Raport Badawczy

Research Report

RB/12/2016

Determining of the estimate of the equivalence relation for moderate and large size sets

L. Klukowski

Instytut Badań Systemowych
Polska Akademia Nauk

Systems Research Institute
Polish Academy of Sciences

POLSKA AKADEMIA NAUK

Instytut Badań Systemowych

ul. Newelska 6
01-447 Warszawa
tel.: (+48) (22) 3810100
fax: (+48) (22) 3810105

Kierownik Zakładu zgłaszający pracę:
Prof. dr hab. inż. Zbigniew Nahorski

Warszawa 2016

DETERMINING OF THE ESTIMATE OF THE EQUIVALENCE RELATION FOR MODERATE AND LARGE SIZE SETS

Leszek Klukowski*

*Systems Research Institute Polish Academy of Sciences, Newelska 6, 01-447 Warsaw, e-mail: Leszek.Klukowski@ibspan.waw.pl

The paper presents two approaches for determining of estimates of the equivalence relation on the basis of pairwise comparisons with random errors. Obtaining of the estimates requires an optimal solution of a discrete programming problem which minimizes sum of differences between relation form and comparisons. The problem is NP hard and can be solved with the use of exact algorithms for moderate size of sets, i.e. about 50 elements. In the case of larger sets, i.e. at least 200 comparisons for each element, it is necessary to apply heuristic algorithms. The paper presents the results (a statistical preprocessing), which allow determining of the optimal or suboptimal solution with acceptable computational cost. These comprise: development of a statistical tests producing comparisons with low probabilities of errors and a heuristic algorithm based on the comparisons. Thus, the approach proposed guarantees applicability of the estimators for any size of set.

Key words: estimation of the equivalence relation, pairwise comparisons with random errors, nearest adjoining order idea

1. Introduction

The estimators of equivalence relation based on multiple pairwise comparisons with random errors, proposed in Klukowski (2011, 2012), require optimal solutions of a discrete programming problem. The problem minimizes differences between relation form, determined in appropriate way, and comparisons. The estimates are consistent, under non-restricted assumptions about comparisons errors; the speed of convergence is of exponential type (see Klukowski 2011) - for increasing number of comparisons of each pair. The optimization problems can be solved with the use of appropriate algorithms: the complete enumeration – for sets including not more than several elements, discrete mathematical programming – up to 50 elements (assuming single comparison of each pair), heuristic approach - for sets exceeding 50, especially in the case of multiple comparisons of each pair. Heuristic algorithms reduce computational costs, but can provide questionable solutions in the case of probabilities of comparisons errors not close to zero. However, large number of comparisons of any element, i.e. at least 200 single comparisons or 100 multiple comparisons, can be advantageous. It is so, because such size of set allows some preprocessing - obtaining of new single comparisons with significantly reduced probabilities of errors. The comparisons can be generated with the use of statistical tests proposed in the paper. Such the results can be used as the base of a proposed heuristic algorithm and also as a starting point for an exact discrete algorithm. The computational cost of "combined" approach is typically acceptable. This features make the approach proposed, based on nearest adjoining order idea (Slater 1961), highly efficient and applicable for any size of a set.

The paper consists of five sections. The second section presents the estimation problem, assumptions about pairwise comparisons and the form of estimator. In the third section are described concisely well-known exact optimization problems for equivalence relation, suitable for the sets with moderate number of elements. Next section presents statistical tests generating pairwise comparisons with reduced probabilities of errors, based on large number of initial comparisons, and the algorithm proposed. Last section summarizes the results.

2. Estimation problem, assumptions about comparisons, form of estimators

2.1. Estimation problems

We are given a finite set of elements $\mathbf{X}=\{x_1,\ldots,x_m\}$ ($3\le m<\infty$). It is assumed that there exists in the set $\mathbf{X}$ the equivalence relation, i.e.: reflexive, transitive, symmetric. The relation generates some family of subsets $\overset{*}{\chi}_1,\ldots,\overset{*}{\chi}_n$ $(n\geq 2)$; each subset includes equivalent elements only.

The family $\overset{*}{\chi}_1,\ldots,\overset{*}{\chi}_n$ have the following properties:

$$\bigcup_{q=1}^{n}\overset{*}{\chi}_q=\mathbf{X}, \qquad (1)$$

$$\overset{*}{\chi}_r\cap\overset{*}{\chi}_s=\{\mathbf{0}\}, \qquad (2)$$

where:

$\mathbf{0}$ – the empty set,

$$x_i,x_j\in\overset{*}{\chi}_r\equiv x_i,x_j-\text{equivalent elements}, \qquad (3)$$

$$(x_i\in\overset{*}{\chi}_r)\wedge(x_j\in\overset{*}{\chi}_s)\equiv x_i,x_j\quad (i\neq j,r\neq s)\text{- non-equivalent elements.}$$

(4)

The relation defined by (1) - (4) can be defined, alternatively, by the values $T(x_i,x_j)$ $((x_i,x_j)\in\mathbf{X}\times\mathbf{X}))$:

$$T(x_i,x_j)=\begin{cases}0\ \mathit{if\ exists}\ \ r\ \mathit{such\ that}\ \ (x_i,x_j)\in\chi_r^*,\\ 1\ \mathit{otherwise.}\end{cases} \qquad (5)$$

2.2. Assumptions about pairwise comparisons

The relation $\chi_1^*,\ldots,\chi_n^*$ is to be determined (estimated) on the basis of N ($N\geq 1$) comparisons of each pair $(x_i,x_j)\in\mathbf{X}\times\mathbf{X}$; any comparison $g_k(x_i,x_j)$ $(k=1,\ldots,N)$ evaluates the actual value of $T(x_i,x_j)$ and can be disturbed by a random error.

The following assumptions are made:

A1. The number of subsets n is unknown.

A2. The probabilities of errors $g_k(x_i,x_j)-T(x_i,x_j)$ $(k=1,\ldots,N)$ have to satisfy the following assumptions:

$$P(g_k(x_i,x_j)-T(x_i,x_j)=\kappa_{ij}\,\big|\,T(x_i,x_j)=\kappa_{ij})\geq 1-\delta$$
$$(\kappa_{ij}\in\{0,1\},\ \delta\in(0,\tfrac{1}{2})), \qquad (6)$$

$$P(g_k(x_i,x_j)-T(x_i,x_j)=\kappa_{ij}\,\big|\,T(x_i,x_j)=\kappa_{ij})+$$
$$P(g_k(x_i,x_j)-T(x_i,x_j)=\kappa_{ij}\,\big|\,T(x_i,x_j)\neq\kappa_{ij.})=1. \qquad (7)$$

A3. The comparisons $g_k(x_i,x_j)$; $((x_i,x_j)\in\mathbf{X}\times\mathbf{X}$; $k=1,\ldots,N)$ are independent random variables.

The assumptions A2 – A3 reflect the following properties of distributions of comparisons errors: • each probability of a correct comparison is greater than of incorrect one (inequalities (6), (7)), • zero is the median (in "sharp" form) and mode of each distribution of comparison error, • the comparisons are realizations of independent random variables, • the expected value of any error can differ from zero.

2.3. The form of estimator

The estimator presented in Klukowski (2011 Chap. 3, 2012), is based on the total sum of absolute differences between relation form (values $T(x_i,x_j)$) and comparisons $g_k(x_i,x_j)$ $((x_i,x_j)\in\mathbf{X}\times\mathbf{X})$. The estimates will be denoted $\hat{\chi}_1,\ldots,\hat{\chi}_{\hat{n}}$ or $\hat{T}(x_i,x_j)$. They are obtained on the basis of the discrete minimization problem:

$$\min_{\chi_r,\ldots,\chi_r\in F_X}\ \{\sum_{<i,j>\in R_m}\ \sum_{k=1}^{N}\left|g_k(x_i,x_j)-t(x_i,x_j)\right|\}, \qquad (8)$$

where:

$F_{\mathbf{X}}$ - the feasible set: the family of all relations $\chi_1, \ldots, \chi_r$ in the set **X**,

$t(x_i, x_j)$ - the values describing any relation $\{\chi_1, \ldots, \chi_r\}$ from $F_{\mathbf{X}}$,

R_m - the set of the form $R_m = \{< i, j> \mid 1 \leq i, j \leq m;\ j > i\}$.

The number of estimates, resulting from the criterion function (8) can exceed one, the minimal value of the function (8) is equal zero.

2.4. Properties of estimators

The analytical properties of the estimates, resulting from (8) are based on the random variables: $\Sigma_{R_m}\Sigma_k\left|g_k(x_i,x_j) - T(x_i,x_j)\right|$. The following results have been obtained by the author (Klukowski 2011):

(i) the expected values: $E(\Sigma_{R_m}\Sigma_k\left|g_k(x_i,x_j)-T(x_i,x_j)\right|)$ and $E(\Sigma_{R_m}\ \Sigma_k\left|g_k(x_i,x_j)-\widetilde{T}(x_i,x_j)\right|)$, i.e. corresponding – respectively - to actual and to any other relation $\widetilde{T}(x_i,x_j)$, satisfy the inequality:

$$E(\sum_{<i,j>\in R_m}\sum_{k=1}^{N}\left|g_k(x_i,x_j)-T(x_i,x_j)\right|) < E(\sum_{<i,j>\in R_m}\sum_{k=1}^{N}\left|g_k(x_i,x_j)-\widetilde{T}(x_i,x_j)\right|);$$

(9)

(ii) the variances of the above random variables divided by the number of comparisons N converge to zero, as $N \to \infty$, i.e.:

$$\lim_{N\to\infty} Var(\frac{1}{N}\sum_{<i,j>\in R_m}\sum_{k=1}^{N}\left|g_k(x_i,x_j)-T(x_i,x_j)\right|) = 0,$$

$$\lim_{N\to\infty} Var(\frac{1}{N}\sum_{<i,j>\in R_m}\sum_{k=1}^{N}\left|g_k(x_i,x_j)-\widetilde{T}(x_i,x_j)\right|) = 0; \qquad (10)$$

(iii) the probability of the inequality $\Sigma_{R_m}\Sigma_k\left|g_k(x_i,x_j)-T(x_i,x_j)\right| < \Sigma_{R_m}\Sigma_k\left|\ g_k(x_i,x_j)-\widetilde{T}(x_i,x_j)\right|$ converges to one, as $N \to \infty$, i.e.:

$$\lim_{N\to\infty} P(\sum_{<i,j>\in R_m}\sum_{k=1}^{N}\left|g_k(x_i,x_j)-T(x_i,x_j)\right| < \sum_{<i,j>\in R_m}\sum_{k=1}^{N}\left|g_k(x_i,x_j)-\widetilde{T}(x_i,x_j)\right|) = 1,$$

(11)

moreover:

$$P(\sum_{<i,j>\in R_m}\sum_{k=1}^{N}\left|g_k(x_i,x_j)-T(x_i,x_j)\right|<\sum_{<i,j>\in R_m}\sum_{k=1}^{N}\left|g_k(x_i,x_j)-\widetilde{T}(x_i,x_j)\right|\geq$$

$$1-\exp\{-2N(\tfrac{1}{2}-\delta)^2\} \qquad (12)$$

(inequality (12) is based on the Hoeffding (1963) inequality).

The relationships (i) - (iii) guarantee consistency and fast convergence to actual relation.

3. Optimization problems for the equivalence relation

The optimal solutions of the problem (8) can be obtained with the use of the discrete optimization algorithms, applied also in cluster analysis. They are usually formulated for fixed number n (because there exist methods for determining this number – see e.g. Gordon 1999, point 3.5). The discrete algorithms are presented: in Hansen et al (1994),
Hansen, P., Jaumard, B. (1997), Chopra, R. Rao, M.R. (1993), Gordon (1999, Chap. 3).

An initial approach (Rao 1971) has a form:

$$\min\{\sum_{j=1}^{n}\sum_{k=1}^{m}\sum_{l=1}^{m} d_{kl}z_{kj}z_{lj}\} \qquad (13)$$

$$\sum_{j=1}^{n} z_{kj}=1 \quad (k=1,\ldots,m), \qquad (14)$$

$$z_{kj}\in\{0,1\} \quad (j=1,\ldots,n;\ k=1,\ldots,m), \qquad (15)$$

where:

d_{kl} - distance (dissimilarity) between elements x_k, x_l,

z_{kj} - decision variable equal 1 if an element x_k is assigned to j-th cluster, zero otherwise.

The problem (13) – (15) has quadratic criterion function, linear constraints and {0, 1} variables. It can be applied for single comparison of each pair in the following way: distances d_{kl} ought to be replaced by comparisons $g_k(x_i,x_j)$ and optimal solution z_{kj}^* determines the form of n subsets. The problem can be applied also for the case N>1 using median from comparisons $g_1(x_i,x_j),\ldots,g_N(x_i,x_j)$.

The problem (13) – (15) is hard to solve in original form and, therefore, is linearized by assuming $y_{klj}=x_{kj}+x_{lj}-1$ and adding constraints $y_{klj}\leq x_{kj}$, $y_{klj}\leq x_{kj}$. The modified problem has also some drawbacks, especially large number of variables. Therefore, others approaches have been proposed - for the problem of the minimum weight equivalence relation (Hansen P. et al 1994, Hansen, P., Jaumard, B. 1997):

$$\min\{\sum_{k=1}^{m-1}\sum_{l=k+1}^{m} d_{kl} z_{kl}\} \qquad (16)$$

$$z_{kj} + z_{lq} - z_{kq} \leq 1 \quad (k = 1, ..., m), \qquad (17)$$

$$- z_{kj} + z_{lq} + z_{kq} \leq 1 \quad (l = k+1, ..., m-1), \qquad (18)$$

$$z_{kl} + z_{lq} + z_{kq} \leq 1 \quad (q = l+1, ..., m-1), \qquad (19)$$

$$z_{kj} \in \{0, 1\} \quad (k = 1, ..., n-1;\ l = k+1, ..., m). \qquad (20)$$

The problem (16) – (20) can be solved with the use of the dual linear relaxation and revised simplex algorithm. However, the approach need not always provide optimal solution and other approaches have been developed too (see Hansen et al 1994, Hansen, P., Jaumard, B. 1997); in general, they can be used for the number of elements not (significantly) greater than 50.

4. The algorithm based on test reducing probabilities of errors

The problem (8) can be effectively solved with the use of heuristic algorithms in the case of probabilities of errors close to zero. Such the probabilities indicate low fraction of incorrect comparisons - their expected value is equal $(m(m-1)/2)\delta N$. Large number of elements, i.e. $m \geq 100$, together with multiple comparisons ($N > 1$) or $m \geq 200$, allows obtaining of "new" comparisons with significantly lower probabilities of errors than δ. The base for such comparisons are statistical tests which verify identity of distributions of parallel comparisons: $g_k(x_i, x_1)$ and $g_k(x_r, x_1)$, …, $g_k(x_i, x_m)$ and $g_k(x_r, x_m)$ $(k = 1, ..., N; r \neq i)$.

The null hypothesis has a form H_0: all comparisons $g_k(x_i, x_j)$ and $g_k(x_r, x_j)$ $(k = 1, ..., N;\ r \neq i, j;\ i \neq j)$ have the same distributions, under alternative H_1: some of these comparisons have different distributions. The hypotheses can be replaced by: H_0: x_i, x_r are equivalent and H_1: x_i, x_r are not equivalent. The test statistic, proposed below, is based on values of the comparisons $g_k(x_i, x_j)$ and $g_k(x_r, x_j)$ $(k = 1, ..., N; r \neq i, j)$; it has, for $(m-1)N \geq 200$, Gaussian limiting distribution. The test allows determining of both errors; it is proper to fix them on similar level. It is clear that such the test reduces significantly the probability of error δ.

4.1. The test for equivalency of elements

The test proposed is based on random variables:

$$\eta_{irjk} = \begin{cases} 1 \ if \ g_k(x_i, x_j) = g_k(x_r, x_j), \\ 0 \ if \ g_k(x_i, x_j) \neq g_k(x_r, x_j). \end{cases} \qquad (k = 1, ..., N;\ r \neq i, j) \ (21)$$

The parameters of these (zero-one) variables are as follows: the expected value assumes, under H_0, the form:

$$E(\eta_{irjk}|\ H_0) = (1-\delta)^2 + \delta^2 \quad (r \neq i, j;\ \ j \neq i), \tag{22}$$

the variance – the form:

$$Var(\eta_{irjk}|\ H_0) = 2\delta(1-3\delta\ +4\delta^2 - 2\,\delta^3). \tag{23}$$

If H_1 is true the parameters of the variable η_{irjk} assume the form:

$$E(\eta_{irjk}\ |H_1) = 2\delta(1-\delta) \text{ and } Var(\eta_{irjk}\ |H_1) = 2\delta(1-3\delta\ +4\delta^2 - 2\,\delta^3).$$

(24)

It is obvious that:

$$E(\eta_{irjk}\ |H_0) = (1-\delta)^2 + \delta^2 > E(\eta_{irjk}\ |H_1) = 2\delta(1-\delta) \tag{25}$$

and that the difference of both expressions is equal: $1-4\delta(1-\delta)$.

The same parameters can be determined for the variables η_{irik} $(k = 1, ..., N)$, i.e. for *j=i*. They assume the form:

$$E(\eta_{irik}\ |H_0) = 1-\delta, \tag{26}$$

$$Var(\eta_{irik}\ |H_0) = \delta(1-\delta), \tag{27}$$

$$E(\eta_{irik}\ |H_1) = \delta, \tag{28}$$

$$Var(\eta_{irik}\ |H_1) = \delta(1-\delta). \tag{29}$$

The variables η_{irik} have higher expected value and lower variance than the variables η_{irjk} $(j \neq i)$.

The above results show that the distributions of the variables:

$$\frac{1}{(m-1)N} \sum_{r\neq i,j} \sum_{k=1}^{N} E(\eta_{irjk}\ |H_0) \text{ and } \frac{1}{(m-1)N} \sum_{r\neq i,j} \sum_{k=1}^{N} E(\eta_{irjk}\ |H_1) \tag{30}$$

are not the same: the expected value of the variable corresponding to H_1 is lower, while the variances of both variables are the same. Thus, the null hypothesis can be formulated in the form:

$$H_0:\ \sum_{r\neq i,j} \sum_{k=1}^{N} E(\eta_{irjk}) = N(m-1)((1-\delta)^2 + \delta^2) + N(1-\delta), \tag{31}$$

the alternative:

$$H_1: \sum_{r\neq i,j} \sum_{k=1}^{N} E(\eta_{irjk}) < N(m-1)((1-\delta)^2+\delta^2)+N(1-\delta)\,; \qquad (32)$$

The variance of both variables is equal:

$$Var(\sum_{r\neq i,j} \sum_{k=1}^{N} \eta_{irjk} \mid H_0) = Var(\sum_{r\neq r,j} \sum_{k=1}^{N} \eta_{irjk} \mid H_1) = \\ 2(m-1)N\delta(1-3\delta+4\delta^2-2\delta^3)+N\delta\,(1-\delta). \qquad (33)$$

In the case of large mN, the hypotheses (31), (32) can be replaced by:

$$H_0^{'}: \frac{1}{(m-1)N} \sum_{r\neq i,j} \sum_{k=1}^{N} E(\eta_{irjk}) = (1-\delta)^2+\delta^2, \qquad (34)$$

$$H_1^{'}: \frac{1}{(m-1)N} \sum_{r\neq i,j} \sum_{k=1}^{N} E(\eta_{irjk}) < (1-\delta)^2+\delta^2. \qquad (35)$$

The test statistics for null hypothesis assumes the form:

$$N((1-\delta)^2+\delta^2+\tfrac{1-\delta}{m-1}, \ \tfrac{1}{(m-1)N}2\delta(1-3\delta\ +4\delta^2-2\delta^3)+\tfrac{\delta(1-\delta)}{(m-1)^2N}). \ (36)$$

The test has one sided rejection region, i.e. values lower than the value, corresponding to assumed significance level α.

The example.

Let us examine the example: $\delta=0{,}1$; m=100; N=3. The difference of expected values of individual statistics $E(\eta_{irjk} \mid H_0)-E(\eta_{irjk} \mid H_1)$ equals 0,64, the statistics $E(\eta_{irik} \mid H_0)-E(\eta_{irik} \mid H_1)$ equals 0,8, the variance of the distribution (36) is equal 0,0005 (standard deviation 0,02236). In the case of elements $x_i \in \chi_p^*$ and $x_r \in \chi_q^*$, $(i \neq r,\ p \neq q)$ included in different subsets, each with 10 elements, the difference of statistics (31) and (32) is equal 0,1244. Therefore, the test based on Gaussian distribution guarantees both probabilities of errors lower than 0,003 and the expected value of incorrect comparisons lower than 15 (total number of comparisons equals 4550).

It is clear that, before estimation, the actual form of the relation is not known; therefore, the probabilities of the second type errors, cannot be determined precisely. Thus, it is rational to determine this probability for the subset $\chi_{\min}^*$ with minimal possible number of elements and to use the number for evaluation of the probabilities of both errors in the test. The minimal subset ought to include at least several percent of number m. Thus, "small" subsets (outliers), e.g. including less than 5-10% of elements of the set **X**, ought to be detected and excluded; their elements can be associated with an estimate based on a reduced set, as a next step. The detection of elements from small subsets can be done also on the basis of a statistical test. The null hypothesis assumes the form: $\sum_{j \neq i} T(x_i, x_j) \geq m - \nu - 1$ $(i = 1, ..., m)$ under alternative: $\sum_{j \neq i} T(x_i, x_j) < m - \nu - 1$, where: ν natural number satisfying $\nu \leq \zeta(m-1)$, where $\zeta \geq 0{,}05$. The test can be based on the properties of the statistics: $\frac{1}{N} \sum_{j \neq i} \sum_{k=1}^{N} g_k(x_i, x_j)$. Its expected value and variance can be determined under null hypothesis – they are equal, respectively, $(m-1-\nu)(1-\delta)$ and $(m-1-\nu)\delta(1-\delta)/N$; for the alternative the expected value is lower than $(m-1-\nu)(1-\delta)$. In the case $mN \geq 200$ the Gaussian asymptotic distribution can be applied. Rejecting of the null hypothesis, for an element x_i, means that it does not belong to a small subset; rejecting it to whole set **X** indicates lack of small subsets.

The comparisons obtained after above preprocessing (with low probabilities of errors and without small subsets) are satisfactory for heuristic algorithms, performing partitioning or agglomeration of elements. The algorithm proposed below belongs to the second group.

4.2 The form of the algorithm

The comparisons obtained on the basis of the hypotheses H_0 and H_1 are denoted $\mathbf{\Gamma} = \gamma(x_i, x_j)$ $(<i, j> \in R_m)$. The result $\gamma(x_i, x_j) = 0$ corresponds to H_0, while $\gamma(x_i, x_j) = 1$ - to H_1. The comparisons $\gamma(x_i, x_j)$ allow determining, for each element $x_i \in \mathbf{X}$, two sets: the first one $\Psi(x_i)$ comprises indexes of equivalent elements (acceptance of H_0), the second $\Omega(x_i)$ - indexes of non-equivalent elements (H_1). It is clear that equivalent elements x_i, x_j (H_0), have the same sets $\Omega(x_i) = \Omega(x_j)$; the sets $\Psi(x_i)$, $\Psi(x_j)$ satisfy the relationship $\Psi(x_i) - \{j\} = \Psi(x_j) - \{i\}$. Thus, the algorithm minimizing the function (8) can be based on detection of subsets $\hat{\chi}_r$ $(r = 1, ..., \hat{n})$ with these features or close to them.

START

1^0. To verify the null hypothesis H_0 (defined by (31)) for $(x_i, x_j) \in \mathbf{X} \times \mathbf{X}$ under alternative H_1 (32) , on the basis of comparisons $g_k(x_i, x_j)$, $g_k(x_r, x_j)$ $(k = 1, ..., N; r \neq i, j)$ assuming equal probabili-ties of both errors (the results $\Gamma = \gamma(x_i, x_j)$ $(i = 1, ..., m, j \neq i)$).

To determine the sets $\Omega(x_i)$, $\Psi(x_i)$ for each element $x_i \in \mathbf{X}$.

2^0. To merge elements of the set **X** having (exactly) the same sets $\Omega(x_i)$ and sets $\Psi(x_i)$ satisfying conditions: $\Psi(x_i)-\{j\}=\Psi(x_j)-\{i\}$; remaining elements assume, temporarily, as single element subsets. The family of subsets created in this way is denoted $\breve{\chi}_q$ $(q=1,\ldots,\breve{n})$ (or $\breve{t}(x_i,x_j)$); the subsets are indexed accordingly to number of elements; ordering of subsets having the same number of elements is optional.

To determine the value of the criterion function (8) after this operation; the value is denoted F_{cur}.

To determine the upper limit m_d of a difference $\sum_{j\neq i}\left|\gamma(x_i,x_j)-\gamma(x_r,x_j\right|$ $(r\neq i)$:

$m_d = \text{int}[(2\alpha(1-\alpha)(m-1)+3((2\alpha(1-\alpha)(1-2\alpha(1-\alpha))(m-1)))^{0,5}+0,5]$ where: α - significance level in the test H_0, int[z] – integer part of z;

and assume a value of the current limit $v_d=1$.

3^0. To determine the set Λ including elements x_i of the set **X**, satisfying the conditions:

- $x_i \in \breve{\chi}_q \quad (1\leq q\leq \breve{n})$,
- $\sum_{j\neq i}\left|\gamma(x_i,x_j)-\gamma(x_r,x_j)\right|\leq v_d$ for each element $x_r \in \breve{\chi}_s \quad (s\neq q)$,
- $x_r \in \breve{\chi}_s \Rightarrow x_r \notin \Lambda$.

4^0. Check the number of elements of the set Λ (denoted $\#\Lambda$).

If $\#\Lambda$=0 and $v_d<m_d$ then increase v_d by one $(v_d := v_d+1)$ and go to 3^0.

If $\#\Lambda$=0 and $v_d=m_d$ go to 7^0.

5^0. To determine elements of the set Λ which decrease the value F_{cur}, after joining to a subset $\breve{\chi}_s$ selected in point 3^0; to determine the value of the criterion function (8), corresponding to each joined element.

To remove from the set Λ all elements, which do not decrease the criterion function (8).

If the set Λ is empty then go to 7^0.

To identify an element x_i guaranteeing maximal decrease of the criterion function; if there exists other element x_r $(r\neq i)$ (or elements) with the same comparisons $g_k(x_r,x_j)$ $(r\neq i,\ k=1,\ldots,N)$ join all these elements to $\breve{\chi}_s$ and to determine the value F_{cur}.

In the case of multiple elements providing the same decrease of the criterion function (8), apply the sequence of criterions: • maximal power of (absorbing) set $\#\breve{\chi}_s$, • minimal power of the set $\#\breve{\chi}_q$ including an element x_i, • random selection.

6^0. To check the value F_{cur}.

If $F_{cur} = 0$ then go to 10^0.

If $F_{cur} > 0$ then to exclude from the set Λ the element(s) joined in point 5^0; if the set Λ is not empty, after the exclusion, then go to 5^0.

If the set Λ is empty and $\nu_d < m_d$ then increase the value ν_d by one $(\nu_d := \nu_d + 1)$ and go to 3^0.

7^0. To determine the set Δ, comprising elements of the set **X**, having significant contribution to the value of criterion function F_{cur}, i.e. all elements x_i satisfying the inequality:

$$\sum_{j\neq i} \left|\breve{t}(x_i,x_j)-\gamma(x_i,x_j)\right| > m_h,$$

where:

$$m_h = (m-1)\alpha + 3(\alpha(1-\alpha)(m-1))^{0,5}. \qquad (*)$$

If the set Δ is empty go to 10^0.

8^0. To determine the best relocation of each element of the set Δ, i.e. into a subset $\breve{\chi}_q$ $(1 \le q \le \breve{n})$ (also "new" subset $\breve{\chi}_{\breve{n}+1}$) providing maximal decrease of the criterion function (8); remove from the set Δ the elements which do not decrease the criterion function.

If the set Δ is empty then go to 10^0.

If the set Δ is not empty then select an element x_i providing maximal decrease of the value F_{cur}. In the case of multiple elements having the same comparisons $g_k(x_r,x_j)$ $(r \neq i,\ k = 1,\dots,N)$ select all these elements. In the case of different elements providing the same decrease of the value F_{cur} assume the sequence of criterions: • maximal power of (absorbing) set $\#\breve{\chi}_s$, • minimal power of the set including selected element (elements), random selection.

9^0. To relocate the element (or the elements having the same comparisons $g_k(x_r,x_j)$ $(r \neq i,\ k = 1,\dots,N)$) selected in point 8^0 and to determine the value of F_{cur}.

To remove the relocated elements from the set Δ. If the set Δ is not empty go to 8^0.

If the set Δ is empty, but previous relocation has decreased the value of the criterion function go to 7^0.

10^0. Assume $\breve{\chi}_q$ $(q = 1, \ldots, \breve{n})$ as the estimate $\hat{\chi}_q$ $(q = 1, \ldots, \hat{n})$.

END

The above algorithm is composed of two phases. The first phase is agglomeration of elements with similar sets $\Omega(\cdot)$, $\Psi(\cdot)$. Initially, the elements with the same sets $\Omega(\cdot)$ and sets $\Psi(\cdot)$ satisfying the conditions $\Psi(x_i) - \{j\} = \Psi(x_j) - \{i\}$, obtained as a result of verifying of the hypotheses H_0 and H_1, are agglomerated (point 2^0). Next, remaining elements are examined: the elements with a difference $\sum_{j \neq i} \left| \gamma(x_i, x_j) - \gamma(x_r, x_j) \right|$ not greater than one, for each x_r from a subset $\breve{\chi}_s$ $(1 \leq s \leq \breve{n})$, are detected; next they are agglomerated in the case of decreasing of the criterion function (8). Such the agglomeration is repeated for next values of the difference; its maximal value m_d is determined on the basis of the sum of: expected value of the variable $(\sum_{j \neq i} \left| \gamma(x_r, x_j) - \gamma(x_i, x_j) \right|)$, $((x_i, x_r) \in \overset{*}{\chi}_q \;\; (1 \leq q \leq n))$ and its three standard deviations. The formula determining m_d results from binomial distribution with probability α (significance level in the test).

The next steps of the first phase lead to estimates decreasing the criterion function (8). The phase is finished after exhaustion of elements with the difference $\sum_{j \neq i} \left| \gamma(x_i, x_j) - \gamma(x_r, x_j) \right|$ not exceeding m_d.

The second phase is oriented at "improvement" of the estimate obtained. The elements x_i of the current estimate $\breve{\chi}_q$ $(1 \leq q \leq \breve{n})$ having significant contribution to the criterion function (8) are detected. The threshold value of the contribution m_h is determined on the basis of expected value $E(\sum_{j \neq i} \left| \breve{t}(x_i, x_j) - \gamma(x_i, x_j) \right|)$ $((x_i, x_r) \in \overset{*}{\chi}_q \;\; (1 \leq q \leq n))$ and three standard deviations of the variable. The elements with significant contribution $\sum\limits_{j \neq i} \left| \breve{t}(x_i, x_j) - \gamma(x_i, x_j \right|$ are relocated to subsets guaranteeing a decrease of the criterion function. The phase is finished after exhaustion of such elements.

The estimate with the criterion function equal to zero is equivalent to exact optimal solution, while with low value – can be close or equal to exact. It is clear that comparisons $\gamma(x_i, x_j)$ having very low probabilities of errors (not greater than 10^{-3}) are also usable for discrete programming algorithms for the sets **X** having more than 50 elements. A computations cost may be acceptable in this case.

The literature of the subject contains many other heuristic algorithms - see Hansen, et al (1994). The estimate obtained in such a way can be verified with the use of tests stating existence of the relation against randomness of comparisons or equivalency of all elements – see e.g. Klukowski (2011), Gordon (1999, Chapt. 7); verification of individual subsets $\hat{\chi}_r\ (1 \le r \le \hat{n})$ can be done with the use of the (e.g.) Cochran test.

5. Concluding remarks

The paper presents the algorithms for solving of the optimization problem necessary for obtaining the estimates of the equivalence relation, on the basis of pairwise comparisons with random errors. The criterion function of the problem expresses the difference between relation form and comparisons. They are applicable for moderate (about 50 elements) and large sets (at least 100 elements with multiple comparisons). The moderate case can be solved with the use of well-known exact algorithms. Large number of comparisons indicates another approach - it allows construction of the tests generating "new" comparisons with significantly reduced probabilities of errors. Such the comparisons allow applying of the heuristic algorithm proposed in the paper. The result of such algorithm can be final, if the value of the criterion function approaches zero or close to zero, or provides starting point for exact algorithms. Thus, the approach based on minimization of differences between comparisons and relation form is useful, computationally efficient and reliable for any size of the set.

References

[1] CHOPRA, R., RAO, M.R., *The partition Problem.* Mathematical Programming, 1993, 59, 87-115.

[2] DAVID, H. A., *The Method of Paired Comparisons*, 2nd ed. Ch. Griffin, London 1988.

[3] GORDON, A. D., *Classification*, 2nd ed. Chapman&Hall/CRC, 1999.

[4] HANSEN, P., JAUMARD, B. *Cluster analysis and mathematical programming.* Mathematical Programming, 1997, 79, 191–215.

[5] HANSEN, P., JAUMARD, B., SANLAVILLE, E., *Partitioning Problems in Cluster Analysis: A Review of Mathematical Programming Approaches*. Studies in Classification, Data Analysis, And Knowledge Organization, Springer-Verlag, 1994.

[6] HOEFFDING, W., *Probability inequalities for sums of bounded random variables*. JASA, 1963, 58, 13–30.

[7] KLUKOWSKI L., *Some probabilistic properties of the nearest adjoining order method and its extensions*. Annals of Operational Research, 1994, 51, 241–261.

[8] KLUKOWSKI L., *The nearest adjoining order method for pairwise comparisons in the form of difference of ranks*. Annals of Operations Research, 2000, 97, 357-378.

[9] KLUKOWSKI, L., *Methods of Estimation of Relations of: Equivalence, Tolerance, and Preference in a Finite Set*. IBS PAN, Series: Systems Research, Vol. 69, Warsaw 2011.

[10] KLUKOWSKI, L., *Estimators of the Relations of Equivalence, Tolerance and Preference Based on Pairwise Comparisons with Random Errors*. Operations Research and Decisions, 2012, 22, 15-34.

[10] RAO, M.R., *Cluster Analysis and Mathematical Programming*. Journal of American Statistical Association, 1971, 66, 622-626.

[11] SLATER P., *Inconsistencies in a schedule of paired comparisons.* Biometrika, 1961, 48, 303–312.

The paper is extended version of the text presented at BOS'2016 Conference, 13-14 Oct. Warsaw, Poland