56/2006

Raport Badawczy
Research Report

RB/39/2006

Estimation of Tolerance Relation on the Basis of Multiple Pairwise Comparisons with Random Errors

Leszek 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) 8373578

fax: (+48) (22) 8372772

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

Warszawa 2006

ESTIMATION OF TOLERANCE RELATION ON THE BASIS OF MULTIPLE PAIRWISE COMPARISONS WITH RANDOM ERRORS

by

Leszek Klukowski

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

Abstract: The methods of tolerance relation estimation on the basis of pairwise comparisons with random errors, in the case of multiple comparisons of each pair are proposed in the paper. Each comparison expresses the number of common features of both elements or the number of their missing features. The assumptions made about distributions of comparisons errors are very weak, especially they may be unknown. Two approaches are discussed: the first one based on averaging of each pair comparisons and the second based on the median from comparisons. Results of estimation are obtained on the basis of discrete programming tasks. The properties of estimators are based on some probabilistic inequalities. An example of application of the estimators proposed is presented too.

Keywords: tolerance relation estimation, multiple pairwise comparisons, comparisons expressing the number of common features

1. Introduction

The tolerance relation is a relaxed form of the equivalence relation, i.e. without transitivity property. It divides a set of elements into a family of subsets with at least one non-empty intersection. The relation is a model of many real-life phenomena, e.g. analysis of marketing data (purchasing patterns of customers, when comparisons are applied to some number of independent purchases of each customer and the number of patterns in not known); some other example - analysis of empirical functions shapes - is presented in Klukowski (2006).

The methods of tolerance relation estimation presented in the paper are extensions of the approach introduced in Klukowski (2002) section 4, for the case of $N>1$ independent comparisons. The methods exploit the idea of *nearest adjoining order* introduced by Slater (1961) for the preference relation (see also David (1988)). Two approaches are examined in the paper: the first one - based on averaged comparisons of each pair and the second - based on the median from comparisons. In both cases it is assumed, that each comparison expresses the number of subsets of an intersection including both elements (in other words a number of common features of both elements) or the number of subsets, which do not comprise both elements (a number of missing features of both elements). The estimated form of the relation is obtained on the basis of optimal solution of some discrete optimization problems. The properties of the estimators proposed are based on well known probabilistic inequalities: Hoeffding (see Hoeffding (1963)), Chebyshev (for expected value) and properties of the order statistics (see David (1970)). An important property of the first estimator - based on the averaging approach – is the fact, that the probability of errorless estimation result converges exponentially to one, for $N\to\infty$. The effectiveness of the median approach (evaluation of the probability of errorless result) is worse than those corresponding to the averaging approach, but optimization problem for this case is easier to solve. Empirical experience and some asymptotic properties of the median indicate, that the advantage may be not significant for some type of distributions of comparisons errors. For both approaches it is possible to obtain some approximations of the probability of errorless solution obtaining in the case of unknown distributions of comparisons errors.

The papers consists of six sections. The second section presents basic definitions, assumptions and notations. In the third section the averaging approach is examined. The fourth section presents the median approach and an algorithm for determination the probability function of the median. In the fifth section an example of application of both approaches is discussed; the example is based on stochastic simulations. Last section sums up the results presented.

2. Basic definitions, assumptions and notations

It is assumed, that there exists an (unknown) tolerance relation (reflexive, symmetric) in a finite set $\mathbf{X}=\{x_1, \ldots, x_m\}$ $(m \geq 3)$; the relation divides the set $\mathbf{X}$ into a family of subsets $\chi_1^*, \ldots, \chi_n^*$, $1<n<m$, with the following properties:

$$\bigcup_{q=1}^{n} \chi_q^* = \mathbf{X}, \qquad \chi_q^* \neq \{\varnothing\}, \qquad \exists\, q, s\ (q \neq s)\colon\ \chi_q^* \cap \chi_s^* \neq \{\varnothing\}. \tag{1}$$

Moreover, in purpose to avoid "degenerated" form of the relation it is assumed additionally, that in each subset $\chi_q^* \subset \mathbf{X}$ there exists an element x_i, which belong to the subset χ_q^* only, i.e.: $x_i \in \chi_q^*$ and $x_i \notin \chi_s^*$ for $s \neq q$.

The basis for further considerations are two functions $T_1(\cdot)$ and $T_2(\cdot)$, defined as follows $T_1 : \mathbf{X} \times \mathbf{X} \rightarrow D$, $T_2 : \mathbf{X} \times \mathbf{X} \rightarrow D$, $D = \{0, 1, \ldots, n\}$, where:

$$T_1(x_i, x_j) = \#(\Omega_i^* \cap \Omega_j^*), \tag{2}$$

$$T_2(x_i, x_j) = \#(\Psi_i^* \cap \Psi_j^*), \tag{3}$$

where:

Ω_i^* - the set of the form $\Omega_i^* = \{s \mid x_i \in \chi_s^*\}$,

Ψ_i^* - the set of the form $\Psi_i^* = \{1, \ldots, n\} - \Omega_i^*$,

$\#(\Xi)$ - number of elements of the set Ξ.

Under the assumption about "non-degeneration" of the relation, each function $T_1(\cdot)$ and $T_2(\cdot)$ characterize the relation form.

If an element x_i is included in some subset χ_q^*, it can be interpreted, that it posses some feature; if it is included in a conjunction $\bigcap_{q \in R} \chi_q^*$, then the element posses some set of features. Thus, the function $T_1(\cdot)$ express the number of common features of elements x_i and

x_j, while the function $T_2(\cdot)$ express the number of lacking features of both elements, from the features existing in the set **X**.

It is assumed, that the basis for estimation of the relation are results of comparisons $g_k^{(1)}(x_i, x_j)$ or/and $g_k^{(2)}(x_i, x_j)$ ($1 \le k \le N$; $(x_i, x_i) \in \mathbf{X} \times \mathbf{X}$; $j \ne i$), corresponding to the form of the functions $T_1(\cdot)$ and $T_2(\cdot)$ respectively. The comparisons $g_k^{(f)}(x_i, x_j)$ observed instead of (unknown) values $T_f(x_i, x_j)$ are disturbed with random errors; they can be obtained, as a result of an application of statistical tests, expert opinions or other decisions functions.

The comparisons are defined in the following way:

$$g_k^{(1)}(x_i, x_j) = d_{ijk}^{(1)}, \qquad d_{ijk}^{(1)} \in D, \tag{4}$$

$$g_k^{(2)}(x_i, x_j) = d_{ijk}^{(2)}, \qquad d_{ijk}^{(2)} \in D, \tag{5}$$

where:

$d_{ijk}^{(f)}$ (f=1, 2) is the assessment of the value $T_f(x_i, x_i)$, obtained in k-th comparison.

The probabilities of random errors of each comparison are determined with the use of the probability function:

$$P(T_f(x_i, x_j) - g_k^{(f)}(x_i, x_j) = l)) = \alpha_{ijk}^{(f)}(l) \qquad ((x_i, x_i) \in \mathbf{X} \times \mathbf{X}; f=1, 2; -n \le l \le n). \tag{6}$$

It is assumed, that comparisons $g_\kappa^{(f)}(x_i, x_j)$ and $g_\iota^{(f)}(x_q, x_s)$ ($\kappa \ne \iota$) are independent, i.e.:

$$P((g_\kappa^{(f)}(x_i, x_j) = d_{ij\kappa}^{(f)}) \cap (g_\iota^{(f)}(x_q, x_s) = d_{qs\iota}^{(f)})) = P(g_\kappa^{(f)}(x_i, x_j) = d_{ij\kappa}^{(f)}) P(g_\iota^{(f)}(x_q, x_s) = d_{qs\iota}^{(f)}) \tag{7}$$

and the probabilities $\alpha_{ijk}^{(f)}(l)$ satisfy the conditions:

$$\sum_{l \le 0} \alpha_{ijk}^{(f)}(l) > \tfrac{1}{2}, \qquad \sum_{l \ge 0} \alpha_{ijk}^{(f)}(l) > \tfrac{1}{2}, \tag{8}$$

$$\left.\begin{array}{ll}\alpha_{ijk}^{(f)}(l) \geq \alpha_{ijk}^{(f)}(l+1), & l \geq 0, \\ \\ \alpha_{ijk}^{(f)}(l) \geq \alpha_{ijk}^{(f)}(l-1), & l \leq 0.\end{array}\right\} \qquad (9)$$

The conditions (8) – (9) guarantee, that: zero is the median of each distribution (on the basis of median definition), each probability function is unimodal and assumes maximum in zero. The expected value of the any comparison error $E(T_f(\cdot)\text{-}g_k^{(f)}(\cdot))$ can differ from zero; it is typical for $T_f(\cdot)=0$ or $T_f(\cdot)=n$.

Both types of comparisons $g_k^{(1)}(\cdot)$ and $g_k^{(2)}(\cdot)$ can be used as a base of estimation of the relation form - separately or simultaneously. In the second case it is assumed, that comparisons $g_k^{(1)}(\cdot)$ and $g_k^{(2)}(\cdot)$ are not correlated, i.e. $Cov(g_k^{(1)}(\cdot), g_k^{(2)}(\cdot))=0$. Correlation of such comparisons means, that their content is similar.

It should be emphasized, that comparisons of different pairs $g_k^{(f)}(x_i, x_j)$ and $g_k^{(f)}(x_r, x_s)$ ($<i, j> \neq <r, s>$, k, f - fixed) are not assumed independent (in stochastic sense).

For simplification of further considerations it is assumed, that the distributions of comparisons $g_k^{(f)}(\cdot)$ are the same for each k ($1 \leq k \leq N$); an extension for the case of different distributions for individual k is not difficult.

Let us define for any tolerance relation $\chi_1,\ldots,\chi_r$ in the set **X**, the following sets of indices $I(\chi_1,\ldots,\chi_r)$ and $J(\chi_1,\ldots,\chi_r)$:

$I(\chi_1,\ldots,\chi_r)=\{<i, j> \mid \exists q, s$ ($q=s$ not excluded) such, that $x_i, x_j \in \chi_q \cap \chi_s$; $j>i\}$, (10)

$J(\chi_1,\ldots,\chi_r)=\{<i, j> \mid$ there do not exist q, s such, that $x_i, x_j \in \chi_q \cap \chi_s$; $j>i\}$. (11)

The set $I(\chi_1,\ldots,\chi_r)$ includes such pairs of indexes $<i, j>$, that there exists an intersection $\chi_q \cap \chi_s$ of some subsets comprising both elements (x_i, x_j); in the case $q=s$ both elements belong to the same subset. The set $J(\chi_1,...,\chi_r)$ includes such pairs (x_i, x_j), that both elements belong to different subsets χ_q, χ_s and the pair do not belong to the intersection $\chi_q \cap \chi_s$.

It is obvious, that:

$$I(\chi_1,...,\chi_r) \cap J(\chi_1,...,\chi_r) = \{\varnothing\} \quad \text{and} \quad I(\chi_1,...,\chi_r) \cup J(\chi_1,...,\chi_r) = \{<i,j> \mid 1 \le i, j \le m; j > i\}. \tag{12}$$

For any relation $\chi_1,...,\chi_r$ in the set **X** the functions $t_1(x_i, x_j)$ and $t_2(x_i, x_j)$ characterizing this relation are defined ($T_f(\cdot)$ relates to the "true" relation $\chi_1^*,...,\chi_n^*$):

$$t_1(x_i, x_j) = \#(\Omega_i \cap \Omega_j), \tag{13}$$

$$t_2(x_i, x_j) = \#(\Psi_i \cap \Psi_j), \tag{14}$$

where:

$$\Omega_i = \{s \mid x_i \in \chi_s\} \text{ and } \Psi_i = \{1, ..., r\} - \Omega_i. \tag{15}$$

The properties of the estimators proposed below are based on the properties of random variables $U_{fij}^{(k)}(\chi_1,...,\chi_r)$ and $W_f^{(k)}(\chi_1,...,\chi_r)$ defined, as follows:

$$U_{fij}^{(k)}(\chi_1,...,\chi_r) = \left| t_f(x_i, x_j) - g_k^{(f)}(x_i, x_j) \right|, \tag{16}$$

$$W_f^{(k)}(\chi_1,...,\chi_r) = \sum_{<i,j> \in I(\chi_1,...,\chi_r) \cup J(\chi_1,...,\chi_r)} U_{fij}^{(k)}(\chi_1,...,\chi_r). \tag{17}$$

For simplification of the notation, the symbols corresponding to the relation $\chi_1^*,...,\chi_n^*$ will be denoted with asterisks (i.e. $U_{fij}^{(k)*}$, I^*, J^*, etc.) while corresponding to any other relation $\tilde{\chi}_1,...,\tilde{\chi}_r$ - with waves, e.g.:

$$U_{fij}^{(k)*} = \left| T_f(x_i, x_j) - g_k^{(f)}(x_i, x_j) \right|, \tag{18}$$

$$\tilde{U}_{fij}^{(k)} = \left| \tilde{T}_f(x_i, x_j) - g_k^{(f)}(x_i, x_j) \right|. \tag{19}$$

It follows from (6) and (16), that the distribution function of each comparison error satisfies the conditions:

$$P(U_{fij}^{(k)*} = l) = \alpha_{ij}^{(f)}(-l) + \alpha_{ij}^{(f)}(l) \quad (l>0). \tag{20}$$

3. The averaging approach

In the case of the averaging approach, the basis for the problem of estimation of the relation are the averages of the random variables $U_{fij}^{(k)}(\chi_1, ..., \chi_r)$, $U_{fij}^{(k)*}$, $\tilde{U}_{fij}^{(k)}$, $W_f^{(k)}(\chi_1, ..., \chi_r)$, $W_f^{(k)*}$ and $\tilde{W}_f^{(k)}$, i.e.: the variables:

$$\overline{U}_{fij}(\chi_1, ..., \chi_r) = \frac{1}{N}\sum_{k=1}^{N} \left|t_f(x_i, x_j) - g_k^{(f)}(x_i, x_j)\right|, \tag{21}$$

$$\overline{U}_{fij}^{*} = \frac{1}{N}\sum_{k=1}^{N} \left|T_f(x_i, x_j) - g_k^{(f)}(x_i, x_j)\right|, \tag{22}$$

$$\overline{\tilde{U}}_{fij} = \frac{1}{N}\sum_{k=1}^{N} \left|\tilde{t}_f(x_i, x_j) - g_k^{(f)}(x_i, x_j)\right|, \tag{23}$$

$$\overline{W}_f^{*} = \sum_{<i,j> \in I^* \cup J^*} \overline{U}_{fij}^{*}, \tag{24}$$

$$\overline{\tilde{W}}_f = \sum_{<i,j> \in \tilde{I} \cup \tilde{J}} \overline{\tilde{U}}_{fij}. \tag{25}$$

The probabilistic properties of the difference: $\overline{W}_f^{*} - \overline{\tilde{W}}_f$ - the basis for the properties of estimation results - are determined on the basis of Hoeffding inequality (see Hoeffding (1963)):

$$P(\sum_{k=1}^{N} Y_i - \sum_{k=1}^{N} E(Y_i) \geq Nt) \leq \exp\{-2Nt^2/(b-a)^2\}, \tag{26}$$

where:

Y_i (i=1, …, N) – independent random variables satisfying the conditions: $P(a \leq Y_i \leq b)=1$,

a, b, t – constants satisfying the conditions: $t>0$, $b>a$.

They are determined in the following

Theorem 1.

The random variables $\overline{W}_f^{*}$ and $\overline{\tilde{W}}_f$, defined in (24) and (25) respectively, satisfy the conditions:

$$E(\overline{W}_f^{*} - \overline{\tilde{W}}_f) < 0, \tag{27}$$

$$P(\overrightarrow{W}_f - \tilde{\overline{W}}_f < 0) \geq 1 - \exp\{-\frac{N(\sum_{T_f(\cdot)\neq\tilde{T}_f(\cdot)} E(|T_f(\cdot) - g_1^{(f)}(\cdot)| - |\tilde{T}_f(\cdot) - g_1^{(f)}(\cdot)|))^2}{2(m-1)^2}\}, \quad (28)$$

where:

$T_f(\cdot)\neq\tilde{T}_f(\cdot)$ is the set of the form $\{<i, j> \mid T_f(x_i, x_j)\neq\tilde{T}_f(x_i, x_j)\}$.

The proof of the inequality (27) for f=1, under assumption, that the distributions of comparisons errors (see (6)) are the same for each k (k=1, ..., N).

The difference: $U_{1ij}^{(k)*} - \tilde{U}_{1ij}^{(k)}$ can be expressed in the following way:

$$U_{1ij}^{(k)*} - \tilde{U}_{1ij}^{(k)} = |T_1(x_i, x_j) - g_k^{(1)}(x_i, x_j)| - |\tilde{T}_1(x_i, x_j) - g_k^{(1)}(x_i, x_j)|. \quad (29)$$

The fact $T_1(\cdot)\neq\tilde{T}_1(\cdot)$ indicates: $T_1(\cdot)>\tilde{T}_1(\cdot)$ or $T_1(\cdot)<\tilde{T}_1(\cdot)$. In the case $T_1(\cdot)>\tilde{T}_1(\cdot)$ each random variable $g_k^{(1)}(\cdot)$ can assume values, which satisfy the conditions:

(i) $g_k^{(1)}(\cdot)\geq T_1(\cdot)$;

(ii) $\tilde{T}_1(\cdot)<g_k^{(1)}(\cdot)< T_1(\cdot)$;

(iii) $g_k^{(1)}(\cdot)\leq\tilde{T}_1(\cdot)$.

For the values $g_k^{(1)}(\cdot)\geq T_1(\cdot)$ (the case (i)) the difference $U_{1ij}^{(k)*} - \tilde{U}_{1ij}^{(k)}$ equals: $-T_1(\cdot)+\tilde{T}_1(\cdot)$; the last value is negative, its probability satisfy the inequality (see (8)): $\sum_{l\leq 0} P(T_1(\cdot) - g_k^{(1)}(\cdot) = l) > ½$. In the case (iii) the difference (29) is equal to: $T_1(\cdot)-\tilde{T}_1(\cdot)>0$ with probability (see (8) and (9)) $\sum_{l\geq T_1(\cdot)-\tilde{T}_1(\cdot)} P(T_1(\cdot) - g_k^{(1)}(\cdot) = l) < ½$. The inequality (ii) indicates $T_1(\cdot)-\tilde{T}_1(\cdot)\geq 2$ and the difference (29) is equal to: $T_1(\cdot)+\tilde{T}_1(\cdot)-2g_k^{(1)}(\cdot)$. Moreover, the values $T_1(\cdot)+\tilde{T}_1(\cdot)-2g_k^{(1)}(\cdot)$ ($\tilde{T}_1(\cdot)<g_k^{(1)}(\cdot)<T_1(\cdot)$) satisfy the condition:

$$-T_1(\cdot)+\tilde{T}_1(\cdot)< T_1(\cdot)+\tilde{T}_1(\cdot)-2g_k^{(1)}(\cdot)< T_1(\cdot)-\tilde{T}_1(\cdot) \quad (30)$$

and assume the values from the set $\{-T_1(\cdot)+\tilde{T}_1(\cdot)+2, \ldots, T_1(\cdot)-\tilde{T}_1(\cdot)-2\}$ with probabilities $P(T_1(\cdot)+\tilde{T}_1(\cdot)-2g_k^{(1)}=\iota)=P(g_k^{(1)}=(T_1(\cdot)+\tilde{T}_1(\cdot)-\iota)/2)$. The expression $T_1(\cdot)+\tilde{T}_1(\cdot)-2g_k^{(1)}(\cdot)$

($\widetilde{T}_1(\cdot)<g_k^{(1)}(\cdot)<T_1(\cdot)$) assumes values placed symmetrically around zero; their probabilities satisfy the conditions:

$$P(T_1(\cdot)+\widetilde{T}_1(\cdot)-2g_k^{(1)}=-\iota)\geq P(T_1(\cdot)+\widetilde{T}_1(\cdot)-2g_k^{(1)}=\iota) \qquad (\iota>0);$$

last inequality results from the fact, that in the case $T_1(\cdot)+\widetilde{T}_1(\cdot)-2g_k^{(1)}(\cdot)=-\iota$ the value of the difference $T_1(\cdot)-g_k^{(1)}(\cdot)$ is smaller (closer to zero), than in the case $T_1(\cdot)+\widetilde{T}_1(\cdot)-2g_k^{(1)}(\cdot)=\iota$.

Assembling the facts concerning the case: $T_1(\cdot)>\widetilde{T}_1(\cdot)$, i.e.:

$$\sum_{l\leq 0}P(T_1(\cdot)-g_k^{(1)}=l)>½, \tag{31}$$

$$\sum_{l\geq T_1(\cdot)-\widetilde{T}_1(\cdot)}P(T_1(\cdot)-g_k^{(1)}=l)<½, \tag{32}$$

$$P(T_1(\cdot)+\widetilde{T}_1(\cdot)-2g_k^{(1)}=-\iota)\geq P(T_1(\cdot)+\widetilde{T}_1(\cdot)-2g_k^{(1)}=\iota) \qquad (\iota>0), \tag{33}$$

one can obtain:

$$E(U_{1ij}^{(k)*}-\widetilde{U}_{1ij}^{(k)}\mid T_1(\cdot)>\widetilde{T}_1(\cdot))<0. \tag{34}$$

The inequality:

$$E(U_{1ij}^{(k)*}-\widetilde{U}_{1ij}^{(k)}\mid T_1(\cdot)<\widetilde{T}_1(\cdot))<0 \tag{35}$$

corresponding to the case $T_1(\cdot)<\widetilde{T}_1(\cdot)$ is proved in similar way.

The inequalities (34) and (35) indicate - for each k (k=1, …, N) - the inequality:

$$E(U_{1ij}^{(k)*}-\widetilde{U}_{1ij}^{(k)})<0; \tag{36}$$

which is sufficient for (27).

Proof of the inequality (28).

The inequality (28) is proved on the basis of Hoeffding inequality (26). The difference: $\overline{W}_f^{*}$ - $\widetilde{\overline{W}}_f$ can be expressed in the following way:

$$\overline{W}_1^{*}-\widetilde{\overline{W}}_1=\frac{1}{N}\sum_{k=1}^{N}\sum_{T_1(x_i,x_j)\neq\widetilde{T}_1(x_i,x_j)}\left(\left|T_1(x_i,x_j)-g_k^{(1)}(x_i,x_j)\right|-\left|\widetilde{T}_1(x_i,x_j)-g_k^{(1)}(x_i,x_j)\right|\right). \tag{37}$$

The probability $P(\overline{W}_1^{*}-\widetilde{\overline{W}}_1<0)$ can be expressed in the form:

$$P(\overline{W}_1^{*} - \tilde{\overline{W}}_1 < 0) = 1 - P(\overline{W}_1^{*} - \tilde{\overline{W}}_1 \geq 0). \tag{38}$$

The probability $P(\overline{W}_1^{*} - \tilde{\overline{W}}_1 \geq 0)$ can be evaluated in the following way. It follows from (29), that:

$$P(\overline{W}_1^{*} - \tilde{\overline{W}}_1 \geq 0) =$$

$$= P\Big(\frac{1}{N}\sum_{k=1}^{N} \sum_{T_1(x_i,x_j)\neq \tilde{T}_1(x_i,x_j)} \big|T_1(x_i,x_j) - g_k^{(1)}(x_i,x_j)\big| - \big|\tilde{T}_1(x_i,x_j) - g_k^{(1)}(x_i,x_j)\big| \geq 0\Big). \tag{39}$$

Introducing the notations:

$$D_k^{(1)}(x_i, x_j) = \big|T_1(x_i,x_j) - g_k^{(1)}(x_i,x_j)\big| - \big|\tilde{T}_1(x_i,x_j) - g_k^{(1)}(x_i,x_j)\big| \tag{40}$$

one can express the probability (39) in the form:

$$P(\overline{W}_1^{*} - \tilde{\overline{W}}_1 \geq 0) = P\Big(\frac{1}{N}\sum_{k=1}^{N} \sum_{T_f(\cdot)\neq \tilde{T}_f(\cdot)} D_k^{(1)}(\cdot) \geq 0\Big) =$$

$$= P\Big(\frac{1}{N}\sum_{k=1}^{N} \sum_{T_1(\cdot)\neq \tilde{T}_1(\cdot)} D_k^{(1)}(\cdot) - \frac{1}{N}\sum_{k=1}^{N} \sum_{T_1(\cdot)\neq \tilde{T}_1(\cdot)} E(D_k^{(1)}(\cdot)) \geq -\frac{1}{N}\sum_{k=1}^{N} \sum_{T_1(\cdot)\neq \tilde{T}_1(\cdot)} E(D_k^{(1)}(\cdot))\Big) =$$

$$= P\Big(\sum_{k=1}^{N} \sum_{T_1(\cdot)\neq \tilde{T}_1(\cdot)} D_k^{(1)}(\cdot) - N \sum_{T_1(\cdot)\neq \tilde{T}_1(\cdot)} E(D_1^{(1)}(\cdot)) \geq N\big(-\sum_{T_1(\cdot)\neq \tilde{T}_1(\cdot)} E(D_1^{(1)}(\cdot))\big)\Big). \tag{41}$$

The probability (41) can be evaluated on the basis of the inequality (26), in the following way:

$$P\Big(\sum_{k=1}^{N} \sum_{T_1(\cdot)\neq \tilde{T}_1(\cdot)} D_k^{(1)}(\cdot) - N \sum_{T_1(\cdot)\neq \tilde{T}_1(\cdot)} E(D_1^{(1)}(\cdot)) \geq N\big(-\sum_{T_1(\cdot)\neq \tilde{T}_1(\cdot)} E(D_1^{(1)}(\cdot))\big)\Big) \leq \exp\Big\{-\frac{2N\big(\sum_{T_1(\cdot)\neq \tilde{T}_1(\cdot)} E(D_1^{(1)}(\cdot))\big)^2}{(2(m-1))^2}\Big\}. \tag{42}$$

The expression in exponent results from the fact, that: each value $D_1^{(1)}(x_i,x_j)$ satisfy the condition

$-(m-1) \leq D_1^{(1)}(x_i,x_j) \leq m-1$ (because $n<m$ and therefore the number of subsets generating any conjunction in the tolerance relation cannot exceed m-1), the expected values $E(D_k^{(1)}(x_i,x_j))$ are equal for $1 \leq k \leq N$ and equal to $E(D_1^{(1)}(x_i,x_j))$; the last component $\sum_{T_1(\cdot)\neq \tilde{T}_1(\cdot)} E(D_1^{(1)}(x_i,x_j))$ is negative and therefore the term: $-\sum_{T_1(\cdot)\neq \tilde{T}_1(\cdot)} E(D_1^{(1)}(x_i,x_j))$ is positive. The inequality (42) is equivalent to the proved inequality (28). The proof for f=2 is similar.

□

The inequality (27) shows, that the expected value of the random variable $\overline{W}_f^{*}$ is less, than expected value of any other variable $\tilde{\overline{W}}_f$. Moreover, the evaluation (28) indicates, that probability $P(\overline{W}_f^{*}<\tilde{\overline{W}}_f)$ exceeds or is equal to the right hand side of the inequality (28). Thus, it is rational to estimate the relation $\chi_1^{*},...,\chi_n^{*}$ with the relation $\hat{\chi}_1, ..., \hat{\chi}_{\hat{n}}$, which minimizes the value of the random variable $\overline{W}_f(\chi_1, ..., \chi_n)$. It is meaningful, that the evaluation of the lower bound of the probability $P(\overline{W}_f^{*}<\tilde{\overline{W}}_f)$ converges exponentially to zero, for $N\rightarrow\infty$. In the case of non-identical distributions of comparisons errors (for different k) the expected value $E(D_1^{(1)}(x_i,x_j))$ have to be replaced with $\min_k\{E(D_k^{(1)}(x_i,x_j))\}$. The probability $P(\overline{W}_1^{*}-\tilde{\overline{W}}_1<0)$ can be also evaluated with the use of other probabilistic inequalities.

The estimated form $\hat{\chi}_1, ..., \hat{\chi}_{\hat{n}}$ of the relation $\chi_1^{*},...,\chi_n^{*}$ can be obtained on the basis of the solution of optimization tasks:

$$\min_{F_X}\Big[\sum_{k=1}^{N}\sum_{X\times X}\left|t_f(x_i,x_j)-g_k^{(f)}(x_i,x_j)\right|\Big], \qquad (f=1 \text{ or } 2) \qquad (43)$$

or

$$\min_{F_X}\Big[\sum_{k=1}^{N}\sum_{X\times X}\Big(\left|t_1(x_i,x_j)-g_k^{(1)}(x_i,x_j)\right|+\left|t_2(x_i,x_j)-g_k^{(2)}(x_i,x_j)\right|\Big)\Big], \qquad (44)$$

where:

F_X – the feasible set of the problem (the set including all tolerance relations satisfying the conditions (1) and "non-degeneration" condition)).

The feasible set of each problem (43) and (44) is finite and the optimal solution always exist; however the number of solutions of each task may exceed one. In the case of multiple solutions the evaluation (28) relates to whole set of solutions (estimates).

The evaluation of the probability of errorless solution obtaining (28) can be determined in the case of known probability distributions of the comparisons errors. In opposite case, it is possible to determine some approximations of the evaluation. As the basis of the approximation can be used:

- the estimated form of the relation $\hat{\chi}_1, ..., \hat{\chi}_{\hat{n}}$ (it allows to determine the estimates $\hat{T}_f(\cdot)$ and $\hat{n}$), the formulas (31) – (33) together with the conditions (8) – (9) or
- the estimated form of the probability functions $\alpha_{ij}^{(f)}(l)$ obtained on the basis of comparisons $g_1^{(f)}(\cdot), ..., g_N^{(f)}(\cdot)$.

The first approach can be used for any value of N; however for N close to one such approximation may be of rough type. The second approach requires – for purpose of realistic estimates – an appropriate number of comparisons N (N>>n).

Let us notice, that the right-hand side of the inequality (28) is based on the constraint -(m-1)$\leq D_1^{(1)}(x_i, x_j) \leq$($m$-1). Typically the value ±(m-1) is excessive (significantly greater, than n); especially in the case m-1> $\max_{X \times X}\{T_1(x_i, x_j)\}$ the constraint ±(m-1) negatively influences (decreases) the evaluation (28). Therefore, it is rational to replace the value m–1 with the estimate $\hat{n}$ or $\max_{X \times X}\{\hat{T}_1(x_i, x_j)\}$.

4. The median approach

In the case of median approach the basis for estimation are the medians from comparisons of each pair and it is assumed, that N=2τ+1 (τ=1, ...,). More precisely, each set of comparisons $g_1^{(f)}(x_i, x_j)$, ... , $g_N^{(f)}(x_i, x_j)$ ($(x_i, x_j) \in \mathbf{X} \times \mathbf{X}$) is replaced with their median $g_{me,N}^{(f)}(x_i, x_j)$ and the variables $U_{fij}^{(k)}(\chi_1, ..., \chi_r)$, $U_{fij}^{(k)*}$, $\tilde{U}_{fij}^{(k)}$, $W_f(\chi_1, ..., \chi_r)$, W_f^*, $\tilde{W}_f$ (f=1, 2) are replaced - respectively - with the variables:

$$U_{fij}^{(me,N)}(\chi_1, ..., \chi_r) = \left| t_f(x_i, x_j) - g_{me,N}^{(f)}(x_i, x_j) \right| \tag{45}$$

$$U_{fij}^{(me,N)*} = \left| T_f(x_i, x_j) - g_{me,N}^{(f)}(x_i, x_j) \right|, \tag{46}$$

$$\widetilde{U}_{fij}^{(me,N)} = \left| \widetilde{T}_f(x_i, x_j) - g_{me,N}^{(f)}(x_i, x_j) \right|, \tag{47}$$

$$W_f^{(me,N)*} = \sum_{<i,j> \in I^* \cup J^*} U_{fij}^{(me,N)*}, \tag{48}$$

$$\widetilde{W}_f^{(me,N)} = \sum_{<i,j> \in \widetilde{I} \cup \widetilde{J}} \widetilde{U}_{fij}^{(me,N)}. \tag{49}$$

where:

$g_{me,N}^{(f)}(x_i, x_j)$ - the median from comparisons $g_1^{(f)}(x_i, x_j)$, ..., $g_N^{(f)}(x_i, x_j)$, i.e. the $\frac{N+1}{2}$-th order statistics $g_{((N+1)/2)}^{(f)}(x_i, x_j)$ ($g_{(1)}^{(f)}(x_i, x_j)$, ..., $g_{(N)}^{(f)}(x_i, x_j)$ - non-decreasing ordered results of comparisons).

4.1. The form of the estimator and its properties

This problem considered in this point is similar to the single comparison case. However, the probability functions of the medians $g_{me,N}^{(f)}(x_i, x_j)$ from comparisons $(x_i, x_j) \in \mathbf{X} \times \mathbf{X}$ are not the same, as the probability functions of individual comparisons $g_k^{(f)}(x_i, x_j)$ ($1 \le k \le N$); therefore the properties of the tolerance relation estimated on the basis of the medians are also not the same, as in the single comparison case. The properties of the estimator based on medians are presented in the following

Theorem 2.

The random variables $W_f^{(me,N)*}$ and $\widetilde{W}_f^{(me,N)}$ defined in (48) and (49) satisfy the conditions:

$$E(W_f^{(me,N)*} - \widetilde{W}_f^{(me,N)}) < 0, \tag{50}$$

$$P(W_f^{(me,N)*} < \widetilde{W}_f^{(me,N)}) \ge -\frac{1}{\nu(m-1)} E\Big(\sum_{T_1(\cdot) \ne \widetilde{T}_1(\cdot)} \left| T_1(x_i, x_j) - g_{me,N}^{(1)}(x_i, x_j) \right| - \left| \widetilde{T}_1(x_i, x_j) - g_{me,N}^{(1)}(x_i, x_j) \right| \Big), \tag{51}$$

where:

ν - the number of elements of the set $\{(x_i, x_j) \in \mathbf{X} \times \mathbf{X} \mid T_1(x_i, x_j) \ne \widetilde{T}_1(x_i, x_j); j > i\}$.

Proof of the inequality (50) for f=1, assuming the same distributions $g_k^{(1)}(\cdot)$ for each k (k=1, ..., N).

The inequality (50) is true for N=1 (it results from the Theorem 1, for N=1). For $N=2\tau+1$ (τ=1, ...,) it can be shown, that the probability function $P(T_1(x_i,x_j)-g_{me,N}^{(1)}(x_i,x_j))=l)$ ($N=2\tau+1$; τ=0, 1, ...,) satisfies for each pair $(x_i, x_j)\in \mathbf{X}\times\mathbf{X}$ the inequalities:

$$P(T_1(x_i,x_j)-g_{me,N+2}^{(1)}(x_i,x_j)=0)>P(T_1(x_i,x_j)-g_{me,N}^{(1)}(x_i,x_j)=0)\,; \tag{52a}$$

$$P(T_1(x_i,x_j)-g_{me,N+2}^{(1)}(x_i,x_j)=l)<P(T_1(x_i,x_j)-g_{me,N}^{(1)}(x_i,x_j)=l) \qquad (l\neq 0). \tag{52b}$$

The inequalities (52a) and (52b) result from the following facts. The probabilities: $P(T_1(x_i,x_j)-g_{me,N}^{(1)}(x_i,x_j)=l)$ can be expressed in the form (see David (1970), section 2.4):

$$P(T_1(x_i,x_j)-g_{me,N}^{(1)}(x_i,x_j)=0)=$$

$$=P(T_1(x_i,x_j)-g_{me,N}^{(1)}(x_i,x_j)\leq 0)\ -P(T_1(x_i,x_j)-g_{me,N}^{(1)}(x_i,x_j)\leq -1)=$$

$$=\frac{N!}{(((N-1)/2)!)^2}\int_{G(-1)}^{G(0)} t^{(N-1)/2}(1-t)^{(N-1)/2}dt\,, \tag{53a}$$

$$P(T_1(x_i,x_j)-g_{me,N}^{(1)}(x_i,x_j)=l)=$$

$$=P(T_1(x_i,x_j)-g_{me,N}^{(1)}(x_i,x_j)\leq l)\ -P(T_1(x_i,x_j)-g_{me,N}^{(1)}(x_i,x_j)\leq l-1)=$$

$$=\frac{N!}{(((N-1)/2)!)^2}\int_{G(l-1)}^{G(l)} t^{(N-1)/2}(1-t)^{(N-1)/2}dt\,, \tag{53b}$$

where:

$$G(l)=P(T_1(x_i,x_j)-g_k^{(1)}(x_i,x_j)\leq l)\,.$$

The expressions (53a) and (53b) are determined on the basis of beta distribution $B(p, q)$, with parameters $p=q=(N+1)/2$. The expected value and variance of the distribution assume the form – respectively: ½ and $((N+1)/2)^2/(N+1)^2(N+2)=1/4(N+2)$. The variance of the distribution converges to zero for $N\to\infty$ and the integrand in integrals (53a), (53b) is symmetric around ½. These facts guarantee, that: the distributions of the random variables:

$T_1(x_i,x_j)-g^{(1)}_{me,N}(x_i,x_j)$ $((x_i, x_j)\in\mathbf{X}\times\mathbf{X})$ are for each N unimodal, their probability functions assume maximum in zero (i.e. for $T_1(x_i,x_j)-g^{(1)}_{me,N}(x_i,x_j)=0$) and satisfy the inequalities (52a), (52b). Last two conditions are sufficient (see the assumptions (8), (9) and inequality (27) from the theorem 1) for the inequality (50).

Proof of the inequality (51).

Let us introduce the notations similar to those in the Theorem 1:

$$D^{(1)}_{me}(x_i, x_j)=|T_1(x_i,x_j)-g^{(1)}_{me,N}(x_i,x_j)|-|\tilde{T}_1(x_i,x_j)-g^{(1)}_{me,N}(x_i,x_j)|\,. \qquad (54)$$

Thus, the difference (54) can be expressed in the form:

$$P(W_1^{(me,N)*}<\tilde{W}_1^{(me,N)})=1-P(W_1^{(me,N)*}-\tilde{W}_1^{(me,N)}\geq 0)$$

and the probability $P(W_1^{(me,N)*}-\tilde{W}_1^{(me,N)}\geq 0)$ can be evaluated on the basis of Chebyshev inequality for expected value, in the following way:

$$P(W_1^{(me,N)*}-\tilde{W}_1^{(me,N)}\geq 0)=P(\sum_{T_1(\cdot)\neq\tilde{T}_1(\cdot)} D^{(1)}_{me}(\cdot)\geq 0)=$$

$$=P(\sum_{T_1(\cdot)\neq\tilde{T}_1(\cdot)}(D^{(1)}_{me}(\cdot)+m-1)\geq \nu(m-1)) \qquad (55)$$

(ν - the number of components of the sum $\sum_{T_j(\cdot)\neq\tilde{T}_j(\cdot)} D^{(1)}_{me}(\cdot)$).

The probability (55) can be evaluated with the use the Chebyshev inequality:

$$P(\sum_{T_1(\cdot)\neq\tilde{T}_1(\cdot)}(D^{(1)}_{me}(\cdot)+m-1)\geq \nu(m-1))\leq\frac{1}{\nu(m-1)}E(\sum_{T_1(\cdot)\neq\tilde{T}_1(\cdot)}(D^{(1)}_{me}(\cdot)+(m-1)))=1+\frac{1}{\nu(m-1)}E(\sum_{T_1(\cdot)\neq\tilde{T}_1(\cdot)} D^{(1)}_{me}(\cdot))\,. \qquad (56)$$

The last expression in (56) is equal to the right-hand side of the in equality (51).

The proof for $f=2$ is similar.

□

The expression $\frac{1}{\nu(m-1)}E(\sum_{T_1(\cdot)\neq\tilde{T}_1(\cdot)} D^{(1)}_{me,N}(\cdot))$ (in the right-hand side of the equality (51)) is not positive, more precisely – it is included in the interval (-1, 0). Its numerical value can be determined in the case of known distributions of comparison errors $P(T_1(\cdot)-g^{(1)}_{me,N}(\cdot))$. In opposite case they can be approximated in some way. The approximation procedure based on the relationships (53a), (53b) (see David (1970), section 2.4) and some additional assumptions (quasi-uniform distribution with symmetry of tails) is proposed in the point 4.2 below.

The evaluation (56) is typically significantly weaker, than (28). However, some asymptotic properties of the estimator based on the medians can be determined too. They result from the properties of beta distribution for $N \to \infty$ (see relationships (53a), (53b)). They indicate, that the median $g^{(1)}_{me,N}(\cdot)$ converges in stochastic sense to $T_1(\cdot)$, i.e. for any $\varepsilon > 0$ it is valid: $\lim_{N\to\infty} P(|g^{(1)}_{me,N}(\cdot) - T_1(\cdot)| > \varepsilon) = 0$ and the difference $E(W_1^{(me,N)*}) - E(\tilde{W}_1^{(me,N)})$ converges to some negative value. The speed of the convergences is the problem for future investigations.

The right-hand side of the inequality (51) is based on the fact, that $-(m-1) \leq D^{(1)}_{me}(x_i, x_j) \leq m-1$. Such constraint is typically (i.e. for $m-1 > \max_{X \times X}\{T_1(x_i, x_j)\}$) too excessive. Therefore, it is rational to replace the value m-1 (in the right-hand side of inequality (51)) with the estimate $\hat{n}$.

The optimization problems for the median approach, are similar to those formulated for the case of single comparison of each pair (see Klukowski (2002)), with that difference, that individual comparison $g_k^{(f)}(x_i, x_j)$ ($(x_i, x_j) \in \mathbf{X} \times \mathbf{X}$) is replaced with the median $g^{(f)}_{me,N}(x_i, x_j)$ from N comparisons:

$$\min_{F_X}[\sum_{X \times X} |t_f(x_i, x_j) - g^{(f)}_{me,N}(x_i, x_j)|], \qquad (f=1 \text{ or } 2) \tag{57}$$

or

$$\min_{F_X}[\sum_{X \times X}(|t_1(x_i, x_j) - g^{(1)}_{me,N}(x_i, x_j)| + |t_2(x_i, x_j) - g^{(2)}_{me,N}(x_i, x_j)|)] \tag{58}$$

(F_X – the feasible set of the problem).

The problems (57) and (58) are simpler to solve in comparison to the problems (43) and (44); the number of solutions may exceed one.

4.2. The procedure for an approximation of the distribution function $P(T_1(x_i, x_j)- g^{(1)}_{me,N}(x_i, x_j)=l)$

The approximation procedure proposed in this point is especially useful for moderate N, namely N=5, 7, 9, 11; for N>10 the Gaussian approximation can be also used (see David (1970), point 2.5).

The procedure is based on: some kind of "upper bound" distribution, the formulas (53a, b) and the estimated form of the relation. The "upper bound" distribution (a kind of evaluation) is obtained on the basis of: the conditions (8) – (9), some quasi-uniform (discrete) distribution and an assumption, that the values of tails of comparisons errors are symmetric, i.e. $P(T_f(\cdot)- g_k^{(f)}(\cdot)<0)=P(T_f(\cdot)- g_k^{(f)}(\cdot)>0)$ - with except of extreme values of $T_f(\cdot)$ (minimum and maximum). Estimated form of the relation, i.e. $\hat{\chi}_1, ..., \hat{\chi}_{\hat{n}}$, allows to determine the values $\hat{T}_f(x_i,x_j)$ $((x_i, x_j)\in \mathbf{X}\times\mathbf{X})$ and $\hat{n}$. The estimates can be also used for determination the extreme values $\hat{T}_f(x_i,x_j)$ and the sets of admissible values (range) of comparisons $g_k^{(f)}(x_i,x_j)$. It is suggested to determine the range of comparisons in the following way: to assume minimum equal to zero and maximum equal to $\hat{n}$. The minimum is natural – because any comparison result cannot be negative. The maximum can be assumed in many ways, e.g.: $\max\{\hat{T}_f(x_i,x_j) \mid (x_i, x_j)\in \mathbf{X}\times\mathbf{X}\}$ or $\hat{n}$ or m-1. The "compromise" value is the estimate $\hat{n}$, because $\max\{\hat{T}_f(x_i,x_j) \mid (x_i, x_j)\in \mathbf{X}\times\mathbf{X}\} \leq \hat{n} \leq m-1$. The assumptions about symmetry of tails and quasi–uniform distribution of each tail allow to determine the distributions completely. Symmetry of tails is not unrealistic, because zero is the median of each distribution of comparison error $T_f(\cdot)-g_k^{(f)}(\cdot)$. The relationships (53a, b) allows to determine of the distributions functions of medians of comparisons errors for N>1.

The quasi-uniform distribution is constructed for f=1 in the following way. The estimates $\hat{T}_1(\cdot)$ and $\hat{n}$ are used instead of the actual values $T_1(\cdot)$ and n (i.e. they are assumed constants, not realizations of the random variables). The probabilities $P(T_f(\cdot)- g_k^{(f)}(\cdot)<0)$ and

$P(T_f(\cdot) - g_k^{(f)}(\cdot) > 0)$ are assumed equal (for $\hat{T}_f(\cdot) \neq 0$ and $\hat{T}_f(\cdot) \neq \hat{n}$); the probabilities $P(T_f(\cdot) - g_k^{(f)}(\cdot) = -l)$ are assumed equal for each (integer) $l>0$ and the probabilities $P(T_f(\cdot) - g_k^{(f)}(\cdot) = l)$ are assumed equal for each (integer) $l>0$ (quasi-uniform distribution). For the case: $\hat{T}_1(\cdot) \neq 0$, $\hat{T}_1(\cdot) \neq \hat{n}$, $\hat{n}>2$, $\hat{n}$ - odd and $\hat{T}_1(\cdot) < \hat{n}/2$ the "upper bound" distribution of comparisons errors is obtained for each pair (x_i, x_j) from the system of equations:

$$
\left.
\begin{aligned}
&P_b(\hat{T}_1(\cdot) - g_k^{(1)}(\cdot) = \hat{T}_1(\cdot) - \hat{n}) = \ldots = P_b(\hat{T}_1(\cdot) - g_k^{(1)}(\cdot) = -1), && (59)\\
&P_b(\hat{T}_1(\cdot) - g_k^{(1)}(\cdot) = 1) = \ldots = P_b(\hat{T}_1(\cdot) - g_k^{(1)}(\cdot) = \hat{T}_1(\cdot)), && (60)\\
&P_b(\hat{T}_1(\cdot) - g_k^{(1)}(\cdot) < 0) = P_b(\hat{T}_1(\cdot) - g_k^{(1)}(\cdot) > 0), && (61)\\
&P_b(\hat{T}_1(\cdot) - g_k^{(1)}(\cdot) = 0) = P_b(\hat{T}_1(\cdot) - g_k^{(1)}(\cdot) = 1), && (62)\\
&\sum_{l=\hat{T}_1(\cdot)-\hat{n}}^{\hat{T}_1(\cdot)} P_b(\hat{T}_1(\cdot) - g_k^{(1)}(\cdot) = l) = 1. && (63)
\end{aligned}
\right\}
$$

In the case $\hat{T}_1(\cdot) > \hat{n}/2$, the equation (62) is replaced with the equation:

$$P_b(\hat{T}_1(\cdot) - g_k^{(1)}(\cdot) = 0) = P_b(\hat{T}_1(\cdot) - g_k^{(1)}(\cdot) = 1) \qquad (64)$$

(the probability $P_b(\hat{T}_1(\cdot) - g_k^{(1)}(\cdot) = 0)$ is equal to $\max\{P_b(\hat{T}_1(\cdot) - g_k^{(1)}(\cdot) = 1), P_b(\hat{T}_1(\cdot) - g_k^{(1)}(\cdot) = -1)\}$).

In the case $\hat{T}_1(\cdot) = 0$ the system assumes the simple form:

$$
\left.
\begin{aligned}
&P_b(\hat{T}_1(\cdot) - g_k^{(1)}(\cdot) = 0) = 1/2,\\
&P_b(\hat{T}_1(\cdot) - g_k^{(1)}(\cdot) = -l) = 1/2\hat{n} \qquad (l=1, \ldots, \hat{n}),
\end{aligned}
\right\} \qquad (65a)
$$

while in the case $\hat{T}_1(\cdot) = \hat{n}$, the second relationship in (65a) is replaced with:

$$P_b(\hat{T}_1(\cdot) - g_k^{(1)}(\cdot) = l) = 1/2\hat{n} \qquad (l=1, \ldots, \hat{n}). \qquad (65b)$$

In the case of even $\hat{n}$ ($\hat{n}$>2) it is necessary to take into account the equity $\hat{T}_1(\cdot)=\hat{n}/2$. In this case the distribution of comparison errors is assumed in the form (an uniform discrete distribution):

$$P_b(\hat{T}_1(\cdot)-g_k^{(1)}(\cdot)=l)=1/(\hat{n}+1) \quad (l=-\hat{n}/2, ..., 0, ..., \hat{n}/2). \tag{66}$$

In the case $\hat{n}$=2, the system assumes the simplest form:

$$\left.\begin{array}{ll} P_b(\hat{T}_1(\cdot)-g_k^{(1)}(\cdot)=0)=1/2\,;\; P_b(\hat{T}_1(\cdot)-g_k^{(1)}(\cdot)=1)=P_b(\hat{T}_1(\cdot)-g_k^{(1)}(\cdot)=2)=1/4\,; & \text{for } \hat{T}_1(\cdot)=2 \\ P_b(\hat{T}_1(\cdot)-g_k^{(1)}(\cdot)=1)=P_b(\hat{T}_1(\cdot)-g_k^{(1)}(\cdot)=0)=P_b(\hat{T}_1(\cdot)-g_k^{(1)}(\cdot)=-1)=1/3\,; & \text{for } \hat{T}_1(\cdot)=1 \\ P_b(\hat{T}_1(\cdot)-g_k^{(1)}(\cdot)=0)=1/2\,;\; P_b(\hat{T}_1(\cdot)-g_k^{(1)}(\cdot)=-1)=P_b(\hat{T}_1(\cdot)-g_k^{(1)}(\cdot)=-2)=1/4\,; & \text{for } \hat{T}_1(\cdot)=0. \end{array}\right\} \tag{67}$$

Each probability function generated by above systems of equations can be considered, as a kind of a "conservative approximation" of the actual distribution function, because any other distribution (based on estimated relation form $\hat{\chi}_1, ..., \hat{\chi}_{\hat{n}}$ and a distribution function with symmetric tails) is more concentrated (its variance is smaller). If there exists some knowledge about asymmetry of tails, then the equation systems (59) – (67) ought to be modified, especially the equality (61).

The distribution function obtained on the basis of the equation systems (59) - (67) allows to use the relationships (53a, b) for determination the "upper bound" approximation of the probability function $P_b^{(me,N)}(T_1(x_i,x_j)-g_{me,N}^{(1)}(x_i,x_j)=l)$ of the median in the following way:

$$P_b^{(me,N)}(T_1(x_i,x_j)-g_{me,N}^{(1)}(x_i,x_j)=l)=$$

$$=P_b^{(me,N)}(T_1(x_i,x_j)-g_{me,N}^{(1)}(x_i,x_j)\le l)\;-P_b^{(me,N)}(T_1(x_i,x_j)-g_{me,N}^{(1)}(x_i,x_j)\le l-1)=$$

$$=\frac{N!}{(((N-1)/2)!)^2}\int_{G_b(l-1)}^{G_b(l)} t^{(N-1)/2}(1-t)^{(N-1)/2}dt\,, \tag{68}$$

where: $G_b(l)=P_b(\hat{T}_1(x_i,x_j)-g_k^{(1)}(x_i,x_j)\le l)$, $\quad G_b(l-1)=P_b(\hat{T}_1(x_i,x_j)-g_k^{(1)}(x_i,x_j)\le l-1)$.

The approach presented above allows to determine some upper bound of the right-hand side of inequality (51).

In the case $N >> \hat{n}$ the upper bound distributions functions can be replaced with estimated distribution functions; especially nonparametric estimators can be used.

5. Example of application of algorithms proposed

A simple (simulated) example of an application of the estimators proposed is considered below. The relation under examination assumes the form $\chi_1^* = \{x_1, x_2, x_3, x_4\}$, $\chi_2^* = \{x_3, x_4, x_5\}$, $\chi_3^* = \{x_4, x_6\}$, $\chi_4^* = \{x_7\}$. Each pair (x_i, x_j) is compared five times (comparisons are independent); the results of comparisons (a result of stochastic simulation) are presented in the Table 1, while the distribution functions of the comparisons are presented in the Table 2. The function $T_1(\cdot)$ assumes the following values:

$T_1(x_1, x_5)=T_1(x_1, x_6)=T_1(x_1, x_7)=T_1(x_2, x_5)=T_1(x_2, x_6)=T_1(x_2, x_7)=T_1(x_5, x_6)=T_1(x_5, x_7)=$
$T_1(x_6, x_7)=0$;

$T_1(x_1, x_2)=T_1(x_1, x_3)=T_1(x_1, x_4)=T_1(x_2, x_3)=T_1(x_2, x_4)=T_1(x_3, x_5)=T_1(x_4, x_5)=T_1(x_4, x_6)=1$;

$T_1(x_3, x_4)=2$.

5.1. The algorithm based on averaging approach

The estimated form of the relation $\chi_1^*, \ldots, \chi_4^*$ is obtained on the basis of the optimisation task (43), for $f=1$. It assumes the form $\hat{\chi}_1 = \{x_1, x_2, x_3, x_4\}$, $\hat{\chi}_2 = \{x_3, x_4, x_5\}$, $\hat{\chi}_3 = \{x_4, x_6\}$, $\hat{\chi}_4 = \{x_7\}$, i.e. is the same, as the errorless one; therefore $\hat{n}=n=4$. The minimal value of the function (43) equals 23, the solution is not multiple. The evaluation of the probability (28) is determined for the relation $\tilde{\chi}_1 = \{x_1, x_2, x_3, x_4\}$, $\tilde{\chi}_2 = \{x_1, x_3, x_4, x_5\}$, $\tilde{\chi}_3 = \{x_1, x_4, x_6\}$, $\tilde{\chi}_4 = \{x_7\}$ - similar to errorless one. The difference between the relations $\chi_1^*, \ldots, \chi_4^*$ and $\tilde{\chi}_1, \ldots, \tilde{\chi}_4$ concerns the element x_1; in errorless form of the relation it belongs (exclusively) to the set χ_1^*, while in the relation $\tilde{\chi}_1, \ldots, \tilde{\chi}_4$ it belongs to the intersection $\bigcap_{r=1}^{3} \tilde{\chi}_r$.
The value of the function (43) corresponding to the relation $\tilde{\chi}_1, \ldots, \tilde{\chi}_4$ equals 41. The

inequalities $T_1(\cdot)\neq\tilde{T}_1(\cdot)$ appear for the pairs: (x_1, x_3), (x_1, x_4), (x_1, x_5), (x_1, x_6); the values $\tilde{T}_1(\cdot)$ for these pairs are equal: $\tilde{T}_1(x_1,x_3)=2$, $\tilde{T}_1(x_1,x_4)=3$, $\tilde{T}_1(x_1,x_5)=1$, $\tilde{T}_1(x_1,x_6)=1$.

The evaluation (28) requires the probabilities functions of comparisons errors and the values $T_1(x_i,x_j)$ ($(x_i, x_j)\in\mathbf{X}\times\mathbf{X}$). In the case of unknown distributions and N=5 it is rational the use the approximation of probabilities functions $P_b(\hat{T}_1(\cdot)-g_k^{(1)}(\cdot)=l)$, described in the point 4.2 (see (59) – (67)). For the pair (x_1, x_3) the system of equations assumes the form (the distribution functions $P(g_k^{(1)}(\cdot)=i)$ for all pairs satisfying the inequality $T_1(\cdot)\neq\tilde{T}_1(\cdot)$ are presented in the Table 3):

$$\left.\begin{aligned}&P_b(\hat{T}_1(x_1,x_3)-g_k^{(1)}(x_1,x_3)=-3)=P_b(\hat{T}_1(x_1,x_3)-g_k^{(1)}(x_1,x_3)=-2)=P_b(\hat{T}_1(x_1,x_3)-g_k^{(1)}(x_1,x_3)=-1),\\&\sum_{l=1}^{3}P_b(\hat{T}_1(x_1,x_3)-g_k^{(1)}(x_1,x_3)=-l)=P_b(\hat{T}_1(x_1,x_3)-g_k^{(1)}(x_1,x_3)=1),\\&P_b(\hat{T}_1(x_1,x_3)-g_k^{(1)}(x_1,x_3)=0)=P_b(\hat{T}_1(x_1,x_3)-g_k^{(1)}(x_1,x_3)=1),\\&\sum_{l=-3}^{1}P_b(\hat{T}_1(x_1,x_3)-g_k^{(1)}(x_1,x_3)=l)=1.\end{aligned}\right\}$$

(69a)

The solution of the above system assumes the form:

$$\left.\begin{aligned}&P_b(\hat{T}_1(x_1,x_3)-g_k^{(1)}(x_1,x_3)=-3)=P_b(\hat{T}_1(x_1,x_3)-g_k^{(1)}(x_1,x_3)=-2)=P_b(\hat{T}_1(x_1,x_3)-g_k^{(1)}(x_1,x_3)=-1)=1/9,\\&P_b(\hat{T}_1(x_1,x_3)-g_k^{(1)}(x_1,x_3)=0)=P_b(\hat{T}_1(x_1,x_3)-g_k^{(1)}(x_1,x_3)=1)=1/3.\end{aligned}\right\}$$

(69b)

The expected value $E(D_{kb}^{(1)}(x_1, x_3))$ corresponding to above "upper bound" distribution is equal:

$$E(D_{kb}^{(1)}(x_1, x_3))=E_b(|\hat{T}_1(x_1,x_3)-g_k^{(1)}(x_1,x_3)|)-E_b(|\tilde{T}_1(x_1,x_3)-g_k^{(1)}(x_1,x_3)|)=$$

$$\tfrac{1}{3}(|1-0|+|1-1|)+\tfrac{1}{9}(|1-2|+|1-3|+|1-4|)-\tfrac{1}{3}(|2-0|+|2-1|)-\tfrac{1}{9}(|2-2|+|2-3|+|2-4|)=-\tfrac{1}{3}.$$

Table 1. The results of comparisons (simulation)

Pair (i, j)	$g_{(1)}^{(1)}(\cdot)$	$g_{(2)}^{(1)}(\cdot)$	$g_{(3)}^{(1)}(\cdot)$	$g_{(4)}^{(1)}(\cdot)$	$g_{(5)}^{(1)}(\cdot)$	$g_{me}^{(1)}(\cdot)$	$\frac{1}{5}\sum_{k=1}^{5} g_k^{(1)}(\cdot)$
(1, 2)	1	1	1	1	1	1	1
(1, 3)	1	1	2	2	3	2	1,8
(1, 4)	1	1	1	2	2	1	1,4
(1, 5)	0	0	0	0	0	0	0
(1, 6)	0	0	0	1	1	0	0,4
(1, 7)	0	0	0	0	0	0	0
(2, 3)	1	1	1	1	1	1	1
(2, 4)	0	1	1	1	2	1	1
(2, 5)	0	0	0	0	0	0	0
(2, 6)	0	0	0	0	0	0	0
(2, 7)	0	0	0	0	1	0	0,2
(3, 4)	2	2	2	2	2	2	2
(3, 5)	1	1	1	1	2	1	1,2
(3, 6)	0	0	0	1	1	0	0,4
(3, 7)	0	0	0	0	0	0	0
(4, 5)	0	1	1	1	2	1	1
(4, 6)	0	1	2	2	2	2	1,4
(4, 7)	0	0	0	0	1	0	0,2
(5, 6)	0	0	0	0	0	0	0
(5, 7)	0	0	0	0	0	0	0
(6, 7)	0	0	0	1	1	0	0,4

The expected values of remaining pairs are determined in similar way and their is equal:

$$\sum_{T_1(\cdot)\neq\tilde{T}_1(\cdot)} E(D_{k,b}^{(1)}(x_1, x_3)) = -1\tfrac{2}{9}.$$

Thus the evaluations of the right-hand side of inequality (28) equals:

$$\exp\{-\frac{2N(\sum_{T_1(\cdot)\neq\tilde{T}_1(\cdot)} E(D_{k,b}^{(1)}(\cdot)))^2}{(2(m-1))^2}\} = \exp\{-0,1037\} = 0,9015$$

and the evaluation of the probability corresponding to m-1 (denoted $P^{(av)}_{b,m-1}(\overline{W}^*_1 - \tilde{\overline{W}}_1 < 0)$)

assumes the form:

$$P^{(av)}_{b,m-1}(\overline{W}^*_1 - \tilde{\overline{W}}_1 < 0) \geq 1\text{-}0{,}9015 = 0{,}0985. \quad (70)$$

If the value m-1 is replaced with the estimate $\hat{n}$=4, then:

$$P^{(av)}_{b,\hat{n}}(\overline{W}^*_1 - \tilde{\overline{W}}_1 < 0) \geq 1 - \exp\{-0{,}1910\} = 0{,}1739. \quad (71)$$

Table 2. The probability distributions functions $P(g^{(l)}_k(x_i, x_j)=l)$ - the basis for simulations

Pair	$P(g^{(l)}_k(x_i, x_j)=l)$				
(i,j)	L=0	l=1	l=2	l=3	l=4
(1, 2)	0,2	0,6	0,1	0,1	0,0
(1, 3)	0,2	0,5	0,2	0,1	0,0
(1, 4)	0,1	0,6	0,3	0,0	0,0
(1, 5)	0,7	0,2	0,1	0,0	0,0
(1, 6)	0,8	0,2	0,0	0,0	0,0
(1, 7)	0,9	0,1	0,0	0,0	0,0
(2, 3)	0,1	0,8	0,05	0,05	0,0
(2, 4)	0,2	0,75	0,05	0,0	0,0
(2, 5)	0,75	0,25	0,0	0,0	0,0
(2, 6)	0,65	0,35	0,0	0,0	0,0
(2, 7)	0,9	0,05	0,05	0,0	0,0
(3, 4)	0,0	0,1	0,7	0,1	0,1
(3, 5)	0,0	0,7	0,2	0,1	0,0
(3, 6)	0,8	0,2	0,0	0,0	0,0
(3, 7)	0,9	0,1	0,0	0,0	0,0
(4, 5)	0,3	0,6	0,1	0,0	0,0
(4, 6)	0,3	0,4	0,3	0,0	0,0
(4, 7)	0,9	0,1	0,0	0,0	0,0
(5, 6)	0,85	0,1	0,05	0,0	0,0
(5, 7)	0,95	0,05	0,0	0,0	0,0
(6, 7)	0,6	0,3	0,1	0,0	0,0

The evaluations obtained on the basis of the actual probability functions (see Table 2) assumes the form – for m-1 and $\hat{n}$ respectively:

$$P^{(av)}_{m-1}(\overline{W}^{*}_{f} - \tilde{\overline{W}}_{f} < 0) \geq 1 - \exp\{-0{,}5444\} = 0{,}4198, \qquad (72)$$

$$P^{(av)}_{\hat{n}}(\overline{W}^{*}_{f} - \tilde{\overline{W}}_{f} < 0) \geq 1 - \exp\{-1{,}2250\} = 0{,}7062. \qquad (73)$$

The evaluations (70) and (71) assume low values (closer to zero, than one), but the relations $\chi^{*}_{1}, \ldots, \chi^{*}_{4}$ and $\tilde{\chi}_{1}, \ldots, \tilde{\chi}_{4}$ are similar and differences between the "upper bound" and actual distributions are not negligible (see tables (2) and (3)). If the relations $\tilde{\chi}_{1}, \ldots, \tilde{\chi}_{r}$ and $\chi^{*}_{1}, \ldots, \chi^{*}_{n}$ are "more distant" (i.e. the set $\{T_1(\cdot) \neq \tilde{T}_1(\cdot)\}$ includes more elements), then the expression $(\sum_{T_1(\cdot)\neq\tilde{T}_1(\cdot)} E(D^{(1)}_{k}(x_1, x_3)))^2$ increases and the evaluation of the probability (28) also rises. The evaluations (72) and (73) based on actual probability functions are much better (especially (73)), than those "conservative".

Table 3. The "upper bound" distributions functions $P(g^{(1)}_{k}(\cdot) = \iota)$

Pair $<i, j>$	$P(g^{(1)}_{k}(\cdot) = \iota)$				
	$\iota = 0$	$\iota = 1$	$\iota = 2$	$\iota = 3$	$\iota = 4$
<1, 3>, <1, 4>	1/3	1/3	1/9	1/9	1/9
<1, 5>, <1, 6>	½	1/8	1/8	1/8	1/8

The empirical results confirm usefulness of the averaging approach. They suggests also, that the evaluation (71) based on the estimate $\hat{n}$ and "upper bound" distribution can be used in practice, as a rough approximation, albeit its value may be significantly below the probability resulting from the actual distributions.

5.2. The algorithm based on median approach

The medians from comparisons $g^{(1)}_{(1)}(x_i, x_j), \ldots, g^{(1)}_{(5)}(x_i, x_j)$ $((x_i, x_j) \in \mathbf{X} \times \mathbf{X})$ are presented in the Table 1. The optimal solution of the task (57), for f=1, is the same, as errorless one and those based on averaging approach. The minimal value of the function (57) equals 2, the solution is not multiple. The approximation of the right-hand of the probability (51) is determined with the use of algorithm described in the point 4.2 (see (59) – (67) and (68)). The first step of the median approach - determination of the probabilities $P_b(\hat{T}_1(\cdot) - g_k^{(1)}(\cdot) = l)$ - is described in the point 5.1. The second step - determination the values of the formula (68) – is performed, as follows. The expression for the distribution of the median (for conservative distributions) of comparisons errors assumes the form (see (53a, b)):

$$P_b^{(me,5)}(T_1(x_i,x_j) - g_{me,5}^{(1)}(x_i,x_j) = l) =$$

$$= \frac{5!}{(((5-1)/2)!)^2} \int_{G_b(l-1)}^{G_b(l)} t^{(5-1)/2}(1-t)^{(5-1)/2} dt = 30 \int_{G_b(l-1)}^{G_b(l)} t^2(1-t)^2 dt = 30t^3\left(\tfrac{1}{3} - t/2 + t^2/5\right) \Big|_{G_b(l-1)}^{G_b(l)}. \qquad (74)$$

For the pair (x_1, x_3) the distribution of the median of comparisons errors obtained on the basis of: the probability function resulting from the relationships (69a), the estimate $\hat{T}_1(x_1,x_3)$ and the expression (74) assumes the following form:

$P_b^{(me,5)}(\hat{T}_1(x_1,x_3) - g_{me,5}^{(1)}(x_1,x_3) = -3) = 0{,}0112,$

$P_b^{(me,5)}(\hat{T}_1(x_1,x_3) - g_{me,5}^{(1)}(x_1,x_3) = -2) = 0{,}0632,$

$P_b^{(me,5)}(\hat{T}_1(x_1,x_3) - g_{me,5}^{(1)}(x_1,x_3) = -1) = 0{,}1305,$

$P_b^{(me,5)}(\hat{T}_1(x_1,x_3) - g_{me,5}^{(1)}(x_1,x_3) = 0) = 0{,}5901,$

$P_b^{(me,5)}(\hat{T}_1(x_1,x_3) - g_{me,5}^{(1)}(x_1,x_3) = 1) = 0{,}2050.$

Thus, the expected value $E_b(|\hat{T}_1(x_1,x_3) - g_{me,5}^{(1)}(x_1,x_3)| - |\tilde{T}_1(x_1,x_3) - g_{me,5}^{(1)}(x_1,x_3)|)$ assumes the form:

$E_b(|\hat{T}_1(x_1,x_3) - g_{me,5}^{(1)}(x_1,x_3)| - |\tilde{T}_1(x_1,x_3) - g_{me,5}^{(1)}(x_1,x_3)|) =$

=0,0112(|1-4|-|2-4|)+0,0632(|1-3|-|2-3|)+0,1305(|1-2|-|2-2|)+

+0,5901(|1-1|-|2-1|)+0,2050(|1-0|-|2-0|)=-0,5901.

The remaining components of the sum $E(\sum_{T_1(\cdot)\neq\tilde{T}_1(\cdot)} D_{me,b}^{(1)}(\cdot))$ are determined in similar way and:

$E(\sum_{T_1(\cdot)\neq\tilde{T}_1(\cdot)} D_{me,b}^{(1)}(\cdot)) = -2{,}1866.$

The evaluation of the probability corresponding to the value m-1 (denoted $P_{b,m-1}^{(me)}(W_1^{(me,5)*} - \widetilde{W}_1^{(me,5)} < 0)$) assumes the form:

$$P_{b,m-1}^{(me,5)}(W_1^{(me,5)*} < \widetilde{W}_1^{(me,5)}) \geq -\frac{1}{\nu(m-1)} E\Big(\sum_{T_1(\cdot) \neq \hat{T}_1(\cdot)} D_{me,b}^{(1)}(\cdot)\Big) = -(1/(4*6))*(-2,1866) = 0,0911. \quad (75)$$

If the value m-1 is replaced with $\hat{n}$, then:

$$P_{b,\hat{n}}^{(me,5)}(W_1^{(me,5)*} < \widetilde{W}_1^{(me,5)}) \geq -\frac{1}{\nu\hat{n}} E\Big(\sum_{T_1(\cdot) \neq \hat{T}_1(\cdot)} D_{me,b}^{(1)}(\cdot)\Big) = -(1/4*4)*(-2,1866) = 0,1367. \quad (76)$$

Both evaluations (75) and (76) are rather poor, but they are based on rough probabilistic inequality (51) and "conservative" distributions functions $P_b(\hat{T}_1(\cdot) - g_k^{(1)}(\cdot) = l)$. However, in the example under consideration, both approaches (averaging and median) indicate the same estimation result and therefore the evaluations of the probabilities obtained for the averaging approach are valid also in the median case.

6. Summary and conclusions

The methods of the tolerance relation estimation presented in the paper are often essential in practice, but seldom discussed in the literature of the subject. The idea of the methods proposed is the same, as in earlier author's papers in this area (Klukowski (1990, 1994, 2000, 2002, 2006)). The results obtained are especially meaningful in the case of averaging approach, when $N \to \infty$; they indicate, that the probability of errorless result of estimation converges exponentially to one. The estimator based on the comparisons median also posses some asymptotic stochastic properties and is simpler from computational point of view. The range of statistical properties of both estimators can be extended.

The important features of the estimators proposed are weak assumptions about stochastic properties of the comparisons. Especially, the distributions functions of comparisons errors may be unknown, the comparisons of different pairs may be not independent in stochastic sense and the number of the subsets in the relation is not required.

Such features of comparisons are typical, when they are obtained with the use of statistical tests or other decision functions, generating random errors.

The estimated form of the relation is obtained on the basis of the optimal solution of appropriate discrete programming tasks. Therefore, the number of solutions may exceed one; each of them can be regarded, as estimated form of the relation. It is not a negative feature of the methods proposed; the unique estimate can be selected with the use of additional criteria, e.g. number of elements in intersections (maximal or minimal).

Empirical experience confirms usefulness of both estimators proposed, however it seems rational to perform also a broader simulation experiment.

Literature

David H.A. (1970) *Order Statistics*, J. Wiley, New York.

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

Hoeffding W. (1963) *Probability inequalities for sums of bounded random variables.* J. Amer. Statist. Assoc., vol. 58, pp. 13-30.

Klukowski L. (1990) *Algorithm for classification of samples in the case of unknown number variables generating them* (in Polish). Przegląd Statystyczny *R. XXXVII*, pp. 167 – 177.

Klukowski L. (1994) *Some probabilistic properties of the nearest adjoining order method and its extensions.* Annals of Operational Research, vol. 51, pp. 241 - 261.

Klukowski L. (2000) *The nearest adjoining order method for pairwise comparisons in the form of difference of ranks.* Annals of Operational Research, vol. 97, pp. 357 -378.

Klukowski L. (2002) *Estimation of the tolerance relation on the basis of pairwise comparisons with random errors* (in Polish). In.: Bubnicki Z., Hryniewicz O., Kulikowski R. eds. *Methods and Technics of Data Analysis and Decision Support.* Academic Publishing House EXIT, Warsaw, pp. V-21 – V-35.

Klukowski L. (2006) *Tests for relation type - equivalence or tolerance - in finite set of elements.* Control and Cybernetics, vol. 35, pp. 369 - 384.

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