Impact of Data Division on the Adequacy of External Criterion of Unbiasedness Errors in GMDH Algorithms

Nina Kondrashova

International Research and Training Center for Information Technologies and Systems of the National Academy of Sciences of Ukraine, 03680, Kiev, av. Glushkova, 40, Ukraine

NKondrashova@ukr.net

Abstract. *The article demonstrates the validity of the joint application of a data partitioning method and a criterion of external additions in finding the most accurate models. If the condition of data proportionality is satisfied, the theoretical substantiations were obtained that the criterion of unbiasedness errors is not the criterion "adequate" to noise while minimizing quasioptimal criterion of the sample division but only when the criterion is maximized or if the scheme of repeated experiment is performed.*

Keywords

Group method of data handling (GMDH), external supplement criterion, data partitioning method, ρ-optimal sample division, quasi-optimal partitions.

1 Introduction

The paper is devoted to solving the so-called "problem of sample partitioning" in the group method of data handling (GMDH). This problem has 45-year history of the consideration in the literature about GMDH starting from the publication the partitioning “by variance” in monograph [1] and idea of article [2]. In [3], based on the results known from [4-7], for the criterion of regularity it has investigated the problem of partitioning for GMDH criteria belonging to the class criteria of the minimum bias (or unbiased criteria), namely, to the unbiasedness criterion of solutions. In this article the soundness of the joint application of the data partitioning method and criterion of unbiasedness errors in finding the most accurate models is investigated. Geometric illustration is given of samples partitioning by the variance conventionally used in different proportions of their sizes.

2 Statement of the problem

In modeling of an object it is assumed that 1) the model is linear in parameters; 2) "ideal" model exists, i.e. $\mathbf{y}^0 = \mathbf{X}^0\Theta^0$; 3) uncorrelated noise ξ on the output $\mathbf{y} = \mathbf{y}^0 + \xi$ is present with zero mathematical expectation, finite variance $\sigma^2 < \infty$ and diagonal covariance matrix $\xi\xi^T = \sigma^2\mathrm{I}$, where $\mathbf{y}^0$ is the output variable of the "ideal" model, $\mathbf{X}^0$ is matrix of full set of input variables (arguments), Θ^0 is a vector of parameters, $\mathbf{I}$ is the identity matrix. It is also believed noncorrelatedness of different realisations of noise among themselves and with useful signal. Estimates of model parameters are determined according to the subsample A when $B \neq \varnothing$ or vice

versa, according to the sample B at $A \neq \varnothing$, $\mathbf{X}_A$ is a matrix of full-column rank ($n_A \geq s$); additive noise with the above specified properties is present at the output.

Formulation of diversity of species partitioning criteria is represented. It should be noted that two of four criteria unbiasedness, except unbiasedness criterion coefficients and of absolute noise immunity criterion, can be calculated by criteria of accuracy: regularity criterion

$$AR_{Q|G}(s) = \min_{s=1,m} \left\| \mathbf{y}_Q - f(\mathbf{X}_{Qs}, \widehat{\Theta}_{Gs}) \right\|^2, \quad (1)$$

and residual sum of squares criterion

$$RSS_Q(s) = AR_{Q|Q} = \min_{s=1,m} \left\| \mathbf{y}_Q - f(\mathbf{X}_{Qs}, \widehat{\Theta}_{Qs}) \right\|^2, \quad (2)$$

wherein as G and Q samples can be the original sample W and any of its subsamples A and B.

Unbiasedness criterion coefficients has the form

$$n^2_{см(1)}(s) = \left\| \widehat{\Theta}_{As} - \widehat{\Theta}_{Bs} \right\|^2, \ \dim \widehat{\Theta}_{As} = \dim \widehat{\Theta}_{Bs} = s \times 1,$$

where $\widehat{\Theta}_{As}$, $\widehat{\Theta}_{Bs}$ are coefficients estimated on subsamples A and B, s is model complexity or number unequal zero parameters, and absolute noise immunity criterion:

$$n^2_{см(3)}(s) = \left(\hat{\mathbf{y}}_W(\widehat{\Theta}_{Ws}) - \hat{\mathbf{y}}_W(\widehat{\Theta}_{As})\right)^T \left(\hat{\mathbf{y}}_W(\widehat{\Theta}_{Bs}) - \hat{\mathbf{y}}_W(\widehat{\Theta}_{Ws})\right) = \left(\widehat{\Theta}_{Ws} - \widehat{\Theta}_{As}\right)^T \mathbf{X}^T_{Ws}\mathbf{X}_{Ws}\left(\widehat{\Theta}_{Bs} - \widehat{\Theta}_{Ws}\right),$$

where $\hat{\mathbf{y}}_W(\widehat{\Theta}_{As})$, $\hat{\mathbf{y}}_W(\widehat{\Theta}_{Bs})$ and $\hat{\mathbf{y}}_W(\widehat{\Theta}_{Ws})$ are extrapolation outputs of the model, calculated on the W, with $\widehat{\Theta}_{As}$, $\widehat{\Theta}_{Bs}$ coefficients estimated on subsamples A and B; $\mathbf{X}_{Ws}$ is an initial matrix with s arguments.

Criterion of solutions unbiasedness and the criterion of unbiasedness errors can be written and calculated by criteria (1) and (2). Consider the last of them, notably criterion of unbiasedness errors as a sum of the components: regularity criterion and the residual sum of squares

$$n_{см(4)} = \left| AR_{W|A} - AR_{W|B} \right| = \left| AR_{B|A} + RSS_A - AR_{A|B} - RSS_B \right|.$$

3 Research adequacy of unbiasedness errors criterion at ρ-optimal sample division

Let us analyze the criterion of unbiasedness errors. It is a minimum of absolute value the sum of the structural and the noise component

$$n_{см(4)} = \left| n_{см(4)}{}^b + n_{см(4)}{}^v \right|.$$

The adequacy of this criterion external addition will be researched when selecting the best model structure. Consider the function of the numerical values s, standing under the sign module.

Structural component looks like

$$n_{см(4)}(s)^b = \left\| \mathbf{y}^0_W - \mathbf{X}_{Ws}\overline{\Theta}_{As} \right\|^2 - \left\| \mathbf{y}^0_W - \mathbf{X}_{Ws}\overline{\Theta}_{Bs} \right\|^2.$$

Noise components using the criteria RSS_Q and $AR_{Q \mid G}$, where Q, G are any of the subsamples A and B, averaged over the set of realizations of noise [5], can be written as

$$n_{см(4)}(s)^v = \sigma^2_B n_B + \sigma^2_A tr(\mathbf{X}^T_{As}\mathbf{X}_{As})^{-1}\mathbf{X}^T_{Bs}\mathbf{X}_{Bs}) + \sigma^2_A n_A - \sigma^2_A n_A - \sigma^2_B tr(\mathbf{X}^T_{Bs}\mathbf{X}_{Bs})^{-1}\mathbf{X}^T_{As}\mathbf{X}_{As}) - \sigma^2_B n_B . \quad (3)$$

ρ^2-proportional sample partitioning for matrices $\mathbf{X}_{As}$ and $\mathbf{X}_{Bs}$ of full-column rank $k = s$:

$$\mathbf{X}_{As}^T\mathbf{X}_{As} = \rho_B^2\mathbf{X}_{Bs}^T\mathbf{X}_{Bs}, \tag{4}$$

where $\dim\mathbf{X}_A \neq \dim\mathbf{X}_B$, $n_A > s$, $n_B > s$, $n_A \neq n_B$, $\rho_B^2 \neq 0$.

Then (3) with condition (4) write like

$$n_{CM(4)}(s)^v = \sigma_A^2\frac{1}{\rho_B^2}tr(\mathbf{X}_{As}^T\mathbf{X}_{As})^{-1}\mathbf{X}_{As}^T\mathbf{X}_{As}] - \sigma_B^2\rho_B^2 tr(\mathbf{X}_{Bs}^T\mathbf{X}_{Bs})^{-1}\mathbf{X}_{Bs}^T\mathbf{X}_{Bs}] =$$

$$= \frac{1}{\rho_B^2}\sigma_A^2 s - \rho_B^2\sigma_B^2 s = \frac{1}{\rho_B^2}s(\sigma_A^2 - \sigma_B^2\rho_B^4).$$

Noise component is linearly increasing from the numbers *s*, if $\sigma_A^2 > \sigma_B^2\rho_B^4$. If $\sigma^2 = \sigma_A^2 = \sigma_B^2$ is fulfilled, then

$$n_{CM(4)}(s)^v = \frac{1}{\rho_B^2}s\sigma^2(1-\rho_B^4) = \frac{1}{\rho_B^2}s\sigma^2(1+\rho_B^2)(1-\rho_B^2).$$

In order to, the noise component is linearly increasing from the numbers *s* is necessary and sufficient, that $0 < \rho_B^2 < 1$ (see Figure 1). This condition will provide adequacy to criterion unbiasedness errors when monotonically decreases to zero the structural component.

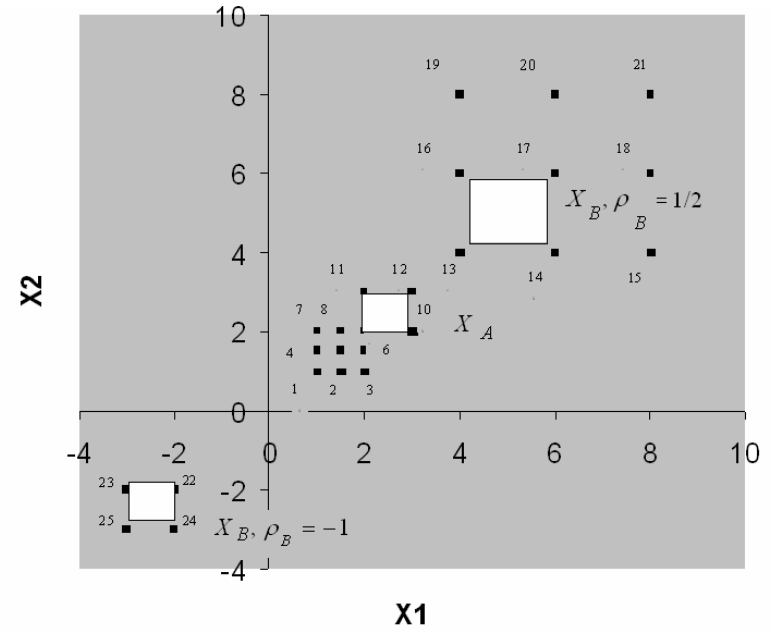

Fig 1. Location of points set $\mathbf{X}_A$ relative to the points $\mathbf{X}_B$ in the plan of two variables which give examples of linear (quadratic) dependent divisions.

The absolute value always has a minimum and it corresponds on the axis of the complexity of the structures *s* to case of equality of structures and parameters of the models obtained on the data of subsamples *A* and *B*. If there are present all "true" arguments, the global minimum (zero)

the structural component corresponds to the "true" model for achieving the value of s^0. When not present all the "true" arguments, i.e. $s^0_{par} < s^0$, and in the matrix $\mathbf{X}$ are some false arguments $s > s^0_{par}$, then the minimal value of the structural component is achieved at presence in model of s^0_{par} arguments for matrices of full rank. In order to the structural component was falling is needed model error caused by structural component on the *A* was much larger, than the errors on the *B* at $s = 1$ (or vice versa) and when the growing *s*, it declined. This is possible, if the variance of the values of "true" variables on the *A* would be a greater variance of their values on the *B*. Otherwise (at equality of errors on the subsamples *A* and *B*) we get 0 or close to zero the value of structural component. Then, if the variance of the noise will increase, it will always be chosen a model trivial structure: $y = \theta$.

In order to the noise component has been strictly increasing, need $0 < \rho_B^2 < 1$, which will be satisfied if points on *B* will have greater dispersion, the dispersion on *A* (or vice versa, on *A* and on *B*). As the Figure 1 shows, last requirements are fulfilled in the case of ρ-optimal partition, i.e. ρ-proportional, ρ^2-proportional and to quasi optimal partitioning. Then the criterion of partitioning is the minimum of norm $E1_\ell$. In the last case the best division is

$$\ell^* = \arg \min_{\ell=\overline{1,L}} E1_\ell = \arg \min_{\rho^2_{B\ell} \neq 0,\ \ell=\overline{1,L}} \left| (\mathbf{X}^T_{A\ell}\mathbf{X}_{A\ell} - \rho^2_{B\ell}\mathbf{X}^T_{B\ell}\mathbf{X}_{B\ell}) \right| \qquad (5)$$

or minimization of trace of the information matrixes difference

$$l^*_d = \arg \min_{\rho^2_{B\ell} \neq 0,\ \ell=\overline{1,L_d}} tr\left| (\mathbf{X}^T_{A\ell}\mathbf{X}_{A\ell} - \rho^2_{B\ell}\mathbf{X}^T_{B\ell}\mathbf{X}_{B\ell}) \right|. \qquad (6)$$

But simultaneously these requirements (5) and (6) together with changing of structural components are contradictory.

Therefore, the *first case* would then be when the samples have an equal number of observation points and variances on *A* and *B* are closed or equal. This corresponds to the condition of repeated experiments. This case matches to the building of the subsample close to the planning an experiment. At repeated plan ($\rho \approx 1$), and, if $\mathbf{X}_{Al}$, $\mathbf{X}_{Bl}$ are matrixes of full column rank, we find partitioning

$$\ell^* = \arg \min_{,\ \ell=\overline{1,L}} \left| (\mathbf{X}^T_{A\ell}\mathbf{X}_{A\ell} - \mathbf{X}^T_{B\ell}\mathbf{X}_{B\ell}) \right|,$$

or from conditions of minimum of information matrixes trace

$$l^*_d = \arg \min_{l=\overline{1,2L_d}} tr\left[(\tilde{\mathbf{X}}^T_{Al}\tilde{\mathbf{X}}_{Al} - \tilde{\mathbf{X}}^T_{Bl}\tilde{\mathbf{X}}_{Bl}) \right] \qquad (7)$$

where $\tilde{\mathbf{X}}_{Al}$, $\tilde{\mathbf{X}}_{Bl}$ are matrixes with centered variables $\tilde{x}_{ij} = x_{ij} - \bar{x}_j$, $i = \overline{1, n_G}$, $j = \overline{1, s}$, $\bar{x}_j = {1}/{n_G} \sum_{i=1}^{n_G} x_{ij}$, $G = A \vee B$.

The above mentioned method of partitioning data is recommended if nothing is known about the structure and parameters of the object, as well as an input matrix has a "false" arguments since otherwise we have reason to get a trivial model.

Using results [8], rewrite (7) using correlation coefficients r_{i,j_A} and r_{i,j_B} of *i*-th and *j*-th variables in different partitioning of the total number of points on points the *A* and points the *B*.

$$\ell^* = \min_{\ell \in [1,L]} \left\| \begin{matrix} 0 & (r_{1,2_A} - r_{1,2_B}) & \dots & (r_{1,m-1_A} - r_{1,m-1_B}) & (r_{1,m_A} - r_{1,m_B}) \\ & & \dots & \dots & \dots \\ & & & 0 & (r_{m-1,m_A} - r_{m-1,m_B}) \\ & & & & 0 \end{matrix} \right\|.$$

The *second way* to solve problem of the joint application of the method data partitioning and criterion of unbiasedness errors is the choose the best division next view

$$\ell^* = \arg \max_{\rho^2_{B\ell} \neq 0,\ \ell=\overline{1,L}} \left\| \mathbf{X}^T_{A\ell} \mathbf{X}_{A\ell} - \rho^2_{B\ell} \mathbf{X}^T_{B\ell} \mathbf{X}_{B\ell} \right\|$$

or use subsamples *A* and *B* " dissimilar by the variance"

$$l^* = \arg \max_{l=\overline{1,L_d}} tr(\tilde{\mathbf{X}}^T_{Al} \tilde{\mathbf{X}}_{Al} - \tilde{\mathbf{X}}^T_{Bl} \tilde{\mathbf{X}}_{Bl}) = \arg \max_{l=\overline{1,L_d}} \left[tr(\tilde{\mathbf{X}}^T_{Al} \tilde{\mathbf{X}}_{Al}) - tr(\tilde{\mathbf{X}}^T_{Bl} \tilde{\mathbf{X}}_{Bl}) \right].$$

The *third variant* is maximizing norm of difference of information matrices when the best partitioning find as

$$\ell^* = \arg \max_{\ell=\overline{1,L}} \left\| \mathbf{X}^T_{A\ell} \mathbf{X}_{A\ell} - \mathbf{X}^T_{B\ell} \mathbf{X}_{B\ell} \right\|.$$

Analysis of *ρ*-optimal partitioning and of it practicable analogue – quasi-optimal partitioning of samples gives three variants of data division methods corresponding to selection of the best (nontrivial) model by unbiasedness errors criterion.

4 Geometric illustration of the variance partitioning of samples

The paper [9] shows the relationship between the criteria of ρ-proportional partitioning and "similar by the variance" partitioning. In order to satisfy the balance condition of the sets size, the most commonly subsamples ratios are used:

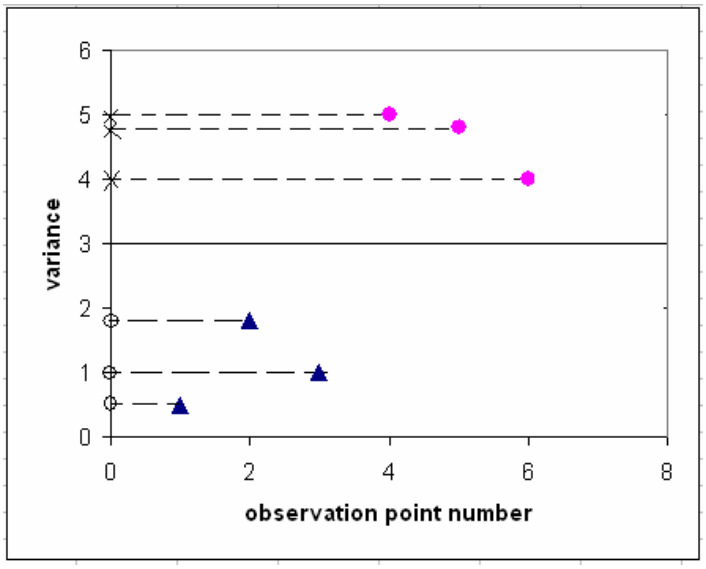

Fig. 2. Design observation data on the variance values axle

1) 1: 1, which is associated with the selection of sub-samples an equal volume of *A* and *B* at the design of observation points on a straight line;

2) 2: 1, for searching the median of the distance (midpoint) between the sets S_1, S_2 and S_3 at the design of their points on the plane and presenting them in the form of three equipotent sets.

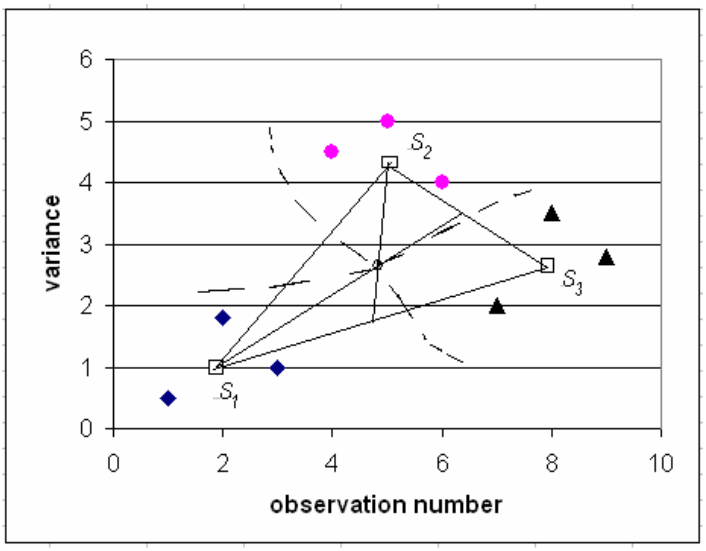

Fig. 3. Partitioning of the observation points on two sets *A* and *B* in the ratio 2: 1 by placement of sets S_1, S_2 and S_3.

Figure 2 shows how at the design observation point on the variance values axle can be obtained sub-samples *A* and *B* in the ratio of their sizes 1: 1. The value of variance of 3 divides the data into two equal subsamples A and B with equidistance from the point of division.

Figure 3 illustrates how in the plane of two coordinates it can be found partitioning of the observation points on two sets in the ratio 2: 1, where $A = S_i \cup S_j$, $B = S_k$, $i \neq j \neq k$.

In the case of "similar by variance" with the partitioning ratio of 1: 1, there is reason to expect better conditions in order to achieve the minimum criteria using for the selection of the model the criterion of regularity. «Dissimilar by the variance" partition expediently at using criterion of unbiasedness errors, the better result can be expected when the ratio is 2: 1 instead of a ratio = 1: 1.

5 Conclusion

We investigated criteria of unbiasedness errors, theoretically justified their use in certain types of optimal way of partitioning the sample. If the condition of proportionality of data is satisfied, the theoretical substantiation were obtained that the criterion of unbiasedness errors, is not "adequate" to the noise criterion while minimizing the criterion of the sample division, but only when criterion is maximized or if is fulfilled scheme of repeated experiments.

References

[1] Ivakhnenko A. G. Systems of heuristic self-organization in technical cybernetics. *Tekhnika*, Kiev, USSR. – 1971. – 372p. (In Russian).

[2] Pavlov A. A. Criterion for ranking of self-selection variables using the threshold in GMDH algorithms. *Avtomatika*. Kiev, USSR. – 1969. – N 4. – p. 89–91. (In Ukrainian)

[3] Kondrashova N. V. Matching of external criterion and method of sample partitioning for solving problem of structural-parametric identification by GMDH. *International scientific and technical journal problems of control and informatics*. NASU. Institute of Cybernetics. Space Research Institute of NCAU. Kiev, Ukraine. – 2015. – N 4, – p.20–32.

[4] Vysotsky V. N. About the best division of initial data in GMDH algorithms. *Avtomatika* – Kiev, USSR. –1976. – N 3 – p.71–74. (In Ukrainian)

[5] Ivakhnenko A. G., Stepashko V.S. Noise-immunity of modeling. *Naukova dumka*, Kiev, USSR. – 1985. – 214p. (In Russian).

[6] Ivakhnenko A. G., Yurachkovsky Yu. P. Modeling of complex systems on experimental data. *Radio and Communications*. Moscow, USSR – 1987. – 119 p. (In Russian)

[7] Sarychev A. P. Solution of partitioning problem in GMDH at calculation criterion of regularity in conditions of active experiment. *Avtomatika*. Kiev, USSR. – 1989. – N 4. – p. 19–27. (In Russian).

[8] Stepashko V. S., Kondrashova N. V. Evaluation of the transformation of geometric shapes. *Proc. 1st International workshop on inductive modeling*. (IWIM 2005). Lviv Politechnic National University. Institute of CS and IT. Lviv, Ukraine.– p.294–301. (In Russian).

[9] Kondrashova N. V. Influence of a way for partitioning sample in GMDH algorithms on the adequacy of criterion external supplement. *Upravlyayushchiye sistemy i mashiny (CS&C)*. NASU. IRTC IT&S. Institute of Cybernetics. Kiev, Ukraine. – 2015. –N 4. – p. 20 -32. (In Russian).