Journal of Classification 4:7-31 (1987)

Journal of
Classification

Additive Clustering and Qualitative Factor Analysis Methods for Similarity Matrices

B. G. Mirkin

Central Economics-Mathematics Institute of USSR's Academy

Abstract: We review methods of qualitative factor analysis (QFA) developed by the author and his collaborators over the last decade and discuss the use of QFA methods for the additive clustering problem. The QFA method includes, first, finding a square Boolean matrix in a fixed set of Boolean matrices with "simple structures" to approximate a given similarity matrix, and, second, repeating this process again and again using residual similarity matrices. We present convergence properties for three versions of the method, provide "cluster" interpretations for results obtained from the algorithms, and give formulas for the evaluation of "factor shares" of the initial similarities variance.

Keywords: Additive Clustering; Qualitative factor analysis; Macrostructures; Average compactness.

Introduction

A model and corresponding algorithms for the additive clustering of similarity matrices (ADCLUS) have been described by Shepard and Arabie (1979), Arabie and Carroll (1980), and Arabie, Carroll, DeSarbo, and Wind (1981). A three way generalization of this model (INDCLUS) and a corresponding method have also been described by Carroll and Arabie (1983). The additive clustering model assumes an approximate representation of the initial similarity matrix in the form of a weighted sum of Boolean (0, 1) matrices with simple structures that correspond to clusters. A similar idea motivates the methods of qualitative factor analysis elaborated by the

I am indebted to Professor P. Arabie and the referees for valuable comments and editing of the text.

Author's Address: Boris G. Mirkin, Central Economics-Mathematics Institute, Krasikova str. 32, Moscow 117418, U.S.S.R.

author and his collaborators (Mirkin 1976; Kupershtokh, Mirkin, and Trofimov 1976a; Trofimov 1976; Mirkin, Kupershtokh, and Trofimov 1979; Mirkin 1980; Trofimov 1981a, 1981b). The purpose of the present article is to review qualitative factor analysis (QFA) methods and to discuss their possible use in additive clustering.

The paper consists of eight parts. The first deals with the additive clustering model itself. Parts 2-4 contain an account of three versions of the qualitative factor analysis method. Part 5 investigates the main properties of these algorithms, namely convergence and interpretation of factor variances. Part 6 deals with the description of the algorithms for additive clustering in the framework of QFA methods and a discussion of properties of the clusters fitted. The results of the QFA algorithms applied to data from Arabie and Shepard (1979) and Breiger, Boorman, and Arabie (1975) are described in Part 7, and then the main distinctions of our approach are briefly summarized.

1. Additive Clustering Model

Using notation from Mirkin (1976, 1980), consider a finite set of N units, denoted by the indices $i, j = 1, \ldots, N$. The basic data are a symmetric nonnegative similarity (proximity, correlation) matrix $\mathbf{A} = \{a_{ij}\}$ $(i, j = 1, \ldots, N)$. Similarities a_{ii} of i with itself are not considered, so all further constructions will use $N(N-1)$-dimensional vectors presented as square $N \times N$ matrices with the main diagonal excluded. In addition, an N-dimensional Boolean vector $\mathbf{z}^t = \{z_i^t\}$ is associated with each subset of units, S_t $(t = 1, \ldots, m)$, so that $z_i^t = 1$ if $i \in S_t$ and $z_i^t = 0$, otherwise. Each subset or cluster S_t is characterized by a Boolean $N \times N$ matrix $\mathbf{R}^t = \mathbf{z}^t (\mathbf{z}^t)' = \{r_{ij}^t\}$, where $r_{ij}^t = z_i^t z_j^t$, so that $r_{ij}^t = 1$ iff $i, j \in S_t$. The matrix $\mathbf{R}^t$ is an indicator of the binary relation $S_t \times S_t$.

If λ_t denotes a nonnegative real number expressing the weight of a subset S_t, the total association, b_{ij}, between units i and j, defined by the set of clusters $S_1, \ldots, S_m$ with given weights $\lambda_1, \ldots, \lambda_m$ is equal to

$$b_{ij} = \sum_t \lambda_t z_i^t z_j^t = \sum_t \lambda_t r_{ij}^t \tag{1}$$

Thus, the association between two units is equal to the sum of the weights of those subsets, S_t, that contain both units simultaneously. In this sense, the subsets S_t are referred to as additive clusters.

The additive clustering model requires specifying both the subsets S_t and the corresponding weights λ_t minimizing the discrepancy between the similarities a_{ij} and the (predicted) associations b_{ij} measured by the sum of squared differences

$$\Delta = \sum_{i,j} (a_{ij} - b_{ij})^2 = \sum_{i,j} (a_{ij} - \sum_{t} \lambda_t z_i^t z_j^t)^2 \tag{2}$$

The additive clustering problem includes the following requirement (Shepard and Arabie 1979): one of the subsets S_t must consist of all N units or, equivalently, the divergence of a_{ij} and b_{ij} may be defined by the following formula

$$\tilde{\Delta} = \sum_{i,j} (a_{ij} - \sum_{t} \lambda_t z_i^t z_j^t - \mu)^2 \tag{3}$$

which has to be minimized by the choice of λ_t, μ, and S_t.

The problem thus involves parameters that are continuous (the weights λ_t and μ) and discrete parameters (the subsets S_t). For a given set of subsets S_t, optimal λ_t are found by ordinary least-squares regression. Minimization of (3) by the choice of S_t is a combinatorially explosive problem of discrete programming. Discrete (but suboptimal) methods based on graph-theoretic concepts are used by Shepard and Arabie (1979) to fit the general ADCLUS model defined in equation (1).

Arabie and Carroll (1980) employ a gradient-based method of optimization utilizing a penalty function approach which asymptotically produces z_i^t's which are 0 or 1 (but which during intermediate stages take on general real values). Their MAPCLUS algorithm uses this penalty function approach for fitting a single $\mathbf{z}^t$ as a component of an overall alternating least squares algorithm for fitting the ADCLUS model. (A final combinatorial optimization stage is used to further improve these estimates.) The Carroll and Arabie (1983) INDCLUS algorithm generalizes this approach to the three-way case.

Our approach to fitting the ADCLUS model must confront four main problems:

(i) determination of the number of clusters, m,

(ii) determination of an initial set of subsets, S_t,

(iii) avoidance of local minima in optimizing the objective function in (3),

(iv) ensuring positivity for the weights, λ_t.

2. Qualitative Factor Analysis: QFA-0

The qualitative factor analysis method for proximity matrices, proposed by Mirkin (1976), can be characterized as follows.

Fix a set $\bar{E}$ of $N \times N$ Boolean matrices with simple structures that are admissible as approximate descriptions of the given proximity matrix $\mathbf{A} = \{a_{ij}\}$. Nearly ten such sets for $\bar{E}$ have been used in the literature,

including the set M of all macrostructures (partitions with structure), the set P of all partitions, the set Q of all ordered partitions (rankings), the set $\overset{v}{S}$ of all subsets, and so on (Mirkin 1974, 1976; Kupershtokh, Mirkin, and Trofimov 1976a; Trofimov 1976; Mirkin, Kupershtokh, and Trofimov 1979; Mirkin 1980; Trofimov 1981a, 1981b). To define these sets $\overline{E}$ precisely, consider the concept of macrostructure (R, K) (Mirkin 1974), where $R = \{R_1, \ldots, R_p\}$ is a partition of the set of units with an arbitrary number p of classes R_s, and $K \subseteq \{1, \ldots, p\}^2$ is an association graph on the set of classes R_s; in English this notion in sociometry is sometimes referred to as a blockmodel (Arabie, Boorman, and Levitt 1978). The Boolean matrix $\mathbf{R} = \{r_{ij}\}$ corresponding to (R, K) is defined by the following condition: $r_{ij} = 1$ iff $i \in R_s$ and $j \in R_t$ for $(s, t) \in K$. It is clear that the set M of such matrices coincides with the set of all Boolean $N \times N$ matrices.

The structure (R, K) corresponds to the partition R itself for $K = \{(s, s): s = 1, \ldots, p\}$ (nominal attributes); here, $\mathbf{R} = \{r_{ij}\}$ is an indicator of the corresponding equivalence relation. For $K = \{(s, t): s \leqslant t\}$, (R, K) determines a non-strict ranking (ordinal variable) and the matrix $\mathbf{R}$ is an indicator of the corresponding linear quasi-order relation. The structure (R, K) corresponds to the subset R_s if $K = \{(s, s)\}$ for a fixed s; then $\mathbf{R}$ becomes an indicator of the binary relation $R_s \times R_s$ and coincides with the matrix of the set R_s that was defined above. Later on we shall deal mainly with the set $\overset{v}{S}$ of matrices defining fixed subsets.

As concrete examples, consider a set of seven units with the partition $R = \{R_1, R_2, R_3\}$, where $R_1 = \{1, 2\}$, $R_2 = \{3, 4, 5\}$, $R_3 = \{6, 7\}$. For $K_1 = \{(1, 1), (2, 2), (3, 3)\}$, the pair (R, K_1) corresponds to a nonordered (nominal) partition with matrix $\mathbf{R}^1 \in P$; for $K_2 = \{(1, 1), (2, 2), (3, 3), (1, 2), (1, 3), (2, 3)\}$, the pair (R, K_2) corresponds to an ordered partition with matrix $\mathbf{R}^2 \in Q$; for $K_3 = \{(2, 2)\}$, the pair (R, K_3) corresponds to a subset with matrix $\mathbf{R}^3 \in \overset{v}{S}$; and for $K_4 = \{(1, 1), (1, 2), (2, 3), (3, 1)\}$, the pair (R, K_4) corresponds to a macrostructure with matrix $\mathbf{R}^4 \in M$. Explicitly,

$$\mathbf{R}^1 = \begin{pmatrix} * & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & * & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & * & 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & * & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 & * & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & * & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 & * \end{pmatrix}, \qquad \mathbf{R}^2 = \begin{pmatrix} * & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & * & 1 & 1 & 1 & 1 & 1 \\ 0 & 0 & * & 1 & 1 & 1 & 1 \\ 0 & 0 & 1 & * & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 & * & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & * & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 & * \end{pmatrix}$$

$$\mathbf{R}^3 = \begin{pmatrix} * & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & * & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & * & 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & * & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 & * & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & * & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & * \end{pmatrix}, \qquad \mathbf{R}^4 = \begin{pmatrix} * & 1 & 1 & 1 & 1 & 0 & 0 \\ 1 & * & 1 & 1 & 1 & 0 & 0 \\ 0 & 0 & * & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & * & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & * & 1 & 1 \\ 1 & 1 & 0 & 0 & 0 & * & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & * \end{pmatrix}.$$

If the researcher's goal is to reveal the macrostructure of all the important associations in **A**, one must choose $\bar{E} = M$. Similarly, when a partitioning of the set of units is of interest, $\bar{E} = P$, and for ordered partitions, $\bar{E} = Q$. Finally, when $\bar{E} = \overset{\vee}{S}$ the problem of additive clustering is identified. Concrete examples are given, for example, in Mirkin (1976), Mirkin and Rodin (1977), and Trofimov (1981b).

Qualitative factor analysis methods can be characterized by the iterative reconstruction of the input matrix **A** by a subset of admissible matrices $\mathbf{R} \in \bar{E}$, and in some ways, are analogous to the principal axes method. Each step of the method consists of minimizing the discrepancy

$$\Delta(\mathbf{R},\lambda) = \sum_{i,j} (a_{ij} - \lambda\, r_{ij})^2 \tag{4}$$

over all possible real λ and $\mathbf{R} \in \bar{E}$, using, for example, alternating "coordinate" minimization of (4) by choosing λ and **R** in turn through an iterative process. For a given **R**, an optimal λ may be derived in closed form if the derivative of $\Delta(\mathbf{R},\lambda)$ with respect to λ, $\sum_{i,j} (a_{ij} - \lambda\, r_{ij})\, r_{ij}$, is set equal to 0. Taking into account that $r_{ij}\, r_{ij} = r_{ij}$,

$$\lambda = \sum_{i,j} a_{ij}\, r_{ij} \Big/ \sum_{i,j} r_{ij} \ . \tag{5}$$

Thus, the optimal value of λ is the mean of the proximities a_{ij} that are internal to (R, K), i.e., those corresponding to a unitary value of r_{ij}. In particular, when **R** corresponds to a subset S, the value of λ is equal to the mean of a_{ij} for all $i, j \in S$.

For a fixed value of λ, the minimization of (4) through the choice of $\mathbf{R} \in \bar{E}$ is reduced to the problem of finding extreme values for

$$f(\mathbf{R},\pi) = \sum_{i,j}^{N} (a_{ij} - \pi)\, r_{ij} \ . \tag{6}$$

Specifically, $\Delta(\mathbf{R},\lambda) = \sum_{i,j} a_{ij}^2 - 2\lambda \sum_{i,j} (a_{ij} - \lambda/2)\, r_{ij}$, where $\pi = \lambda/2$ plays the role of a proximity value threshold. For $\lambda > 0$, it is advantageous to define $r_{ij} = 1$ and include the pair (i, j) in (R, K) when $a_{ij} - \pi > 0$ and (6) is to be maximized; $r_{ij} = 0$ is appropriate when $a_{ij} - \pi < 0$. Unlike the usual application of the threshold in clustering (eliminations of all proximities that are less than π), the value of π is used in a "soft" manner here. We shall see this explicitly in Part 6 when methods for solving (6) for $\bar{E} = \overset{\vee}{S}$ are discussed (see also Mirkin 1980, Ch. 5, for other concrete examples of $\bar{E}$).

When $\lambda < 0$, we minimize (6), including in (R, K) the pairs (i, j) with $a_{ij} < \pi$ and excluding those with $a_{ij} > \pi$. However, the same algorithm used for the maximization of (6) can be applied because the minimum of $f(\mathbf{R}, \pi)$ corresponds to the maximum of $f(\mathbf{R}, -\pi)$ for $-\mathbf{A} = \{-a_{ij}\}$. In general, the final choice between those $\mathbf{R}$'s obtained for $\lambda > 0$ and for $\lambda < 0$ may be made according to the value of (4). We shall consider here only the solutions for $\mathbf{R} \in \bar{E}$ when $\lambda > 0$.

An iterative algorithm may be adopted for the minimization of (4). Beginning with some fixed value of the threshold π (usually the mean of a_{ij}), $f(\mathbf{R}, \pi)$ in (6) is minimized by the choice of $\mathbf{R} \in \bar{E}$ using a (typically) suboptimal algorithm. Then, λ is recalculated according to (5) and π is set equal to $\lambda/2$ and used for the next step. The computations are finished when λ or, equivalently $\mathbf{R}$, is stabilized. The process must converge in a finite number of iterations because the value of (4) decreases each time and the set $\bar{E}$ consists of a finite number of elements.

We note that minimizing (4) is equivalent to maximizing

$$g(\mathbf{R}) = \frac{\left(\sum_{i,j} a_{ij}\, r_{ij}\right)^2}{\sum_{i,j} r_{ij}} \tag{7}$$

which depends only on $\mathbf{R}$, although not linearly. To prove this equivalence, the optimal value of λ is substituted from (5) into (4). It would be possible to search for an optimal $\mathbf{R} \in \bar{E}$ using direct optimization of (7) (see Mirkin 1985, pp. 210-211).

A possibly suboptimal solution to (4), denoted by $(\mathbf{R}^1, \lambda_1)$, forms the first qualitative factor obtained. A residual matrix

$$\mathbf{A}^1 = \mathbf{A} - \lambda_1 \mathbf{R}^1$$

is calculated, and again, $\Delta(\mathbf{R}, \lambda)$ is minimized using the residuals $\mathbf{A}^1$ instead

of **A**. Continuing to the $(m+1)$-st iteration, the qualitative factors $(\mathbf{R}^1, \lambda_1), \ldots, (\mathbf{R}^m, \lambda_m)$ found sequentially on preceding steps are used to compute the residual proximity matrix

$$\mathbf{A}^m = \mathbf{A} - \sum_t \lambda_t \mathbf{R}^t = \mathbf{A}^{m-1} - \lambda_m \mathbf{R}^m \tag{8}$$

and $\Delta(\mathbf{R}, \lambda)$ is minimized for $\mathbf{A} = \mathbf{A}^m$. The resulting solution forms the $(m+1)$-st qualitative factor $(\mathbf{R}^{m+1}, \lambda_{m+1})$. Conceivably the process has an infinite number of steps, but it empirically converges in finite time (see Part 5).

Application of the algorithm to matrices with positive entries usually leads to results that are difficult to interpret. W. L. Kupershtokh and W. A. Trofimov (private communication) conjectured that the difficulties arose because of "noise effects" of other factors on subsequent steps of the algorithm, and proposed a modification of the process considered in the next section. In the present author's opinion (which is confirmed empirically), the "noise background" may be eliminated by beginning the iterative process not with **A**, but rather with the matrix $\{a_{ij} - a\}$, where a is equal to the mean of all initial similarities a_{ij}. This matrix is a result of subtracting the "initial" factor $\lambda_0 \mathbf{R}^0$ from **A**, where $r^0_{ij} = 1$ for all $i, j = 1, \ldots, N$ (that is, $\mathbf{R}^0$ corresponds the subset S^0 consisting of all units), and $\lambda_0 = a$ by (5). The additive constant μ in (3) is equal to $\lambda_0 = a$ (supposing that $S^t \neq S^0$ for all $t > 0$). We shall refer to the exhaustion method with this initial subtraction of the mean from all proximities as method QFA-0.

3. Qualitative Factor Analysis with an Additive Constant: QFA-1

This second method eliminates the "noise" on each iterative step of the process with the help of a matrix $\lambda \mathbf{R} + \mu$, where μ is a real number added for simulation of the noise background to each element of $\lambda \mathbf{R}$. Thus,

$$\Delta(\mathbf{R}, \lambda, \mu) = \sum_{i,j} (a_{ij} - \lambda r_{ij} - \mu)^2 \tag{9}$$

is minimized on each step. A theoretical and experimental investigation of this modified method was carried out by Trofimov (1976, 1981a,b; see also Kupershtokh, Mirkin, and Trofimov 1976a).

It is not difficult to prove that optimal values of λ and μ for fixed **R** are defined by formulae:

$$\mu = \sum_{i,j} a_{ij} (1 - r_{ij}) / \sum_{i,j} (1 - r_{ij}) \tag{10}$$

$$\lambda + \mu = \sum_{i,j} a_{ij} r_{ij} / \sum_{i,j} r_{ij} \tag{11}$$

Thus, μ is an average "external" proximity for $\mathbf{R}$, i.e., the mean of a_{ij} for the pairs (i, j) not present in (R, K) (for which $r_{ij} = 0$), and λ is a difference between the "internal" average proximity (11) and "external" average proximity (10). Thus, λ characterizes a contrast between these two kinds of proximities.

An optimal $\mathbf{R} \in \bar{E}$ for given λ and μ may be found by optimizing the criterion $f(\mathbf{R}, \pi)$ in (6) for $\pi = \frac{\lambda}{2} + \mu$. It is necessary to maximize $f(\mathbf{R}, \pi)$ for $\lambda > 0$ and to minimize it for $\lambda < 0$. Thus, the same iterative algorithm from the preceding section can be used. Beginning from some threshold π (e.g., the mean of a_{ij}, corresponding to a trivial one-element subset or the trivial partition with N one-element classes according to (10) and (11)), the criterion in (6) is maximized ($\lambda > 0$ is assumed). For the given $\mathbf{R} \in \bar{E}$, λ and μ are computed from (10) and (11) and combined as $\pi = \frac{\lambda}{2} + \mu$ for use in the next step of the iteration.

Making a substitution into (9) of λ and μ taken from (10) and (11), $\Delta(\mathbf{R}, \lambda, \mu)$ can be represented as $c_1 - c_2 f(\mathbf{R})$, where $c_1, c_2 > 0$ and

$$f(\mathbf{R}) = \frac{\left(\sum_{i,j} a_{ij}(1 - r_{ij})\right)^2}{\sum_{i,j} (1 - r_{ij})} + \frac{\left(\sum_{i,j} a_{ij} r_{ij}\right)^2}{\sum_{i,j} r_{ij}} \tag{12}$$

This expression depends only on $\mathbf{R}$. It is possible to optimize this squared criterion directly without recalculation of λ and μ, but we do not consider this possibility here.

In short, the general step of the QFA-1 method is as follows. Using a residual matrix $\mathbf{A}^{m-1}$ from the preceding step, minimize (9) for $\mathbf{A}^{m-1}$ and use the obtained solution $\lambda_m \mathbf{R}^m + \mu_m$ in the next step based on the residual matrix

$$\mathbf{A}^m = \mathbf{A}^{m-1} - \lambda_m \mathbf{R}^m - \mu_m$$

In typical situations the QFA-1 method tends to the equality $\mathbf{A} = \sum_t \lambda_t \mathbf{R}^t + \sum_t \mu_t$ (see Part 5).

The additive constant μ can be considered more than a technical detail and is an essential part of the quantitative representation of qualitative information (Mirkin, Kupershtokh, and Trofimov 1979; Mirkin 1980). It is somewhat arbitrary to evaluate the inclusion of a pair (i, j) in (R, K) by the arbitrary weight λ, while the opposite is always evaluated by zero. Suppose qualitative information contained in a macrostructure (R, K) is represented by a Boolean matrix $\mathbf{S} = \{s_{ij}\}$ with two arbitrary values, α and β, for entries, so that $s_{ij} = \alpha$ for $(i, j) \in (R, K)$ (i.e., $(s, t) \in K$ for $i \in R_s$ and $j \in R_t$) and $s_{ij} = \beta$, otherwise. The value α is a weight of inclusion and β is a weight for noninclusion of a pair (i, j) in (R, K). The matrix $\lambda \mathbf{R} + \mu$ has exactly the form of an (α, β)-matrix $\mathbf{S}$ for $\alpha = \lambda + \mu, \beta = \mu$. According to formulae (10) and (11), an optimal β is a mean of external proximities and an optimal α is a mean of internal proximities for the macrostructure (R, K) corresponding to the matrix $\mathbf{R}$. The value π in (6) is $\pi = (\alpha + \beta)/2$.

Thus, the QFA-1 method decomposes the matrix $\mathbf{A}$ into the sum

$$\mathbf{A} = \sum_t \mathbf{S}^t \ , \tag{13}$$

where $\mathbf{S}^t = \lambda_t \mathbf{R}^t + \mu_t$. This representation does not use any external quantitative parameters such as λ_t and μ in (2) and (3). The values λ and μ are the weights of inclusion or noninclusion, respectively, for an arbitrary pair of units in (R, K) but are not indications of factor importance. The signs of λ and μ influence the interpretation of the factor but not its salience (see Part 5).

4. Recalculation of the Weights in Method QFA

Trofimov (1981a) proposed another modification of the QFA method typically leading to a finite series of factors. This modification (the QFA-2 method) is based on the recalculation of the weights for all previously obtained factors at each step. The method has been formulated using centered and normed matrices S (Trofimov 1981). Here, formula (13) based on (α, β)-representation of the factors is used.

The factors $\mathbf{S}^1 = \lambda_1 r^1 + \mu_1, \ldots, \mathbf{S}^m = \lambda_m r^m + \mu_m$, obtained during the first m steps, are used in a least squares minimization. A vector $\nu = (\nu_1, \ldots, \nu_m)$ minimizing the sum of squared differences

$$\sum_{i,j} (a_{ij} - \nu_1 s_{ij}^1 - \cdots - \nu_m s_{ij}^m)^2 \tag{14}$$

is determined by the usual least squares formula

$$\nu = (\mathbf{S}'\mathbf{S})^{-1}\mathbf{S}'\mathbf{A} \tag{15}$$

where $\mathbf{S}$ is the $N(N-1) \times m$ matrix that consists of $N(N-1)$-dimensional vectors $\mathbf{S}^t$ as its columns; $\mathbf{A}$ is also considered an $N(N-1)$-dimensional vector; $\mathbf{S}'\mathbf{S}$ is a matrix of scalar products $(\mathbf{S}^t, \mathbf{S}^q) = \sum_{i,j} s_{ij}^t s_{ij}^q$ $(t, q = 1, \ldots, m)$. According to assumptions of Part 5 (see Theorem 3), the matrix $\mathbf{S}'\mathbf{S}$ has rank m and its inverse always exists.

Though the weights of the factors are now changed ($\bar{\mathbf{S}}^t = \nu_t \mathbf{S}^t$ instead of $\mathbf{S}^t = \lambda_t r^t + \mu_t$), the ratios, α_t / β_t, are invariant. Moreover, the linear space in which the matrix $\mathbf{A}$ is projected does not change, so the matrices $\mathbf{S}^t$ of the first m factors are unchanged. A residual matrix for the next step is determined as follows:

$$\mathbf{A}^m = \mathbf{A} - \sum_{t=1}^{m} \bar{\mathbf{S}}^t = \mathbf{A} - \sum_{t=1}^{m} \nu_t \mathbf{S}^t \tag{16}$$

The $(m+1)$-st factor $\mathbf{S}^{m+1} = \lambda_{m+1} \mathbf{R}^{m+1} + \mu_{m+1}$ is found by minimizing (9) for $\mathbf{A} = \mathbf{A}^m$, defined by (16). The process is then repeated.

5. Properties of QFA Methods

Convergence of the QFA method, summarized by Theorems 1 and 2, is ensured by the diminishing of residual proximities to zero as the number of factors increases. For the interpretation of a solution, the problem of assessing factor salience is most important (see Theorem 3).

Let $\mathbf{R}^{ij}$ be an $N \times N$ matrix with all zero entries except for $r_{ij} = 1$ (for $\bar{E}$ consisting only of symmetric matrices, r_{ij} and r_{ji} must both be nonzero). Require for each i, j

$$\mathbf{R}^{ij} \in \bar{E} \tag{17}$$

The sets M, P, $\overset{\vee}{S}$ evidently satisfy requirement (7). (Recall that the diagonal entries of matrices are not considered). For example, for $\bar{E} = \overset{\vee}{S}$, matrix $\mathbf{R}^{ij}$ corresponds to the two-element subset $S = \{i, j\}$.

The matrix $\mathbf{A}$ is nonnegative and nonzero as in Part 1. For symmetric $\bar{E}$, matrix $\mathbf{A}$ is assumed to be symmetric. We assume also that the algorithm used for suboptimization of $f(\mathbf{R}, \pi)$ leads to a solution not worse than that given by each of the $\mathbf{R}^{ij}$. Given these assumptions, the following statement of convergence can be made:

Theorem 1 *Methods QFA-0 and QFA-1 have the property that* $\mathbf{A}^m \to \mathbf{0}$ for $m \to \infty$.

Proof We first consider the process (8) of the QFA-0 and then explore the value $\Delta(\mathbf{R}^t, \lambda_t)$ for $\mathbf{A} = \mathbf{A}^{t-1}$. Clearly,

$$\Delta(\mathbf{R}^t, \lambda_t) = (\mathbf{A}^{t-1} - \lambda_t \mathbf{R}^{t-1}, \mathbf{A}^{t-1} - \lambda_t \mathbf{R}^{t-1}) = (\mathbf{A}^t, \mathbf{A}^t),$$

where $(\mathbf{A}, \mathbf{B}) = \sum_{i,j} a_{ij} b_{ij}$. Choose an element $\hat{a}_{ij}$ in matrix $\mathbf{A}^{t-1}$ with maximal absolute value, which may be assumed nonzero because otherwise $\mathbf{A}^{t-1} = 0$. Consider the matrix $\mathbf{R}^{ij}$, belonging to $\bar{E}$ by (17), and let $\lambda = \hat{a}_{ij}$. The fact of suboptimality of $(\mathbf{R}^t, \lambda_t)$ implies

$$(\mathbf{A}^t, \mathbf{A}^t) = \Delta(\mathbf{R}^t, \lambda_t) \leqslant \Delta(\mathbf{R}^{ij}, \lambda) = (\mathbf{A}^{t-1}, \mathbf{A}^{t-1}) - 2(\mathbf{A}^{t-1}, \lambda \mathbf{R}^{ij}) + (\lambda \mathbf{R}^{ij}, \lambda \mathbf{R}^{ij}) \ .$$

Clearly, $(\mathbf{A}^{t-1}, \lambda \mathbf{R}^{ij}) = \hat{a}_{ij}^2 = (\lambda \mathbf{R}^{ij}, \lambda \mathbf{R}^{ij})$, thus

$$(\mathbf{A}^t, \mathbf{A}^t) \leqslant (\mathbf{A}^{t-1}, \mathbf{A}^{t-1}) - \hat{a}_{ij}^2 \ .$$

In addition, the fact of maximality of $|\hat{a}_{ij}|$ in $\mathbf{A}^{t-1}$ implies

$$(\mathbf{A}^{t-1}, \mathbf{A}^{t-1}) / (N(N-1)) \leqslant \hat{a}_{ij}^2 \ .$$

Combining these inequalities, we find that

$$(\mathbf{A}^t, \mathbf{A}^t) \leqslant (\mathbf{A}^{t-1}, \mathbf{A}^{t-1})\, \epsilon$$

where $\epsilon = 1 - 1 / (N(N-1))$, implying

$$(\mathbf{A}^t, \mathbf{A}^t) \leqslant (\mathbf{A}, \mathbf{A})\, \epsilon^t \ ,$$

where $0 < \epsilon < 1$; thus $\epsilon^t \to 0$ for $t \to \infty$, proving convergence of the QFA-0 method.

The QFA-1 method can be treated in analogous manner taking $\lambda \mathbf{R} + \mu$ instead of $\lambda \mathbf{R}$. For $\mathbf{R}^{ij}$ we let $\mu = 0$ and $\lambda = \hat{a}_{ij}$, proving the theorem.

The possibility of an inequality $\lambda < 0$ (for $\hat{a}_{ij} < 0$) is essential in this proof.

The finiteness of the number of iterations may be proved for method QFA-2 (Trofimov 1981).

Theorem 2 *The matrices* $\mathbf{S}^1, \ldots, \mathbf{S}^m, \ldots$ found by the QFA-2 method form a linearly independent set of $N(N-1)$-dimensional linear space elements, and so it is finite.

Proof We use an induction in the number m of factors. For $m = 1$ the statement is trivial. Let the system $\mathbf{S}^1, \ldots, \mathbf{S}^m$ be linearly independent. We have to prove now that equality $\mathbf{S}^{m+1} = \sum_{t=1}^{m} \alpha_t \mathbf{S}^t$ is impossible for any set of nonzero α_t.

According to properties of the orthogonal projection operation, the residual matrix $\mathbf{A}^m$ (16) is orthogonal to every $\mathbf{S}^t$ $(t = 1, \ldots, m)$. Thus, $(\mathbf{A}^m, \mathbf{S}^1) = \cdots = (\mathbf{A}^m, \mathbf{S}^m) = 0$. If $\mathbf{A}^m = 0$, then the process is already complete, with the matrices $\mathbf{S}^1, \ldots, \mathbf{S}^m$ being linearly independent while completely accounting for $\mathbf{A}$. So, consider a case $\mathbf{A}^m \neq 0$. If we choose a maximal (by absolute value) entry $\hat{a}_{ij}$ in $\mathbf{A}^m$ and set the matrix $\mathbf{S} = \lambda \mathbf{R}^{ij} + \mu$, where $\lambda = \hat{a}_{ij}$ and $\mu = 0$, then the value Δ (9) for $\mathbf{A} = \mathbf{A}^m$ is equal to

$$(\mathbf{A}^m - \mathbf{S}, \mathbf{A}^m - \mathbf{S}) = (\mathbf{A}^m, \mathbf{A}^m) - 2(\mathbf{A}^m, \mathbf{S}) + (\mathbf{S}, \mathbf{S}) = (\mathbf{A}^m, \mathbf{A}^m) - \hat{a}_{ij}^2 ,$$

because $(\mathbf{A}^m, \mathbf{S}) = (\mathbf{S}, \mathbf{S}) = \hat{a}_{ij}^2$.

Therefore, the $(m+1)$-st step's optimal value Δ^* of the $\Delta(\mathbf{R}, \lambda, \mu)$ for $\mathbf{A} = \mathbf{A}^m$ satisfies the inequality

$$\Delta^* = (\mathbf{A}^m - \mathbf{S}^{m+1}, \mathbf{A}^m - \mathbf{S}^{m+1}) \leqslant \Delta(\mathbf{R}^{ij}, \lambda, \mu) < (\mathbf{A}^m, \mathbf{A}^m) .$$

Assuming $\mathbf{S}^{m+1} = \sum_t \alpha_t \mathbf{S}^t$ we find $(\mathbf{A}^m, \mathbf{S}^{m+1}) = \sum_t \alpha_t (\mathbf{A}^m, \mathbf{S}^t) = 0$. In this case

$$\begin{aligned} \Delta^* &= (\mathbf{A}^m - \mathbf{S}^{m+1}, \mathbf{A}^m - \mathbf{S}^{m+1}) \\ &= (\mathbf{A}^m, \mathbf{A}^m) - 2(\mathbf{A}^m, \mathbf{S}^{m+1}) + (\mathbf{S}^{m+1}, \mathbf{S}^{m+1}) > (\mathbf{A}^m, \mathbf{A}^m) , \end{aligned}$$

which contradicts the preceding inequality $\Delta^* < (\mathbf{A}^m, \mathbf{A}^m)$. Q.E.D.

Since the maximal number of linearly independent vectors equals the dimensionality of the space, Theorem 2 ensures that method QFA-2 finds an exact representation of a matrix $\mathbf{A}$ as the sum of a finite number of

factors, always, independently of the fact that the methods of resolving problem (9) may be approximate or heuristic.

The problem of salience of factors is approached using the optimality properties of the weights λ and μ. Substituting (5) for (4) and performing the necessary multiplications, one can find for every $\mathbf{R} \in \bar{E}$

$$\Delta(\mathbf{R}, \lambda) = (\mathbf{A} - \lambda\mathbf{R}, \mathbf{A} - \lambda\mathbf{R}) = (\mathbf{A}, \mathbf{A}) - \lambda^2(\mathbf{R}, \mathbf{R}) \tag{18}$$

This result implies that for each factor $\mathbf{R}^t$ found using method QFA-0 the following equality holds:

$$(\mathbf{A}^t, \mathbf{A}^t) = (\mathbf{A}^{t-1}, \mathbf{A}^{t-1}) - \lambda_t^2 (\mathbf{R}^t, \mathbf{R}^t) \tag{19}$$

Summing (19) over t we find

$$(\mathbf{A}, \mathbf{A}) = \sum_t \lambda_t^2 (\mathbf{R}^t, \mathbf{R}^t) \tag{20}$$

which is a decomposition of the sum of the squares of initial proximities a_{ij} into a sum of terms $\lambda_t^2 (\mathbf{R}^t, \mathbf{R}^t)$ corresponding to separate factors and characterizing their degree of salience.

In case the mean of proximities is first subtracted from all entries in $\mathbf{A}$ (i.e., the universal factor u is subtracted) we obtain the following by dividing (20) by $N(N-1)$:

$$\sigma^2 (\mathbf{A}) = \sum_t \lambda_t^2 \epsilon (\mathbf{R}^t) , \tag{21}$$

where $\epsilon (\mathbf{R})$ is a proportion of ones in $\mathbf{R}$. In particular, $\epsilon (\mathbf{R}) = \sum_s p_s^2$ with negligible error in case $\mathbf{R}$ corresponds to some partition $R = \{R_1, \ldots, R_q\}$ with distribution $p_1, \ldots, p_q$. This quantity is a mean of correct predictions (probability) in a model of proportional prediction of the classes R_s (with probabilities p_s) (see Mirkin 1980). We have $\epsilon (\mathbf{R}) = p^2$ for $\mathbf{R}$ corresponding to a subset S containing pN units.

Thus, formula (21) shows that the share of the t-th factor in the variance accounted for in the initial proximities is equal to

$$v_t = \lambda_t^2 \epsilon (\mathbf{R}^t) \tag{22}$$

This term increases as the number of ones in $\mathbf{R}^t$ and/or the mean of proximities a_{ij} taken into account in $\mathbf{R}^t$ increases.

According to formula (21), the salience of a factor is characterized by the value v_t, and not by the value of λ_t alone. Using the factor variance v_t we can make decisions on selection of the most informative factors and, moreover, on a sufficient number of factors: we can stop the iterative process when the total factor variance $\sum_{t=1}^{m} v_t$ exceeds some fixed share of the total variance $\sigma^2(\mathbf{A})$.

In general, the values λ_t are not monotonic in t. But for positive $\mathbf{A}$ the initial steps of method QFA-0 will lead to the monotonicity of the λ_t because all $\lambda_t > 0$. It follows from (18) that $\lambda_t < \lambda_{t+1}$ implies that we could obtain a better result in the t-th step by substituting $\lambda_{t+1}\mathbf{R}_{t+1}$ for $\lambda_t \mathbf{R}^t$. Thus, the nonmonotonic behavior of λ_t's implies that during initial steps the algorithm used for optimizing $f(\mathbf{R}, \pi)$ gives a local optimum.

Method QFA-1 possesses an analogous characteristic of factor variance. For given $\mathbf{R}$, the problem of minimizing (9) is equivalent to linear regression of $\mathbf{A}$ onto $\mathbf{R}$. As is well known in this case, the minimal value of $\Delta(\mathbf{R}, \lambda, \mu)$ can be represented as follows:

$$\Delta^*/(N(N-1)) = \sigma^2(\mathbf{A})(1-\rho^2),$$

where ρ is a correlation coefficient of $N(N-1)$-dimensional vectors $\mathbf{A}$ and $\mathbf{R}$, which is linked to the optimal λ by the formula

$$\rho = \lambda \frac{\sigma(\mathbf{R})}{\sigma(\mathbf{A})}$$

Thus, we have the following expression associating the minimal value of Δ with the optimal value of λ:

$$\Delta^*/(N(N-1)) = \sigma^2(\mathbf{A}) - \lambda^2\sigma^2(\mathbf{R}).$$

For $\mathbf{R} = \mathbf{R}^1$, the left part of the equality is equal to the variance of residuals computed from the proximity data, $\mathbf{A}^1 = \mathbf{A} - \lambda_1 \mathbf{R}^1 - \mu_1$; thus

$$\sigma^2(\mathbf{A}) = \sigma^2(\mathbf{A}^1) + \lambda_1^2\sigma^2(\mathbf{R}^1).$$

Continuing the reasoning we obtain

$$\sigma^2(\mathbf{A}) = \sigma^2(\mathbf{A}^{m+1}) + \sum_{t=1}^{m} \lambda_t^2 \sigma^2(\mathbf{R}^t) \tag{23}$$

The formula shows that the factor variance accounted for by method QFA-1 is equal to

$$w_t = \lambda_t^2 \sigma^2(\mathbf{R}^t) . \tag{24}$$

It is not difficult to demonstrate that

$$\sigma^2(\mathbf{R}^t) = \epsilon(\mathbf{R}^t)(1 - \epsilon(\mathbf{R}^t)) . \tag{25}$$

The quantity $1 - \epsilon(\mathbf{R})$ is refered to as the qualitative variance of the factor by Mirkin (1976, 1980). For example, it is equal to $\sum_s p_s(1 - p_s)$ for a partition with distribution (p_s). It is clear that $\epsilon(1 - \epsilon)$ increases when ϵ tends to 1/2. In particular, for fixed λ the factor variance of a subset S grows when the number of its elements tends to $N/\sqrt{2}$.

Method QFA-2 changes the factor variances step by step, so that the only criterion of goodness-of-fit for the method is a residual proximities (16) variance.

We have proved the following assertion.

Theorem 3 *For optimal weights of the factors, the share of the factor in the general variance of initial proximities (factor variance) in either method QFA-0 or QFA-1 is determined by the value* v_t *(22) for QFA-0 and by* w_t *(24) for QFA-1.*

6. Additive Clustering by QFA Methods

Consider the problem of maximizing the criterion in (6) when applied to the additive clustering problem, i.e., for $\bar{E} = \overset{v}{S}$ (for cases $\bar{E} = P, Q, M$, see Mirkin 1976, 1980).

For a subset S corresponding to $\mathbf{R} \in \overset{v}{S}$, the criterion (6) may be rewritten as

$$f(S, \pi) = \sum_{i,j \in S} (a_{ij} - \pi) = \sum_{i,j \in S} a_{ij} - \pi\, n(n-1) \tag{26}$$

where n is the number of units in subset S.

We have already described an interpretation of π as a threshold for gauging salience: it is advantageous to place units i and j together in S for $a_{ij} > \pi$ and disadvantageous for $a_{ij} < \pi$. The last expression in (26) generates another interpretation of π: it is a compromise coefficient for two conflicting criteria, $f_1 = \sum_{i,j \in S} a_{ij}$ and $f_2 = -n(n-1)$. Thus, for $\mathbf{A} \geqslant 0$ the use of f_1 requires the union of all units in S; in contrast, criterion f_2 requires a smaller number of units in S. For small π the differences $a_{ij} - \pi$ are positive and the criterion in (26) leads to the complete set of units as the optimal one; after increasing π to the point where all $a_{ij} - \pi$ are negative, the criterion in (26) leads to one-element subsets (i.e., clusters consisting of singletons) as optimal.

Does the number of units in an optimal cluster decrease when π increases? The answer is "yes."

Theorem 4 *Let $\pi_1 > \pi_2$ and suppose S_i is a solution to the $f(S, \pi_i)$ maximization problem with n_i elements. Then $n_1 \leqslant n_2$.*

Proof By definition, $f(S, \pi) = f(S, 0) - \pi\, n(n-1)$. And after the definition of S_i, $f(S_1, \pi_1) \geqslant f(S_2, \pi_1)$, and $f(S_2, \pi_2) \geqslant f(S_1, \pi_2)$. Thus,

$$f(S_1, 0) - f(S_2, 0) \geqslant \pi_1 (n_1(n_1-1) - n_2(n_2-1)) ,$$

and

$$f(S_1, 0) - f(S_2, 0) \leqslant \pi_2 (n_1(n_1-1) - n_2(n_2-1)) .$$

Combining these two inequalities, we find

$$(\pi_2 - \pi_1)(n_1(n_1-1) - n_2(n_2-1)) \geqslant 0 .$$

Thus $n_1(n_1-1) \leqslant n_2(n_2-1)$ because $\pi_2 < \pi_1$. Also $n_1 \leqslant n_2$, since the function $g(x) = x(x-1)$ is strictly monotone for $x \geqslant 1$. Q.E.D.

In general, the problem of maximizing (26) has exponential complexity, and so for practical reasons we must limit ourselves to suboptimal solutions only. Traditionally, suboptimal algorithms are based on a neighborhood approach. We begin by considering the set of all clusters obtained from S by adding to or eliminating one unit from S as the neighborhood for the set S (cf. the approach to combinatorial optimization given by Borodkin et al. 1978; Arabie and Carroll 1980, p. 225; Mirkin 1980.) Our suboptimal algorithm begins with an arbitrary S and step by step adds to or eliminates from S that unit giving a maximal increment of the $f(S, \pi)$. Computations end when all increments are nonpositive.

To be more precise, we use notation from Borodkin et al. (1978) and characterize a set S by an N-dimensional $(1, -1)$-vector $\mathbf{z} = (z_1, \ldots, z_N)$, where $z_i = 1$ for $i \in S$ and $z_i = -1$ for $i \notin S$. A change of sign for z_i means a change of the state of membership for unit i vis-a-vis S: transition from $z_i = -1$ to $z_i = 1$ means adding i to S and the reverse transition means elimination of i from S. Denote the operator of sign changes by 0_i. For symmetric $\mathbf{A}$ the increment to (26) after an application of the sign change operation is given by

$$\Delta_i(S) = f(0_i S, \pi) - f(S, \pi) = -2 z_i \sum_{j \in S} (a_{ij} - \pi) . \tag{27}$$

If S is optimal, then $\Delta_i(S) \leqslant 0$ for all $i = 1, \ldots, N$, and this state implies a "compactness" of S. Specifically, let us refer to the set S as "compact in average" iff the average proximity for all pairs of units mutually included in S is more than the average proximity over all pairs such that one is included in and the other is excluded from S.

Theorem 5 *If S maximizes criterion (26), then S is compact in average.*

Proof According to (27), an optimal S satisfies

$$z_i \sum_{j \in S} (a_{ij} - \pi) \geqslant 0$$

for each i. For $i \in S$ this implies $\sum_{j \in S} (a_{ij} - \pi) \geqslant 0$, or equivalently $\sum_{j \in S} a_{ij} \geqslant \pi(n-1)$, and, therefore, $(1/(n-1)) \sum_{j \in S} a_{ij} \geqslant \pi$. Similarly, we find that $(1/n) \sum_{j \in S} a_{ij} \leqslant \pi$ for $i \notin S$. These inequalities together imply the statement of the theorem.

The proofs of the last two Theorems follow proofs of analogous statements concerning optimal partitions (for $\bar{E} = P$) (Kupershtokh, Mirkin, and Trofimov 1976b; Kupershtokh 1976).

Theorem 5 reveals a certain quality or type of cluster corresponding to an optimal subset. Different definitions of a neighborhood would lead to other properties for optimal subsets.

The proof of Theorem 5 generates a simple algorithm ADDI for construction of a suboptimal subset S that is compact in average.

The ADDI algorithm. Begin with $S = 0$. At first, we select a pair i_o and j_o corresponding to maximal a_{ij} in $\mathbf{A}$. If $a_{i_o j_o} > \pi$, then we include i_o and j_o in S. Otherwise, each one-element set is a solution.

For nonempty S we consider all units $i \notin S$ and add to S that i with maximal average proximity with elements of S if this average proximity is more than π. If such unit i does not exist, then the process is finished, and the resulting S is a suboptimal set.

Though the S so constructed satisfies the condition that $\Delta_i \leqslant 0$, the algorithm itself uses only a_{ij} and requires no recourse to criterion (26). In this sense, ADDI is a typical clustering algorithm.

Alternatively, we can use only $\Delta_i(S)$ for generating the subset S. Then the state of unit i that maximizes $\Delta_i(S)$ is changed if $\max_i \Delta_i(S) > 0$. This process can be sequenced so that values Δ_i are computed by Δ_i's from the preceding step, without recourse to formula (27) (Borodkin et al. 1978). These calculations are based on the equality

$$\Delta_i(0_k S) = \Delta_i(S) + \Delta_{ik} \tag{28}$$

where

$$\Delta_{ik} = \begin{cases} 4z_i z_k(a_{ik} - \pi), & i \neq k \\ -2\Delta_i(S), & i = k. \end{cases} \tag{29}$$

Indeed, for $k \neq i$, (27) implies

$$\Delta_i(0_k S) = f(0_i 0_k S, \pi) - f(0_k S, \pi) = -2 z_i \sum_{j \in 0_k S} (a_{ij} - \pi),$$

and subtracting $\Delta_i(S)$, we have

$$\Delta_{ik} = -2 z_i \left[\sum_{j \in 0_k S} (a_{ij} - \pi) - \sum_{j \in S} (a_{ij} - \pi) \right] = 4 z_i z_k (a_{ik} - \pi).$$

For $i = k$, $0_i 0_k S = S$, so that $\Delta_i(0_k S) = -\Delta_k(S)$, and the proof is complete.

The quantity Δ_{ik} plays the role of a second-order increment for the objective function in (26) during the process of movement across the system of neighborhoods. A universal concept of second-order increments in suboptimization processes was proposed by Borodkin and Muchnik (1977). It is clear that formulae (28) and (29) simplify the computation because they greatly reduce the number of times the elements of A are used. The method of second-order increments can also be developed for more complex criteria such as (7) and (12) (Mirkin and Rostovtsev 1978; Mirkin 1980). It

is clear from the construction of algorithm ADDI that solutions obtained are not worse than $\mathbf{R}^{ij}$ from (17), so that in using ADDI, Theorems 1-3 are true.

The proof of theorem 5 allows us similarly to characterize properties of suboptimal clusters for criteria (7) and (12) which include optimal values of λ and μ.

Criterion (7) for $\bar{E} = \overset{v}{S}$ has the form:

$$g(S) = (\sum_{i,j \in S} a_{ij})^2 / (n(n-1)) . \tag{30}$$

It is generated for the QFA-0 method with optimal $\pi = \frac{\lambda}{2}$, which is equal to half the average internal similarity in S by (5). By the proof of theorem 5 for $\lambda > 0$, each $i \notin S$ satisfies the inequality $(1/n) \sum_{j \in S} a_{ij} \leqslant \pi$ for suboptimal S. So the following property is true: the average proximity of each $i \notin S$ with the elements of S is less than half the average proximity of units mutually included in S. We refer to such an S as a "strict cluster." Analogously, for $\lambda < 0$ a suboptimal S will be a "strict anti-cluster" in the following sense: the average proximity of each $i \notin S$ with S is more than half the average internal proximity in S. So the following assertion is proved (Mirkin 1985):

Theorem 6 *If S maximizes criterion (30), then S is either a strict cluster or a strict anti-cluster.*

The theorem may be proved also by direct analysis of conditions insuring inequalities $g(S) - g(O_i S) \leqslant 0$ for all i. It leads to a suboptimal algorithm ADDI-S for construction of strict clusters by consecutive maximization of $g(O_i S) - g(S)$ (Mirkin 1985).

The ADDI-S algorithm. Begin with $S = 0$. At first, we select a pair i and j corresponding to maximal a_{ij} in **A**.

For nonempty S we consider all $i \notin S$ and add to S just that i which has maximal average proximity with elements of S, if this value is more than half the average internal proximity in S. If such unit i does not exist, then the process is finished, and the resulting S is a suboptimal cluster for the QFA-0 method.

The ADDI-S algorithm resembles the heuristic B-coefficient method for correlation matrices (Harman 1960) very closely and may be considered as a more precise version and justification of the previous one. We omit the details of recomputations of the average values from step to step.

TABLE 1

W1	100													
W2	30	100												
W3	58	18	100											
S1	34	5	35	100										
W4	46	17	38	56	100									
W5	7	46	−4	1	3	100								
W6	−12	−12	−20	9	3	11	100							
S2	−5	−5	−6	21	22	−7	22	100						
W7	−8	−26	−10	8	3	−4	33	19	100					
W8	−23	−23	−22	7	9	1	33	21	45	100				
W9	−24	−24	−15	5	−9	7	38	20	50	58	100			
S4	−19	−19	−24	−8	−7	11	38	−5	30	36	43	100		
I1	41	27	17	37	27	27	−7	−4	0	3	2	−3	100	
I3	−14	41	−17	−8	−7	27	5	36	0	−8	−9	−15	−11	100
	W1	**W2**	**W3**	**S1**	**W4**	**W5**	**W6**	**S2**	**W7**	**W8**	**W9**	**S4**	**I1**	**I3**

TABLE 2

Rank by λ	Weight λ	Members
1	.415	S2,I3
2	.385	W2,W5,I3
3	.363	W6,W7,W8,W9,S4
4	.342	W1,W2,W3,W4,S1,I1
5	.302	W1,W3,W4,S1
6	.287	W2,I3
7	.270	W1,W2,W5,I1
8	.235	W1,W3
9	.201	W4,W5,W6,W7,W8,W9,S1,S2,S4,I1,I3
10	.172	W6,W7,W8,W9,S2

TABLE 3

Order t	Weight λ_t	Members S_t	Share of total squares of proximities, %	Factor variance, %
0	9.26	All workers	14.75	–
1	32.40	WI,W2,W3,W4,S1,I1	29.74	34.88
2	40.40	W6,W7,W8,W9,S4	30.83	36.16
3	38.00	W2,W5,I3	8.18	9.60
4	12.66	W4,W6,S1,S2,I3	3.03	3.55
5	5.76	W1,W3,W4,W5,S1,I1	.94	1.10
6	12.01	W7,W8,W9,S1,S2	2.73	3.20
7	23.34	S2,I3	1.03	1.21

For QFA-0 method the ADDI-S algorithm is more effective than the ADDI algorithm, because it does not require continued reconstruction of S for recomputed π.

The criterion (12) of the QFA-1 method may be analyzed analogously. For example, for $\lambda > 0$, any set S, which is optimal by (12), has to satisfy the condition: the average proximity of each $i \notin S$ with elements of S is less than half the sum of the average external and average internal proximities μ (10) and $\lambda + \mu$ (11). This condition does not lead to such intuitively clear cluster structure as do previous conditions of strict clusteredness or compactness (the results reported in Section 7 support this conclusion).

7. An Illustrative Example

We use a matrix of correlations between all pairs of 14 workers with four values corrected according to Shepard and Arabie (1979, p. 112) (see Table 1; the letters W, S, I denote professions of the workers; all quantities are multiplied by 100). The matrix was computed by Breiger, Boorman, and Arabie (1975), using sociometric data from Roethlisberger and Dickson (1939). The given set of workers, according to the observer who collected the sociometric data, consisted of two "cliques": $S_1 = \{W1, W3, W4, S1, I1, W2\}$ and $S_2 = \{W7, W8, W9, S4, W6\}$ and three isolated individuals $W5$, $S2$, $I3$. (More detailed interpretations are given in Breiger et al. 1975; Shepard and Arabie 1979).

A set of 10 clusters was found in Shepard and Arabie (1979) which accounts for 89.0% of the variance (with an additive constant of .12) (see Table 2; all parameters are expressed in terms of the Table 1 entries divided by 100).

We do not quote interpretations as given in Shepard and Arabie (1979), but notice that Clusters 3 and 4 coincide with the cliques S_1 and S_2, though their weights are not the largest.

Table 3 presents the solution obtained by method QFA-0 (with preliminary subtraction of the universal cluster (first row); the last two columns correspond to formulae (20) and (21)). The first two clusters in the table correspond to the cliques S_1 and S_2 and account for the greatest shares of the variance. The cluster with the highest factor variance (36.16%) was found on the second step (not on the first) because of the local nature of algorithm ADDI. Taken together, the seven clusters presented in Table 3 account for 89.7% of the variance of input proximities, though only the first three subsets can be considered reliable. The rest of the clusters account for a too small share of the variance and may be considered the result of "a noise effect." These three clusters account for 80.6% of the variance and are present in Table 2 as well, though the additive clustering model itself

TABLE 4

Order	Weights λ_t	Weights μ_t	Members S_t	Factor variance, %
1	27.7	4.7	W1,W2,W3,W4,S1,I1	21.3
2	40.1	−4.4	W6,W7,W8,W9,S4	31.7
3	10.2	−4.0	W1,W2,W4,W5,W6,S1,S2,I1,I3	5.0
4	10.4	−5.1	W4,W5,W6,W7,W8,W9,S1,S2,S4,I1	5.4
5	37.9	−1.2	W2,W5,I3	9.3
6	10.4	−5.1	W1,W3,W4,W7,W8,W9,S1,S2,I3	5.2
7	10.6	−1.7	W1,W3,W4,W5,S1,I1	3.1
8	7.9	−3.1	W5,W6,W7,W8,W9,S2,S4,I1,I3	3.0
9	11.3	−1.2	W1,W2,W3,W5,I1	2.5
10	6.0	−1.8	W4,W6,W7,W8,W9,S1,S2,I3	1.5

TABLE 5

Order	Weights on 10th step			Members	Residual variance, %
t	λ_t	μ_t	ν_t	S_t	
1	27.7	4.7	1.35	W1,W2,W3,W4,S1,I1	78.7
2	40.1	−4.4	1.04	W6,W7,W8,W9,S4	46.5
3	39.6	−2.3	1.17	W2,W5,I3	36.0
4	13.5	−6.1	1.22	W4,W5,W6,W7,W8,W9,S1,S2,I1	26.2
5	10.7	−3.3	1.55	W1,W3,W4,W7,S1,S2,I3	21.7
6	7.2	−2.8	1.41	W5,W6,W8,W9,S1,S2,S4,I1,I3	18.6
7	13.3	−1.3	1.40	W2,W6,S2,I3	16.1
8	12.9	−1.8	1.41	W1,W2,W3,W5,I1	12.2
9	10.0	−0.3	1.17	W1,W4,W5,W6,S4	9.9
10	7.6	−1.4	1.16	W4,W7,W8,W9,S2,S4,I3	7.5

does not allow an analysis of their salience, because the weights λ_t can not be used for that purpose.

The results of applying methods QFA-1 and QFA-2 to the data in Table 1 are presented in Tables 4 and 5, respectively. Nonmonotonicity of the decreasing factor variances in the Tables is caused both by the suboptimality of algorithm ADDI and the essence of the QFA methods as techniques of projection onto a discrete cone of Boolean (α, β)-matrices. Three clusters with maximal factor variance, (1, 2, 5) in Table 4 and (1, 2, 3) in Table 5, coincide with three main clusters of the preceding Tables. Yet the total factor variance for these clusters here is a bit smaller: 62.3% in Table 4 and 64.0% in Table 5 while in Table 3 it is equal to 80.6%. The latter result

depends on an additive constant μ in S, because it diminishes λ, and on multiplying the factor variance by $1 - \epsilon$ (**R**). In Table 5, Cluster 4 with a solid share 9.8% of the variance is worth noting.

Conclusion

We have discussed the QFA methods and demonstrated the possibilities for using them for additive clustering. The QFA methods are intended for more general situations when qualitative factors are not limited to being clusters; they might be partitions, orderings, or macrostructures (blockmodels) represented by square Boolean matrices. In the special case of clusters, the QFA methods may serve as methods for fitting additive clustering models. Here, each step of the QFA computations has a strict "cluster" interpretation. A salience weight of the cluster turned out not to be λ_f itself, but the factor variance $\lambda_f^2 \in (\mathbf{r}')$ for QFA-0 or $\lambda_f^2 \in (\mathbf{R}')(1 - \epsilon\ (\mathbf{R}'))$ for QFA-1. These factor variances allow selection of the more important clusters and suggest an appropriate number of clusters.

Another aspect of the QFA methods compared to other algorithms for fitting the ADCLUS model (Shepard and Arabie 1979; Arabie and Carroll 1980) is a considerable restriction of freedom in searching for solutions. First, only one cluster is constructed on each step and no clusters have to be declared a priori. Second, some necessary conditions we proved clarify the nature of clusters resulting from such algorithms as ADDI. Third, these algorithms allow control of a sign in the weight coefficient λ_f such that selection of units with proximities whose averages are larger than the threshold π ensures $\lambda > 0$, while selection of units with proximities smaller than π leads to $\lambda < 0$. These facts help suggest solutions to problems $(i) - (iv)$ of additive clustering (see Part 1) and to some extent eliminate them. A technique for the qualitative factor analysis of many qualitative variables is given by Mirkin (1980, Ch. 4).

It is evident that there are many different versions of the QFA method in addition to the three considered. For example, after each iteration an algorithm may consecutively revise all or some of clusters found before, but we have no theoretical nor practical experience in such modifications.

References

ARABIE, P., BOORMAN, S. A., and LEVITT, P. R. (1978), "Constructing Block Models: How and Why," *Journal of Mathematical Psychology, 17*, 21-63.

ARABIE, P., and CARROLL, J. D. (1980), "MAPCLUS: A Mathematical Programming Approach to Fitting the ADCLUS Model," *Psychometrika, 45*, 211-235.

ARABIE, P., CARROLL, J. D., DeSARBO, W., and WIND, J. (1981), "Overlapping Clustering: A New Method for Product Positioning," *Journal of Marketing Research, 18*, 310-317.

BORODKIN, A. M., KRIEGMAN, I. E., MUCHNIK, I. B., and TELKOV, U. K. (1978), "Revealing of Design Jobs Set for Purposes of Design Automatization," *Automation and Remote Control, 5*, 97-105 (in Russian). "Identification of Industrial-design Activities for Computerization," *Automation and Remote Control, 39*, 5, 701-708 (in English).

BORODKIN, A. M., and MUCHNIK, I. B. (1977), "Approximate Algorithm Based on Second Increments Method for the Problem of Approximation of Graphs," *Automation and Remote Control, 4*, 114-120 (in Russian). "An Approximate Algorithm for Graph Approximations Based on the Method of Second Difference," *Automation and Remote Control, 38*, 4, Part 2, 548-553 (in English).

BREIGER, R. L., BOORMAN, S. A., and ARABIE, P. (1975), "An Algorithm for Clustering Relational Data with Applications to Social Network Analysis and Comparison with Multidimensional Scaling," *Journal of Mathematical Psychology, 12*, 328-383.

CARROLL, J. D., and ARABIE, P. (1983), "INDCLUS: An Individual Differences Generalization of the ADCLUS Model and the MAPCLUS Algorithm," *Psychometrika, 48*, 157-169.

HARMAN, H. H. (1960), *Modern Factor Analysis*, Chicago: University of Chicago Press.

KUPERSHTOKH, W. L. (1976), "On Threshold of Significance of Proximities in Classification Problem," In *Problems of Discrete Data Analysis*, ed. B. G. Mirkin, Novosibirsk, 70-74 (in Russian).

KUPERSHTOKH, W. L., MIRKIN, B. G., and TROFIMOV, W. A. (1976a), "Least Squares Method in Qualitative Variables Analysis," In *Problems of Discrete Data Analysis*, ed. B. G. Mirkin, Novisibirsk, 4-23 (in Russian).

KUPERSHTOKH, W. L., MIRKIN, B. G., and TROFIMOV, W. A. (1976b), "Sum of Internal Proximities as a Criterion of Classification Quality," *Automation and Remote Control, 3*, 91-98 (in Russian). "The Sum of Internal Couplings as an Index of Classification Quality," *Automation and Remote Control, 37, 3, Part 2*, 548-553 (in English).

MIRKIN, B. G. (1974), "Approximation Problems in Space of Binary Relations and Nonquantitative Variables Analysis," *Automation and Remote Control*, 51-61 (in Russian). "Approximation Problems in a Relative Space and the Analysis of Nonnumeric Methods," *Automation and Remote Control, 35, 9, Part 2*, 1424-1431 (in English).

MIRKIN, B. G. (1976), *Analysis of Qualitative Variables*, Moscow: Statistika Publishers (in Russian).

MIRKIN, B. G. (1980), *Analysis of Qualitative Attributes and Structures*, Moscow: Statistika Publishers (in Russian).

MIRKIN, B. G. (1985). *Classifications in Social and Economic Research*, Moscow: Finansy and Statistika Publishers (in Russian).

MIRKIN, B. G., KUPERSHTOKH, W. L., and TROFIMOV, W. A. (1979), "On Models of Discrete Data Aggregation," In *Mathematical Models and Methods in Economics Study*, eds. L. E. Bierland and B. G. Mirkin, Novosibirsk, 158-214 (in Russian).

MIRKIN, B. G., and RODIN, S. N. (1977), *Graphs and Genes*, Moscow, 1977 (English translation: Springer-Verlag, 1984).

MIRKIN, B. G., and ROSTOVTSEV, P. S. (1978), "A Method for Revealing of Associated Sets of Variables," In *Models of Social and Economics Data Aggregation*, ed. B. G. Mirkin, Novosibirsk, 107-112 (in Russian).

ROETHLISBERGER, F. J., and DICKSON, W. J. (1939), *Management and the Worker*, Cambridge, Mass.: Harvard University Press.

SHEPARD, R. N., and ARABIE, P. (1979), "Additive Clustering: Representation of Similarities as Combinations of Discrete Overlapping Properties," *Psychological Review, 86*, 2, 87-123.

TROFIMOV, W. A. (1976), "Models and Methods of Qualitative Factor Analysis for Proximity Matrices," In *Problems of Discrete Data Analysis*, ed. B. G. Mirkin, Novosibirsk, 24-35 (in Russian).

TROFIMOV, W. A. (1981a), "A Finite Method of Qualitative Factor Analysis," In *Methods of Multidimensional Economics Data Analysis*, ed. B. G. Mirkin, Novosibirsk, 12-29 (in Russian).

TROFIMOV, V. A. (1981b), "Experimental Base of Qualitative Factor Analysis Methods," In *Methods of Multidimensional Economics Data Analysis*, ed. B. G. Mirkin, Novosibirsk, 30-48 (in Russian).