

The fuzzy algorithm of GMDH and its expedient modifications with combinatorial-selective generation of particular models

Alexander Pavlov, Vladimir Pavlov

The chair of Descriptive Geometry and Engineering Drawing, National Technical University «KPI», Victory A venue 38, ,Kiev, Ukraine

The chair of Computer Economic and Ecological Monitoring, University «Ukraine»,Horevaja str. 1,Kiev, Ukraine

Alexander_mk@ukr.net, vpavlo@bk.ru

Abstract. *The GMDH algorithm is offered in the paper for synthesizing the fuzzy models of optimal complication and models of error corridor. The algorithm uses a combinatorial-selective method of forming the particular models and uses linear programming for receiving estimations of model parameters on a learning sample. The possibilities for generalizing designed algorithm for a problem of synthesizing of a error corridor model are considered. The method of piecewise approximating of curves and processes is offered, that is based on a usage of the models family from the corridor that was received*

Keywords

GMDH, fuzzy model, error corridor, linear programming, approximation, models family

1 Introduction

The engaging of possibilities of fuzzy modeling to the extension of varieties of algorithms GMDH is called

1. Expediency of carrying of equivocation of the relations between the model and object on structure of coefficients of the model in some practical applications of the simulation theory

2. By necessity for a number of problems of analysis and simulation to receive not only well-defined model of process (we shall term as CM the central model Y of process, as introduced by significances of records $-Y_T$), but also

boundary models (curves Y^+ and Y_- - fig. 1), which one in the best way driving "above" and "lower" of process and the appropriate CM, in sense of significance of some criterion. These curves should derivate a corridor, in which one the model, process and, thus, error of simulation hits. Than higher is the quality of the synthesized central model and than more flexible models of boundary curves managed to be constructed, the already narrow corridor of the error will be received.

3. It is obvious the possibility to diminish the error of simulation for the problems, where usage of the discontinuous composite models is allowed. At the expense of usage on some paths section, as its well-defined model not central, but other, from the set received, in a corridor of the models such outcome can be carried out. A method of piecewise approximating of curves or processes for problems thus can be realized, where the models accuracy requirements are more essential, than requirements to a continuity and smoothness in a connection records in composite model

2 Theoretical Part

Let we have the observational data of process $Y_T: (y_1, \dots, y_m)$ and significances of its inputs $\bar{x}: (x_{11}, \dots, x_{n1}), (x_{12}, \dots, x_{n2}), \dots, (x_{1m}, \dots, x_{nm})$, where n - dimension of the vector \bar{x} , m - amount of training sample.

The fuzzy models can be received as fuzzy functions by the way of bundle of trends and by the way of fuzzy discrete schemas, as in case of the fuzzy forecasting models. On figure 1 is exhibited Y_T - significances of process, $Y = \Phi(\bar{r}, \bar{x}) = r_0 + \sum_{i=1}^N r_i f_i(\bar{x})$ - central model, $Y_{\pm} = r_0 + \sum_{i=1}^N r_i f_i(\bar{x}) \pm S(\bar{c}, \bar{x})$ - extreme high and low models of an fuzzy corridor, where $S(c, x) = c_0 + \sum_{i=1}^N c_i |f_i(\bar{x})|$.

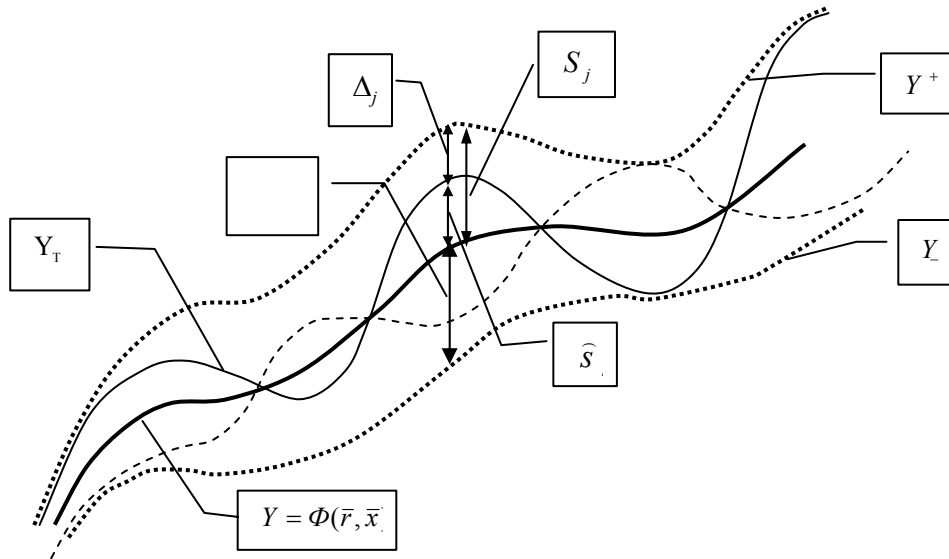


Fig. 1

The synthesizing of the linear fuzzy models with given basis is realized as a solution of a problem of a linear programming (LPP) in paper [1]. The indicated approach was utilized for synthesizing the nonlinear fuzzy models of optimal complication in the present paper. The GMDH fuzzy algorithm with combinatorial-selective forming of particular models (FACS) was developed for this purpose. A combinatorial-selection way [2] for a forming of the particular models could be realized by two paths. In case of the former the particular model looks like $Y_k = A_k + B_k \cdot Y_{k-1} + C_k \cdot Z_k$ on a stage of selection k , in the second case - $Y_k = A_0 + A_1 \cdot Z_1 + \dots + A_k \cdot Z_k$, here Z_k - generalized variable, that was selected on a stage k , $B_k, C_k, A_i, i = 0, \dots, k$ - symmetric triangular fuzzy coefficients. As is known, such fuzzy numbers $A = (r, c)_\Delta$ may be defined by a pair of numbers - r and c and a membership function

$$\mu_{A_i}(a) = \Lambda\left(\frac{a-r_i}{c_i}\right), c_i > 0$$

where $\Lambda(x) = \Lambda(-x); \Lambda(0) = 1 \quad \Lambda(x) = \max(0, 1 - |x|)$

Here numbers r and c mean centre and deviation, accordingly. The generalized variable Z_k is reshaped of source variables $x_j \in \{x_j\} = X$ as follows:

$$z_i \in Z, Z = T \cup T \otimes T \cup \dots \cup T \otimes T \dots \otimes T, t_i \in T = \Omega \cup X, \Omega = \{1/x_j\} \forall j$$

It is expedient to select the second approach - $Y_k = A_0 + A_1 \cdot Z_1 + \dots + A_k \cdot Z_k$ from the point of view of accuracy of the model, if the anticipated optimal complication of the model is not so high (6-8 stages of selection) and the amount of records in a learning sequence allows to calculate arguments of the model of the last stage. Consideration outcomes [1], it is possible to apply LPP (2) at the stage N of algorithm for the receiving estimations of triangular fuzzy coefficients $A_i, i = 0, \dots, N$ of the particular models (1) with nonlinear basis $Z_i = f_i(\bar{x}), i = 1, \dots, N$

$$Y_N = A_0 + A_1 f_1(\bar{x}) + \dots + A_N f_N(\bar{x}) \quad (1)$$

$$\left\{ \begin{array}{l} L = \min \left\{ mc_0 + \sum_{j=1}^{m_A} \sum_{i=1}^N c_i \cdot |f_i(\bar{x}_j)| \right\} \\ y_j - \left(r_0 + \sum_{i=1}^N r_i \cdot f_i(\bar{x}_j) \right) \geq -(1-h) \left(c_0 + \sum_{i=1}^N c_i |f_i(\bar{x}_j)| \right), \quad j = 1, 2, \dots, m_A, \\ y_j - \left(r_0 + \sum_{i=1}^N r_i \cdot f_i(\bar{x}_j) \right) \leq +(1-h) \left(c_0 + \sum_{i=1}^N c_i |f_i(\bar{x}_j)| \right), \quad j = 1, 2, \dots, m_A, \\ c_i \geq 0, \quad i = 0, 1, 2, \dots, N \end{array} \right. \quad (2)$$

We determine here the degree h to which we wish the given data to be included in the inferred number Y_j , that is,

$$\mu_Y(y_j) \geq h \quad j = 1, \dots, m_A \quad 0 \leq h < 1$$

The quality of fuzzy model (1) and received simulation error corridor is characterized by parameters

$$L_\phi = \sum_{j=1}^{m_A} |y_j - (r_0 + \sum_{i=1}^N r_i \cdot f_i(\bar{x}_j))| - \text{the error of CM}, \quad L_s = m_A c_0 + \sum_{j=1}^{m_A} \sum_{i=1}^N c_i \cdot |f_i(\bar{x}_j)| - \text{parameter of}$$

width of error corridor, their difference $L_{\Delta_s} = L_s - L_\phi$ mirrors a degree of approximation by extreme curves of a corridor the modulus of the simulation error of a CM, the relative parameter $I_{\Delta_s} = L_{\Delta_s} / L_\phi$ demonstrates relative efficiency of simulation of an error corridor, m_A - amount of records in a learning sample.

Criterion of selection is combined: $J_c = \alpha \cdot J_A + \beta \cdot J_B$, here J_A, J_B characterizes the width of a simulation error corridor on learning and testing records, α, β - weighting coefficients accordingly for parts of criterion on learning and checking samples. $J_A = L_s$, and we form J_B as follows -

$$J_B = \sum_{j=1}^{m_B} J_j^B,$$

$$\text{denote } S_j = S_j(\bar{c}, \bar{x}) = c_0 + \sum_{i=1}^N c_i |f_i(\bar{x}_j)| \quad \hat{S}_j = y_j - \left(r_0 + \sum_{i=1}^N r_i f_i(\bar{x}_j) \right), \quad j = 1, \dots, m_B \text{ (fig. 1) and}$$

having that always we have $S_j \geq 0$ [3], we form $J_j^B = |\hat{S}_j| + |S_j - |\hat{S}_j||$.

Possibilities of the offered approach (the account is made at $h=0$) are demonstrated on an example of FACS simulation of an index of the prices in Ukraine in 1997-1999 years (see figure 2) on monthly average data of 21 economic variable in 44 points. The received optimal fuzzy model of the prognosis of an index of the prices for three months forward below is recorded [4]:

$$\begin{aligned} Y(t+3) = & (99.6883, 0) + (824221115.8, 82637006.20) * X_{14}(t) / (X_{21}^2(t-4) * X_{21}(t-5)) + \\ & (-530492.89, 0) * X_{14}(t-1) / (X_{22}(t-3) * X_{07}(t-4) * X_{08}(t-2)) + (2398760.351, 0) * X_{14}(t) / (X_{14}(t-2) * \\ & X_{20}(t-5) * X_{21}(t-1)) + (-775504.2, 63036.6) * X_{20}(t-3) / (X_{22}(t-4) * X_{20}(t-5) * X_{02}(t-3)) + \\ & (82654.58154, 0) * X_{21}(t-3) / (X_{20}(t-5) * X_{14}(t) * X_{08}(t-3)) + (4557496.35, 400296.35) * X_{14}(t-5) / \\ & (X_{14}(t-3) * X_{20}(t) * X_{20}(t-3)) + (-1.551129, 0) * (X_{02}(t-4) * X_{18}(t-5)) / (X_{02}(t-3) * X_{19}(t-1)) \end{aligned}$$

9 of 21 variables have come in received fuzzy model, the significances of delays, are indicated in brackets:

X02 - cash outlay and savings of the population (million. gr.), X07 - production volume of an industry (million. gr.), X08 - retail turnover (million. gr.), X14 - established average refinancing rate of business banks. (% per annum), X18 - accounts receivable between the enterprises. (million. gr.), X19 - bill payable between the enterprises (million. gr.), X20 - costs of the Summary budget, all (million. gr.), X21 - incomes of the Summary budget, all (million. gr.), X22 - volume of production of a light industry (million. gr.)

Significance of criterions for CM: NMSE on training records $\Delta_n = 0.35960$, $L_\phi = 16.70840$, the relative reduced error of examination

$$RE_{ex} = \sum_i^{m_{ex}} |y_i^E - y_i^T| / m_{ex} |y_{max}^T - y_{min}^T| = 0,13$$

Significance of criterions for an error corridor: $L_s = 17.89$, $I_{\Delta_s} = 0.07$

So, it was possible to receive enough satisfactory CM and a narrow (small significance $I_{\Delta_s} = 0.07$) simulation error corridor and allowable bundle of solutions on A .

Let's consider further possibilities of generalizing of an offered formulation of synthesizing of the fuzzy models. Apparently, that the existing definition of the fuzzy models dictates necessity of bound nature of CM curves and corridor. It is a condition of causing of bundle of solutions. However, if to refuse receiving a solution of a problem synthesizing of fuzzy models by the way records (1), but to save up expediency of the approach from the point of view of minimization of a CM error corridor and receiving of the set of the models in this corridor, it is possible to offer generalizing of algorithm on the basis LP problem (2) as follows.

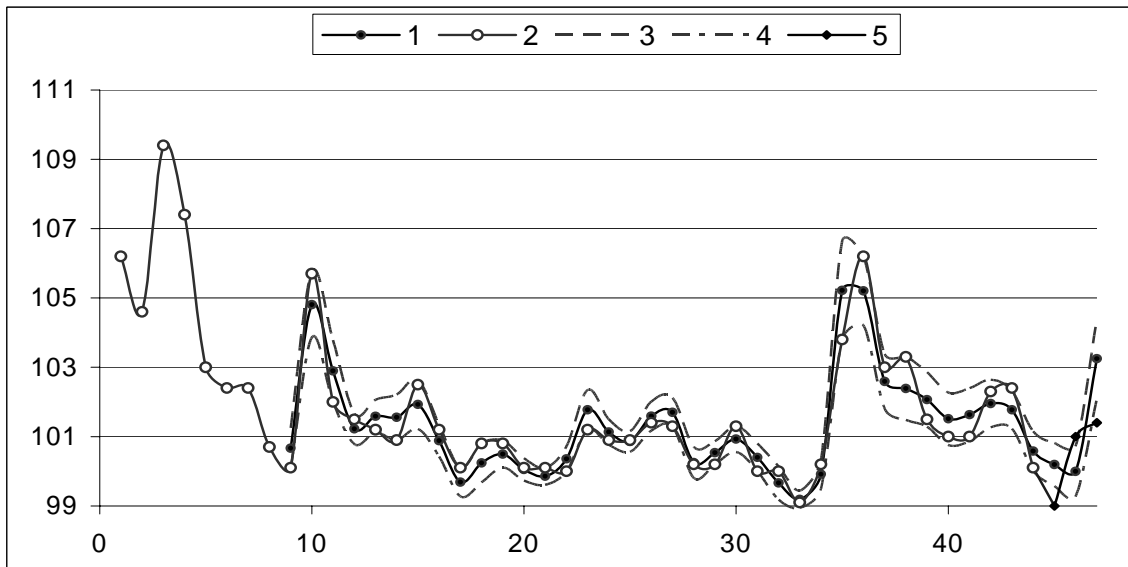


Fig. 2 The fuzzy model of the prognosis of an index of the prices in Ukraine in 1997-1999 yy
1- model curve, 2 – process curve, 3,4 – boundary curves, 5 – records of examination

It is possible to show [3], that at an arbitrary kind of the function $S(\bar{c}, \bar{x}) = c_0 + \sum_{i=1}^N c_i f_i(\bar{x})$, if there is allowable solutions applicable LP problem on points of set A , we shall have always deviation $S_j = S_j(\bar{c}, \bar{x}) \geq 0$ on this set. However outside set A , a corridor can not exist. A unique condition of existence of a corridor in given existence domain is the fulfillment

$$S(\bar{c}, \bar{x}) = c_0 + \sum_{i=1}^{N_2} c_i \cdot f_i'(\bar{x}) \geq 0, \quad \forall \bar{x} \quad (3)$$

in this area. Thus we receive indispensable limitation on a kind of functions of the boundary form. Under a condition (3), it is possible to offer generalizing of algorithm FACS for a case of introduction of different bases for the central model $\Phi(\bar{r}, \bar{x})$ and limiting form $S(\bar{c}, \bar{x})$. Is offered LPP (4), with criterion $L = \sum_{j=1}^{m_A} \alpha' \cdot \hat{s}_j + \beta' \cdot \Delta_j$ permitting by means of a variation of factors α' and β' to influence on relative velocity convergences of algorithm for the models

$\Phi(\bar{r}, \bar{x})$ and $S(\bar{c}, \bar{x})$ and, thus, to receive different alternatives of the central and boundary models below. The significance of gated in variables \hat{S}_j and Δ_j is apparent from a figure 1.

$$\begin{cases} L = \sum_{j=1}^{m_A} \alpha' \cdot \hat{s}_j + \beta' \cdot \Delta_j \\ y_j - \Phi_j \geq -S_j, \quad \Phi_j = r_0 + \sum_{i=1}^{N_1} r_i f_i(\bar{x}_j), \quad S_j = c_0 + \sum_{i=1}^{N_2} c_i \cdot f'_i(\bar{x}_j) \\ y_j - \Phi_j \leq S_j, \quad y_j - \Phi_j \geq -\hat{s}_j, \quad y_j - \Phi_j \leq \hat{s}_j, \quad \Delta_j = S_j - \hat{s}_j, \quad j = 1, 2, \dots, m_A \end{cases} \quad (4)$$

As an example for $S(\bar{c}, \bar{x})$, meeting condition (3) we shall reduce a kind

$$S(\bar{c}, \bar{x}) = c_0 + \sum_{i=1}^{N_2} c_i \cdot |f'_i(\bar{x})| \geq 0, \quad c_0, c_i \geq 0$$

The corridor, that was received as a result of job of algorithm on a ground LPP (2) or (4) can be utilized for receiving different composite, more precise, than CM, well-defined approximative model. One of apparent alternatives - following:

The generalized composite model $F(\bar{r}, \bar{c}, \bar{x}) = \Phi(\bar{r}, \bar{x})$ only on parts, where a CM is closer to records of process, than border of a corridor. On remaining parts:

$$\begin{aligned} &F(\bar{r}, \bar{c}, \bar{x}) = \Phi(\bar{r}, \bar{x}) + S(\bar{c}, \bar{x}) \quad \text{in neighborhood of a record } j, \text{ under} \\ &y_j - (r_0 + \sum_{i=1}^{N_1} r_i \cdot f_i(\bar{x}_j)) > 0 \quad \text{and} \quad F(\bar{r}, \bar{c}, \bar{x}) = \Phi(\bar{r}, \bar{x}) - S(\bar{c}, \bar{x}) \quad \text{in neighborhood} \\ &\text{of a record } j, \text{ under} \quad y_j - (r_0 + \sum_{i=1}^{N_1} r_i \cdot f_i(\bar{x}_j)) < 0 \end{aligned}$$

The preference is desirable (at the expense of a choice of factors significances α' and β') for faster convergence of the central model in the problems of the prognosis.

Conclusion

1. The fuzzy GMDH algorithms as FACS can supply narrow enough the corridor of simulation error.
2. Improvement of properties of an error corridor is possible at utilization of different bases for the central model and borders of a corridor. The condition for the boundary models is offered for providing existence of a corridor.
3. The method of the piecewise approximating is offered on the basis of usage of the models from the corridor of simulation error.

References

1. Paradopoulos B. K., Sirpi M. A., Similarities in fuzzy regression models., *Journal of optimization theory and applications*: 1999, Vol. 102, № 2, pp. 374- 383.
2. Ivakhnenko A.G., Stepashko V.S., Noise immunity of simulation. Kiev, *Naukova dumka*, 1985, - 216 p. (in Russian).
3. Pavlov A.V. Algorithms of selforganizing in problems of increase of selfdescriptiveness of the geometrical models of processes, given dot frame. *Ph.D. thesis, K., NTUU "KPI"*, 2006, 197pp. (in Russian).
4. Pavlov A.V. Multi-stage combinatorial - selection GMDH algorithm for synthesizing the fuzzy regression models. / *Applied geometry and engineering drawing*, Numb. 75., K.: KNUBA, 2005. pp. 188-192. (in Ukrainian).