Exact Sequential Algorithms for Additive Clustering

Pierre Hansen

GERAD andÉcole des Hautes Études Commerciales
3000 chemin de la Côte-Sainte-Catherine
Montréal, Canada H3T 2A7

Brigitte Jaumard

GERAD and École Polytechnique de Montréal
Département de Mathématiques et de Génie Industriel
C. P. 6079, succ. Centre-ville
Montréal (Québec) Canada H3C 3A7

Christophe Meyer

Institut für Mathematik B
Technische Universität Graz
Steyrergasse 30
A-8010 Graz, Austria

March, 2000

Les Cahiers du GERAD

G–2000–06

Abstract

Mirkin's additive clustering (or qualitative factor analysis) algorithm explains similarities between entities by sequentially finding a cluster of entities and a weight which minimizes the residual sum of squared errors. Simple heuristics are used to find the cluster at each step. We show that this step can be solved exactly by reducing it to a series of quadratic problems in 0-1 variables with a cardinality constraint. These in turn can be reduced to unconstrained quadratic 0-1 programs. Problems with up to 40 entities can be solved exactly. Both the case where weights must be positive and the case where they are unrestricted in sign are considered.

Keywords: additive clustering, quadratic $0-1$ programming.

Résumé

L'algorithme de Mirkin pour la classification additive (ou analyse factorielle qualitative) explique les similarités entre objets en trouvant séquentiellement une classe d'objets et un poids qui minimisent la somme résiduelle des erreurs au carré. Des heuristiques simples sont utilisées pour trouver la classe à chaque étape. Nous montrons que cette étape peut être résolue exactement en la ramenant à une série de problèmes quadratiques en variables 0-1 avec une contrainte de cardinalité. Ceux-ci sont à leur tour ramenés à des programmes quadratiques 0-1 non contraints. Des problèmes avec jusqu'à 40 objets peuvent être résolus exactement. Le cas où les poids doivent être positifs et le cas où ils sont non contraints en signe sont tous deux considérés.

Mots clés: classification additive, programmation quadratique $0-1$.

Acknowledgments: Work of the first and the second authors have been supported by FCAR (Fonds pour la Formation de Chercheurs et l'Aide à la Recherche) grant 95ER1048, and by NSERC (National Sciences and Engineering Research Council of Canada) grants GP0105574 and GP0036426.

1 Introduction

The additive clustering model (ADCLUS) has been independently developed in the USA by Shepard and Arabie [30], Arabie and Carroll [1] and in the former USSR by Mirkin and his collaborators [16, 20, 21, 26, 32, 33, 34] (under the name of Qualitative Factor Analysis). Generalizations of the model and algorithms to fit them have been proposed by numerous authors: MLADCLUS [13], INDCLUS [4, 6, 15], GENNCLUS [7], weak hierarchies additive clustering (which is intermediate between the standard additive clustering and the standard hierarchical model) [3], SEFIT [23], fuzzy additive clustering [27, 28], additive two-mode clustering (also known as box additive clustering) [25]. Applications of the additive clustering model can be found in various areas, like phonetics: the well known 16 consonant phoneme problem [1, 30, 31] of Miller and Nicely [19], the 26 letters confusion problem [35]; sociology: the 14 worker problem [22, 30] of Roethlisberger and Dickson [29]; marketing [2]; psychology [14], to cite only a few. Extended models of additive clustering have also applications in marketing [7] and in biology [3][9]. Recently, the additive clustering model has been discussed in the context of neural computation: see Lee [17]. Recent surveys on clustering methods can be found, e.g., in Mirkin [24], Hansen and Jaumard [11].

The additive clustering model represents inter-objects similarities as combinations of discrete and overlapping properties. An ADCLUS representation consists of a set of m (possibly overlapping) clusters, each having an associated nonnegative weight $w_k, k = 1, \ldots, m$. For any pair of objects, the predicted similarity is the sum of the weights of the clusters containing the given pair of objects. In matrix notation, the ADCLUS model can be written as

$$\hat{S} = PWP^t \tag{1}$$

where $\hat{S}$ is an $n \times n$ symmetric matrix of reconstructed similarities $\hat{s}_{ij}$ (assuming that n denotes the total number of objects), W is an $m \times m$ nonnegative diagonal matrix with the weights $w_k, k = 1, \ldots, m$ in the principal diagonal, P is an $n \times m$ rectangular matrix of binary values p_{ik} (with p_{ik} equal to 1 if object i is in cluster k, 0 otherwise) and P^t denote the transpose of P. Given a symmetric $n \times n$ matrix of similarities S, the problem is to estimate P and W in order to minimize the sum of the squared errors $\|\hat{S} - S\|^2$. (the matrix "norm" used in this paper is defined by $\|B\|^2 = \sum_{i \neq j} b_{ij}^2$ for a matrix $B = (b_{ij})$; in particular we do not care for the diagonal elements of the matrices).
Shepard and Arabie [30] impose that the cluster containing all objects must be part of the solution. Particularizing this cluster, the reconstructed similarities can be written as

$$\hat{S} = PWP^t + C \tag{2}$$

where P and W are as above and C is an $n \times n$ matrix (not necessarily nonnegative) having zeroes in its principal diagonal and the additive constant c (which has to be determined) in all the remaining entries.

The additive clustering problem is a particular case of the qualitative factor analysis developed by Kupershtokh, Mirkin and Trofimov (see, e.g., [16][22]). The qualitative factor analysis for proximity matrix aims to reconstruct a given proximity matrix by using a predefined set of boolean matrices with simple structures. No restriction is made on the sign of the matrix weights. If the set of boolean matrices is the set of matrices M that can be written $M = pp^t$ where p is a boolean vector, and if the weights are restricted to be positive, we obtain the additive clustering problem (in the additive clustering problem, only the cluster containing all objects can have a negative weight; see Shepard and Arabie [30, p.101] for a discussion about the sign of the weights).

Several methods have been suggested for fitting the ADCLUS model. In their original paper, Shepard and Arabie [30] first determine candidates for the clusters by considering maximal cliques (with ponderation on the edges corresponding to the similarity values) in the complete graph whose vertices are the n objects, and in a second step determine iteratively the associated weights while reducing the number of clusters (the clusters whose associated weight is less than a prespecified ε are eliminated). Arabie and Carroll [1] propose the MAPCLUS algorithm which combines a penalty approach with a gradient-based method of optimization. The obtained solution is then improved by using some exchange heuristics. Mirkin [22] searches for the clusters and their associated weight sequentially, i.e. at each iteration looks for the cluster that reduces the most the objective function (this problem will be referred to as the 1-cluster problem). Such a cluster is found by a neighborhood approach: beginning with an arbitrary cluster, objects are removed or added one by one until no amelioration of the objective function can be obtained. In [15], Kiers proposes a sequential algorithm for the more general model INDCLUS. Assuming initial values for the m clusters, their weight and the additive constant, one iteration of the algorithm goes as follows: choose a weight, and assuming all the rest is fixed, determine its optimal value. Then let the corresponding cluster vary, and determine the optimal one by solving an unconstrained quadratic $0-1$ problem. Finally do the same for the additive constant. The algorithm stops when no improvement is possible.

In this paper, we reconsider Mirkin's sequential approach and solve exactly the 1-cluster problem at each step. The paper is organized as follows. In Section 2, we present formally the sequential approach. In the four next sections, we consider the 1-cluster problems corresponding to the four combinations of the two properties "presence of an additive constant"/"absence of an additive constant" and "positive weights"/"weights unrestricted in sign". In Section 7, we prove convergence of the corresponding four sequential algorithms for the m-clusters problem. Finally, Section 8 presents some computational results.

2 The sequential approach

Given a symmetric similarity matrix $A = (a_{ij})$, the 1-cluster additive clustering problem in its most general form is to find the cluster (defined by its characteristic vector p) together with its weight w and an additive constant c such that the sum of squared errors between the given and explained similarities is minimized. Mathematically this amounts to solving the problem

$$\min \quad \sum_{i \neq j} (a_{ij} - wp_ip_j - c)^2$$

$$\text{s.t.} \quad \begin{cases} p \in \{0,1\}^n \\ w \geq 0. \end{cases}$$

The sequential approach for the m-clusters problem is as follows. We begin by solving the 1-cluster problem with $A = S$, where S is the initial similarity matrix, obtaining an optimal solution (p^1, w_1, c_1). We then compute the residual matrix $S^1 = S - w_1(p^1)(p^1)^t - c_1ee^t$ (where e is the vector with all components equal to 1) and solve again the 1-cluster problem for $A = S^1$, obtaining the solution (p^2, w_2, c_2), and so on. See Table 1 for a more formal description.

$S^0 \leftarrow S$
$k \leftarrow 0$
While (stopping condition not met)
Solve the 1-cluster problem with $A = S^k$, obtaining the optimal solution $(p^{k+1}, w_{k+1}, c_{k+1})$.
Compute the residual matrix $S^{k+1} = S^k - w_{k+1}(p^{k+1})(p^{k+1})^t - c_{k+1}ee^t$
$k \leftarrow k+1$
EndWhile

Table 1: Sequential algorithm for the m-cluster additive clustering problem

After k iterations, the residual matrix is

$$S^k = S - \sum_{\ell=1}^{k} w_\ell(p^\ell)(p^\ell)^t - \left(\sum_{\ell=1}^{k} c_\ell\right) ee^t$$

or, equivalently, the decomposition of the initial similarity matrix S is

$$S = \sum_{\ell=1}^{k} w_\ell(p^\ell)(p^\ell)^t + \left(\sum_{\ell=1}^{k} c_\ell\right) ee^t + S^k.$$

Note that since a same cluster can be obtained at several iterations, it could take more than m iterations to have m distinct clusters. One possible stopping rule is to stop when $\|S^k\|^2 \leq \varepsilon$ with ε sufficiently small.

Let $\Phi(S^k)$ be the norm of S^k as defined in Section 1, i.e., $\Phi(S^k) = \sum_{i \neq j} (S^k_{ij})^2$. At iteration k, we have to find $S^{k+1} = S^k - wpp^t - cee^t, w \geq 0, p \in B^n$ that minimizes $\Phi(S^{k+1})$. Since $\Phi(S^k)$ does not depend on w, p and c, this is equivalent to maximizing $\Delta\Phi(S^k) = \Phi(S^k) - \Phi(S^k - wpp^t - cee^t)$. Hence the 1-cluster problem at iteration k can be formulated as

$$\max \quad \delta\phi(S^k, p, w, c) \tag{3}$$

$$\text{s.t.} \quad \begin{cases} p \in B^n = \{0,1\}^n & (4) \\ w \geq 0 & (5) \end{cases}$$

with

$$\delta\phi(A, p, w, c) \quad = \quad \sum_{i \neq j} (a_{ij})^2 - \sum_{i \neq j} (a_{ij} - wp_ip_j - c)^2$$

$$= \quad 2\sum_{i\neq j} a_{ij}(wp_ip_j + c) - \sum_{i\neq j}(wp_ip_j + c)^2. \tag{6}$$

Note that contrary to what does Kiers, we optimize simultaneously the weight, the cluster and the additive constant.

In Sections 3 and 4, we consider the case where there is no additive constant in the formulation of the 1-cluster problem. In Section 3, we assume in addition that w is unrestricted in sign. This last assumption is also made in Section 5 where however the additive constant is present. Finally, in Section 6 we consider the 1-cluster problem as described in (3)-(5). In Section 7, we show that the corresponding sequential algorithms converge all to the null (residual) matrix (in a finite number of iterations for the positive additive clustering problem without additive constant).

3 The sign-unrestricted 1-cluster additive problem without additive constant

In this section, we assume that there is no additive constant c (or equivalently that c is fixed to 0). We do not however exclude the cluster containing all elements but rather handle it as any other cluster. We assume further that the weight w is not restricted to be positive, i.e., we do not consider constraint (5). The 1-cluster problem is then the following (we omit the parameters A and c for conciseness)

$$\max_{p\in B^n} \quad \delta\phi(p, w) = 2w\sum_{i\neq j} a_{ij}p_ip_j - w^2\sum_{i\neq j} p_ip_j. \tag{7}$$

3.1 Mathematical properties

If $0 \leq \sum_{i=1}^{n} p_i \leq 1$, $\delta\phi(p, w) = 0$. If $\sum_{i=1}^{n} p_i \geq 2$, an optimal w may be derived in closed form by writing that the derivative of $\delta\phi(p, w)$ with respect to w vanishes. We find

$$w(p) = \frac{\sum_{i\neq j} a_{ij}p_ip_j}{\sum_{i\neq j} p_ip_j} \tag{8}$$

which, after replacement in (7), gives the optimal value:

$$\delta\phi(p, w(p)) = \frac{\left(\sum_{i\neq j} a_{ij}p_ip_j\right)^2}{\sum_{i\neq j} p_ip_j}. \tag{9}$$

Hence problem (7) is equivalent to

$$\max_{p\in B^n; e.p\geq 2} \quad g_1^2(p) \tag{10}$$

with

$$g_1(p) \;=\; \frac{\sum\limits_{i\neq j} a_{ij}p_ip_j}{\sqrt{\sum\limits_{i\neq j} p_ip_j}}. \tag{11}$$

3.2 Solution

Assume that the matrix A has strictly negative and positive components (this may be the case starting from the second iteration of the sequential algorithm). Then problem (10) can be decomposed into two problems

$$\max_{p\in B^n;e.p\geq 2} g_1(p) \tag{12}$$

and

$$\min_{p\in B^n;e.p\geq 2} g_1(p). \tag{13}$$

By the assumption we have just made, the optimal value of (12) is positive while the optimal value of (13) is negative. It follows that the optimal value of problem (10) is the maximum of the square of these two values. If A is positive, we need only consider problem (12) and if A is negative only problem (13).

Noting that $\sum\limits_{i\neq j} p_ip_j = (\sum_i p_i)(\sum_i p_i - 1)$, problem (12) amounts to solving the parametric $0-1$ quadratic problems for $s = 2,\ldots,n$.

$$\begin{aligned} \max \quad & \frac{1}{\sqrt{s(s-1)}}\sum_{i\neq j} a_{ij}p_ip_j \\ \text{s.t.} \quad & \begin{cases} \sum\limits_i p_i = s \\ p\in B^n \end{cases} \end{aligned} \tag{14}$$

while problem (13) amounts to solving

$$\begin{aligned} \min \quad & \frac{1}{\sqrt{s(s-1)}}\sum_{i\neq j} a_{ij}p_ip_j \\ \text{s.t.} \quad & \begin{cases} \sum\limits_i p_i = s \\ p\in B^n \end{cases} \end{aligned} \tag{15}$$

(note that the latter is essentially the former with A replaced by $-A$). For fixed s, these constrained quadratic $0-1$ problems can be reduced respectively to the unconstrained ones (see for example Hansen [10]):

$$\max_{p\in B^n} \quad \sum_{i\neq j} a_{ij}p_ip_j - M\left(\sum_i p_i - s\right)^2$$

and

$$\min_{p\in B^n} \quad \sum_{i\neq j} a_{ij}p_ip_j + M\left(\sum_i p_i - s\right)^2$$

with M sufficiently large to ensure the constraint is satisfied by the optimal solution, e.g., $\sum_{i \neq j} |a_{ij}| + 1$.
The procedure to solve problem (7) is given in Table 2.

$M \leftarrow \sum_{i \neq j} |a_{ij}| + 1$
$z^* \leftarrow 0$
$s \leftarrow 2$
While $s \leq n$
  Solve the 0-1 quadratic problem

$$\max_{p \in B^n} \sum_{i \neq j} a_{ij} p_i p_j - M \left(\sum_i p_i - s \right)^2 .$$

  Let ρ^s be the optimal value and p^s an optimal solution
  If $\dfrac{(\rho^s)^2}{s(s-1)} > z^*$ **Then**

$$z^* \leftarrow \frac{(\rho^s)^2}{s(s-1)};\ p^* \leftarrow p^s;\ w^* \leftarrow \tfrac{\rho^s}{s(s-1)}$$

  EndIf
  $s \leftarrow s + 1$
EndWhile

$s \leftarrow 2$
While $s \leq n$
  Solve the 0-1 quadratic problem

$$\min_{p \in B^n} \sum_{i \neq j} a_{ij} p_i p_j + M \left(\sum_i p_i - s \right)^2 .$$

  Let ρ^s be the optimal value and p^s an optimal solution
  If $\dfrac{(\rho^s)^2}{s(s-1)} > z^*$ **Then**

$$z^* \leftarrow \frac{(\rho^s)^2}{s(s-1)};\ p^* \leftarrow p^s;\ w^* \leftarrow \tfrac{\rho^s}{s(s-1)}$$

  EndIf
  $s \leftarrow s + 1$
EndWhile

Return p^*, w^*

Table 2: Procedure for the sign-unrestricted 1-cluster additive problem without additive constant

This procedure requires to solve $2(n-1)$ quadratic 0-1 programs. This number can be reduced by computing lower and upper bounds on the optimal value of s. Appendix I shows how to compute such bounds for the maximization problem (12).

3.3 Convergence rate

Recall that the residual matrix A' is given by $A' = A - w^* p^* (p^*)^t$.

Proposition 1

$$\frac{\Phi(A')}{\Phi(A)} \leq 1 - \frac{2}{n(n-1)}.$$

Proof:

By definition of $\Delta\Phi$,

$$\Delta\Phi(A) = \Phi(A) - \Phi(A') = \delta\phi(p^*, w(p^*)) = \max_{p \in B^n; e.p \geq 2} \delta\phi(p, w(p)).$$

Let a_{uv} $(u \neq v)$ be an element in matrix A of maximal absolute value and consider the feasible solution $(\hat{p}, \hat{w})$ with $\hat{p}$ defined by $\hat{p}_u = \hat{p}_v = 1$, $\hat{p}_i = 0$ for $i \neq u, i \neq v$, and $\hat{w} = |a_{uv}|$. Then

$$\Phi(A) - \Phi(A') \geq \delta\phi(\hat{p}, \hat{w}) = 2\hat{w}^2.$$

By definition of a_{uv},

$$\Phi(A) = \sum_{i \neq j} a_{ij}^2 \leq n(n-1) a_{uv}^2.$$

Combining these two inequalities, we obtain

$$\Phi(A') \leq \left(1 - \frac{2}{n(n-1)}\right) \Phi(A)$$

which proves the result. ■

Note that this result was also obtained by Mirkin [22], despite the fact that he does not solve the 1-cluster problem exactly. Indeed, by looking carefully at the proof, we observe that the fact that p^* is optimal is not used but rather the weaker condition that p^* is at least as good as the solution $\hat{p}$ corresponding to a_{uv}.

4 The positive 1-cluster additive problem without additive constant

In this section, we consider the case in which the weights of the clusters are restricted to be positive. This restriction implies that, if at a given iteration of the sequential algorithm an element s_{ij}^{k+1} of the residual similarity matrix becomes negative, this element will remain negative at all further iterations, and hence the residual matrix will not tend to the null matrix even if an infinite number of iterations are performed. In order to avoid this bad behavior, we require in addition that the residual matrix remains positive. This leads to the constraints

$$a_{ij} - w p_i p_j \geq 0 \qquad \forall i, j \mid i \neq j$$
$$w \geq 0.$$

The 1-cluster problem to solve is then the following

$$\max \quad \delta\phi(p,w) = 2w\sum_{i\neq j} a_{ij}p_ip_j - w^2\sum_{i\neq j} p_ip_j \tag{16}$$

$$\text{s.t.} \quad \begin{cases} wp_ip_j \le a_{ij}, & \forall i,j \mid i\neq j \\ w \ge 0 \\ p\in B^n. \end{cases}$$

4.1 Solution

The solution method is based on the following observation.

Observation 1 *Let (p^*, w^*) be an optimal solution. Then $w^* = a_{ij}$ for some i, j such that $p_i^* = p_j^* = 1$.*

Indeed, for p fixed we have $w \le a_{ij}$ for all i, j such that $p_i = p_j = 1$, and the function $w \mapsto \delta\phi(p,w)$ is non-decreasing on $[0,\overline{w}]$ where $\overline{w} = \min_{(i,j)\mid p_ip_j=1}\{a_{ij}\}$.

This suggests to solve problem (16) with w equal to $a_{k\ell}$ and the additional constraints $p_k = p_\ell = 1$ for each (k,ℓ) such that $k\neq\ell$. After division by $a_{k\ell}$, the problem corresponding to (k,ℓ) is the following:

$$\max \quad \sum_{i\neq j}(2a_{ij} - a_{k\ell})p_ip_j$$

$$\text{s.t.} \quad \begin{cases} p_k = p_\ell = 1 \\ p_ip_j = 0, & \forall (i,j) \quad | \quad a_{ij} < a_{k\ell} \\ p\in B^n \end{cases}$$

or

$$\max \quad \sum_{(i,j)\mid a_{ij}\ge a_{k\ell}} (2a_{ij} - a_{k\ell})p_ip_j - M \sum_{(i,j)\mid a_{ij}< a_{k\ell}} p_ip_j \tag{17}$$

$$\text{s.t.} \quad \begin{cases} p_k = p_\ell = 1 \\ p_i = 0 & \forall i \quad | \quad a_{ik} < a_{k\ell} \text{ or } a_{i\ell} < a_{k\ell} \\ p\in B^n \end{cases}$$

with $M = \sum_{(i,j)\mid a_{ij}\ge a_{k\ell}} |a_{k\ell} - 2a_{ij}| + 1$.

The procedure to solve problem (16) is summarized in Table 3.

4.2 Convergence rate

The same result that for the sign-unrestricted variant holds:

Proposition 2

$$\frac{\Phi(A')}{\Phi(A)} \le 1 - \frac{2}{n(n-1)}.$$

Proof:

Similar to that of Proposition 1, noting that since A is positive, $\hat{w} = a_{uv}$ satisfies the positivity constraint. ∎

$z^* \leftarrow 0$
For all (k,ℓ) such that $a_{k\ell} > 0$
  Solve problem (17), obtaining an optimal value $F^{k\ell}$ and solution $p^{k\ell}$.
  If $a_{k\ell}F^{k\ell} > z^*$ **Then**
    $z^* \leftarrow a_{k\ell}F^{k\ell}$; $p^* \leftarrow p^{k\ell}$; $w^* \leftarrow a_{k\ell}$
  EndIf

EndFor

return p^*, w^*

Table 3: Procedure for the positive 1-cluster additive problem without additive constant

However, in this case, we have a more interesting result.

Proposition 3 *The residual matrix A' has at least one more null element than A.*

Proof:
Follows from the fact that w^* is equal to an element of A. ■

This result will be used in Section 7 to show that after at most $\frac{n(n-1)}{2}$ iterations of the corresponding sequential algorithm, the residual matrix is null.

Another interesting property of the algorithm is the following:

Proposition 4 *Let A be a similarity matrix, and A' be the residual matrix after application of the 1-cluster procedure to A. Then*

$$\Delta\Phi(A') \leq \Delta\Phi(A).$$

Proof:
Assume that $A' = A - w^*p^*(p^*)^t$, where (w^*, p^*) is an optimal solution of

$$\Delta\Phi(A) = \max \quad 2w\sum_{i\neq j} a_{ij}p_ip_j - w^2\sum_{i\neq j} p_ip_j \tag{18}$$

$$\text{s.t.} \quad \begin{cases} wp_ip_j \leq a_{ij}, & \forall i,j \mid i \neq j \\ w \geq 0 \\ p \in B^n, \end{cases}$$

and let $(\hat{w}, \hat{p})$ be an optimal solution of

$$\Delta\Phi(A') = \max \quad 2w\sum_{i\neq j}(a_{ij} - w^*p_i^*p_j^*)p_ip_j - w^2\sum_{i\neq j} p_ip_j$$

$$\text{s.t.} \quad \begin{cases} wp_ip_j \leq a_{ij} - w^*p_i^*p_j^*, & i \neq j \\ w \geq 0 \\ p \in B^n. \end{cases}$$

Clearly $(\hat{w}, \hat{p})$ is also a feasible solution to problem (18), hence

$$\Delta\Phi(A) \quad \geq \quad 2\hat{w}\sum_{i\neq j} a_{ij}\hat{p}_i\hat{p}_j - \hat{w}^2\sum_{i\neq j}\hat{p}_i\hat{p}_j$$

$$\geq \quad 2\hat{w}\sum_{i\neq j}(a_{ij} - w^* p_i^* p_j^*)\hat{p}_i\hat{p}_j - \hat{w}^2\sum_{i\neq j}\hat{p}_i\hat{p}_j = \Delta\Phi(A').$$

■

5 The sign-unrestricted 1-cluster additive problem with additive constant

In this section, the 1-cluster problem to solve is the following

$$\max_{p\in B^n, c\in \mathbb{R}} \quad \delta\phi(p,w,c) = 2\sum_{i\neq j} a_{ij}(wp_ip_j + c) - \sum_{i\neq j}(wp_ip_j + c)^2. \tag{19}$$

5.1 Centering of the residual matrix

We say that a matrix $B = (b_{ij})$ is centered if $\sum_{i\neq j} b_{ij} = 0$, i.e., if the sum of its off-diagonal elements is null. We have the following result.

Proposition 5 *Let (p^*, w^*, c^*) be an optimal solution of problem (19). Then the residual matrix $A' = A - w^*(p^*)(p^*)^t - c^*ee^t$ is centered.*

Proof:

The partial derivative of $\delta\phi$ with respect to c is

$$\frac{\partial\delta\phi}{\partial c} \quad = \quad 2\sum_{i\neq j}(a_{ij} - wp_ip_j - c) \tag{20}$$

which vanishes for $c^*(p,w) = \frac{\sum_{i\neq j}(a_{ij} - wp_ip_j)}{n(n-1)}$. Hence for (p^*, w^*, c^*) solving (19), the residual matrix $A' = (a'_{ij})$ is defined by

$$a'_{ij} = a_{ij} - w^* p_i^* p_j^* - \frac{\sum_{u\neq v}(a_{uv} - w^* p_u^* p_v^*)}{n(n-1)}, \qquad i \neq j.$$

But then $\sum_{i\neq j} a'_{ij} = 0$, which show that the matrix A' is centered. ■

Since the residual matrix is the input matrix A for the next call of the 1-cluster problem, we can assume that the matrix A is centered, i.e., that $\sum_{i\neq j} a_{ij} = 0$ (the initial matrix A is not necessarily centered; however we can show that performing first the centering and then solving the 1-cluster problem with centered matrix yields the same result than solving directly a general 1-cluster problem).

Let $s = \sum_{i=1}^{n} p_i$. Then c^* is related to p and w by the simple formula

$$c^*(p,w) = -w\frac{s(s-1)}{n(n-1)}. \tag{21}$$

Replacing c by its expression (21) in (19), we obtain

$$\delta\phi_2(p,w) \;=\; \delta\phi(p,w,c^*(p,w))$$

$$= 2w \sum_{i \neq j} a_{ij} \left(p_i p_j - \frac{s(s-1)}{n(n-1)} \right) - w^2 \sum_{i \neq j} \left(p_i p_j - \frac{s(s-1)}{n(n-1)} \right)^2$$

$$= 2w \sum_{i \neq j} a_{ij} p_i p_j - w^2 \left[\sum_{i \neq j} \left(1 - 2\frac{s(s-1)}{n(n-1)} \right) p_i p_j + n(n-1) \left(\frac{s(s-1)}{n(n-1)} \right)^2 \right]$$

$$= 2w \sum_{i \neq j} a_{ij} p_i p_j - w^2 s(s-1) \left(1 - \frac{s(s-1)}{n(n-1)} \right)$$

where we used the facts that $\sum_{i \neq j} a_{ij} = 0$ (centering), $p_i^2 = p_i$ for all i, and $\sum_{i \neq j} p_i p_j = s(s-1)$.

5.2 Mathematical properties

Note that $\delta\phi_2(p, w) = 0$ for p such that $0 \leq \sum_{i=1}^{n} p_i \leq 1$ or $\sum_{i=1}^{n} p_i = n$. If $2 \leq \sum_{i=1}^{n} p_i \leq n-1$, we have

$$\frac{\partial \delta\phi_2}{\partial w} = 2 \left(\sum_{i \neq j} a_{ij} p_i p_j - s(s-1) \left(1 - \frac{s(s-1)}{n(n-1)} \right) w \right)$$

which vanishes for

$$w = w^*(p) = \frac{\sum_{i \neq j} a_{ij} p_i p_j}{s(s-1) \left(1 - \frac{s(s-1)}{n(n-1)} \right)}.$$

Hence

$$\delta\phi(p) = \delta\phi_2(p, w^*(p)) = \frac{\left(\sum_{i \neq j} a_{ij} p_i p_j \right)^2}{s(s-1) \left(1 - \frac{s(s-1)}{n(n-1)} \right)}. \tag{22}$$

Hence maximizing $\delta\phi$ is equivalent to solving the problem

$$\max_{p \in B^n} h_1^2(p) \tag{23}$$

with

$$h_1(p) = \frac{\sum_{i \neq j} a_{ij} p_i p_j}{\sqrt{s(s-1) \left(1 - \frac{s(s-1)}{n(n-1)} \right)}}. \tag{24}$$

5.3 Solution

Problem (23) can be decomposed into two problems

$$\max_{p \in B^n; 2 \leq e.p \leq n-1} h_1(p) \tag{25}$$

and

$$\min_{p \in B^n; 2 \leq e.p \leq n-1} h_1(p). \tag{26}$$

Since the matrix A is centered and assumed to be non null, the optimal value of (25) is positive while the optimal value of (26) is negative. It follows that the optimal value of problem (23) is the maximum of the squares of these two values.

Problems (25) and (26) amount to solving some parametric $0-1$ quadratic problems for $s = 2, \ldots, n-1$, which can be reduced respectively to

$$\max_{p \in B^n} \quad \sum_{i \neq j} a_{ij} p_i p_j - M \left(\sum_i p_i - s \right)^2, \qquad s = 2, \ldots, n-1$$

and

$$\min_{p \in B^n} \quad \sum_{i \neq j} a_{ij} p_i p_j + M \left(\sum_i p_i - s \right)^2, \qquad s = 2, \ldots, n-1$$

with M equal to $\sum\limits_{i \neq j} |a_{ij}| + 1$.

The procedure to solve problem (19) is given in Table 4. Again, the computation of lower and upper bounds on the optimal value of s (see the Appendix) can reduce the number of quadratic $0-1$ problems to solve.

5.4 Convergence rate

Proposition 6

$$\frac{\Phi(A')}{\Phi(A)} \leq 1 - \frac{2}{(n-2)(n+1)}.$$

Proof:

By (22) and by definition of $\Delta\Phi$, we have

$$\Delta\Phi(A) = \Phi(A) - \Phi(A') = \frac{\left(\sum\limits_{i \neq j} a_{ij} p_i^* p_j^* \right)^2}{s^*(s^*-1)\left(1 - \frac{s^*(s^*-1)}{n(n-1)}\right)}$$

with $s^* = \sum\limits_{i=1}^{n} p_i^*$.

Let again a_{uv} be an element in matrix A with maximal absolute value. Consider the solution $(\hat{p}, \hat{w}, \hat{c})$ with $\hat{p}_i$ equal to 1 for $i \in \{u, v\}$ and to 0 otherwise, $\hat{w} = a_{uv}$ and $\hat{c} = -\frac{2}{n(n-1)} a_{uv}$. By optimality of p^*, we have

$$\frac{\left(\sum\limits_{i \neq j} a_{ij} p_i^* p_j^* \right)^2}{s^*(s^*-1)\left(1 - \frac{s^*(s^*-1)}{n(n-1)}\right)} \geq \frac{4a_{uv}^2}{2\left(1 - \frac{2}{n(n-1)}\right)}.$$

On the other hand, by definition of a_{uv}, we have $\Phi(A) = \sum\limits_{i \neq j} a_{ij}^2 \leq n(n-1)a_{uv}^2$.

Noting that $n(n-1) - 2 = (n-2)(n+1)$, we obtain

$$\Phi(A') \leq \left(1 - \frac{2}{(n-2)(n+1)}\right) \Phi(A).$$

$z^* \leftarrow 0$
$M \leftarrow \sum_{i \neq j} |a_{ij}| + 1$

$s \leftarrow 2$
While $s \leq n-1$
Solve the unconstrained quadratic 0-1 program

$$\max_{p \in B^n} \quad \sum_{i \neq j} a_{ij} p_i p_j - M \left(\sum_i p_i - s \right)^2 .$$

Let ρ^s be the optimal value and p^s an optimal solution
If $\dfrac{(\rho^s)^2}{s(s-1)\left(1 - \frac{s(s-1)}{n(n-1)}\right)} > z^*$ **Then**

$$z^* \leftarrow \frac{(\rho^s)^2}{s(s-1)\left(1 - \frac{s(s-1)}{n(n-1)}\right)};\ p^* \leftarrow p^s;\ w^* \leftarrow \frac{z^*}{\rho^s};\ c^* \leftarrow -w^* \frac{s(s-1)}{n(n-1)}$$

EndIf
$s \leftarrow s + 1$
EndWhile

$s \leftarrow 2$
While $s \leq n-1$
Solve the unconstrained quadratic 0-1 program

$$\min_{p \in B^n} \quad \sum_{i \neq j} a_{ij} p_i p_j + M \left(\sum_i p_i - s \right)^2 .$$

Let ρ^s be the optimal value and p^s an optimal solution
If $\dfrac{(\rho^s)^2}{s(s-1)\left(1 - \frac{s(s-1)}{n(n-1)}\right)} > z^*$ **Then**

$$z^* \leftarrow \frac{(\rho^s)^2}{s(s-1)\left(1 - \frac{s(s-1)}{n(n-1)}\right)};\ p^* \leftarrow p^s;\ w^* \leftarrow \frac{z^*}{\rho^s};\ c^* \leftarrow -w^* \frac{s(s-1)}{n(n-1)}$$

EndIf
$s \leftarrow s + 1$
EndWhile
return p^*, w^*, c^*

Table 4: Procedure for the sign-unrestricted 1-cluster additive problem with additive constant

■

6 The positive 1-cluster additive problem with additive constant

In this section, the 1-cluster problem to solve is the following

$$\max_{p\in B^n, w\geq 0, c\in\mathbb{R}} \quad \delta\phi(p,w,c) = 2\sum_{i\neq j} a_{ij}(wp_ip_j + c) - \sum_{i\neq j}(wp_ip_j + c)^2. \tag{27}$$

As in Section 5, we can assume that the matrix A is centered.

From Section 5.2 we deduce that for p fixed, $\delta\phi_2(p,w) = \delta\phi(p,w,c^*(p,w))$ is maximized for $w = \max\left\{0, \frac{\sum_{i\neq j} a_{ij}p_ip_j}{s(s-1)\left(1-\frac{s(s-1)}{n(n-1)}\right)}\right\}$ (where $s = \sum_{i=1}^{n} p_i$). Formula (22) implies that the solution $w = 0$ is dominated by all solutions $w = \frac{\sum_{i\neq j} a_{ij}p_ip_j}{s(s-1)\left(1-\frac{s(s-1)}{n(n-1)}\right)}$ corresponding to a $p \in B^n$ satisfying $2 \leq e.p \leq n-1$ and $\sum_{i\neq j} a_{ij}p_ip_j \geq 0$ (note that a p satisfying these constraints always exists because the matrix A is centered and not null). Hence our problem is the following

$$\begin{array}{ll} \max & h_1^2(p) \\ \text{s.t.} & \begin{cases} \sum_{i\neq j} a_{ij}p_ip_j \geq 0 \\ 2 \leq e.p \leq n-1 \\ p \in B^n \end{cases} \end{array}$$

which is equivalent to

$$\max_{p\in B^n} h_1(p).$$

This is nothing else than problem (25) of the previous section (see that section for its solution).

6.1 Solution

The procedure to solve the positive 1-cluster additive problem with additive constant is given in Table 5. Again, the computation of lower and upper bounds on the optimal value of s (see the Appendix) can reduce the number of quadratic $0-1$ problems to solve.

6.2 Convergence rate

Proposition 7

$$\frac{\Phi(A')}{\Phi(A)} \leq 1 - \frac{4}{(n-2)^2(n+1)^2}.$$

$z^* \leftarrow 0$
$M \leftarrow \sum_{i\neq j} \lvert a_{ij}\rvert + 1$
 	
$s \leftarrow 2$
While $s \leq n-1$
Solve the unconstrained quadratic 0-1 program
$$\max_{p\in B^n} \quad \sum_{i\neq j} a_{ij}p_ip_j - M\left(\sum_i p_i - s\right)^2.$$
Let ρ^s be the optimal value and p^s an optimal solution
If $\dfrac{(\rho^s)^2}{s(s-1)\left(1-\frac{s(s-1)}{n(n-1)}\right)} > z^*$ **Then**
$$z^* \leftarrow \frac{(\rho^s)^2}{s(s-1)\left(1-\frac{s(s-1)}{n(n-1)}\right)};\ p^* \leftarrow p^s;\ w^* \leftarrow \frac{z^*}{\rho^s};\ c^* \leftarrow -w^*\frac{s(s-1)}{n(n-1)}$$
EndIf
$s \leftarrow s+1$
EndWhile
 	
Return p^*, w^*, c^*

Table 5: Procedure for the positive 1-cluster additive problem with additive constant

Proof:

Using again (22), we have

$$\Delta\Phi(A) = \Phi(A) - \Phi(A') = \frac{\left(\sum\limits_{i\neq j} a_{ij}p_i^*p_j^*\right)^2}{s^*(s^*-1)\left(1-\frac{s^*(s^*-1)}{n(n-1)}\right)} \tag{28}$$

with $s^* = \sum\limits_{i=1}^{n} p_i^*$.

Let a_{uv} be a greatest element of A (note that a_{uv} is not necessarily the greatest in absolute value). Consider the solution $(\hat{p}, \hat{w}, \hat{c})$ with $\hat{p}_i$ equal to 1 for $i \in \{u, v\}$ and to 0 otherwise, $\hat{w} = a_{uv}$ and $\hat{c} = -\frac{2}{n(n-1)}a_{uv}$ (note that we have $\hat{w} \geq 0$). By optimality of p^*, we have

$$\frac{\left(\sum\limits_{i\neq j} a_{ij}p_i^*p_j^*\right)^2}{s^*(s^*-1)\left(1-\frac{s^*(s^*-1)}{n(n-1)}\right)} \geq \frac{2a_{uv}^2}{1-\frac{2}{n(n-1)}}. \tag{29}$$

Let $N_+ = \{(i,j) \mid a_{ij} > 0\}$ and $N_- = \{(i,j) \mid a_{ij} \leq 0\}$. Since A is centered and by definition of a_{uv}, we have

$$\sum_{(i,j)\in N_-} |a_{ij}| = \sum_{(i,j)\in N_+} |a_{ij}| \leq |N_+| a_{uv}.$$

Decomposing $\Phi(A)$, we obtain

$$\Phi(A) = \sum_{(i,j)\in N_-} |a_{ij}|^2 + \sum_{(i,j)\in N_+} |a_{ij}|^2.$$

Using Karush-Kuhn-Tucker conditions (see e.g., Luenberger [18]), it is easy to show that $\sum_{\ell=1}^{|N_-|} y_\ell^2$ attains its maximum over the set $\{y \geq 0 \mid \sum_{\ell=1}^{|N_-|} y_\ell \leq |N_+| a_{uv}\}$ for y satisfying $y_\ell = \frac{|N_+|}{|N_-|} a_{uv}\ \forall \ell = 1, \ldots, |N_-|$. Hence

$$\sum_{(i,j)\in N_-} |a_{ij}|^2 \leq \frac{|N_+|^2}{|N_-|} a_{uv}^2.$$

Moreover, by definition of a_{uv},

$$\sum_{(i,j)\in N_+} |a_{ij}|^2 \leq |N_+| a_{uv}^2.$$

Using the fact that $|N_+| + |N_-| = n(n-1)$, we obtain

$$\Phi(A) \leq \left(\frac{|N_+|^2}{n(n-1) - |N_+|} + |N_+| \right) a_{uv}^2 = \left(\frac{n(n-1)|N_+|}{n(n-1) - |N_+|} \right) a_{uv}^2.$$

Since A is symmetric, centered and non null, $|N_+| \leq n(n-1) - 2 = (n-2)(n+1)$. Hence

$$\Phi(A) \leq \frac{n(n-1)(n-2)(n+1)}{2} a_{uv}^2.$$

Comparing with (28) and (29), we obtain

$$\Phi(A') \leq \left(1 - \frac{4}{(n-2)^2(n+1)^2} \right) \Phi(A).$$

■

7 Convergence

We denote by ADDCLUS-U the sequential algorithm for the sign-unrestricted additive clustering problem without additive constant whose corresponding 1-cluster procedure was described in Section 3, by ADDCLUS the algorithm for the positive clustering problem without additive constant described in Section 4, by ADDCLUS-UC the algorithm for the sign-unrestricted additive clustering problem with additive constant presented in Section 5 and finally by ADDCLUS-C the algorithm for the positive clustering problem with additive constant considered in Section 6.

Recall that S^{k+1} denotes the residual matrix at end of iteration k. We have the following result:

Theorem 1 *For algorithms ADDCLUS-U, ADDCLUS, ADDCLUS-UC and ADDCLUS-C,* $\Phi(S^k) \to 0$ *when* $k \to +\infty$.

Proof:

By Propositions 1, 2, 6 and 7, we have

$$\frac{\Phi(S^{k+1})}{\Phi(S^k)} \leq \rho(n)$$

where $\rho(n) \in (0,1)$ depends uniquely on n. Hence $\lim_{k\to+\infty} \Phi(S^k) = 0$. ■

For algorithm ADDCLUS, we have a stronger result:

Theorem 2 *If algorithm ADDCLUS is used,* $\Phi(S^k) = 0$ *for* $k \geq \frac{n(n-1)}{2}$*.*
Moreover, the clusters are found in non-increasing order of their contribution to the sum of squared errors.

Proof:

By Proposition 3, S^{k+1} has at least one more element equal to 0 than S^k. Since the matrices S^k are symmetric, after at most $\frac{n(n-1)}{2}$ iterations, all entries of the residual matrix S^k are null (excepted the diagonal elements but we do not care of them). The second part of the theorem is a consequence of Proposition 4. ■

8 Computational results

We have tested algorithms ADDCLUS and ADDCLUS-C both on artificial data and on data of the literature. The programs have been implemented in C and run on a Sun-Sparc20/514MP (128Mram). The quadratic 0-1 programs are solved by Hansen *et al.* algorithm [12].

As a measure of the quality of the solution, we considered the *sum of squares accounted for* (in %):

$$s2af = 100\left(1 - \frac{\Phi(S)}{\Phi(\tilde{S})}\right)$$

and the *variance accounted for* (in %):

$$vaf = 100\left(1 - \frac{\sum_{i\neq j}(s_{ij} - \overline{s})^2}{\sum_{i\neq j}(\tilde{s}_{ij} - \overline{\tilde{s}})^2}\right)$$

where $\tilde{S}$ and S are respectively the initial similarity matrix and the residual one, and where $\overline{s}$ and $\overline{\tilde{s}}$ are the means of the entries of these 2 matrices.

8.1 Artificial data

We generated two sets of artificial data. In the first, 20 objects are distributed randomly into 4 possibly overlapping clusters (each object is assigned to each cluster with a probability of $\frac{1}{2}$). Weights of the clusters are chosen randomly in [0.1,0.6] (these parameters are taken from Tenenbaum [31]). Instances with clusters containing 0 or 1 object and/or with objects belonging to no cluster were rejected. A total of 10 instances was generated. The second set is constructed similarly, but with 40 objects distributed into 8 clusters.

With algorithm ADDCLUS, it takes less than 1s to solve each instance of the first set. For all of them, the 4 clusters were recovered with the correct weights. For the second set, only for two instances were the 8 clusters correctly found. For the 8 remaining instances, the explained sum of squares ranges from 83.51% to 97.50% with an average of 89.94%. The average CPU time is 37.3s.

With ADDCLUS-C, the CPU time is much longer: 153s in average for the first set, and for the second one, we were not able at all to solve them (program stopped after 12h). Moreover, the quality of the recovering is worse: in average 87.40% of the variance was recovered (in the worst case: 78.55%). This bad quality can be explained by the centering that introduces at the first iteration a positive additive constant. Although this initial constant could be compensated for by the additive constants of the subsequent iterations to give the correct value 0, this was never observed in practice. The longer time is due to the $0-1$ quadratic problems with cardinality constraint that have to be solved for each instance and that seem to be more difficult than the simple $0-1$ quadratic problems that arise in ADDCLUS. The bounding procedure that aims to reduce the number of quadratic problems to solve is rapid but unfortunately allows the elimination of a very small number of quadratic problems.

A more profound reason that explains that the algorithms cannot explain 100% of the variance despite the fact that no noise was introduced is that these are greedy algorithms: at the first iteration, the cluster that explains the greatest possible variance is found, but this may not be the optimal choice when explaining the variance with more than 1 cluster.

8.2 The 14 workers' problem

This problem is described in details in Shepard and Arabie [30] and in the references therein. The similarity matrix can be found e.g., in Mirkin [22]. Since algorithm ADDCLUS needs a similarity matrix with nonnegative entries, we performed a linear transformation on the matrix in [22] so that all its entries are in $[0, 1]$. Algorithm ADDCLUS-C was run with the original data.

Table 7 gives the results obtained by algorithm ADDCLUS. Table 6 corresponds to Table 3 of Mirkin [22], in which we have added the sum of squares accounted for and the variance accounted for after each iteration. These results were obtained by Mirkin's QFA-0 method, which is similar to our ADDCLUS procedure, except that there is no constraint on the positivity of the residual matrix. QFA-0 stops when the residual matrix is negative. Also note that QFA-0 first centers the matrix, which amounts to take the universal cluster with a negative weight.

Because of the transformation we made on the matrix, comparisons between QFA-0 and ADDCLUS must be viewed with caution. We can nevertheless say the following. The three first clusters found by QFA-0 are also found by ADDCLUS, in the same order but at rank 2 to 4. The first cluster of ADDCLUS, which is the universal cluster minus workers $W1$, $W2$, $W3$ and $I3$ can be thought off as corresponding to the universal cluster of QFA-0 resulting from centering. Starting from rank 5, the clusters found by QFA-0 and ADDCLUS differ. Note that contrary to what happens with QFA-0, ADDCLUS find its clusters in

	Weight	Members	$s2af$	vaf	$\Delta\Phi$
1	32.40	W1,W2,W3,W4,S1,I1	29.74	20.69	29.74
2	40.40	W6,W7,W8,W9,S4	60.57	53.81	30.83
3	38.00	W2,W5,I3	68.15	63.98	8.18
4	12.66	W4,W6,S1,S2,I3	71.78	68.92	3.03
5	5.76	W1,W3,W4,W5,S1,I1	72.72	71.41	0.94
6	12.01	W7,W8,W9,S1,S2	75.45	77.15	2.73
7	23.34	S2,I3	76.48	78.93	1.03

Table 6: Mirkin's results for the 14 workers' problem (QFA-0 method)

non-increasing order of $\Delta\Phi$: see Theorem 2. Also with 7 clusters, we explain a greater part of the sum of squares (89.79% versus 76.48%) and of the variance (86.63% versus 78.93%). Finally 7 clusters is the maximum number of clusters that QFA-0 can found because the residual matrix is negative, whereas ADDCLUS can explain 100% of the sum of squares and of the variance with a sufficiently large number of clusters (we explain 100% of the variance and of the sum of squares with 77 clusters). The time needed by ADDCLUS to find the 10 clusters is 0.76s.
In [30], Shepard and Arabie give a solution with 10 clusters that explain 91% of the variance. We do slightly better with 92.62% of the variance explained.

	Weight	Members	$s2af$	vaf	$\Delta\Phi$
1	0.2024	W4,W5,W6,W7,W8,W9,S1,S2,S4,I1	32.06	7.14	32.06
2	0.3690	W1,W2,W3,W4,S1,I1	55.27	38.46	23.21
3	0.4643	W6,W7,W8,W9,S4	70.00	56.25	14.73
4	0.6310	W2,W5,I3	77.53	70.99	7.53
5	0.2381	W1,W3,W4,S1,S2	83.14	77.71	5.61
6	0.1786	W7,W8,W9,S2,I3	86.89	82.43	3.75
7	0.1429	W1,W2,W3,W5,I1	89.79	86.63	2.90
8	0.3690	W6,S2,I3	92.24	91.08	2.45
9	0.1071	W1,W3,W4,W7,S1,I3	93.97	91.91	1.73
10	0.1190	W5,W8,W9,S1,I1	95.12	92.62	1.15

Table 7: Results for the 14 workers' problem with ADDCLUS

Tables 8 and 9 show the results obtained by Mirkin and by our method when there is an additive constant (Table 8 corresponds to Table 4 in Mirkin [22]). We observe that the two main clusters which explain more than 53% are found by the two methods, although not in the same order. For all number of clusters, ADDCLUS-C gives slightly better result in terms of the sum of squares accounted for and the variance accounted for. ADDCLUS-C took 6.20s to perform this computation.

	Weight	c	Members	$s2af$	vaf	Δvaf
1	27.7	4.7	W1,W2,W3,W4,S1,I1	32.91	21.30	21.30
2	40.1	-4.4	W6,W7,W8,W9,S4	59.96	53.03	31.73
3	10.2	-4.0	W1,W2,W4,W5,W6,S1,S2,I1,I3	64.22	58.03	5.00
4	10.4	-5.1	W4,W5,W6,W7,W8,W9,S1,S2,S4,I1	68.86	63.47	5.44
5	37.9	-1.2	W2,W5,I3	76.75	72.73	9.26
6	10.4	-5.1	W1,W3,W4,W7,W8,W9,S1,S2,I3	81.20	77.95	5.22
7	10.6	-1.7	W1,W3,W4,W5,S1,I1	83.87	81.08	3.13
8	7.9	-3.1	W5,W6,W7,W8,W9,S2,S4,I1,I3	86.45	84.10	3.02
9	11.3	-1.2	W1,W2,W3,W5,I1	88.59	86.62	2.52
10	6.0	-1.8	W4,W6,W7,W8,W9,S1,S2,I3	89.91	88.16	1.54

Table 8: Mirkin's results for the 14 workers' problem (QFA-1 method)

	Weight	c	Members	$s2af$	vaf	Δvaf
1	34.98	5.42	W6,W7,W8,W9,S4	35.33	24.13	24.13
2	32.30	-5.32	W1,W2,W3,W4,S1,I1	60.02	53.10	28.97
3	39.20	-1.29	W2,W5,I3	68.44	62.98	9.88
4	14.08	-5.57	W4,W5,W6,W7,W8,W9,S1,S2,I1	76.59	72.54	9.56
5	14.67	-1.61	W1,W3,W4,S1,S2	80.21	76.78	4.24
6	44.87	-0.49	S2,I3	83.97	81.19	4.41
7	12.51	-1.37	W7,W8,W9,S2,S4	86.60	84.28	3.09
8	15.93	-1.05	W1,W2,W5,I1	89.29	87.44	3.16
9	5.78	-2.29	W4,W5,W6,W8,W9,S1,S2,S4,I3	90.66	89.05	1.61
10	24.88	-0.27	W1,W3	91.82	90.40	1.35

Table 9: Results for the 14 workers' problem with ADDCLUS-C

8.3 The 16 consonant phonemes confusion problem

This problem originates from Miller-Nicely [19] study of subjects' errors in identifying 16 (English) consonant phonemes under different conditions of noise. The data can be found in Carroll and Wish [5].

Table 10 summarizes the result from the literature (in column c, n.a. means that the value of the constant was not available).

Tables 11 and 12 show the results for algorithms ADDCLUS and ADDCLUS-C. The CPU time is respectively 2.23s and 38.46s. In this case the obtained clusterings account for slightly less variance than the cited methods from the literature. This appears to be due to the sequential character of the algorithms used.

#clusters	c	vaf	Reference
16	n.a.	98.1	Arabie and Carroll [1]
16	0.057	94.5	Shepard and Arabie [30]
12	n.a.	95.6	Arabie and Carroll [1]
10	n.a.	93.7	Arabie and Carroll [1]
8	n.a.	90.7	Arabie and Carroll [1]
8	0.049	89.6	Arabie and Carroll [1]
8	0.047	90.2	Tenenbaum [31]

Table 10: Results from the literature for the 16 phonemes' problem

	Weight	Members	$s2af$	vaf	$\Delta\Phi$
1	0.2290	1,2,3	22.96	25.66	22.96
2	0.4230	4,5	39.03	45.10	16.07
3	0.0540	8,9,10,11,12,13	54.78	59.52	15.75
4	0.2880	9,10	62.23	67.97	7.45
5	0.2840	11,12	69.47	76.34	7.24
6	0.0380	1,3,5,6,7	74.53	80.43	5.06
7	0.0130	1,3,5,6,10,12,13,14,15,16	78.58	81.43	4.05
8	0.0770	1,3,4	81.54	84.23	2.96
9	0.0720	10,13,14	84.11	86.65	2.57
10	0.1560	8,11	86.30	88.99	2.19
11	0.1380	15,16	88.01	90.75	1.71
12	0.0130	2,4,6,7,9,12,13	89.52	91.38	1.51
13	0.0370	2,5,6	90.63	92.47	1.11
14	0.1080	13,15	91.68	93.52	1.05
15	0.0200	9,12,13,14	92.56	94.27	0.88
16	0.0330	1,3,5	93.34	94.99	0.78

Table 11: Results for the 16 phonemes' problem with ADDCLUS

	Weight	c	Members	$s2af$	vaf	Δvaf
1	0.2479	0.0528	1,2,3	53.62	25.81	25.81
2	0.3733	-0.0031	4,5	66.03	45.66	19.85
3	0.0979	-0.0122	8,9,10,11,12,13	77.32	63.72	18.06
4	0.2084	-0.0017	9,10	81.19	69.90	6.18
5	0.2061	-0.0017	11,12	84.97	75.95	6.05
6	0.0674	-0.0056	1,3,4,5,6	88.71	81.94	5.99
7	0.0452	-0.0038	10,13,14,15,16	90.39	84.63	2.69
8	0.0498	-0.0025	2,5,6,7	91.66	86.66	2.03
9	0.0954	-0.0008	1,3	92.47	87.95	1.29
10	0.0916	-0.0008	8,11	93.22	89.15	1.20
11	0.0860	-0.0007	15,16	93.87	90.20	1.05
12	0.0496	-0.0012	9,10,14	94.52	91.23	1.03
13	0.0448	-0.0011	12,13,14	95.05	92.07	0.84
14	0.0152	-0.0027	1,4,5,8,11,12,15	95.41	92.65	0.58
15	0.0131	-0.0039	1,3,5,6,7,9,10,13,14	95.79	93.27	0.62
16	0.0655	-0.0005	13,15	96.18	93.88	0.61

Table 12: Results for the 16 phonemes' problem with ADDCLUS-C

References

[1] Arabie, P., & Carroll, J. D. MAPCLUS: A Mathematical Programming Approach to Fitting the ADCLUS Model. *Psychometrika*, 1980, *45-2*, 211–235.

[2] Arabie, P., Carroll, J. D., Desarbo, W., & Wind, J. Overlapping Clustering: a New Method for Product Positioning. *Journal of Marketing Research*, 1981, *18*, 310–317.

[3] Bandelt, H.-J., & Dress, A. W. M. Weak Hierarchies Associated with Similarity Measures - An Additive Clustering Technique. *Bulletin of Mathematical Biology*, 1989, *51-1*, 133–166.

[4] Carroll, J. D., & Arabie, P. INDCLUS: An Individual Differences Generalization of the ADCLUS Model and the MAPCLUS Algorithm. *Psychometrika*, 1983, *48*, 157–169.

[5] Carroll, J. D., & Wish, M. Multidimensional Perceptual Models and Measurement Methods, in *Handbook of Perception*, vol. 2, Carterette and Friedman, eds., Academic Press, New-York 1974.

[6] Chaturvedi, A., & Carroll, J. D. An Alternating Combinatorial Optimization Approach to Fitting the INDCLUS and Generalized INDCLUS Models. *Journal of Classification*, 1994, *11*, 155-170.

[7] Desarbo, W. S. GENNCLUS: New Models for General Nonhierarchical Clustering Analysis. *Psychometrika*, 1982, *47-4*, 449–475.

[8] Dinkelbach, W. On Nonlinear Fractional Programming. *Management Science*, 1967, *13*, 492–498.

[9] Faith, D. P. A Model of Immunological Distances in Systematics. *Journal of Theoretical Biology*, 1985, *114*, 511–526.

[10] Hansen, P. Methods of Nonlinear 0-1 Programming. *Annals of Discrete Mathematics*, 1979, *5*, 53–70.

[11] Hansen, P., & Jaumard, B. Cluster Analysis and Mathematical Programming. *Mathematical Programming*, 1997, *79*, *1–3, Ser. B*, 191–215.

[12] Hansen, P., Jaumard, B., & Meyer, C. A Simple Enumerative Algorithm for Unconstrained $0-1$ Quadratic Programming. *Les Cahiers du GERAD*, 2000 (forthcoming).

[13] Hojo, H. A Maximum Likelihood Method for Additive Clustering and its Applications. *Japanese Psychological Research*, 1983, *25-4*, 191–201.

[14] Kazuhisa, T., & Osamu, T. Psychological Factors in the "ijime" Phenomenon: Negative Attitude Toward a Deviator and Conformity to Majority. *Japanese Journal of Educational Psychology*, 1988, *36-1*, 57–62 (in Japanese).

[15] Kiers, H. A. L. A Modification of the SINDCLUS Algorithm for Fitting the ADCLUS and INDCLUS. *Journal of Classification*, 1997, *14-2*, 297–310.

[16] Kupershtokh, W. L., Mirkin, B. G., & Trofimov, W. A. Least Squares Method in Qualitative Variables Analysis. In *Problems of Discrete Data Analysis*, ed. B. G. Mirkin, Novosibirsk, 4–23 (in Russian) 1976.

[17] Lee, M. D. Neural Feature Abstraction from Judgments of Similarity. *Neural Computation*, 1998, *10*, 1815–1830.

[18] Luenberger, D. G. *Linear and Nonlinear Programming*. Addison-Wesley Publishing Company, Reading (Massachusetts), second edition, 1989.

[19] Miller, G. A., & Nicely, P. E. An Analysis of Perceptual Confusions among some English Consonants. *Journal of the Acoustical Society of America*, 1955, *27*, 338–352.

[20] Mirkin, B. G. *Analysis of Qualitative Variables*. Moscow: Statistika Publishers (in Russian) 1976.

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

[22] Mirkin, B. G. Additive Clustering and Qualitative Factor Analysis Methods for Similarity Matrices. *Journal of Classification*, 1987, *4*, 7–31 (Erratum, *Journal of Classification*, 1989, *6*, 271–272).

[23] Mirkin, B. G. A Sequential Fitting Procedure for Linear Data Analysis Models. *Journal of Classification*, 1990, *7-2*, 167–195.

[24] Mirkin, B. *Mathematical Classification and Clustering. Nonconvex Optimization and its Applications, 11*. Kluwer Academic Publishers, Dordrecht, 1996.

[25] Mirkin, B. G., Arabie, P., & Hubert, L. J. Additive Two-Mode Clustering: the Error-Variance Approach Revisited. *Journal of Classification*, 1995, *12*, 243–263.

[26] Mirkin, B. G., Kupershtokh, W. L., & Trofimov, W. A. 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) 1979.

[27] Mirkin, B. G., & Satarov, G. A. The Method of Fuzzy Additive Types for the Analysis of Multidimensional Data. II. A Model and its Properties (Russian). *Avtomat. i Telemekh*, 1990, *5*, 134–140; translation in *Automat. Remote Control*, 1990, *51*, 683–688.

[28] Mirkin, B. G., & Satarov, G. A. The Method of Fuzzy Additive Types for the Analysis of Multidimensional Data. II. An Algorithm and its Properties (Russian). *Avtomat. i Telemekh*, 1990, *6*, 112–117; translation in *Automat. Remote Control*, 1990, *51*, 817–821.

[29] Roethlisberger, F. J., & Dickson, W. J. *Management and the Worker*. Cambridge, Mass.: Harvard University Press, 1939.

[30] Shepard, R. N., & Arabie, P. Additive Clustering: Representation of Similarities as Combinations of Discrete Overlapping Properties. *Psychological Review*, 1979, *86-2*, 87–123.

[31] Tenenbaum, J. B. Learning the Structure of Similarity. *Advances in Neural Information Processing Systems 8* p. 3–9, D. S. Touretzky, M. C. Mozer and M. E. Hasselmo (eds.), MIT Press, 1996.

[32] Trofimov, W. A. 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) 1976.

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

[34] Trofimov, W. A. Experimental Base of Qualitative Factor Analysis Methods. In *Methods of Multidimensional Economics Data Analysis*, ed. B. G. Mirkin, Novosibirsk, 30-48 (in Russian) 1981.

[35] Walker, M. E. An Analysis of Auditory Alphabet Confusions. *Perception & Psychophysics*, 1989, *45-4*, 315–322.

Appendix

I Computation of bounds on the parameter s for the quadratic parametric problems

When solving the weight-unrestricted 1-cluster problem without additive constant, we have to solve the problems

$$\max_{p\in B^n; s(p)\geq 2} \quad g_1(p) = \frac{|\sum\limits_{i\neq j} a_{ij}p_ip_j|}{\sqrt{s(p)\,(s(p)-1)}} \tag{30}$$

whereas when solving the 1-cluster problems with additive constant (both positive and weight-unrestricted), we have to consider the problems

$$\max_{p\in B^n; n-1\geq s(p)\geq 2} \quad h_1(p) = \frac{|\sum\limits_{i\neq j} a_{ij}p_ip_j|}{\sqrt{s(p)\,(s(p)-1)\left(1-\frac{s(p)(s(p)-1)}{n(n-1)}\right)}} \tag{31}$$

where $s(p) = \sum\limits_i p_i$. In this appendix we give lower and upper bound $\underline{s}$ and $\overline{s}$ on the optimal s^* for all of these problems. They are of the form

$$\max_{p\in P} \frac{|pAp|}{D(s(p))} \tag{32}$$

where P is some subset of B^n and D is a strictly positive function on P. The main result is the following

Proposition 8 *Let p^* be an optimal solution of problem (32) and $\hat{p}$ an optimal solution of problem*

$$\max_{p\in P} \quad \frac{|pAp|}{D_1(s(p))}$$

where D_1 is a strictly positive function on P. Assume furthermore that p^*Ap^* and $\hat{p}A\hat{p}$ are non-null. Then if $t \mapsto R(t) = \frac{D(t)}{D_1(t)}$ is non-decreasing (respectively non-increasing) on $s(P) = \{s(p) \,:\, p\in P\}$, we have $s^* \leq \hat{s}$ (respectively $s^* \geq \hat{s}$) with $s^* = s(p^*)$ and $\hat{s} = s(\hat{p})$.

Proof:

We prove the Proposition when R is non-decreasing. The case where R is non-increasing is handled similarly. By optimality of p^* and $\hat{p}$, we have

$$\frac{|p^*Ap^*|}{D(s^*)} \geq \frac{|\hat{p}A\hat{p}|}{D(\hat{s})}$$

$$\frac{|\hat{p}A\hat{p}|}{D_1(\hat{s})} \geq \frac{|p^*Ap^*|}{D_1(s^*)}.$$

Since all terms are strictly positive, we deduce

$$\frac{D(s^*)}{D_1(s^*)} \leq \frac{D(\hat{s})}{D_1(\hat{s})}$$

which, as R is non-decreasing, implies $s^* \leq \hat{s}$. ∎

I.1 Bounds for the weight-unrestricted 1-cluster problem without additive constant

In the case of the weight-unrestricted 1-cluster problem without additive constant, we have $D(t) = \sqrt{t(t-1)}$, $P = \{p \in B^n \,|\, s(p) \geq 2\}$ and $s(P) = [2, n]$. We are looking for a function D_1 of the form $D_1(t) = t + a$. We have then

$$R(t) = \frac{\sqrt{t(t-1)}}{t+a}.$$

By easy manipulations,

$$R'(t) = \frac{(2a+1)t - a}{2(t+a)^2\sqrt{t(t-1)}}.$$

It follows immediately that $R'(t) \leq 0$ for all t in $[2, n]$ (i.e., R is non-increasing) iff $a \leq -\frac{2}{3}$ and $R'(t) \geq 0$ for all $t \in [2, n]$ iff $a \geq -\frac{n}{2n-1}$. By Proposition 8, a lower bound $\underline{s}$ is obtained by solving the problem

$$\max_{p \in B^n; e.p \geq 2} \quad \underline{g}(p) = \frac{|\sum_{i \neq j} a_{ij} p_i p_j|}{\sum_i p_i - \frac{2}{3}} \tag{33}$$

while an upper bound $\overline{s}$ is obtained by considering the problem

$$\max_{p \in B^n; e.p \geq 2} \quad \overline{g}(p) = \frac{|\sum_{i \neq j} a_{ij} p_i p_j|}{\sum_i p_i - \frac{n}{2n-1}}. \tag{34}$$

Since the value of the objective function of these two problems is 0 for $p \in B^n$ satisfying $\sum_i p_i = 1$ (because there are no linear terms in the quadratic function of the numerator), the problems (33) and (34) can be reduced respectively to $\max\limits_{p \in B^n \setminus \{0\}} \underline{g}(p)$ and $\max\limits_{p \in B^n \setminus \{0\}} \overline{g}(p)$ which can be solved by Dinkelbach [8] iterative procedure for fractional quadratic 0-1 problem. Briefly, Dinkelbach's procedure for the fractional problem $\min\limits_{x \in X} \frac{h_1(x)}{h_2(x)}$ with $h_2(x) > 0$ for all $x \in X$ consists in solving the problem $(P_\lambda) \quad \min\limits_{x \in X} h_1(x) - \lambda h_2(x)$ where $\lambda = \frac{h_1(\overline{x})}{h_2(\overline{x})}$ for some $\overline{x} \in X$. If the optimal value of this problem is positive, then $\overline{x}$ is optimal; otherwise λ is set to $\frac{h_1(x^*)}{h_2(x^*)}$ where x^* is the optimal solution of problem (P_λ) and the procedure iterates.

I.2 Bounds for the 1-cluster problem with additive constant

In the case of the 1-cluster problem with additive constant, we have $D(t) = \sqrt{t(t-1)\left(1 - \frac{t(t-1)}{n(n-1)}\right)}$, $P = \{p \in B^n | n-1 \geq s(p) \geq 2\}$. Let $X = s(s-1)$. We have $X(P) = [2, (n-2)(n-1)]$. We are looking for a function D_1 of the form $D_1(X) = X + a$. We have then

$$R(X) = \frac{\sqrt{X\left(1 - \frac{X}{n(n-1)}\right)}}{X+a}.$$

Easy manipulations show that

$$R'(X) = \frac{-\left(1 + \frac{2a}{n(n-1)}\right) X + a}{2(X+a)^2 \sqrt{X\left(1 - \frac{X}{n(n-1)}\right)}}.$$

The numerator of $R'(X)$ is negative for $X \in [2, (n{-}2)(n{-}1)]$ for all $a \in [-\frac{(n-2)(n-1)n}{n-4}, \frac{2}{1-\frac{4}{n(n-1)}}]$. In order that $X + a$ be positive, we need in addition that $a > -2$. By Proposition 8, a lower bound $\underline{X}$ on $X = s(s-1)$ is obtained by solving the problem

$$\max_{p \in B^n; n-1 \geq e.p \geq 2} \quad \underline{h}(p) = \frac{|\sum_{i \neq j} a_{ij} p_i p_j|}{\sum_{i \neq j} p_i p_j + \frac{2}{1-\frac{4}{n(n-1)}}} \tag{35}$$

while an upper bound $\overline{X}$ is obtained by considering the problem

$$\max_{p \in B^n; n-1 \geq e.p \geq 2} \quad \overline{h}(p) = \frac{|\sum_{i \neq j} a_{ij} p_i p_j|}{\frac{(n-2)(n-1)n}{n-4} - \sum_{i \neq j} p_i p_j}. \tag{36}$$

$\underline{X}$ and $\overline{X}$ give in turn lower and upper bounds on s.