Self-organising Predictive Modelling of Complex Systems

Overview Paper

Abstract

Developed in 1968 by A.G. Ivakhnenko, the self-organising modelling approach introduced the idea of *external information* into modelling and a class of algorithms which inductively, systematically and autonomously evolve an *optimal complex model* from noisy observation data by employing *parameter and structure identification*. Later, many important analytical results were obtained forming the theory of *noise immunity modelling*. It was shown that inductively self-organised "non-physical" models of optimal complexity that match the unknown noise-to-signal ratio are usually more accurate for interpolation and forecasting purposes than "physical" models obtained deductively. Recent research and development resulted in *Networks of Active Neurons*, *parallel implementations* of diverse algorithms, and in *multi-level self-organisation* for modelling high-dimensional data sets and *interdependent systems of equations*.

Current efforts is targeting at self-organising modelling for decision-making of complex real-world problems. We think that integration of soft systems approaches, learning from data, and inductive modelling paradigms will lead to a *framework of new theories, methodologies, algorithms and technologies* for modelling, describing, predicting and controlling complex systems and interrelated processes in near real-time.

History of Self-organising Modelling

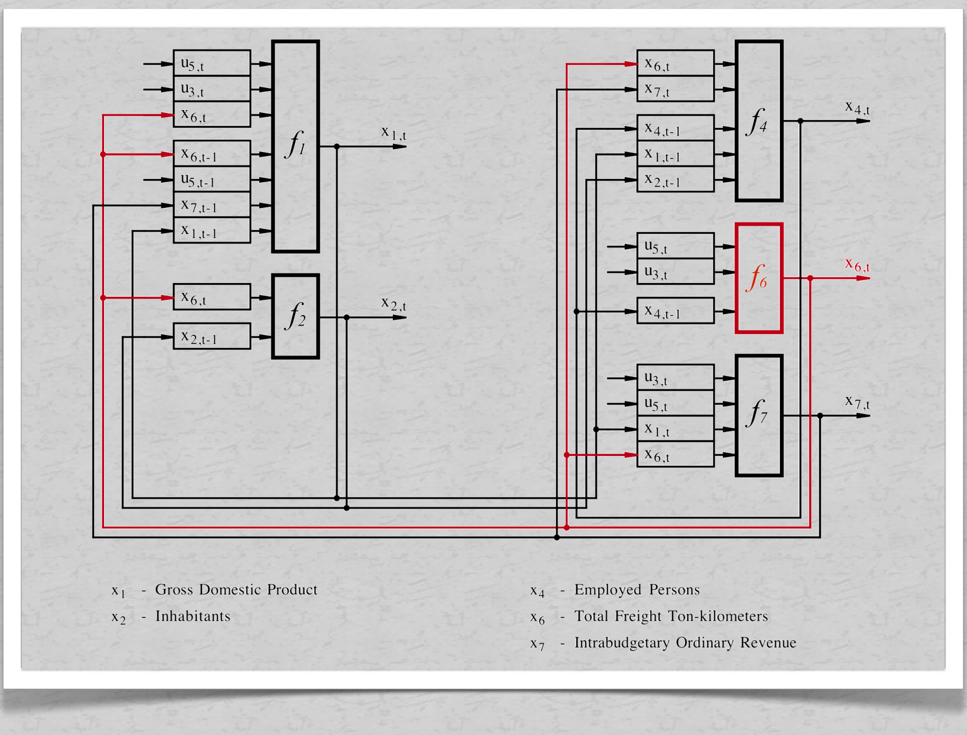
The idea of developing models from noisy data in a self-organising way has a long history. In 1968 the Ukrainian scientist A.G. Ivakhnenko based on statistical learning, perceptron, and pattern recognition concepts developed the first version of the Group Method of Data Handling (GMDH) algorithm as the core of self-organising modelling. He introduced the idea of *external information* into modelling by subdividing a dataset into training and testing data sets for model evaluation, and he used linear or second-order polynomial functions in two or three variables as elementary neuron transfer functions which inductively, systematically and autonomously evolve to an *optimal complex model* by employing *parameter and structure identification*. An optimal complex model is a model that optimally balances model fit on a given learning data set and its generalisation power on new, not previously seen data with respect to the data's noise level and the task of modelling (prediction, classification, modelling, etc.). It thus solves the basic problem of experimental systems analysis of systematically avoiding overfitted models based on the data's information only. Additionally, the self-organised model is available analytically in an explicit form of algebraic (static) or difference (dynamic) equations.

In the past more than 40 years of its history self-organising modelling technologies have been improving continuously. So the problem of modelling noisy data under incomplete information has been solved. Multi-criteria model selection and utilization of well-known a priori information was proposed. Experiments have shown that in this way models can be reliably self-organised from data sets that show a noise-to-signal ratio of up to 10. In the 1980s many important analytical results were obtained forming the theory of noise immunity modelling. It was shown, for example, that “physical models” couldn’t be used for long-term forecasting on noisy data. It was proven that self-organised “non-physical” models of optimal complexity that match the unknown noise-to-signal ratio are usually more accurate for interpolation and forecasting purposes than physical models whose parameters are estimated by regression analysis. Non-parametric modelling algorithms for describing fuzzy objects have been developed.

In the 1990s an original and efficient concept of neural networks with *active neurons* was introduced. The core idea here is in self-organising not only the network structure but, simultaneously, the individual optimal transfer function of each elementary neuron. This twice self-organised network of active neurons also results in various network topologies with heterogeneous neurons. They are based on competing types of transfer functions by choosing the optimal activation function for every neuron.

Recent research and development resulted in parallel implementations of diverse algorithms, multi-level self-organisation for modelling high-dimensional data sets of many thousand input variables and for modelling interdependent systems of equations (multi-input/multi-output systems), cost-sensitive modelling, and new model evaluation techniques to improve reliability and applicability of models to the user.

Today, self-organising modelling is a proven and highly efficient knowledge extraction technology and there exist powerful software that has been applied in various fields from image recognition over biomarker detection and QSAR modelling, wastewater management and reuse questions to Global Warming and micro and macro economic forecasting problems.



Self-organised predictive system of difference equations with endogenous and exogenous variables. Each model $f_i(x, u)$ is available analytically and represents a twice self-organised network of active neurons.

The Self-organising Modelling Approach

The traditional inductive approach of self-organising modelling was described by A.G. Ivakhnenko in 1968. This GMDH method has been further developed since then. However, one principle that share all GMDH variants and that makes it different from

other well known modelling and data mining methods is that of induction. The concept of induction is composed of three ideas:

- the principle of self-organisation for adaptively evolving a network model without subjective points given;
- the principle of external information to allow objective selection of a model of optimal complexity, and
- the principle of regularization of ill-posed tasks.

The GMDH algorithm is based on adaptive networks. Self-organisation is considered in identifying connections between the network units by a learning mechanism to represent discrete items. For this approach, the objective is to estimate networks of relevant and sufficient size with a structure evolving during the estimation process. A process is said to undergo self-organisation if identification emerges through the system's environment.

To realise a self-organisation of models from a finite number of input-output data samples the following conditions must exist to be fulfilled:

First condition: *There is a very simple initial organisation that enables the description of a large class of systems through the organisation's evolution.*

A common class often used is that of dynamic systems, which can be described by Volterra functional series. Discrete analogues of the Volterra functional series describing systems with a finite memory are higher-order polynomials of the Kolmogorov-Gabor form. For one input variable *x* it is:

$$y_t^M = k_0[t] + \sum_{s=0}^{g} a_s x[t-s] + \sum_{s_1}\sum_{s_2} a_{s_1} a_{s_2} x[t-s_1]x[t-s_2] + \ldots,$$

where k_0[t] is some trend function and g memory depth. If there is more than one input variable $x_j[t]$, $x_j[t\text{-}1]$, ..., $x_j[t\text{-}g_j]$ (t=1(1)$N\text{-}g$, g = max g_j) with j = 1(1) m+1 and $x_{m+1}[t\text{-}k+1] = y[t\text{-}k]$ (k=1(1)g_{m+1}) new variables v_{it} can be introduced, and the description will be:

$$y_t^M = \sum_{j=0}^{M} b_j v_{jt} + \sum_{j=1}^{M}\sum_{i\le j} b_{ij} v_{it} v_{jt} + \ldots = f_q\left(v_{0t}, v_{1t}, v_{2t}, \ldots, v_{Mt}\right),$$

where

$$v_{0t} = k_0[t], M = \sum_{j=1}^{m+1} g_j + m$$

and q order of the highest polynomial. This general model structure can be developed by networks of initial elementary models like:

$$f_1\left(v_i,v_j\right) = a_0 + a_1 v_i + a_2 v_j ,or$$
$$f_2\left(v_i,v_j\right) = f_1\left(v_i,v_j\right) + a_3 v_i v_j + a_4 v_i^2 + a_5 v_j^2 .$$

This is equivalent to describing a system by means of differential or difference equations.

Second condition: *There is an algorithm for mutation of the initial or already evolved organisations of a population.*

Genetic Algorithms are working on more or less stochastic mutations of the model structure by means of crossover, stochastic changes of characteristic parameters, and others. In the GMDH approach, a gradual increase of model complexity is used as a basic principle. The successive combination of many variants of mathematical models with increasing complexity has proven to be a universal solution in the theory of self-organisation presenting mutation in a way like that in biological selection processes. To apply this principle, a system of basic functions (first condition) is needed. Their appropriate choice, and the way the elementary models are combined to more complicated models decide the success of self-organisation. In most self-organising modelling algorithms, pairwise combination of *M* inputs is used to develop model candidates of growing complexity.

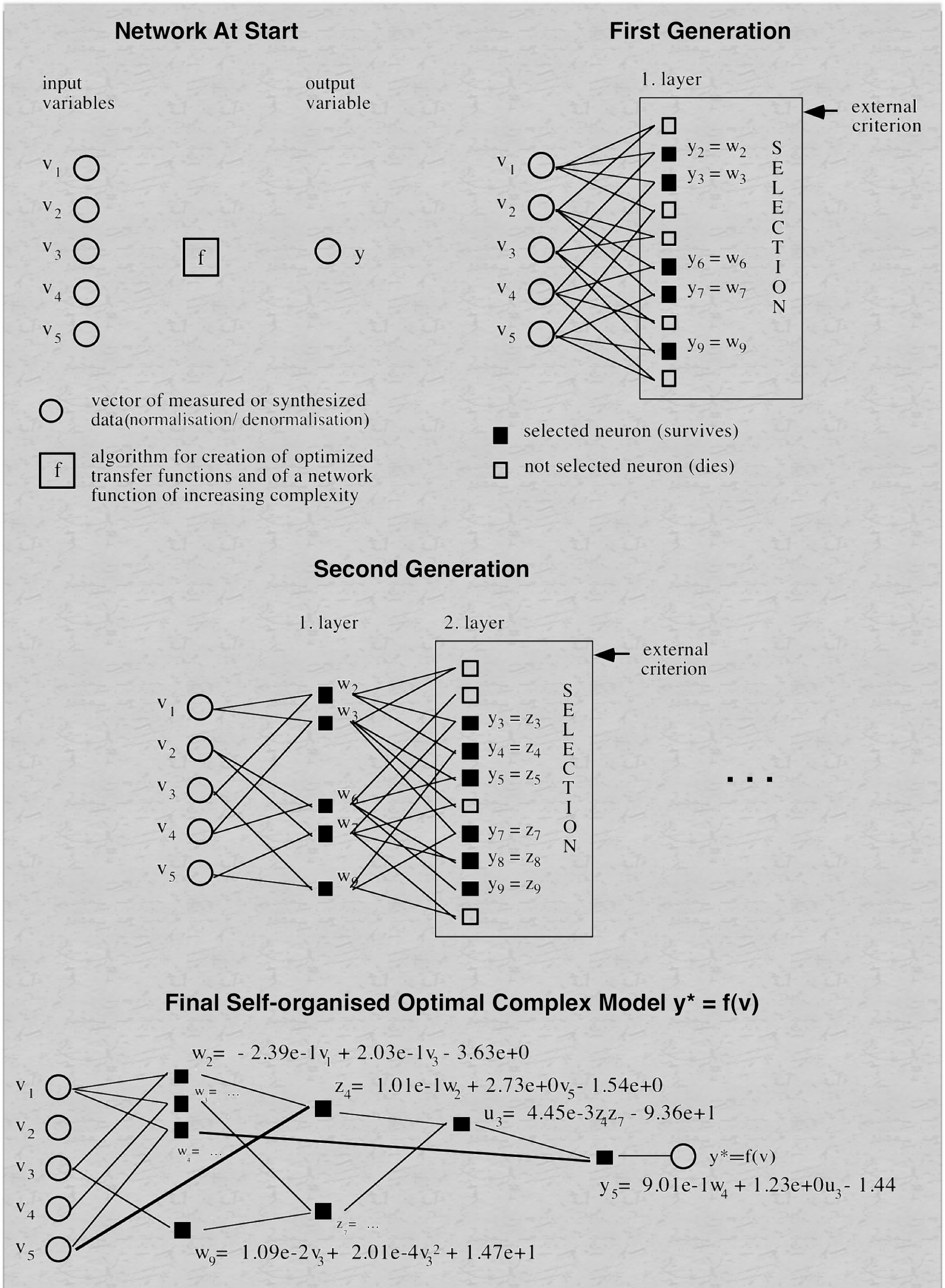
Third condition: *There is a selection criterion for measuring and validating the usefulness of a model relative to the intended task of modelling.*

According to this condition, several best model candidates are ranked and selected by an external selection criterion. The selected models survive, and they are in turn used as inputs for the following layer(s) to develop a new, more complex generation of models while the non-selected models die.

The principle of selection is closely linked to the principle of self-organisation in biological evolution; it is very important for the evolution of species. In the case of self-organising modelling, it is applied when the number of all possible model candidates in one generation is going to become too large for a complete induction. Using a threshold value, those model candidates are selected that are best in the sense of a given quality function, and they are stored as inputs for the next generation's model evolution.

The overall procedure of inheritance, mutation and selection stops automatically if a new generation of models provides no further model quality improvement. Then, a final optimal complex analytical model is obtained.

The process of self-organising modelling is summarised in the image below.



Network At Start
input variables
output variable
v 1
v 2
v 3
v 4
v 5
f
y
vector of measured or synthesized data(normalisation/ denormalisation)
f
algorithm for creation of optimized transfer functions and of a network function of increasing complexity
First Generation
1. layer
external criterion
y2 = w2
y3 = w3
y6 = w6
y7 = w7
y9 = w9
SELECTION
selected neuron (survives)
not selected neuron (dies)
Second Generation
1. layer
2. layer
external criterion
w2
w3
w6
w7
w9
y3 = z3
y4 = z4
y5 = z5
y7 = z7
y8 = z8
y9 = z9
SELECTION
. . .
Final Self-organised Optimal Complex Model y* = f(v)
w2= - 2.39e-1v1 + 2.03e-1v3 - 3.63e+0
w3 = ...
w4 = ...
z4= 1.01e-1w2 + 2.73e+0v5- 1.54e+0
u3= 4.45e-3z4z7 - 9.36e+1
y*=f(v)
y5= 9.01e-1w4 + 1.23e+0u3- 1.44
z7 = ...
w9= 1.09e-2v3+ 2.01e-4v3² + 1.47e+1

- *Self-selected input variables*: v_1, v_3, v_4, v_5
- *Self-organised Transfer Functions*: Active Neurons
- *Self-organised Network Topology*: Analytical Optimal Complex Model

Self-organisation of a Network of Active Neurons.

Challenges & Vision

Towards Self-organising Models from Data for Decision-Making of Complex Real-World Problems and Processes

Decision-making, whatever the field of human endeavour, requires formulation and a good understanding of what the problem is. To predict what may happen to a system under certain circumstances is often very difficult even for the simplest of systems, especially if they are not man-made. Humans have for centuries been seeking proxies for real processes. A substitute that can generate reliable information about a real system and its behaviour is called a model and they form the basis for any decision. It is worth building models to aid decision-making, because models make it possible to:

- *process* the time series involved, eliminate outliers and examine the integrity of the data (the subject of sensor data),
- *identify* the relationships between cause and effect (the subject of identification). This leads to a deeper understanding of the problem at hand by deriving an analytical relationship between them,
- *predict* the respective objects can expect over a finite future time span (the subject of prediction), but also to experiment with models. Exactly the ability to make predictions about the future forms the core of intelligence at all. Our brain uses a memory-prediction model to make continuous prediction of future events in parallel across all our senses,
- *simulate* the objects' behaviour by experiment with models, and thus answer "what-if" questions (subject of simulation) essential to decision-making,
- *control* the objects by finding suitable means to affect the objects (the subject of control) and enforce a specific behaviour.

The world around us is getting more complex, more interdependent, more connected and global. We can observe it, but we cannot understand it because of its complexity and the myriad interactions that are impossible to know let alone foresee. Uncertainty and vagueness, coupled with rapid developments radically affect humanity. Though we observe these effects, we most often do not understand the consequences of any actions, the dynamics involved and the inter-dependencies of real-world systems in which system variables are dynamically related to many others, and where it is usually difficult to differentiate which are the *causes* and which are the *effects*.

There are many cases in practice where it is impossible to create analytical models using classical theoretical systems analysis since there is incomplete knowledge of the processes involved. Environmental, medical and socio-economic systems are but three examples. We are facing complex problems, which do need decision-making, but the

means – the models – for understanding, predicting, simulating, and where possible controlling such systems are absent. This is an increasingly common situation in many real-world problems. To fill this increasing gap, new and appropriate inductive self-organising modelling methods have been theoretically and practically developed as powerful tools in revealing the missing implicit relationships within complex systems.

Recent research using new nature-inspired, inductive learning modelling and knowledge mining technologies based on self-organizing concepts shows breakthrough results and leads to an *intelligent inductive modelling technology* which is fundamental and key for the long-term vision. The concepts and results of this effort it is hoped will offer a new viewpoint to computational intelligence and lead to new lines of research that should provide answers to many open questions.

Mathematical modelling is at the core of many decision support systems. However, many problems in economics, ecology, biology, biochemistry, sociology, and life sciences, to name but a few, are ill-defined and can be characterized by:

- insufficient a priori information about the system for adequately describing the inherent system relationships,
- possessing a large number of variables, many of which are unknown and/or cannot be measured,
- noisy data available in very small to very large data sets,
- vague and fuzzy objects whose variables have to be described adequately.

Common to all modelling problems this means to:

- apply a systematic, holistic approach to modelling,
- take into account the incompleteness and inadequacy of *a priori* information about the real-world system,
- describe the vagueness and uncertainties of variables and, consequently, uncertainty of results and
- handle very small to very large sets of noisy data.

For ill-defined systems the classical hard approach that is based on the assumption that the world can be understood objectively and that knowledge about the world can be validated through empirical means needs to be replaced by a soft systems paradigm which can better describe vagueness and imprecision. This approach is based on the observation that humans only have an incomplete and rather vague understanding of the nature of the world but nevertheless are able to solve unexpected problems in uncertain situations.

Systems can be modelled through *deductive* logical-mathematical methods (theory-driven approach) or by *inductive* modelling methods (data-driven approach). Deductive methods have been used to advantages in cases of well-understood

problems and that obey well-known principles (*microscopic* approach). The spectacular results in aerospace are prime examples of this approach. Here, the theory of the object being modelled is well known and obeys known physical laws.

In contrast, *inductive* methods are used when *macroscopic* models (sometimes termed black box models) are the only alternative. These models are derived from real physical data and represent the relationships implicit within the system without or with only little knowledge of the physical processes or mechanisms involved.

In the real world there is a vast treasure trove of data that is being continuously amassed that contains useful information about the behaviour of systems. This is priceless information, which only needs to be trawled and suitably mined so as to transform it into useful knowledge that will expose the causal relationships between the principal variables. Theory-driven approaches to modelling are unduly restrictive to this end because of insufficient *a priori* knowledge, complexity and the uncertainty of the objects, as well as the exploding time and computing demands.

We maintain that there is a definite demand for a modelling methodology that can tackle real-world problems and which addresses such key questions like:

- Why a system behaves in a certain manner?
- Which causes lead to the observed effects?
- What will happen if certain causes are subject to changes?
- How are system variables inter-related?
- How can an object be manipulated to obtain desired effect?

From extensive experience with knowledge mining software developed by us based on concepts of inductive self-organisation and from applications in numerous fields, it is clear that reliable answers to these questions can only be obtained if the resulting model has sufficient validity, fidelity, transparency, and reproducibility to adequately explain observed phenomena and predict them over a finite future horizon. We expect that integration of soft systems approaches, learning from data, computational intelligence, and inductive modelling paradigms will lead to a new synergetic *framework of new theories, methodologies, algorithms and technologies* for modelling, describing, predicting and controlling complex systems and interrelated processes in near real-time.

We have been working in this direction for many years putting the focus not only on research but also always on real applications and cooperation as essential and invaluable part for development, innovation, and progress.

About us

KnowledgeMiner Software, Berlin, Germany

http://knowledgeminer.eu

KnowledgeMiner Software is a privately held company in the field of research, consulting, development, and application of unique self-organising, inductive, statistical learning modelling and knowledge discovery technologies for more than 15 years. The company is the developer of the KnowledgeMiner® software package, a distinguished commercial self-organising modelling and knowledge extraction tool. It implements a unique and innovative set of parallel algorithms for modelling, validation, and workflow processing of complex, high-dimensional systems to allow knowledge extraction from noisy data in a most objective, automated, and fast way.

The company cooperates with several international researchers well known in the field of self-organising modelling including the research group of the founder of the inductive modelling approach, A.G. Ivakhnenko, and scientists from Czech republic, China, and Greece.

One area where KnowledgeMiner Software has been active in more intensely in the past years is about modelling and prediction of toxicological and eco-toxicological hazards and risks of chemical compounds for regulatory purposes within REACH (http://ec.europa.eu/environment/chemicals/reach/reach_intro.htm) with the goal to increase consumer safety and to minimize and substitute animal tests by computer models. The company participated in three EC research projects related to QSAR modelling for regulatory purposes. Other fields of activity have been climate change related modelling and prediction problems (http://climateprediction.eu) sales and demand predictions, macro- and micro-economic modelling problems like national economy and balance sheet prediction, energy consumption analysis and prediction, medical diagnosis of diseases, traffic prediction, and wastewater reuse problems.

Since 1993 the company is run by Frank Lemke. He developed and implemented a number of innovative algorithms for validation of inductively built models including new approaches for application domain definition, cost-sensitive and ensemble modelling with per-compound prediction uncertainty, and self-organising knowledge extraction from high-dimensional variables space of noisy data.

Selected Publications

Original Work

Ivakhnenko A.G.: Group Method of Data Handling as a Rival of Stochastic Approximation Method, *Journal "Soviet Automatic Control",* No. 3 (1968), pp. 58-72. [Eng, USA]

Ivakhnenko A.G.: Heuristic Self-Organization in Problems of Automatic Control, *Automatica (IFAC),* No 6 (1970), pp. 207-219

Ivakhnenko A.G.: Polynomial theory of complex systems, *IEEE Trans. Sys., Man and Cyb.,* 1 (1971), No 4, pp. 364-378.

Books

Farlow, S.J. (ed.): **Self-Organizing methods in Modeling. GMDH Type Algorithm.** Marcel Dekker. New York, Basel. 1984, ISBN: 0-8247-7161-3

Madala, H.R., Ivakhnenko, A.G.: **Inductive Learning Algorithms for Complex Systems Modelling.** CRC Press Inc..Boca Raton, Ann Arbor, London, Tokyo. 1994, ISBN: 0-8493-4438-7

Müller, J.-A., Lemke, F.: **Self-Organising Data Mining**. Libri, Hamburg, 2000, ISBN: 3-89811-861-4

Papers and online sources

Global Warming Prediction Project: Open, transparent, objective, continued modelling and prediction of global temperature anomalies through self-organizing knowledge extraction from noisy data. (2010), http://climateprediction.eu

Lemke, F., Benfenati, E., Mueller, J.-A.: Data-driven Modeling and Prediction of Acute Toxicity of Pesticide Residues, *SIGKDD Explorations Special Issue*, 8(2006)1, pp. 71-79;
see also QSAR models for REACH: http://www.insilico.eu

Stathaki A and King R E : Mining the causality relationships of the pre-precipitation stage of a wastewater treatment process using self-organizing networks, *Int Journal of Environmental Engineering Science*, Vol 1, No 2, pp 159-168, 2010.

Stepashko, V.S.: Method of Critical Variances as Analytical Tool of Theory of Inductive Modeling, *Journal of Automation and Information Sciences*, 40, No. 3, (2008), pp. 4-22.