Computer tests as an instrument for effectiveness investigation of modeling algorithms

Yefimenko, S.M., and Stepashko, V.S.

International Research and Training Centre of Information Technologies and Systems of the National Academy of Sciences and Ministry of Education and Sciences of Ukraine

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

Abstract. *Technology of testing algorithms effectiveness for structural models identification with the use of statistical computer tests is developed. The recurrent bordering algorithm is investigated. Recurrent modifications of well known Gauss and Gram-Schmidt methods are developed.*
The comparative testing of recurrent methods for solving of linear equations system in the parameters estimation problem as well as structural identification criteria of model construction is carried out. Effectiveness of parallel computation are explored with the purpose of extension of modeling possibilities from statistical data.
Real world problems are solved for modeling of the ferromolybdenum market price and upper sedimentary layer density of the Caspian Sea bottom.

Keywords

Computer tests, modeling from data,
recurrent parameter estimation, bordering method

1 Introduction

Currently there are various methods of modeling from data observed and variants of their software implementation. Such variety of methods complicates the rational choice the optimum one of them for a concrete practical problem. An optimum in this case can mean higher accuracy of the result, higher performance of the method, or less requirements to computer memory etc.

Thus, the successful solving a modelling problem for complex objects, processes and systems from data substantially depends on the choice of optimum method and consequently from getting acquainted with modelling methods. It is of importance first of all for an expert in data-based modeling domain which should make decisions on what method will be the most effective in a concrete case and also for a user wanting to apply the available modelling methods.

Growing processor speed, wide applying of computers for mathematical modelling, sufficiently powerful theoretical and experimental base allow using computer tests as an effective technology of knowledge discovery on effectiveness of modelling methods and their constituent elements.

2 Recurrent algorithms of parameters estimation as the basis of effective modelling methods

It is wise to use algorithms recurrent in the number of parameters in structural identification problems for the parameters estimation of model structures being sequentially complicated. The inefficiency of nonrecurrent methods is explained by the necessity to recalculate the extended matrix of normal systems (in the Gauss elimination method) or to orthogonalize new design matrix (in the Gramm-Schmidt orthogonalization method) every time with addition of a new regressor.

Recurrent bordering algorithm and features of it. The bordering method [1] is a recurrent method. The idea of the bordering algorithm (modification of a least-squares method) consists in a sequential correction of parameter estimations by the recurrent calculation of inverse matrix elements in the estimation process. Such algorithm in the form suggested in [2] has a number of useful features enabling to use additional information being obtained during the algorithm execution without need of any additional calculations [3].

At the same time the algorithm has a disadvantage: it is not numerically stable in ill-conditioned problems. That is why the construction task of recurrent algorithms for classical numerically stable ones is actual.

Recurrent modification of the Gauss elimination algorithm [4]. From the complete system of conditional equations $X\theta = y$ corresponding to the sample volume n for an object with m inputs and one output we proceed to the normal system $H\theta = g$ with the elements $H = X^T X = \left\{H_{ij};\, i, j = \overline{1,m}\right\}$ and $g = X^T y = \left\{g_i, i = \overline{1,m}\right\}$. Formulas on step s (when including an argument s to the model containing s-1 arguments) will look like as follows. For the forward motion:

$$H^1_{is} = \frac{H^0_{is} - \sum_{j=1}^{i-1} H^1_{ij} H^1_{js}}{H^1_{ii}}, \quad i = \overline{1, s-1}, \quad H^1_{si} = H^0_{si} - \sum_{j=1}^{i-1} H^1_{sj} H^1_{ji}, \quad i = \overline{1, s}, \quad g^1_s = \frac{g^0_s - \sum_{i=1}^{s-1} H^1_{si} g^1_i}{H^1_{ss}}.$$

For the backward motion:

$$\theta^s_i = g^s_i - \sum_{j=i+1}^{s} H^s_{ij} \theta^s_j, \quad i = \overline{s, 1}.$$

The modification consists in that at every step s the preliminary calculated column vector and row vector are added to the matrix H_{s-1} which is not recalculated. Reverse motion calculates the estimation of parameters θ_s.

Recurrent modification of Gramm-Schmidt orthogonalization algorithm [4]. Let us have the system of vectors v_i, i=1, 2, ..., s-1, at a step s-1 (columns of matrix V obtained by orthogonalization of the matrix X). At the step s, when adding to the model of a new argument (that is by solving of the equations system containing s arguments), we will get the vector v_s:

$$v^0_s = x_s - \sum_{j=1}^{s-1} (x_s^T v_j) v_j, \quad v_s = \frac{v^0_s}{\left\| v^0_s \right\|}.$$

After that we calculate the estimation of coefficients of the new system (taking into account an argument s):

$$\theta_i = \frac{y^T v_i - \sum_{j=i+1}^{m} (v_i^T x_j)\theta_j}{v_i^T x_i}, \; i = s, s-1, \ldots, 1.$$

An obvious modification of the algorithm consists in that at a step s we do not reorthogonalize matrix X_{s-1} but use previously orthogonalized matrix V_{s-1} by adding the column vector calculated at this step. After that we estimate parameters θ_s with the use of V_s.

Figure 1 shows a comparison of theoretical indices of 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.

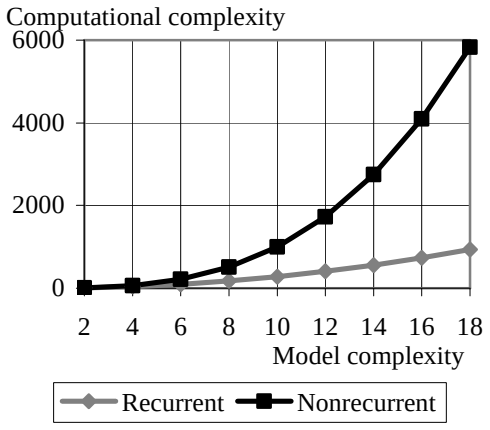

Figure 1

Such a dependence of the computational complexity is similar for both the Gauss and Gramm-Schmidt algorithms and is proportional to the second degree of model complexity for a recurrent algorithm and to the third degree for a nonrecurrent one.

3 The results of experiments

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.

Testing of the regularity criterion. Design matrix X of the size 8×12 was generated for the system of conditional equations. Vector y was formed as a linear combination of the first five regressors so that the true model looked like $y=5x_1+4x_2+3x_3+2x_4+x_5$ with addition of noise. Structural identification in the class of nested structures was executed (i.e. parameters estimation of the complicated structures which contain one argument at first, then two, and so on to the complete including of all five arguments in our case). Complexity of model corresponding to the minimum of the criterion was determined and the value of this criterion was calculated. The results were averaged by 500 repetitions.

Dependences of the criterion value on model complexity for different values of the noise variance are illustrated in Figure 2. The minimum of a criterion (lower curve) correspond to a model containing 5 true regressors under the absence of the noise. When the noise variance grows the minimum of the criterion is shifted toward more simple models comparing to the true model. The graph shows that the model containing one true argument is the best by the criterion value under the noise level of 100% from the signal. Marked points indicate the minimum of every curve.

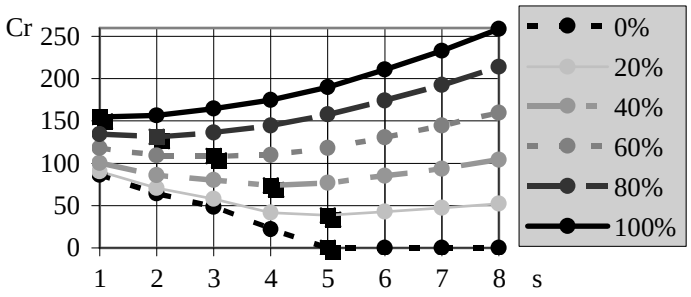

Figure 2

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: $y=10x1+9x2+8x3+7x4+6x5+5x6+4x7+3x8+2x9+x10+\xi$. 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.

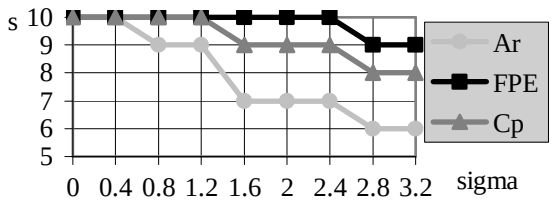

Figure 3

Parallel computation applications in the modeling problems. With the purpose of extension of modeling resources from data observed, the experiments with parallelization of combinatorial algorithm were executed. Run-time of the program (axis Z in Figure 4) was measured for different amounts of processors (from 1 to 8, axis X) at the different number of factors (number of matrix columns, from 20 to 24, axis Y). The results show the good degree of parallelization of combinatorial algorithm.

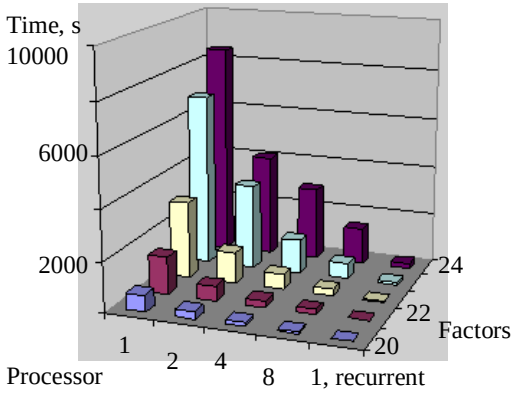

Figure 4

The last row of diagram corresponds to the run-time of the combinatorial scheme with the bordering algorithm [2] in the case of only one processor. This result obviously shows a considerable advantage of the recurrent procedure against parallelization with on 8 processors.

4 Solving of real modeling problems

Modeling of a density of the upper sedimentary layer of sea-bottom is important in respect to the search of oilfields as a result of the hydroacoustic monitoring of the Caspian Sea [5]. The recurrent bordering algorithm was used for the modeling. Arguments were included in a model according to a minimum value of the multiple correlation coefficients [3]. A simple mathematical dependence of density from three of all five factors, namely temperature, viscosity, and echo intensity, was obtained. Figure 5 illustrates the high quality of the model where the comparison of experimental values of the bottom upper sedimentary layer density and corresponding modeling results are presented. 18 records were used for the model training and 2 points for checking of the model accuracy.

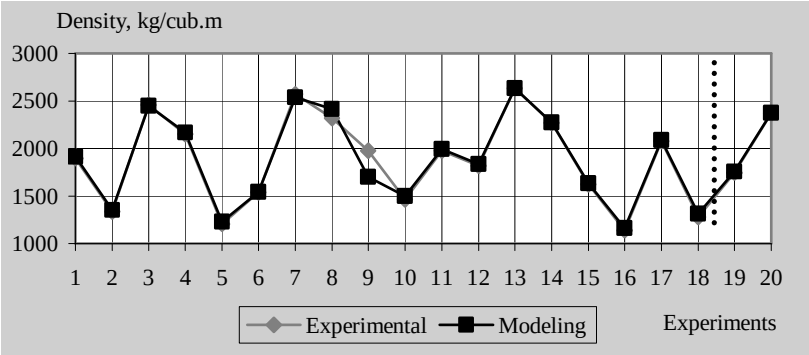

Figure 5

Forecasting the ferromolybdenum market price. The cost of ferromolybdenum forms almost one third of total costs of all metal consumption volumes for any engineering plant. That is why the task of the price forecasting is so important. 9 factors were chosen for the price modeling. The combinatorial algorithm was used in the class of linear models with using the Akaike and regularity criteria for selection of the best models. According to the results represented in Figure 6 the model built by using the regularity criterion Ar was more precise in testing points (three years, from 1996 to 1998). It was used for extrapolation of price value on the concentrate of molybdenum for two years.

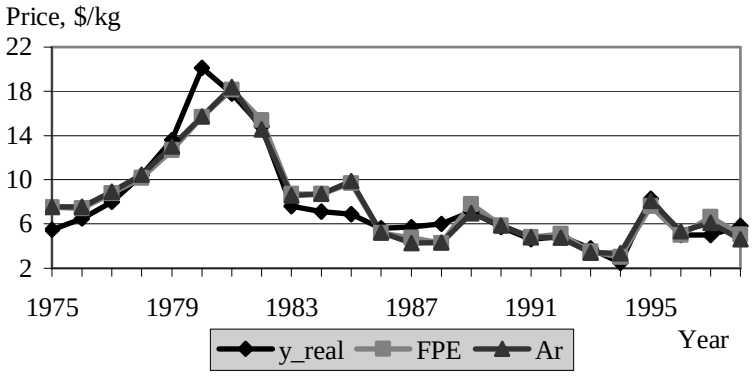

Figure 6

5 Conclusions

Due to the test examples it is shown that the use of recurrent algorithms is effective for parameter estimation in the problem of modeling from observation data because of their considerably higher computation speed comparing to nonrecurrent algorithms. Using the recurrent algorithms in the combinatorial procedure is much more effective than the parallelization on 8 processors.

The comparative testing of criteria of models structure identification demonstrates simplification of the selected models with increasing of noise level and the higher effectiveness of the regularity criterion.

Solving the real world modeling problems allows obtaining the reliable information on natural and technogenic anomalies in a water environment and prediction the market price on ferromolybdenum.

References

[1] Seber, G.A.F. Linear Regression Analysis. John Wiley and Sons, New York – London – Sydney – Toronto, 1977

[2] 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

[3] 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.

[4] V. S. Stepashko, and S. N. Efimenko. Sequential Estimation of the Parameters of Regression Model // Cybernetics and Systems Analysis, Springer New York, July, 2005, Vol. 41, Num. 4, pp.631-634.

[5] T. I. Nizamov, S. R. Ibrahimova, R. K. Quluzade, A. I. Isayev, V. S. Stepashko, S. N. Yefimenko. 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.