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 1

THE CONCEPT OF THE MONOGRAPH

Methods of estimation of relations of: equivalence, tolerance and preference in a finite set

1.1. Introduction

Estimation of the relations of equivalence, tolerance, or preference, on the basis of multiple pairwise comparisons with random errors, is aimed at determination of the actual structure of data. It also provides the properties of estimates. The properties comprise: consistency, distributions of errors, efficiency of estimators, etc. They allow for the statistical validation of estimates - including the assumptions concerning the comparison errors and existence of a relation.

The approach applied in the work rests on a statistical paradigm: to determine the relation form, which minimizes the inconsistencies (differences) with a sample - in the form of multiple pairwise comparisons. Such sample can be obtained with the use of statistical tests, expert opinions or other procedures, prone to generating random errors. The estimates are obtained on the basis of optimization tasks. However, the approach enables also extraction of some knowledge about relation form, which is a priori unknown. The example is determination of the type of a relation - equivalence or tolerance. Moreover, the entire estimation process: identification of relation type, estimation and validation of estimates, can be computerized. Therefore, such the approach can be included in the data mining techniques.

The approach presented here is an original contribution of the author to the subject. The main components of the contribution comprise: determination

of weak assumptions on comparison errors, definition of two types of estimators and two types of data – qualitative and quantitative, properties of the estimators, and validation of estimates. The assumptions allow for the extension of the application sphere of pairwise techniques. The estimators considered have different efficiency and computational cost – the results of the work allow for choosing the best approach. The estimators allow for combining of both types of comparisons. The results of the work provide a comprehensive solution to an important statistical problem.

The problems, which require estimation of relations of equivalence, tolerance, or preference, on the basis of pairwise comparisons with random errors, appear in many disciplines of knowledge: economy, finance, medicine, etc. The examples of such problems are:

- the preference relation: selection of the best form of econometric (or time series) model(s) (e.g. for inflation forecasting) on the basis of tests comparing the accuracy of two models (Vuong, 1989);

- the equivalence relation: partition of the set of functions, expressing hormonal profiles into non-overlapping subsets with the same shapes (see Klukowski, 1991);

- the tolerance relation: partition of a set of functions expressing profitability of the treasury securities into subsets with homogenous shapes, taking into account the existence of evolutionary shapes (overlapping partition) (see Klukowski, 2008c).

The work presents the synthesis of the results obtained by the author in this area, i.e. the estimators of the relations considered - based on different forms of pairwise comparisons. The idea of estimators is based on the concept of the nearest adjoining order (NAO – Slater, 1961, David, 1988): to minimize the inconsistencies with the given sample of comparisons. The estimators proposed have two basic forms:

- minimization of the sum of inconsistencies between relation the form and the comparisons,

- minimization of the sum of inconsistencies between the relation form and the medians from multiple comparisons of each pair.

The comparisons are assumed also in two basic forms:

- binary – expressing qualitative features of a pair, e.g. the direction of preference, and

- multivalent – expressing quantitative features of a pair, e.g. the difference of ranks of elements.

The errors of pairwise comparisons are realizations of some random variables. The assumptions about distributions of errors are weaker than those commonly used in the literature (David, 1988); they are satisfied in the case of each rational scientific investigation. The estimators can be applied in the case of unknown distributions of comparison errors.

The estimators proposed by the author have good statistical (analytical) properties, obtained on the basis of: • properties of random variables expressing differences between the relation form and the pairwise comparisons, probabilistic inequalities (Hoeffding, Chebyshev), properties of order statistics (David, 1970). The properties guarantee convergence of estimates to the actual relation form for the number of independent comparisons of each pair approaching infinity. Thus, the estimators are consistent. The study of the properties of the estimators has been complemented with the use of a simulation survey. It confirms their high efficiency, also for a finite number of comparisons, and allows for determination of parameters guaranteeing high precision of the estimates. The results of estimation can be verified with the use of statistical tests. The estimates are obtained on the basis of appropriate optimization tasks. Some of them can be solved with the use of existing algorithms. Remaining – complete enumeration of the feasible solutions or by heuristic approaches. The results of the work fill the gap between the methods, which require strong assumptions, and the methods based on heuristic rules, not vested with formal properties.

The work offers a significant contribution to the discipline of pairwise comparisons in the area of statistics and data mining. The results obtained by the author have been published in numerous publications (see Bibliography); some of the papers have been published or accepted for publication as a result of the project.

1.2. Purpose of the project

The main purpose of the project, whose results are reported here, is to synthesize the findings of the author, concerning estimation of relations of

equivalence, tolerance, and preference on the basis of pairwise comparisons with random errors. The results are based on the original concepts of the author.

Thus, the aims of the work comprise the following components:

1. Formulation of the estimation problems and the form of estimators (discrete optimization tasks).

2. Determination of the analytical properties of estimators, based on the sum of inconsistencies between the given comparisons on the one hand, and the relation form and the medians from comparisons of individual pairs, on the other hand - for binary and multivalent comparisons.

3. Examination of the properties of estimators, especially distributions of errors and the speed of convergence to the actual relation, with the use of the simulation approach.

4. Development of statistical tests for validation of the estimates.

1.3. Results obtained by the author

The following results have been obtained by the author in the framework of the project:

- determination of two main estimators of the relations considered, and examination of their properties; the estimators have different efficiency and computational cost; the first one (sum of inconsistencies) has higher efficiency and cost; the second (median) estimator entails lower computational cost and can be more robust in the case of outliers in comparisons;

- the estimators of the equivalence relation – based on binary comparisons; the estimators have the simplest form and the widest spectrum of properties;

- the estimators of the tolerance relation – based on binary and multivalent comparisons; the multivalent comparisons can express the number of common features or lacking features of two elements;

- the estimators of the preference relation – based on binary and multivalent comparisons; a multivalent comparison expresses the difference of ranks for a pair of elements; the estimators allow for combining of binary and multivalent comparisons in the case of the preference and tolerance relations;

- the two-stage estimators and combined estimators; the two-stage estimators are based, first, on binary comparisons and then on multivalent comparisons resulting from binary estimates, in the case of the preference and tolerance relations; combined estimators can use both type of comparisons;

- conclusions from the simulation survey, especially the evaluation of precision of estimates and the frequency of errorless estimates; the results indicate good efficiency of estimators, in particular – very fast convergence to errorless estimate – for a moderate number of comparisons;

- formulation of the assumptions about comparison errors, weaker than used in the literature of the subject, especially: • non-zero expected values of errors, • distributions of multivalent comparisons requiring only: unimodality and mode and median equal zero, • allowing for the probability of errorless comparison lower than ½ in case of multivalent comparisons, • possibility of application of estimators in the case of unknown distributions of errors, • possibility of non-independent comparisons of pairs containing common element, • possibility of cycles and reciprocal preference in comparisons - in the case of the preference relation; the conjunction of the above assumptions is significantly weaker than encountered in the literature.

The results constitute an important contribution to the area of pairwise comparisons. The work comprises main theoretical results; examples of application, while some proofs of theorems and detailed results have been presented in articles and conference papers.

1.4. Literature of the subject

The literature on pairwise comparisons with random errors concerns mainly ranking problems – classical results are presented in: David (1988), Bradley (1976, 1984), Davidson (1976) (bibliography), Brunk (1960), Davoodzadeh,

Beaver (1982). The authors mentioned present and discuss a complete range of existing methods: assumptions, estimators and their properties, tests for validation of results. In general, the assumptions required by the methods impose significant restrictions on probabilistic properties of comparisons; these assumptions constrain the application sphere. In particular, the comparisons can assume only the binary form, indicating the direction of the preference; some methods do not allow ties (equivalent elements). The basic methods are based on the linear model and the combinatorial models (David, 1988).

The linear model assumes the existence of a true ranking and imposes restrictive assumptions concerning preference probabilities, especially their distributions have to be known and comparisons have to be independent. The special cases are the well-known models: Thurstone-Mostller (David, 1988, Ch. 4, Maydeu-Olivares, 1999) and Bradley-Terry. The models have many different extensions – see David (1988, Ch. 4).

The combinatorial methods (David, 1988, Ch. 2) are based on minimization of inconsistencies between the ranking and the comparisons; a special case is constituted by the nearest adjoining order method – the idea is exploited and developed extensively in this work. The classical approaches have also complete theory: assumptions about distributions of pairwise comparisons, construction of estimators and tests for validation of results. Estimation procedures are based on the optimization approach: maximum likelihood, least squares, or discrete optimization. The nearest adjoining order approach, in its original form, corresponds to the Hamiltonian path in a graph (David, 1988, Ch. 2). There exist a number of discrete binary algorithms for solving the respective optimization problems (e.g. DeCani, 1972, Philips, 1969, Remage and Thompson, 1966); the results of optimization can be non-unique. It is clear that the classical approaches – minimization of the likelihood functions or the least squares method are very often not applicable, because they are based on assumptions which are not satisfied. Especially, probability distributions of comparisons can be unknown, comparisons not independent (in stochastic sense), expected values of errors not equal zero. This work presents estimators, which are both applicable and efficient in such cases.

The problems of estimation of the equivalence relation and tolerance relation are presented mainly in the literature concerning classification methods, e.g. Gordon (1999), Hand (1986), McLachlan (1992), Intelligent Data Analysis (2003), Webb (2003), Cormack (1971), and cluster analysis, Hartigan (1975), Kaufman, Rousseeuw (1990), Diday et al. (1994). The second

approach (clustering) rests typically on deterministic models; such approach is not examined in the present work. The main approaches are based on measures of similarity and dissimilarity of the classified elements – especially different kinds of distances. The key term of *pairwise comparisons* is not used in these problems, although distances, similarity or dissimilarity are features of pairs of elements. Thus, the methods presented in the literature are not equivalent to the approach presented here. In particular, in the present work it is assumed that distance equal zero (excluding random error) implies equivalent elements, while distance different from zero - different elements; in the case of qualitative data zero-one distances are applied. However, comparisons disturbed by random errors do not have these features. Both approaches here mentioned, i.e. classification and the nearest adjoining order, have some common features, in particular: • minimization of the sum of distances (several types of distances are used Gordon, 1999, Ch. 3), • unknown number of classes, • validation of the results with the use of statistical methods.

The literature concerning classification methods is extremely extensive. However, it should be emphasized that existing approaches do not cover entirely the problems presented in the work.

The scope of the work is limited by the purposes of the project. It comprises a slight part of the huge field of statistical methods for classification, discrimination and ranking of elements. These methods have been developed on the grounds of various disciplines: classical statistics, statistical learning (Hastie et al., 2002, Vapnik, 1998, Koronacki and Mielniczuk, 2005), data mining (Witten, 2005), neural networks (Ripley, 1996, Kohonen, 1995), fuzzy sets and other methods (Intelligent Data Analysis - Berthold, Hand, 2003), rough sets (Pawlak, 1982), immune systems (deCastro, 2002), swarm intelligence (Abraham, Grosan, 2006). The results developed by the author provide significant contribution to this area, with data in the form of pairwise comparisons. The main features of the approach proposed are:

- various forms of comparisons, i.e. binary (qualitative), multivalent (quantitative), and combined (allowing both forms),
- various form of estimators, i.e.: based on the sum of differences between comparisons and the relation form, based on medians from comparisons of each pair, and two-stage estimators based on binary and multivalent comparisons,
- weak (non-restrictive) assumptions about the comparison errors and similar optimization algorithms for all the problems considered; the assumptions admitted are satisfied in typical scientific investigations,

- possibility of evaluating the precision of estimates in the case of unknown distributions of comparison errors; the approach is based on the own concept of the author,
- possibility of versatile verification of estimates, guaranteeing high level of reliability,
- possibility of recognition of relation type on the basis of pairwise comparisons; this is possible in the case of: • equivalence relation and tolerance relation based on binary data and • different forms of preference relation – weak or strict.

1.5. Plan of the work

The book consists of eleven chapters covering the contents of the project. The theoretical issues and simulation results are presented in Chapters 2 – 11, devoted to the following topics:

- main ideas of estimation and validation, including: assumptions, forms of estimators and optimization algorithms (Ch. 2),
- definition of the estimators based on binary comparisons and their properties (Ch. 3, 4, 7, 9, 10),
- development of tests for determination of the relation type – equivalence or tolerance, i.e. extraction of information from pairwise comparisons (Ch. 5),
- definition of the estimators based on multivalent comparisons and their properties (Ch. 6, 8, 9, 10),
- simulation survey determining the efficiency of estimators of the preference relation (Ch. 9),
- methods of validation of estimates, including the tests for determination of the preference relation properties (Ch. 10),
- summary and conclusions (Ch. 11).

The presentation has a concise form – details and applications are presented in articles and conference papers. Applications of special (economic) importance are financial problems, i.e. concerning forecasting and optimization of public debt management (Klukowski, 2008a, Klukowski, Kuba, 2002). Medical applications, concerning hormonal profiles, are also important (Klukowski, 1991).

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