MSC 00X00

About nonparametric modeling of multidimensional noninertial systems with delay

A.V. Medvedev, Siberian Federal University, Krasnoyarsk, Russian Federation**,**
E.A. Chzhan, Siberian Federal University, Krasnoyarsk, Russian Federation, ekach@list.ru

We consider the problem of noninertial objects identification under nonparametric uncertainty when prior knowledge of parametric structure of the object is not available. In many applications there is a situation, when measurements of various output variables are made through significant period of time and it can substantially exceed the constant object of time. In this context, we must consider the object as the noninertial with delay. In fact, there are two basic approaches to solve the problems of identification: one of them is identification in "narrow" sense or parametric identification. However, it is natural to apply the local approximation methods when we have not enough prior knowledge to select the parameter structure. These methods deal with qualitative properties of the object. If the source data of the object sufficiently representative, the nonparametric identification gives a satisfactory result but if there are "sparsity" or "gaps" in the space of input and output variables the quality of nonparametric models is significantly reduced. This article is devoted to the method of filling or generation of training samples based on current available information. This can significantly improve the accuracy of identification of nonparametric models of noninertial systems with delay. Conducted computing experiments have confirmed that the quality of nonparametric models of noninertial systems can be significantly improved as a result of original sample "repair". At the same time it helps to increase the accuracy of the model at the border areas of the process input-output variables definition.

Keywords: nonparametric identification, data analysis, computational modeling

Introduction

Simulation of discrete-continuous systems is of considerable interest due to the practice prevalence of situations when some components of the output variables are measured through significant periods of time which are substantially bigger then the time constant of the object. For example, transient process of dynamic object can be completed in 20 minutes, but output variable measurements are carried out after 2 hours. Here we consider the problem of identification of multidimensional static systems with delay under nonparametric uncertainty, when the model parametric structure of the process is not known. In other words, the a priori information about the process under study is not enough to more or less objectively define the model of the process within the parameter vector. In this case, the identification of the problem can be considered in the framework of nonparametric system [1]. It should be noted that further used nonparametric Nadaraya-Watson estimation of regression function refers to the category of local approximation methods as opposed to parametric methods.

In nonparametric identification of multidimensional static objects with delay quality of the resulting model depends heavily on the initial data. Sample of observations of the input and output variables can have a number of disadvantages [2]. They can be of different nature, come out of measurement error, the functioning of the investigated process and

various control discreteness of input and output variables. Here we consider the problem of identification of stochastic systems when the sample of observations contains "sparsity" in the regulated area of the process.

Note that for solving the problem of identification within the parametric approach, the problem is not so acute. But in nonparametric identification it requires special attention. In this case, in case of nonparametric identification the situation can arise when the forecast of output variable is inaccurate or even can not be calculated because of an uncertainty type [0/0]. This is typical of the nonparametric regression function estimation on observations that is used to solve this problem. To some extent, this is "payment" for the absence of the stage of parametric structure definition of the investigated process model. The selection of parametric structure within the parameter vector is quite difficult task and it is required significant research efforts.

The distribution of the observations sample in the space of input and output variables plays an important role in nonparametric estimation. Often there is a need to supplement the initial learning sample in order to eliminate the "sparsity" in certain subregions of the investigated process. The following we discuss methods, techniques of supplemeny the initial learning sample, which, ultimately, leads to models improvement of the object under nonparametric identification.

1. The Problem Statement

Consider a multi-dimensional static object with delay, its general scheme is shown in Fig. 1 [3, 4].

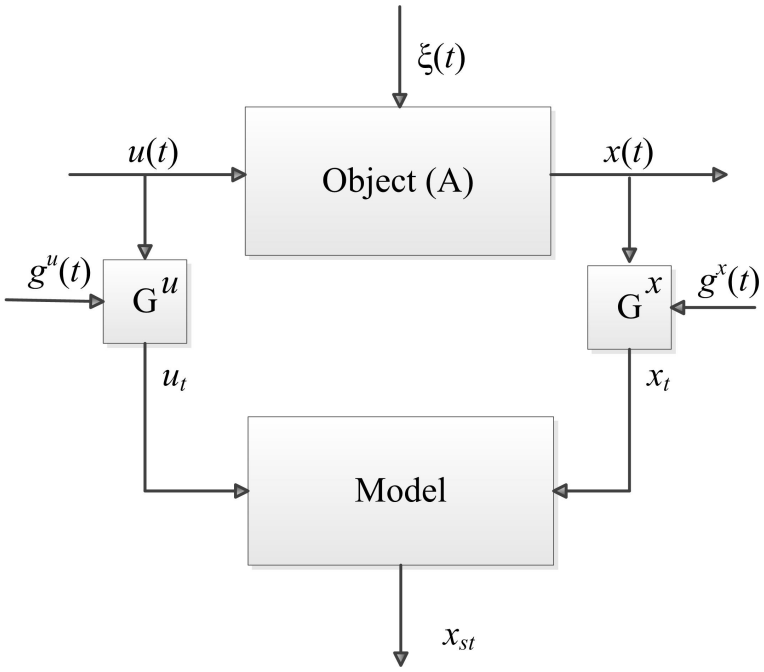

Рис. 1. The General Scheme of the Investigated Object

In Fig. 1 there is accepted the following notation: A is an unknown object operator, the input vector of the object $u(t) = (u_1(t), u_2(t), ..., u_m(t)) \in \Omega(u) \subset R^m$ has dimension

m, the output variable vector of the object $x(t) = (x_1(t), x_2(t), ..., x_n(t)) \in \Omega(x) \subset R^n$ has dimension n, (t) is continuous time, Δt is control discreteness of input and output variables of the process; $\xi(t)$ is random noise vector; G^u, G^x are control units of input and output variables with random noise $g^u(t), g^x(t)$ which have zero mathematical expectations and limited variances; u_t and x_t are measurement of variables $u(t)$ and $x(t)$ at discrete time. Thus, by measuring the values of input and output variables, we obtain a sample $\{u_i, x_i, i = \overline{1, s}\}$, where s is sample size, which is said to be the initial sample of observations.

It is clear that, if we conduct another experiment on the same object, we get different sample with another distribution in space of observations of input and output variables. In particular, the subregion with large amount of observations may be replaced by sparsity.

The content of the identification problem is to construct a non-parametric model of the investigated process based on available learning sample $\{u_i, x_i, i = 1, \ldots, s\}$ under conditions of incomplete information, when it is difficult to parameterize the model. The asymptotic properties of non-parametric estimations of regression functions have been studied in detail in [5]. Analysis of smoothing properties of nonparametric estimations of regression functions have been considered in sufficient detail in the monographs [6, 7]. Note that the sparsity of the sample in the space of input and output variables should not be confused with the gaps in the data. It is necessary to restore the sparsity areas due to the internal properties of nonparametric Nadaraya-Watson estimation. However, it should be clearly understood that generated elements included in the initial learning sample do not contain information about the object, because they are not actual data obtained on-site. We use these generated elements in computing nonparametric estimation to eliminate the uncertainty of type [0/0]. It should be noted that we supplement the sample of observations with new elements that are not measured on a real object, however, they are generated on the basis of the initial learning sample which reflects properties of the real object.

2. Parametric Identification

In the construction of models of discrete-continuous processes there dominates parametric identification or identification in "narrow" sense [3, 4]. The parametric identification of stochastic systems is based on two main phases: parameterization of the model and estimation of the parameter vector with the sample of observations of input and output variables of the process. In other words, in the first stage we select the parametric structure of the model, for example:

$$x_\alpha(t) = f(u(t - \tau), \alpha), \tag{1}$$

where f is a certain function, α is a parameter vector, τ is delay. At the second stage, we estimate the parameters α on the basis of the available sample $\{u_i, x_i, i = \overline{1, s}\}$. There are many methods and algorithms to get these estimations [3].

In this way, the main difficulty lies in the choice of a parametric structure of the object (1). This is the most difficult stage for researchers. Here it would be appropriate to recall the phrase of Democritus: "Even a slight deviation from the truth, in the future leads to infinite error".

3. Nonparametric Identification

In most cases, we have little information about the object and only a few qualitative properties of the investigated object, such as uniqueness or ambiguity, the linearity of the dynamic object or non-linearity and others. In this case, the a priori information is not sufficient to select the parametric structure of the object. It is proposed to use identification methods in a "broad" sense. On this occasion, professor N.S. Raybman in the preface to the book [4] mentions: "priori information about the object in the identification in a broad sense is absent or very poor, so we have to previously solve the large number of additional tasks. These tasks include: system parametric structure choices and model class assignment ...". As a model of the object we take the nonparametric estimation of the regression function Nadaraya-Watson [1, 8]:

$$x_s(u) = \frac{\sum_{i=1}^{s} x_i \prod_{j=1}^{m} \Phi(c_s^{-1}(u^j - u_i^j))}{\sum_{i=1}^{s} \prod_{j=1}^{m} \Phi(c_s^{-1}(u^j - u_i^j))}, \tag{2}$$

where bell-shaped function $\Phi(c_s^{-1}(u^j - u_i^j))$, $i = \overline{1,s}$ and smoothing parameter c_s satisfy the convergence conditions [4, 5].

Parameter c_s is determined by solving the problem of minimizing a quadratic criterion of difference between the object and model output, based on a sliding exam when in the model (2) i-th variable is excluded:

$$R(c_s) = \sum_{k=1}^{s} (x_k - x_s(u_k, c_s))^2 = \min_{c_s}, k \neq i. \tag{3}$$

Estimation $x_s(u)$ (2) based on sample $\{u_i, x_i, i = \overline{1,s}\}$ belongs to the class of local approximations. Note that the function $\Phi(c_s^{-1}(u^j - u_i^j))$, $i = \overline{1,s}$, $j = \overline{1,m}$ has the following property:

$$\Phi(c_s^{-1}(u^j - u_i^j)) = \begin{cases} > 0, \text{ if } c_s^{-1}|u^j - u_i^j| < \eta, \\ = 0, \text{ else,} \end{cases} \tag{4}$$

where $i = \overline{1,s}$, $j = \overline{1,m}$, η is a constant depending on the choice of a particular bell-shaped function $\Phi(c_s^{-1}(u^j - u_i^j))$, the argument is determined by the values of $(u^j - u_i^j)$ and the smoothing parameter c_s. Value of argument $(c_s^{-1}(u^j - u_i^j))$ of bell-shaped function with random value of u^j depends on the value of c_s. For example, if we select a triangular kernel as the bell-shaped function:

$$\Phi(c_s^{-1}(u^j - u_i^j)) = \begin{cases} 1 - c_s^{-1}|u^j - u_i^j|, \text{ if } c_s^{-1}|u^j - u_i^j| \leq 1, \\ 0, \text{ else,} \end{cases} \tag{5}$$

then $\eta = 1$. Below, we discuss c_s-neighborhood of the point $u = u^j, j = \overline{1,m}$ for fixed c_s. In the analysis of nonparametric estimation of the regression function from observations (2) may arise situations when none of the elements of the learning sample $\{u_i, x_i, i = \overline{1,s}\}$ belongs to c_s-neighborhood of $u = u^j, j = \overline{1,m}$, which lead, in view (4), to uncertainty (2) of the form [0/0].

Estimation $x_s(u')$ at the point $u' = (u'_1, u'_2, ..., u'_m)$ is restored on the basis of the sample elements that are in the c_s- neighborhood of the point u'. The obvious is the fact that

the accuracy of estimation depends on the number of items on which this estimation is computed. In case, if there are no elements of learning sample in c_s-neighborhood of the point u', it impossible to give the estimation. In this case there is a problem of uncertainty of the form [0/0]. One possible way to get an estimation is to increase the value of smoothing parameter c_s. In some cases, it helps to provide a forecast (to avoid uncertainty), but the forecast $x_s(u')$ can be inaccurate.

The accuracy of nonparametric estimation (2) depends on the sample of observations $\{u_i, x_i, i = \overline{1,s}\}$. In many practical problems, even for the same process under investigation sample $\{u_i, x_i, i = \overline{1,s}\}$ in different time intervals may differ significantly, which affect on the accuracy of forecast $x_s(u)$. Hence, there is a problem of generating the working learning sample based on the initial sample $\{u_i, x_i, i = \overline{1,s}\}$. In the initial learning sample there are sparsity and subregions with large amount of observations of the domain $\Omega(x,u)$ (Fig. 2).

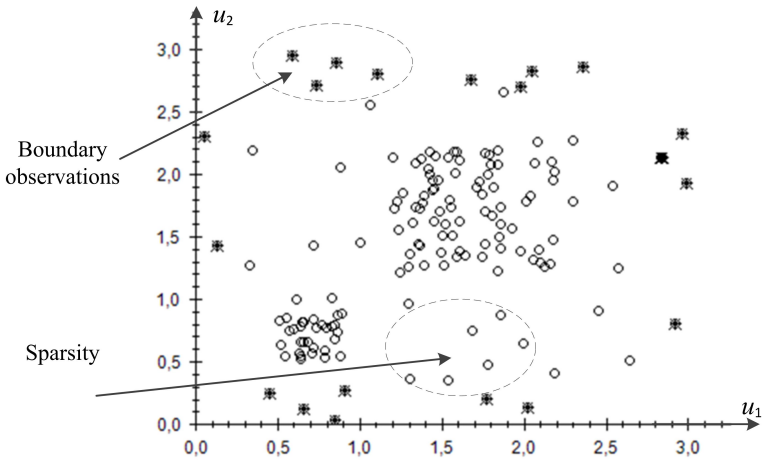

Рис. 2. Correlation Field of Input Variables for Initial Sample

In Figure 2 it is shown subdomains with sparsity and boundary points are marked with asterisks where Nadaraya-Watson nonparametric estimation (2) is inaccurate. Our task is to algorithmically convert the initial sample shown in Figure 2 into a working sample, for example, shown in Figure 3.

Note that although we are interested in the case of m-dimensional vectors $u \in \Omega(u) \subset R^m$, in the interests of simplicity of visualization two-dimensional vector $u \in \Omega(u) \subset R^2$ is presented in the figures.

4. Method of Working Sample Generation

The main idea of generation of a working sample $\{\tilde{u}_i, \tilde{x}_i, i = \overline{1,N}\}, N > s$ based on the initial $\{u_i, x_i, i = \overline{1,s}\}$ lies in the fact that the sparsity of the field are complemented by new sample items that are included in the working sample according to a particular algorithm.

The idea of sample generation is the following: we generate working samples based on initial observations using different methods. In particular, this idea is used in the bootstrap

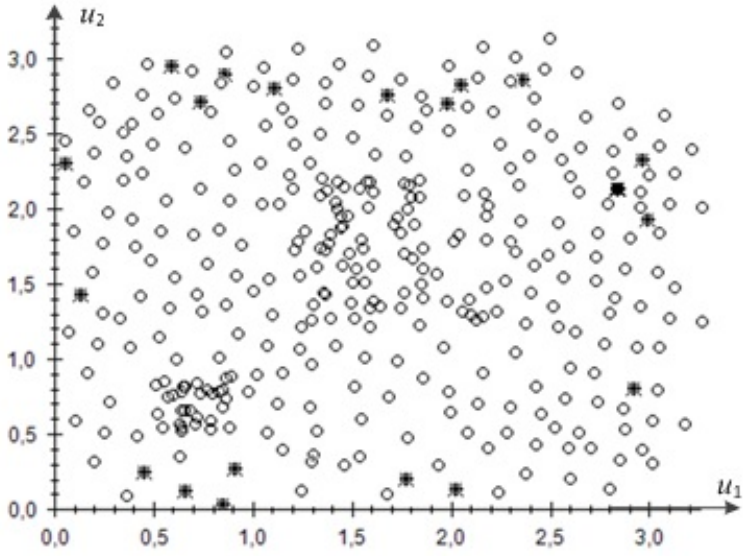

Рис. 3. Correlation Field of Input Variables for Working Sample

methods [9]. Bootstrap methods are widely used in statistical analysis in estimating the distribution parameters and hypothesis testing [10, 11].

Generally speaking, it should be noted that only the initial sample of observations of input and output variables $\{u_i, x_i, i = \overline{1,s}\}$ contains information on the investigated process. We need the newly created elements only to improve the efficiency of nonparametric models, because they are based on local approximation methods. It should be understood that the generated new points do not contain information about the object. Over time, when there are new real measurements of object variables, they are naturally included in the initial learning sample.

Algorithm of generating new working sample is the following.

- Using initial sample $\{u_i, x_i, i = \overline{1,s}\}$ find value of smoothing parameter c_s by minimizing the criterion (3) for c_s.

- Let denote ρ_k as the number of the sample elements $\{u_i, x_i, i = \overline{1,s}\}$, which are located in c_s-neighborhood of the k-th element. In other words, choose those points of the sample $\{u_i, x_i, i = \overline{1,s}\}$, which satisfy the inequality: if $\prod_{j=1}^{m} \Phi(c_s^{-1}(u_k^j - u_i^j)) > 0$, then the element u_i is in c_s-neighborhood of u_k, otherwise - no.

- Calculate the average number of elements $\rho_{av.}$ in c_s-neighborhoods of the original sample elements using the following formula:

$$\rho_{av.} = s^{-1} \sum_{i=1}^{s} \rho_i. \tag{6}$$

- Calculate the Euclidean distance between all elements of the sample $\{u_i, x_i, i = \overline{1,s}\}$:

$$d(u_i, u_j) = \sqrt{\sum_{l=1}^{m} (u_i^l - u_j^l)^2}. \tag{7}$$

Let Ω' be a set of all distances $d(u_i, u_j), i = \overline{1,s}, j = \overline{1,s}, i > j$.

- Select elements of the initial sample $\{u_i, x_i, i = \overline{1,s}\}$ between which the distances d are minimal. For example, $\{u_i, x_i, i = \overline{1,s_1}\}, s_1 < s$. As it is shown by numerous numerical studies, the size of new sample s_1 ranges from 1/2 to 2/3 of the initial sample (size s). Among all the elements of the set $\Omega' \subset \Omega$ we find the minimum value of $d_{\min}(u_i, u_j)$ and begin to form a new sample set $\tilde{\Omega} \subset \Omega$, including in it the couple members u_i, u_j, and excluding from the set Ω'. Then find in the set Ω' the following minimum distance between elements of the initial sample, which is also included in the sample set. If $d_{\min}(u_i, u_j) = 0$, then the elements u_i and u_j coincide, we include in the sample set only one element which was not previously included. Repeat this procedure until a new set $\tilde{\Omega}$ will not contain about 70% of elements of the initial sample. We suggest to select 70% of the sample due to numerous computational experiments. So, $\tilde{\Omega}$ is a set of all elements that make up domains with a sufficient number of elements.

- Check the following condition for all the elements of sample set $\tilde{\Omega}$:

$$\rho_k < \rho_{av.}, k = \overline{1; s_1}. \tag{8}$$

If the condition 8 is satisfied, then we exclude the k-th element from the sample set $\tilde{\Omega}$.

- Thus, all elements of the sample set $\tilde{\Omega}$ exclude from the original sample $\{u_i, x_i, i = \overline{1,s}\}$. We obtain the sample $\{u_i, x_i, i = \overline{1,s'}\}, s' < s$. All the elements of the sample are located in sparsity subdomains. This sample also contains the boundary elements.

Next, consider the generation of new elements to be included in the working sample $\{\tilde{u}_i, \tilde{x}_i, i = \overline{1,N}\}, N > s$.

- Check the following condition for the sample size s': $\rho_i < \rho_{av.}, i = \overline{1,s'}$. If the condition is satisfied, then we generate k elements in c_s-neighborhood of the point u_i , where $k_i = \rho_{av.} - \rho_i, i = \overline{1,s'}$. For example, if there are 7 elements ($\rho_i = 7$) of the initial sample $\{u_i, x_i, i = \overline{1,s}\}$ in c_s-neighborhood of the element u_i and an average is 6 ($\rho_{av.} = 6$), we do not generate any new element, but if $\rho_i = 4$, then we generate 2 additional elements. These elements $\tilde{u}_k$ are generated according to the following rule:

$$\tilde{u}^j_{ki} = u^j_i + \zeta^j_k c_s, j = \overline{1,m}, i = \overline{1,s'}, k = \overline{1,k_i}, \tag{9}$$

where ζ^j_k is a random variable distributed according to a uniform law in the interval $[-1; 1]$, u^j_i is a value of input variables u_i. In c_s-neighborhood of the point u_i we generate elements $\tilde{u}_k$.

- For generated elements $u = \tilde{u}$ object output values $x(t)$ are not known, so for these elements we calculate the estimation of the output variable $x_s(\tilde{u})$ (2) based on the initial sample $\{u_i, x_i, i = \overline{1,s}\}$. Thus, for each value of $\tilde{u}$, obtained by (9), we calculate the estimation of $x_s(\tilde{u})$. If there is a situation of uncertainty, i.e. in c_s-neighborhood there are no elements, then we increase the value of smoothing parameter c_s. Generated elements and the initial learning sample form a new working sample $\{\tilde{u}_i, \tilde{x}_i, i = \overline{1,N}\}, N > s$. So, new working sample consists of observations $\{u_i, x_i, i = \overline{1,s}\}$ and generated elements.

- New elements are randomly generated in c_s-neighborhood of the initial sample elements, so some of them may be located too close to each other, for example, the distance between them is less than or equal to a sufficiently small value ε, defined experimentally, or even coincide. Such generated elements are not of interest, and they should be removed.

Note that we remove the artificially generated sample elements. To do this, we calculate the value of the average distance between the points of the main sample $\{u_i, x_i, i = \overline{1,s}\}$:

$$d_{av.} = \frac{2}{s(s-1)} \sum_{i=1}^{s} \sum_{j=1}^{s} d(u_i, u_j), i < j, \tag{10}$$

where $d(u_i, u_j)$ is the distance between the elements u_i and u_j, which is calculated by the formula (7).

- Determine the distance from the artificially generated element to all elements of the new sample. Next, we find the value of the minimum distance $d_{\min}$, if $d_{\min} < \varepsilon$, where $\varepsilon = ad_{av.}$, then this item is removed from the sample. The value of the coefficient a is determined experimentally so that the size of the working sample in 1.5 - 2 times bigger then the size of the initial sample $\{u_i, x_i, i = \overline{1,s}\}$. Thus, all the extra are deleted. By selecting different values of the coefficient a, we can adjust the size of the generated sample.

- After the working sample is generated, calculate the estimation (2), defining a new value of smoothing parameter c_s according to the criteria (3). In the estimation (2) we use instead of values x_i the object output for the observations of the initial sample $\{u_i, x_i, i = \overline{1,s}\}$ and model output value for the generated elements, as for them there is no way to calculate the output value of the object.

Thus, the new generated sample elements are located in sparsity subdomains (Fig. 3). The sample size is increased, on average, in 1.5 - 2 times, depending on the original sample size. In calculating the non-parametric estimation of regression function there are a large number of observations in the c_s-neighborhood of the initial observations sample, so we improve the accuracy of modeling and eliminate uncertainty in calculations.

This method of the working sample generation can improve the accuracy of recovery of nonparametric estimation of the regression function (2) for the boundary points due to the fact that in c_s-neighborhood of these points we generate some new elements.

The peculiarity of the above described method is that there is no need to specifically allocate the boundary elements of the initial sample. However, it must be done in order to estimate how accuracy of modeling is changed. This is easily done by statistical modeling methods.

5. Computer Modeling

In computer modeling of supplement of the initial sample $\{u_i, x_i, i = \overline{1,s}\}$ we get the working sample $\{\tilde{u}_i, \tilde{x}_i, i = \overline{1,N}\}, N > s$. In the interests of simplicity of visualization let consider two-dimensional vector $u \in \Omega(u) \subset R^2$. Without loss of generality, let the investigated object be described:

$$x(u) = u_1 + u_2 + \xi, \tag{11}$$

where ξ is uniformly distributed random noise:

$$\xi = k\Theta x, \tag{12}$$

where coefficient k determines the level of interference, Θ is a random variable distributed according to a uniform law with zero expectation in the range of $[-1; 1]$.

It should be noted that we take the uniform distribution law to toughen the simulation conditions, because normal and similar law of distribution is more natural in practice. The coefficient k, in fact, determines the percentage of interference. Thus, for example, for 5% noise: $k = 0,05$.

Let the values of the input variables be distributed in the range of $[0;3]$. Thus we have a sample of observations $\{u_i, x_i, i = \overline{1,100}\}$. The initial observations sample is generated such a way that in the space of input and output variables there are subdomains of sparsity with a small number of items (Fig. 2).

We construct nonparametric estimation of the regression function $x_s(u)$ (2) on the basis of the initial sample of observations. Note again that the dependence of (11) is unknown, but only a sample $\{u_i, x_i, i = \overline{1,100}\}$ is given. We present the following results when the nature of the relationship is non-linear, and the dimension of the vector x is 10. But first, let us consider the two-dimensional case in more detail.

We use the non-parametric model (2). If there is a situation of uncertainty of the forecast at $u = u'$ based on the original sample of observations $\{u_i, x_i, i = \overline{1,100}\}$, in c_s-neighborhood of the point $u = u'$ there are no sample elements, the forecast value $x_s(u')$ is assigned the expectation of the output variable x.

As a result of the above-described methods we receive new sample $\{\tilde{u}_i, \tilde{x}_i, i = \overline{1,281}\}$, which includes elements of the original sample and generated, which now is called the new learning sample. Then we conduct an experiment with this sample $\{\tilde{u}_i, \tilde{x}_i, i = \overline{1,281}\}$ in the sliding test mode. In addition, we generate from (11) a new sample $\{u_i, x_i, i = \overline{1,100}\}$ uniformly distributed in the space of input and output observations and use it as the examining sample.

The relative approximation error has the following form:

$$W = \sqrt{\frac{\frac{1}{s}\sum\limits_{i=1}^{s}(x_i - x_{si})^2}{\frac{1}{s-1}\sum\limits_{i=1}^{s}(x_i - \hat{m}_x)^2}}, \tag{13}$$

where $\hat{m}_x$ is the estimation of expectation of the output variable x, x_i is the result of measurement of output variable x when $u = u_i$, x_{si} is the value of nonparametric estimation when $u = u_i$.

In Table 1 the following notation is accepted: "before"is the relative error for the initial sample, "after"is the relative error for the working sample, which includes the elements of the original sample and generated with the proposed method, A is the number of elements of examining sample for which it is impossible (because of an uncertainty [0/0]) to receive the forecast based on the initial sample, B - based on the working sample.

As can be seen from Table 1, the use of the working sample leads to improvement of estimation accuracy in average to 2 times. Furthermore, the use of the working sample allows getting forecast for all examining sample points.

As we use the nonparametric models so parameterization is not required, these models are robust to the type of nonlinearity. Consider the results of modeling the nonlinear object. Let the object be described by the following equation:

$$x(u) = u_1^2 - 2\sin u_2 + \xi, \tag{14}$$

Таблица 1

Results of modeling of the object (11)

Sample	Error "before"	Error "after"	A	B
Examining sample	0,363	0,141	28	0
Initial sample	0,136	0,096	4	0
Border points of the initial sample	0,135	0,084	3	0

where ξ is uniformly distributed noise (12), input variables $u_1, u_2 \in [0;3] \times [0;3]$. We use the equation (14) to generate the initial learning sample $\{u_i, x_i, i = \overline{1,s}\}$.

The sample also contains sparsity. The size of the initial sample is 200 elements. Then, using the above algorithm we generate new elements. The size of the working sample is 453 element. We carry out series of experiments similar to the case of simulation of linear object. The results are shown in Table. 2.

Table 2

Results of modeling of the object (14)

Sample	Error "before"	Error "after"	A	B
Examining sample	0,838	0,237	4	0
Initial sample	0,212	0,113	0	0
Border points of the initial sample	0,347	0,155	0	0

As can be seen from the table 2, we can not get the estimation for 4 elements of the examining sample in the case of using the original sample. If we use the working sample, generated using the proposed method, there is no uncertainty and we can get the estimation for all sample elements. In addition, if we use the working sample, the nonparametric estimation $x_s(u)$ is two times more accurate.

Let consider results of the above experiments for the high dimensional vector u. Assume that the investigated object has the form (15):

$$x(u) = 0,5u_1 - \sin u_2 + 0,3u_3^2 + u_4 - 0,3u_5 + u_6 + 2u_7 + 2\cos u_8 + u_9 + u_{10} + \xi. \quad (15)$$

The size of the initial and examining sample is 300 elements. In the sample, as well as in previous experiments, there are subdomains of sparsity and the lack of observations. The results of similar experiments are shown in Table 3. As seen from the experiment results, the use of the new working sample increases the accuracy of identification.

6. A Practical Example

Let consider the results of applying the proposed method by the example of the oxygen-converter steel smelting process simulation. The process is described by the

Table 3

Results of modeling of the object (15)

Sample	Error "before"	Error "after"	A	B
Examining sample	0,812	0,612	51	0
Initial sample	0,427	0,277	1	0

controlled and uncontrolled variables. The controlled input variables are the following:

- material consumption, t: raw iron (u_1), scrap (u_2), lime (u_3), broken electrodes (u_4), flux (u_5), agglomerate fluxed (u_6), coal (u_7),
- oxygen blowdown, m^3 (u_8),
- heating oxygen, m^3 (u_9);

uncontrolled variables:

- the chemical composition of raw iron, %: silicon Si (mu_1), magnesium Mn (μ_2), sulfur S (mu_3), phosphorus P (μ_4),
- temperature of iron, °C (μ_5),
- converter load, t (μ_6);

and output variables, which are responsible for the quality of the finished steel:

- metal turndown temperature, ° C (x_1),
- the chemical composition of the metal on turndown, %: aluminum Al (x_2), carbon C (x_3), magnesium Mn (x_4), sulfur S (x_5), phosphorus P (x_6).

Thus, there are 15 input and 6 output variables which describe the investigated object. We have 176 the oxygen-converter steel heats. It is necessary to get a model of the process. Due to the fact that the a priori information is not sufficient, it is proposed to use a nonparametric estimation 2.

The simulation, as in the previous case, has two stages. At the first stage we use the initial sample of observations, obtained by measuring the input and output variables of the process, as a learning sample. At the second stage, using the proposed method we generate new elements. The initial sample and generated elements form the working sample. The simulation results are presented in Table 4.

г

Таблица 3

Resulta of Modeling of the oxygen-converter steel heats

Output variable	Error "before"	Error "after"	A	B
The metal turndown temperature (x_1)	0,99	0,51	19	0
Aluminum, Al (x_2)	1	0,63	30	0
Carbon, C (x_3)	1	0,59	24	0
Magnesium, Mn (x_4)	0,95	0,64	18	0
Sulfur, S (x_5)	0,85	0,35	15	0
Phosphorus, P (x_6)	1	0,49	18	0

We use the proposed methodology to supplement the initial sample of observations. Using the new learning sample leads to improvement of modeling accuracy. It should be noted that we make the estimation for all elements of the initial sample. For example, for the variable x_3 (carbon concentration) model provides a forecast for the whole sample.

7. Conclusion

The main purpose of this article is to improve the accuracy of nonparametric identification by supplement the initial sample obtained on the real object with new elements. It should be noted once again that the identification of noninertial process with delay is carried out in conditions of nonparametric uncertainty, when it is impossible to get the parametric model due to lack of a priori information. In many practical problems the distribution of measurements of input and output variables of the object is often substantially non-uniform, there can be subdomains of sparsity. Use of nonparametric identification algorithms, based on Nadaraya-Watson estimation, leads to a rather rough models, if size of the initial sample of observations is small. Earlier, we noted that the new generated elements of working sample do not replace observations on the object, but from a computational point of view, significantly improve the accuracy of nonparametric identification algorithms. We should also keep in mind that the new elements of the working sample are generated on the basis of available initial observations, so they are indirectly related to the object under investigation. In conclusion, we present the results of modeling the oxygen-converter steel smelting process. We apply the method of working sample generation which allow to significantly increase the accuracy of the model.

References

1. Medvedev A.V. Osnovy teorii adaptivnyh sistem [Fundamentals of the theory of adaptive systems]. Krasnoyarsk, SibGAU Publ., 2015, 525 p.

2. Zagoruyko N. G. Prikladnye metody analiza dannykh i znaniy [Applied methods of data and knowledge analysis]. Novosibirsk, IM SO RAN Publ., 1999. 264 p.

3. Cypkin Ja. Z. Osnovy informacionnoj teorii identifikacii [The foundation of information identification theory]. Moscow, Nauka Publ., 1984, 320 p.

4. Eykhoff P. Osnovy identifikacii sistem upravleniya [System identification parameter and state estimation]. Moscow, Mir Publ., 1975, 683 p.

5. Vasilev V. A., Dobrovidov A. V., Koshkin G. M. Neparametricheskoe ocenivanie funkcionalov ot raspredelenij stacionarnyh posledovatelnostej [Nonparametric estimation of functionals of distributions of stationary sequences]. Moscow, Nauka Publ., 2004, 512 p.

6. Hardle V. Applied nonparametric regression. Moscow, Mir Publ., 1993, 349 p.

7. Katkovnik V. Ya. Neparametricheskaja identifikacija i sglazhivanie dannyh: metod lokalnoj approksimacii [Non-parametric identification and data smoothing: local approximation method]. Moscow, Nauka Publ., 1985, 336 p.

8. Nadaraya E. A. Neparametricheskie ocenki plotnosti veroyatnosti i krivoj regressii [Non-parametric estimation of the probability density and the regression curve]. Tbilisi, Tbilisi University Publ., 1983, 194 p.

9. Bradley E. Bootstrap Methods: Another Look at the Jackknife. Annals of Statistics, 1979, vol. 7, no. 1, pp. 1 - 26.

10. Garcia-Soidan P., Menezes R., Rubinos O. Bootstrap approaches for spatial data. Stoch Environ Res Risk Assess, 2014, no. 28, pp. 1207-1219.

11. Loh J., Stein M. L. Spatial bootstrap with increasing observations in a fixed domain. Statistica Sinica, 2008, no. 18, pp. 667-688.

12. Medvedev A. V. Neparametricheskie sistemy adaptacii [Nonparametric adaptation systems]. Novosibirsk, Nauka Publ., 1983, 174 p.

О НЕПАРАМЕТРИЧЕСКОМ МОДЕЛИРОВАНИИ МНОГОМЕРНЫХ БЕЗЫНЕРЦИОННЫХ СИСТЕМ С ЗАПАЗДЫВАНИЕМ

А.В. Медведев, Е.А. Чжан

Рассматривается задача идентификации безынерционных объектов с запаздыванием в условиях непараметрической неопределенности, т.е. когда априорные сведения о параметрической структуре исследуемого объекта отсутствуют. Во многих приложениях возникает ситуация, когда измерение тех или иных выходных переменных осуществляется через значительные промежутки времени и могут существенно превышать постоянную времени объекта. В этой связи приходится рассматривать объект как безынерционный с запаздыванием. В сущности, для решения задач идентификации используются два основных подхода: один из них - это идентификация в «узком»смысле или параметрическая идентификация либо при недостатке априорных сведений для выбора параметрической структуры естественно применить методы локальной аппроксимации, которые в последнем случае используют в качестве априорных сведений лишь качественные свойства исследуемого объекта. В случае если исходные данные об объекте достаточно представительны, то непараметрическая идентификация дает удовлетворительный результат, если же в пространстве входных и выходных переменных имеют места разреженности, то качество непараметрических моделей существенно снижается. Настоящая статья посвящена методике заполнения или генерации обучающих выборок на основании имеющейся текущей информации. Это позволяет существенно повысить точность непараметрических моделей при идентификации безынерционных систем с запаздыванием. Проведенные вычислительные эксперименты подтвердили, что качество непараметрических моделей безынерционных систем может быть существенно улучшено в результате «ремонта»исходной выборки. Одновременно значительно повышается точность модели на границе областей определения входных-выходных переменных процесса.

Ключевые слова: непараметрическая идентификация, анализ данных, выборка, компьютерное моделирование.

Литература

1. Медведев, А.В. Основы теории адаптивных систем / А.В. Медведев. Красноярск: изд-во Сиб. гос. аэрокосмич. ун-та, 2015, 525 с.

2. Загоруйко, Н.Г. Прикладные методы анализа данных и знаний / Н.Г. Загоруйко. Новосибирск: Издательство ИМ СО РАН, 1999, 264 с.
3. Цыпкин, Я.З. Основы информационной теории идентификации / Я.З. Цыпкин. М.: Наука, 1984, 320 с.
4. Эйкхофф, П. Основы идентификации систем управления / П. Эйкхофф. М.: Мир, 1975. 681 с.
5. Васильев, В. А. Непараметрическое оценивание функционалов от распределений стационарных последовательностей / В. А. Васильев, А. В. Добровидов, Г. М. Кошкин. М. : Наука, 2004, 512 с.
6. Хардле, В. Прикладная непараметрическая регрессия. М.: Мир, 1993. - 349 с.
7. Катковник, В. Я. Непараметрическая идентификация и сглаживание данных: метод локальной аппроксимации / В. Я. Катковник. М.: Наука, Глав. ред. физико-математической лит-ры, 1985, 336 с.
8. Надарая, Э.А. Непараметрические оценки плотности вероятности и кривой / Э.А. Надарая. Тбилиси : издательство Тбилисского университета, 1983, 194 с.
9. Bradley, E. Bootstrap Methods: Another Look at the Jackknife / E. Bradley. // Annals of Statistics. 1979, Vol. 7, No 1, P. 1 - 26.
10. Garcia-Soidan, P. Bootstrap approaches for spatial data /P. Garcia-Soidan, R. Menezes, O. Rubinos // Stoch Environ Res Risk Assess. 2014, No 28, P. 1207-1219.
11. Loh, J. Spatial bootstrap with increasing observations in a fixed domain / J. M. Loh, M. L. Stein // Statistica Sinica. 2008, No 18, P. 667-688.
12. Медведев, А.В. Непараметрические системы адаптации / А.В. Медведев. Новосибирск: Наука, 1983, 173 с.

Александр Васильевич Медведев, доктор технических наук, профессор, кафедра «Информационные системы», Сибирский федеральный университет (г. Красноярск, Российская Федерация).

Чжан Екатерина Анатольевна, ассистент, кафедра «Информатика», Сибирский федеральный университет (г. Красноярск, Российская Федерация), e-mail: ekach@list.ru.

Поступила в редакцию