Raport Badawczy

Research Report

RB/9/2016

Estimation of trees on the basis of pairwise comparisons with random errors

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

Estimation of trees on the basis of pairwise comparisons with random errors

Leszek Klukowski

Systems Research Institute,

Polish Academy of Sciences,

Newelska 6, 01-447 Warszawa, Poland

Abstract

The estimators of the trees on the basis of multiple pairwise comparisons, with random errors, are proposed in the paper. The estimators are based on the idea of the nearest adjoining order (see Slater, 1961; Klukowski 2011). Two kinds of trees are examined: non-directed and directed. The approach is similar to estimation of the preference relation with incomparable elements on the basis of binary comparisons. The estimates are obtained on the basis of discrete optimization problems; their properties, especially accuracy, are similar to those for the preference relation. Such the trees can be applied to modelling of many phenomena, e.g. biological evolution, decision problems, etc.
Keywords: estimation of trees, non-directed and directed trees, pairwise comparisons with random errors

1 Introduction

The problems of estimation of the relations of: preference (complete and with partial order), equivalence and tolerance, on the basis of multiple pairwise comparisons with random errors, has been examined in Klukowski (1994, 2011 Chapt. 7 – 11, 2013, 2014a, b). The same approach can be applied to the trees – non-directed and directed; they are more general objects than the relations mentioned. The non-directed tree can be defined as a graph – non-directed, acyclic and complete; in other words: it doesn't exist a path (sequence of edges) from a fixed node (element) to this node and each pair of nodes is connected with a path. The directed tree can be defined as a graph directed, acyclic and complete. The directed graph has a root (initial node), paths leading in one direction and leafs (final nodes of a tree).

The problem of estimation of a tree, non-directed or directed, can be expressed as follows:

- it is given a finite set of nodes (elements) with unknown paths (system of edges);

- instead of system of edges, it is known a set of pairwise comparisons, which evaluate unknown paths, with random errors; any comparison states existence or non-existence of a connection between two elements – in the case of non-directed tree connection means an edge in the case of directed tree a connection mean a path and its direction;

- a random error means that a result of any comparison can be true or not with a probability satisfying some weak assumptions; any pair is compared N times ($N \geq 1$), all comparisons are assumed independent in stochastic way;

- the form of a tree, i.e. the system of its paths, has to be determined (estimated) on the basis of the set of pairwise comparisons characterized above.

The idea of estimation consists in minimization of differences between the form of a tree, expressed in appropriate way, and a given set of pairwise comparisons with random errors (Slater 1961, Klukowski 2011, 2013). The estimates are obtained as the optimal solutions of the discrete programming problems defined below; the number of solutions can exceed one.

The approach rested on the statistical paradigm provides the properties of estimates and the possibility of verification of the results obtained. The main property is consistency, for the number of comparisons N (for each pair) converging to infinity, under non-restricted assumptions about comparison errors. In general it is assumed that probability of correct comparison is greater than ½ and that multiple comparisons of each pair are independent random variables. The estimators can be also applied in the case of unknown distributions of comparison errors, which have to satisfy the assumptions made.

The idea of the estimators was introduced firstly by Slater (1961) - for the case of single, binary comparisons and the complete preference relation; some other ideas, in the area of pairwise comparisons, have been presented in: David (1988), Bradley (1984), Flinger and Verducci (1993), Gordon (1999), Klukowski (2011, 2013) .

The paper consists of four sections. The second section presents the definitions, notations and assumptions about comparison errors. The next sections consider the form of estimators, for both kinds of trees, and their properties. The last section summarizes the results. The Appendix presents proofs of some relationships determining properties of the estimators proposed.

2 Definitions, notations and assumptions about comparisons errors

2.1 Definitions and notations

The problem of estimation of the non-directed tree on the basis of pairwise comparisons can be stated as follows.

We are given a finite set of elements $\mathbf{X} = \{x_1, \ldots, x_m\}$ ($3 \leq m < \infty$). The elements of the set $\mathbf{X}$ (nodes) are connected with edges generating a non-directed tree (non-directed, acyclic and complete). Each pair of elements (x_i, x_j) can have an edge or not; thus the set of pairs of indices:

$$R_m = \{<i,j> \mid i=1,\ldots,m-1,\ j=i+1,\ldots,m\} \tag{1}$$

can be divided into two disjoint subsets – the first one I_o include pairs connected with an edge, the second one I_v pairs not connected with an edge, and $R_m = I_o \cup I_v$. Any pair $<i,j>$ is not ordered, i.e. is the same, as $<j, i>$.

The (non-directed) tree can be expressed with a use of values $T_v(x_i, x_j)$ $(<i,j> \in R_m)$, indicating existence or non-existence of an edge:

$$T_v(x_i, x_j) = \begin{cases} 1 \textit{ if } x_i \textit{ and } x_j \textit{ are connected with an edge,} \\ 0 \textit{ if } x_i \textit{ and } x_j \textit{ are not connectedw wit an edge.} \end{cases} \tag{2}$$

The values $T_v(x_i, x_j)$ define the non-directed tree in the unique way.

The similar considerations relate to the directed tree (directed, acyclic and complete). Such the tree can be expressed with a use of values $T_d(x_i, x_j)$ $(<i,j> \in R_m)$, indicating existence or non-existence of a path between elements (nodes) and direction of the path:

$$T_d(x_i, x_j) = \begin{cases} -1 \textit{ if there exists a path from } x_i \textit{ to } x_j, \\ 1 \textit{ if there exists a path from } x_j \textit{ to } x_i, \\ 2 \textit{ if there not exists a path between } x_i \textit{ and } x_j. \end{cases} \tag{3}$$

The set of indices R_m can be expressed as the alternative of the subsets $R_m = I_{\pm 1} \cup I_v$, where: $I_{\pm 1}$ includes indices of pairs of elements connected with a path and I_v includes indices of non-connected pairs; any pair of indices $<i,j> \in I_{\pm 1}$ of connected elements is ordered, i.e. shows that the direction of a path between x_i and x_j. It is clear that $T_d(x_i, x_j) = -T_d(x_j, x_i)$ for $<i,j> \in I_{\pm 1}$.

The values $T_d(x_i, x_j)$ define the directed tree in the unique way.

<u>Examples</u>

The non-directed tree - the set $\mathbf{X} = \{x_1, x_2, x_3, x_4, x_5, x_6\}$ with pairs connected with an edge: (x_1, x_2), (x_1, x_3), (x_3, x_4), (x_3, x_5), (x_3, x_6); the sets I_0 and I_v assume the forms:

$I_o = \{<1,2>, <1,3>, <3,4>, <3,5>, <3,6>\}$,

$I_v = \{<1,4>, <1,5>, <1,6>, <2,3>, <2,4>, <2,5>, <2,6>, <4,5>, <4,6>, <5,6>\}$. The values $T_v(x_i, x_j)$ assume the form:

$$T_v(x_i,x_j) = \begin{bmatrix} \times & 1 & 1 & 0 & 0 & 0 \\ & \times & 0 & 0 & 0 & 0 \\ & & \times & 1 & 1 & 1 \\ & & & \times & 0 & 0 \\ & & & & \times & 0 \\ & & & & & \times \end{bmatrix}.$$

The directed tree - the set $\mathbf{X}=\{x_1,x_2,x_3,x_4,x_5,x_6\}$ with the following order: x_1 precedes x_2, x_1 precedes x_3, x_2 not connected (incomparable) with x_3, x_3 precedes x_4, x_3 precedes x_5, x_3 precedes x_6, x_4 not connected with x_5, x_4 not connected with x_6, x_5 not connected with x_6. The sets $I_{\pm 1}$ and I_v assume the forms:

$I_{\pm 1} = \{<1,2>, <1,3>, <1,4>, <1,5>, <1,6>, <3,4>, <3,5>, <3,6>\}$,

$I_v = \{<2,3>, <2,4>, <2,5>, <2,6>, <4,5>, <4,6>, <5,6>\}$.

The values $T_d(x_i,x_j)$:

$$T_d(x_i,x_j) = \begin{bmatrix} \times & -1 & -1 & -1 & -1 & -1 \\ & \times & 2 & 2 & 2 & 2 \\ & & \times & -1 & -1 & -1 \\ & & & \times & 2 & 2 \\ & & & & \times & 2 \\ & & & & & \times \end{bmatrix}$$

2.2 Assumptions about distributions of comparisons errors

The form of both types of the trees, expressed - respectively - by $T_v(x_i,x_j)$ or $T_d(x_i,x_j)$, has to be determined (estimated) on the basis of N $(N \geq 1)$ comparisons of each pair (x_i,x_j) $(<i,j> \in R_m)$, evaluating the values $T_v(x_i,x_j)$ or $T_d(x_i,x_j)$, disturbed by random errors. The comparisons evaluating $T_v(x_i,x_j)$ and $T_d(x_i,x_j)$ will be denoted – respectively - $g_k^{(v)}(x_i,x_j)$ and $g_k^{(d)}(x_i,x_j)$ $(k=1,\ldots,N)$. The comparison errors - respectively $\phi_k^{(v)*}(x_i,x_j)$ and $\phi_k^{(d)*}(x_i,x_j)$ - can be expressed in the following form:

$$\phi_k^{(v)*}(x_i,x_j) = \begin{cases} 0 \textit{ if } g_k^{(v)}(x_i,x_j) \textit{ and } T_v(x_i,x_j) \textit{ are the same,} \\ 1 \textit{ if } g_k^{(v)}(x_i,x_j) \textit{ and } T_v(x_i,x_j) \textit{ are not the same,} \end{cases} \quad (4)$$

$$\phi_k^{(d)*}(x_i,x_j) = \begin{cases} 0 \textit{ if } g_k^{(d)}(x_i,x_j) \textit{ and } T_d(x_i,x_j) \textit{ are the same,} \\ 1 \textit{ if } g_k^{(d)}(x_i,x_j) \textit{ and } T_d(x_i,x_j) \textit{ are not the same.} \end{cases} \quad (5)$$

The distributions of comparison errors have to satisfy the following assumptions.

A1. Any comparison $g_k^{(\upsilon)}(x_i,x_j)$ $(\upsilon \in \{\nu, d\};\ k=1,\dots,N;\ <i,j>\in R_m)$ is an evaluation of the value $T_\upsilon(x_i,x_j)$; the probabilities of errors $P(\phi_k^{(\upsilon)*}(x_i,x_j)=l)$ $(\upsilon \in \{\nu, d\},\ l\in\{0,1\})$ have to satisfy the following assumptions:

$$P(\phi_k^{(\upsilon)*}(x_i,x_j)=0) \geq 1-\delta_\upsilon \quad (\delta_\upsilon \in (0, ½)), \qquad (6)$$

$$P(\phi_k^{(\upsilon)*}(x_i,x_j)=0) + P(\phi_k^{(\upsilon)*}(x_i,x_j)=1) = 1, \qquad (7)$$

A2. The comparisons: $g_k^{(\upsilon)}(x_i,x_j)$ $(k=1,\dots,N;\ <i,j>\in R_m)$ are independent random variables.

The assumptions about comparisons errors reflect the following facts. The probability of a correct comparison is greater than incorrect one (assumptions (6), (7)). The comparisons errors are independent in the stochastic way. The assumption can be relaxed in such a way that (multiple) comparisons of the same pair are independent and comparisons of pairs comprising different elements are independent.

The random variables $\phi_k^{(\nu)}(x_i,x_j)$ and $\phi_k^{(d)}(x_i,x_j)$, corresponding to any tree (non-directed or directed) - denoted, respectively, by $t_\nu(x_i,x_j)$ and $t_d(x_i,x_j)$, expressing differences between relation form and comparisons, assume the form:

$$\phi_k^{(\upsilon)}(x_i,x_j) = \begin{cases} 0 \textit{ if } g_k^{(\upsilon)}(x_i,x_j) \textit{ and } t_\upsilon(x_i,x_j) \textit{ are the same}, \\ 1 \textit{ if } g_k^{(\upsilon)}(x_i,x_j) \textit{ and } t_\upsilon(x_i,x_j) \textit{ are not the same}. \end{cases} \qquad (8)$$

3 Estimation problems and properties of estimates

The idea of the nearest adjoining order estimators is to minimize the absolute differences between a set of comparisons and the tree, expressed by the values $T_\nu(x_i,x_j)$ or $T_d(x_i,x_j)$. Thus, the estimates $\hat{T}_\nu(x_i,x_j)$ or $\hat{T}_d(x_i,x_j)$ $(<i,j>\in R_m)$ are the optimal solutions of the discrete programming problems – respectively:

$$\min_{F_X^{(\nu)}} \{ \sum_{<i,j>\in R_m} \sum_{k=1}^{N} \phi_k^{(\nu)}(x_i,x_j) \}, \qquad (9)$$

$$\min_{F_X^{(d)}} \{ \sum_{<i,j>\in R_m} \sum_{k=1}^{N} \phi_k^{(d)}(x_i,x_j) \}, \qquad (10)$$

where:

$F_X^{(\upsilon)}$ $(\upsilon \in \{\nu, d\}$ - feasible set, i.e. family of all trees (non-directed or directed) determined on the set **X**,

$\phi_k^{(\upsilon)}(x_i,x_j)$ $(\upsilon \in \{\nu, d\})$ - defined in (8).

In the earlier works of the author Klukowski (2011, 2012, 2013, 2014a, b), about estimation of the relations mentioned (equivalence, tolerance and preference), it was proved that the estimators based on the optimal solutions of the problems (9), (10) have good statistical properties, especially they are consistent, as $N \to \infty$. Moreover, the speed of the convergence can be determined - it is of the exponential type. The precision of estimates can be evaluated with the use of simulation approach. The proofs of the consistency are based on the following facts.

Firstly, the expected values of the random variables:

$$W_\nu^* = \sum_{<i,j>\in R_m} \sum_{k=1}^{N} \phi_k^{(\nu)*}(x_i,x_j), \tag{11}$$

$$W_d^* = \sum_{<i,j>\in R_m} \sum_{k=1}^{N} \phi_k^{(d)*}(x_i,x_j), \tag{12}$$

expressing the differences between the comparisons and the actual tree, ($T_\nu(x_i,x_j)$ or $T_d(x_i,x_j)$ $(<i,j>\in R_m)$), are lower than the expected values of the variables:

$$\widetilde{W}_\nu = \sum_{<i,j>\in R_m} \sum_{k=1}^{N} \widetilde{\phi}_k^{(\nu)}(x_i,x_j), \tag{13}$$

$$\widetilde{W}_d = \sum_{<i,j>\in R_m} \sum_{k=1}^{N} \widetilde{\phi}_k^{(d)}(x_i,x_j), \tag{14}$$

expressing differences between comparisons and any other relation, denoted by $\widetilde{T}_\nu(x_i,x_j)$ or $\widetilde{T}_d(x_i,x_j)$.

Secondly, the variances of the variables, divided by N, i.e.: $Var(\frac{1}{N}W_\nu^*)$, $Var(\frac{1}{N}W_d^*)$, $Var(\frac{1}{N}\widetilde{W}_\nu)$, $Var(\frac{1}{N}\widetilde{W}_d)$, converge to zero, as $N \to \infty$.

Thirdly, the probabilities: $P(W_b^* < \widetilde{W}_b)$ and $P(W_\mu^* < \widetilde{W}_\mu)$ converge to one, as $N \to \infty$; the speed of convergence is determined by the exponential subtrahend, obtained on the basis of Hoeffding (1963) inequality for bounded random variables. These relationships can be formulated shortly in the following

Theorem. *The following relationships hold true*:

$$E(W_\nu^*) < E(\widetilde{W}_\nu), \tag{15}$$

$$E(W_d^*) < E(\widetilde{W}_d), \tag{16}$$

$$\lim_{N\to\infty} Var(\tfrac{1}{N}W_v^*) = 0, \qquad \lim_{N\to\infty} Var(\tfrac{1}{N}\widetilde{W}_v) = 0, \tag{17}$$

$$\lim_{N\to\infty} Var(\tfrac{1}{N}W_d^*) = 0, \qquad \lim_{N\to\infty} Var(\tfrac{1}{N}\widetilde{W}_d) = 0, \tag{18}$$

$$P(W_v^* < \widetilde{W}_v) \geq 1 - \exp\{-2N(\tfrac{1}{2} - \delta_v)^2\}, \tag{19}$$

$$P(W_d^* < \widetilde{W}_d) \geq 1 - \exp\{-2N(\tfrac{1}{2} - \delta_d)^2\}. \tag{20}$$

Proofs of the relationships (15) – (20) are similar to the case of the relations mentioned (see Klukowski 1994, 2011 Chapt. 7, 8, 2014a, b), their idea is presented in the Appendix.

The relationships (15) – (20) are the theoretical basis for establishing the estimators $\hat{T}_v(x_i,x_j)$ and $\hat{T}_d(x_i,x_j)$ - they indicate consistency. This is so, because the random variables $\frac{1}{N}W_v^*$ or $\frac{1}{N}W_d^*$, corresponding to the actual trees, have minimal expected values in the family $F_X^{(v)}$ or $F_X^{(d)}$ and variances converging to zero. The optimal solutions of the problems (9) and (10), determining trees with minimal values of differences with respect to comparisons, converge to "true" trees with probability converging to one. The simulation experiments, concerning the preference relation (see Klukowski 2011, Chapt. 9), show that for $\delta_\upsilon = 0{,}1$ and $N \geq 3$ the frequency of correct estimates exceeds 75%. Moreover, estimation errors, i.e.:

$$\sum_{<i,j>\in R_m} \left|\hat{T}_\upsilon(x_i,x_j) - T_\upsilon(x_i,x_j)\right| \qquad (\upsilon \in \{v,d\}),$$

are close to zero. The values of N equal 7 provides frequency of correct results exceeding 95%.

Known properties of the estimates allow verification the results of estimation, i.e. checking if the tree is true model of the data, with the use of statistical tests (see Klukowski 2011, Chapt. 10, Gordon 1999, Chapt. 7).

The approach can be applied also in the case of unknown probabilities of comparison errors; it is necessary to satisfy the condition $\delta_\upsilon < \frac{1}{2}$. In such a case, and the number of N at least several, the values of δ_υ $(\upsilon \in \{v,d\})$ can be estimated. The precision of the estimates of the relations can be determined with the use of simulation approach, in a similar way, as in Klukowski, 2011, Chapt. 9.

Minimization of the problems (9), (10) requires discrete programming methods or heuristic algorithms. For a low number m of elements of the set **X**, i.e. several, the minimization can be performed simply by complete enumeration. For the moderate values of m, i.e. $m \leq 50$, and N=1 the problem can be solved with the use of known discrete methods. They are similar to the approach presented in Hansen, P., Jaumard. B., Sanlaville E. (1994) (for the equivalence relation) and in David, (1988) (the preference relation). In the case of N equal at least several the comparisons of each pair can be replaced by the median – it reduces N times number of variables in optimization problem. In the case of m>50 and multiple comparisons (N>1), heuristic algorithms are necessary. They can have low computational cost and good efficiency (close or identical result as exact methods). The reason of this fact are explained in Klukowski (2016).

4 Concluding remarks

The paper presents the estimators of the trees, non-directed and directed, which are based on the pairwise comparisons, in the binary form, disturbed by random errors. They have similar properties to the estimators of the relations of: equivalence, tolerance and preference (also including incomparable elements); in particular – consistency and speed of convergence. The statistical properties together with possibility of verification of estimates produce results which are trustworthy and reliable.

Appendix

The idea of the proof of the Theorem (relationships (15) – (20)).

The proof of the inequality (15), i.e. $E(W_\upsilon^*) < E(\widetilde{W}_\upsilon)$; the expected value of the difference $W_\upsilon^* - \widetilde{W}_\upsilon$ assumes the form:

$$E(W_\upsilon^* - \widetilde{W}_\upsilon) =$$

$$E(\sum_{<i,j>\in R_m} \sum_{k=1}^{N} \phi_k^{(\upsilon)*}(x_i,x_j) - \sum_{<i,j>\in R_m} \sum_{k=1}^{N} \widetilde{\phi}_k^{(\upsilon)}(x_i,x_j)) = \qquad \text{(A1)}$$

$$\sum_{k=1}^{N} \sum_{<i,j>\in R_m} E(\phi_k^{(\upsilon)*}(x_i,x_j) - \widetilde{\phi}_k^{(\upsilon)}(x_i,x_j)).$$

It is clear that each component $E(\phi_k^{(\upsilon)*}(x_i,x_j) - \widetilde{\phi}_k^{(\upsilon)}(x_i,x_j))$ can be either zero or negative; the value of zero corresponds to the case $T_\upsilon(x_i,x_j) = \widetilde{T}_\upsilon(x_i,x_j)$, negative – to the case of $T_\upsilon(x_i,x_j) \neq \widetilde{T}_\upsilon(x_i,x_j)$. The negative value results from the fact that any correct comparison, i.e. $\phi_k^{(\upsilon)*}(x_i,x_j) = 0$, indicates $\widetilde{\phi}_k^{(\upsilon)}(x_i,x_j) = 1$ and probability of the event equals $1-\delta_\upsilon$, i.e. greater than ½; the opposite case - incorrect comparison - $\phi_k^{(\upsilon)*}(x_i,x_j) = 1$ indicates $\widetilde{\phi}_k^{(\upsilon)}(x_i,x_j) = 0$ with probability δ_υ. Thus, the inequality $T_\upsilon(x_i,x_j) \neq \widetilde{T}_\upsilon(x_i,x_j)$ indicates existence of negative components; this fact is sufficient for the inequality (15).

The prof of the inequality (16) is similar.

The proof of the inequality (17) is obvious: the variable W_v^* is the sum of N iid. random variables $\sum_{<i,j>\in R_m} \phi_k^{(v)*}(x_i, x_j)$ $(k = 1, ..., N)$ with finite expected value and variance. Therefore the variance of the variable $\frac{1}{N}W_v^*$ converges to zero for $N \to\infty$. The convergence to zero of variances of the variables $\frac{1}{N}\widetilde{W}_v$, $\frac{1}{N}W_d^*$, $\frac{1}{N}\widetilde{W}_d$ of the remaining random variables is proved in similar way.

The inequalities (19) - (20) can be proved on the basis of Hoeffding's (1963) inequality for a sum of independent, binary random variables. The inequality applied in the case under consideration assumes the form:

$$P(\sum_{k=1}^{N} Y_k - \sum_{k=1}^{N} E(Y_k) \geq Nt) \leq \exp(-2Nt^2), \qquad (*)$$

where:

Y_k $(k = 1, ..., N)$ iid. random variables, satisfying: $P(0 \leq Y_k \leq 1) = 1$, $E(Y_k) < \frac{1}{2}$,

t – positive constant.

The inequality (*) can be applied to the random variables $\sum_{k=1}^{N} \sum_{<i,j>\in R_m} (\phi_{\upsilon k}^*(x_i, x_j) - \widetilde{\phi}_{\upsilon k}(x_i, x_j))$ $(\upsilon \in \{b, \mu\})$, after a following transformations:

$$P(\sum_{k=1}^{N} \sum_{<i,j>\in R_m} (\phi_k^{(\upsilon)*}(x_i, x_j) - \widetilde{\phi}_k^{(\upsilon)}(x_i, x_j)) < 0) =$$

$$1 - P(\sum_{k=1}^{N} \sum_{<i,j>\in R_m} (\phi_k^{(\upsilon)*}(x_i, x_j) - \widetilde{\phi}_k^{(\upsilon)}(x_i, x_j)) \geq 0).$$

Moreover:

$$P(\sum_{k=1}^{N} \sum_{<i,j>\in R_m} (\phi_k^{(\upsilon)*}(x_i, x_j) - \widetilde{\phi}_k^{(\upsilon)}(x_i, x_j)) \geq 0) =$$
$$P(\sum_{k=1}^{N} \sum_{<i,j>\in R_m} (2\phi_k^{(\upsilon)*}(x_i, x_j) - 1) \geq 0). \qquad (A2)$$

Assuming that the sum $\sum_{<i,j>\in R_m}(2\phi_k^{(\upsilon)*}(x_i, x_j) - 1)$ includes τ zero-one variables, the inequality (A2) can be transformed as follows:

$$
\begin{aligned}
&P(\sum_{k=1}^{N} \sum_{<i,j>\in R_m} (2\phi_k^{(\upsilon)*}(x_i,x_j)-1)\geq 0)=\\
&P(\sum_{k=1}^{N} \sum_{<i,j>\in R_m} \phi_k^{(\upsilon)*}(x_i,x_j)\geq N\tau/2)=\\
&P(\sum_{k=1}^{N} \sum_{<i,j>\in R_m} \phi_k^{(\upsilon)*}(x_i,x_j)-N\tau\,\delta_\upsilon\geq N\tau/2-N\tau\,\delta_\upsilon)=\\
&P(\sum_{k=1}^{N} \tfrac{1}{\tau} \sum_{<i,j>\in R_m} \phi_k^{(\upsilon)*}(x_i,x_j)-N\,\delta_\upsilon\geq N(\tfrac{1}{2}-\delta_\upsilon)).
\end{aligned}
\tag{A3}
$$

The last expression in (A3) can be evaluated on the basis of Hoeffding inequality:

$$
\begin{aligned}
&P(\sum_{k=1}^{N} \tfrac{1}{\tau} \sum_{<i,j>\in R_m} \phi_k^{(\upsilon)*}(x_i,x_j)-N\,\delta_\upsilon\geq N(\tfrac{1}{2}-\delta_\upsilon))\leq\\
&\exp(-2N(\tfrac{1}{2}-\delta_\upsilon)^2).
\end{aligned}
\tag{A4}
$$

The evaluation (A4) is equivalent to the inequality (19).

The inequality (20) can be proved in a similar way.

References

[1] Bradley, R. A. (1984) Paired comparisons: some basic procedures and examples. Handbook of Statistics Vol. 4, P. R. Krishnaiah and P. K. Sen, eds. Amsterdam: North-Holland, 299 – 326.

[2] David H. A. (1988) The Method of Paired Comparisons, 2nd ed. Ch. Griffin, London.

[3] Flinger A. F., Verducci J. S., eds. (1993) Probability Models and Statistical Analyses for Ranking Data. Lecture Notes in Statistics **80**, Springer-Verlag, New York, Berlin, Heidelberg.

[4] Gordon A.D. (1999) Classification, 2nd Ed., Chapman &Hall/CRC, Boca Raton London New York Washington, D.C.

[5] Hansen, P., Jaumard. B., Sanlaville E. (1994) Partitioning Problems in Cluster Analysis: A Review of Mathematical Programming Approaches. Studies in Classification, Data Analysis, and Knowledge Organization. Springer - Verlag.

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

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

[8] Klukowski, L. (2011) *Methods of Estimation of Relations of: Equivalence, Tolerance, and Preference in a Finite Set.* IBS PAN, Series: Systems Research, Volume 69, Warsaw.

[9] Klukowski L. (2012) Properties of estimators of the preference relation based on pairwise comparisons – simulation survey. In: K. T. Atanassov et al., eds., *New Developments in Fuzzy Sets, Intuitionistic Fuzzy Sets, Generalized Nets and Related Topics*, 75 - 90. *Volume II: Applications*, SRI PAS, Warsaw.

[10] Klukowski L. (2013) Estimations of the Relations of: Equivalence, tolerance and Preference on the Basis of Pairwise Comparisons. In: Burduk (et al, eds.) Proceedings of the 8th In Conference on Computer Recognition Systems CORES 2013, Springer Cham Heidelberg New York Dordrecht London.

[11] Klukowski L. (2014) Estimation of the partial order on the basis of pair-wise comparisons in binary and multivalent form. In: K. T. Atanassov et al., eds., *Modern Approaches in Fuzzy Sets, Intuitionistic Fuzzy Sets, Generalized Nets and Related Topics*, 21 - 41. *Volume II: Applications*, SRI PAS, Warsaw.

[12] Klukowski L. (2014) Estimation of the Relations of: Equivalence, Tolerance and Preference on the Basis of Pairwise Comparisons with Random Errors. In: Angelov P. et al. eds., *Intelligent Systems'2014. Proc. Of the 7th IEEE Int. Conf. Intelligent Systems IS'2014, Vol. 1: Mathematical Foundations, Theory, Analyses.* Springer Cham Heidelberg New York Dordrecht London.

[13] Klukowski L. (2016) Determining of the Estimate of the Equivalence Relation for Moderate and Large Size Sets. In: *Proc. of the XVI BOS'2016 Conference (*13 -14 Oct., 2016 year*)*, Systems Research Institute, Polish Academy of Sciences, Warsaw.

[14] Slater P. (1961) Inconsistencies in a schedule of paired comparisons. *Biometrika*, **48**, 303 - 312.