Investigation of Selective Properties of the Error Bias Criterion

Savchenko E.A., Stepashko V.S., Somina L.P.

International Research and Educational Centre for Information Technologies and Systems of NAS of Ukraine, Prospect Akademika Glushkova, 40, Kyiv, 03680, Ukraine

savchenko@irtc.org.ua

Abstract. *Selective properties of a new error bias criterion are investigated first analytically. It is shown that the same optimal model can be chosen by the new criterion as by classical solution bias criterion. This criterion is more sensitive to change of model complexity then classical. The numerical analysis of the error bias criterion for different levels of noise in data is carried out. The example shows that error bias criterion has the main property of effective external criterion: property of noise-immunity. It is shown that sequential using of the regularity criterion firstly and the error bias criterion secondly are the most reasonable for selection of an optimal complexity model.*

Keywords

Inductive modelling, GMDH, regularity criterion,
noise-immunity, error bias criterion, Pareto region.

1. Introduction

The inductive modelling is directed to an optimal model construction using external criteria. The regularity criterion is the most used external criterion in GMDH algorithms but there is a problem of optimum model choice when using it: as a rule some set of models is chosen having practically the some value of this criterion.

There are combined criteria taking into account values of several criteria simultaneously for an optimal model choice. However computation of a combined criterion has some failings. First, calculation of such criterion requires large time because values of a few criteria must be calculated for all models being sorted out simultaneously. Second, it is required to assign the weights of the individual criteria in a combined criterion.

These problems can be solved by application of a sequence of criteria. If we use only two criteria, at the beginning some set of the best models is selected by the first criterion, then optimal model is found by the minimum of the second criterion from the set of selected models. The reasonable sequence of external GMDH criteria is using the regularity criterion in combination with the error bias criterion.

2. Error bias criterion and its comparison with other GMDH criteria

In [1] the following form of an error bias criterion is offered:

$$BS = \left| AR_{W/A} - AR_{W/B} \right|, \tag{1}$$

where $AR_{W/A}$ and $AR_{W/B}$ are the summarized errors on the joint sample $W = A \cup B$ of model of the same structure with parameters estimated on samples A and B accordingly:

$$AR_{W|A} = \left\| y_W - X_W \hat{\theta}_A \right\|^2 = AR_{B|A} + RSS_A \text{ and } AR_{W|B} = \left\| y_W - X_W \hat{\theta}_B \right\|^2 = AR_{A|B} + RSS_B . \tag{2}$$

where RSS_A and RSS_B - residual sum of squares on the sample *A* and *B*.

The idea of bias criteria consists in the fact that the properties of the best model should be minimally different when estimating parameters on both learning and testing samples A and B [2].

Relation between the error bias criterion and other GMDH criteria may be expressed using known relation of the stability criterion AS and the bias criterion CB [2]:

$$AS = CB + 2(RSS_A + RSS_B)\,. \tag{3}$$

Using (1), (2) and (3), in [3] the following relation was obtained:

$$BS_{B|A} = \left| CB - 2(AR_{A|B} - RSS_A) \right|. \tag{4}$$

In [3] the error bias criterion was investigated analytically and it has shown that value of this criterion BS is always less than values of solution bias criterion CB due to the following inequality:

$$AR_{A|B} \geq RSS_A\,,$$

where $AR_{A|B}$ is the minimum of regularity criterion on the sample B for a model with parameters estimated on sample A. Numerical example of the next section 3 confirms analytical result.

3. Numerical example of comparison of solution and error bias criteria

The aim of this example is to show that the same model will be selected by the minimum of bias criteria of solution and error. A data sample was generated randomly and contains 8 variable and 30 observations (Tab.1).

Tab.1. Input data sample

№	X1	X2	X3	X4	X5	X6	X7	X8	Y
1	0.98	4.70	-1.20	-1.60	0.94	-2.50	4.70	-1.70	37.00
2	-3.00	3.00	4.10	4.00	0.40	2.00	1.90	-3.30	1.10
3	0.03	-2.50	2.70	-3.00	-3.90	-0.74	-1.00	2.50	-15.00
...	...	...	...	...	...	...	...	...	...
30	0.98	4.70	-1.20	-1.60	0.94	-2.50	4.70	-1.70	37.00

The output variable contains 10% uniform noise and was calculated by the formula:

$$y = -3x_3 + 7x_7 + \varepsilon \tag{5}$$

It is necessary to find the function (5) by the noisy data sample using the combinatorial GMDH algorithm. We will find an optimum model at first by solution bias criterion CB and then by error bias criterion BS and compare values of these criteria. 128 models-candidates are sorted out by combinatorial GMDH algorithm. Four best models are selected by minimum of criterion regularity AR (Tab.2).

Tab.2. Best models selected by minimum of regularity criterion

s	*AR*	*BS*	*CB*	Models
3	0,254	0,03	7,1	$y = 0{,}055 – 2{,}948x_3–6{,}980\ x_7$
4	0,228	0,135	7,48	$y = 0{,}068 – 2{,}960x_3 + 6{,}982\ x_7 – 0{,}022\ x_8$
5	0,227	0,16	8,38	$y = 0{,}05 – 2{,}95\ x_3 – 0{,}032x_5 + 6{,}987\ x_7 – 0{,}022x_8$
6	0,255	0,54	12,4	$y = 0{,}035 – 2{,}94\ x_3 – 0{,}452x_4 – 0{,}285x_5 + 6{,}97\ x_7 – 0{,}026x_8$

The result of comparison of both value of the error *CB* and solution *BS* bias criteria is shown on the Fig.1.

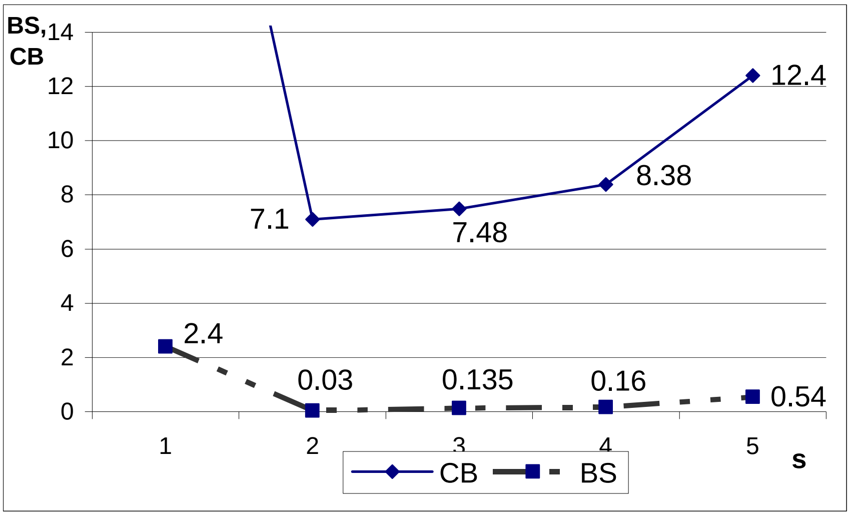

Fig.1. Dependences of error and solution bias criterion from model complexity

The same model is selected by minimum of error or solution bias criteria hence both criteria can be used equally. But the error bias criterion is more sensitive to change of model complexity as the numerical example shows.

4. Numerical research of noise-immunity of the error bias criterion

In [4] the noise-immunity of error bias criterion was investigated. Input data sample was generated randomly and contained 7 arguments and 30 observations. For each of experiments output variable was generated separately by expression

$$y = 5x_1 + 4x_2 + 3x_3 + 2x_4 + x_5 + \varepsilon \tag{6}$$

with the different level of noise beginning from 0% to 120% with step 10%

The best model for every level of noise was got by combinatorial algorithm using of external error bias criterion. 13 experiments were carried out. The best model was selected by the minimum of error bias criterion as a result of each experiment. Fig. 2 shows dependences of error bias criterion from model complexity for different noise levels.

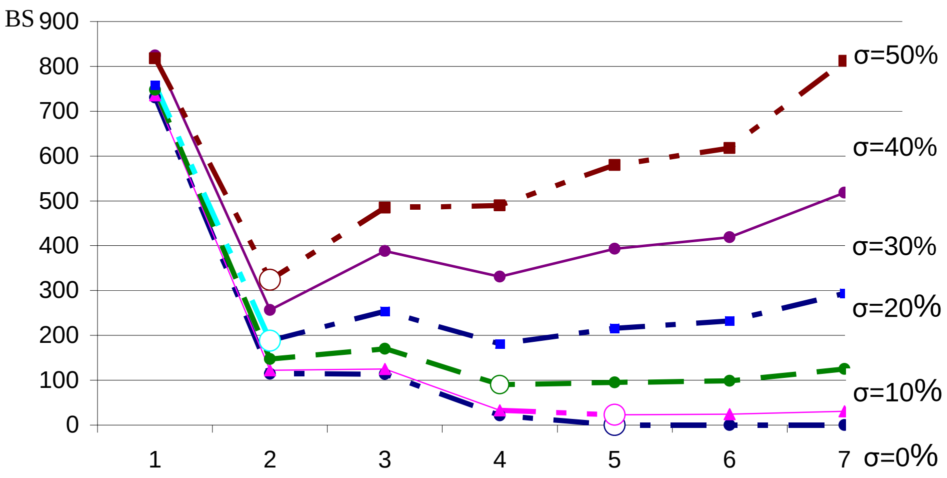

Fig.2. Dependences of error bias criterion from model complexity for a different level of noise

The character of dependence of error criterion from noise does not change when dispersion of noise is more than 50%. But when dispersion of noise is 120% then minimum will be for the model complexity $s = 1$. Minima of values criterion for every level of noise are marked in the Fig.2 by circles. The true model ($s_0 = 5$) can be obtained only at dispersion of noise less than 20%. Increasing of noise simplifies the optimal model: firstly model containing four members is selected then two ones and finally only one argument.

It was considered the same dependence but in other coordinates: dependence of error bias criterion from noise dispersion for models of different complexity (from $s = 1$ to $s = 7$). The model which minimizes the value of external criterion is selected for every level of noise dispersion separately [5]:

$$s^* = \arg \min_{i=\overline{1,m}} BS(s) . \quad (7)$$

The points of «switching» from one complexity to another are obtained: «switching» from $s^* = s_0 = 5$ to $s^* = 4$ takes place at dispersion of noise is 20% and from $s^* = 4$ to $s^* = 2$ at the dispersion 30%. If the level of noise increased to 120%, a model with $s^* = 1$ will be selected.

5. Method of model choose determination as a method of two-criterion search of an optimal model

External criteria which are used for model construction by GMDH have some drawback. First of all it is the problem of ambiguity. The method of two-criterion model search using regularity and bias criteria was proposed in [1] in order to avoid the ambiguity of the model choice. External criteria are used successively: firstly an accuracy criterion is used and then a bias criterion.

The noise-immunity of optimal model is increased when successive application of two external criteria is used. We will give a numerical example of application of the method for the search of optimum model. Data sample generated randomly and contains 6 variable and 40 observations.

The aim of this example is to find an optimal model sequence of using the external criteria of regularity and bias. The output variable contains 10% of noise in the data and present the model:

$$y = 0{,}3 + 0{,}7x_2 - 1{,}3x_4 + 0{,}8x_5 \quad (8)$$

Dependence of regularity criterion from model complexity is built by GMDH combinatorial algorithm (Fig.3).

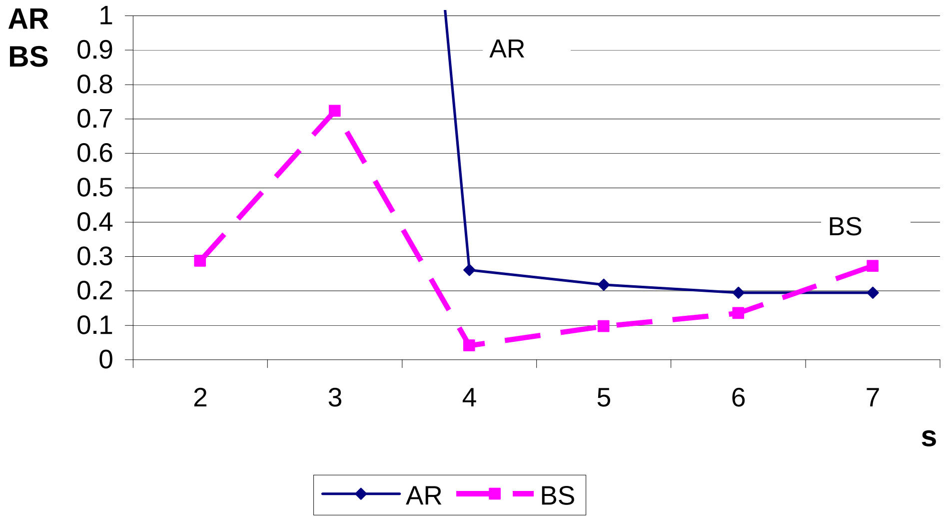

Fig. 3. Dependence of regularity and bias criteria from model complexity

Four models had the least and almost identical values of the regularity criterion, namely for $s = 4,5,6,7$ as shown in the Fig.3. The model III is selected by the minimum of regularity criterion (Tab.3).

Tab. 3. Best models selected by minimum of regularity criterion

s	*AR*	*BS*	Models	Number of models
4	0,2607	0,041	$y = 0,2576 + 0,6780x_2 – 1,3032x_4 + 0,7595x_5$	I
5	0,2180	0,097	$y = 0,3137 + 0,6608x_2 – 0,0678x_3 – 1,2976\ x_4 +$ $+ 0,7772\ x_5$	II
6	0,1945	0,135	$y =0,2847 + 0,6614\ x_2 – 0,0647\ x_3 – 1,3014\ _{x4} +$ $+ 0,7761\ x_5 – 0,0293\ x_6$	III
7	0,1953	0,272	$y =0,2850 + 0,0024x_1 + 0,6608x_2 – 0,0648x_3$ - $– 1,301\ x_4 + 0,776\ x_5 – 0,0294\ x_6$	IV

However values of regularity criterion for four best models are almost identical so the error bias criterion is calculated for these models. The model I is chosen by the minimum of the error bias criterion. Fig.4 shows the set of eight best models selected by the regularity criterion value, each of them contains all the true arguments x_2, x_4 and x_5.

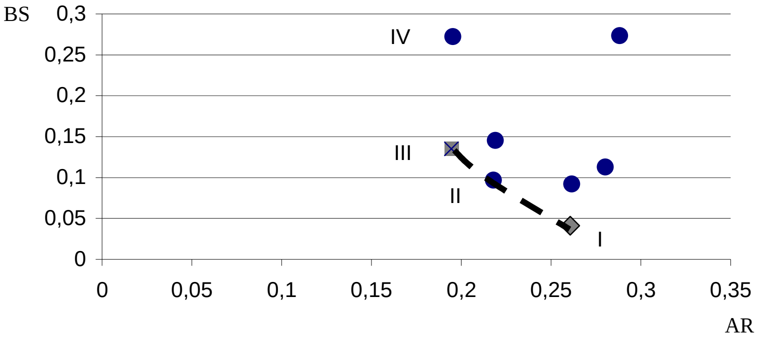

Fig.4. Models which have the least value of criteria

The Pareto region is marked in the Fig.4 by the dotted line and consists of three best models (I, II, III). This region contains the best model by the two criteria. This means that there is no any model which would minimise value of one of criteria without worsening value of another. The model I is an estimation of the true model (8).

Conclusion

The analytical and numerical investigation of the error bias criterion showed that this is an adequate GMDH criterion. Moreover the criterion has the main property of an effective external criterion, i.e. noise-immunity property. The error bias criterion is much more sensitive to the change of model structure and is preferable for applying in GMDH algorithm, especially by two-criterion choice of optimal model.

References

[1] Ivakhnenko A.G., Ivakhnenko G.A., Savchenko E.A. GMDH Algorithm for Optimal Model Choice by the External Error Criterion with the Extension of Definition by Model Bias and its Applications to the Committees and Neural Networks // Pattern Recognition and Image Analysis.– 2002.– V.12, N 4.– P.347-353.
[2] Ivakhnenko A.G., Stepashko V.S. Noise-immunity of modelling: Kiev.- Naukova Dumka. – 1985. – 216 p. (in Russian).
[3] Ivakhnenko A.G., Savchenko E.A. Investigation of Efficiency of Additional Determination Method of the Model Selection in the Modeling Problems by Application of the GMDH Algorithm // Journal of Automation and Information sciences. Begell House., Inc. Publishers.– 40, No 3, 2008.– pp.47-58.
[4] Savchenko I. Double-criterion Choice of an Optimal Model in GMDH Algorithms // Proceedings of 2nd International Conference on Inductive Modelling, September 15-19, 2008, Kyiv, p. 23-27.
[5] Stepashko, V.S. Method of Critical Variances as Analytical Tool of Theory of Inductive Modeling // Journal of Automation and Information Sciences, 2008, 40, No. 3, pp. 4-22.