique produces a modified clustering approach.

2 Bilinear partition model and scatter decomposition

2.1 Bilinear model and standardization of mixed data

Let us consider an entity-to-variable data table in which a quantitative variable k is represented by a quantitative N-dimensional column-vector x_k of its values on N entities under consideration. A binary variable (category) k is basically a question admitting only answers Yes or No for each of the entities; the values are coded 1 (Yes) and 0 (No), which produces a zero-one N-dimensional column vector x_k. A nominal variable k is coded by a zero-one $N \times \#k$ submatrix $x_k = (x_{iv})$ where $\#k$ is the number of categories $v \in k$ and x_{iv} equals 1 when entity i belongs to category v of k, and 0 otherwise.

Encoded this way, the data matrix will be denoted by $X = (x_{iv})$ where $i \in I$ are entities and $v \in V$ are variables/categories corresponding to columns. These data are preprocessed into matrix $Y = (y_{iv})$ by the stan-

dard preliminary transformation (standardization) so that

$$y_{iv} = \frac{x_{iv} - a_v}{b_v},\ i \in I,\ v \in V \tag{1}$$

assuming change of both the scale factor (dividing by b) and the origin (adding of a) in the original column x_v. Choice of a and b as well as their meaning for categories will be discussed below after introducing the bilinear clustering model.

Let the entities be assigned into groups (clusters) presented by an additive type cluster structure which is a set of m clusters, any cluster t, $t = 1, ..., m$, being defined with two objects: 1) its membership function $z_t = (z_{it}), i \in I$, where z_{it} is 0 or 1 characterizing thus a cluster set $S_t = \{i \in I : z_{it} = 1\}$, 2) its standard point, or centroid vector, $c_t = (c_{tv}), v \in V$, to be combined in an $N \times |V|$ cluster-type matrix with elements $\sum_{t=1}^{m} c_{tv} z_{it}$. ($|V|$ is the number of columns in X.)

The cluster-type matrix models the given matrix Y via equations

$$y_{iv} = \sum_{t=1}^{m} c_{tv} z_{it} + e_{iv} \tag{2}$$

where residual values e_{iv} show difference between the data and the clusters. When clusters are not given a priori, they can be found in such a way that the residuals are made as small as possible, thus minimizing $\Phi(\{|e_{iv}|\})$ where Φ is an increasing monotone function of its arguments. The equations in (2) along with criterion Φ to be minimized by unknown parameters, c_{tv}, z_{it}, e_{iv}, for y_{iv} given, will be referred to as the *bilinear clustering model.* This model was suggested by the author as an extension of a version of the principal component analysis technique in Mirkin (1987) and updated in Mirkin (1990). It was considered also in Chaturvedi and Carroll (1994). A detailed account of the model and its use in hard and fuzzy clustering and machine learning can be found in Mirkin (1996).

Though the model is quite similar to that of the principal component analysis (the only difference is that the "components" z_t are Boolean, not arbitrary, vectors), it has a meaning on its own, just as a clustering model. When the clusters are required to be nonoverlapping, the type-cluster matrix $\sum_{t=1}^{m} c_{tv} z_{it}$ has especially simple structure considered also by Van Buuren and Heiser (1989): its rows are the vectors $c_t = (c_{tv})$ so that every i-th row equals c_t for that specific cluster t which contains the entity $i \in I$.

Two Minkowski forms of criterion Φ for minimizing the residuals are $L_2 = \sum_{i \in I} \sum_{v \in V} e_{iv}^2$ and $L_1 = \sum_{i \in I} \sum_{v \in V} |e_{iv}|$. With the non-overlapping

restriction, the criteria become especially simple:

$$L_p = \sum_{i \in I} \sum_{v \in V} |y_{iv} - \sum_t c_{tv} z_{it}|^p = \sum_{v \in V} \sum_{t=1}^{m} \sum_{i \in S_t} |y_{iv} - c_{tv}|^p \qquad (3)$$

which shows that L_p, actually, is $L_p = \sum_{t=1}^{m} \sum_{i \in S_t} d^p(y_i, c_t)$ where d^p is the p-th power of the Minkowski distance.

Due to formula (3), when the membership functions are given, the optimal c_{tv} is determined only by the values y_{iv} within S_t. In particular, the least-squares ($p = 2$) optimal c_{tv} is the average of y_{iv} in S_t, $c_{tv} = \sum_{i \in S_t} y_{iv}/|S_t|$, while the least-moduli (p=1) optimal c_{tv} is a median of $y_{iv}, i \in S_t$.

The criterion Φ, when its argument is the data matrix $Y = (y_{iv})$ itself, $\Phi(\{|y_{iv}|\})$, may be considered as a measure of the scatter of the data while $\Phi(\{|e_{iv}|\})$ as a measure of the "unexplained" scatter. Indeed, their difference, $\bar{\Phi} = \Phi(\{|y_{iv}|\}) - \Phi(\{|e_{iv}|\})$ will be nonnegative for any appropriate minimizer of Φ since $e_{iv} = y_{iv}$ (for all i, v) and thus $\bar{\Phi} = 0$ when all $c_{tv} = 0$ which is not an optimal solution. Value $\bar{\Phi}$ can be interpreted as the "explained" part of the data scatter $\Phi(\{|y_{iv}|\})$, which gives a decomposition of the data scatter in the two parts, $\Phi(\{|y_{iv}|\}) = \bar{\Phi} + \Phi(\{|e_{iv}|\})$.

In this setting, it is the data scatter which is decomposed into explained and unexplained parts due to the bilinear model; moreover, the unexplained part is nothing but the minimized criterion of the model. This is why the present author considers the data scatter as the base for choosing the data standardization parameters in (1).

Let us require that all the variables are standardized so that their contributions to the data scatter are equal to each other. The principle should be considered as an adequate formalization of the requirement of equal weight of the variables in numerical taxonomy (Sneath and Sokal, 1973). The choice of parameter a_v does not affect the model (2) for a non-overlapping cluster structure, however when the bilinear model is set forth in a sequential way with the "component" axes z_t identified one-by-one, not simultaneously, the solution heavily depends on the origin of the variable/category space. To adjust to this kind of principal/correspondence analysis methods, let us postulate an analogue to the law of minimum moment of inertia in mechanics: the origin of the variable space should be a minimizer of the data scatter.

The two scatter-based principles make the parameters defined unanimously for L_1 and L_2. When $p = 2$, they lead to the usual z-score standardization rule: the origin is the grand mean while the standard deviation is the scale factor, which will be referred to as *square-scatter standardization*. When $p = 1$, the origin must be grand median while the scale

factor is the absolute deviation, which will be referred to as *module-scatter standardization.*

In the case of mixed data, the average of a category $v \in V$ column vector is equal, obviously, to the relative frequency of the category in I, p_v, while its median may be 1, 1/2, or 0 depending on whether p_v is larger than, equal to, or smaller than 1/2, respectively. To satisfy the principle of equal contribution with $a_v = p_v$, the L_2-based scale factor of a category v can be taken as $b_v = \sqrt{1 - \sum_v p_v^2}$ where summation is made by all the categories of a variable k, $v \in k$ (the square root of Gini index). There can be also other standardizing options suggested as, for instance, $b_v = \sqrt{(\#k - 1)p_v}$ which is category-specific.

The absolute deviation of the values of a binary column vector from the median is equal to p_v or $1 - p_v$ depending on whether p_v is less than 1/2 or not.

2.2 Decomposition of the least-squares criterion

With the least-squares criterion, the following decomposition holds (see, for instance, Jain and Dubes, 1988).

Statement 1 *If values c_{tv} are optimal for a partition $S = \{S_t\}$ of I, then*

$$\sum_{i \in I} \sum_{v \in V} y_{iv}^2 = \sum_{t=1}^{m} \sum_{v \in V} c_{tv}^2 |S_t| + \sum_{i \in I} \sum_{v \in V} e_{iv}^2, \tag{4}$$

Usually the equation in (4) is interpreted in terms of analysis of variance. In cluster analysis, interpretation of (4) in terms of the contributions to data scatter seems more helpful. The contribution of a pair variable-cluster (v, t) to the explained part of the data scatter is $c_{tv}^2 |S_t|$: it is proportional to the cluster cardinality and to the squared distance from the grand mean of the variable to its mean (standard value) within the cluster. The contribution of an entity-cluster pair can be evaluated as (y_i, c_t) because $c_{tv}^2 |S_t| = (\sum_{i \in S_t} y_{iv}/|S_t|) c_{tv} |S_t| = \sum_{i \in S_t} y_{iv} c_{tv}$. These cluster-specific salience weights of the variables and entities can be employed for concept learning and feature selection in machine learning (Mirkin, 1997).

To analyze the contributions of nominal variables and their categories to the scatter part explained via cluster partition S, let us denote the relative frequency (proportion of ones) of category v in set I by p_v and the proportion of entities simultaneously having category v and belonging to cluster S_t, by p_{vt}. Then, for any category v standardized by formula (1), its mean within cluster S_t is equal to $c_{tv} = (p_{vt} - p_t a_v)/(p_t b_v)$. The contribution of a category-cluster pair (v, t) to the explained part of the

data scatter is equal to

$$s(v,t) = c_{tv}^2|S_t| = N(p_{vt} - p_t a_v)^2/(p_t b_v^2), \tag{5}$$

which can be considered a measure of association between category v and cluster t.

Since every nominal variable k is considered as the set of its categories v, the joint contribution of k and the set of the clusters S_t to the scatter of the data is equal to $F(k,S) = \sum_t \sum_{v \in k} s(v,t)$ which is

$$F(k,S) = N \sum_{t=1}^{m} \sum_{v \in k} \frac{(p_{vt} - p_t a_v)^2}{p_t b_v^2} \tag{6}$$

by (5). Substituting the appropriate values of $a_v = p_v$ and b_v, we arrive at the following.

Statement 2 *For criterion L_2, the contribution of a nominal variable $k \in K$ to the part of the square scatter of the square standardized data that is explained by the (sought or found or expert-given) cluster partition $S = \{S_1, ..., S_m\}$, is equal to*

$$\Delta(S/k) = N \sum_{v \in k} \sum_{t=1}^{m} \frac{(p_{vt} - p_v p_t)^2}{p_t} \tag{7}$$

when $b_v = 1$ (no normalization), or

$$W(R/k) = N \sum_{v \in k} \sum_{t=1}^{m} \frac{(p_{vt} - p_v p_t)^2/p_t}{1 - \sum_{v \in k} p_v^2} \tag{8}$$

when $b_v = \sqrt{1 - \sum_{v \in k} p_v^2}$ (a standardizing option suggested), or

$$M(S/k) = \frac{N}{\#k - 1} \sum_{v \in k} \sum_{t=1}^{m} \frac{(p_{vt} - p_v p_t)^2}{p_v p_t} \tag{9}$$

when $b_v = \sqrt{p_v(\#k - 1)}$ (another standardizing option).

All three of the coefficients relate to well known indices of contingency between the nominal variables: $M(S/k)$ is a normalized version of the Pearson chi-square coefficient, $\Delta(R/k)$ is proportional to the coefficient of reduction of the error of proportional prediction, and $W(R/k)$ is the Wallis coefficient. Amazingly, it is the method of data standardization which determines which of the coefficients is produced as the contribution-to-scatter.

The contribution of a quantitative variable into the explained part of the L_2 data scatter is also meaningful. When the variable k is standardized, it is exactly $N\eta^2(k,S)$ where $\eta^2(k,S)$ is the so-called correlation ratio (squared).

2.3 Least-moduli decomposition

Similar decomposition can be done for the least-moduli criterion (the contents of this section is a corrected version of section 6.1.4 in Mirkin, 1996).

Let S_t be an entity subset, and $S_{tv} = \{i \in S_t : |y_{iv}| < |c_{tv}| \ \& \ sgn\ y_{iv} = sgn\ c_{tv}\}$ where, as usual, $sgn\ x$ is 1 if $x > 0$, 0 if $x = 0$, and -1 if $x < 0$. This means that $S_{tv} = \{i \in S_t : 0 \leq y_{iv} < c_{tv}\}$ if c_{tv} is positive or $S_{tv} = \{i \in S_t : c_{tv} < y_{iv} \leq 0\}$ if c_{tv} is negative. Having a value c_{tv} fixed, the set S_t is partitioned into three subsets by the variable/category v depending on relations between y_{iv}, $i \in S_t$, and c_{tv}. For $c_{tv} > 0$, let us denote the cardinalities of the subsets where y_{iv} is larger than, equal to or less than c_{tv} by n_{tv1}, n_{tv2} and n_{tv3}, respectively. Then, let $n_{tv} = n_{tv1} + n_{tv2} - n_{tv3}$. For $c_{tv} < 0$, the symbols n_{tv1} and n_{tv3} are interchanged. If c_{tv} is the median of values y_v in S_t and all the values y_{iv}, $i \in S_t$, are different, then $n_{tv1} = n_{tv3}$ and $n_{tv} = n_{tv2} = 0$ or $= 1$ depending on the cardinality of S_t (even or odd, respectively).

Statement 3 *When values c_{tv} are L_1-optimal for a partition $S = \{S_t\}$ of I, the following decomposition of the module data scatter holds:*

$$\sum_{i \in I} \sum_{v \in V} |y_{iv}| = \sum_{v \in V} \sum_{t=1}^{m} (2 \sum_{i \in S_{tv}} |y_{iv}| + n_{tv}|c_{tv}|) + \sum_{i \in I} \sum_{v \in V} |e_{iv}|. \qquad (10)$$

The proof is based on the following equation, $|a-b| = |a|+|b|-|sgn\ a+sgn\ b| \min(|a|,|b|)$, which holds for any real a and b.

Let us denote the contribution of a variable-cluster pair (v,t) to the module scatter in (10) by $s(t,v) = 2\sum_{i \in S_{tv}} |y_{iv}| + n_{tv}|c_{tv}|$. Based on this, various relative contribution measures can be defined: (a) variable to scatter, $w(v) = \sum_t s(t,v) / \sum_{i,v} |y_{iv}|$; (b) cluster to scatter, $w(t) = \sum_v s(t,v) / \sum_{i,v} |y_{iv}|$; (c) variable to cluster, $w(v/t) = s(t,v) / \sum_v s(t,v)$; (d) entity to cluster, $w(i/t) = |sgn\ y_{iv} + sgn\ c_{tv}| \min(|y_{iv}|, |c_{tv}|) - |c_{tv}|$.

Let us consider the case when v is a category.

Statement 4 *For any category v standardized (with arbitrary a_v and b_v), its median, c_{tv}, in cluster S_t is equal to $-a_v/b_v$, $(1-2a_v)/2b_v$, or $(1-a_v)/b_v$ depending on p_{tv} is smaller than, equal to, or greater than $0.5p_t$, respectively. The contribution of a category-cluster pair, (v, S_t), to the module data scatter is equal to*

$$s(t,v) = N|2p_{vt} - p_t||c_{tv}|.$$

Proof: The formulas for c_{tv} are evident. To derive the formula for $s(t,v)$, let us see that $S_{tv} = \emptyset$ since the values c_{tv} and y_{iv} must have different signs

if they are not equal to each other (y_{iv} may have one of two values only since v is a category). Then, $n_{tv1} + n_{tv2} = Np_{tv}$ and $n_{tv3} = N(p_t - p_{tv})$ when $c_{tv} = (1 - a_v)/b_v > 0$ where a_v, b_v are the values used in the module-scatter standardization rule. Analogously, $n_{tv1} + n_{tv2} = N(p_t - p_{tv})$ and $n_{tv3} = Np_{tv}$ when $c_{tv} = -a_v/b_v < 0$. □

Putting the module-scatter based a_v and b_v into $s(t,v)$, we have:

Statement 5 *The contribution of a nominal variable k to the absolute scatter of the module-scatter standardized data, as explained by the partition $S = \{S_1, ..., S_m\}$, is equal to*

$$A(S/k) = \frac{N}{\#k}\left(\sum_{(v,t)\in A_+} \frac{2p_{vt} - p_t}{p_v} + \sum_{(v,t)\in A_-} \frac{p_t - 2p_{vt}}{1 - p_v} + \sum_{(v,t)\in A_=} |2p_{vt} - p_t|\right) \tag{11}$$

where $A_+ = \{(v,t) : p_v < 0.5$ and $p_{vt}/p_t > 0.5\}$, $A_- = \{(v,t) : p_v > 0.5$ and $p_{vt}/p_t < 0.5\}$, and $A_= = \{(v,t) : p_v = 0.5\}$.

The coefficient $A(S/k)$ takes into account the situations when the patterns of occurrences of the categories $v \in k$ in the clusters t differ from those in the entire set I. Such a difference appears when v is frequent in S_t ($p(v/t) > 0.5$) and rare in I ($p_v < 0.5$), or, conversely, v is rare in S_t and frequent in I.

3 K-Means and bilinear partitioning

3.1 Principal clustering and K-Means

Following the standard strategy of sequential extraction of factors in principal component analysis, the clusters can be extracted one by one in (2), which constitutes the method of principal cluster analysis (PCL) (applicable in both overlapping and non-overlapping cluster cases).

1. Set $t = 1$ and define data matrix Y_t as the initial data matrix standardized according to the criterion, L_1 or L_2, chosen as described in Section 2.1. Choose whether the clusters to be found are required to be nonoverlapping or they may overlap each other.

2. For $Y = Y_t$ find a principal cluster minimizing L_1 or L_2 as described in the algorithm for Single Cluster Clustering (SCC) below. Define z_t, c_t as the cluster solution found (membership function and the standard point, respectively); compute its contribution to the data scatter.

3. Stop-Condition. If there must be nonoverlapping hard clusters, check whether there are yet unclustered entities remaining. (In the other case,

check the standard contribution-based stopping rule of the principal component analysis). If yes, go to 4; else end.

4. Compute the residual data $y_{iv}^{(t+1)} = y_{iv}^t - c_{tv}z_{it}$. In the nonoverlapping case, set Y_{t+1} by removing from Y_t all the rows corresponding to the previously found cluster t. Increase t by 1, and go to 2.

The PCL algorithm can be rephrased in terms of K-Means method (MacQueen (1967), Jain and Dubes (1988)), which, in its "parallel" version, starts with m somehow selected tentative standard points or "seeds", c_t. Then the algorithm repeatedly performs the following two-step iteration : (1) update the partition based on the standard points : given c_t, make each S_t the set of y_i that are nearest to c_t, $t = 1, ..., m$; (2) update the standard points: when all S_t are given, compute c_t as the mean (or median, for L_1) of the within-cluster vectors. This algorithm is, in fact, a version of the alternating minimization for criteria L_p : (1) given c, find optimal z; (2) given z, find optimal c.

The principal cluster analysis can be considered as a technique that exploits many of the same mechanisms, but which mitigates the need for prior knowledge, and separates clusters from the set of instances one by one. First, an initial cluster $S_1 \subset I$ is extracted with its standard point c_1; the complementary set represents the main "body" of entities, which serves as the source for separating additional clusters one by one. This is reflected in that fact that the main body's standard point is fixed at 0, given the square or module scatter standardization, and it is not changed during the entire clustering computation. The algorithm SCC (Single Cluster Clustering) for separating a principal cluster at the Step 2 of PCL is as follows (the data matrix is denoted by Y, not Y_t):

Step 1. (Selection of an extreme point). Pick a point, y_{i^*}, maximizing distance $d(0, y_i)$, $i \in I$, from the origin (the distance is taken according to the criterion, L_1 or L_2, selected : city-block or Euclidean squared). Take $c = y_{i^*}$ as the initial center (seed) of the cluster to be found.

Step 2. (Updating of the cluster). Define cluster S of points y_i around the center c as $S = \{i : d(y_i, c) < d(y_i, 0)\}$.

Step 3. (Updating of the standard point). Compute the center of S, $c = c(S)$, which is the median or average vector, depending on L_1 or L_2 is utilized.

Step 4. (Stop condition) Compare S with that at the previous iteration. If there is no difference, the process ends: S and $c(S)$ are the result. Else go to Step 2.

A general property is that the size of an SCC-designed cluster depends

on its distance from the origin (which is just a reference point): the nearer to that point, the less the diameter of the cluster! Thus, SCC could be modified to allow the user to specify the reference-point origin based on the user's knowledge of the variable space: the better the knowledge, the smaller the classes.

The principal cluster analysis method can be used as an option in extending the K-Means method for a wider class of situations when the user can fix a few or none tentative centers even if she/he does not know the total number of the clusters or the total number is larger than the number of tentative centers the user is able to specify (Mirkin (1996)).

3.2 How K-Means parameters should be chosen

The user of the K-Means method faces, usually, problems in choosing the following five important kinds of parameter associated with the method: 1) preliminary transformation of the raw data X into matrix Y to be processed; 2) entity-to-center distance $d(x, c)$; 3) centroid concept; 4) number of clusters; 5) initial centers. Traditionally, these parameters are considered

Criterion	Data Scatter	Metric	Centroid	Scale Parameter	Shift Parameter
Least Squares	Square	Euclidean	Average	Standard Deviation	Average
Least Moduli	Absolute Value	City-Block	Median	Absolute Deviation	Median
Least Maximum	Maximal Range	Chebyshev	Midrange	Half-range	Midrange

Table 1: Correspondence between clustering parameters due to the bilinear model.

as completely independent except for the obvious equality of the numbers of clusters and centers. The bilinear clustering model suggests that there is no independence anymore: the parameters are associated to the criterion for model fitting. The correspondence is presented in Table 1; The Chebyshev least-maximum criterion, $L_\infty = \max_{i,v} |y_{iv} - \sum_t c_{tv} z_{it}|$, also is included since all these parameters can be derived from it, too.

Table 1 can be used for determining all of the six parameters when the user is able to choose at least one of them. If, for instance, the user prefers medians as the centroids, she/he is restricted, due to the bilinear model, with the least-moduli criterion along with the city-block distance, etc.

4 Bilinear hierarchic clustering

4.1 Linear embedding binary hierarchies

To discuss hierarchic clustering, we consider a binary hierarchy as a set of subsets $S_W = \{S_w : S_w \subseteq I, w \in W\}$ called clusters containing all singletons and I so that the clusters S_w, $w \in W$, are nested and every non-singleton cluster S_w, $w \in W$, is a union of its two children clusters $S_{w1}, S_{w2} \in S_W$.

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* $\phi_w = (\phi_{iw})$ is defined by $\phi_{iw} = a_w$ if $i \in S_{w1}$, $= -b_w$ if $i \in S_{w2}$, and $= 0$ if $i \notin S_w$, where the values a_w and b_w satisfy the two conditions: (1) vector ϕ_w is centered; (2) vector's ϕ_w norm is 1. It is easy to see that

$$a_w = \sqrt{\frac{n_{w2}}{n_{w1} n_w}}, \text{ and } b_w = \sqrt{\frac{n_{w1}}{n_{w2} n_w}} \tag{12}$$

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 follows:

$$Y = \Phi C \tag{13}$$

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

Since $\Phi^T \Phi$ is the identity matrix, multiplying equality in (13) by Φ^T leads to $C = \Phi^T Y$, that is,

$$c_{wk} = \sqrt{\frac{n_{w1} n_{w2}}{n_w}} (y_{w1v} - y_{w2v}) = \sqrt{\frac{n_{w1} n_w}{n_w 2}} (y_{w1v} - y_{wv}), \tag{14}$$

where y_{wv}, y_{w1v} and y_{w2v} are the averages of the variable/category $v \in V$ in S_w, 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 m-dimensional vector of the averages of the variables in a subset S_w, $w \in W$. The equality in (14) implies that both

L_1 and L_2 norms of vector $c_w = (c_{wk})$ can be expressed as

$$\mu_w = \sqrt{\frac{n_{w1}n_{w2}}{n_w}} d(y_{w1}, y_{w2}) \tag{15}$$

where $d(x, y)$ is the city-block/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 in decomposition (13) considered as an analogue of the singular-value decomposition.

Another useful property of decomposition (13) is that $Y^T Y = C^T C$, which is a decomposition of the between-variable covariance (or correlation) coefficients by clusters of the hierarchy S_W.

Thus, in the case when all the cluster hierarchy is available, there is not much difference between L_1 and L_2 cases, just the choice of distance must be done accordingly while the average serves as the center in both cases.

4.2 Least-squares hierarchy fitting

When the hierarchy is partly unknown so that only upper clusters are given, the exact equality in (13) must be changed for the bilinear model equation, in this case $Y = \Phi C + E$, where the residuals E are to be minimized. It is not difficult to prove that, when some columns of Φ are given (as a part of a basis), the least-squares estimator for corresponding C still satisfies equation (4.1). Moreover, the data scatter is decomposed as follows:

$$\sum_{i \in I, v \in V} y_{iv}^2 = \sum_{t=1}^{m} \mu_t^2 + \sum_{i \in I, v \in V} e_{iv}^2 \tag{16}$$

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

In the framework of the principal component analysis-like sequential fitting strategy, splitting are to be done sequentially, starting with the all set I, each time maximizing corresponding μ_w^2. It is exactly the criterion

$$\mu_w^2 = \frac{n_{w1}n_{w2}}{n_w} d^2(y_{w1}, y_{w2}), \tag{17}$$

suggested by Edwards and Cavalli-Sforza (1965) for divisive clustering, to be maximized by splitting a cluster S_w into S_{w1} and S_{w2}. The step of taking residual data in the principal component analysis-like strategy can be skipped here since it doesn't affect the results, as is not difficult to prove. The standard K-Means method (with two clusters) can be applied as an alternating maximization technique since criterion (17) is equivalent to the least-squares clustering criterion.

The other expression in (14) leads to the same criterion expressed as $\mu_w^2 = N_w N_{w1} d^2(y_{w1}, y_w)/N_{w2}$ which implies a different, SCC-like algorithm, because the center y_w of S_w does not vary in the splitting process.

4.3 Least-moduli hierarchic clustering

The least moduli estimator of C in the bilinear equation $Y = \Phi C + E$ (with E minimized) does not satisfy (14) anymore. Let us discuss the sequential fitting strategy with the criterion, $L_1 = \sum_{i \in I} \sum_{v \in V} |y_{iv} - c_v \phi_i|$, to be minimized by unknown c_v, ϕ_i (index w is omitted, for notational simplicity). Let us denote $c = (c_v)$ and apply definition of ϕ; the criterion becomes:

$$L_1(S_1, S_2, c) = \sum_{i \in S_1} d(y_i, c') + \sum_{i \in S_2} d(y_i, c'') \tag{18}$$

where d is the city-block distance and $c' = (n_2/nn_1)^{1/2}c$, $c'' = -(n_1/nn_2)^{1/2}c$. The latter equations give $n_1 c' = -n_2 c''$, which makes the alternating minimization algorithm for criterion (18) different of the standard K-Means. More precisely, one step, updating of the clusters, is performed exactly as in K-Means: just collecting the entities around centers, c' and c'', to give S_1 and S_2, respectively. The other, updating of the center, is somewhat more subtle and based on an equivalent form of (18):

$$(nn_1n_2)^{1/2} L_1 = \sum_{i \in S_1} n_2 d(n_1 y_i, \bar{c}) + \sum_{i \in S_2} n_1 d(-n_2 y_i, \bar{c}) \tag{19}$$

where $\bar{c} = (n_1 n_2/n)^{1/2} c$. The center updated has c_v equal to $(n/n_1 n_2)^{1/2} \bar{c}_v$ where $\bar{c}_v$ is the weighted median in the set of reals consisting of $n_1 y_{iv}$ (for $i \in S_1$, weight n_2) and $-n_2 y_{iv}$ (for $i \in S_2$, weight n_1), for every $v \in V$. The weighted median for a set of decreasing $x_1, ...x_N$ having $p_1, ..., p_N$ as their respective weights is defined as the value x_a ($a = 1, ..., N$) for which $\sum_{i<a} p_i = \sum_{i>a} p_i$; or, when this equality cannot be achieved, it is an intermediate between those x_a and x_{a+1} for which the best approximation of the weight equality is reached.

Regretfully, the L_1 criterion is irrelevant with regard to nominal data.

Statement 6 *The value of L_1 criterion (18) does not depend on the module-scatter standardized nominal variables.*

Proof: Indeed, in a typical case when $p_v \neq 0.5$, the weighted median of the column y_v is zero, which makes the corresponding components in c, c', c'' in (18) be zero thus giving the same contribution in either of terms of the criterion. □

Acknowledgements

The research was supported by the Office of Naval Research under grants number N00014-93-1-0222 and N00014-96-1-0208 to Rutgers University.

The author thanks Willem Heiser for his support of this work and the referee for comments.

References

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

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

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

[4] J.B. MacQueen (1967). Some methods for classification and analysis of multivariate observations. In *Proc. of 5th Berkeley Symposium,* Vol. 2, pp. 281-297.

[5] B.G. Mirkin (1987). Method of principal cluster analysis. *Automation and Remote Control* **48** 1379-1388.

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

[7] B. Mirkin (1996). *Mathematical Classification and Clustering.* Boston-Dordrecht: Kluwer Academic Press.

[8] B. Mirkin (1997). Concept learning and feature selecting based on square-error clustering. *Machine Learning.* To appear.

[9] P.H.A. Sneath and R.R. Sokal (1973). *Numerical Taxonomy.* San Francisco: W.H. Freeman.

[10] S. Van Buuren and W.J. Heiser (1989). Clustering N objects into K groups under optimal scaling of variables. *Psychometrika* **54** 699-706.