УДК 512.54

On Non-parametric Models of Multidimensional Non-inertial Processes with Dependent input Variables

Alexander V. Medvedev[*]
Siberian State Aerospace University,
Krasnoyarsky Rabochy, 31, Krasnoyarsk, 660014
Russia

Ekaterina A. Chzhan[†]
Institute of Information and Space Technology,
Siberian Federal University,
Svobodny, 79, Krasnoyarsk, 660041
Russia

Received 08.11.2016, received in revised form 12.03.2017, accepted 20.07.2017

The problem of identification of multidimensional non-inertial systems with delay is considered. Components of the input vector are stochastically related, and this relationship is unknown a priori. Such processes have "tubular" structure in the space of the input and output variables. In this situation methods of identification theory of non-inertial systems are not applicable. In general, it is not known a priori whether the process has "tubular" structure or not. To clear up this question the problem of estimation of the volume of a subdomain where "tubular" process takes place is considered. The initial data for this problem follows from the measurement of input-output variables. An algorithm for estimating the volume of the "tubular" subdomain in relation to the volume of the investigated process is suggested. The volume of the investigated process is always known from a priori information or production schedules. Numerical experiments are carried out with the use of the method of statistical modeling. They show high effectiveness of the proposed algorithm.

Keywords: non-parametric modeling, non-inertial processes with delay, indicator function, H-process.
DOI: 10.17516/1997-1397-2017-10-4-514-521.

Introduction

One of the key factors of identification processes in various sectors of human activity (economy, production) is the use of a priori information about the process under investigation. An appropriate sample of observations of input-output variables can be obtained on site in experiment. In various practical problems, these variables can be stochastically related. The nature of this relationship is often unknown a priori. Ragnar Firsh drew attention to this fact in creating economic models [1]. He introduced the term multicollinearity – stochastic relationships between input variables. The close linear correlation between input variables leads to the loss in accuracy of coefficients of the estimated model or even makes impossible to obtain estimates [2]. This phenomenon is typical of many industries. Thus, the correlation between the world financial indicators was found [3]. The authors suggested to use in predictive models only such variables that are not linearly related. A linear model of net profit based on the actual data of financial statements of the "Svyaz" company was obtained using input variables that are not linearly

[*] medvedev@sibgau.ru
[†] ekach@list.ru

related [4]. The phenomenon of multicollinearity is typical of processes in genetics [5] and ecology [6]. There are parametric linear models that are traditionally applied to such processes. We consider the situation when there are stochastic non-linear relationships between input variables. We propose models of "tubular" process.

We consider dynamic processes. In practice, input variables are often measured at sufficiently small intervals Δt, for example, with electrical sensors (current, frequency, temperature, humidity, etc.). Some output variables can be measured at a substantially longer time interval ΔT, $\Delta T >> \Delta t$ (chemical analysis, physical and mechanical testing, etc.). Thus, the duration of the investigated process may be considerably less than the interval ΔT. In this case, the main idea is to treat such channel as non-inertial with delay and formulate appropriate problem of identification and control.

1. Problem statement

General scheme of the identification process is shown in Fig. 1. The input vector $\boldsymbol{u}(t) = (u_1(t), u_2(t), \ldots, u_m(t)) \in \Omega(\boldsymbol{u}) \subset \mathbb{R}^m$ has dimension m, the output variable vector $\boldsymbol{x}(t) = (x_1(t), x_2(t), \ldots, x_n(t)) \in \Omega(\boldsymbol{x}) \subset \mathbb{R}^n$ has dimension n. For simplicity, let us consider the case of scalar output variable $x(t)$. System response channels $G^{u_1}, G^{u_2}, \ldots, G^{u_m}, G^x$ correspond to input and output variables, and they include control tools. Random error of variable measurements has zero mean value and bounded dispersion. The object can be described as follows:

$$x(t+\tau) = A(\boldsymbol{u}(t)) + \xi(t), \tag{1}$$

where A is unknown object operator, τ is the value of delay, $\xi(t)$ is random disturbance with zero mean value and bounded dispersion.

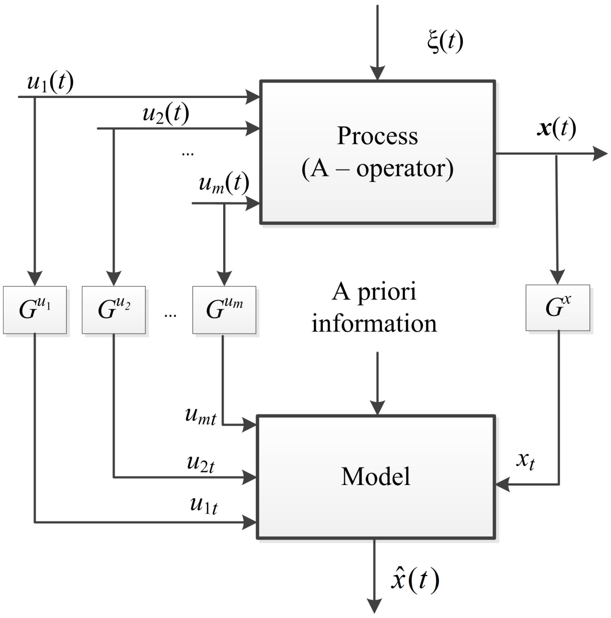

Fig. 1. General scheme of the identification process

The input and output variables are continuous because of the nature of the process but measurements are carried out at discrete times Δt due to control tools. There is an initial sample of observations $\{\boldsymbol{u}_i, x_i, i = 1, 2, \ldots, s\}$, where s is the sample size.

Current information about the process (sample of observations) as well as available a priori information are supplied to the unit "model", where $\hat{x}(t+\tau)$ is the model output. This block contains a certain class of models. Thus, it is necessary to formulate the model of the process.

The peculiarity of these processes is the presence of a stochastic relationship between input variables [7]. Such processes are called "tubular" or H-processes. Conventional identification algorithms do not give satisfactory results in simulation of such processes so it is proposed to use the H-model.

2. "Tubular" processes

For reasons of simplicity and without the loss of generality, we consider the process with two input variables u_1, u_2 and one output variable x. Let us assume that values of input and output variables u_1, u_2 and output variable x are distributed in the range $[0, 1]$. The domain of each variable is the interval, so the process takes place in the unit hypercube (Fig. 2). If there is a relationship between input variables:

$$u_1(t) = f(u_2(t)), \tag{2}$$

then the process proceeds along the line in the three-dimensional space. Thus, in Fig. 2 unit cube $\Omega(\boldsymbol{u}, x)$ is the domain of the process. The process observations belong not to the whole unit cube but only to the line which is located inside it. It should be noted that form (2) may be either linear or nonlinear.

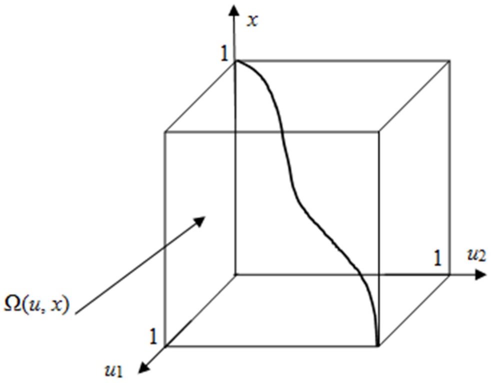

Fig. 2. Simple scheme of a process with functionally related input variables

The relationship between input variables can be stochastic:

$$u_1(t) = f(u_2(t)) + \mu(t), \tag{3}$$

where $\mu(t)$ is the random disturbance with zero mean value and bounded dispersion. Let us define the domain where "tubular" process proceeds as $\Omega^H(\boldsymbol{u}, x)$. Volume of this subdomain $\Omega^H(\boldsymbol{u}, x)$ is less then the volume of the hypercube $\Omega(\boldsymbol{u}, x)$ (Fig. 3).

Peculiarity of this process is that it proceeds not in the whole domain $\Omega(\boldsymbol{u}, x)$, but only in subdomain $\Omega^H(\boldsymbol{u}, x)$. This must be taken into account in solving the identification problem of the "tubular" process. Thus, the use of conventional parametric models [8–10] for identifying the "tubular" processes can lead to unsatisfactory results. To make a prediction we set values of input variables u_1 and u_2 which belong to the regulated domain $\Omega(\boldsymbol{u}, x)$ but they do not belong

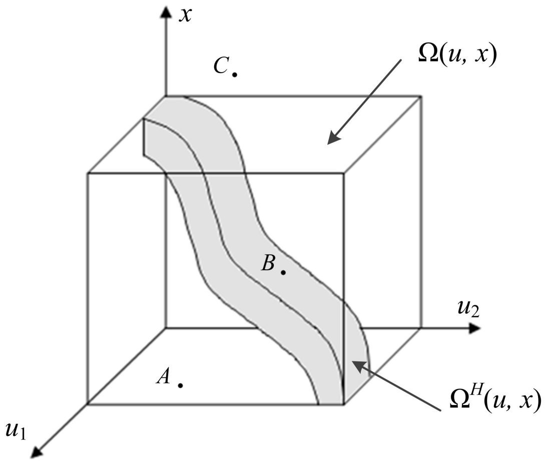

Fig. 3. Scheme of "tubular" process

to the "tubular" subdomain $\Omega^H(\boldsymbol{u}, x)$. The value of the output variable may not belong to the domain $\Omega(\boldsymbol{u}, x)$ (point C in Fig. 3). This value is easily eliminated because the boundary of $\Omega(\boldsymbol{u}, x)$ is always known. Another situation occurs when the value of the output variable x is in regulated area but it does not belong to $\Omega^H(\boldsymbol{u}, x)$ (point A). In this case, it is problematic to eliminate this value. Only point B belongs to the domain of the "tubular" process, i.e., $(\boldsymbol{u}_B, x_B) \in \Omega^H(\boldsymbol{u}, x) \subset \Omega(\boldsymbol{u}, x)$.

2. Model of "tubular" process

Let us consider the use of conventional parametric models for identification of stochastic process when input variables are related. In particular, let H-process be defined with the linear equation in the three-dimensional space. The schedule of the process has the form of line in the absence of noise. The conventional parametric model for this process has the form

$$\hat{x}(t+\tau) = A^{\alpha}(\boldsymbol{u}(t), \boldsymbol{\alpha}), \tag{4}$$

where A^{α} is the selected class of parametric models, $\boldsymbol{\alpha}$ is the parameter vector.

If we use several samples of observations we get different values of the estimated coefficients $\boldsymbol{\alpha}$. Also, every model has the form of plane. Thus, estimations that determine the position of the plane in the space of input and output variables can vary significantly depending on the particular sample. It is obvious that such model can not adequately describe investigated processes.

It is proposed to supplement the conventional parametric models with indicator function $I(\boldsymbol{u})$. Then model (4) can be reworked as follows:

$$\hat{x}(t+\tau) = A^{\alpha}(\boldsymbol{u}(t), \boldsymbol{\alpha}) I_s(\boldsymbol{u}), \tag{5}$$

where indicator $I_s(\boldsymbol{u})$ can be taken in the following form:

$$I_s(\boldsymbol{u}) = \begin{cases} 1, \text{ if } \sum\limits_{i=1}^{s} \prod\limits_{j=1}^{m} \Phi\left(c_s^{-1}\left(u^j - u_i^j\right)\right) \neq 0, \\ 0, \text{ if } \sum\limits_{i=1}^{s} \prod\limits_{j=1}^{m} \Phi\left(c_s^{-1}\left(u^j - u_i^j\right)\right) = 0. \end{cases} \tag{6}$$

The smoothing parameter c_s is defined as a solution of minimization problem for the quadratic criterion which shows the equivalence between object and model outputs compliance. Solution of minimization problem is based on the method of "sliding examination" [10]. Parameter c_s and bell function $\Phi\left(c_s^{-1}\left(u^j - u_i^j\right)\right)$ satisfy the convergence conditions [11].

The initial sample of observations $\{\boldsymbol{u}_i, x_i, i = 1, 2, \ldots, s\}$ is obtained by measuring the input and output variables. It is used in the calculation of parameters $\boldsymbol{\alpha}$ in model (5). The initial sample acts also as a learning sample when we calculate the estimation of indicator function (6). If input variables are related then this relationship is contained in the initial sample, that is, all sampling points belong to the "tubular" domain. So, if we have to deal only with real data obtained from the object then the estimation of the indicator function (6) of output model (5) is equal to one. If we use model (5) with the value $\boldsymbol{u}' \in \Omega(\boldsymbol{u}, x)$ that does not belong to the field of "tubular" process then the indicator function is equal to zero. This indicates that the process at this value $\boldsymbol{u}'$ does not exist. If there is no relationship between input variables then model (5) coincides with the standard parametric model (4).

3. Volume estimation of "tubular" process domain

It is unknown a priori whether the process has "tubular" structure or not. Restoration of the relationship between input variables is a complex and time-consuming process, especially if the vector of input variables $\boldsymbol{u}(t)$ has high dimension. It is proposed to estimate the volume of "tubular" process domain $\Omega^H(\boldsymbol{u}, x)$ using the following algorithm.

Algorithm.

Step 1. Generate initial learning sample $\{\boldsymbol{u}_i, x_i, i = 1, 2, \ldots, s\}$. In practice, we measure input-output variables and use these observations as a learning sample.

Step 2. For every variable $u_j, j = 1, \ldots, m$ find the minimal $\underline{u}_j$ and the maximum $\overline{u}_j$ values: $u_j \in [\underline{u}_j, \overline{u}_j], j = 1, \ldots, m$.

Step 3. Generate test sample $\{\boldsymbol{u}'_i, 1 = 1, \ldots, s'\}$ in the interval $[\underline{\boldsymbol{u}}, \overline{\boldsymbol{u}}]$.

Step 4. Define sampling points $\{\boldsymbol{u}'_i, 1 = 1, \ldots, s'\}$ that belong to the "tubular" subdomain. To do this, calculate the estimation of the indicator function (6).

Step 5. Then find the ratio of number of sample points s_1 that belong to "tubular" subdomain (the indicator function for such elements is equal to 1) to the total size of the test sample s':

$$v = \frac{s_1}{s'}. \tag{7}$$

Accordingly, the stronger is the relationship between input variables, the smaller is the value of v. If the process is not "tubular", i.e., all input variables are independent with each other then the value of v is close to 1.

4. Computer experiment

We carry out series of simulations of the "tubular" process. Let us assume that the object is described with the following equation:

$$x(t + \tau) = 2u_1^2(t) + \sin u_2(t) - 0.5u_3^2(t) + u_4(t) - 0.3u_5^2(t). \tag{8}$$

This equation simulates the behavior of some real process. The initial sample is $\{x_{i+\tau}, \boldsymbol{u}_i, i = 1, 2, \ldots, s\}$, where the value of τ is a multiple of discreteness Δt. Later a shift in the output variable x in the observation matrix of input and output variables is introduced so the delay in subsequent expressions is omitted. The sample is $\{x_i, \boldsymbol{u}_i,\ i = 1, 2, \dots, s\}$. Equation (8) is not known. When measuring the output variable x random noise is introduced as

$$\xi(t) = a\zeta(t), \tag{9}$$

where $\zeta(t)$ is a random variable uniformly distributed on the interval [–1, 1], a is the interference value. For example, if the interference is 10% then $a = 0.1$.

The investigated object has "tubular" structure due to the relationship between input variables. H-models describe such objects. Any a priori information on the form of the relationship between input variables is not available. The relationships between input variables are described as follows:

$$\begin{cases} u_1(t) \in [0, 3], \\ u_2(t) = u_1(t) + \mu_1(t), \\ u_3(t) = \sin(u_1(t) + u_2(t)) + \mu_2(t), \\ u_4(t) = 0.3u_1(t)u_2(t) + \mu_3(t), \\ u_5(t) = u_1(t) - u_4(t) + \mu_4(t), \end{cases} \tag{10}$$

here variable $u_1(t)$ is the random number uniformly distributed on the interval [0, 3], $\mu_i(t)$, $i = 1, \dots, 4$ are random values generated according to the following formula:

$$\mu(t) = b\varsigma(t), \tag{11}$$

where $\varsigma(t)$ is the random number uniformly distributed on the interval [–1, 1], b is the value of the interference. Let us note again that the form of equation (8) and system (10) is not known. System (10) is needed to construct a model of the object based on observations of input and output variables. First, let us consider the traditional way of identification. Taking into account the identification theory, we assume the following parametric model for object (8) using a priori information [8]:

$$\hat{x}(t + \tau) = \alpha_1 u_1^2(t) + \alpha_2 \sin u_2(t) + \alpha_3 u_3^2(t) + \alpha_4 u_4(t) + \alpha_5 u_5^2(t). \tag{12}$$

where $\alpha_i, i = 1, \dots, 5$ are unknown parameters.

Let us generate sample $\{u_{1i}, u_{2i}, u_{3i}, u_{4i}, u_{5i}, x_i, i = 1, \dots, s\}$ and estimate coefficients $\alpha_i, i = 1, \dots, 5$ of model (13), using the least squares method [8] with various values of interference a, b and the sample size s. Results are presented in the Tab. 1.

There is a small refinement of parameters of model (13) with the growth of sample size. The accuracy of the simulation at $s = 500,\ a = 0,5,\ b = 0,5$ is 0.07. However, stochastic relationship between input variables is not included in model (13). Consequently, this parametric model can not be used for prediction because the process does not exist in the regulated area. It is necessary to modify the parametric model with the indicator function:

$$\hat{x}(t + \tau) = \left(\alpha_1 u_1^2(t) + \alpha_2 \sin u_2(t) + \alpha_3 u_3^2(t) + \alpha_4 u_4(t) + \alpha_5 u_5^2(t)\right) I_s(\boldsymbol{u}(t)). \tag{13}$$

where $I_s(\boldsymbol{u})$ is the indicator function (6).

We present the results of estimation of the volume of "tubular" process subdomain (Tab. 2). The sample $\{\boldsymbol{u}', i = 1, 2, \dots, s'\}$ is generated, using the proposed algorithm.

About 2% of elements of test sample belong to "tubular" subdomain in the case of 5% noise level and 3% of elements of test sample belong to "tubular" subdomain in the case of 10%. The results indicate that the region of the "tubular" process is much less than the regulated area. This means that the investigated process has "tubular" structure.

Table 1. Coefficients estimations of the model (13)

s	a	b	α_1	α_2	α_3	α_4	α_5
500	0.5	0.5	2.01	2.99	–0.49	0.98	–0.24
1000	0.5	0.5	1.99	3	-0.5	1.01	–0.3
500	0.1	0.1	1.98	3	–0.51	1.09	–0.23
1000	0.1	0.1	2	3	–0.49	0.99	–0.21

Table 2. Volume estimation of "tubular" subdomain

s'	a	b	s_1	v
500	0.5	0.5	11	0.022
1000	0.5	0.5	24	0.024
500	0.1	0.1	16	0.032
1000	0.1	0.1	30	0.03

Conclusion

An algorithm for the identification of "tubular" processes with stochastic relationships between input variables is proposed. The form of these relationships is not known a priori. It is shown that the dynamical system with significant discrete control of output variable should be treated as non-inertial with delay. In this case, conventional models of the identification theory can not be used. The introduction of appropriate indicators is required. The H-model with stochastically independent input variables coincides with well-known models.

The problem of volume estimation of the region $\Omega^H(u, x)$ is also discussed. The method of calculating this volume is based on the Monte-Carlo method. This algorithm with the existing initial learning sample allows us to find out the presence or absence of a "tubular" structure of the process. Some numerical results of the implementation of proposed algorithms are presented.

References

[1] K.J.Arrow, The work of Ragnar Frisch, econometrician, *Econometrica: Journal of the Econometric Society*, **8**(1960), no. 2, 175–192.

[2] S.A.Ayvazyan, I.S.Enukov, L.D.Meshalkin, Applied Statistics: Basics of modeling and primary data processing, M., Finansy i Statistika, 1983 (in Russian).

[3] A.V.Koltyshev, The methods of forecasting the financial condition of the oil and gas company, Problems of Geology and Mineral Resources Development: Proceedings of the XIX International Symposium of Academician M.A.Usov, **2**(2015), Tomsk, 664–670 (in Russian).

[4] I.V.Orlova, E.S.Filonova, Selection of exogenous factors in the regression model with data multicollinearity, *The International Journal of Applied and Basic Research*, **5**(2015), 108–116 (in Russian).

[5] J.G.Prunier et al., Multicollinearity in spatial genetics: separating the wheat from the chaff using commonality analyses, *Molecular ecology*, **24**(2015), 263–283.

[6] S.F.Spear, N.Balkenhol, M.J.Fortin, B.H.McRae, K.Scribner, Use of resistance surfaces for landscape genetic studies: considerations for parameterization and analysis, *Molecular Ecology*, **19**(2010), 3576–3591.

[7] A.V.Medvedev, E.D.Mihov, O.V.Nepomnyashchiy, Mathematical Modeling of H-processes, *Journal of Siberian Federal University. Mathematics & Physics*, 9(2016), no. 3, 338–346.

[8] Ya.Z.Tcypkin, Foundation of theory identification, M., Nauka, 1984 (in Russian).

[9] A.Fournier, D.Fussell, L.Carpenter, Computer rendering of stochastic models, *Communications of the ACM*, **25**(1982), no. 6, 371–384.

[10] B.Peeters, G. De Roeck, Stochastic system identification for operational modal analysis: a review, *Journal of Dynamic Systems, Measurement and Control*, **123**(2001), 659–667.

[11] E.A.Nadaraya, On estimating regression, *Theory of Probability and its Applications*, **9**(1964), 141–142.

О непараметрических моделях безынерционных многомерных процессов с зависимыми входными переменными

Александр В. Медведев
Сибирский государственный аэрокосмический университет
Красноярский рабочий, 31, Красноярск, 660014
Россия

Екатерина А. Чжан
Институт космических и информационных технологий
Сибирский федеральный университет
Свободный, 79, Красноярск, 660041
Россия

Рассматривается задача идентификации многомерных безынерционных систем с запаздыванием при стохастической зависимости компонент вектора входных воздействий, причем характер этой зависимости априори неизвестен. Подобные процессы имеют «трубчатую» структуру в пространстве входных-выходных переменных. Методы теории идентификации для построения моделей безынерционных систем оказываются неприменимыми. Вообще априори неизвестно, является ли интересующий нас процесс "трубчатым". Для анализа этого обстоятельства специально рассматривается задача вычисления объема подобласти, в которой протекает "трубчатый" процесс. Исходными данными являются результаты наблюдений входных-выходных переменных. Приведен алгоритм вычисления объема этой подобласти по отношению к объему исследуемого процесса, который всегда известен из априорных сведений или технологического регламента. Проведены объемные численные исследования средствами метода статистического моделирования, которые свидетельствуют о достаточно высокой эффективности предложенных моделей.

Ключевые слова: непараметрическое моделирование, безынерционный объект с запаздыванием, индикаторная функция, H-процесс.