СИСТЕМНИЙ АНАЛІЗ І ТЕОРІЯ ПРИЙНЯТТЯ РІШЕНЬ

SYSTEM ANALYSIS AND DECISION-MAKING THEORY

DOI: 10.20998/2079-0023.2023.01.01
UDC 004:519.24:681.3.06

A. A. PAVLOV, Doctor of Technical Sciences, Full Professor, National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute", Kyiv, Ukraine, Professor of Informatics and Software Engineering Department; e-mail: pavlov.fiot@gmail.com; ORCID: https://orcid.org/0000-0002-6524-6410

M. N. HOLOVCHENKO, National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute", Kyiv, Ukraine, Senior Lecturer of Informatics and Software Engineering Department; e-mail: ma4ete25@ukr.net; ORCID: https://orcid.org/0000-0002-9575-8046

V. V. DROZD, National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute", Bachelor of Informatics and Software Engineering Department, Kyiv, Ukraine, e-mail: drozdllera@gmail.com, ORCID: https://orcid.org/0000-0003-0418-1139

EFFICIENCY SUBSTANTIATION FOR A SYNTHETICAL METHOD OF CONSTRUCTING A MULTIVARIATE POLYNOMIAL REGRESSION GIVEN BY A REDUNDANT REPRESENTATION

In recent years, the authors in their publications have developed two different approaches to the construction of a multivariate polynomial (in particular, linear) regressions given by a redundant representation. The first approach allowed us to reduce estimation of coefficients for nonlinear terms of a multivariate polynomial regression to construction of a sequence of univariate polynomial regressions and solution of corresponding nondegenerate systems of linear equations. The second approach was implemented using an example of a multivariate linear regression given by a redundant representation and led to the creation of a method the authors called a modified group method of data handling (GMDH), as it is a modification of the well-known heuristic self-organization method of GMDH (the author of GMDH is an Academician of the National Academy of Sciences of Ukraine O. G. Ivakhnenko). The modification takes into account that giving a multivariate linear regression by redundant representation allows for construction of a set of partial representations, one of which has the structure of the desired regression, to use not a multilevel selection algorithm, but an efficient algorithm for splitting the coefficients of the multivariate linear regression into two classes. As in the classic GMDH, the solution is found using a test sequence of data. This method is easily extended to the case of a multivariate polynomial regression since the unknown coefficients appear in the multivariate polynomial regression in a linear way. Each of the two approaches has its advantages and disadvantages. The obvious next step is to combine both approaches into one. This has led to the creation of a synthetic method that implements the advantages of both approaches, partially compensating for their disadvantages. This paper presents the aggregated algorithmic structure of the synthetic method, the theoretical properties of partial cases and, as a result, the justification of its overall efficiency.

Keywords: univariate polynomial regression, multivariate polynomial regression, redundant representation, least squares method, test sequence, repeated experiment.

Introduction. Multivariate linear and non-linear regressions, constructed based on the results of active or passive experiments, are widely used in modern diagnostic information systems, in particular, medical ones, and in information management systems with a wide range of applications [1–10]. Universal methods for multivariate regressions construction vary from classical statistical methods to heuristics, such as the group method of data handling (GMDH) or genetic algorithms. But none of them, due to the complexity of the problem, dominates the others. Existing methods complement each other. Therefore, scientific research in this field is still relevant.

We give in the abstract, at a qualitative level, the characteristics of a synthetic method of constructing a multivariate polynomial regression (MPR) given by a redundant representation. In this paper, we explain the aggregated algorithmic structure of the synthetic method, substantiate the logic of its construction, its theoretical properties, the properties of partial cases of redundant representations of MPRs that lead to the finding of coefficient estimates for nonlinear members of the MPR with acceptable accuracy.

1. General theoretical provisions that we use. *1.1. Univariate polynomial regression (UPR)* [11]. Let a UPR be given in the form:

$$Y(x) = \theta_0 + \theta_1 x + ... + \theta_r x^r + E\ , \tag{1}$$

where E is a random variable with an arbitrary distribution, its mathematical expectation $ME = 0$, its variance $DE = \sigma^2 < \infty$, the variance or its upper bound is known.

© Pavlov A. A., Holovchenko M. N., Drozd V. V., 2023

Research Article: This article was published by the publishing house of ***NTU "KhPI"*** in the collection "Bulletin of the National Technical University "KhPI" Series: System analysis, management and information technologies." This article is distributed under a Creative Common Creative Common Attribution (CC BY 4.0). **Conflict of Interest:** The author/s declared no conflict of interest.

According to the results of the experiment $\left(x_i \to y_i,\ i=\overline{1,n}\right)$, $r<n$, $x_i \in [c,d]$, where

$$y_i = \theta_0 + \theta_1 x_i + \ldots + \theta_r x_i^r + \delta_i\,, \tag{2}$$

where δ_i is an implementation of a random variable E in the i-th test, estimates of unknown coefficients can be found using normalized orthogonal polynomials of Forsythe (NOPFs) constructed from the values of the input variable $x_i, i=\overline{1,n}$.

$$Q_j(x) = q_{j0} + q_{j1}x + \ldots + q_{jj}x^j,\ j=\overline{0,r}\,.$$

In this case, the UPR (1) has the form

$$Y(x) = \sum_{j=0}^{r} \omega_j Q_j(x) + E\,, \tag{3}$$

$$\hat{\omega}_j = \sum_{i=1}^{n} y_i Q_j(x_i),\ j=\overline{0,r}\,, \tag{4}$$

$$\hat{\theta}_j = \hat{\omega}_r q_{rj} + \ldots + \hat{\omega}_j q_{jj},\ j=\overline{0,r}\,, \tag{5}$$

$$M\hat{\omega}_j = \omega_j, D\hat{\omega}_j = \sigma^2, M\hat{\theta}_j = \theta_j, D\hat{\theta}_j = \sigma^2 \sum_{j=1}^{r} q_{jj}^2\,. \tag{6}$$

1.2. A repeated experiment. A. Pavlov and D. Kovalenko proved that the results of a repeated experiment with k repetitions of the main experiment $\left(x_i \to y_i,\ i=\overline{1,n}\right)$ are equivalent to the experiment

$$\left(x_i \to \frac{\sum_{p=0}^{k-1} y_{p\cdot n+i}}{k},\ i=\overline{1,n}\right),$$

and the variances of estimates $\hat{\theta}_j$, $j=\overline{0,r}$, decrease by the factor of k.

1.3. Estimation of coefficients for nonlinear members of a UPR using a single set of NOPFs. It was shown in [12] that the estimates of the coefficients for nonlinear terms of a UPR (1) at arbitrary values of $c<d$ can be found using only a single set of NOPFs based on the results of the following virtual experiment. We introduce a virtual deterministic scalar variable z and for its values

$$z_1 < z_2 < \ldots < z_n \tag{7}$$

we find with the specified accuracy a set of NOPFs $Q_j(z)$, $j=\overline{0,r}$. Having put

$$a = \frac{d-c}{z_n - z_1} > 0,\ b = c - \frac{d-c}{z_n - z_1} z_1\,,$$

$$x_i = az_i + b,\ i=\overline{1,n}\ \left(x_1 = c,\ x_n = d\right), \tag{8}$$

substituting $x = az+b$, we reduce the UPR (1) to a virtual UPR

$$Y(z) = \theta_0 + \theta_1(az+b) + \ldots + \theta_r(az+b)^r + E =$$

$$= \gamma_0 + \gamma_1 z + \ldots + \gamma_r z^r + E\,. \tag{9}$$

The coefficients θ_j, γ_j, $j=\overline{0,r}$, are in a mutually unambiguous correspondence

$$\begin{pmatrix} 1 & & & \\ & a & & a_{ij} \\ & & a^2 & \\ & 0 & & \ddots \\ & & & & a^r \end{pmatrix} \begin{pmatrix} \theta_0 \\ \theta_1 \\ \vdots \\ \theta_r \end{pmatrix} = \begin{pmatrix} \gamma_0 \\ \gamma_1 \\ \vdots \\ \gamma_r \end{pmatrix}. \tag{10}$$

According to results of the main experiment $\left(x_i \to y_i,\ i=\overline{1,n}\right)$, where $x_i, i=\overline{1,n}$, satisfy (8), we design a virtual experiment $\left(z_i \to y_i,\ i=\overline{1,n}\right)$. Based on the set of NOPFs constructed for $z_i, i=\overline{1,n}$, we find estimates $\hat{\gamma}_j$, $j=\overline{0,r}$, and based on the system of equations (10) we find estimates $\hat{\theta}_j$, $j=\overline{0,r}$. It is shown in [12] that

$$\sum_{i=1}^{n}\left(y_j - \sum_{j=0}^{r}\hat{\theta}_j x_i^j\right)^2 = \min_{\theta_j,\ j=\overline{0,r}} \sum_{i=1}^{n}\left(y_j - \sum_{j=0}^{r}\theta_j x_i^j\right)^2.$$

2. Estimation with acceptable accuracy of the coefficients for nonlinear members of a UPR (1) using a single set of NOPFs. *2.1. Justification of the number of tests n of the main experiment and the choice of values of $z_i, i=\overline{1,n}$, of the virtual scalar variable z.* With the appropriate selection of the values of the virtual scalar variable z, expression (10) significantly simplifies the preliminary analysis and the input data formation for the main active experiment to obtain, with acceptable accuracy, estimates of the coefficients for nonlinear terms of the UPR and, as will be shown later, an MPR. Such a requirement is for $r_{\max} < n$ a compromise between the number of tests n of the main experiment and the value of variances of the estimates of the coefficients for nonlinear members of the virtual UPR (9). As a result of the analysis of the conducted experiments, the following compromise solution is proposed for $r_{\max} \le 5$: $n=10$, $z_1=-50$, $z_{10}=50$, $\Delta z = (z_i - z_{i-1}) = \text{const}$. In this case,

$$D\hat{\gamma}_2 = 4.26\cdot 10^{-6}\sigma^2,\ D\hat{\gamma}_3 = 7.55\cdot 10^{-9}\sigma^2\,,$$

$$D\hat{\gamma}_4 = 1.4\cdot 10^{-12}\sigma^2,\ D\hat{\gamma}_5 = 1.28\cdot 10^{-15}\sigma^2\,, \tag{11}$$

that is, with an increase of j by one, $D\hat{\gamma}_j$ is decreased by three orders of magnitude, starting from $D\hat{\gamma}_2$.

2.2. Justification of conditions for obtaining estimates for nonlinear members of a UPR (1) with acceptable accuracy. In [12], analytical expressions are given for variances of estimates $\hat{\theta}_j$, $j \ge 2$, for a UPR (1). The expressions were obtained from the results of estimation by a virtual UPR (9). This makes it possible to determine, depending on the values of c, d, a, b (8), the number k of repetitions of the main experiment (variances of the coefficient estimates $\hat{\theta}_j$ are reduced by the factor of k) that, using the three-sigma rule, leads to obtaining the estimates $\hat{\theta}_j$, $j \geq 2$, with acceptable accuracy. Thus, when the repeated experiment turned out to be practically feasible, the problem has a solution.

Remark 1. The three-sigma rule for an arbitrary random variable X is formulated as follows. With a probability of 0,89, any realization of X belongs to the segment $[MX - 3\sigma, MX + 3\sigma]$, where $\sigma = \sqrt{DX}$.

Remark 2. Due to the roughness of the three-sigma rule, the actual number of repetitions k is significantly less.

3. A decomposition method of estimating the coefficients of an MPR given by a redundant representation. *3.1. The problem formulation*. Let an MPR be given by the following redundant representation [13]:

$$Y(\bar{\boldsymbol{x}}) = \sum_{\forall (i_1,\ldots,i_t) \in K,\, \forall (j_1,\ldots,j_t) \in K(i_1,\ldots,i_t)} b_{i_1\ldots i_t}^{j_1\ldots j_t} \left(x_{i_1}\right)^{j_1} \cdots \left(x_{i_t}\right)^{j_t} + E, \quad (12)$$

where $\bar{\boldsymbol{x}} = (x_1,\ldots,x_m)^{\mathrm{T}}$ is a deterministic vector of input variables,

$$x_i \in [c_i, d_i],\ c_i > 0,\ i = \overline{1,m}, \qquad (13)$$

where E is a random variable, $ME = 0$, $DE = \sigma^2 < \infty$. The values of the coefficients $b_{i_1\ldots i_t}^{j_1\ldots j_t}$ are unknown (b_0^0 is a constant).

Remark 3. The condition $c_i > 0$, $i = \overline{1,m}$, corresponds to the most common practical case. All the results obtained below are trivially extended to arbitrary real values of the numbers $c_i < d_i$, $i = \overline{1,m}$.

3.2. The decomposition method [13] implements the methodology of reducing the estimation of nonlinear members of an MPR (12) to the sequential construction of UPRs and the solution of the corresponding systems of linear equations, the variables of which are the estimates for the nonlinear members of the MPR (12).

The general algorithmic scheme for obtaining estimates for nonlinear members of the MPR (12) consists of two sub-algorithms [13].

3.2.1. Aggregated modified algorithmic scheme of the first sub-algorithm [13]. The l-th step ($l \leq L_1$, where L_1 is the total number of nonlinear components in (12), each of which contains at least one scalar variable raised to a power greater than or equal to two) is implemented for the next nonlinear term of the MPR (12), whose coefficient was not estimated at the previous steps of the first sub-algorithm and that contains a scalar variable to the maximum power. Let's denote it as x_{i_p}. In the MPR (12), the scalar variable x_{i_p} is replaced with a virtual scalar variable z: $x_{i_p} = a_{i_p} z + b_{i_p}$, where a_{i_p}, b_{i_p} are found according to (8) for $d = d_{i_p}$, $c = c_{i_p}$. In the real main experiment, the scalar variable x_{i_p} takes the value according to (7), (8), and the other scalar variables in all tests take fixed values. In this case, the MPR (12) is transformed into a UPR, and the data of the main virtual experiment are found based on the main real experiment

$$\left(x_{i_p,i} \ \underset{j \neq i_p}{\forall}\ x_{j,i} = x_j^{\mathrm{F}} \rightarrow y_i,\ i = \overline{1,n} \right),$$

namely:

$$\left(z_i \rightarrow y_i - \sum_{\forall b_{i_1\ldots i_t}^{j_1\ldots j_t} \in \bigcup_{m=1}^{l-1}\{J_m\}} b_{i_1\ldots i_t}^{j_1\ldots j_t} \left(x_{i_1,i}\right)^{j_1} \cdots \left(x_{i_t,i}\right)^{j_t},\ i = \overline{1,n} \right), \quad (14)$$

where $\bigcup_{m=1}^{l-1}\{J_m\}$ is the set of coefficients estimated with acceptable accuracy at the previous steps of the first sub-algorithm.

Remark 4. In a similar way, we found the data of a repeated virtual experiment, in which the number of output data is $y_i, i = \overline{1,kn}$, and the input data of the main experiment is repeated k times.

Remark 5. The maximum degree of a UPR is j_p.

The number of UPRs constructed at the l-th step can be more than one if the corresponding coefficient(s) of the first UPR is (are) expressed linearly by several coefficients for the nonlinear terms of the MPR (12) not evaluated at the previous steps of the first sub-algorithm.

The right-hand parts of the obtained nondegenerate systems of linear equations are the estimates for nonlinear terms of the UPRs. Their solutions are the estimates of the corresponding coefficients for nonlinear terms of the MPR (12).

3.2.2. Aggregated modified scheme of the second sub-algorithm [13]. The l-th step ($l \leq L_2$, where L_2 is the number of nonlinear components in (12) of the form $b_{i_1\ldots i_t}^{1\ldots 1} \times$ $\times x_{i_1} \cdots x_{i_t}$) is implemented for a nonlinear coefficient $b_{i_1\ldots i_{t_l}}^{1\ldots 1}$ that was not evaluated at the previous steps and has a maximum value of t_l. Each input variable x_{i_j}, $j = \overline{1,t_l}$, is expressed linearly by a virtual variable z: $x_{i_j} = a_{i_j} z + b_{i_j}$ according to (8) for $c = c_{i_j}$, $d = d_{i_j}$, $j = \overline{1,t_l}$. In the main experiment, the variables x_{i_j}, $j = \overline{1,t_l}$, vary according to (7), (8). Other scalar input variables take fixed values in each test. The data of the virtual main experiment are determined based on the data of the main experiment

$$\left(x_{i_1,i},\ldots,x_{i_{t_l},i}\ \underset{j}{\forall}\ x_{j,i} = x_j^{\mathrm{F}},\ j \notin \left\{i_1,\ldots,i_{t_t}\right\} \rightarrow y_i,\ i = \overline{1,n} \right),$$

namely:

$$\left(z_i \rightarrow y_i - \sum_{\forall b_{i_1\ldots i_t}^{j_1\ldots j_t} \in \bigcup_{m=1}^{K_1}\{J_m\}} \hat{b}_{i_1\ldots i_t}^{j_1\ldots j_t} \left(x_{i_1,i}\right)^{j_1} \cdots \left(x_{i_t,i}\right)^{j_t} - \right.$$

$$\left. - \sum_{\forall b_{i_1\ldots i_t}^{j_1\ldots j_t} \in \bigcup_{m=1}^{l-1}\{G_m\}} \hat{b}_{i_1\ldots i_t}^{1\ldots 1} \prod_{l=1}^{t_l} x_{i_l,i},\ i = \overline{1,n} \right), \qquad (15)$$

where K_1 is the number of steps of the first sub-algorithm; G_m is a set of coefficients estimated with sufficient accuracy at the previous steps of the second sub-algorithm.

Remark 6. The data for the repeated virtual experiment is found in a similar way.

Remark 7. The maximum degree of a UPR is t_l .

3.3. Classes of redundant representations for which the coefficients for nonlinear terms of an MPR (12) are estimated completely or partially with acceptable accuracy. 3.3.1. The redundant representation of (12) satisfies the following three conditions.

1) $\forall \left(x_{i_l}\right)^{j_l}$, $j_l \geq 2$ can be included in only one component of (12);

2) for any two nonlinear components of (12) with coefficients $b_{i_1 \ldots i_{t_i}}^{j_1 \cdots j_{t_i}}$, $b_{l_1 \ldots l_{t_l}}^{p_1 \cdots p_{t_l}}$ the following is true:

$$\left\{i_1, \ldots, i_{t_i}\right\} \neq \left\{l_1, \ldots, l_{t_l}\right\} ; \tag{16}$$

3) $$0 \in \left[c_i, d_i\right], i = \overline{1, m} , \tag{17}$$

where m is the number of input variables.

Then the l -th step of the first sub-algorithm of the decomposition method implemented for the input scalar variable $\left(x_{i_p}\right)^{j_p}$, $j_p \geq 2$, of the nonlinear component of (12)

$$b_{i_1 \ldots i_t}^{j_1 \cdots j_t} \left(x_{i_1}\right)^{j_1} \cdots \left(x_{i_t}\right)^{j_t} \tag{18}$$

must satisfy the following additional conditions. In all tests of the active experiment, in each component of (12) $b_{i_1 \ldots i_t}^{j_1 \cdots j_t} \left(x_{i_1}\right)^{j_1} \cdots \left(x_{i_t}\right)^{j_t}$, $b_{i_1 \ldots i_t}^{j_1 \cdots j_t} \in \bigcup_{m=1}^{l-1} \left\{J_m\right\}$ (see (14)), one input scalar variable whose index is not included in the set $\left\{i_1, \ldots, i_{t_i}\right\}$ (18) is equal to zero. In this case, the linear equation for estimating $b_{i_1 \ldots i_t}^{j_1 \cdots j_t}$ (18) has the form

$$b_{i_1 \ldots i_t}^{j_1 \cdots j_t} \left(a_{i_p}\right)^{j_p} \prod_{\substack{m=1 \\ m \neq p}}^{t} x_{i_m}^{\mathrm{F}} = \hat{\gamma}_{j_p} = \gamma_{j_p} \pm \left|\varepsilon_{\hat{\gamma}_{j_p}}\right| , \tag{19}$$

where $\hat{\gamma}_{j_p}$ is the estimate of γ_{j_p} — the coefficient of the corresponding virtual UPR (depending on its meaning, it is a realization or a random variable), and $\varepsilon_{\hat{\gamma}_{j_p}}$ is the realization of a random variable with zero mathematical expectation and the variance equal to $D\hat{\gamma}_{j_p}$ (if $\hat{\gamma}_{j_p}$ is considered a random variable). Let $\varepsilon\left(\sum_{l=1}^{t} j_l\right) > 0$ (this is an expertly set upper bound on the value of $\left|b_{i_1 \ldots i_t}^{j_1 \cdots j_t}\right|$ below which the corresponding component (18) is excluded from the redundant representation (12) as not essential). Then the values of $x_{i_m}^{\mathrm{F}}$ $\forall i_m$, $i_m \neq p$, are set equal to $x_{i_m}^{\mathrm{F}} = d_{i_m} > 0, m \neq p$, and thereby the number of repetitions of the main experiment that guarantees the fulfillment of the inequality

$$\left|\varepsilon_{\hat{\gamma}_{j_p}}\right| \leq 10^{-1} \left(a_{i_p}\right)^{j_p} \varepsilon\left(\sum_{l=1}^{t} j_l\right) \prod_{\substack{m=1 \\ m \neq p}}^{t} d_{i_m} \tag{20}$$

(with the corresponding probability, using the three-sigma rule), is the minimum possible.

Remark 8. To obtain the estimate (20), the three-sigma rule is applied to the random variable $\hat{\gamma}_{j_p}$, which realization is used to find the estimate of the coefficient $\hat{b}_{i_1 \ldots i_t}^{j_1 \cdots j_t}$ (18).

If

$$\left|\hat{\gamma}_{j_p}\right| < \left(a_{i_p}\right)^{j_p} \varepsilon\left(\sum_{l=1}^{t} j_l\right) \prod_{\substack{m=1 \\ m \neq p}}^{t} d_{i_m} , \tag{21}$$

then the component (18) is excluded from the redundant representation (12). In the opposite case, we have found an estimate of the coefficient for the component (18) with an acceptable accuracy (20).

The l -th step of the second sub-algorithm of the decomposition method implemented for the component

$$b_{i_1 \ldots i_t}^{1 \ldots 1} \prod_{l=1}^{t} x_{i_l} , \ t = \max , \tag{22}$$

must satisfy the following additional conditions. In all tests of the active experiment in each component of (12) of the form $b_{i_1 \ldots i_t}^{j_1 \cdots j_t} \left(x_{i_1}\right)^{j_1} \cdots \left(x_{i_t}\right)^{j_t}$, $b_{i_1 \ldots i_t}^{j_1 \cdots j_t} \in \left\{\bigcup_{m=1}^{K_1} \left\{J_m\right\}\right\} \bigcup \left\{\bigcup_{m=1}^{l-1} \left\{G_m\right\}\right\}$ (see (15)), one input scalar variable whose index does not belong to the set $\left\{i_1, \ldots, i_t\right\}$ (22) is equal to zero. The linear equation for estimating $b_{i_1 \ldots i_t}^{1 \ldots 1}$ (22) has the form

$$b_{i_1 \ldots i_t}^{1 \ldots 1} \prod_{l=1}^{t} a_{i_l} = \hat{\gamma}_t = \gamma_t \pm \left|\varepsilon_{\hat{\gamma}_t}\right| , \tag{23}$$

where $\hat{\gamma}_t$ is the estimate of γ_t — the coefficient of the corresponding virtual UPR. Let the number of repetitions of the main experiment be feasible to fulfill, based on the three-sigma rule, the limitation

$$\left|\varepsilon_{\hat{\gamma}_t}\right| \leq 10^{-1} \varepsilon(t) \prod_{l=1}^{t} a_{i_l} . \tag{24}$$

Then, if

$$\left|\hat{\gamma}_t\right| < \varepsilon(t) \prod_{l=1}^{t} a_{i_l} , \tag{25}$$

then the component (22) is excluded from the redundant representation (12). In the opposite case, we have found an estimate of the coefficient $b_{i_1 \ldots i_t}^{1 \ldots 1}$ (22) with an acceptable accuracy (24).

Thus, if all the repeated experiments are feasible, the decomposition method allows you to find all estimates for nonlinear terms of the MPR (12) with acceptable accuracy.

Corollary. If for some steps of the first or second sub-algorithms the required number of repetitions of the main experiment is unfeasible, then the corresponding coefficients are not estimated, and conditions (15), (16) allow such components of the redundant representation (12) to be excluded from other steps of the first and second sub-algorithms.

Remark 9. Each step of the first and second sub-algorithms estimates only one coefficient of the MPR (12).

3.3.2. The redundant representation (12) is given by the following four conditions.

The first condition: all input variables included in the redundant representation (12) of the form $b_{i_1\ldots i_t}^{j_1\ldots j_t} \times \times \left(x_{i_1}\right)^{j_1} \cdots \left(x_{i_t}\right)^{j_t}$, $\sum_{l=1}^{t} j_l \ge 2$, have a range of acceptable values $\left\lfloor 1, d_{i_l} \right\rfloor$, $l = \overline{1,t}$.

The second condition: if an arbitrary input variable is included in an arbitrary component of (12) to a power greater than or equal to two, it is no longer included in any other component of (12) of the form $b_{i_1\ldots i_t}^{j_1\ldots j_t}\left(x_{i_1}\right)^{j_1} \cdots \left(x_{i_t}\right)^{j_t}$, $\sum_{l=1}^{t} j_l \ge t+1$.

The third condition: all nonlinear components of the form $b_{i_1\ldots i_t}^{1\ldots 1}\prod_{l=1}^{t} x_{i_l}$ do not have common input variables.

The fourth condition: no input variable included in any component of (12) that has $\exists j_p \ge 2$, is included in any component of (12) whose coefficient has the form $b_{i_1\ldots i_t}^{1\ldots 1}$, $t \ge 2$.

Let the number of repetitions of the main experiment for an arbitrary step of the first and second sub-algorithms be feasible for obtaining, using the three-sigma rule, the estimates:

a) for the first sub-algorithm:

$$\left|\varepsilon_{\hat{\gamma}_p}\right| \le 10^{-1}\left(a_{i_p}\right)^{j_p} \varepsilon\left(\sum_{l=1}^{t} j_l\right); \tag{26}$$

b) for the second sub-algorithm: estimate (24).

Then, if during the implementation of an arbitrary step of the first and second sub-algorithms, the fixed input variables included in the nonlinear terms of the MPR in expressions (14), (15) are set equal to one, then in expressions (14), (15) from the values of y_i will be subtracted the values of $\sum\limits_{\forall(\cdot)} \hat{b}_{i_1\ldots i_t}^{j_1\ldots j_t}$, which will differ in modulus from the values of $\sum\limits_{\forall(\cdot)} b_{i_1\ldots i_t}^{j_1\ldots j_t}$, due to estimates (24), (26) and the roughness of the three-sigma rule, by values that can be practically neglected. Then all estimates for non-linear terms of the MPR (12) are found with acceptable accuracy.

Remark 10. The number of repetitions of the main experiment can be real if $a_i > 1/3$, $i = \overline{1,m}$. This is also true for the second sub-algorithm for the class 3.3.1.

Remark 11. The number of repetitions of the main experiment at some steps of the first sub-algorithm can be significantly smaller, if for this step the estimate for $\left|\varepsilon_{\hat{\gamma}_p}\right|$ is of the form (20). That is, some input variables that are not included in (14), (15) can take maximum values.

Remark 12. If we exclude the third and fourth conditions imposed on the redundant representation (12), then the decomposition method implements only the first sub-algorithm.

Remark 13. Each step of the decomposition method for the redundant representation (12) that satisfies the four or the first two conditions of the class 3.3.2, estimates only one coefficient of the redundant representation (12).

3.3.3 (generalization of the class 3.3.2). A redundant representation (12) satisfies the first condition of the class 3.3.1. $c_i < d_i$, $c_i > 0$, $i = \overline{1,m}$, are arbitrary numbers. The second condition is set based on the results of the implementation of the following version of the decomposition method. When performing the l-th ($l \ge 2$) step of the first sub-algorithm for the term $b_{i_1\ldots i_{t_l}}^{j_1\ldots j_{t_l}}\left(x_{i_1}\right)^{j_1} \cdots \left(x_{i_{t_l}}\right)^{j_{t_l}}$, the fixed values of the input variables in (14), which are not part of the set $\left\{x_{i_1}, \ldots, x_{i_{t_l}}\right\}$, are set modulo the minimum possible. When implementing an arbitrary step of the second sub-algorithm, all fixed variables included in (15) take modulo minimum values.

Let the number of repetitions of the main experiment for each step of the first and second sub-algorithms be feasible to fulfill constraints (20), (24). Then the second condition imposed on the redundant representation (12) is

$$\forall\left|\hat{b}_{i_1\ldots i_t}^{j_1\ldots j_t}\right| \ge \varepsilon\left(\sum_{l=1}^{t} j_l\right)\cdot 10^{p},\ p \ge 2,\ \text{or}$$

$$\left|\hat{b}_{i_1\ldots i_t}^{j_1\ldots j_t}\right| < \varepsilon\left(\sum_{l=1}^{t} j_l\right).$$

It is obvious that for any real values of $c_i < d_i$, $i = \overline{1,m}$, $r_{\max}$ (the maximum degree of the virtual UPR) there is such a small enough value of the natural number p that the replacement of exact values of $b_{i_1\ldots i_t}^{j_1\ldots j_t}$ in expressions (14), (15) with their estimates is statistically guaranteed to have practically no effect on the values of coefficient estimates for nonlinear terms of the MPR (12).

Remark 14. Only one coefficient of the MPR (12) is estimated at each step of the decomposition algorithm.

3.3.4. The redundant representation of the MPR (12) is arbitrary, the random variable $E = 0$. Such a case occurs when E is a measurement error and is neglected having sufficient measurement accuracy. In this case, the general algorithmic procedure of the decomposition method accurately finds the values of all coefficients for the nonlinear terms of (12). The exact values of the coefficients for linear terms, including the constant, are found by the usual interpolation procedure, which consists in solving a system of linear equations with the number of variables $m+1$, where m is the number of input variables.

The advantage of the decomposition method over an arbitrary interpolation method lies is two-fold: the coefficients for nonlinear terms of the redundant representation are estimated using only a single set of NOPFs found with a given accuracy; the needed-to-solve nondegenerate systems of linear equations have a significantly smaller dimension than the number of nonlinear components of (12), and in most cases the dimension is equal to one.

3.3.5. The redundant representation (12) is arbitrary. In general case, the detailed formalization of the first and second sub-algorithms for finding estimates with sufficient accuracy is inefficient. A visual analysis of the specific redundant representation (12) using the above theoretical results allows for each specific case to create an efficient sequence of steps of an individual algorithm of the decomposition method (which can be a step of the first or second sub-algorithm in any order), which leads to estimation with acceptable accuracy of the maximum number of coefficients for nonlinear terms of the MPR (12).

4. A modified group method of data handling (MGMDH). As a result of the implementation of the decomposition method, we obtained a set $\{J\}$ of coefficients for nonlinear terms of the MPR (12) estimated with acceptable accuracy.

Remark 15. Insignificant coefficients are excluded from the redundant representation (12) and the set $\{J\}$.

To obtain the final result, we propose to use the MGMDH described in [14] to construct a multivariate linear regression (MLR) given by a redundant representation. Indeed, the problem of constructing an MPR given by a redundant representation (12) was reduced to the following:

$$Y(\bar{x}) = \sum_{\forall\left\{b_{i_1\cdots i_t}^{j_1\cdots j_t}\right\}\setminus\{J\}} b_{i_1\cdots i_t}^{j_1\cdots j_t}\left(x_{i_1}\right)^{j_1}\cdots\left(x_{i_t}\right)^{j_t} + f(\bar{x}) + E, \qquad (27)$$

where $f(\bar{x}) = \sum_{\forall b_{i_1\cdots i_t}^{j_1\cdots j_t}\in\{J\}} \hat{b}_{i_1\cdots i_t}^{j_1\cdots j_t}\left(x_{i_1}\right)^{j_1}\cdots\left(x_{i_t}\right)^{j_t}$.

The regression problem without any changes (to an accuracy of the content of the columns of the matrix A in the expression $\left(A^{\mathrm{T}}A\right)^{-1}A^{\mathrm{T}}$ of the general formula of the least squares method) can be solved by the method [14], since all unknown coefficients are included linearly in expression (27).

Remark 16. As shown in [14], the use of k repetitions of the main experiment significantly increases the computational efficiency of the MGMDH. Namely, the variance of estimates is reduced by a factor of k, and only the matrices of the main experiment are used in the general formula of the least squares method.

Remark 17. The MGMDH can solve the problem of constructing an MPR given by a redundant representation based also on the results of a passive experiment. In this case, disappear only the advantages of an active experiment: the possibility of using repetitions of the main experiment.

Conclusions. 1. We considered the aggregated algorithmic scheme of the universal synthetic method of constructing an MPR given by a redundant representation. The synthetic method organically combines the decomposition method of estimating coefficients for nonlinear terms of the MPR with acceptable accuracy and the modified group method of data handling.

2. We presented theoretically justified classes of redundant representation of MPR, which allow to estimate fully or partially coefficients for nonlinear terms of an MPR with acceptable accuracy.

3. We substantiated the possibility of extending the modified group method of data handling, created for estimating the coefficients of an MLR given by a redundant representation, to the case of estimating the coefficients of an MPR given by a redundant representation.

References

1. Yu L. Using negative binomial regression analysis to predict software faults: a study of Apache Ant. *International Journal of Information Technology and Computer Science (IJITCS)*. 2012. Vol. 4, No. 8. P. 63–70. DOI: 10.5815/ijitcs.2012.08.08.
2. Shahrel M.Z., Mutalib S., Abdul-Rahman S. PriceCop – price monitor and prediction using linear regression and LSVM-ABC methods for e-commerce platform. *International Journal of Information Engineering and Electronic Business (IJIEEB)*. 2021. Vol. 13, no. 1. P. 1–14. DOI: 10.5815/ijieeb.2021.01.01.
3. Satter A., Ibtehaz N. A regression based sensor data prediction technique to analyze data trustworthiness in cyber-physical system. *International Journal of Information Engineering and Electronic Business (IJIEEB)*. 2018. Vol. 10, no. 3. P. 15–22. DOI: 10.5815/ijieeb.2018.03.03.
4. Isabona J., Ojuh D. O. Machine learning based on kernel function controlled gaussian process regression method for in-depth extrapolative analysis of Covid-19 daily cases drift rates. *International Journal of Mathematical Sciences and Computing (IJMSC)*. 2021. Vol. 7, no. 2. P. 14–23. DOI: 10.5815/ijmsc.2021.02.02.
5. Sinha P. Multivariate polynomial regression in data mining: methodology, problems and solutions. *International Journal of Scientific & Engineering Research*. 2013. Vol. 4, iss. 12. P. 962–965
6. Kalivas J. H. Interrelationships of multivariate regression methods using eigenvector basis sets. *Journal of Chemometrics*. 1999. Vol. 13 (2). P. 111–132. DOI: 10.1002/(SICI)1099-128X(199903/04)13:2<111::AID-CEM532>3.0.CO;2-N.
7. Ortiz-Herrero L., Maguregui M. I., Bartolomé L. Multivariate (O)PLS regression methods in forensic dating. *TrAC Trends in Analytical Chemistry*. 2021. Vol. 141. 116278. DOI: 10.1016/j.trac.2021.116278.
8. Guo G., Niu G., Shi Q., Lin Q., Tian D., Duan Y. Multi-element quantitative analysis of soils by laser induced breakdown spectroscopy (LIBS) coupled with univariate and multivariate regression methods. *Analytical Methods*. 2019. Vol. 11, iss. 23. P. 3006–3013. DOI: 10.1039/C9AY00890J.
9. Настенко Е., Павлов В., Бойко Г., Носовец О. Многокритериальный алгоритм шаговой регрессии. *Біомедична інженерія і технологія*, 2020. № 3. С. 48–53. DOI: 10.20535/2617-8974.2020.3.195661.
10. Babatunde G., Emmanuel A. A., Oluwaseun O. R., Bunmi O. B., Precious A. E. Impact of climatic change on agricultural product yield using *k*-means and multiple linear regressions. *International Journal of Education and Management Engineering (IJEME)*. 2019. Vol. 9, no. 3. P. 16–26. DOI: 10.5815/ijeme.2019.03.02.
11. Худсон Д. *Статистика для физиков: Лекции по теории вероятностей и элементарной статистике*. Москва: Мир, 1970. 296 с.
12. Pavlov A. A. Holovchenko M. N., Drozd V. V. Construction of a multivariate polynomial given by a redundant description in stochastic and deterministic formulations using an active experiment. *Вісник Нац. техн. ун-ту «ХПІ»: зб. наук. пр. Темат. вип.: Системний аналіз, управління та інформаційні технології*. Харків: НТУ «ХПІ», 2022. № 1 (7). С. 3–8. DOI: 10.20998/2079-0023.2022.01.01.
13. Pavlov A., Holovchenko M., Mukha I., Lishchuk K., Drozd V. A Modified Method and an Architecture of a Software for a Multivariate Polynomial Regression Building Based on the Results of a Conditional Active Experiment. *Advances in Computer Science for Engineering and Education VI (ICCSEEA 2023)*. 2023. (у друці)
14. Pavlov A. A., Holovchenko M. N. Modified method of constructing a multivariate linear regression given by a redundant description. *Вісник Нац. техн. ун-ту «ХПІ»: зб. наук. пр. Темат. вип.: Системний аналіз, управління та інформаційні технології*. Харків: НТУ «ХПІ», 2022. № 2 (8). С. 3–8. DOI: 10.20998/2079-0023.2022.02.01.

References (transliterated)

1. Yu L. Using negative binomial regression analysis to predict software faults: a study of Apache Ant. *International Journal of Information Technology and Computer Science (IJITCS)*. 2012, vol. 4, no. 8, pp. 63–70. DOI: 10.5815/ijitcs.2012.08.08.
2. Shahrel M.Z., Mutalib S., Abdul-Rahman S. PriceCop – price monitor and prediction using linear regression and LSVM-ABC methods for e-commerce platform. *International Journal of Information Engineering and Electronic Business (IJIEEB)*. 2021, vol. 13, no. 1, pp. 1–14. DOI: 10.5815/ijieeb.2021.01.01.

3. Satter A., Ibtehaz N. A regression based sensor data prediction technique to analyze data trustworthiness in cyber-physical system. *International Journal of Information Engineering and Electronic Business (IJIEEB).* 2018, vol. 10, no. 3, pp. 15–22. DOI: 10.5815/ijieeb.2018.03.03.
4. Isabona J., Ojuh D. O. Machine learning based on kernel function controlled gaussian process regression method for in-depth extrapolative analysis of Covid-19 daily cases drift rates. *International Journal of Mathematical Sciences and Computing (IJMSC).* 2021, vol. 7, no. 2, pp. 14–23. DOI: 10.5815/ijmsc.2021.02.02.
5. Sinha P. Multivariate polynomial regression in data mining: methodology, problems and solutions. *International Journal of Scientific & Engineering Research*. 2013, vol. 4, iss. 12, pp. 962–965.
6. Kalivas J. H. Interrelationships of multivariate regression methods using eigenvector basis sets. *Journal of Chemometrics*. 1999, vol. 13 (2), pp. 111–132. DOI: 10.1002/(SICI)1099-128X(199903/04)13:2<111::AID-CEM532>3.0.CO;2-N.
7. Ortiz-Herrero L., Maguregui M. I., Bartolomé L. Multivariate (O)PLS regression methods in forensic dating. *TrAC Trends in Analytical Chemistry*. 2021, vol. 141, 116278. DOI: 10.1016/j.trac.2021.116278.
8. Guo G., Niu G., Shi Q. et al. Multi-element quantitative analysis of soils by laser induced breakdown spectroscopy (LIBS) coupled with univariate and multivariate regression methods. *Analytical Methods*. 2019, vol. 11, iss. 23, pp. 3006–3013. DOI: 10.1039/C9AY00890J.
9. Nastenko E., Pavlov V., Boyko G., Nosovets O. Mnogokriterial'nyj algoritm shagovoj regressii. *Biomedychna inzheneriya i tekhnolohiya* [Biomedical ingeneering and technology]. 2020, no. 3, pp. 48–53. DOI: 10.20535/2617-8974.2020.3.195661.
10. Babatunde G., Emmanuel A. A., Oluwaseun O. R., Bunmi O. B., Precious A. E. Impact of climatic change on agricultural product yield using *k*-means and multiple linear regressions. *International Journal of Education and Management Engineering (IJEME).* 2019, vol. 9, no. 3, pp. 16–26. DOI: 10.5815/ijeme.2019.03.02.
11. Hudson D. J. Statistics Lectures, Volume 2: Maximum Likelihood and Least Squares Theory. CERN Reports 64(18). Geneva, CERN, 1964. (Russ. ed.: Hudson D. *Statistika dlja fizikov: Lekcii po teorii verojatnostej i jelementarnoj statistike.* Moscow, Mir Publ., 1970. 296 p.). DOI: 10.5170/CERN-1964-018.
12. Pavlov A. A. Holovchenko M. N., Drozd V. V. Construction of a multivariate polynomial given by a redundant description in stochastic and deterministic formulations using an active experiment. *Visnyk Nats. tekhn. un-tu "KhPI": zb. nauk. pr. Temat. vyp.: Systemnyy analiz, upravlinnya ta informatsiyni tekhnologiyi* [Bulletin of the National Technical University "KhPI": a collection of scientific papers. Thematic issue: System analysis, management and information technology]. Kharkov, NTU "KhPI" Publ., 2022, no. 1 (7), pp. 3–8. DOI: 10.20998/2079-0023.2022.01.01.
13. Pavlov A., Holovchenko M., Mukha I. et al. A Modified Method and an Architecture of a Software for a Multivariate Polynomial Regression Building Based on the Results of a Conditional Active Experiment. *Advances in Computer Science for Engineering and Education VI (ICCSEEA 2023).* 2023. (to appear)
14. Pavlov A. A., Holovchenko M. N. Modified method of constructing a multivariate linear regression given by a redundant description. *Visnyk Nats. tekhn. un-tu "KhPI": zb. nauk. pr. Temat. vyp.: Systemnyy analiz, upravlinnya ta informatsiyni tekhnologiyi* [Bulletin of the National Technical University "KhPI": a collection of scientific papers. Thematic issue: System analysis, management and information technology]. Kharkov, NTU "KhPI" Publ., 2022, no. 2 (8), pp. 3–8. DOI: 10.20998/2079-0023.2022.02.01.

Received 15.04.2023

УДК 004:519.24:681.3.06

О. А. ПАВЛОВ, доктор технічних наук, професор, Національний технічний університет України «Київський політехнічний інститут імені Ігоря Сікорського», м. Київ, Україна, професор кафедри інформатики та програмної інженерії; e-mail: pavlov.fiot@gmail.com; ORCID: https://orcid.org/0000-0002-6524-6410

М. М. ГОЛОВЧЕНКО, Національний технічний університет України «Київський політехнічний інститут імені Ігоря Сікорського», м. Київ, Україна, старший викладач кафедри інформатики та програмної інженерії; e-mail: ma4ete25@ukr.net; ORCID: https://orcid.org/0000-0002-9575-8046

В. В. ДРОЗД, Національний технічний університет України «Київський політехнічний інститут імені Ігоря Сікорського», м. Київ, Україна, бакалавр кафедри інформатики та програмної інженерії; e-mail: drozdllera@gmail.com, ORCID: https://orcid.org/0000-0003-0418-1139

ОБҐРУНТУВАННЯ ЕФЕКТИВНОСТІ СИНТЕТИЧНОГО МЕТОДУ ПОБУДОВИ БАГАТОВИМІРНОЇ ПОЛІНОМІАЛЬНОЇ РЕГРЕСІЇ, ЗАДАНОЇ НАДЛИШКОВИМ ОПИСОМ

Протягом останніх років автори в своїх публікаціях паралельно розвивали два різних підходи до побудови багатовимірних поліноміальних, зокрема, лінійних регресій, заданих надлишковим описом. Перший підхід дозволяв знаходження оцінок коефіцієнтів при нелінійних членах багатовимірної поліноміальної регресії зводити до побудови послідовності одновимірних поліноміальних регресій та розв'язання відповідних невироджених систем лінійних рівнянь. Другий підхід був реалізований на прикладі багатовимірної лінійної регресії, заданої надлишковим описом, і привів до створення методу, названого авторами модифікованим методом групового урахування аргументів (МГУА), так як він є модифікацією широко відомого методу евристичної самоорганізації МГУА (автор МГУА – академік НАН України О. Г. Івахненко). Модифікація полягає в тому, що завдання багатовимірної лінійної регресії надлишковим описом дозволяє для побудови множини часткових описів, один з яких має структуру шуканої регресії, використовувати не багаторівневий селекційний алгоритм, а ефективний алгоритм розбиття коефіцієнтів багатовимірної лінійної регресії на два класи. Як і в класичному МГУА, розв'язок знаходиться за допомогою перевірочної послідовності даних. Цей метод легко поширюється на випадок багатовимірної поліноміальної регресії, так як невідомі коефіцієнти в багатовимірну поліноміальну регресію входять лінійно. Кожен з двох підходів має свої переваги і недоліки. Очевидним наступним кроком є поєднання обох підходів в один. Це призвело до створення синтетичного методу, який реалізує переваги обох підходів, частково компенсуючи їх недоліки. В цій роботі наведена агрегована алгоритмічна структура синтетичного методу, теоретичні властивості часткових випадків і, як наслідок, обгрунтування його ефективності в цілому.

Ключові слова: одновимірна поліноміальна регресія, багатовимірна поліноміальна регресія, надлишковий опис, метод найменших квадратів, перевірочна послідовність, повторний експеримент.

Повні імена авторів / Author's full names

Автор 1 / Author 1: Павлов Олександр Анатолійович, Pavlov Alexander Anatolievich
Автор 2 / Author 2: Головченко Максим Миколайович, Holovchenko Maxim Nikolaevich
Автор 3 / Author 3: Дрозд Валерія Валеріївна, Drozd Valeriia Valeriivna