Nonparametric modelling of multidimensional technological processes with dependent variables

E A Chzhan[1]**, A V Medvedev**[1]**, V V Kukartsev**[1,2]

[1] Siberian Federal University, 79, Svobodny, Krasnoyarsk, 660041, Russia
[2] Reshetnev Siberian State University of Science and Technology, 31, Krasnoyarskiy rabochiy, Krasnoyarsk 660037, Russia

E-mail: ekach@list.ru

Abstract. The article deals with the features of simulation of multidimensional discrete-continuous processes with delay, the input variables of which are connected by stochastic dependence. These processes are typical for metallurgy and the mining industry. The type of dependence is unknown to the researcher. A series of computational experiments on the construction of nonparametric models in the conditions of incompleteness of a priori information, small samples and in the presence of interference shows the effectiveness of proposed nonparametric identification algorithm.

1. Introduction

Discrete-continuous processes with delay are common in various industries (metallurgy, petrochemistry, construction industry) [1 – 3]. Technological processes proceed in time continuously, but measurements of input and output characteristics of the process occur discretely. With discreteness of control, the following feature is associated, which must be taken into account when solving the identification problem. Input and output variables can be measured at different time intervals. For example, if the values of some characteristics can be obtained with electrical controls (temperature), then the discreteness can be arbitrarily small. We can get the values of other variables only after chemical or physical-mechanical analysis. An example of such characteristics is cement activity. Activity or compressive strength determines cement brand andб therefore, its cost. The activity value according to GOST 310.4-81 can be obtained only after 28 days. So, there is no way to use this indicator for control purposes. In some cases, the discreteness of the control significantly exceeds the time constant of the object, so the dynamic processes must be considered as static with delay.

Often in practice, when investigating and modelling real processes, the researcher encounters a situation when there is a relationship between the input variables of the process [4]. The form of this dependence remains unknown. The identification and restoration of these dependencies are a complex and time-consuming process. This article is devoted to modelling this kind of processes.

2. The problem statement

Processes with stochastic dependence of the input vector components are called "tubular" or H-processes. The general scheme of the "tubular" process identification is shown in figure 1.

In figure 1 the following notation is accepted: A is an unknown operator, $x(t) \in \Omega(x) \subset R^1$ is an output variable of the process; $u(t) = \left(u_1(t), u_2(t), ..., u_m(t) \right)$; $u(t) \in \Omega(u) \subset R^m$ is vector input action;

$\xi(t)$ is vector random noise; (t) is continuous time; $G^{u_i}, i = \overline{1,m}$, G^x are communication channels corresponding to different variables, including control tools. In the communication channels of input and output variables random, there are random noises with zero mathematical expectations and bounded variances.
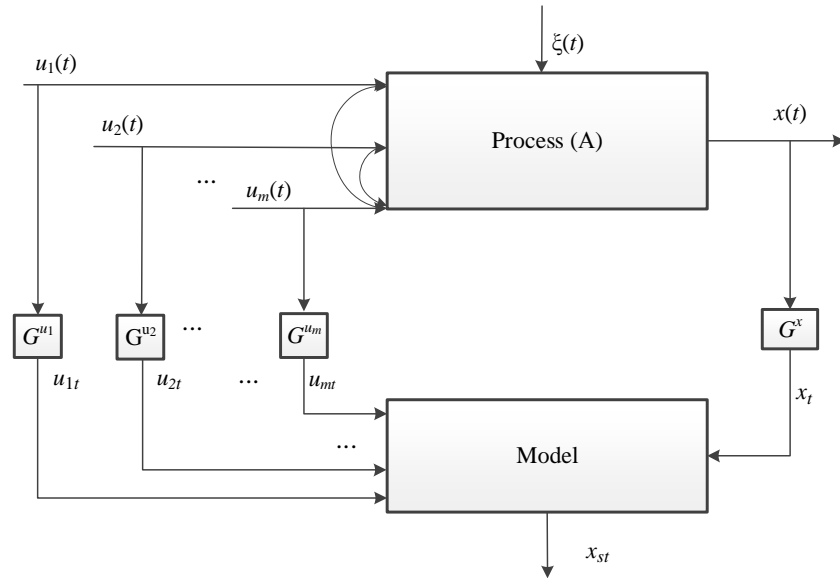

Figure 1. The general identification scheme of a "tubular" process

The task of identification is to restore the relations between the input and output variables of the process. As already mentioned above, the distinctive feature of H-processes is that there are not only relations between the input and output variables of the process, but the input variables are stochastically related. Let us turn to Fig. 1, where arc-shaped arrows indicate a possible version of the relationship of the input vector components. The figure shows the possible version of the dependence, in each case the position of the arrows can be different.

3. Nonparametric identification

In the conditions of incompleteness of a priori information, when it is not possible to determine the parametric structure of the object under investigation up to a parameter vector, it is appropriate to use nonparametric statistics tools [5]. The nonparametric Nadaraya-Watson estimation of the regression function on observations was used [5]:

$$x_s(u) = \frac{\sum\limits_{i=1}^{s} x_i \prod\limits_{j=1}^{m} \Phi\left(c_s^{-1}\left(u^j - u_i^j\right)\right)}{\sum\limits_{i=1}^{s} \prod\limits_{j=1}^{m} \Phi\left(c_s^{-1}\left(u^j - u_i^j\right)\right)}, \tag{1}$$

where the bell-shaped function $\Phi\left(c_s^{-1}\left(u^j - u_i^j\right)\right), i = \overline{1,s}, j = \overline{1,m}$ and the kernel blur factor c_s satisfy the convergence conditions [5]:

$$c_s > 0; \qquad \lim_{s \to \infty} c_s = 0;$$

$$\Phi(c_s^{-1}(u - u_i)) \geq 0; \qquad \int_{\Omega(u)} \Phi(c_s^{-1}(u - u_i))du < \infty; \qquad (2)$$

$$\lim_{s \to \infty} s c_s = \infty; \qquad \lim_{s \to \infty} c_s^{-1}\Phi(c_s^{-1}(u - u_i)) = \delta(u - u_i).$$

In the computational experiments, a parabolic function was used:

$$\Phi\left(c_s^{-1}\left(u^j - u_i^j\right)\right) = \begin{cases} 0.75\left(1 - \left(c_s^{-1}\left(u^j - u_i^j\right)\right)^2\right), \text{if } \left|c_s^{-1}\left(u^j - u_i^j\right)\right| < 1, \\ 0, \text{if } \left|c_s^{-1}\left(u^j - u_i^j\right)\right| \geq 1. \end{cases} \qquad (3)$$

The accuracy of the simulation is estimated from the relative error of approximation:

$$W = \left(\frac{s^{-1}\sum_{i=1}^{s}(x_{si} - x_i)^2}{(s-1)^{-1}\sum_{i=1}^{s}(x_i - \hat{m}_x)^2} \right)^{1/2}. \qquad (4)$$

The error (3) is normalized and its values belong to the interval [0, 1]. The error (3) can be interpreted as follows: the smaller the value (3), the more accurate the model (2) describes the object under consideration.

4. Computational experiment

The nonparametric estimate (1) can be referred to the class of local approximation. The H-process does not occur in the entire regulated area, but only in some of its similarity, because of the dependence between the input variables of the process. In this case, we can use data-based approach [5]. In contrast to the parametric approach, when the model is a certain surface in a multidimensional feature space, the restoration occurs only in those areas where the process actually takes place [4].

Below are the results of modelling a multidimensional "tubular" process, processes of this type are often encountered in practice [6]. In the framework of the computational experiment, we consider various types of dependencies between input variables. Let the object be described by the following equation:

$$x(u) = 2u_1u_2 - u_1^2 + 1.5u_3u_2^2 + 2u_3^3u_4 - u_5u_6^3 + u_7^2u_8^3 - 2u_9u_{10}^4 + 1.5u_9 + \xi, \qquad (5)$$

where the input variables $u_i, i = \overline{1,10}$ are uniformly distributed in the interval of [0; 3], ξ is the uniformly distributed interference:

$$\xi = k\sigma x(u), \qquad (6)$$

where k is the level of interference, σ is a random variable distributed according to a uniform law in the interval [-1; 1], $x(u)$ is the output of the object.

The object under consideration is "tubular", i.e. its components of the vector of the input variable are related by some dependence. In practice, this dependence remains unknown, but for the experiments we specify the form of the dependence. It is worthwhile to note that the form of the dependence is given only for generation of samples of observations and is subsequently considered to be unknown. Let the input variables be related by the following relations (table 1). In table 1, u_1 is an independent variable whose values are uniformly distributed in the interval of [0; 3]. Values are generated using the built-in pseudorandom number sensor on the .Net platform (C # language). All other variables $u_2 - u_{10}$ are interconnected.

Table 1. The type of dependence between the input variables of the object (4)

The variable	Functional relationship
u_1	$[0;3]$
u_2	$\sqrt{u_1}$
u_3	$\sin u_1$
u_4	$u_2 + u_3$
u_5	u_4
u_6	$0.4u_1^2$
u_7	$u_4 + u_6$
u_8	$\sqrt{u_4}$
u_9	$u_4^{1.5}$
u_{10}	$u_1 + u_9$

For example, a tubular object may have the following description:

$$\begin{cases} x = f(u_1, u_2, ..., u_{10}), \\ u_2 = f(u_1), \\ u_3 = f(u_1), \\ \dots \\ u_{10} = f(u_1, u_9). \end{cases} \qquad (7)$$

In a series of computational experiments, we increased the number of independent variables. In the first experiment, all variables except u_1 were functionally related, in particular, $u_2 = f(u_1)$, $u_3 = f(u_1)$, ..., $u_{10} = f(u_1)$. In the second experiment, u_1 and u_2 were independent, and all others depended on them, i.e. the values of the variables u_1 and u_2 were uniformly distributed in the interval of $[0; 3]$, and $u_3 = f(u_1)$, $u_4 = f(u_1, u_2)$, ..., $u_{10} = f(u_1, u_2)$. In subsequent experiments, the number of independent variables increased and, finally, in 10 experiments all variables were not dependent on each other. In each experiment, a sample of observations $\left\{ u_i, x_i, i = \overline{1,1000} \right\}$ was generated.

Table 2. Results of H-process modelling

	Independent variables	Error of approximation W
1	u_1	0.034
2	u_1, u_2	0.054
3	u_1, u_2, u_3	0.109
4	u_1, u_2, u_3, u_4	0.120
5	u_1, u_2, u_3, u_4, u_5	0.136
6	$u_1, u_2, u_3, u_4, u_5, u_6$	0.183
7	$u_1, u_2, u_3, u_4, u_5, u_6, u_7$	0.202
8	$u_1, u_2, u_3, u_4, u_5, u_6, u_7, u_8$	0.212
9	$u_1, u_2, u_3, u_4, u_5, u_6, u_7, u_8, u_9$	0.371
10	$u_1, u_2, u_3, u_4, u_5, u_6, u_7, u_8, u_9, u_{10}$	0.621

As the number of independent variables increases, the approximation error grows. With the increase in the number of independent variables from 1 to 10, the error increased almost 20-fold. With 10 independent variables, much larger sample sizes are required. In the conditions of small samples, it is possible to obtain an adequate model only if the process is tubular.

5. Conclusion

The results of modelling a multidimensional "tubular" process are presented in the article. It can be concluded that as the interference grows, the quality of the model decreases slightly with the same sample size. With a small sample size and a large number of variables describing the process, the exact model can be obtained only if there is a strong dependence between the input variables. As the number of dependent variables increases, the accuracy of the simulation increases.

6. References

[1]   Lucia Figuli *et al* 2016 *IOP Conf. Ser.: Earth Environ. Sci.* **44**
[2]   Nadia Rahmah and Imas Sukaesih Sitanggang 2016 *IOP Conf. Ser.: Earth Environ. Sci.* **31**
[3]   Medvedev A V, Kornet M E, Chzhan E A 2016 *Steel in Translation* **46(12)** 855–859
[4]   Medvedev A V, Chzhan E A 2017 *Bulletin of the South Ural State University. Series: Mathematical modeling and programming* **10(2)** 124-136
[5]   Härdle W 1990 *Applied nonparametric regression* (Cambridge university press)
[6]   Chzhan E A 2017 *Applied Methods of Statistical Analysis. Nonparametric Methods in Cybernetics and System Analysis* 82-87