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

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 5

Tests for relation type – equivalence or tolerance – for binary comparisons

5.1. Introduction

This chapter presents tests for detection of relation type, equivalence or tolerance, in the case of binary comparisons (see Klukowski 2006, 2008c). Such problem appears, e.g., when the set **X** can comprise brothers and sisters (equivalence relation) or stepbrothers and stepsisters (tolerance relation) and the actual relation is not known. The comparisons indicate only the existence or lack of any kind of consanguinity. In other words the comparison indicates inclusion both elements to common subset or not, while a relation type is unknown.

5.2. Tests based on the estimator in the form of sums of inconsistencies

The tests are constructed under the assumption that one of the relations considered – equivalence or tolerance - exists in the set **X**, but is a priori unknown. The basis for the tests are comparisons and estimates of both relations. The comparisons, denoted $g_{bk}^{(\cdot)}(x_i, x_j)$, indicate belonging to the same subset $\chi_q^{(\ell)*}$ $(\ell \in \{e, \tau\},\ 1 \le q \le n)$ or not. The estimates are based on sum of inconsistencies, i.e. tasks (3.2), (4.2) and will be denoted – respectively $\hat{T}_{bN}^{(e)}(x_i, x_j)$, $\hat{T}_{bN}^{(\tau)}(x_i, x_j)$. The test statistic proposed expresses differences between comparisons and values $\hat{T}_{bN}^{(\ell)}(x_i, x_j)$ $(\ell \in \{e, \tau\})$ for $\hat{T}_{bN}^{(e)}(x_i, x_j) \ne \hat{T}_{bN}^{(\tau)}(x_i, x_j)$. The statistic is a realization of some mixture of random variables (for details see Klukowski 2006).

The tests based on medians are similar to the case of *N*=1, but the distribution of the median has lower variance than the individual comparison.

The hypotheses verified assume the form:

H_0: the tolerance relation exists in the set **X**,
H_1: the equivalence relation exists in the set **X**,

or vice versa.

It is rational to assume the null hypothesis corresponding to the relation with the estimate providing lower value of the criterion function.

The set of pairs $I_s(x_i, x_j)$ with different values of estimates $\hat{T}_{bN}^{(e)}(x_i, x_j)$, $\hat{T}_{bN}^{(\tau)}(x_i, x_j)$ assumes the form:

$$I_s = \{(x_i, x_j) \mid \ \hat{T}_{bN}^{(e)}(x_i, x_j) \neq \hat{T}_{bN}^{(\tau)}(x_i, x_j)\}\,. \tag{5.1}$$

The test statistic is defined as follows:

$$S_N^{(e\tau)} = \frac{1}{N*card(I_s)} \sum_{k=1}^{N} \sum_{<i,j>\in I_s} \Big(\left| g_{bk}^{(\cdot)}(x_i, x_j) - \hat{T}_{bN}^{(e)}(x_i, x_j) \right| - \left| g_{bk}^{(\cdot)}(x_i, x_j) - \hat{T}_{bN}^{(\tau)}(x_i, x_j) \right| \Big). \tag{5.2}$$

The test statistic is based on comparisons corresponding to different estimates $\hat{T}_{bN}^{(e)}(x_i, x_j)$ and $\hat{T}_{bN}^{(\tau)}(x_i, x_j)$ of the same pair (x_i, x_j); the index of relation type is omitted in symbols $g_{bk}^{(\cdot)}(x_i, x_j)$ denoting comparisons.

The distributions of the variables: $\left| g_{bk}^{(\cdot)}(x_i, x_j) - \hat{T}_{bN}^{(e)}(x_i, x_j) \right|$ and $\left| g_{bk}^{(\cdot)}(x_i, x_j) - \hat{T}_{bN}^{(\tau)}(x_i, x_j) \right|$ $((x_i, x_j) \in I_s)$ are determined initially under the assumption that the estimate of the tolerance relation is errorless and the probability of such estimate $P(\hat{\chi}_1^{(\tau)}, ..., \hat{\chi}_{\hat{n}}^{(\tau)} \equiv \chi_1^{(\tau)*}, ..., \chi_n^{(\tau)*})$ is known. In the case, when all probabilities of incorrect comparisons are equal exactly δ, the distributions of differences $\left| g_{bk}^{(\cdot)}(x_i, x_j) - \hat{T}_{bN}^{(\tau)}(x_i, x_j) \right|$ assume the form (Klukowski, 2006, 2008c):

$$P\left(\left| g_{bk}^{(\cdot)}(x_i, x_j) - \hat{T}_b^{(\tau)}(x_i, x_j) \right| = 0 \ \middle| \ \hat{\chi}_1^{(\tau)}, ..., \hat{\chi}_{\hat{n}}^{(\tau)} \equiv \chi_1^{(\tau)*}, ..., \chi_n^{(\tau)*} \right) = 1 - \delta\,, \tag{5.3}$$

$$P(\left|g_{bk}^{(\cdot)}(x_i,x_j)-\hat{T}_b^{(\tau)}(x_i,x_j)\right|=1 \mid \hat{\chi}_1^{(\tau)},...,\hat{\chi}_{\hat{n}}^{(\tau)}\equiv\chi_1^{(\tau)*},...,\chi_n^{(\tau)*})=\delta\,. \tag{5.4}$$

Under the assumption $\hat{T}_{bN}^{(e)}(x_i,x_j)\neq\hat{T}_{bN}^{(\tau)}(x_i,x_j)$, the probability function of the random variable $\left|g_{bk}^{(\cdot)}(x_i,x_j)-\hat{T}_{bN}^{(e)}(x_i,x_j)\right|$ assumes the form:

$$P(\left|g_{bk}^{(\cdot)}(x_i,x_j)-\hat{T}_b^{(e)}(x_i,x_j)\right|=0 \mid \hat{\chi}_1^{(\tau)},...,\hat{\chi}_{\hat{n}}^{(\tau)}\equiv\chi_1^{(\tau)*},...,\chi_n^{(\tau)*})=\delta\,, \tag{5.5}$$

$$P(\left|g_{bk}^{(\cdot)}(x_i,x_j)-\hat{T}_b^{(e)}(x_i,x_j)\right|=1 \mid \hat{\chi}_1^{(\tau)},...,\hat{\chi}_{\hat{n}}^{(\tau)}\equiv\chi_1^{(\tau)*},...,\chi_n^{(\tau)*})=1-\delta\,. \tag{5.6}$$

Therefore:

$$\begin{aligned}E(&\left|g_{bk}^{(\cdot)}(x_i,x_j)-\hat{T}_b^{(e)}(x_i,x_j)\right|-\left|g_{bk}^{(\cdot)}(x_i,x_j)-\hat{T}_b^{(\tau)}(x_i,x_j)\right|\\&\mid \hat{\chi}_1^{(\tau)},...,\hat{\chi}_{\hat{n}}^{(\tau)}\equiv\chi_1^{(\tau)*},...,\chi_n^{(\tau)*})=1-2\delta.\end{aligned} \tag{5.7}$$

It can be shown, in a similar way, that in the case of the equivalence relation existing in the set **X** and errorless estimate $\hat{\chi}_1^{(e)},...,\hat{\chi}_{\hat{n}}^{(e)}$ the expected value of the test statistics (5.2) assumes the form:

$$\begin{aligned}E(&\left|g_{bk}^{(\cdot)}(x_i,x_j)-\hat{T}_b^{(e)}(x_i,x_j)\right|-\left|g_{bk}^{(\cdot)}(x_i,x_j)-\hat{T}_b^{(e)}(x_i,x_j)\right|\\&\mid \hat{\chi}_1^{(e)},...,\hat{\chi}_{\hat{n}}^{(e)}\equiv\chi_1^{(e)*},...,\chi_n^{(e)*})=2\delta-1.\end{aligned} \tag{5.8}$$

Moreover:

$$\begin{aligned}Var(&\left|g_{bk}^{(\cdot)}(x_i,x_j)-\hat{T}_b^{(e)}(x_i,x_j)\right|-\left|g_{bk}^{(\cdot)}(x_i,x_j)-\hat{T}_b^{(\tau)}(x_i,x_j)\right|\\&\mid \hat{\chi}_1^{(\tau)},...,\hat{\chi}_{\hat{n}}^{(\tau)}\equiv\chi_1^{(\tau)*},...,\chi_n^{(\tau)*})=4\delta(1-\delta),\end{aligned} \tag{5.9}$$

$$\begin{aligned}Var(&\left|g_{bk}^{(\cdot)}(x_i,x_j)-\hat{T}_b^{(e)}(x_i,x_j)\right|-\left|g_{bk}^{(\cdot)}(x_i,x_j)-\hat{T}_{bN}^{(\tau)}(x_i,x_j)\right|\\&\mid \hat{\chi}_1^{(e)},...,\hat{\chi}_{\hat{n}}^{(e)}\equiv\chi_1^{(e)*},...,\chi_n^{(e)*})=4\delta(1-\delta).\end{aligned} \tag{5.10}$$

Therefore, the standard deviation of the test statistics (5.2) assumes the form:

$$SD(S_N^{(e\tau)})=(Var(S_N^{(e\tau)})^{1/2}=2(\tfrac{1}{N\times card(I_s)}\delta(1-\delta))^{\frac{1}{2}}\,. \tag{5.11}$$

Thus, the hypothesis that the tolerance relation exists in the set **X** assumes the form:

H$_0$: $E(S_N^{(e\tau)}) = 1 - 2\delta$,
H$_1$: $E(S_N^{(e\tau)}) = 2\delta - 1$.

The test for such hypotheses can be constructed on the basis of Chebyshev inequality rested on the expected value and variance of the test statistics (5.2) (Fisz, 1969):

$$P(\left|S_N^{(e\tau)} - (1-2\delta)\right| > \lambda_\tau SD(S_N^{(e\tau)}) \mid \hat{\chi}_1^{(\tau)}, ..., \hat{\chi}_{\hat{n}}^{(\tau)} \equiv \chi_1^{(\tau)*}, ..., \chi_n^{(\tau)*}) < \tfrac{1}{\lambda_\tau^2}, \quad (5.12)$$

where:
λ_τ - positive constant.

The inequality (5.12) indicates the critical region:

$$\Lambda_\tau = \{S_N^{(e\tau)} \mid S^{(e)} < 1 - 2\delta - \lambda_\tau SD(S_N^{(e\tau)})\}. \quad (5.13)$$

The evaluation of the significance level, equal $\frac{1}{\lambda_\tau^2}$, resulting from the inequality (5.12), is valid in the case of errorless estimate. Taking into account the probability of such the estimate $P(\hat{\chi}_1^{(\tau)}, ..., \hat{\chi}_{\hat{n}}^{(\tau)} \equiv \chi_1^{(\tau)*}, ..., \chi_n^{(\tau)*})$, the significance level assumes the form:

$$\alpha_s^{(\tau)} = 1 - (1 - \tfrac{1}{\lambda_\tau^2}) P(\hat{\chi}_1^{(\tau)}, ..., \hat{\chi}_{\hat{n}}^{(\tau)} \equiv \chi_1^{(\tau)*}, ..., \chi_n^{(\tau)*}). \quad (5.14)$$

The value $\alpha_s^{(\tau)}$ is greater than $\frac{1}{\lambda_\tau^2}$. Typically, it exceeds the actual significance level, because results from the two-sided inequality (5.12). Let us note that for $N \to \infty$, the variance $SD(S_N^{(e\tau)})$ converges to zero and the probability $P(\hat{\chi}_1^{(\tau)}, ..., \hat{\chi}_{\hat{n}}^{(\tau)} \equiv \chi_1^{(\tau)*}, ..., \chi_n^{(\tau)*})$ converges to one.

The probability of the second type error is obtained in a similar way. Under the assumption that the estimate of the equivalence relation is errorless, the probability is evaluated by the relationship:

$$P(S_N^{(e\tau)} \geq 1 - 2\delta - \lambda_\tau SD(S_N^{(e\tau)}) \mid \hat{\chi}_1^{(e)}, ..., \hat{\chi}_{\hat{n}}^{(e)} \equiv \chi_1^{(e)*}, ..., \chi_n^{(e)*}) =$$

$$P(S_N^{(e\tau)} - (2\delta - 1) \geq 1 - 2\delta - (2\delta - 1) - \lambda_e SD(S_N^{(e\tau)}) \mid \hat{\chi}_1^{(e)}, ..., \hat{\chi}_{\hat{n}}^{(e)} \equiv \chi_1^{(e)*}, ..., \chi_n^{(e)*}) =$$

$$P(S_N^{(e\tau)} - (2\delta - 1) \geq \varsigma_\tau SD(S_N^{(e\tau)}) \mid \hat{\chi}_1^{(e)}, ..., \hat{\chi}_{\hat{n}}^{(e)} \equiv \chi_1^{(e)*}, ..., \chi_n^{(e)*}) \leq \tfrac{1}{\varsigma_\tau^2}, \qquad (5.15)$$

where:

$$\varsigma_\tau = \frac{2(1-2\delta) - \lambda_\tau SD(S_N^{(e\tau)})}{SD(S_N^{(e\tau)})}.$$

Thus, taking into account the probability of errorless estimate, the probability of the second type error is evaluated by:

$$\beta_e = 1 - (1 - \tfrac{1}{\varsigma_\tau^2}) P(\hat{\chi}_1^{(e)}, ..., \hat{\chi}_{\hat{n}}^{(e)} \equiv \chi_1^{(e)*}, ..., \chi_n^{(e)*}). \qquad (5.16)$$

It is clear that for $N \to \infty$ the probability $P(\hat{\chi}_1^{(e)}, ..., \hat{\chi}_{\hat{n}}^{(e)} \equiv \chi_1^{(e)*}, ..., \chi_n^{(e)*})$ converges to one and the standard deviation $SD(S_N^{(e\tau)})$ converges to zero. Therefore, the test is consistent.

The test for the equivalence relation is constructed in the same way (see Klukowski, 2006).

5.3. Tests based on the median estimator

The tests based on the median estimator exploit the same idea. They are similar to the case of one comparison for each pair, but the probability of the inequality:

$$P(g_b^{(\cdot,me)}(x_i, x_j) \neq \hat{T}_b^{(\ell)}(x_i, x_j)) = \eta_N^{(\ell)}, \qquad (5.17)$$

($g_b^{(\cdot,me)}(x_i, x_j)$ - symbol denoting, in this section, the median from comparisons of the same pair)

is determined by

$$\eta_N^{(\ell)} = \begin{cases} P(\sum_{k=1}^{N} g_b^{(\cdot,me)}(x_i,x_j) > \tfrac{N}{2};\ T_b^{(\ell)}(x_i,x_j)=0); \\ P(\sum_{k=1}^{N} g_b^{(\cdot,me)}(x_i,x_j) < \tfrac{N}{2};\ T_b^{(\ell)}(x_i,x_j)=1), \end{cases}$$

instead of δ. Therefore, the variance of each component $\left|g_b^{(\cdot,me)}(x_i,x_j) - \hat{T}_b^{(\ell)}(x_i,x_j)\right| - \left|g_b^{(\cdot,me)}(x_i,x_j) - \hat{T}_b^{(\ell)}(x_i,x_j)\right|$ of the test statistic – assumes the form:

$$\begin{aligned} &Var(\left|g_b^{(\cdot,me)}(x_i,x_j) - \hat{T}_b^{(\ell)}(x_i,x_j)\right| - \left|g_b^{(\cdot,me)}(x_i,x_j) - \hat{T}_b^{(\ell)}(x_i,x_j)\right| \\ &\quad \mid \hat{\chi}_1^{(\ell)}, \ldots, \hat{\chi}_{\hat{n}}^{(\ell)} \equiv \chi_1^{(\ell)*}, \ldots, \chi_n^{(\ell)*}) = 4\eta_N^{(\ell)}(1-\eta_N^{(\ell)}) \quad \ell \in \{e,\tau\}. \end{aligned} \tag{5.18}$$

The test statistic assumes the form:

$$\begin{aligned} S_N^{(e\tau,me)} = \tfrac{1}{card(I_s)} &\sum_{<i,j>\in I_s}(\left|g_b^{(\cdot,me)}(x_i,x_j) - \hat{T}_{bN}^{(e,me)}(x_i,x_j)\right| - \\ &\left|g_b^{(\cdot,me)}(x_i,x_j) - \hat{T}_{bN}^{(\tau,me)}(x_i,x_j)\right|); \end{aligned} \tag{5.19}$$

the critical region, for the hypotheses considered in the previous point, assumes the form:

$$\Lambda_N^{(me)} = \{S_N^{(e\tau,me)} \mid S_N^{(e\tau)} < 1 - 2\eta_N^{(\ell)} - \lambda_\tau^{(me)} \sigma_S^{(me,N)}\}, \tag{5.20}$$

where:

$$SD(S_N^{(e\tau,me)}) = (Var(S_N^{(e\tau,me)})^{1/2} = 2(\tfrac{1}{card(I_s)}\eta_N^{(\ell)}(1-\eta_N^{(\ell)}))^{½}.$$

The evaluation of significance level of the test assumes the form:

$$\alpha_{me}^{(\tau)} = 1-(1-\tfrac{1}{\lambda_\tau^2})P(\widehat{\chi}_1^{(\tau)}, \ldots, \widehat{\chi}_{\hat{n}}^{(\tau)} \equiv \chi_1^{(\tau)*}, \ldots, \chi_n^{(\tau)*}), \tag{5.21}$$

And evaluation of the probability of the second type error – the form:

$$\beta_{me}^{(\tau)} = 1-(1-\tfrac{1}{\lambda_\tau^2})P(\widehat{\chi}_1^{(e)}, \ldots, \widehat{\chi}_{\hat{n}}^{(e)} \equiv \chi_1^{(e)*}, \ldots, \chi_n^{(e)*}). \tag{5.22}$$

The formulas (5.19) – (5.22), defining the test, are similar to the formulas obtained for the case of sums of inconsistencies and *N*=1; the differences concern the probabilities (5.17), the variances (5.19) and the probabilities: $P(\widehat{\chi}_1^{(\tau)},...,\widehat{\chi}_{\hat{n}}^{(\tau)} \equiv \chi_1^{(\tau)*},...,\chi_n^{(\tau)*})$, $P(\widehat{\chi}_1^{(e)},...,\widehat{\chi}_{\hat{n}}^{(e)} \equiv \chi_1^{(e)*},...,\chi_n^{(e)*})$. The formulas indicate the consistency of the test. However, the simulation survey (Chapter 9) shows that the efficiency of the estimators based on medians is lower than of those based on the sum of inconsistencies. The same property characterizes the tests for the relation type.

5.4. Summary

The tests for determination of the relation form are a useful tool for the recovery of an actual data structure. An inappropriate model of data induces an incorrect estimate. The tests are based on weak assumptions about pairwise comparisons. The statistic is a realization of a mixture of random variables (Klukowski, 2006), because any estimate can be different than actual relation. An important feature of the tests is their consistency for $N \to \infty$.

An example of test application is presented in Klukowski (2006); the tests have been used for verification of the relation type in a set comprising empirical functions describing profitability of treasury bonds, resulting from auctions.

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