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 2

Estimation of relations – the main ideas

2.1. Introduction

The chapter presents the main ideas of the work in a concise way. The new, important features and properties of the approach presented in next chapters are introduced in this chapter.

2.2. Definitions and notations

The problem of estimation of relation on the basis of pairwise comparisons can be stated as follows.

We are given a finite set of elements $\mathbf{X} = \{x_1, ..., x_m\}$ ($3 \le m < \infty$). There exists in the set $\mathbf{X}$: the equivalence relation $\mathbf{R}^{(e)}$ (reflexive, transitive, symmetric), or the tolerance relation $\mathbf{R}^{(\tau)}$ (reflexive, symmetric), or the preference relation $\mathbf{R}^{(p)}$ (alternative of the equivalence relation and strict preference relation). Each relation generates some family of subsets $\chi_1^{(\ell)*}, ..., \chi_n^{(\ell)*}$ ($\ell \in \{p, e, \tau\}; n \ge 2$).

The equivalence relation generates the family $\chi_1^{(e)*}, ..., \chi_n^{(e)*}$ having the following properties:

$$\bigcup_{q=1}^{n} \chi_q^{(e)*} = \mathbf{X}, \tag{2.1}$$

$$\chi_r^{(e)*} \cap \chi_s^{(e)*} = \{\mathbf{0}\}, \tag{2.2}$$

where:
$\mathbf{0}$ – the empty set,

$$x_i, x_j \in \chi_r^{(e)*} \equiv x_i, x_j - \text{equivalent elements}, \tag{2.3}$$

$(x_i \in \chi_r^{(e)*}) \cap (x_j \in \chi_s^{(e)*}) \equiv x_i, x_j$ – non-equivalent elements for $i \neq j, r \neq s$. (2.4)

The tolerance relation generates the family $\chi_1^{(\tau)*}, ..., \chi_n^{(\tau)*}$ with the property (2.1), i.e. $\bigcup_{q=1}^{n} \chi_q^{(\tau)*} = \mathbf{X}$, and the properties:

$\exists r, s\ (r \neq s)$ *such that* $\chi_r^{(\tau)*} \cap \chi_s^{(\tau)*} \neq \{\mathbf{0}\}$,

$x_i, x_j \in \chi_r^{(\tau)*} \equiv x_i, x_j -$ equivalent elements, (2.5)

$(x_i \in \chi_r^{(\tau)*}) \cap (x_j \in \chi_s^{(\tau)*}) \equiv x_i, x_j$ – non-equivalent elements for $i \neq j$ and

$$(x_i, x_j) \notin \chi_r^{(\tau)*} \cap \chi_s^{(\tau)*}, \tag{2.6}$$

each subset $\chi_r^{(\tau)*}\ (1 \leq r \leq n)$ includes an element x_i such that

$x_i \notin \chi_s^{(\tau)*}\ (s \neq r)$. (2.7)

The preference relation generates the family $\chi_1^{(p)*}, ..., \chi_n^{(p)*}$ with the properties (2.1), (2.2) and the property:

$(x_i \in \chi_r^{(p)*}) \cap (x_j \in \chi_s^{(p)*}) \equiv x_i$ is preferred to x_j for $r<s$. (2.8)

The relations defined by the conditions (2.1) - (2.8) can be expressed, alternatively, by the values (functions) $T_{\upsilon}^{(\ell)}(x_i, x_j)$ $((x_i, x_j) \in \mathbf{X} \times \mathbf{X}$; $\ell \in \{p, e, \tau\}$, $\upsilon \in \{b, \mu\}$; symbols b, μ denote – respectively – the binary and multivalent comparisons), defined as follows:

$$T_b^{(e)}(x_i, x_j) = \begin{cases} 0\ \textit{if exists } r \textit{ such that } (x_i, x_j) \in \chi_r^{(e)*}, \\ 1\ \textit{otherwise}; \end{cases} \tag{2.9}$$

• the function $T_b^{(e)}(x_i, x_j)$, describing the equivalence relation, assuming binary values, expresses the fact if a pair (x_i, x_j) belongs to a common subset or not;

$$T_b^{(\tau)}(x_i, x_j) = \begin{cases} 0\ \textit{if exists } r, s\ (r = s \textit{ not excluded}) \textit{ such that} \\ \quad (x_i, x_j) \in \chi_r^{(\tau)*} \cap \chi_s^{(\tau)*}, \\ 1\ \textit{otherwise}; \end{cases} \tag{2.10}$$

• the function $T_b^{(\tau)}(x_i, x_j)$, describing the tolerance relation, assuming binary values, expresses the fact if a pair (x_i, x_j) belongs to any conjunction of subsets (also to the same subset) or not; the condition (2.7) guarantees uniqueness of the description;

$$T_\mu^{(\tau)}(x_i, x_j) = \#(\Omega_i^* \cap \Omega_j^*), \tag{2.11}$$

where:

Ω_l^* - the set of the form $\Omega_l^* = \{s \mid x_l \in \chi_s^{(\tau)*}\}$,

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

• the function $T_\mu^{(\tau)}(x_i, x_j)$, describing the tolerance relation, assuming multivalent values, expresses the number of subsets of conjunction including both elements; condition (2.7) guarantees the uniqueness of the description;

$$T_b^{(p)}(x_i, x_j) = \begin{cases} 0 \text{ if there exists } r \text{ such that } (x_i, x_j) \in \chi_r^{(p)*}, \\ -1 \text{ if } x_i \in \chi_r^{(p)*}, x_j \in \chi_s^{(p)*} \text{ and } r < s; \\ 1 \text{ if } x_i \in \chi_r^{(p)*}, x_j \in \chi_s^{(p)*} \text{ and } r > s; \end{cases} \tag{2.12}$$

• the function $T_b^{(p)}(x_i, x_j)$, describing the preference relation, assuming binary values, expresses the direction of preference in a pair or the equivalence of its elements;

$$T_\mu^{(p)}(x_i, x_j) = d_{ij} \Leftrightarrow x_i \in \chi_r^{(p)*},\ x_j \in \chi_s^{(p)*},\ d_{ij} = r - s; \tag{2.13}$$

• the function $T_\mu^{(p)}(x_i, x_j)$, describing the preference relation, assuming multivalent values, expresses the difference of ranks of elements x_i and x_j.

2.3. Assumptions about pairwise comparisons

The relation $\chi_1^{(\ell)*}, ..., \chi_n^{(\ell)*}$ is to be determined (estimated) on the basis of N ($N \geq 1$) comparisons of each pair $(x_i, x_j) \in \mathbf{X} \times \mathbf{X}$; any comparison $g_{\upsilon k}^{(\ell)}(x_i, x_j)$ evaluates the actual value of $T_\upsilon^{(\ell)}(x_i, x_j)$ and can be disturbed by a random error. The following assumptions are made concerning the comparison errors:

A1. The relation type, i.e.: equivalence or tolerance or preference, is known, the number of subsets n - unknown.

A2. Any comparison $g_{\upsilon k}^{(\ell)}(x_i, x_j)$ $(\ell \in \{e, \tau, p\};\ \upsilon \in \{b, \mu\};\ k = 1, ..., N)$, is the evaluation of the value $T_{\upsilon}^{(\ell)}(x_i, x_j)$, disturbed by a random error. The probabilities of errors $g_{\upsilon k}^{(\ell)}(x_i, x_j) - T_{\upsilon}^{(\ell)}(x_i, x_j)$ have to satisfy the following assumptions:

$$P(g_{bk}^{(\ell)}(x_i, x_j) - T_b^{(\ell)}(x_i, x_j) = 0 \,\big|\, T_b^{(\ell)}(x_i, x_j) = \kappa_{bij}^{(\ell)}) \geq 1 - \delta$$
$$(\kappa_{bij}^{(\ell)} \in \{-1, 0, 1\},\quad \delta \in (0, 1/2)), \tag{2.14}$$

$$\sum_{r \leq 0} P(g_{\mu k}^{(\ell)}(x_i, x_j) - T_{\mu}^{(\ell)}(x_i, x_j) = r \,\big|\, T_{\mu}^{(\ell)}(x_i, x_j) = \kappa_{\mu ij}^{(\ell)}) > 1/2$$

$$(\kappa_{\mu ij}^{(\ell)} \in \{0, ..., \pm m\},\ \ r - \text{zero or an integer number}), \tag{2.15}$$

$$\sum_{r \geq 0} P(g_{\mu k}^{(\ell)}(x_i, x_j) - T_{\mu}^{(\ell)}(x_i, x_j) = -r \,\big|\, T_{\mu}^{(\ell)}(x_i, x_j) = \kappa_{\mu ij}^{(\ell)}) > 1/2$$

$$(\kappa_{\mu ij}^{(\ell)} \in \{0, ..., \pm m\},\ r - \text{zero or an integer number}), \tag{2.16}$$

$$P(g_{\mu k}^{(\ell)}(x_i, x_j) - T_{\mu}^{(\ell)}(x_i, x_j) = r) \geq P(g_{\mu k}^{(\ell)}(x_i, x_j) - T_{\mu}^{(\ell)}(x_i, x_j) = r + 1 \,\big|\, T_{\mu}^{(\ell)}(x_i, x_j) = \kappa_{\mu ij}^{(\ell)}) \quad (\kappa_{\mu ij}^{(\ell)} \in \{0, ..., m\},\ r > 0), \tag{2.17}$$

$$P(g_{\mu k}^{(\ell)}(x_i, x_j) - T_{\mu}^{(\ell)}(x_i, x_j) = r) \geq P(g_{\mu k}^{(\ell)}(x_i, x_j) - T_{\mu}^{(\ell)}(x_i, x_j) = r - 1 \,\big|\, T_{\mu}^{(\ell)}(x_i, x_j) = \kappa_{\mu ij}^{(\ell)}) \quad (\kappa_{\mu ij}^{(\ell)} \in \{0, ..., m\},\ r < 0), \tag{2.18}$$

A3. The comparisons $g_{\upsilon k}^{(\ell)}(x_i, x_j)$ $(\ell \in \{e, \tau, p\};\ \upsilon \in \{b, \mu\};\ k = 1, ..., N)$ are independent random variables.

The assumption A3 makes it possible to determine the distributions of estimation errors of estimators proposed in this work. However, determination of the exact distributions of the (multidimensional) errors, in an analytic way, is complicated and in practice unrealizable. The main properties of the estimators, especially their consistency, are valid without the assumption.

The assumption A3 can be relaxed in the following way: the comparisons $g_{vk}^{(\ell)}(x_i, x_j)$ and $g_{vl}^{(\ell)}(x_r, x_s)$ $(l \neq k; r \neq i, j; s \neq i, j)$, i.e. including different elements, have to be independent.

In the case of the preference relation including equivalent elements, the condition (2.14) can be relaxed to the form (2.15) – (2.16).

The assumptions A2 – A3 reflect the following properties of distributions of comparisons errors:

- the probability of correct comparison is greater than of the incorrect one - in the case of binary comparisons (inequality (2.14));
- zero is the median of each distribution of comparison error (inequalities (2.14) – (2.16)),
- zero is the mode of each distribution of comparison error (inequalities (2.14) – (2.18));
- the set of all comparisons comprises the realizations of independent random variables;
- the expected value of any comparison error can differ from zero.

The assumptions about comparisons errors are not restricted. Especially, the errors can have non-zero expected values; the probabilities of errorless results have to satisfy the mode and median condition. These features guarantee broad spectrum of applications and protects against incorrect results.

2.4. The main idea of estimation – minimization of differences with comparisons

The main idea of the estimators proposed, i.e. minimization of differences between the relation and the pairwise comparisons, refers to a well-known principle. However, in the case under consideration, it does not indicate analytical properties, because it is not associated with minimization of the likelihood function or the sum of error squares. In our case, the properties of the estimators have been obtained on the basis of differences between the properties of the errorless estimate (actual form of the relation) and the estimates different from the errorless one. The properties have been proven

by the author on the basis of the well-known probabilistic inequalities (see Hoeffding, 1963, Chebyshev - for variance), properties of order statistics (David, 1970), and convergence of variances. The theoretical properties have been verified through the simulation survey.

Two forms of estimators are examined.

The estimates based on the total sum of differences, denoted $\hat{\chi}_1^{(\ell)}, ..., \hat{\chi}_n^{(\ell)}$ (or $\hat{T}_\upsilon^{(\ell)}(x_i, x_j)\ \ <i, j> \in R_m$), resulting from the minimization problem:

$$\min_{\chi_1^{(\ell)}, ..., \chi_r^{(\ell)} \in F_{\mathbf{X}}^{(\ell)}} \{ \sum_{<i,j> \in R_m} \sum_{k=1}^{N} \left| g_{\upsilon k}^{(\ell)}(x_i, x_j) - t_\upsilon^{(\ell)}(x_i, x_j) \right| \}, \tag{2.19}$$

where:

$F_{\mathbf{X}}^{(\ell)}$ - the feasible set, i.e. the family of all relations $\chi_1^{(\ell)}, ..., \chi_r^{(\ell)}$ of ℓ - th type in the set **X**,

$t_\upsilon^{(\ell)}(x_i, x_j)$ - the function describing any relation $\{\chi_1^{\ell)}, ..., \chi_r^{(\ell)}\}$ of ℓ -th type,

R_m - the set of the form $R_m = \{< i, j > \ \big| 1 \le i, j \le m;\ j > i\}$

(symbol $g_{\upsilon k}^{(\ell)}(x_i, x_j)$ is used for both random variables and realizations, because this does not lead to misunderstanding).

In the case of the preference relation and binary comparisons the following transformation is also applied:

$$\theta(g_{\upsilon k}^{(\ell)}(x_i, x_j) - t_\upsilon^{(\ell)}(x_i, x_j)) = \begin{cases} 0 \ \ if \ \ g_{\upsilon k}^{(\ell)}(x_i, x_j) = t_\upsilon^{(\ell)}(x_i, x_j); \\ 1 \ \ if \ g_{\upsilon k}^{(\ell)}(x_i, x_j) \ne t_\upsilon^{(\ell)}(x_i, x_j). \end{cases} \tag{2.19a}$$

The optimization problem, with the use of the transformation (2.19a), expresses the number of differences between the comparisons and the function $T_b^{(p)}(x_i, x_j)$. It is simpler from the computational point of view, because the variables $\theta(g_{\upsilon k}^{(\ell)}(x_i, x_j) - t_\upsilon^{(\ell)}(x_i, x_j))$ assume binary values (zero or one), while the difference $\left| g_{\upsilon k}^{(\ell)}(x_i, x_j) - t_\upsilon^{(\ell)}(x_i, x_j) \right|$ assumes values from the set $\{0, \pm 1, \pm 2\}$. The properties of both approaches (error measures) are similar (Klukowski, 1990b).

The estimate based on medians, denoted $\widehat{\chi}_1^{(\ell)}, ..., \widehat{\chi}_r^{(\ell)}$ (or $\widehat{T}_\upsilon^{(\ell)}(x_i, x_j)$), is obtained on the basis of the following minimization problem:

$$\min_{\chi_1^{(\ell)},...,\chi_r^{(\ell)} \in F_X} \{ \sum_{<i,j> \in R_m} \left| g_\upsilon^{(\ell,me)}(x_i,x_j) - t_\upsilon^{(\ell)}(x_i,x_j) \right| \}, \tag{2.20}$$

where:

$g_\upsilon^{(\ell,me)}(x_i,x_j)$ - the sample median (a middle value) in the set $\{g_{\upsilon,1}^{(\ell)}(x_i,x_j),...,g_{\upsilon N}^{(\ell)}\}$.

The estimate, resulting from the criterion (2.19) or (2.19a) will be denoted with symbols $\hat{\chi}_1^{(\ell)},...,\hat{\chi}_r^{(\ell)}$ or $\hat{T}_\upsilon^{(\ell)}(x_i,x_j)$, while the estimate resulting from the criterion (2.20) - with symbols $\tilde{\chi}_1^{(\ell)},...,\tilde{\chi}_r^{(\ell)}$ or $\tilde{T}_\upsilon^{(\ell)}(x_i,x_j)$.

In the case of the preference relation and medians from comparisons, the following transformation is also applied:

$$\theta(g_\upsilon^{(\ell,me)}(x_i,x_j) - t_\upsilon^{(\ell)}(x_i,x_j)) = \begin{cases} 0 \ \ if \ \ g_\upsilon^{(\ell,me)}(x_i,x_j) = t_\upsilon^{(\ell)}(x_i,x_j); \\ 1 \ \ if \ \ g_\upsilon^{(\ell,me)}(x_i,x_j) \neq t_\upsilon^{(\ell)}(x_i,x_j), \end{cases} \tag{2.20a}$$

instead of the difference $\left| g_\upsilon^{(\ell,me)}(x_i,x_j) - t_\upsilon^{(\ell)}(x_i,x_j) \right|$.

The transformation (2.20a) sums up the number of inconsistencies between the comparisons and the relation form, while the difference $\left| g_\upsilon^{(\ell,me)}(x_i,x_j) - t_\upsilon^{(\ell)}(x_i,x_j) \right|$ takes also into account the opposite direction of preference in a comparison. The optimization based on transformation (2.20a) is simpler to solve; and both approaches have similar efficiency (see Klukowski, 1990b).

It is clear that the number of estimates, resulting from the criterion functions (2.19), (2.19a), (2.20), (2.20a) can exceed one; the unique estimate can be determined in a random way or as a result of validation. Multiple estimates can appear also in other methods (see David 1988, Ch. 2). The minimal values of the respective functions are equal zero.

The assumptions A1 – A3 allow for inference about distributions of errors of estimates. Let us discuss first the estimator based on of the criterion (2.19). For each relation type one can determine a finite set including all possible realizations of comparisons

$$g_{\upsilon k}^{(\ell)}(x_i, x_j),\ (\ell \in \{e, \tau, p\}, \upsilon \in \{b, \mu\}, k = 1, ..., N; < i, j > \in R_m)$$

and the probability of each realization. The use of the criterion (2.19) determines: the estimate, its probability and estimation error. The error has the form: $\{\hat{T}_{\upsilon}^{(\ell)}(x_i, x_j) - T_{\upsilon}^{(\ell)}(x_i, x_j);\ < i, j > \in R_m\}$, i.e. it is a multidimensional random variable. The analysis of such error is, in fact, unrealizable and it is suggested to replace it with one-dimension error:

$$\hat{\Delta}_{\upsilon}^{(\ell)} = \sum_{< i, j > \in R_m} \left| \hat{T}_{\upsilon}^{(\ell)}(x_i, x_j) - T_{\upsilon}^{(\ell)}(x_i, x_j) \right|. \tag{2.21}$$

The estimate with the error $\hat{\Delta}_{\upsilon}^{(\ell)} = 0$ is the errorless estimate. The probability of such error can be determined in the analytic way – as a sum of probabilities of all realizations of comparisons indicating the errorless estimate. It is clear that its value (probability) depends on the number of comparisons N and the variance of comparison errors; increase of N decreases the probability of such error and decreases the variance of the estimator. The probabilities of errors different from zero can be determined in a similar way; all possible errors and their probabilities determine the distribution function of the estimation error. Determination of the probability function in the analytic manner is complicated and involves huge computational cost - even for moderate m. Therefore, simulation approach has to be used for this purpose (see Chapter 9). Simulation study provides complementary (to analytic results) knowledge about efficiency of estimators, especially useful in applications.

Similar considerations apply for the criteria (2.19a), (2.20), (2.20a).

2.5. Properties of estimators

The analytical properties of the estimators, established by the author, have mainly asymptotic character, i.e. they apply to the case $N \to \infty$. The properties guarantee the basic feature of the estimators - consistency. It is clear that errorless estimates can be also obtained for finite N, with probability close to one, because the number of variants (in optimization problems) is huge, but finite. In general, precision of estimates depends not only on N, but also on distributions of comparison errors and some features of the form of relation, e.g. the number of subsets n and the number of elements in each subset. The precision level is also not the same for both estimators considered. Simulation survey (Chapter 9) gives indications about the necessary number of N for given distributions of comparison errors.

The analytical properties of the estimators are based on properties of random variables expressing differences between pairwise comparisons and the relation form (expressed by $T_{\upsilon}^{(\ell)}(x_i, x_j)$). It has been demonstrated in the papers of the author that the variables corresponding to the actual relation form have different properties than the variables corresponding to any other relation. The following results have been obtained:

(i) the expected values of the variables, corresponding to actual relation form are lower than the expected values of variables corresponding to any other relation;

(ii) the variances of the variables expressing differences between comparisons and the relation form, both - actual and different than actual, divided by the number of comparisons N in the case of sum of differences, converge to zero for $N \to \infty$;

(iii) the probability of the event that the variable corresponding to actual relation assumes a value lower than the variable corresponding to a relation other than actual converges to one for $N \to \infty$; the speed of convergence guarantees good efficiency of the estimates.

Properties (i) - (iii) provide the basis for construction of estimators; these properties have been complemented with some additional features and a simulation study. An important result of the simulation survey consists in the fact that efficiency of the estimator based on the sum of inconsistencies is higher than of the median estimator; the latter estimator is, though, simpler from computational point of view and more robust with respect to outliers.
Let us illustrate these considerations by the simplest case, i.e. equivalence relation and the estimator resulting from the criterion (2.20). The differences between any comparison $g_{bk}^{(e)}(x_i, x_j)$ and the value $T_b^{(e)}(x_i, x_j)$ assume the form:

$$U_{bk}^{(e)*}(x_i, x_j) = \begin{cases} 0 \;\; if \;\; g_{bk}^{(e)}(x_i, x_j) = T_b^{(e)}(x_i, x_j);\; T_b^{(e)}(x_i, x_j) = 0; \\ 1 \;\; if \;\; g_{bk}^{(e)}(x_i, x_j) \neq T_b^{(e)}(x_i, x_j);\; T_b^{(e)}(x_i, x_j) = 0, \end{cases} \tag{2.22}$$

$$V_{bk}^{(e)*}(x_i, x_j) = \begin{cases} 0 \;\; if \;\; g_{bk}^{(e)}(x_i, x_j) = T_b^{(e)}(x_i, x_j);\; T_b^{(e)}(x_i, x_j) = 1; \\ 1 \;\; if \;\; g_{bk}^{(e)}(x_i, x_j) \neq T_b^{(e)}(x_i, x_j);\; T_b^{(e)}(x_i, x_j) = 1. \end{cases} \tag{2.23}$$

The sum of differences assumes, for any k $(1 \le k \le N)$, the form:

$$\sum_{<i,j>\in I^{(e)*}} U_{bk}^{(e)*}(x_i, x_j) + \sum_{<i,j>\in J^{(e)*}} V_{bk}^{(e)*}(x_i, x_j), \tag{2.24}$$

where:

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

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

The total sum of the differences between the relation form and the comparisons is equal:

$$W_{bN}^{(e)*} = \sum_{k=1}^{N} \left(\sum_{<i.j>\in I^{(e)*}} U_{bk}^{(e)*}(x_i, x_j) + \sum_{<i,j>\in J^{(e)*}} V_{bk}^{(e)*}(x_i, x_j) \right). \tag{2.25}$$

Under the assumptions A1, A2, A3, the expected values of the variables $U_{bk}^{(e)*}(x_i, x_j)$, $V_{bk}^{(e)*}(x_i, x_j)$ satisfy the inequalities: $E(U_{bk}^{(e)*}(x_i, x_j)) \le \delta$, $E(V_{bk}^{(e)*}(x_i, x_j)) \le \delta$. Therefore, the expected value of the variable $W_{bN}^{(e)*}$ satisfies the inequality $E(W_{bN}^{(e)*}) \le \frac{Nm(m-1)}{2}\delta$. Assumptions A1 – A3 allow for determining the variance $Var(W_{bN}^{(e)*})$; its value is finite and satisfies the inequality $Var(W_{bN}^{(e)*}) \le \frac{Nm(m-1)}{2}\delta(1-\delta)$.

Obviously:

$$E\left(\tfrac{1}{N} W_{bN}^{(e)*}\right) \le \tfrac{m(m-1)}{2}\delta, \tag{2.26}$$

$$\lim_{N\to\infty} Var\left(\tfrac{1}{N} W_{bN}^{(e)*}\right) = 0. \tag{2.27}$$

Let us consider any relation $\tilde{\chi}_1^{(e)}, ..., \tilde{\chi}_{\tilde{n}}^{(e)}$ different than $\chi_1^{(e)*}, ..., \chi_n^{(e)*}$; this means that there exist pairs (x_i, x_j), such that $\tilde{T}_b^{(e)}(x_i, x_j) \ne T_b^{(e)}(x_i, x_j)$. Define the random variables $\tilde{U}_{bk}^{(e)}(x_i, x_j)$, $\tilde{V}_{bk}^{(e)}(x_i, x_j)$, corresponding to the such values $\tilde{T}_b^{(e)}(x_i, x_j)$:

$$\widetilde{U}_{bk}^{(e)}(x_i,x_j)=\begin{cases}0 \;\; if \;\; g_{bk}^{(e)}(x_i,x_j)=\widetilde{T}_b^{(e)}(x_i,x_j);\;\; \widetilde{T}_b^{(e)}(x_i,x_j)=0;\\ 1 \;\; if \;\; g_{bk}^{(e)}(x_i,x_j)\neq\widetilde{T}_b^{(e)}(x_i,x_j);\;\; \widetilde{T}_b^{(e)}(x_i,x_j)=0,\end{cases} \tag{2.28}$$

$$\widetilde{V}_{bk}^{(e)}(x_i,x_j)=\begin{cases}0 \;\; if \;\; g_{bk}^{(e)}(x_i,x_j)=\widetilde{T}_b^{(e)}(x_i,x_j);\;\; \widetilde{T}_b^{(e)}(x_i,x_j)=1;\\ 1 \;\; if \;\; g_{bk}^{(e)}(x_i,x_j)\neq\widetilde{T}_b^{(e)}(x_i,x_j);\; \widetilde{T}_b^{(e)}(x_i,x_j)=1.\end{cases} \tag{2.29}$$

The expected values $E(\widetilde{U}_{bk}^{(e)}(x_i,x_j))$, $E(\widetilde{V}_{bk}^{(e)}(x_i,x_j))$ assume the form:

$$\begin{aligned}E(\widetilde{U}_{bk}^{(e)}(x_i,x_j))=0*P(g_{bk}^{(e)}(x_i,x_j)=0 \;\big|\; T_b^{(e)}(x_i,x_j)=1)+\\ 1*P(g_{bk}^{(e)}(x_i,x_j)=1 \;\big|\; T_b^{(e)}(x_i,x_j)=1)\geq 1-\delta,\end{aligned} \tag{2.30}$$

$$\begin{aligned}E(\widetilde{V}_{bk}^{(e)}(x_i,x_j))=0\times P(g_{bk}^{(e)}(x_i,x_j)=0 \;\big|\; T_b^{(e)}(x_i,x_j)=0)+\\ 1\times P(g_{bk}^{(e)}(x_i,x_j)=1 \;\big|\; T_b^{(e)}(x_i,x_j)=0)\geq 1-\delta,\end{aligned} \tag{2.31}$$

and:

$$E(\widetilde{W}_{bN}^{(e)})=\sum_{k=1}^{N}(\sum_{\widetilde{I}^{(e)}}E(\widetilde{U}_{bk}^{(e)}(x_i,x_j))+\sum_{\widetilde{J}^{(e)}}E(\widetilde{V}_{bk}^{(e)}(x_i,x_j)))>\frac{m(m-1)}{2}\delta. \tag{2.32}$$

The formulae (2.26)–(2.32) 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 $\widetilde{\chi}_1^{(p)},...,\widetilde{\chi}_{\tilde{n}}^{(p)}$. The variances of both variables converge to zero for $N\to\infty$. The variables $U_{bk}^{(e)*}(x_i,x_j)$, $V_{bk}^{(e)*}(x_i,x_j)$ assume values equal to $\left|g_{bk}^{(e)*}(x_i,x_j)-T_b^{(e)}(x_i,x_j)\right|$, used in the criterion function (2.19). Moreover, it can be also shown (see Klukowski, 1994), that:

$$P(W_{bN}^{(p)*}<\widetilde{W}_{bN}^{(p)})\geq 1-\exp\{-2N(\tfrac{1}{2}-\delta)^2\}. \tag{2.33}$$

The above facts indicate that the estimator $\hat{\chi}_1^{(p)},...,\hat{\chi}_{\hat{n}}^{(p)}$, minimizing the number of inconsistencies with comparisons, guarantees the errorless estimate for $N\to\infty$. The inequality (2.33) shows that the errorless estimate

can be obtained with the probability close to one for finite N. Moreover, the inequality indicates the influence of δ and N on the precision of the estimator. The distribution of an error of the estimator, for given parameters, has to be evaluated with the use of simulation approach.

The properties of the median estimator are based on the fact that the random variables $\frac{1}{N}\sum_{k=1}^{N} U_{bk}^{(e)*}(x_i, x_j)$ and $\frac{1}{N}\sum_{k=1}^{N} V_{bk}^{(e)*}(x_i, x_j)$ converge, with probability one, to a limit equal or lower than δ, for $N \to \infty$. Therefore, the median $g_b^{(e,me)}(x_i, x_j)$ converges to the actual value $T_b^{(e)}(x_i, x_j)$. As a result, minimization of (2.20) guarantees that the estimate $\hat{\chi}_1^{(e)}, ..., \hat{\chi}_{\bar{n}}^{(e)}$ converges to $\chi_1^{(e)*}, ..., \chi_n^{(e)*}$. Moreover, it can be shown (see Klukowski, 1994) that:

$$P(W_{bN}^{(p,me)*} < \tilde{W}_{bN}^{(me,p)}) \geq 1 - 2\exp\{-2N(\tfrac{1}{2} - \delta)^2\}. \tag{2.34}$$

Inequality (2.34) gives some evaluation of precision of the median estimator; the evaluation of error of the estimator has been obtained with the use of simulation (see Chapter 9).

The results presented in Klukowski (1994) include some additional inequalities and evaluations, especially for the case of single comparison for each pair. They are not repeated in this work, which concentrates on multiple comparisons. Moreover, simulation survey covers and completes some of these results.

The above considerations are valid also in the case of the tolerance and preference relations, estimated with the use of binary comparisons.

The case of multivalent comparisons, can be analyzed in a similar way. However, the considerations are more complicated from the analytical point of view – the details are presented in Chapters 6 and 8.

2.6. Validation of estimates

The estimators of relations are based on the assumptions A1–A3. The crucial assumption A1 states that the relations exist and their type is known, the assumptions A2 and A3 establish the properties of pairwise comparisons. These assumptions can be verified with the use of statistical tests; the positive result of verification validates the estimate obtained.

The first step of validation is to verify the assumptions on comparison errors. The assumptions A2 and A3 can be verified with the use of the well-known tests for independence, randomness, unimodality, and values of mode and median (see Daniel, 1990, Sheskin, 1997, Siegel and Castellan, 1988, Domański, 1979, 1990, Fraser, 1957, Hollander, Wolfe, 1973, Randles, Wolfe, 1979, Sachs, 1978). Such hypotheses can be tested on the basis of comparisons:

$$g_{\upsilon,1}^{(\ell)}(x_i,x_j),\ \ldots,\ g_{\upsilon N}^{(\ell)}(x_i,x_j)\ \ (\upsilon\in\{b,\mu\},\ <i,j>\in R_m,\ \ell\in\{e,\tau,p\})$$

or differences:

$$g_{\upsilon k}^{(\ell)}(x_i,x_j)-\hat{T}_{\upsilon}^{(\ell)}(x_i,x_j),\ g_{\upsilon k}^{(\ell)}(x_i,x_j)-\widehat{T}_{\upsilon}^{(\ell)}(x_i,x_j)\ (k=1,\ldots,N);$$

with the details given in Chapter 10.

The assumption of independence of the whole set of comparisons is difficult to verify; it seems more reliable to verify the assumption about independence of comparisons of individual pairs.

Verification of existence of a relation has to be done after the positive results of tests verifying the assumptions A2, A3 and has to be based on the estimates of the relation. Typical hypotheses verify the fact that the estimate is valid, i.e. the relation exists, under alternatives about the equivalency of all elements of the set **X** or randomness of comparisons or other data structure. Another basis for the verification is constituted by the optimal values of the functions (2.19), (2.19a) or (2.20), (2.20a); large values indicate significant differences with comparisons and suggest rejection of estimates. Critical values of such tests have to be obtained on the basis of simulations.

Some other features of estimates of relations can be used as the basis for verification, like, e.g., positive correlation of ranks of individual elements obtained on the basis of sequential subsets of comparisons:

$$g_{\upsilon,1}^{(\ell)}(x_i,x_j),\ \ldots,\ g_{\upsilon N}^{(\ell)}(x_i,x_j)\ \ (\ell\in\{e,\tau,p\},\ \upsilon\in\{b,\mu\},<i,j>\in R_m).$$

The tests for verification of relation type, i.e. equivalence or tolerance, and the weak or strong form of the preference relation have also been developed by the author (see Chapter 10).

2.7. Optimization problems

Minimization of the functions (2.19), (2.20) is, in general, not an easy problem, because of the dimensions of the feasible set. Currently, the algorithms are available only for ranking problems based on binary single comparisons (see David, 1988, Ch. 2); they refer to the dynamic programming or branch-and-bound algorithms. Some of them can be used for known n (see Cormen et al., 2001). The algorithms are efficient for the moderate number of elements m. In the case of large m, the problems can be also solved with the use of heuristic algorithms: genetic (Falkenauer, 1998), artificial neural networks, random search (Ripley, 2006), swarm intelligence (Abraham and Grosan, 2006), etc.

In the case of multivalent comparisons the exact algorithms are not available now. The problems with moderate number of elements m, i.e. 3–12, can be solved with the use of complete enumeration. Problems with higher number of elements can be solved using heuristic algorithms, mentioned above.

It is obvious that the estimators based on multivalent comparisons require more computations than those based on binary comparisons. However, speed of computers increases quickly and computational problems will disappear in a near future.

It seems that computers based on new quantum technology will allow for solving the problems without significant restrictions on the number of elements m. New optimization algorithms have to be developed for such computers.

2.8. Summary

This chapter shows the main ideas of estimation and validation, based on the concept of minimal inconsistencies, NAO. In general, the estimators proposed have a simple form, good statistical properties and are based on weak assumptions concerning the comparison errors. The properties of estimates, especially their precision, can be evaluated also in the case of unknown distributions of comparisons errors (see Chapter 10). In the case of appropriate number of comparisons (at least several) for each pair, the distributions can be estimated. The estimates of any relation can be

effectively validated and that in a variety of manners. In the case of doubt about relation type – equivalence or tolerance – it can be determined with the use of statistical tests developed by the author. Similar tests allow for distinction between the strict and weak form of the preference relation. Simulation survey allows for evaluation of precision of estimates and confirms their practical value. Moreover, it allows for determining the number of comparisons for given distributions of comparison errors. These features are of special importance from both theoretical and practical points of view. The approach proposed provides good estimates when other methods can produce incorrect results.

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