Identification of a Model of Individual Multifactor Estimation Using GMDH Element

Konstantin Petrov

Dept. of Applied Mathematics, Kharkov National University of Internal Affairs, 50-letiya SSSR Avenue, 27, 61080, Kharkov, Ukraine

kept@mail.ru

Abstract. *A method is proposed for decision a problem of formal structurally-parametric identification the individual multifactor estimation models within the framework of a decision making theory, which is based on the use of ideas of comparative identification theory. The model of multifactor estimation invariant to the dimension of factors, characterizing alternatives in the class of Kolmogorov-Gabor polynomials is synthesized. The methods of comparative parametric and structurally-parametric identification of a estimation models are developed on the basis of qualitative expert information about the relative preference of alternatives. The number of the applied problems is solving, particularly, identification of consumer marketing preferences, ordinal classification, social group behavioral management, quality estimation.*

Keywords

Decision-making, comparative identification, utility function, weight coefficient, Kolmogorov-Gabor polynomial.

1 Introduction

The intellectual procedure of a decision-making can be presented as a sequence of the next basic stages: the formation of a set of alternative decisions (ways of attainment of the objective); the estimation of the "quality" of possible decisions; the choice of the "best" decision. Thus, estimation is among the most important stages and the formalization of this intellectual procedure is the conceptual problem of synthesizing a model of the individual decision-maker's (DM) choice.

Estimates of alternative decisions formed by an individual are subjective, and they cannot be measured directly by any physical devices. In such cases, classical methods of direct identification of an estimation model are inapplicable. Alternative methods of indirect identification, for example, the comparative identification method [1] can be used.

2 Problem Statement

Let there be a set of alternatives (decisions) $X = \{ x_1, x_2, ..., x_m \}$ each of which is characterized by a collection of partial criteria (characteristics) $k_i(x_j)$, $i = \overline{1,n}$, $j = \overline{1,m}$, whose values are uniquely defined. Based on the analysis of this information, an individual can

1) choose the most preferable decision, for example, x_t from X;

2) rank all the decisions $x_j \in X$, $j = \overline{1,m}$, in decreasing (increasing) order of their preferability, i.e., to establish a strict (nonstrict) order relation, for example, $x_1 \; \mathbf{f}\,(\geq) x_2 \; \mathbf{f}\,(\geq) \mathbf{K} \; \mathbf{f}\,(\geq) x_m$, or a partial linear order relation, for example, $x_1 \; \mathbf{f}\,(\geq) x_2 \sim x_3 \; \mathbf{f}\,(\geq) \mathbf{K} \; \mathbf{f}\,(\geq) x_m$.

This means that, according to utility theory [2] postulating the existence of some scalar quantitative estimate (function) of preferability of any alternative $x_j \in X$, $j=\overline{1,m}$, we can write

$$P(x_t) > (\geq) P(x_j),\ \forall j \neq t,\ j=\overline{1,m}, \tag{1}$$

for the first situation and, accordingly,

$$P(x_1) > (\geq) P(x_2) > (\geq) \mathbf{K} > (\geq) P(x_m)$$

or

$$P(x_1) > (\geq) P(x_2) = P(x_3) > (\geq) \mathbf{K} > (\geq) P(x_m), \tag{2}$$

for the second situation, where $P(x_j)$ is an unknown scalar individual estimate of the utility of an alternative. Based on this information, it is necessary to synthesize a mathematical model of the individual DM's choice, i.e., a model of formation of the generalized utility $P(x_j)$.

The problem of synthesis of any mathematical model is connected with the solution of two problems, namely, the structural problem of definition of the form of the operator P and the parametric problem of definition of quantitative values of the model. The solution of the first problem is mostly based on the advancement of some hypothesis and postulation of the form of the model structure at the axiomatic level.

At the present time, two axiomatic forms of utility functions are used in utility theory, namely, the additive and multiplicative functions.

Of great theoretical and practical interest is the solution of the general problem of structurally-parametric identification of an individual estimation model under wider restrictive assumptions on the structure of the model.

To this end, we propose to use a fragments of a Kolmogorov-Gabor polynomials as a class of possible structures of the model. This polynomial is written as

$$P(x_j) = a_0 + \sum_{i=1}^{n} a_i k_i^N(x_j) + \sum_{i=1}^{n}\sum_{q=1}^{n} a_{iq} k_i^N(x_j) k_q^N(x_j) + \sum_{i=1}^{n}\sum_{q=1}^{n}\sum_{r=1}^{n} a_{iqr} k_i^N(x_j) k_q^N(x_j) k_r^N(x_j) + \ldots,$$

where a_j are relative dimensionless weight coefficients for which the following constraints are satisfied:

$$0 \leq a_j \leq 1,\ \sum_{j=1}^{v} a_j = 1,$$

and $k_i^N(x_j)$ are normalized (i.e., reduced to an isomorphic form) partial criteria. These criteria are normalized by the formula [3]

$$k_i^N(x_j) = \left(\frac{k_i(x_j) - k_i^w(x_j)}{k_i^b(x_j) - k_i^w(x_j)} \right), \tag{3}$$

where $k_i(x_j)$ is the value of a partial criterion and $k_i^b(x_j)$ and $k_i^w(x_j)$ are, respectively, the best and worst values of the partial criterion that are assumed by it in the domain of its admissible values $x_j \in X$, $j=\overline{1,m}$.

For a solve the general problem of structurally-parametric identification can be using elements of the GMDH [4] and genetic algorithms.

The choice of this polynomial is substantiated by the fact that it contains standard additive and multiplicative models as special cases and also allows one, if necessary, to take into account nonlinear terms of higher orders and all their possible combinations.

3 Estimation of Adequacy of a Multifactor Estimation Model

The specificity of the structurally-parametric identification problem being considered is that it is a Hadamard-incorrect problem since the systems of inequalities (1) and (2) specify some domain of admissible values and information is absent that allows one to generate a criterion for the choice of the uniquely best solution. Thus, as a criterion for the choice of a variant model, the number of satisfied inequalities (1) and (2) can be used. This means that any solution that is obtained during the sequential complication of the model and satisfies the system of inequalities is assumed to be the sought-for model.

The total amount of inequalities (1) and (2) is determined by the cardinality of the initial set of solutions in which an alternative preferable for the DM is chosen. If this set is sufficiently large, some number of inequalities can be selected from inequalities (1) and (2) as an external addition (a test set). In this case, the argumentativeness of the choice of a multifactor estimation model increases. However, even in this situation, it is impossible to quantitatively estimate the quality of the model being synthesized.

To analyze the constructibility of the synthesis method being considered in more detail and to determine quantitative estimates of quality of the obtained model, it is proposed to estimate it using a "quasiobjective" test choice situation playing the role of an independent external supplement.

To this end, a collection of alternatives $X = \{ x_1, x_2, ..., x_m \}$ is specified. Each of them is characterized by a tuple of partial characteristics $K(x_j) = \langle k_i(x_j) \rangle$, $i = \overline{1,n}$, $j = \overline{1,m}$, that are specified in the form of exact numerical values. Moreover, the structure of a reference estimation model is specified and numerical values of its parameters (weight coefficients a_i) are assigned. Thus, the choice situation and structurally-parametric characteristics of the reference choice model are completely and uniquely specified. Based on these data, reference values of multifactor estimates of utility of all alternatives $P^E(x_j)$, $j = \overline{1,m}$, are computed and an alternative x_t with the maximal value of reference utility, i.e., $x_t = arg \max\limits_{x_j \in X} P^E(x_j)$, $j = \overline{1,m}$, is determined. This alternative is considered as the DM's choice. A system of inequalities similar to system (1) is formulated for it, then the comparative structurally-parametric identification problem is solved, and, as a result, the structure and parameters of the sought-for model are determined. For this model, the values of model utility of all alternatives $P^M(x_j)$, $j = \overline{1,m}$, are also computed. The quality of the identified model is estimated by the comparison with the following criteria of the reference model:

- the degree of coincidence of structures (of polynomials);
- the average quadratic estimate of the difference between the values of the utility function of alternatives of the form

$$DP = \frac{1}{m} \sum_{i=1}^{m} \left[P^M(x_j) - P^E(x_j) \right]^2 ; \tag{4}$$

- the degree of coincidence between the reference and computational order relations over the test sequence of alternatives.

As the criterion of the correspondence between the reference order and the order of alternatives that is determined as a result of modeling the succession of alternatives according to their preference, it is proposed to use the coefficient of the Spearman rank correlation

$$r_s = 1 - \frac{6 \cdot \sum_{j=1}^{m} (R_j^E - R_j^M)^2}{m \cdot (m^2 - 1)}, \tag{5}$$

where R_j^E and R_j^M are the ranks (positions) of reference and model alternatives in the established order relation and m is the number of the alternatives being analyzed.

In the statement described, the comparative structurally-parametric identification problem can be solved by various methods. However, the common distinctive feature of all of them is the necessity of realization of the following sequence of procedures:

- formation of the model structure;

- determination of quantitative values of its parameters;
- estimation of the quality of the model.

The model structure can be formed using, for example, the method of sequential complication, GMDH, or genetic algorithms.

To solve the problem of parametric identification of an estimation model, one can use one of comparative identification methods (for example, by the definition of the Chebyshev point), random search methods, or genetic algorithms.

Various compositions of algorithms of the first and second stages can be used that differ in accuracy, computational complexity, and universality. To develop estimates of quality and universality of methods, they should be investigated. Algorithms implementing elements of GMDH are synthesized and investigated below with the help of computing experiments.

4 Solution of the Problem of Structurally-Parametric Identification of a Utility Function, Using Elements of GMDH

The GMDH is based on the partitioning of a set of variables into combinatory groups (that mostly consist of pairs). For each group, the simplest polynomial (the model structure) is formed and the problem of parametric identification is solved for it on the basis of an experimental sample. This problem can be solved by approximation of experimental data on numerical values of the input and output of the object being investigated by the least squares method. The parametric identification problem is solved for each local group model on the basis an experimental training sequence, and then its quality is determined from a test sequence with respect to a given criterion [4]. On this basis, a prescribed number of best quality models of the first level are selected that are again grouped and polynomials similar to those of the first level are constructed for them. The procedure is repeated until a polynomial of optimal complexity [4] is obtained.

Let us consider the distinctive features of using the GMDH for the solution of the problem of identification of a multifactor estimation model, taking into account that the model output (the utility of an alternative) cannot be quantitatively measured.

For example, let four partial criteria $k_i^N(x_j)$, $i=\overline{1,4}$, be given. We compose all non-recurring paired combinations (groups) of partial criterions $(k_1^N(x_j),k_2^N(x_j))$; $(k_1^N(x_j),k_3^N(x_j))$; $(k_1^N(x_j),k_4^N(x_j))$; $(k_2^N(x_j),k_3^N(x_j))$; $(k_2^N(x_j),k_4^N(x_j))$; $(k_3^N(x_j),k_4^N(x_j))$, $j=\overline{1,m}$.

We determine the optimal model in the class of truncated Kolmogorov-Gabor polynomials after eliminating the free term from it and restricting ourselves only to linear and quadratic terms (squares and paired combinations of variables).The introduced constraint on the complexity of a polynomial is based on the fact that, after normalization by formula (3), all partial characteristics assume values $0 \le k_i^N(x_j) \le 1$. The raising of such numbers to a power larger than two or the product of several (more than two) of such numbers leads to a fast decrease in values. Moreover, each term of the polynomial is multiplied by a coefficient $0 \le a_i \le 1$. Based on the fact that the computation of utility of alternatives $P(x_j)$ and weight coefficients a_i with an accuracy higher than the second decimal place is inexpedient, one can draw the conclusion that the terms whose degree is more than two can be eliminated from a Kolmogorov-Gabor polynomial. We assume that partial polynomials are functions of the form

$$
\begin{aligned}
&Y_1(x_j)=a_{11}k_1^N(x_j)+a_{12}k_2^N(x_j)+a_{13}k_1^N(x_j)k_2^N(x_j),\\
&Y_2(x_j)=a_{21}k_1^N(x_j)+a_{22}k_3^N(x_j)+a_{23}k_1^N(x_j)k_3^N(x_j),\quad j=\overline{1,m},\\
&\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\\
&Y_6(x_j)=a_{61}k_3^N(x_j)+a_{62}k_4^N(x_j)+a_{63}k_3^N(x_j)k_4^N(x_j).
\end{aligned} \tag{6}
$$

For each of six polynomials of the system of equations (6), the numerical values of weight coefficients a_{pq}, $p=\overline{1,6}$, $q=\overline{1,3}$, are determined by the comparative identification method. To solve this problem, we use the procedure that is proposed in [5] and is based on genetic algorithms.

Based on the computed values of weight coefficients and using criteria (4) and (5), we rank the polynomials of the first level from polynomials (6) according to their quality. We exclude the polynomials that do not satisfy conditions (1) from further consideration. If such polynomials are absent, we eliminate the worst two polynomials, repeat the operation of paired grouping over the remained four polynomials, and obtain $(Y_1(x_j), Y_2(x_j))$; $(Y_1(x_j), Y_3(x_j))$; …; $(Y_3(x_j), Y_4(x_j))$.

For each pair, we form a polynomial of the second level of the form

$$
\begin{aligned}
&Z_1(x_j) = a_{11}Y_1(x_j) + a_{12}Y_2(x_j) + a_{13}Y_1(x_j)Y_2(x_j),\\
&Z_2(x_j) = a_{21}Y_1(x_j) + a_{22}Y_3(x_j) + a_{23}Y_1(x_j)Y_3(x_j),\ j = \overline{1,m},\\
&\quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\\
&Z_6(x_j) = a_{61}Y_3(x_j) + a_{62}Y_4(x_j) + a_{63}Y_3(x_j)Y_4(x_j)
\end{aligned}
$$

and perform parametric comparative identification.

The procedure is repeated until we obtain the model that provides the order relation that is "most coincident" with the reference relation on the set of alternatives. This model is assumed to be the model of optimal complexity.

5 Conclusions

The development of methods for synthesis of models of intelligent activity, in particular, multifactor estimation models that are not based on a subjective expert estimation is of great theoretical and practical importance for multi-objective optimization theory, multidimensional classification, quality estimation, etc. From this viewpoint, the development of "objective" methods of identification is promising. One of them is the method of comparative structurally-parametric identification of multifactor estimation models.

The GMDH ideas used in this work have allowed us to methodologically correctly solve the most complicated problem of generation of the structure of a model with rather small computing expenditures.

The results of numerical modeling substantiate the efficiency of using the GMDH together with other evolutionary methods for solution of the problem of comparative structurally-parametric identification of complicated nonlinear multifactor estimation models.

References

[1] Ovezgel'dyev A. O., Petrov K. E.: Comparative identification of models of intelligent activity. *Cybernetics and Systems Analysis*, No. 5, p. 48-58, 1996.

[2] Neumann J., Morgenstein O.: Theory of Games and Economic Behavior [Russian translation], Nauka, Moscow, 1970.

[3] Ovezgel'dyev A. O., Petrov E. G., Petrov K. E.: Synthesis and Identification of Models of Multifactor Estimation and Optimization [in Russian], Naukova Dumka, Kiev, 2002.

[4] Ivakhnenko A. G., Miiller J. A.: Self-Organization of Predictive Models [in Russian], Tekhnika, Kiev, 1985.

[5] Petrov E. G., Bulavin D. A., Petrov K. E.: Using genetic algorithms for solution of the problem of structurally-parametric identification of a model of individual multifactor estimation. *Probl. Bioniki*, No. 60, p. 17-27, 2004.