About algorithm of robust nonparametric estimation of regression function on observations

M A Denisov¹, E A Chzhan¹, A A Korneeva¹ and V V Kukartsev¹,²
¹Siberian Federal University, Institute of Space and Information Technologies, Krasnoyarsk 660074, Russian Federation
²Siberian State Aerospace University, Krasnoyarskiy Rabochiy Ave., 31, Russian Federation
E-mail: max_denisov00@mail.ru

Abstract. The field of research presented in the paper is aimed at studying the methods of robust statistics for the modeling of multidimensional processes of discrete-continuous type. The model of the investigated process is constructed using identification methods. In one case, it is used parametric methods of identification where priori information about the object of research is sufficient to build the model accurate within vector of parameters. The second case is specific to lack of priori information. The researchers do not know the structure of the model and represent the object in the form of a "black box", therefore, nonparametric identification methods could be used. The accuracy of models is estimated using a relative error of approximation, which shows how much the model output value corresponds to the output value of the object. The paper proposes a new method of outliers’ detection in the initial sample of observations, which is subsequently used for parametric and nonparametric identification of processes. The developed robust algorithm is applied to both types of models in order to determine in which case accuracy of outliers’ detection is higher. In addition, the above algorithm is compared with an algorithm based on the interquartile range.

1. Introduction
Processes modeling is a relevant objective for many fields of science and technology. This is due to quite obvious reasons. Conducting experiments on real objects can be expensive (building mechanisms connected with space vehicles or liquid rocket engines [1]), time-consuming, life-threatening or even impossible (for example, when the studied system exists in a single copy). Thus, the problem of modeling systems is of no small importance.

In practice most often occurs situations when data tables have defects. For example, omissions or outliers. The presence of outliers in the initial sample may be due to wrong data-entry (human factor) or because of any monitoring instruments defects, as well as an inaccurate mathematical model. The mean value, variance value, correlation number - all these things are sufficiently influenced by even at least one outlier. The values of these parameters can be greatly distorted, and the results of modeling and other mathematical experiments can be inaccurate.

Considering this kind of problem, special methods are applied that allow an abnormal sample measurement to be eliminated. Such methods are usually called robust.

Nowadays, there is an entire independent branch of mathematical statistics which is called robust statistics (or robust analysis) that describes a large range of algorithms and detection methods, and also reduces the impact of layouts in data in order to improve the accuracy of the constructed models [2, 3].
Using robust algorithms is significant if the researcher strives to achieve high accuracy of simulation results. For example, for the medical domain, accuracy is one of the most important factors. In the paper [4] author describes a wide class of problems which are solved with methods of robust statistics. In addition, robust statistics methods which have been already developed continue progressing and improving [5].

Thus, the considered class of questions has a high value in a large number of advanced applications. In this paper, an algorithm is proposed for performing detection and elimination of outliers in the initial sample of observations. Previously, this algorithm was used only for nonparametric identification methods [6]. In this paper, we test the algorithm using both parametric and nonparametric modeling methods. In addition, a comparative analysis of the proposed algorithm is implemented with another algorithm of robust statistics.

2. The problem statement

The generally accepted flowchart of the investigated discrete-continuous process is shown in figure 1. The notation of the figure: $A$ is unknown object functional; $y(t)\in \Omega(y)\subset \mathbb{R}^1$ is an output process variable; $u(t)=(u_1(t), u_2(t),...,u_m(t))\in \Omega(u)\subset \mathbb{R}^m$ is a vector of input signal, where $m$ is the number of input signals; $\xi(t)$ is the vector random noise; $t$ is the continuous time; $H^u, H^y$ are communication channels; $h^u(t)$, $h^y(t)$ are stochastic noise measurements; $\{u_{ji}, y_i, i=1,...,s; j=1,...,m\}$ is a training sample, where $s$ is sample size.

![Flowchart of investigated process.](image)

Figure 1. Flowchart of investigated process.

In this paper, under the computational experiment, we consider two situations. The first situation assumes that priori information about the object of research is sufficient, that means we know the parametric structure of the investigated process (figure 1). Therefore, in this situation the modeling is performed using the parametric identification method. In general, the parametric model can be represented as follows:

$$\hat{y}_i(a) = \hat{f}(a,a),$$

where $a = (a_1, a_2, ..., a_i)$ parameters vector of investigated process which should be estimated. In some cases, exist conditions when there is lack priori information about the examined process [7].

The second case is characterized by a situation in which priori information is insufficient. Therefore, the modeling occurs using nonparametric identification methods. We apply the classical nonparametric estimation of Nadaraya-Watson:

$$\hat{y}_i(u) = \frac{\sum_{i=1}^{s} \prod_{j=1}^{m} \Phi \left( \frac{u_i - u_j}{c_i} \right) / \sum_{i=1}^{s} \prod_{j=1}^{m} \Phi \left( \frac{u_i - u_j}{c_i} \right) }{\sum_{i=1}^{s} \prod_{j=1}^{m} \Phi \left( \frac{u_i - u_j}{c_i} \right) }.$$ 

3. Outliers detection algorithm

The paper proposes an algorithm that allows to determine enormously differ measurements in the data sample for the output variable of the object.

At the first stage, for all sampling points $\{u_{ji}, y_i, i = 1,...,s; j = 1,...,m\}$ we check the following condition:
\[ |y_i - \hat{y}_i| > \varepsilon \cdot \Delta_k \quad (k = 1, 2, i = 1, s), \]  
(3)

where \( \varepsilon \) is an adjusted parameter, \( i \) is observation number, \( k \) is the model type: parametric (1) or nonparametric (2) and parameter value \( \Delta_k \) is determined by the following formula

\[ \Delta_k = \frac{1}{s} \sum_{i=1}^{s} |y_i - \hat{y}_i| \quad (k = 1, 2), \]  
(4)

The algorithm is applied for each value from the initial sample of observations. If inequation (3) is met, this sample element can be named as an outlier. It is necessary to add that the outlier detection depends on the adjusted parameter \( \varepsilon \).

4. Computational experiments

Let the object be described:

\[ y = \alpha_1 \cdot \sin(u_1) + \alpha_2 \cdot \sin(u_2) + \xi, \]  
(5)

where \((\alpha_1, \alpha_2) = \alpha\) are coefficients of investigated object; \( \xi \) is centered noise generated as follows:

\[ \xi = y \cdot \psi \cdot b, \]  
(6)

where \( \psi \) is normally distributed random variable on the interval \([-1;1]\), \( b \) is noise percentage.

For parametric identification, the model equation (1) takes the following form:

\[ \hat{y}_i = \hat{\alpha}_1 \cdot \sin(u_1) + \hat{\alpha}_2 \cdot \sin(u_2). \]  
(7)

Further, for equation (7), it is necessary to estimate the parameters \( \hat{\alpha}_1 \) and \( \hat{\alpha}_2 \) based on the initial sample \( \{y_i, u_{1i}, u_{2i}, i = 1, \ldots, s\} \). There is a large number of methods and algorithms for solving this problem [8-10]. A classical least-squares criterion is used to estimate the parameters \( \hat{\alpha}_1, \hat{\alpha}_2 \):

\[ F(\hat{\alpha}) = \frac{1}{s} \sum_{i=1}^{s} (y_i - \hat{y}_i)^2 \rightarrow \min_{\hat{\alpha}}, \]  
(8)

where \( \hat{y}_i \) is an output value of (7), \( \hat{\alpha} = (\hat{\alpha}_1, \hat{\alpha}_2) \) are required coefficients.

To solve the identification problem under nonparametric uncertainty, we take as the model the following nonparametric Nadaraya-Watson estimation of regression function:

\[ \hat{y}_2(u) = \sum_{j=1}^{s} y_j \prod_{j=1}^{2} \Phi \left( \frac{u_j - u_{ji}}{c_j} \right) / \sum_{j=1}^{s} \prod_{j=1}^{2} \Phi \left( \frac{u_j - u_{ji}}{c_j} \right), \]  
(9)

where \( c_j \) is a bandwidth parameter and \( \Phi[(u_j - u_{ji})/c_j] \) is a kernel function satisfy the convergence conditions [11].

The kernel \( \Phi(z) \) has parabolic form:

\[ \Phi(z) = \begin{cases} 0.75 \cdot (1 - |z|)^2, & |z| \leq 1, \\ 0, & |z| > 1, \end{cases} \]  
(10)

where \( z = [(u_j - u_{ji})/c_j] \).

In order to determine how accurately the obtained models, correspond to the object under study we use a mathematical expression to calculate the relative error of approximation, the mathematical structure of which looks as follows:
\[ W = \sqrt{(s-1) \sum_{i=1}^{s} (y_i - \hat{y}_i)^2 / s \sum_{i=1}^{s} (\hat{m}_i - y_i)^2}, \]  (11)

where \( \hat{m}_i \) is estimation of mean, \( k = 1 \), if we use the parametric model (7); \( k = 2 \), if we use the nonparametric model (9).

5. Computational experiment

Under the conditions of the computational experiment, we generate the inputs of the object \( u_1, u_2 \) with pseudo-random numbers distributed uniformly in the interval \([1;4]\). Further, using (5), we generate the output of the object.

It is necessary to generate a sample containing the outliers to demonstrate the effectiveness of the proposed algorithm as we do not have real data from the object. To get a sample, let the minimum and maximum elements of the sample be multiplied by \(-0.5\). This condition allows to create outliers in the area of determining the values of the output of the object.

After making outliers (in this experiment, their number is three) we construct two models for the investigated object – parametric and nonparametric. In this regard, we determine the values of the outputs of these models by the formulas (7) and (9). The sample size equals \( s = 100 \) and noise level equals 5\%. We calculate the bandwidth parameter \( c_s \) for nonparametric model using cross-validation method and for this computational experiment its value equals \( c_s = 0.4 \).

The values of the outputs of the object (5), models (7) and (9) can be visualized in the form of a three-dimensional graph (figure 2).

\[ \begin{align*}
\text{Figure 2.} & \, \text{The values of the outputs of the object (5) and the models (7), (9) from the sample of observations: a) with outliers, b) without outliers.}
\end{align*} \]

It is shown on figure 2 (a) that outliers influenced on the final values of the model outputs. Besides, the parametric model (7) proved to be more stable to the influence of outliers than the nonparametric model (9). These statements can also be confirmed by the values of the relative error of approximation (11). The error is \( W_1 = 0.51 \) for parametric model (7), for nonparametric model (9) it is \( W_2 = 0.55 \).

Next step is to detect and eliminate outliers in the initial observation sample, which has been obtained in the previous stage. To do this, it is necessary to check condition (3) of the algorithm described above: so, find the value \( \Delta \) using (4); create \( \varepsilon \).

In order to understand how accurately the algorithm determines the outliers, it is necessary to focus on the value (11). If its value decreases, we assume that the accuracy of detecting outliers in the data using the algorithm is high. However, due to the fact that under the conditions of computational experiment we have information about what values are considered to be outliers, we can estimate the accuracy of the algorithm by checking the values of the outliers found by the algorithm coincide with those values, that were artificially generated in the output sample of the object (5).
In figure 2 (b) it is seen that the quality of the approximation has increased in comparison with figure 2 (a), which also confirms by error values (11): $W_1 = 0.03$, $W_2 = 0.12$.

Calculated relative error of approximation values in the second experiment are significantly lower than in the first one, what clearly indicates the correctness of the determination of the outliers’ values in the initial sample of observations.

6. Quartile detection algorithm

It should also be noted that there are other algorithms for detecting the outliers in the data, for example, the algorithm proposed by John Tukey [12], based on the value of interquartile range. The algorithm is simple in both usage as implementation in the applied tasks of the researcher. The basic principle is to use the first $Q_1$ and third $Q_3$ quartiles of the sets of variate values. Also, it is necessary to determine the value of the interquartile range, which is as follows:

$$A = Q_3 - Q_1.$$  \hspace{1cm} (12)

Further, the outliers’ detection occurs based on the following condition:

$$Q_i - A \cdot 1.5 \leq y_i \leq Q_i + A \cdot 1.5 \quad (i = 1, s),$$  \hspace{1cm} (13)

where $y_i$ is an object output value and $i$ is observation number. Values that deviate from the first and third quartiles by no more than one and a half interquartile ranges, that is, satisfy the condition (13), are not considered to be outliers. However, if these values do not satisfy the condition (13) they are defined as insignificant outliers, since their deviations from typical values are not very large.

To identify more significant outliers (which are differ enormously from other values), another condition is usually used:

$$Q_i - A \cdot 3 \leq y_i \leq Q_i + A \cdot 3 \quad (i = 1, s).$$  \hspace{1cm} (14)

If $y_i$ satisfies the condition (14), then we assume that there is a significant outlier.

We used the sample from the previous computational experiment. As a result, it was found that the algorithm based on the interquartile range is not able to determine the outliers located in the domain of the object input variables $u_1$, $u_2$. Using this algorithm, the outliers from the sample are not eliminated and obviously the error values (11) for the parametric and nonparametric model are the same as in the case of constructing models with outliers in the previous computational experiment: $W_1 = 0.519$, $W_2 = 0.546$.

However, if we simulate another situation, when the outlier values are not in the domain of the output variables of the object, but, for example, it is 5 times higher than the maximum allowable value. For this, in the initial sample of observations (from the first computational experiment), we artificially create three outliers as follows: multiply by 5 the outputs of the object $x_5$, $x_{30}$ and $x_{80}$. Next, we calculate the error value (11). Afterwards, we apply conditions (13) and (14) for detecting outliers and at the last stage eliminate these values from the sample of observations. Then re-calculate the error value (11). The noise level in this computational experiment equals $b = 5\%$ and sample size is $s = 100$.

The results are shown below.

Table 1. The error values (11) using quartile detection algorithm.

| Model             | Sample with outliers | Sample with excluded outliers by (13) | Sample with excluded outliers by (14) |
|-------------------|----------------------|---------------------------------------|---------------------------------------|
| Parametric (7)    | $W_1 = 0.37$         | $W_1 = 0.053$                         | $W_1 = 0.039$                         |
| Nonparametric (9) | $W_2 = 0.45$         | $W_2 = 0.16$                          | $W_2 = 0.13$                          |
In the course of the computational experiment, it was found that the outliers’ detection using the condition (13) gave less accuracy than the condition (14). These implications are confirmed by the errors in Table. 3. However, it should be noted that this difference is insignificant in this case.

The accuracy of the algorithm based on the quartiles is less than the algorithm proposed in this paper, which can be seen by comparing values of last and current computational experiment. Of course, the difference in the values of relative error of approximation is not so high, however, it should be kept in mind that for the algorithm proposed in this article, a computational experiment was conducted with outliers whose values were in the domain of determining the output values of the object, which significantly complicates the computational experiment. Consequently, the algorithm proposed by John Tukey is not always able to accurately determine outliers, even if they differ enormously from the rest of the sample of observations. The reason for this can be ignoring almost 50% of the initial data, which is caused by the lack of this measure of dispersion (interquartile range).

7. Conclusions

Thus, the paper proposed an algorithm that allows to detect outliers in data. Its performance was confirmed for models of two types: parametric and nonparametric, which had not previously been done. Comparing the results of the algorithm for these two modeling methods, we can state that its accuracy and quality do not depend on the chosen modeling method. It is only necessary to choose the right parameter $c$.

In addition, a comparison was made with John Tukey's algorithm. As a result, the undeniable advantage of the algorithm proposed in this paper was revealed: it is able to detect with high accuracy the outliers in the domain of determining the values of the object's output, which the algorithm based on the quartiles can not do. These results were confirmed in the computational experiment by determining the values of relative error of approximation.

Acknowledgments

The reported study was funded by Russian Foundation for Basic Research, Government of Krasnoyarsk Territory, Krasnoyarsk Regional Fund of Science to the research project: «Creation of efficient distributed networks of temperature sensors for on-board satellite equipment».

References

[1] Kovalev I V, Zelenkov P V and Tsarev M Yu 2015 IOP Conf. Ser.: Mater. Sci. Eng. 70 012008
[2] Huber P J 2011 Robust Statistics (New York: Wiley–Interscience)
[3] Maronna R A, Martin R D and Yohai V J 2006 Robust Statistics (Chichester: John Wiley & Sons)
[4] Kharin Y 2010 Robust statistics and its application The Science and Innovations 8 22–3
[5] Goryainov V B and Ermakov S Yu 2016 Robust estimation in random coefficient autoregression model Science and Education of the Bauman MSTU 9 111–22
[6] Korneeva A A 2014 Nonparametric models and control algorithms for multi-dimensional systems with delay (Krasnoyarsk: Siberian Federal University, Institute of Space and Information Technologies)
[7] Denisov M A and Chzhan E A 2017 About parametric identification algorithms of discrete-continuous processes Siberian Journal of Science and Technology 18 727-35
[8] Yuen K V and Mu H Q 2015 Real-time system identification: an algorithm for simultaneous model class selection and parametric identification Computer-Aided Civil and Infrastructure Engineering 30 785-801
[9] Mikleš J and Fikar M 2007 Process Modelling, Identification, and Control (Berlin: Springer) 221-51
[10] Keesman K J 2011 System identification. An introduction (London: Springer)
[11] Nadaraya E A 1963 On estimating regression Theory of Probability and Its Applications 10 186-96
[12] Tukey J W 1977 *Exploratory Data Analysis* (California: Addison-Wesley Publishing Company)