POLSKA AKADEMIA NAUK
Instytut Badań Systemowych

METHODS OF ESTIMATION OF RELATIONS OF: EQUIVALENCE, TOLERANCE AND PREFERENCE IN A FINITE SET

Leszek Klukowski

Warsaw 2011

SYSTEMS RESEARCH INSTITUTE
POLISH ACADEMY OF SCIENCES

Series: SYSTEMS RESEARCH
Volume 69

==========

Series Editor:
Prof. dr hab. inż. Jakub Gutenbaum

Warsaw 2011

SYSTEMS RESEARCH INSTITUTE
POLISH ACADEMY OF SCIENCES

==========

Leszek Klukowski

METHODS OF ESTIMATION OF RELATIONS OF: EQUIVALENCE TOLERANCE AND PREFERENCE IN A FINITE SET

Warsaw 2011

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

Papers reviewers:

Prof. dr hab. inż. Ignacy Kaliszewski
Prof. dr hab. Tadeusz Trzaskalik

The work has been supported by the grant No N N111434937
of the Polish Ministry of Science and Higher Education

Printed in Polands
Systems Research Institute
Polish Academy of Sciences
Newelska 6, 01-447 Warsaw, Poland
www.ibspan.waw.pl

ISSN 0208-8029
ISBN 9788389475374

Chapter 6

Estimation of the tolerance relation on the basis of multivalent comparisons

6.1 Introduction

Multivalent comparisons reflect, in the case of the tolerance relation, quantitative attributes of elements compared; they can be interpreted as the number of common features of both elements. The approach can be also applied in the case of multiple binary comparisons – as the second step; the estimates obtained in the first step have to satisfy the assumptions required by multivalent comparisons.

6.2. Assumptions about multivalent comparisons

The tolerance relation, denoted $\chi_1^{(\tau)*}, ..., \chi_n^{(\tau)*}$ or $T_\mu^{(\tau)}(x_i, x_j)$, has to be estimated on the basis of comparisons $g_{\mu k}^{(\tau)}(x_i, x_j)$ $(k = 1, ..., N;$ $< i, j > \in R_m)$ defined as follows:

$g_{\mu k}^{(\tau)}(x_i, x_j) = d_{ijk}^{(\tau)}$ $(d_{ijk}^{(\tau)} \in \{0, ..., m\})$ - $d_{ijk}^{(\tau)}$ evaluation of the number $\#(\Omega_i^* \cap \Omega_j^*)$ - defined in (2.11) with random error satisfying the assumptions A1, A2, A3 in Chapter 2. (6.1)

The values of individual comparisons $g_{\mu k}^{(\tau)}(x_i, x_j)$ are determined by the number of elements *m*, because the number of subsets *n* is assumed not known.

6.3. The form of estimators and their properties

The estimators $\hat{\chi}_1^{(\tau)}, ..., \hat{\chi}_n^{(\tau)}$ and $\widehat{\chi}_1^{(\tau)}, ..., \widehat{\chi}_{\widetilde{n}}^{(\tau)}$ of the tolerance relation $\chi_1^{(\tau)}, ..., \chi_n^{(\tau)}$ are obtained on the basis of the following minimization problems:

$$\min_{F_{\mathbf{X}}^{(\tau)}} \{ \sum_{<i,j>\in R_m} \sum_{k=1}^{N} \left| g_{\mu k}^{(\tau)}(x_i, x_j) - t_{\mu}^{(\tau)}(x_i, x_j) \right| \}, \tag{6.2}$$

where:

$F_{\mathbf{X}}^{(\tau)}$ - the feasible set, i.e. the family of all tolerance relations in the set **X**,

$t_{\mu}^{(\tau)}(x_i, x_j)$ - the function describing the elements of the set $F_{\mathbf{X}}^{(\tau)}$, defined in the same way as $T_{\mu}^{(\tau)}(x_i, x_j)$ (Chapter 2),

and

$$\min_{F_{X}^{(\tau)}} \{ \sum_{<i,j>\in R_m} \left| g_{\mu}^{(\tau,me)}(x_i, x_j) - t_{\mu}^{(\tau)}(x_i, x_j) \right| \}, \tag{6.3}$$

where:

$g_{\mu}^{(\tau,me)}(x_i, x_j)$ - the sample median in the set $\{ g_{\mu,1}^{(\tau)}(x_i, x_j), ..., g_{\mu N}^{(\tau)}(x_i, x_j) \}$.

The properties of both estimators are based on properties of the random variables $W_{\mu N}^{(\tau)*}$ and $W_{\mu}^{(\tau,me)*}$, corresponding to actual relation $\chi_1^{(\tau)*}, ..., \chi_n^{(\tau)*}$:

$$W_{\mu N}^{(\tau)*} = \sum_{<i,j>\in R_m} \sum_{k=1}^{N} U_{\mu k}^{(\tau)*}(x_i, x_j), \tag{6.4}$$

where:

$$U_{\mu k}^{(\tau)*}(x_i, x_j) = \begin{cases} 0 \ \ if \ \ g_{\mu k}^{(\tau)}(x_i, x_j) = T_{\mu}^{(\tau)}(x_i, x_j); \\ g_{\mu k}^{(\tau)}(x_i, x_j) - T_{\mu}^{(\tau)}(x_i, x_j) \ \ if \ \ g_{\mu k}^{(\tau)}(x_i, x_j) > T_{\mu}^{(\tau)}(x_i, x_j); \\ T_{\mu}^{(\tau)}(x_i, x_j) - g_{\mu k}^{(\tau)}(x_i, x_j) \ \ if \ T_{\mu}^{(\tau)}(x_i, x_j) > g_{\mu k}^{(\tau)}(x_i, x_j), \end{cases}$$

$$W_{\mu N}^{(\tau,me)*} = \sum_{<i,j>\in R_m} U_{\mu}^{(\tau,me)*}(x_i, x_j), \tag{6.5}$$

where:

$$U_{\mu}^{(\tau,me)*}(x_i,x_j)=\begin{cases}0 \;\; if \;\; g_{\mu}^{(\tau,me)}(x_i,x_j)=T_{\mu}^{(\tau)}(x_i,x_j);\\ g_{\mu}^{(\tau,me)}(x_i,x_j)-T_{\mu}^{(\tau)}(x_i,x_j) \;\; if \;\; g_{\mu}^{(\tau,me)}(x_i,x_j)>T_{\mu}^{(\tau)}(x_i,x_j);\\ T_{\mu}^{(\tau)}(x_i,x_j)-g_{\mu}^{(\tau,me)}(x_i,x_j) \;\; if \;\; T_{\mu}^{(\tau)}(x_i,x_j)>g_{\mu}^{(\tau,me)}(x_i,x_j),\end{cases}$$

and random variables $\widetilde{W}_{\mu N}^{(\tau)}$, $\widetilde{W}_{\mu N}^{(\tau,me)}$ corresponding to a relation $\widetilde{\chi}_1^{(\tau)},...,\widetilde{\chi}_{\tilde{n}}^{(\tau)}$ different from the actual one, defined in the same way:

$$\widetilde{W}_{\mu N}^{(\tau)}=\sum_{<i,j>\in R_m}\sum_{k=1}^{N}\widetilde{U}_{\mu k}^{(\tau)}(x_i,x_j), \tag{6.6}$$

where:

$$\widetilde{U}_{\mu k}^{(\tau)}(x_i,x_j)=\begin{cases}0 \;\; if \;\; g_{\mu k}^{(\tau)}(x_i,x_j)=\widetilde{T}_{\mu}^{(\tau)}(x_i,x_j);\\ g_{\mu k}^{(\tau)}(x_i,x_j)-\widetilde{T}_{\mu}^{(\tau)}(x_i,x_j) \;\; if \;\; g_{\mu k}^{(\tau)}(x_i,x_j)>\widetilde{T}_{\mu}^{(\tau)}(x_i,x_j);\\ \widetilde{T}_{\mu}^{(\tau)}(x_i,x_j)-g_{\mu k}^{(\tau)}(x_i,x_j) \;\; if \;\; \widetilde{T}_{\mu}^{(\tau)}(x_i,x_j)>g_{\mu k}^{(\tau)}(x_i,x_j),\end{cases}$$

$$\widetilde{W}_{\mu N}^{(\tau,me)}=\sum_{<i,j>R_m}\widetilde{U}_{\mu}^{(\tau,me)}(x_i,x_j), \tag{6.7}$$

where:

$$\widetilde{U}_{\mu}^{(\tau,me)}(x_i,x_j)=\begin{cases}0 \;\; if \;\; g_{\mu}^{(\tau,me)}(x_i,x_j)=\widetilde{T}_{\mu}^{(\tau)}(x_i,x_j);\\ g_{\mu}^{(\tau,me)}(x_i,x_j)-\widetilde{T}_{\mu}^{(\tau)}(x_i,x_j) \;\; if \;\; g_{\mu}^{(\tau,me)}(x_i,x_j)>\widetilde{T}_{\mu}^{(\tau)}(x_i,x_j);\\ \widetilde{T}_{\mu}^{(\tau)}(x_i,x_j)-g_{\mu}^{(\tau,me)}(x_i,x_j) \;\; if \;\; \widetilde{T}_{\mu}^{(\tau)}(x_i,x_j)>g_{\mu}^{(\tau,me)}(x_i,x_j).\end{cases}$$

The variables $W_{\mu N}^{(\tau)*}$, $W_{\mu}^{(\tau,me)*}$ and $\widetilde{W}_{\mu N}^{(\tau)}$, $\widetilde{W}_{\mu}^{(\tau,me)}$ can be expressed in the form corresponding to the optimization tasks (6.2), (6.3), i.e.:

$$W_{\mu N}^{(\tau)*}=\sum_{<i,j>\in R_m}\sum_{k=1}^{N}\left|g_{\mu k}^{(\tau)}(x_i,x_j)-T_{\mu}^{(\tau)}(x_i,x_j)\right|,$$

$$W_{\mu}^{(\tau,me)*}=\sum_{<i,j>\in R_m}\left|g_{\mu}^{(\tau,me)}(x_i,x_j)-T_{\mu}^{(\tau)}(x_i,x_j)\right|,$$

$$\widetilde{W}_{\mu N}^{(\tau)} = \sum_{<i,j>\in R_m} \sum_{k=1}^{N} \left| g_{\mu k}^{(\tau)}(x_i, x_j) - \widetilde{T}_{\mu}^{(\tau)}(x_i, x_j) \right|,$$

$$\widetilde{W}_{\mu}^{(\tau,me)} = \sum_{<i,j>\in R_m} \left| g_{\mu}^{(\tau,me)}(x_i, x_j) - \widetilde{T}_{\mu}^{(\tau)}(x_i, x_j) \right|.$$

It can be shown that:

Theorem 3

The following relationships are true:

$$E(W_{\mu N}^{(\tau)*}) < E(\widetilde{W}_{\mu N}^{(\tau)}), \tag{6.8}$$

$$E(W_{\mu N}^{(\tau,me)*}) < (\widetilde{W}_{\mu N}^{(\tau,me)}), \tag{6.9}$$

$$\lim_{N\to\infty} Var(\tfrac{1}{N} W_{\mu N}^{(\tau)*}) = 0, \tag{6.10}$$

$$\lim_{N\to\infty} Var(\tfrac{1}{N} \widetilde{W}_{\mu N}^{(\tau)}) = 0, \tag{6.11}$$

$$\lim_{N\to\infty} Var(W_{\mu}^{(\tau,me)*}) = 0, \tag{6.12}$$

$$\lim_{N\to\infty} Var(\widetilde{W}_{\mu}^{(\tau,me)}) = 0. \tag{6.13}$$

The probability $P(W_{\mu N}^{(\tau)*} < \widetilde{W}_{\mu N}^{(\tau)})$ satisfies the inequality:

$$P(W_{\mu N}^{(\tau)*} < \widetilde{W}_{\mu N}^{(\tau)}) \geq 1 - \exp\{-2N \frac{(\sum_{<i,j>\in R_m} \sum_{k=1}^{N} E(\left| g_{\mu k}^{(\tau)}(x_i, x_j) - T_{\mu}^{(\tau)}(x_i, x_j) \right| - \left| g_{\mu k}^{(\tau)}(x_i, x_j) - \widetilde{T}_{\mu}^{(\tau)}(x_i, x_j) \right|))^2}{(2\vartheta(m-1))^2}\}$$

which implies that

$$\lim_{N\to\infty} P(W_{\mu N}^{(\tau)*} < \widetilde{W}_{\mu N}^{(\tau)}) = 1, \tag{6.14}$$

where: ϑ - the number of elements of the set

$$\{T_{\mu N}^{(\tau,me)*}(x_i,x_j) \neq \widetilde{T}_{\mu N}^{(\tau,me)}(x_i,x_j)\}.$$

In the case of the median estimator the following relationship is true:

$$\lim_{N\to\infty} P(W_{\mu N}^{(\tau,me)*} < \widetilde{W}_{\mu N}^{(\tau,me)}) = 1. \tag{6.15}$$

The proof of inequality (6.14) is given in Klukowski (2007b), based on Hoeffding's (1963) inequality for bounded random variables. The convergence of the variances (6.12), (6.13) and the probability $P(W_{\mu N}^{(\tau,me)*} < \widetilde{W}_{\mu N}^{(\tau,me)})$ results from the formula determining the distribution of the sample median (David, 1970, Klukowski, 2007c):

$$P(g_{\mu}^{(\tau,me)}(x_i,x_j) - T_{\mu}^{(\tau,me)}(x_i,x_j) = 0) = \frac{N!}{(((N-1)/2)!)^2} \int_{G(-1)}^{G(0)} t^{(N-1)/2}(1-t)^{(N-1)/2} dt, \tag{6.16}$$

where:

$G(t)$ - the value of the cumulative distribution function of the comparison error of the pair (x_i,x_j).

Equality (6.16) has been established in Klukowski (2007). The right-hand side of the (6.16) converges to 1 for $N \to \infty$.

The existence of distributions of both estimators (6.2), (6.3) can be proven in the same way as in Chapter 2.

The question of efficiency of both estimators is addressed in the simulation survey (Chapter 9). Higher efficiency of the estimator based on the sum of inconsistencies is shown. The computation cost, required by the median estimator is lower than for the estimator based on the total sum of inconsistencies; it is also more robust with respect to outliers in comparisons.

The evaluation (6.14) requires distributions of comparisons errors and corresponds to the fixed relation form $\widetilde{\chi}_1^{(\tau)}, ..., \widetilde{\chi}_{\widetilde{n}}^{(\tau)}$. However, some

evaluation of the right-hand side of the inequality can be determined on the basis of the concept of the quasi-uniform probability function, proposed in Klukowski (2006b). The quasi-uniform distribution is based on the following assumptions:

- it is determined on the basis of $\hat{n}$ or $\bar{n}$, corresponding to estimates $\hat{\chi}_1^{(\tau)*}, ..., \hat{\chi}_{\hat{n}}^{(\tau)*}$ or $\bar{\chi}_1^{(\tau)*}, ..., \bar{\chi}_{\bar{n}}^{(\tau)*}$,

- the probabilities in the left and right tie of each distribution are equal, e.g. $P(g_{\mu k}^{(\tau)}(x_i, x_j) - \hat{T}_{\mu}^{(\tau)}(x_i, x_j) < 0 = P(g_{\mu k}^{(\tau)}(x_i, x_j) - \hat{T}_{\mu}^{(\tau)}(x_i, x_j) > 0)$ (for $\hat{T}_{\mu}^{(\tau)}(x_i, x_j) \neq 0, \hat{T}_{\mu}^{(\tau)}(x_i, x_j) \neq n$); in the cases $\hat{T}_{\mu}^{(\tau)}(x_i, x_j) = 0, \hat{T}_{\mu}^{(\tau)}(x_i, x_j) = m$) errors assume – respectively positive and negative values),

- the probabilities in the left tie are equal and in the right tie are equal, e.g. $P(g_{\mu k}^{(\tau)}(x_i, x_j) - T_{\mu}^{(\tau)}(x_i, x_j) = -1) = P(g_{\mu k}^{(\tau)}(x_i, x_j) - T_{\mu}^{(\tau)}(x_i, x_j) = -2)$,

- the probabilities $P(g_{\mu k}^{(\tau)}(x_i, x_j) = T_{\mu}^{(\tau)}(x_i, x_j))$ have to imply satisfying of the assumption on the mode and median of the distribution.

The above assumptions determine in a unique way the probability function; an example of such a distribution is presented in Klukowski (2007b) and in Chapter 10.

In the case of appropriate *N*, at least several, the distributions of comparison errors can be estimated.

The minimization problems (6.2), (6.3) require much more computations than binary problems. For $m \leq 12$ full examination can be applied. Validation of estimates is discussed in Chapter 10.

The estimator examined in this chapter is based on the measure of similarity of elements from a pair (the number of common features). In Klukowski (2007, 2008a) it is presented a more general approach, with the use of dissimilarity measure. This measure expresses the number of lacking

features of elements from a pair. The component, including the measure is included into the criterion function. The use of comparisons expressing lacking features can improve the efficiency of estimates, because it increases the number of data.

6.4. Summary

The estimators of the tolerance relation, based on multivalent comparisons have good statistical properties and can exploit the quantitative form of data. It is important that probability of correct comparison of each pair can be lower than 1/2. However, the computational cost of estimation is higher than in the case of binary estimators.

The simulation study (Chapter 9), concerning the estimators of the preference relation, based on multivalent comparisons, demonstrates high efficiency of the approach. Therefore, it is valid to apply multivalent estimators also in the case of multiple binary comparisons – using the two-stage estimation, i.e.:

- N - binary estimates in the first stage,
- N - multivalent comparisons based on binary estimates,
- application of multivalent estimators.

The multivalent comparisons, resulting from binary estimates have to satisfy the assumptions required.

Appendix 2. The idea of proofs of relationships (6.8) – (6.15)

The proof of the inequality (6.8) is based on the fact that the expected value of each difference $U_{\mu k}^{(\tau)*}(x_i, x_j) - \widetilde{U}_{\mu k}^{(\tau)}(x_i, x_j)$ $(\widetilde{T}_{\mu}^{(\tau)}(x_i, x_j) \neq T_{\mu}^{(\tau)}(x_i, x_j))$ is negative, i.e.:

$$\begin{aligned} &E(U_{\mu k}^{(\tau)*}(x_i, x_j) - \widetilde{U}_{\mu k}^{(\tau)}(x_i, x_j)) = \\ &E(\left|g_{\mu k}^{(\tau)}(x_i, x_j) - T_{\mu}^{(\tau)}(x_i, x_j)\right| - \left|g_{\mu k}^{(\tau)}(x_i, x_j) - \widetilde{T}_{\mu}^{(\tau)}(x_i, x_j)\right|) < 0. \end{aligned} \tag{A2.1}$$

The proof of inequality (A2.1) is elementary, but cumbersome (Klukowski 2007b). It results from the fact that the value $T_{\mu}^{(\tau)}(x_i, x_j)$ is the median of distribution of any comparison $g_{\mu k}^{(\tau)}(x_i, x_j)$.

The proof of inequality (6.9) is similar.

The idea of the proofs of inequalities (6.10)–(6.13) is similar to the case of binary comparisons (see Appendix 1). The inequalities (6.12), (6.13), for the median estimator, are obtained with the use of relationship (6.16), resulting from the properties of order statistic (David, 1970, Ch. 2):

$$P(g_{\mu}^{(\tau,me)}(x_i, x_j) - T_{\mu}^{(\tau,me)}(x_i, x_j) \leq 0) = \frac{N!}{(((N-1)/2)!)^2} \int_0^{G(0)} t^{(N-1)/2}(1-t)^{(N-1)/2} dt \,.$$

The proof of inequality $P(W_{\mu N}^{(\tau)*} < \widetilde{W}_{\mu N}^{(\tau)})$ is based on Hoeffding's (1963) inequality:

$$P(\sum_{i=1}^{N} Y_i - \sum_{i=1}^{N} E(Y_i) \geq Nt) \leq \exp\{-2Nt^2 \big/ (b-a)^2\}, \tag{A2.2}$$

where:

$P(a \leq Y_i \leq b) = 1$,

$a < b$, $t > 0$.

It is applied to the random variables:

$$\sum_{T_\mu(x_i,x_j)\neq\widetilde{T}_\mu(x_i,x_j)} U_{\mu k}^{(\tau)*}(x_i,x_j)-\widetilde{U}_{\mu k}^{(\tau)}(x_i,x_j)=$$

$$\sum_{T_\mu(x_i,x_j)\neq\widetilde{T}_\mu(x_i,x_j)} \left|g_{\mu k}^{(\tau)}(x_i,x_j)-T_\mu^{(\tau)}(x_i,x_j)\right|-\left|g_{\mu k}^{(\tau)}(x_i,x_j)-\widetilde{T}_\mu^{(\tau)}(x_i,x_j)\right|$$

$(k=1,...,N).$

It is clear that such random variables are independent and bounded, and therefore provide the basis for application of inequality (A2.2).

The book presents the estimators of three relations: equivalence, tolerance, and preference in a finite set of data items, based on multiple pairwise comparisons, assumed to be disturbed by random errors. The estimators were developed by the author. They can refer to binary (qualitative), multivalent (quantitative) and combined comparisons. The estimates are obtained on the basis of solutions to the discrete programming problems. The estimators have been developed under weak assumptions on the distributions of comparison errors; in particular, these distributions can have non-zero expected values. The estimators have good statistical properties, including, especially importantly, consistency. Therefore, they produce good results in cases when other methods generate incorrect estimates. The precision of the estimators has been established with the use of simulation methods. The estimates can be validated in a versatile way. The whole estimation process, i.e. comparisons, estimation and validation can be computerized. The approach allows also for inference about the relation type – equivalence or tolerance, on the basis of binary data. Thus, it has features of data mining methods.

The estimators have been applied for ranking and grouping of data from some empirical sets. In particular, estimation of the tolerance relation (overlapping classification) was applied for determination of homogenous shapes of functions expressing profitability of treasury securities and was used for forecasting purposes.

ISSN 0208-8029
ISBN 9788389475374

SYSTEMS RESEARCH INSTITUTE
POLISH ACADEMY OF SCIENCES
Phone: (+48) 22 3810246 / 22 3810277 / 22 3810241 / 22 3810273
email: biblioteka@ibspan.waw.pl