

Technologies of Numerical Investigation and Applying of Data-Based Modeling Methods

S.Yefimenko, and V.Stepashko,

International Research and Training Centre of Information Technologies and Systems of the National Academy of Sciences and Ministry of Education and Sciences of Ukraine, Glushkov ave., 40, Kyiv, 03680, Ukraine

astrid@irtc.org.ua, syefim@ukr.net

Abstract. *In the paper a general methodology of investigations of modelling methods from data observed with the use of statistical tests is suggested based on statement of a problem of choosing the most effective modelling algorithm. The software tools that realize the general methodology of investigations are developed. The software tools were used for investigation and applying of modelling methods. The practical problems of the automated models building were solved with the use of developed software tools.*

Keywords

Computer tests, modeling from data,
recurrent parameter estimation

1 Introduction

The problem of modelling from data observed consists in building of dependence between input and output parameters. Because of the presence of variety of the modelling methods and their software implementations there is the problem of choice of the most effective one of them for a concrete problem.

Whereas analytical researches concerning efficiency of modelling methods are still not sufficient it is wise to use numeral computer experiments for comparison of modelling methods and their elements.

Development of general methodology of statistical testing of modeling methods is the purpose of this work.

2 Problem statement of modeling methods testing

Every method of structural identification in an explicit or implicit form contains four components [1]:

- class of models;
- generator of model structures;
- method of parameters estimation;
- criterion of model selection.

Let F be set of classes of models, $F = \{i_k\}, k = \overline{1, K}$;

G - set of generators of structures of models, $G = \{i_h\}, h = \overline{1, H}$;

M - set of methods of parameters estimation, $M = \{i_e\}, e = \overline{1, E}$;

CR - set of criteria of models selection, $CR = \{i_t\}, t = \overline{1, T}$.

Then set of methods of structural identification S can be represented as direct product of sets

$$S = F \times G \times M \times CR.$$

As an algorithm we will understand the certain element of set S :

$$s_j = \{i_k, i_h, i_e, i_t\}, \quad i_k = \overline{1, K}, i_h = \overline{1, H}, i_e = \overline{1, E}, i_t = \overline{1, T}, \quad i = \overline{1, K \times H \times E \times T}$$

After that the problem of testing of modeling methods can be formulated as follows.

Let quality of every algorithm $s \in S$ be characterized by the value of some criterion $C(s)$. Then the best algorithm (in terms of criterion C) will be s^* satisfying condition:

$$s^* = \arg \min_{s \in S} C(s)$$

Such variants of quality criterion of computational algorithm are possible: accuracy, adequacy, processing speed, economy of computer memory, error of the model on independent data etc.

With the purpose of research of influence of modeling methods on the value of quality criterion it is necessary to compare different modeling methods as a whole and their structural elements.

3 Methodology of investigation of modeling methods

The purpose of investigation consists in comparative testing of classes of models, generators of model structures, methods of solving of linear equalizations systems for the problem of parameters estimation, criteria of model selection for determination of their efficiency by statistical experiment.

Thus, we have 4 variants of of experiments in accordance with the amount of basic components of modeling methods.

1. Testing of classes of models.

Determination of modelling efficiency (extrapolation or prognosis capabilities) is the possible purpose of testing with the use of classes of models different from true ones.

2. Testing of generators of models structures.

Purpose of testing is to define influence of such parameters as number of arguments, methods of models complexity increasing, computing power and other on the result of modeling.

Testing of generators of structures (when criterion of models selection is defined) is possible by following criteria of efficiency:

- obtaining of result of exhaustive search (in case of the use of non-combinatorial generators);
- run-time of algorithm (for different schemes and software implementations of combinatorial algorithm).

3. Testing of methods of solving of linear equalizations systems for the problem of parameters estimation when using the least-squares method.

Whereas the algorithms of combinatorial type are examined here, it is expedient to choose processing speed (or time of structural and parametrical identification) as the main criterion of efficiency.

4. Testing of criteria of models selection.

The possible purposes are to investigate:

- dependence of value of the defined criterion of optimum model (model that gives a minimum of the criterion) from its complication (amount of arguments) at the different values of noise variance;
- exactness of the models got with the use of different criteria on control data set.

4 The software tools for investigation and applying of modeling methods

For realization of the proposed methodology the software tools for investigation and applying of modeling methods were developed.

It enables to:

- construct the modelling methods from data observed;
- compare existing methods with the use of determined criteria;
- test different modelling methods and their components;

- develop techniques and plan the statistical tests;
- solve real world modelling problems;
- carry out the simulated experiments (approximation, extrapolation, etc.) with the models built by different modelling methods;
- enrich knowledge on the modelling methods with the use of the toolkit.

The software tools were used for investigation and applying of modeling methods and their components and for solving of practical problems of the automated models building [2].

5 Enhancement of efficiency of modeling methods on the basis of recurrent algorithms of parameters estimation

For the parameters estimation of model structures being sequentially complicated the use of algorithms recurrent in the number of parameters substantially decreases the computing time.

Recurrent bordering algorithm and features of it.

The bordering method is a traditional recurrent method [3].

The algorithm in the form suggested in [4] has a number of useful features [5]. But also it has substantial disadvantage: as investigations show it is not numerically stable in ill-conditioned problems.

Recurrent modifications of Gauss elimination algorithm and Gramm-Schmidt orthogonalization algorithm based on classical numerically stable methods are developed in [6].

The computational complexity (the number of elementary arithmetic operations) for the estimation of parameters by adding an argument s to a model of $s-1$ arguments is proportional to the second degree of model complexity for a recurrent algorithm (both the Gauss and Gramm-Schmidt) and to the third degree for a nonrecurrent one (see figure 1).

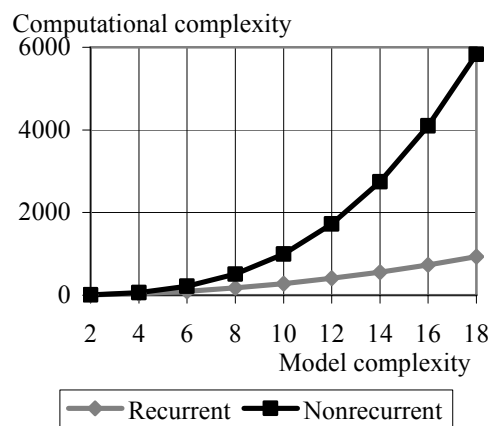


Fig. 1

6 Numerical results

6.1 Testing of classes of models

The modeling of the process $y = \sin \pi t + \xi$ where ξ is realization of noise vector was performed in the class of polynomial, autoregressive, and sinusoidal models. Testing purpose was to investigate extrapolation and approximation properties of the models built with the use of different classes of models.

The variable t was varied from 0 to 10 (81 point with a step 0,125). Vector ξ was generated by evenly distributed with a level 10% from the value of $\sin \pi t$.

Tab.1

Class of models	Autoregressio n $y_t = \sum_{i=1}^L \theta_i y_{t-i},$ $L = 10$	Polynomial $y = \sum_{i=1}^L \theta_i t^i,$ $L = 20$	Trigonometric $y = \sum_{i=1}^L \theta_i \sin(0,25\pi i t),$ $L = 8$
Accuracy of the best model on subsample C, ΔC	0,237	2E+15	0,125

The results represented in a table 1 show high prediction accuracy of autoregressive and trigonometric models. Polynomial model was very exact on training subsample and very inexact on examination subsample.

6.2 Testing of generators of models structures

The purpose is to investigate efficiency (processing speed) of different methods and software implementations of generators of binary structural vectors when modeling with the use of combinatorial algorithm

Run-time of generation of binary structural vector of all possible structures was measured. Number of arguments was varying from 20 to 25. The follows generators were used:

- standard binary;
- modified binary;
- standard successive;
- modified successive (Stepashko's scheme) .

The modified successive generator was the most fast one (Fig. 2).

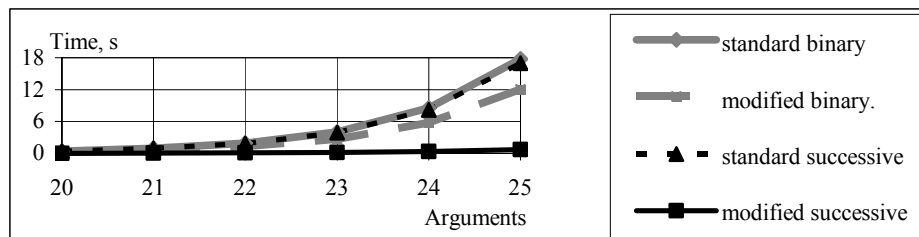


Fig. 2

6.3 Testing of methods of solving of linear equations systems for the problem of parameters estimation

Comparison of performance time of structure and parameter identification for the recurrent and nonrecurrent Gramm-Schmidt and Gauss algorithms

To check the effectiveness of the recurrent algorithms we compared by tests the performance time of structure and parameter identification for the recurrent and nonrecurrent (classical) Gramm-Schmidt and Gauss algorithms. Results of the experiments for the different methods of including regressors in a model confirm the theoretical estimations mentioned above.

6.4 Testing of criteria of models selection

Comparative testing of regularity (Ar), Mallows (Cp), and Akaike (FPE) criteria

Design matrix X of the size 12×15 was generated. Vector of output y was formed as a linear combination of the first ten regressors with addition of noise: $y = 10x_1 + 9x_2 + 8x_3 + 7x_4 + 6x_5 + 5x_6 + 4x_7 + 3x_8 + 2x_9 + x_{10} + \zeta$. In the class of the nested structures we selected the best model with 500 repetitions and averaged the results. Since the Mallows criterion contains the true value of the noise variance, it is possible to consider it as the ideal one. It shows how the optimum model complexity must decrease when increasing the noise level. According to the results represented in Figure 3 it is possible

to consider the regularity criterion to be effective, as opposed to the Akaike criterion which overfits the model by the noise level increasing.

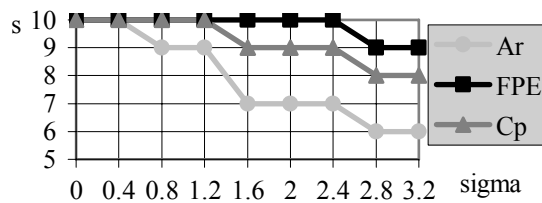


Fig. 3

Comparative testing of criteria of regularity (Ar) and Mallows (Cp)

In this experiment efficiency of the best models built by the criteria of regularity and Mallows was compared for different number of observations (from 10 to 40) and level of noise (from 0 to 80%). Models accuracy was compared on a control subset of ten points.

As table 1 shows both criteria have close accuracy at the small level of noise and a lot of observations.

In case of decreasing of number of points and increasing of noise level the regularity criterion gives more exact models.

Tab. 2

Level of noise %	Number of points							
	10		20		30		40	
	<i>Cp</i>	<i>AR</i>	<i>Cp</i>	<i>AR</i>	<i>Cp</i>	<i>AR</i>	<i>Cp</i>	<i>AR</i>
0	0,0013	0,0013	0,0012	0,0012	0,0011	0,0012	0,0008	0,001
40	94,2	75,7	80,6	69,98	59,6	60,9	45,9	50,6
80	189,6	121,3	166,6	105,4	118,9	82,3	80,6	62,99

7. Conclusion

The use of statistical experiments for investigation of modeling methods from data observed is explored and the general methodology of investigations is developed.

The software tools that realize the general methodology and recurrent parameters estimation are developed. The comparative testing of the modeling methods and their components were carried out and the practical problems of the automated models building were solved with the use of software tools.

References

- [1] Stepashko V.S. GMDH Algorithms as the basis for automation of the process of modeling from experimental data. – Soviet Journal of Automation and Information Sciences.– 1988.– No.4.– P.44-55.
- [2] Nizamov T.I., Ibrahimova S.R., Quluzade R.K., Isayev A.I., Stepashko V.S., Yefimenko S.N. Hydro-acoustic monitoring of water environment // Proceedings of Third International Conference on Technical and Physical Problems in Power Engineering, Ankara, Turkey, May 29-31, 2006, pp.1108-1110.
- [3] Seber G.A.F. Linear Regression Analysis. John Wiley and Sons, New York – London – Sydney – Toronto, 1977
- [4] Stepashko V.S. Optimization and Generalization of Model Sorting-out Schemes in Algorithms for the G.M.D.H., Soviet Automatic Control, 12, 4, (1979), Pp. 28-33
- [5] Stepashko V.S., Yefimenko S.M. On the Effectiveness of Recurrent Methods of Parameter Estimation in Macromodeling Problems//Proceedings of V of International Workshop "Computational Problems of Electrical Engineering", Jazleevets, Ukraine, August 26-29, 2003, pp.106-107.
- [6] Stepashko V.S. and Efimenko S.N. Sequential Estimation of the Parameters of Regression Model // Cybernetics and Systems Analysis, Springer New York, July, 2005, Vol. 41, Num. 4, pp.631-634.