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

Editorial Board
Series: SYSTEMS RESEARCH

Prof. Olgierd Hryniewicz - chairman
Prof. Jakub Gutenbaum – series editor
Prof. Janusz Kacprzyk
Prof. Tadeusz Kaczorek
Prof. Roman Kulikowski
Prof. Marek Libura
Prof. Krzysztof Malinowski
Prof. Zbigniew Nahorski
Prof. Marek Niezgódka
Prof. Roman Słowiński
Prof. Jan Studziński
Prof. Stanisław Walukiewicz
Prof. Andrzej Weryński
Prof. Antoni Żochowski

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

Copyright © by Systems Research Institute
Polish Academy of Sciences
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 3

Estimation of the equivalence relation

3.1. Introduction

The problem of estimation of the equivalence relation has been stated for binary comparisons only. Such estimators assume the simplest form considered here.

3.2. Assumptions about binary comparisons

The equivalence relation, expressed by $\chi_1^{(e)*}, ..., \chi_n^{(e)*}$ or $T_b^{(e)}(x_i, x_j)$, has to be estimated on the basis of comparisons $g_{bk}^{(e)}(x_i, x_j)$ $(k = 1, ..., N; < i, j > \in R_m)$, defined as follows:

$$g_{bk}^{(e)}(x_i, x_j) = \begin{cases} 0 & \text{if } k-\text{th comparison indicates that a pair } (x_i, x_j) \\ & \text{belongs to the same subset } \chi_q^{(e)*} \ (1 \le q \le m); \\ 1 & \text{otherwise.} \end{cases} \tag{3.1}$$

The comparisons $g_{bk}^{(e)}(x_i, x_j)$ $(< i, j > \in R_m)$ have to satisfy the assumptions A1 - A3, i.e.: the probability of correct comparison, $1-\delta$, has to be greater than the probability of incorrect comparison, δ, and the comparisons have to be stochastically independent. The number of subsets, *n,* is assumed unknown and, therefore, confined by the number of elements, *m*.

The assumption about independence of all comparisons can be relaxed in such way that comparisons $g_{bk}^{(e)}(x_i, x_j)$ $(1 \le k \le N; < i, j > \in R_m)$ and $g_{bl}^{(e)}(x_r, x_s)$ $(l \ne k; \ < r, s > \in R_m)$ have to be independent. In such a case the main properties of estimators are preserved, but inference about distributions of the estimators gets more complicated.

$$U_{bk}^{(e)*}(x_i,x_j)=\begin{cases}0 \;\; if \;\; g_{bk}^{(e)}(x_i,x_j)=T_b^{(e)}(x_i,x_j);\;\; T_b^{(e)}(x_i,x_j)=0;\\ 1 \;\; if \;\; g_{bk}^{(e)}(x_i,x_j)\neq T_b^{(e)}(x_i,x_j);\;\; T_b^{(e)}(x_i,x_j)=0,\end{cases} \tag{3.6}$$

$$V_{bk}^{(e)*}(x_i,x_j)=\begin{cases}0 \;\; if \;\; g_{bk}^{(e)}(x_i,x_j)=T_b^{(e)}(x_i,x_j);\;\; T_b^{(e)}(x_i,x_j)=1;\\ 1 \;\; if \;\; g_{bk}^{(e)}(x_i,x_j)\neq T_b^{(e)}(x_i,x_j);\;\; T_b^{(e)}(x_i,x_j)=1,\end{cases} \tag{3.7}$$

$I^{(e)*}$ - the set of pairs $\{<i,j> \;\big|\; T_b^{(e)*}(x_i,x_j)=0\}$,

$J^{(e)*}$ - the set of pairs $\{<i,j> \;\big|\; T_b^{(e)*}(x_i,x_j)=1\}$,

$$U_b^{(e,me)*}(x_i,x_j)=\begin{cases}0 \;\; if \;\; g_b^{(e,me)}(x_i,x_j)=T_b^{(e)}(x_i,x_j);\;\; T_b^{(e)}(x_i,x_j)=0;\\ 1 \;\; if \;\; g_b^{(e,me)}(x_i,x_j)\neq T_b^{(e)}(x_i,x_j);\;\; T_b^{(e)}(x_i,x_j)=0,\end{cases} \tag{3.8}$$

$$V_b^{(e,me)*}(x_i,x_j)=\begin{cases}0 \;\; if \;\; g_b^{(e,me)}(x_i,x_j)=T_b^{(e)}(x_i,x_j);\;\; T_b^{(e)}(x_i,x_j)=1;\\ 1 \;\; if \;\; g_b^{(e,me)}(x_i,x_j)\neq T_b^{(e)}(x_i,x_j);\;\; T_b^{(e)}(x_i,x_j)=1\end{cases} \tag{3.9}$$

and

$$\widetilde{W}_{bN}^{(e)}=\sum_{k=1}^{N}\sum_{<i,j>\in\widetilde{I}^{(e)}}\widetilde{U}_{bk}^{(e)}(x_i,x_j)+\sum_{k=1}^{N}\sum_{<i,j>\in\widetilde{J}^{(e)}}\widetilde{V}_{bk}^{(e)}(x_i,x_j)= \sum_{<i,j>\in R_m}\sum_{k=1}^{N}\left|g_{bk}^{(e)}(x_i,x_j)-\widetilde{T}_b^{(e)}(x_i,x_j)\right|, \tag{3.10}$$

$$\widetilde{W}_{bN}^{(e,me)}=\sum_{<i,j>\in\widetilde{I}^{(e)}}\widetilde{U}_b^{(e,me)}(x_i,x_j)+\sum_{<i,j>\in\widetilde{J}^{(e)}}\widetilde{V}_b^{(e,me)}(x_i,x_j) \sum_{<i,j>\in R_m}\left|g_b^{(e,me)}(x_i,x_j)-\widetilde{T}_b^{(e)}(x_i,x_j)\right|, \tag{3.11}$$

where:

$$\widetilde{U}_{bk}^{(e)}(x_i,x_j)=\begin{cases}0 \;\; if \;\; g_{bk}^{(e)}(x_i,x_j)=\widetilde{T}_b^{(e)}(x_i,x_j);\;\; \widetilde{T}_b^{(e)}(x_i,x_j)=0;\\ 1 \;\; if \;\; g_{bk}^{(e)}(x_i,x_j)\neq\widetilde{T}_b^{(e)}(x_i,x_j);\;\; \widetilde{T}_b^{(e)}(x_i,x_j)=0,\end{cases} \tag{3.12}$$

$$\widetilde{V}_{bk}^{(e)}(x_i,x_j)=\begin{cases}0 \;\; if \;\; g_{bk}^{(e)}(x_i,x_j)=\widetilde{T}_b^{(e)}(x_i,x_j);\;\; \widetilde{T}_b^{(e)}(x_i,x_j)=1;\\ 1 \;\; if \;\; g_{bk}^{(e)}(x_i,x_j)\neq\widetilde{T}_b^{(e)}(x_i,x_j);\; \widetilde{T}_b^{(e)}(x_i,x_j)=1\end{cases} \tag{3.13}$$

$$\widetilde{U}_b^{(e,me)}(x_i,x_j)=\begin{cases}0 \;\; if \;\; g_b^{(e,me)}(x_i,x_j)=\widetilde{T}_b^{(e)}(x_i,x_j);\;\; \widetilde{T}_b^{(e)}(x_i,x_j)=0;\\ 1 \;\; if \;\; g_b^{(e,me)}(x_i,x_j)\neq\widetilde{T}_b^{(e)}(x_i,x_j);\;\; \widetilde{T}_b^{(e)}(x_i,x_j)=0,\end{cases} \tag{3.14}$$

$$\widetilde{V}_b^{(e,me)}(x_i,x_j)=\begin{cases}0 \;\; if \;\; g_{bk}^{(e,me)}(x_i,x_j)=\widetilde{T}_b^{(e)}(x_i,x_j);\;\; \widetilde{T}_b^{(e)}(x_i,x_j)=1;\\ 1 \;\; if \;\; g_{bk}^{(e,me)}(x_i,x_j)\neq\widetilde{T}_b^{(e)}(x_i,x_j);\; \widetilde{T}_b^{(e)}(x_i,x_j)=1,\end{cases} \tag{3.15}$$

$\widetilde{I}^{(e)}$ - the set of pairs $\{<i,j> \;\big|\; \widetilde{T}_b^{(e)}(x_i,x_j)=0\}$,

$\widetilde{J}^{(e)}$ - the set of pairs $\{<i,j> \;\big|\; \widetilde{T}_b^{(e)}(x_i,x_j)=1\}$.

It can be shown, in the same way as in Klukowski (1994), that

Theorem 1

The following relationships are true:

$$E(W_{bN}^{(e)*}<\widetilde{W}_{bN}^{(e)})<0\,, \tag{3.16}$$

$$E(W_{bN}^{(e,me)*}<\widetilde{W}_{bN}^{(e,me)})<0\,, \tag{3.17}$$

$$\lim_{N\to\infty} Var(\tfrac{1}{N}W_{bN}^{(e)*})=0\,, \tag{3.18a}$$

$$\lim_{N\to\infty} Var(W_{bN}^{(e,me)*})=0\,, \tag{3.18b}$$

$$\lim_{N\to\infty} Var(\tfrac{1}{N}\widetilde{W}_{bN}^{(e)})=0\,, \tag{3.19a}$$

$$\lim_{N\to\infty} Var(\widetilde{W}_{bN}^{(e,me)})=0\,. \tag{3.19b}$$

Moreover, the probabilities: $P(W_{bN}^{(e)*}<\widetilde{W}_{bN}^{(e)})$, $P(W_{bN}^{(e,me)*}<\widetilde{W}_{bN}^{(e,me)})$ satisfy the inequalities:

$$P(W_{bN}^{(e)*} < \widetilde{W}_{bN}^{(e)}) \geq 1 - \exp\{-2N(\tfrac{1}{2} - \delta)^2\}, \tag{3.20}$$

$$P(W_{bN}^{(e,me)*} < \widetilde{W}_{bN}^{(e,me)}) \geq 1 - 2\exp\{-2N(\tfrac{1}{2} - \delta)^2\}. \tag{3.21}$$

The proof of the relationships (3.18)–(3.21) is the same as in Klukowski (1994); the inequalities (3.20), (3.21) have been proven on the basis of Hoeffding (1963) inequalities. The idea of the proofs is presented in Appendix 1.

The relationships (3.16)–(3.17) indicate that the expected value $E(\frac{1}{N}W_{bN}^{(p)*})$, corresponding to the actual relation $\chi_1^{(e)*}, ..., \chi_n^{(e)*}$, is lower than the expected value $E(\frac{1}{N}\widetilde{W}_{bN}^{(p)})$, corresponding to any other relation $\widetilde{\chi}_1^{(p)}, ..., \widetilde{\chi}_{\tilde{n}}^{(p)}$. A similar property is true for the median estimator. The variances $Var(\frac{1}{N}W_{bN}^{(e)*})$, $Var(\frac{1}{N}\widetilde{W}_{bN}^{(e)})$, $Var(W_{bN}^{(e,me)*})$, $Var(\widetilde{W}_{bN}^{(e,me)})$ converge to zero for $N \to \infty$; the convergence of the variances $Var(W_{bN}^{(e,me)*}), Var(\widetilde{W}_{bN}^{(e,me)})$ results from the convergence of the median of the binary comparisons to the actual value.

The variables $U_{bk}^{(e)*}(x_i, x_j)$, $V_{bk}^{(e)*}(x_i, x_j)$ assume values equal to $\left|g_{bk}^{(e)}(x_i, x_j) - T_b^{(e)}(x_i, x_j)\right|$ and the variables $\widetilde{U}_{bk}^{(e)}(x_i, x_j)$, $\widetilde{V}_{bk}^{(e)}(x_i, x_j)$ - values $\left|g_{bk}^{(e)}(x_i, x_j) - \widetilde{T}_b^{(e)}(x_i, x_j)\right|$. Similarly, the variables $U_b^{(e,me)*}(x_i, x_j)$, $V_b^{(e,me)*}(x_i, x_j)$ assume values equal to $\left|g_b^{(e,me)}(x_i, x_j) - T_b^{(e)}(x_i, x_j)\right|$ and the variables $\widetilde{U}_b^{(e,me)}(x_i, x_j)$, $\widetilde{V}_b^{(e,me)}(x_i, x_j)$ - values $\left|g_b^{(e,me)}(x_i, x_j) - \widetilde{T}_b^{(e)}(x_i, x_j)\right|$. The inequalities (3.20), (3.21) show that the probabilities of the events $\{W_{bN}^{(e)*} < \widetilde{W}_{bN}^{(e)})\}$ and $\{W_{bN}^{(e,me)*} < \widetilde{W}_{bN}^{(e,me)}\}$ converge to 1 for $N \to \infty$. The right-hand sides of both inequalities include an exponential term, guaranteeing fast type of convergence. The evaluation, corresponding to the estimator based on the sum of inconsistencies is better than the evaluation, corresponding to the median estimator. Both evaluations depend on δ. It is obvious that they are more efficient for the value of the parameter close to zero.

The above facts indicate that the estimators, minimizing the number of inconsistencies with comparisons, guarantee the errorless estimate for $N \to \infty$. The inequalities (3.20)–(3.21) show that this guarantees good result

also for finite N; they show the influence of δ and N on precision of the estimator. The errors of the estimators, for given parameters, have to be evaluated using simulation.

Other, analytical properties of the estimators are presented in Klukowski (1994), e.g. corresponding to the case of $m \to \infty$, with n constant and $card(\chi_q^*) \to \infty$ $(q = 1, ..., n)$. In such a case the probability of errorless estimate converges to 1, also in the case of singular comparison (N=1). However, from the practical point of view (m – finite), more important are the properties obtained from the simulation survey (Chapter 9).

Applications of the approach are presented in Klukowski (1990, 2006), Klukowski, Kuba (2002). The first of these papers presents the classification of slopes of a piecewise linear trend, the subsequent papers – functions expressing profitability of treasury securities, resulting from the auctions.

The tasks (3.2), (3.3) can be solved with the use of the algorithms presented in Chapter 2. Validation of estimates is discussed in Chapter 10.

In Klukowski (1994) it is presented a broader set of properties of the estimators, especially evaluations of variance of the random variable $Var(W_{bN}^{(e,me)*})$ and discussion of the case of dependent comparisons.

3.4. Summary

The estimators of the equivalence relation have good statistical properties and a simple form. Moreover, there exist efficient algorithms for solving the respective optimization tasks.

Appendix 1. The idea of the proofs of inequalities (3.16) - (3.21)

Inequality (3.16).

The difference of the random variables $W_{bN}^{(e)*}$, $\widetilde{W}_{bN}^{(e)}$ assumes the form (see Klukowski 1990a):

$$W_{bN}^{(e)*} - \widetilde{W}_{bN}^{(e)} =$$
$$\sum_{k=1}^{N} \sum_{<i,j>\in I^{(e)*}} U_{bN}^{(e)*}(x_i, x_j) + \sum_{k=1}^{N} \sum_{<i,j>\in J^{(e)*}} V_{bN}^{(e)*}(x_i, x_j) -$$
$$(\sum_{k=1}^{N} \sum_{<i,j>\in \widetilde{I}^{(e)}} \widetilde{U}_{bN}^{(e)}(x_i, x_j) + \sum_{k=1}^{N} \sum_{<i,j>\in \widetilde{J}^{(e)}} \widetilde{V}_{bN}^{(e)}(x_i, x_j)) =$$
$$\sum_{k=1}^{N} (\sum_{I^{(e)*}\cap(\widetilde{J}^{(e)} - I^{(e)*})} (U_{bN}^{(e)*}(x_i, x_j) - \widetilde{V}_{bN}^{(e)}(x_i, x_j)) +$$
$$\sum_{J^{(e)*}\cap(\widetilde{I}^{(e)} - J^{(e)*})} (V_{bN}^{(e)*}(x_i, x_j) - \widetilde{U}_{bN}^{(e)}(x_i, x_j))). \tag{A3.1}$$

Clearly, the expected values of the components $U_{bN}^{(e)*}(x_i, x_j) - \widetilde{V}_{bN}^{(e)}(x_i, x_j)$, $V_{bN}^{(e)*}(x_i, x_j) - \widetilde{U}_{bk}^{(e)}(x_i, x_j)$ of the sum (A3.1) are negative and, therefore:

$$E(W_{bN}^{(e)*} - \widetilde{W}_{bN}^{(e)}) < 0.$$

The proof of inequality (3.17) is similar.

The idea of the proofs of relationships (3.18a), (3.18b).

Relationship (3.18a).

The variance $Var(\frac{1}{N} W_{bN}^{(e)*})$ converges to zero for $N \to \infty$, because the variance of each component $\sum_{<i,j>\in I^{(e)*}} U_{bk}^{(e)*}(x_i, x_j) + \sum_{<i,j>\in J^{(e)*}} V_{bk}^{(e)*}(x_i, x_j)$ $(k = 1, ..., N)$ of the variable $W_{bN}^{(e)*}$ is constant and assumes the same value. The variance of the sum of such variables, divided by their number, converges to zero.

The proof of (3.18b) results from the fact that the variance of each random variable $(U_b^{(e,me)*}(x_i,x_j)-T_b^{(e)}(x_i,x_j))$, $(V_b^{(e,me)*}(x_i,x_j)-T_b^{(e)}(x_i,x_j))$ converges to zero for $N\to\infty$.

The proofs of relationships (3.20), (3.21).

The inequality (3.20).

Inequality (3.20) is proven on the basis of the Hoeffding inequality (Hoeffding, 1963) of the form:

$$P(\sum_{k=1}^{N}Y_k-\sum_{k=1}^{N}E(Y_k)\ge Nt)\le\exp\{-2Nt^2\}, \tag{A3.2}$$

where:

Y_k $(k=1,...,N)$ - independent random variables, with the same distributions and finite expected values and variances,

t – positive constant.

Inequality $P(W_{bN}^{(e)*}<\widetilde{W}_{bN}^{(e)})$ can be transformed to the form $1-P(W_{bN}^{(e)*}-\widetilde{W}_{bN}^{(e)}\ge 0)$; the component $P(W_{bN}^{(e)*}\ge\widetilde{W}_{bN}^{(e)})$ can be evaluated on the basis of (A3.2) and the expected values of the components:

$$\sum_{I^{(e)*}\cap(\widetilde{J}^{(e)}-I^{(e)*})}(U_{bk}^{(e)*}(x_i,x_j)-\widetilde{V}_{bk}^{(e)}(x_i,x_j))+$$
$$\sum_{J^{(e)*}\cap(\widetilde{I}^{(e)}-J^{(e)*})}(V_{bk}^{(e)*}(x_i,x_j)-\widetilde{U}_{bk}^{(e)}(x_i,x_j)). \qquad (k=1,\ldots,N)$$

Application of the inequality (A3.2) and of the expectations leads to the inequality (3.20).

Proof of inequality (3.21).

It can be proven that (Klukowski, 1994):

$$P(W_{bN}^{(e,me)*}<\widetilde{W}_{bN}^{(e,me)})\ge 1-2\lambda_N, \tag{A3.3}$$

where:

λ_N - the probability $P(U_b^{(e,me)*}(x_i,x_j)=1)$ $(<i,j>\in R_m)$.

Each probability $P(U_b^{(e,me)*}(x_i, x_j) = 1)$ satisfies the inequality (Zuiev, 1986):

$$P(U_b^{(e,me)*}(x_i, x_j) = 1) < \exp\{-2N(\tfrac{1}{2} - \delta)^2\}. \qquad \text{(A3.4)}$$

The inequalities (A3.3), (A3.4) imply the inequality (3.21).

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