СИСТЕМНИЙ АНАЛІЗ І ТЕОРІЯ ПРИЙНЯТТЯ РІШЕНЬ

SYSTEM ANALYSIS AND DECISION-MAKING THEORY

DOI: 10.20998/2079-0023.2024.01.01
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

AN ADAPTIVE METHOD FOR BUILDING A MULTIVARIATE REGRESSION

We propose an adaptive method for building a multivariate regression given by a weighted linear convolution of known scalar functions of deterministic input variables with unknown coefficients. As, for example, when multivariate regression is given by a multivariate polynomial. In contrast to the general procedure of the least squares method that minimizes only a single scalar quantitative measure, the adaptive method uses six different quantitative measures and represents a systemically connected set of different algorithms which allow each applied problem to be solved on their basis by an individual adaptive algorithm that, in the case of an active experiment, even for a relatively small volume of experimental data, implements a strategy of a statistically justified solving. The small amount of data of the active experiment we use in the sense that, for such an amount, the variances of estimates of unknown coefficients obtained by the general procedure of the least squares method do not allow to guarantee the accuracy acceptable for practice. We also proposed to significantly increase the efficiency of the proposed by O. A. Pavlov. and M. M. Holovchenko modified group method of data handling for building a multivariate regression which is linear with respect to unknown coefficients and given by a redundant representation. We improve it by including some criteria and algorithms of the adaptive method for building a multivariate regression. For the multivariate polynomial regression problem, the inclusion of a partial case of the new version of the modified group method of data handling in the synthetic method proposed by O. A. Pavlov, M. M. Golovchenko, and V. V. Drozd, for building a multivariate polynomial regression given by a redundant representation, also significantly increases its efficiency.

Keywords: multivariate regression, integral measure, adaptive algorithm, regression analysis, expert coefficients, linear programming.

1. Introduction. In recent years, the authors have been working in the field of regression analysis, namely developing efficient methods for building univariate and multivariate regressions (MR) which are linear in relation to unknown coefficients [1, 2, 3]. The conducted critical analysis of existing universal methods for building an MR [4–20] showed that this problem is still relevant in both theoretical and applied aspects. The new approach implemented in this paper consists in that the universal adaptive method proposed by the authors (which includes six different criteria and four algorithms based on the outlined methodology of their use) allows to create an individual algorithm for an efficient solution of each individual applied problem.

2. The adaptive method for building an MR. ***2.1. Formulation of the problem.*** The BR model looks like

$$Y(\bar{x}) = \sum_{j=1}^{L} \theta_j \psi_j(\bar{x}) + E, \qquad (1)$$

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

E is a random variable, its mathematical expectation is $ME = 0$, its variance $\mathrm{Var}(E) = \sigma^2 < \infty$, the value of $\mathrm{Var}(E)$ is known or its efficient estimate is known;

$\psi(\bar{x})$ are known numerical scalar functions of the vector argument $\bar{x}$. In [1], such functions were the components of a multidimensional polynomial.

According to the results of an active experiment $\left(\bar{x}_i \to y_i, i = \overline{1,n}\right)$ we need to estimate the value of the unknown coefficients θ_j, $j = \overline{1,L}$.

2.2. Measures of deviation of experimental data from the regression model used by the adaptive method. *2.2.1. A classical measure implemented by the general scheme of the least squares method (LSM).* A vector of estimates $\hat{\theta}_1$ of unknown components of the vector $\theta = (\theta_1, \ldots, \theta_L)^{\mathrm{T}}$ minimizes the next measure:

$$\min_{\theta_j,\, j=\overline{1,L}} \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{L} \theta_j \psi_j(\bar{x}_i) \right)^2. \qquad (2)$$

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 an international license Creative Common Attribution (CC BY 4.0). **Conflict of Interest:** The author/s declared no conflict of interest.

It is known that

$$\hat{\theta}_1 = \left(A^{\mathrm{T}}A\right)^{-1}A^{\mathrm{T}}y, \qquad (3)$$

where

$$A = \begin{pmatrix} \psi_1(\bar{x}_1) & \cdots & \psi_L(\bar{x}_1) \\ \vdots & \ddots & \vdots \\ \psi_1(\bar{x}_n) & \cdots & \psi_L(\bar{x}_n) \end{pmatrix},\ y = (y_1,\ldots,y_n)^{\mathrm{T}}.$$

As is known [21], the estimates $\hat{\theta}_1$ are linear, unbiased, efficient in the class of linear estimates (Markov's theorem). One can propose a statistically significantly more efficient linear estimate of the vector θ than the estimate obtained by the LSM only in the case when the structure of the algorithm for finding the linear estimate is adaptive, not fixed, that is, it depends on the input data and intermediate results of its execution. This is the main methodological idea that formed the basis of the proposed adaptive method.

2.2.2. Minimization of the sum of modules. Looks like

$$\min_{\theta_j,\ j=\overline{1,L}} \sum_{i=1}^{n} \left| y_i - \sum_{j=1}^{L} \theta_j \psi_j(\bar{x}_i) \right|. \qquad (4)$$

Remark 1. The use of measure (4) in its explicit form is inefficient due to Markov's theorem.

2.2.3. A measure that minimizes the module of the sum of differences. Looks like

$$\min_{\theta_j,\ j=\overline{1,L}} \left| \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{L} \theta_j \psi_j(\bar{x}_i) \right) \right|. \qquad (5)$$

We recommend using it in explicit form only if it is known that the density function $f(x)$ of the random variable E is symmetrical about the ordinate axis.

2.2.4. The measure of MR deviations from experimental data realized with a given probability. For exact values of θ_j, $j=\overline{1,L}$, the following is met:

$$\sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{L} \theta_j \psi_j(\bar{x}_i) \right) = \sum_{i=1}^{n} \varepsilon_i, \qquad (6)$$

where ε_i is the i-th realization of the random variable E.

A random variable $\frac{1}{n}\sum_{i=1}^{n} E_i$ (E_i are independent copies of E) for $n \geq 20$ on the basis of a partial case of the scalar limit theorem practically has a normal distribution with parameters $M\left(\frac{1}{n}\sum_{i=1}^{n} E_i\right) = 0$, $\mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^{n} E_i\right) = \frac{\sigma^2}{n}$. Let us find $t_{n,\alpha,\sigma}$, for which

$$P\left(\left| \frac{1}{n}\sum_{i=1}^{n} E_i \right| \leq t_{n,\alpha,\sigma} \right) = 1-\alpha, \qquad (7)$$

where α is set experimentally, we recommend choosing an $\alpha \leq 0.1$. Then

$$t_{n,\alpha,\sigma} = \frac{\sigma}{\sqrt{n}} \Phi_0^{-1}\left(\frac{1-\alpha}{2}\right), \qquad (8)$$

where $\Phi_0^{-1}(x)$ is the inverse Laplace transform and σ is the arithmetic root of $\mathrm{Var}(E)$ or its efficient estimate. Then the measure 2.2.4 is the condition for θ_j, $j=\overline{1,L}$, which is met with the probability $1-\alpha$:

$$\left| \frac{1}{n}\sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{L} \theta_j \psi_j(\bar{x}_i) \right) \right| \leq t_{n,\alpha,\sigma}. \qquad (9)$$

Remark 2. Additional fifth and sixth measures are implemented when we know the density function $f(x)$ of the random variable E or the $f(x)$ is known with the known with the accuracy of the values of its numerical parameters. From the methodological point of view, it is convenient to introduce them later (see subsections 2.3.5, 2.3.6).

2.3. Algorithms of the adaptive method*. 2.3.1. The first algorithm (the first version).* The first step. Find the following estimate $\hat{\theta}_1$:

$$\hat{\theta}_1 = \left(A^{\mathrm{T}}A\right)^{-1}A^{\mathrm{T}}y.$$

The second step. Solve the following problem of linear programming (LP):

$$\min \sum_{i=1}^{n} z_i, \qquad (10)$$

$$-z_i \leq y_i - \sum_{j=1}^{L} \theta_j \psi_j(\bar{x}_i) \leq z_i,\ \ z_i \geq 0,\ \ i=\overline{1,n},$$

$$-n \cdot t_{n,\alpha,\sigma} \leq \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{L} \theta_j \psi_j(\bar{x}_i) \right) \leq n \cdot t_{n,\alpha,\sigma}. \qquad (11)$$

The variables of the LP problem (10), (11) are θ_j, $j=\overline{1,L}$, z_i, $i=\overline{1,n}$. Let us denote by $\hat{\theta}_2$ the optimal solution of the LP problem (10), (11).

The third step. The solution of the first algorithm is $\hat{\theta}_1$ if

$$\left| \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{L} \hat{\theta}_{1j} \psi_j(\bar{x}_i) \right) \right| \leq n \cdot t_{n,\alpha,\sigma} \qquad (12)$$

and $\hat{\theta}_2$ if condition (12) is not fulfilled.

2.3.2. The first algorithm (the second version). It differs from the algorithm of subsection 2.3.1 in the second step, in which we solve the following LP problem:

$$\min \sum_{i=1}^{n} \left(u_i^+ + u_i^- \right), \qquad (13)$$

$$y_i - \sum_{j=1}^{L} \theta_j \psi_j(\bar{x}_i) = u_i^+ - u_i^-, u_i^+ \geq 0, u_i^- \geq 0, i=\overline{1,n},$$

$$-n \cdot t_{n,\alpha,\sigma} \leq \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{L} \theta_j \psi_j(\bar{x}_i) \right) \leq n \cdot t_{n,\alpha,\sigma} \qquad (14)$$

The variables of the LP problem (13), (14) are θ_j, $j=\overline{1,L}$, u_i^+, u_i^-, $i=\overline{1,n}$.

Remark 3. The LP problem (13), (14) is solved by the simplex method, since only in this case the fulfillment of the conditions $\forall i\ u_i^+ \cdot u_i^- = 0$ is guaranteed. Whence, the optimal functional values for the LP problems (10), (11) and (13), (14) are the same. The advantage of the LP problem (13), (14) over the problem (10), (11) is that in its standard form it has n variables and equations less.

2.3.3. The second algorithm (the first version). The vector $\hat{\theta}_3$ of estimates of unknown components of the vector θ is a solution to the next LP problem:

$$\min z\,, \tag{15}$$

$$-z \le \sum_{i=1}^{n}\left(y_i - \sum_{j=1}^{L}\theta_j\psi_j(\bar{x}_i)\right) \le z,\ z \ge 0 \tag{16}$$

where the variables of the LP problem (15), (16) are θ_j, $j=\overline{1,L}$, z.

2.3.4. The second algorithm (the second version). $\hat{\theta}_3$ is a solution to the following LP problem:

$$\min\left(u^+ + u^-\right), \tag{17}$$

$$\sum_{i=1}^{n}\left(y_i - \sum_{j=1}^{L}\theta_j\psi_j(\bar{x}_i)\right) = u^+ - u^-\,. \tag{18}$$

The variables of the LP problem (17), (18) are θ_j, $j=\overline{1,L}$, u^+, u^-.

Remark 4. We recommend using the second algorithm when it is known that the density function of the random variable E is symmetric about the ordinate axis and $L << n$.

2.3.5. The first hybrid algorithm of the adaptive method. The hybrid method is used only when the density function of the random variable E is known at least with the accuracy of the values of its numerical parameters, and the number of tests n allows sufficiently precise test of the complex hypothesis by the χ^2 statistic.

The first step. Find the estimates $\hat{\theta}_1$, $\hat{\theta}_2$, $\hat{\theta}_3$.

The second step. Find the values sequence $\hat{\theta}_4\left(\omega_{1,l},\omega_2\right)$, $0 < a = \omega_{1,1} < \omega_{1,2} < \ldots < \omega_{1,k_1} = b$, $\omega_{1,l} - \omega_{1,l-1} = \text{const}$, $l = \overline{2,k_1}$, $\omega_2 = \text{const} > 0$ as solutions to the following LP problems (the first version):

$$\min\left(\omega_{1,l}\sum_{i=1}^{n} z_i + \omega_2 z\right), \tag{19}$$

$$-z_i \le y_i - \sum_{j=1}^{L}\theta_j\psi_j(\bar{x}_i) \le z_i,\ i=\overline{1,n},$$

$$-z \le \sum_{i=1}^{n}\left(y_i - \sum_{j=1}^{L}\theta_j\psi_j(\bar{x}_i)\right) \le z \tag{20}$$

or the LP problems (the second version):

$$\min\left\{\omega_{1,l}\sum_{i=1}^{n}\left(u_i^+ + u_i^-\right) + \omega_2\left(z^+ + z^-\right)\right\}, \tag{21}$$

$$y_i - \sum_{j=1}^{L}\theta_j\psi_j(\bar{x}_i) = u_i^+ + u_i^-;\ u_i^+ \ge 0, u_i^- \ge 0, i=\overline{1,n},$$

$$\sum_{i=1}^{n}\left(y_i - \sum_{j=1}^{L}\theta_j\psi_j(\bar{x}_i)\right) = z^+ - z^-;\ z^+, z^- \ge 0\,. \tag{22}$$

The variables of the LP problem (19), (20) are θ_j, $j=\overline{1,L}$, z_i, $i=\overline{1,n}$, z, and for the LP problem (21), (22) θ_j, $j=\overline{1,L}$, u_i^+, u_i^-, $i=\overline{1,n}$, z^+, z^-.

Remark 5. The LP problem (21), (22) can be solved only by the simplex method.

The third step. For each of k_1+3 found vector estimates $\hat{\theta}_1$, $\hat{\theta}_2$, $\hat{\theta}_3$, $\hat{\theta}_4\left(\omega_{1,l},\omega_2\right)$, $l=\overline{1,k_1}$, find estimates of realizations of the random variable E as

$$y_i - \sum_{j=1}^{L}\hat{\theta}_{p,j}\psi_j(\bar{x}_i),\ i=\overline{1,n},\ p=\overline{1,k_1+3}\,.$$

For $p>3$ $\hat{\theta}_p = \hat{\theta}_4\left(\omega_{1,p-3},\omega_2\right)$, $p=\overline{4,k_1+3}$. The estimate of the components of an unknown vector θ is the vector $\hat{\theta}_p$, $p=\overline{1,k_1+3}$, with which the χ^2 statistic for testing the simple or complex hypothesis about the density function of the random variable E has the minimum value.

Remark 6. Let χ_p^2 be the realization of the χ^2 statistic for the estimate found by the hybrid algorithm. Then, if $P\left(\chi^2 \ge \chi_p^2\right) > 0.05$, then the solution is unreliable, otherwise, the realization of the $\chi^2 - \chi_p^2$ statistic belongs to the feasible region with significance level 0.05, and the smaller the value of χ_p^2 (provided that $\chi_p^2 \ge r-2$, where r is the number of degrees of freedom of the χ^2 statistic), the more reliable is the vector θ of estimates of the components of the unknown vector found by the hybrid algorithm.

2.3.6. The second hybrid algorithm of the adaptive method. It is used when the distribution of the random variable E is known. Let us denote $\nu_{|E|} = M|E|$, $\hat{\nu}_{|E|}$ is a practically exact estimate of $\nu_{|E|}$ obtained as a result of simulation modeling, not related to tests on the regression model (1).

Remark 7. $\frac{1}{n}\sum_{i=1}^{n}|E_i|$ is a consistent sampling characteristic for estimating the unknown vector $\nu_{|E|}$, where $E_i, i=\overline{1,n}$, are independent copies of the random variable E. That is, it coincides with $\nu_{|E|}$ with a probability of 1 when $n\to\infty$. Whence, with a sufficiently large one n, $\frac{1}{n}\sum_{i=1}^{n}|\varepsilon_i|$ in absolute magnitude differs from $\nu_{|E|}$ by a sufficiently small value, where ε_i is the result of the i-th test on the random variable E (i-th test on the regression model (1)).

The estimate of the vector θ by the second hybrid algorithm. $\hat{\theta}_5(\omega_l)$, $0 < a_2 = \omega_1 < \omega_2 < \ldots < \omega_{k_2} = b_2$, $\forall\omega_l - \omega_{l-1} = \text{const}$, is the solution to the following LP problem:

$$\min\sum_{i=1}^{n} z_i\,, \tag{23}$$

$$-z_i \le y_i - \sum_{j=1}^{L}\theta_j\psi_j(\bar{x}_i) \le z_i,\ z_i \ge 0,\ i=\overline{1,n},$$

$$-\omega_l \le \frac{1}{n}\sum_{i=1}^{n} z_i - \hat{\nu}_{|E|} \le \omega_l,$$

$$-n\cdot t_{n,\alpha,\sigma} \le \sum_{j=1}^{L}\left(y_i - \theta_j\psi_j(\bar{x}_i)\right) \le n\cdot t_{n,\alpha,\sigma}\,. \tag{24}$$

The variables of the LP problem (23), (24) are z_i, $i=\overline{1,n}$, θ_j, $j=\overline{1,L}$. Further, the description of the second hybrid algorithm coincides with the corresponding description of the first hybrid algorithm.

Remark 8. $b_2 - a_2$ can be a sufficiently small number (see *Remark* 7). That is, for true values of θ_j, $j = \overline{1,L}$, there is small enough ω_l, in which condition (24) is fulfilled (under the assumption that $\hat{\nu}_{|E|} \approx \nu_{|E|}$).

3. Algorithm for reducing the number of components of the expression (1) for an MR given by a redundant representation. A necessary condition for using algorithms of the adaptive method is that L is significantly less than n. We can offer the following algorithm for reducing the value of L for a fixed set of experimental data.

The first step. Regression (1) is fictitiously increased by one $m+1$-th deterministic input variable x_{m+1}, that is,

$$Y(\bar{x}) = \sum_{j=1}^{L} \theta_j \psi_j(\bar{x}) + \theta_{m+1} x_{m+1} + E, \qquad (25)$$

where $\theta_{m+1} = 0$. Data of the experiment $\left(\bar{x}_i \to y_i, i = \overline{1,n}\right)$ are replaced by $\left(\left(x_{1i}, \ldots, x_{mi}, \frac{1}{m}\sum_{j=1}^{m} x_{ji}\right) \to y_i, i = \overline{1,n}\right)$.

The second step. Find the estimates of θ_j by the LSM, $j = \overline{1, L+1}$, at $\theta_{m+1} \equiv 0$.

The third step. By the cluster analysis algorithm [2] the set of coefficients $\left\{\theta_j, j = \overline{1, L+1}\right\}$ is divided into two classes M_1 and M_2 [2], θ_{L+1} should belong to M_2. Otherwise, the algorithm stops.

Remark 9. We can change the values of $x_{L+1,i}$, $i = \overline{1,n}$, and repeat steps 1–3. The necessary condition for the continuation of the algorithm is $\theta_{L+1} \in M_2$.

The fourth step. Exclude all terms, which coefficients belong to M_2, from expression (25).

The fifth and subsequent steps of the algorithm. Repeat steps 1–3 for the new expression of the MR (provided that for each iteration the coefficient for the fictitious input variable belongs to the class M_2) until only the coefficient for the fictitious input variable will remain in the class M_2. The resulting expression of the MR statistically significantly does not contain terms that insignificantly affect the output variable, and the difference between the number of tests n and the number of MR members can be significantly reduced if really the MR was given by a redundant representation. This will lead to an increase in the efficiency of using the adaptive method.

4. Active experiment. Finding the analytical expression of the density function of the random variable E and the estimate of its variance. We consider the case when an active experiment can be carried out in sufficient quantity at a fixed value of the deterministic input variables $\bar{x}_{\mathrm{f}}$. The number of tests should be sufficient to test the complex hypothesis about the density function of an arbitrary random variable by the χ^2 statistic, as well as for efficient estimation of the variance of a random variable.

4.1. Finding the analytical expression of the density function of a random variable E. The first step. Find a sample of the volume n of realizations of the random variable $E + \sum_{j=1}^{L} \theta_j \bar{x}_{\mathrm{f}}$. They are the values of the output variable from the experiment $\left(\bar{x}_{\mathrm{f}} \to y_i, i = \overline{1,n}\right)$.

The second step. By the values of y_i, $i = \overline{1,n}$, build a histogram, the geometric expression of which allows to deduce a complex hypothesis about the density function of the random variable $E + \sum_{j=1}^{L} \theta_j \bar{x}_{\mathrm{f}}$, which is checked by the most "hard" χ^2 statistic.

Remark 10. Obviously, a more complex procedure for testing a complex hypothesis can be proposed. If altering the value of $M\left(E + \forall \mathrm{const}\right)$ affects solely the numerical parameters within the analytical expression that represents the density function for the random variable E, then, in case of acceptance of the complex hypothesis, the problem is considered solved.

4.2. Estimation of $\mathrm{Var}(E)$. As an estimate of the variance of a random variable E, we can take the number

$$\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \frac{1}{n}\sum_{j=1}^{n} y_j\right)^2$$

because $\mathrm{Var}\left(E + \forall \mathrm{const}\right) = \mathrm{Var}\left(E\right)$.

5. The use of some provisions of the adaptive method in the modified group method of data handling (MGMDH) [2]. We propose to make the following changes to the general algorithmic scheme of MGMDH [2].

5.1. Using two regular criteria. In the case when the analytical expression of the density function of the random variable E is known with the accuracy of the values of the numerical parameters and is symmetric with respect to the ordinate axis, then we propose when finding by the test sequence of experimental data, that is, data that were not used to find estimates of unknown coefficients of partial descriptions of the sought regression, to use not the single regular criterion which is the residual sum of squares, but two criteria.

The first regular criterion:

$$\sum_{i=n+1}^{n+n_1}\left(y_i - \sum_{j=1}^{r_l} \hat{\theta}_{l_j} \psi_{l_j}\left(\bar{x}_i\right)\right)^2, \qquad (26)$$

where the set of coefficients $\left\{\hat{\theta}_{l_j}, j = \overline{1, r_l}\right\}$ gives the l-th partial descriptions of the sought MR; n is the amount of empirical data used to find estimates $\underset{l}{\forall} \theta_{l_j}, j = \overline{1, r_l}$, of the coefficients of the l-th partial description; n_1 is the number of experimental data in the test sequence.

The second regular criterion:

$$\left|\sum_{i=n+1}^{n+n_1}\left(y_i - \sum_{j=1}^{r_l} \hat{\theta}_{l_j} \psi_{l_j}\left(\bar{x}_i\right)\right)\right|. \qquad (27)$$

Thus, in the general case, we get two partial descriptions that claim to be used to find the correct structure of the sought MR, and not a single one.

5.2. The use of the hybrid algorithm. The hybrid algorithm (the first or the second one, depending on the avail-

able information) of the adaptive method is used to find estimates of coefficients for the entire ($n + n_1$) experimental data set for the found partial description(s). The final estimate of the unknown MR is the one for which the realization of the χ^2 statistic is the smallest [2].

Remark 11. If the symmetry condition of the density function of the random variable E is not fulfilled, then the use of the second regular criterion (27) is redundant.

Remark 12. More detailed recommendations on the use of criteria and algorithms of the adaptive method will be the result of its careful experimental research. Now it can be stated that finding estimates of the MR coefficients with the simultaneous use of several criteria significantly expands the application possibilities of universal methods of regression analysis, in particular those proposed by the authors of the adaptive method.

Conclusions. 1. To estimate the coefficients of an MR which is linear with respect to unknown coefficients, we proposed a universal adaptive method that implements not a single criterion, as in the least squares method, but several criteria. The adaptation consists in the fact that the structure of the algorithm for the final result obtaining is not fixed but depends on the intermediate results of calculations and input data.

2. We considered the possibility of a statistically correct obtaining, based on the results of a special experiment, an analytical expression of the density function of a random variable that affects additively the output variable.

3. We showed how the use of the adaptive method can significantly increase the efficiency of the modified group method of data handling proposed earlier by the authors.

References

1. Pavlov A. A., Holovchenko M. N., Drozd V. V. Efficiency substantiation for a synthetical method of constructing a multivariate polynomial regression given by a redundant representation. *Вісник Нац. техн. ун-ту «ХПІ»: зб. наук. пр. Темат. вип.: Системний аналіз, управління та інформаційні технології*. Харків: НТУ «ХПІ», 2023. № 1 (9). С. 3–9. DOI: 10.20998/2079-0023.2023.01.01.
2. 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.
3. 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. *Lecture Notes on Data Engineering and Communications Technologies*. 2023. Vol. 181. P. 207–222. DOI: 10.1007/978-3-031-36118-0_19.
4. Abdulrahman A. T., Alshammari N. S. Factor analysis and regression analysis to find out the influencing factors that led to the countries' debt crisis. *Advances and Applications in Statistics.* 2022. Vol. 78. P. 1–16. DOI: 10.17654/0972361722047.
5. Flitman A. M. Towards analysing student failures: neural networks compared with regression analysis and multiple discriminant analysis. *Computers and Operations Research.* 1997. Vol. 24, no. 4. P. 367–377. DOI: 10.1016/s0305-0548(96)00060-3.
6. Johnson R. A., Wichern D. W. *Applied multivariate statistical analysis*, 5th edn. Upper Saddle River: Prentice-Hall, 2002. 767 p.
7. Knowles D., Parts L., Glass D., Winn J. M. Modeling skin and ageing phenotypes using latent variable models in Infer.NET. *Predictive models in personalized medicine workshop, NIPS 2010*. URL: https://www.researchgate.net/publication/241194775 (дата звернення: 18.05.2024).
8. Lio W., Liu B. Uncertain maximum likelihood estimation with application to uncertain regression analysis. *Soft Computing*. 2020. Vol. 24, no. 13. P. 9351–9360. DOI: 10.1007/s00500-020-04951-3.
9. Liu S.S., Zhu Y. Simultaneous maximum likelihood estimation for piecewise linear instrumental variable models. *Entropy*. 2022. Vol. 24, no. 9. P. 1235. DOI: 10.3390/e24091235.
10. Ruff L., Vandermeulen R., Goernitz N., Deecke D., Siddiqui S. A., Binder A., Müller E., Kloft M. Deep one-class classification. *Proceedings of the 35th international conference on machine learning, PMLR 80*. 2018. P. 4393–4402. URL: http://proceedings.mlr.press/v80/ruff18a/ruff18a.pdf (дата звернення: 18.05.2024).
11. Scott J. T. Factor analysis and regression. *Econometrica*. 1966. Vol. 34. No. 3. P. 552–562. DOI: 10.2307/1909769.
12. Buckley J. J., Feuring T. Linear and non-linear fuzzy regression: Evolutionary algorithm solutions. *Fuzzy Sets and Systems*. 2000. Vol. 112. No. 3. P. 381–394. DOI: 10.1016/s0165-0114(98)00154-7.
13. Draper N. R., Smith H. *Applied regression analysis*, 3rd edn. New York: Wiley & Sons, 1998. 736 p. DOI: 10.1002/9781118625590.
14. Ивахненко А. Г. *Моделирование сложных систем.* Київ: Вища школа, 1987. 63 с.
15. Kapanoglu M., Koc I. O., Erdogmus S. Genetic algorithms in parameter estimation for nonlinear regression models: an experimental approach. *Journal of Statistical Computation and Simulation*. 2007. Vol. 77, no. 10. P. 851–867. DOI: 10.1080/10629360600688244.
16. Mohan S. Parameter estimation of nonlinear Muskingum models using genetic algorithm. *Journal of hydraulic engineering*. 1997. Vol. 123, no. 2. P. 137–142. DOI: 10.1061/(asce)0733-9429(1997)123:2(137).
17. Настенко Е. А., Павлов В. А., Бойко А. Л., Носовец Е. К. Многокритериальный алгоритм шаговой регрессии. *Біомедична інженерія і технологія.* 2020. № 3. С. 48–53. DOI: 10.20535/2617-8974.2020.3.195661.
18. Öztürk O. B., Başar E. Multiple linear regression analysis and artificial neural networks based decision support system for energy efficiency in shipping. *Ocean Engineering*. 2022. Vol. 243. P. 110209. DOI: 10.1016/j.oceaneng.2021.110209.
19. Rajković D., Jeromela A. M., Pezo L., Lončar B., Grahovac N., Špika A. K. Artificial neural network and random forest regression models for modelling fatty acid and tocopherol content in oil of winter rapeseed. *Journal of Food Composition and Analysis*. 2023. Vol. 115, P. 105020. DOI: 10.1016/j.jfca.2022.105020.
20. Tam V. W. Y., Butera A., Le K. N., Da Silva L. C. F., Evangelista A. C. J. A prediction model for compressive strength of CO2 concrete using regression analysis and artificial neural networks. *Construction and Building Materials.* 2022. Vol. 324. P. 126689. DOI: 10.1016/j.conbuildmat.2022.126689
21. Худсон Д. *Статистика для физиков: Лекции по теории вероятностей и элементарной статистике*. Москва: Мир, 1970. 296 с.

References (transliterated)

1. Pavlov A. A., Holovchenko M. N., Drozd V. V. Efficiency substantiation for a synthetical method of constructing a multivariate polynomial regression given by a redundant representation. *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., 2023, no. 1 (9), P. 3–9. DOI: 10.20998/2079-0023.2023.01.01.
2. 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), P. 3–8. DOI: 10.20998/2079-0023.2022.02.01.
3. 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. *Lecture Notes on Data Engineering and Communications*

Technologies. 2023. Vol. 181. P. 207–222. DOI: 10.1007/978-3-031-36118-0_19.
4. Abdulrahman A. T., Alshammari N. S. Factor analysis and regression analysis to find out the influencing factors that led to the countries' debt crisis. *Advances and Applications in Statistics.* 2022, vol. 78, pp. 1–16. DOI: 10.17654/0972361722047.
5. Flitman A. M. Towards analysing student failures: neural networks compared with regression analysis and multiple discriminant analysis. *Computers and Operations Research.* 1997, vol. 24, no. 4, pp. 367–377. DOI: 10.1016/s0305-0548(96)00060-3.
6. Johnson R. A., Wichern D. W. *Applied multivariate statistical analysis*, 5th edn. Upper Saddle River, Prentice-Hall, 2002. 767 p.
7. Knowles D., Parts L., Glass D., Winn J. M. Modeling skin and ageing phenotypes using latent variable models in Infer.NET. *Predictive models in personalized medicine workshop, NIPS 2010*. Available at: https://www.researchgate.net/publication/241194775 (accessed 18.05.2024).
8. Lio W., Liu B. Uncertain maximum likelihood estimation with application to uncertain regression analysis. *Soft Computing*. 2020, vol. 24, no. 13, pp. 9351–9360. DOI: 10.1007/s00500-020-04951-3.
9. Liu S.S., Zhu Y. Simultaneous maximum likelihood estimation for piecewise linear instrumental variable models. *Entropy*. 2022, vol. 24, no. 9, pp. 1235. DOI: 10.3390/e24091235.
10. Ruff L., Vandermeulen R., Goernitz N., Deecke D., Siddiqui S. A., Binder A., Müller E., Kloft M. Deep one-class classification. *Proceedings of the 35th international conference on machine learning, PMLR 80*. 2018, pp. 4393–4402. Available at: http://proceedings.mlr.press/v80/ruff18a/ruff18a.pdf (accessed 18.05.2024).
11. Scott J. T. Factor analysis and regression. *Econometrica*. 1966, vol. 34, no. 3, pp. 552–562. DOI: 10.2307/1909769.
12. Buckley J. J., Feuring T. Linear and non-linear fuzzy regression: Evolutionary algorithm solutions. *Fuzzy Sets and Systems*. 2000, vol. 112, no. 3, pp. 381–394. DOI: 10.1016/s0165-0114(98)00154-7.
13. Draper N. R., Smith H. *Applied regression analysis*, 3rd edn. New York, Wiley & Sons, 1998. 736 p. DOI: 10.1002/9781118625590.
14. Ivakhnenko, A.G. *Modelirovanie slozhnykh sistem* [Complex systems modelling]. Kyiv, Vyshcha shkola Publ., 1987. 63 p.
15. Kapanoglu M., Koc I. O., Erdogmus S. Genetic algorithms in parameter estimation for nonlinear regression models: an experimental approach. *Journal of Statistical Computation and Simulation*. 2007, vol. 77, no. 10, pp. 851–867. DOI: 10.1080/10629360600688244.
16. Mohan S. Parameter estimation of nonlinear Muskingum models using genetic algorithm. *Journal of hydraulic engineering*. 1997, vol. 123, no. 2, pp. 137–142. DOI: 10.1061/(asce)0733-9429(1997)123:2(137).
17. Nastenko E., Pavlov V., Boyko G., Nosovets O. Mnogokriterial'nyy algoritm shagovoy regressii [Multicriteria stepwise regression algorithm]. *Biomedychna inzheneriya i tekhnolohiya*. 2020, no. 3, pp. 48–53. DOI: 10.20535/2617-8974.2020.3.195661
18. Öztürk O. B., Başar E. Multiple linear regression analysis and artificial neural networks based decision support system for energy efficiency in shipping. *Ocean Engineering*. 2022, vol. 243, p. 110209. DOI: 10.1016/j.oceaneng.2021.110209.
19. Rajković D., Jeromela A. M., Pezo L., Lončar B., Grahovac N., Špika A. K. Artificial neural network and random forest regression models for modelling fatty acid and tocopherol content in oil of winter rapeseed. *Journal of Food Composition and Analysis*. 2023, vol. 115, p. 105020. DOI: 10.1016/j.jfca.2022.105020.
20. Tam V. W. Y., Butera A., Le K. N., Da Silva L. C. F., Evangelista A. C. J. A prediction model for compressive strength of CO2 concrete using regression analysis and artificial neural networks. *Construction and Building Materials.* 2022, vol. 324, p. 126689. DOI: 10.1016/j.conbuildmat.2022.126689
21. 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.

Received 25.05.2024

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

АДАПТИВНИЙ МЕТОД ПОБУДОВИ БАГАТОВИМІРНОЇ РЕГРЕСІЇ

Запропоновано адаптивний метод побудови багатовимірної регресії, що задається зваженою лінійною згорткою відомих скалярних функцій від детермінованих вхідних змінних, коефіцієнти при яких є невідомими. Як, наприклад, коли багатовимірна регресія задається багатовимірним поліномом. На відміну від загальної процедури методу найменших квадратів, що мінімізує лише одну скалярну кількісну міру, адаптивний метод використовує шість різних кількісних мір і представляє собою системно зв'язану сукупність різних алгоритмів, що дозволяють кожну прикладну задачу розв'язувати на їх основі індивідуальним адаптивним алгоритмом, який в випадку активного експерименту навіть для порівняно невеликого об'єму експериментальних даних реалізує стратегію статистично обґрунтованого розв'язання. Невеликий об'єм даних активного експерименту використаний в тому сенсі, що для нього дисперсії оцінок невідомих коефіцієнтів, отриманих загальною процедурою метода найменших квадратів, не дозволяють гарантувати допустиму для практики точність. Пропонується також суттєво підвищити ефективність запропонованого Павловим О.А. та Головченко М.М. модифікованого методу групового урахування аргументів побудови багатовимірної регресії, лінійної відносно невідомих коефіцієнтів та заданої надлишковим описом, за рахунок включення в нього деяких критеріїв та алгоритмів адаптивного методу побудови багатовимірної регресії. Для випадку завдання регресії багатовимірним поліномом включення часткового випадку нової версії модифікованого методу групового урахування аргументів в синтетичний метод побудови багатовимірної поліноміальної регресії, заданої надлишковим описом, запропонованого Павловим О.А., Головченко М.М. та Дрозд В.В., також суттєво підвищує його ефективність.

Ключові слова: багатовимірна регресія, інтегральна міра, адаптивний алгоритм, регресійний аналіз, експертні коефіцієнти, лінійне програмування.

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