UDC 52-601

Siberian Journal of Science and Technology. 2017, Vol. 18, No. 4, P. 825–832

TO THE PROBLEM OF NONPARAMETRIC ROBUST ESTIMATION OF THE REGRESSION FUNCTION ON OBSERVATIONS

L. N. Sopova[1], S. S. Chernova[2*]

[1]Reshetnev Siberian State University of Science and Technology
31, Krasnoyarsky Rabochy Av., Krasnoyarsk, 660037, Russian Federation
[2]Siberian Federal University, Institute of space and Information Technologies
26b, Academica Kirenskogo Str., Krasnoyarsk, 660074, Russian Federation
[*]E-mail: chsvetlanas@gmail.com

There are parametric and nonparametric statistical models in the literature. These models differ from each other in levels of the prior indeterminacy in the statistical description of observations. The difference in ways these models were created tends to smoothing by introduction of transition models. It is explained by the fact that a statistical model, as well as any other model, is inevitable idealization and it can be only successful approximation of actual processes at its best. Emphasizing this fact, Box writes: "All models are irregular, but some of them are useful".

When using statistical procedures it is desirable to have information about what deviations have a decisive influence on the final conclusion at statistical analysis. In case the true distribution is not normal, there can be questions of normal theory reference procedures applicability. The recent research approach called "robast statistics" and offered as "third generation statistics" after parametric and nonparametric statistics by American mathematician J. Tyyuki is devoted to answer formulated above questions and create statistical procedures insensitive to deviations from assumptions. A number of publications on this approach constantly increases, there are already monographs, among them the first book of Hyubera, the book by F. Hampel and others, educational literature is also available.

The "robust" term, which corresponds to the definition "rough, strong", was introduced into statistical literature by Box in 1953 and since the middle of the sixtieth this term has became conventional for the section of statistics where statistical procedures insensitive to deviations from the accepted model assumptions develop. The robust idea has had a long history, which was described in Stigler's work. It appears in the work of K. Gauss, S. Newcomb, A. Eddington and others. However systematic development of robust ideas began with J. Tyyuki's works and, especially, after the work of Hyuber in 1964.

In this work an estimation of functions with a data outlier problem is given. In case of nonparametric indeterminacy the following steps are used to solve the problem:

1) the type of regression function with input data is set;

2) function estimation is applied.

We suggest the following reliable robust nonparametric estimation approach. The main idea is to exclude the data which can affect estimation.

Keywords: nonparametric regression estimation, nonparametric model, robust estimation procedure.

Сибирский журнал науки и технологий. 2017. Т. 18, № 4. С. 825–832

К ЗАДАЧЕ НЕПАРАМЕТРИЧЕСКОГО РОБАСТНОГО ОЦЕНИВАНИЯ ФУНКЦИИ РЕГРЕССИИ ПО НАБЛЮДЕНИЯМ

Л. Н. Сопова[1], С. С. Чернова[2*]

[1]Сибирский государственный университет науки и технологий имени академика М. Ф. Решетнева
Российская Федерация, 660037, г. Красноярск, просп. им. газ. «Красноярский рабочий», 31
[2]Сибирский федеральный университет, Институт космических и информационных технологий
Российская Федерация, 660074, г. Красноярск, ул. Академика Киренского, 26б
[*]E-mail: chsvetlanas@gmail.com

В литературе рассматриваются параметрические и непараметрические статистические модели. Эти модели отличаются друг от друга уровнями априорной неопределенности в статистическом описании наблюдений. Различие в способах задания этих моделей имеет тенденцию к сглаживанию, достигаемому путем введения промежуточных моделей. Это объясняется тем фактом, что статистическая модель, как и вообще любая модель, является неизбежной идеализацией и может оказаться в лучшем случае лишь удачной аппроксимацией реальных процессов. Подчеркивая этот факт, Бокс пишет: «Все модели неправильные, но некоторые из них полезны».

При использовании статистических процедур желательно иметь информацию о том, какие отклонения оказывают решающее влияние на конечный вывод при статистическом анализе. Могут возникнуть вопросы о применимости стандартных процедур нормальной теории, когда истинное распределение не является нормальным. Ответам на сформулированные вопросы и построению статистических процедур, нечувствительных к отклонениям от предположений, посвящено новое направление, названное робастной статистикой, которое было выделено американским математиком Дж. Тьюки в «статистику третьего поколения» после параметрической и непараметрической статистики. Публикации по этому направлению постоянно увеличиваются, уже имеется ряд монографий, среди них первая книга Хьюбера, книга Ф. Хампеля и др., также имеется и учебная литература.

Термин «робастность» соответствует английскому слову robust, перевод которого – «грубый, сильный, крепкий», в статистическую литературу этот термин был введен Боксом в 1953 году, и с середины шестидесятых годов этот термин стал общепризнанным для раздела статистики, в котором развиваются статистические процедуры, нечувствительные к отклонениям от предположений принятой модели. Отметим, что идеи робастности имеют давнюю историю, которая прослежена в работе Стиглера. Они появляются в работах К. Гаусса, С. Ньюкомба, А. Эдингтона и др. Однако систематическое развитие идей робастности начинается с работ Дж. Тьюки, и особенно после выхода работы Хьюбера в 1964 г.

Дана оценка функций с проблемой выброса данных. В случае непараметрической неопределенности для решения проблемы используются следующие шаги:

1) задан тип функции регрессии с исходными данными;

2) применяется оценка функции.

Предлагается надежный непараметрический подход к оценке. Основная идея состоит в том, чтобы исключить данные, которые могут повлиять на оценку.

Ключевые слова: непараметрические оценки функции регрессии, непараметрическая модель, процедура робастного оценивания.

Introduction. The problem of restitution of regression function on observations with outliers is considered [1–5]. When studying the task [6; 7], we use the suggested robust estimation procedure, which is a correction of the training sample free of outliers [8–11]. Thus, we obtain the function value and its restitution without outliers**.**

In the last decades of the last century, the intensive development and application of nonparametric and robust methods of data processing began [12–17]. The reason is that on the one hand there is the need to control complex economic and social structures without parametric descriptions, as well as technical objects, for which, for example, the applied methods stability is important to failures and noise in the operation of recording equipment, on the other hand, the development of computing technology, which makes it possible to implement laborious algorithms [18].

Nonparametric regression function estimatimation on observations. Nonparametric estimation of regression function on observations for a one-dimensional case is the following [19; 20]:

$$Y_s(x) = \frac{\sum_{i=1}^{s} y_i \Phi\left(\frac{x - x_i}{c_s}\right)}{\sum_{i=1}^{s} \Phi\left(\frac{x - x_i}{c_s}\right)}, \qquad (1)$$

$\Phi(v)$ – is the kernel. The kernel is a finite bell-shaped square integrable function satisfying conditions [19; 20]:

$$0 < \Phi(v) < \infty \quad \forall v \in \cap(v), \ \frac{1}{c_s}\int \Phi\left(\frac{x - x_i}{c_s}\right) dx = 1,$$

$$\lim_{n\to\infty} \frac{1}{c_s}\Phi\left(\frac{x - x_i}{c_s}\right) = \delta(x - x_i), \qquad (2)$$

c_s – blur coefficient which satisfies the following conditions:

$$c_s > 0, \ \lim_{s\to\infty} s(c_s)^k = \infty, \ \lim_{s\to\infty} c_s = 0. \qquad (3)$$

In case of multi-dimensional data (k-dimensional) it is:

$$Y_s(x) = \frac{\sum_{i=1}^{s} y_i \prod_{j=1}^{k} \Phi\left(\frac{x_j - x_j^i}{c_s}\right)}{\sum_{i=1}^{s} \prod_{j=1}^{k} \Phi\left(\frac{x_j - x_j^i}{c_s}\right)}, \qquad (4)$$

x_i, y_i, $i = \overline{1,s}$, – sample of observations; $\Phi(v)$ – bell-shaped function; v – random variable, c_s – blur coefficient.

Robust nonparametric regression function estimation on observations. Step-by-step experiment scheme is as followed:

1. The initial sample on an actual object is obtained.

2. We set up the blur coefficient and choose the bell-shaped function .

3. We check each sample point for estimation quality.

3.1. If the estimation quality is sufficient and inequality "more $2\sigma^2$" is not satisfied, then the initial sample becomes the working sample.

3.2. If the estimation quality is not sufficient and inequality "more $2\sigma^2$" is satisfied, then outliers are excluded from the initial sample and less points will become the working sample.

4. We restitute the regression function by means of nonparametric estimation.

Computing experiment. $y = \sin(x)^2$ is a function chosen for the computing experiment. When forming the training sample, outliers were artificially added.

The triangular kernel is used as a bell-shaped function $\Phi(v)$:

$$\Phi(v) = \begin{cases} 1 - |v|, & |v| \le 1, \\ 0, & |v| > 1. \end{cases} \qquad (5)$$

Further we perform the work with the entire sample constructing the function and its restitution, we find the criterion of accuracy. As the criterion of nonparametric estimation accuracy we use the quadratic criterion:

$$\sigma^2 = \sum_{i=1}^{s} (y_i - y_s(x_i))^2, \quad (6)$$

y_i – a true sample received on the formulas given above; $y_s(x_i)$ – is a nonparametric estimate.

After checking the accuracy criterion, we pay attention to the points at which the restitution error is big and they satisfy criterion (7). Elements of the training sample that satisfy the requirement:

$$\rho_i > 2\sigma^2, \quad (7)$$

where $\rho_i = (y_i - y_s(x_i)), i = \overline{1,s},$ are allocated and excluded from the initial sample.

We consider in fig. 1 – is a training sample, 2 – is nonparametric estimation. The triangular kernel was used as a bell-shaped finite function,

We present the results of the numerical experiment illustrating the effectiveness of an algorithm. We consider restitution of regression function on observations, which has several outliers at a sample size 100.

For illustrative purposes, we will add the perturbation action to some observations:

$$h_i = l y_i \xi, \quad (8)$$

where $\xi \in [-1,1]$, noise level is $l = 5$ %.

There are the elements of the sample, its approximation and two outliers on fig. 1. The restitution accuracy is 0.36. It is the same on fig. 2, except that 5 % noise level is added. The restitution accuracy is 0.40. It should be noted that restitution accuracy depends on whether there is the noise in the function.

There are five outliers on fig. 3. The restitution accuracy has obviously changed and is equal to 0.54.

Using $\rho_i > 2\sigma^2$, we exclude outliers from the initial sample.

Fig. 4 displays algorithm work with regard to robast estimation. In this case the sample size decreased because the program excluded outliers enterfering good restitution. In fig. 5 the 5 % noise level is added to the restituted function, accuracy of restitution decreased – 0.11, that is more than in fig. 4. Note that restitution accuracy significantly increased, not 0.36 and 0.54, but 0.06. It means that the given function was basically completely restituted.

As an experiment, the same function with the same outliers but for a smaller sample size – 60 was considered.

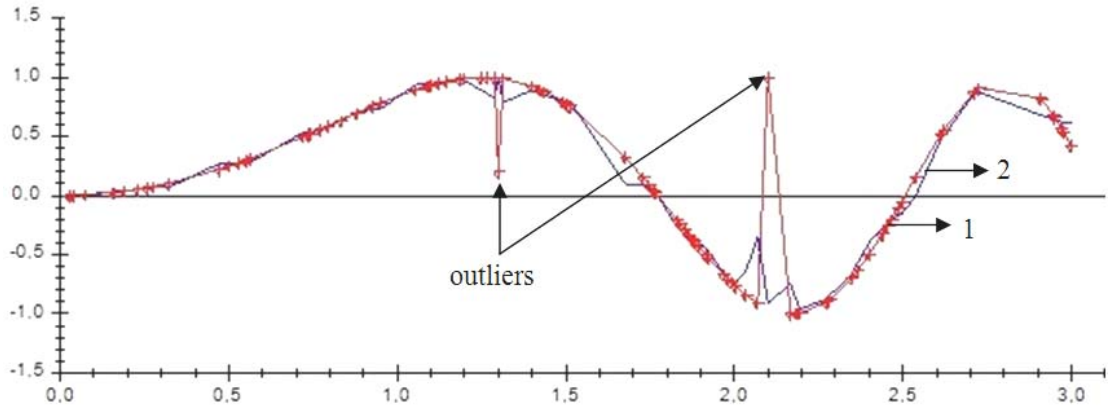

Fig. 1. Two-outlier restituted function

Рис. 1. Восстановленная функция с учетом двух выбросов

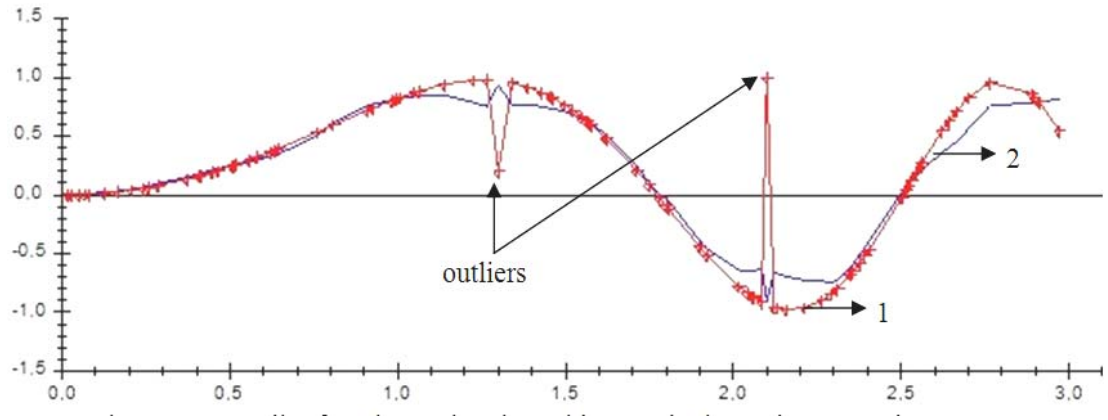

Fig. 2. Two-outlier restituted function with 5 % noise level

Рис. 2. Восстановленная функция с учетом двух выбросов и помехой 5 %

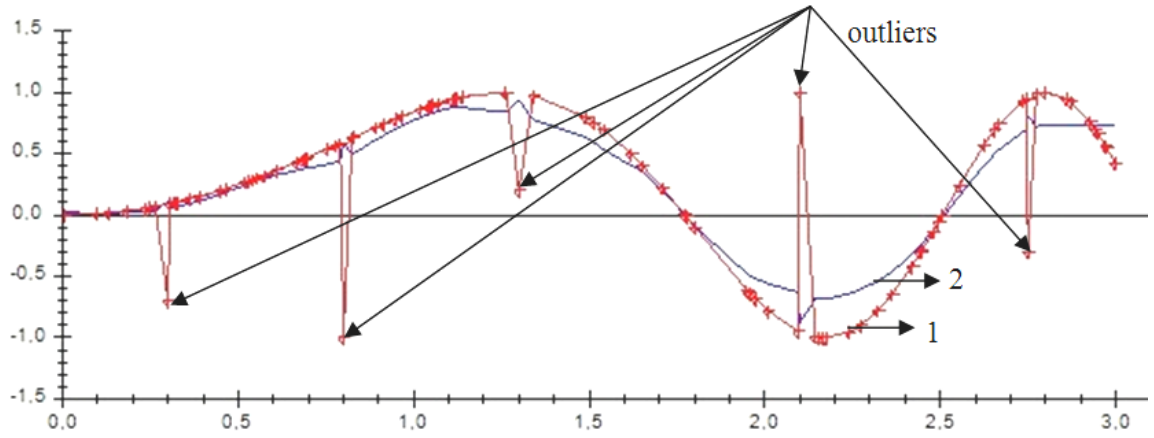

Fig. 3. Five-outlier restituted function

Рис. 3. Восстановленная функция с учетом пяти выбросов

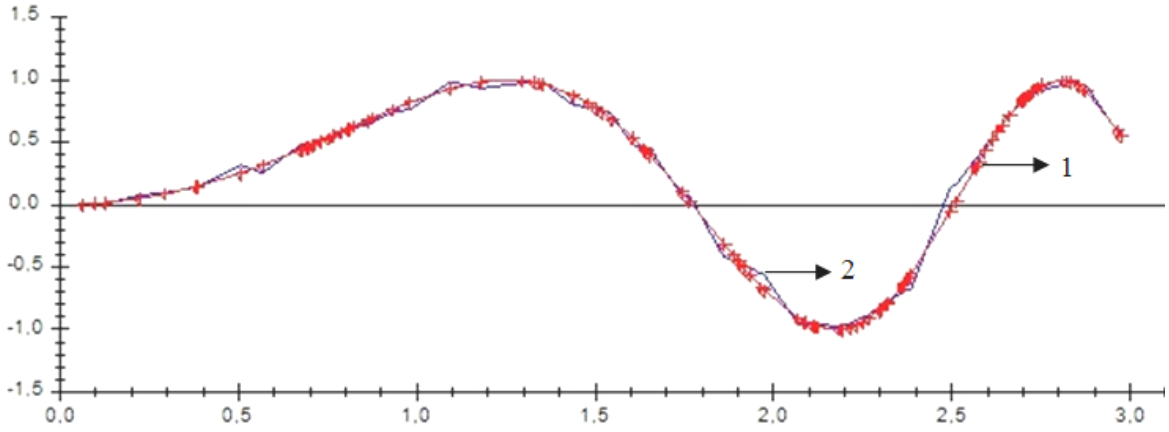

Fig. 4. None-outlier restituted function

Рис. 4. Восстановленная функция без учета выбросов

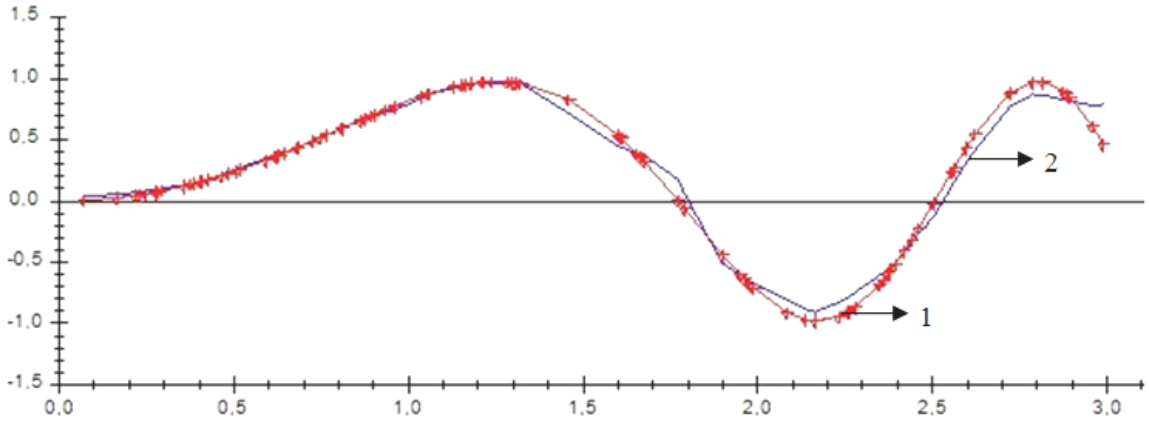

Fig. 5. None-outlier restituted function with 5 % noise level

Рис. 5. Восстановленная функция без учета выбросов, но с помехой 5 %

Sample units, its approximation and two outliers are also given in fig. 6. Restitution accuracy decreased to 0.45. Fig. 7 displays sample units with five outliers, the restitution error is 0.69. In fig. 8 the restituted function without outliers is presented, the accuracy of restitution is 0.14.

It is worth noticing that the sample size considerably influences restitution accuracy. For example, the accuracy of 100 elements sample size with regard to two outliers was 0.36, in the same case of 60 elements sample size it was 0.45.

For descriptive reasons we will consider one more similar function: $y = \cos(x)^2 \cdot \sin(x)$ with 100 elements sample size.

Sample units with one and three outliers respectively are given in fig. 9–11. Restitution accuracy at one outlier is 0.31, and at three – 0.41.

In fig. 10 the 5 % noise level was added to one-outlier restituted function. The restitution accuracy – 0.33. In this case, accuracy of restitution was not strongly affected by the noise.

Fig. 12 shows the restitution of function without outliers, the accuracy of restitution is 0.04. And in fig. 13 there is already 5 % noise level, restitution accuracy is 0.12. In this case, accuracy significantly decreased.

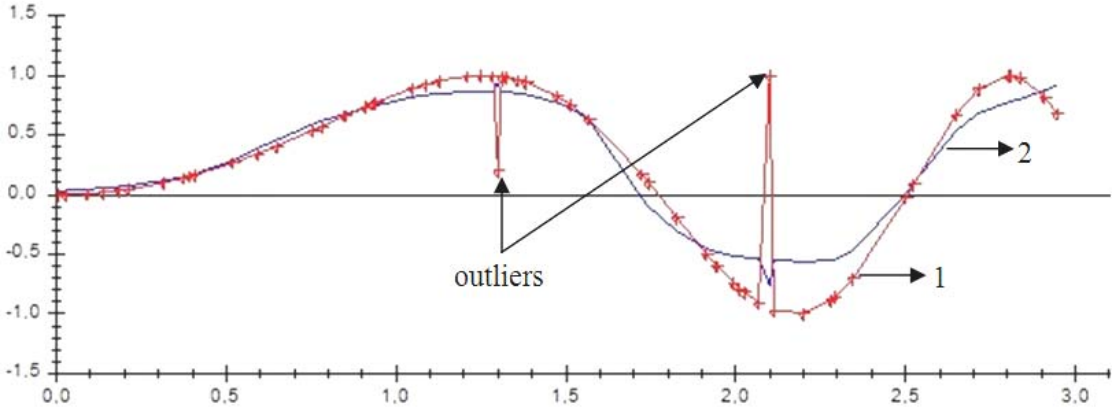

Fig. 2. Two-outlier function estimation

Рис. 6. Восстановленная функция с учетом двух выбросов

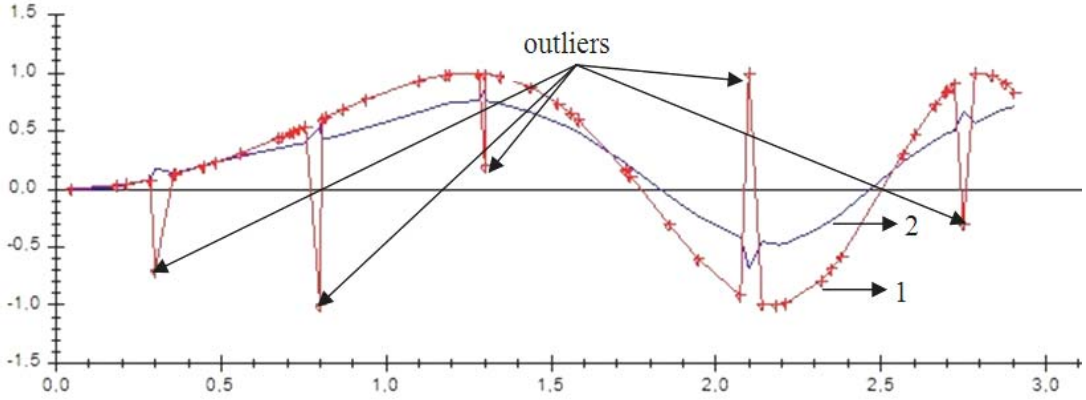

Fig. 7. Five-outlier restituted function

Рис. 7. Восстановленная функция с учетом пяти выбросов

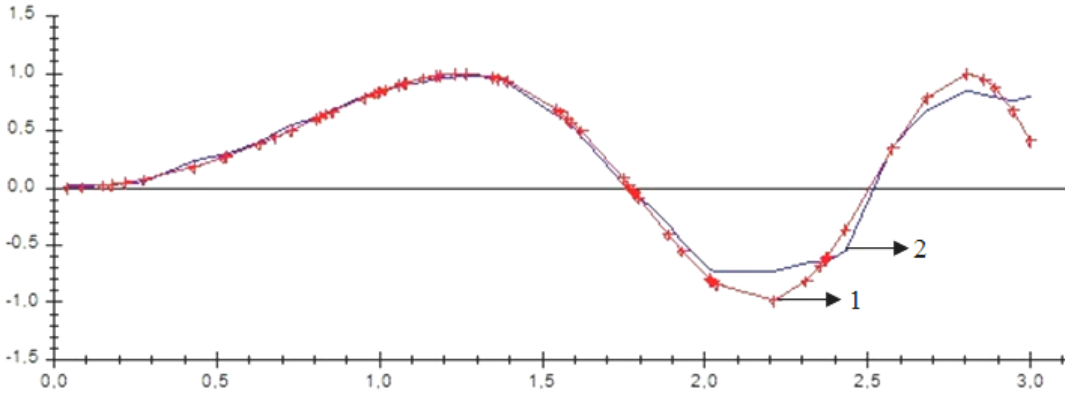

Fig. 8. None-outlier restituted function

Рис. 8. Восстановленная функция без учета выбросов

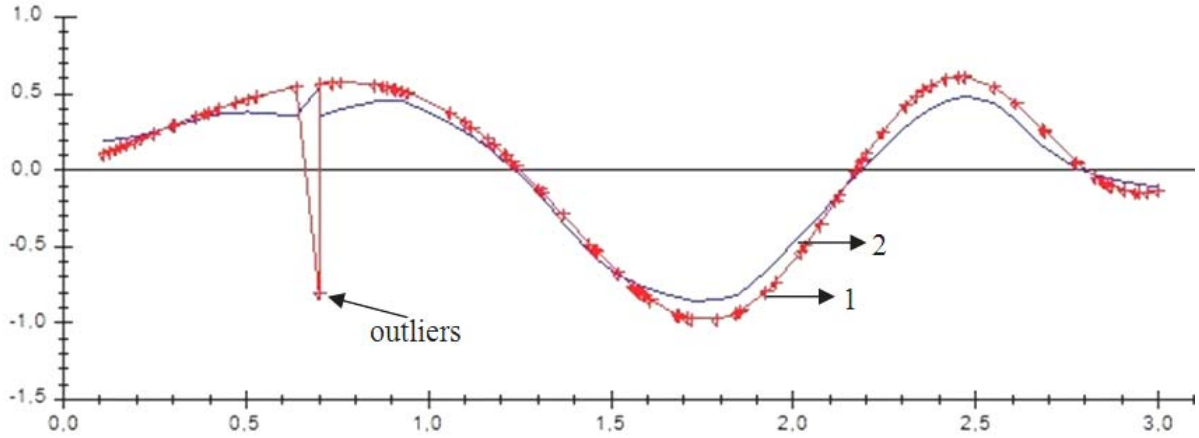

Fig. 9. One-outlier restituted function

Рис. 9. Восстановленная функция с учетом одного выброса

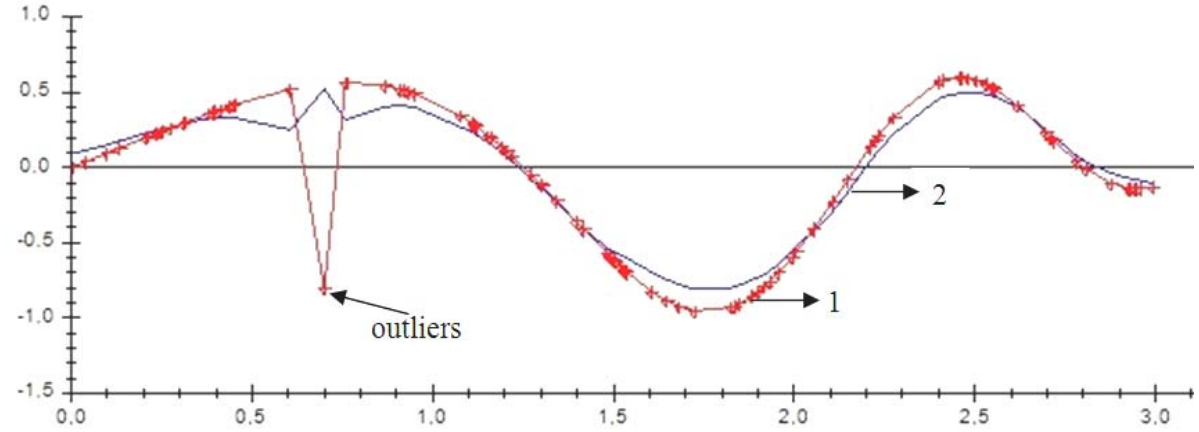

Fig. 10. One-outlier restituted function with 5 % noise level

Рис. 103. Восстановленная функция с учетом одного выброса и помехой 5 %

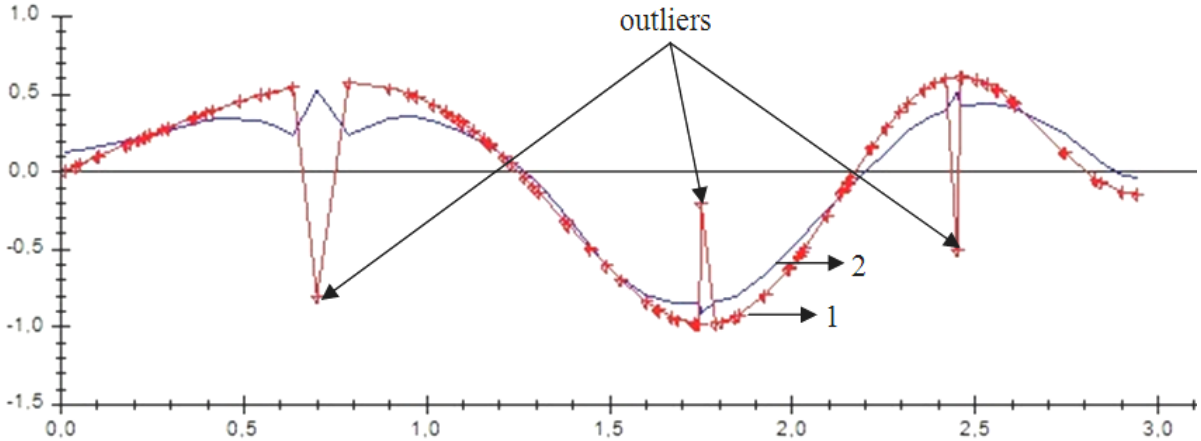

Fig. 11. Three-outlier restituted function

Рис. 11. Восстановленная функция с учетом трех выбросов

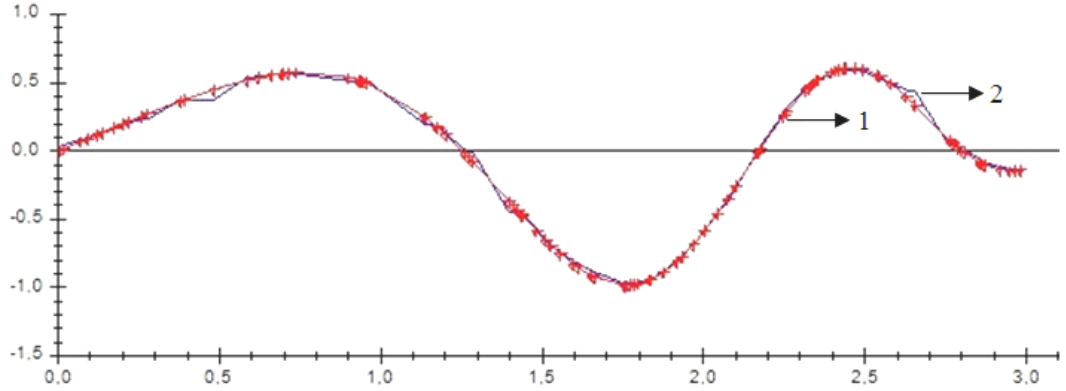

Fig. 12. None-outlier restituted function

Рис. 12. Восстановленная функция без учета выбросов

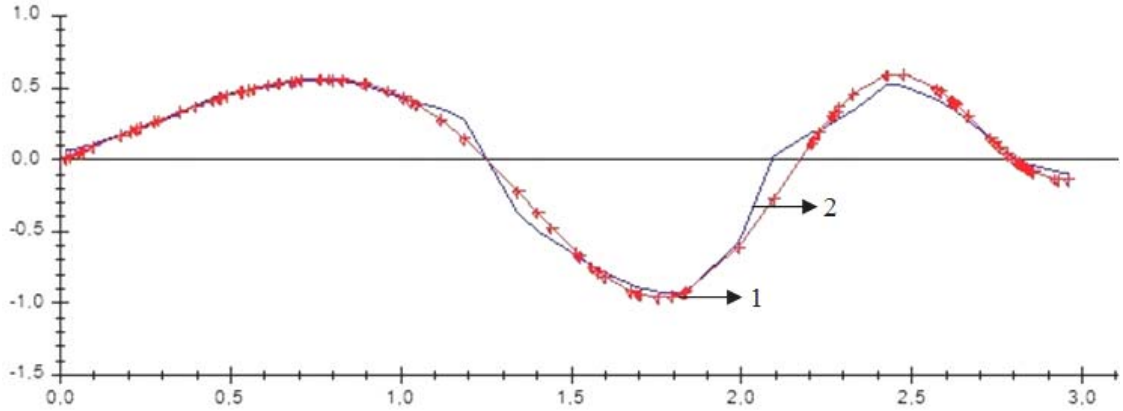

Fig. 13. None-outlier restituted function with 5 % noise level

Рис. 13. Восстановленная функция без учета выбросов, но с помехой 5 %

Conclusion. The main result of the article is that by means of the robast estimation approach it is possible to obtain significantly better function restitution quality on observations. It is worth noticing that restitution accuracy considerably increased after we excluded outliers. For descriptive reasons of the experiment several functions for restitution were considered. For the first function two sample sizes 100 and 60 were considered, we were visually convinced that the sample size has not small value for restitution. The restitution accuracy is significantly higher if the sample size is equal to 100 rather than if it is equal to 60.

References

1. Shulenin V. P. *Robastnye metody matematicheskoy statistiki* [Robust methods of mathematical statistics]. Tomsk, NTL Publ., 2016, 210 p.

2. Tarasenko F. P. *Neparametricheskaya statistika* [Nonparametric statistics]. Tomsk, Izdatel'stvo Tomskogo Universiteta Publ., 1976, 292 p.

3. Kh'yuber P. *Robastnost' v statistike* [Robustness in statistics]. Moscow, Mir Publ., 1984, 304 p.

4. Chernova S. S., Shishkina A. V. [On nonparametric estimation of mutually ambiguous functions from observations]. *Molodoi uchenyi.* 2017, No. 25, P. 13–20 (In Russ.).

5. Korneeva A., Chernova S., Shishkina A. Nonparametric algorithms for recovery of mutually unbeatted functions on observations, Applied Methods of Statistical Analysis. Nonparametric methods in cybernetics and system analysis – AMSA'2017. 18–22 September, Krasnoyarsk, Russia. 64–72 p.

6. Loner R. L., Uilkinson G. N. *Ustoychivye statisticheskie metody ocenki dannyh* [Sustainable statistical methods of data evaluation]. Moscow, Mashinostroenie Publ., 1984, 229 p.

7. Box G. E. P. Non-normality and test on variances. *Biometrika.* 1953, Vol. 40, P. 318–335 p.

8. Khampel' F., Ronchetti E., Raussei P., Shtael' V. *Robastnost' v statistike. Podkhod na osnove funktsii vliyaniya* [Robustness in statistics. The approach based on influence functions]. Moscow, Mir Publ., 1989, 512 p.

9. Shulenin V. P. *Matematicheskaya statistika. Ch. 1. Parametricheskaya statistika* [Mathematical statistics. Part 1. Parametric statistics]. Tomsk, NTL Publ., 2012, 540 p.

10. Shulenin V. P. *Matematicheskaya statistika. Ch. 2. Neparametricheskaja statistika* [Mathematical statistics. Part 2. Nonparametric statistics]. Tomsk, NTL Publ., 2012, 388 p.

11. Shulenin V. P. *Matematicheskaja statistika. Ch. 3. Robastnaja statistika* [Mathematical statistics. Part 3. Robust statistics]. Tomsk, NTL Publ., 2012, 520 p.

12. Stigler S. M. Simon Newcomb, Percy Daniel and history of robust estimations. *Journal of the American Statistical Association.* 1973, Vol. 68, P. 872−879.

13. Tukey J. W. A survey of sampling from contaminated distributions. Contributions to Probability and Statistics. Stanford Univ. Press, 1960, 448–485 p.

14. Tukey J. W. Bias and confidence in not-quite large samples (Abstract). *Annals of Mathematical Statistics.* 1958, Vol. 29, 614 p.

15. Tukey J. W. Data Analysis, Computation and Mathematics. *Quarterly of Applied Mathematics*. 1972, Vol. XXX, P. 51–65 p.

16. Tukey J. W. Exploratory Data Analysis. Reading, Mass., Addison Wesley, 1977.

17. Huber P. J. Robust estimation of location parameter. *Annals of Mathematical Statistics.* 1964, Vol. 35, No. 1, P. 73–101 p.

18. Kitaeva A. V. *Robastnoe i neparametricheskoe ocenivanie harakteristik sluchajnyh posledovatel'nostey. Diss. Dokt. Fiz.-mat. nauk* [Robust and nonparametric estimation of characteristics of random sequences. Doct. Diss.]. Tomsk, 2009, 324 p.

19. Nadaraia E. A. *Neparametricheskoe otsenivanie plotnosti veroiatnostei i krivoi regressii* [Nonparametric estimation of probability density and regression curve]. Tbilisi, TGU Publ., 1983, 194 p.

20. Medvedev A. V. *Osnovy teorii adaptivnykh system* [Fundamentals of the theory of adaptive systems]. Krasnoyarsk, SibGAU Publ., 2015, 526 p.

Библиографические ссылки

1. Шуленин В. П. Робастные методы математической статистики. Томск : НТЛ, 2016. 210 с.

2. Тарасенко Ф. П. Непараметрическая статистика. Томск : Изд-во Том. ун-та, 1976. 292 с.

3. Хьюбер П. Робастность в статистике. М. : Мир, 1989. 304 с.

4. Чернова С. С., Шишкина А. В. О непараметрическом оценивании взаимно неоднозначных функций по наблюдениям // Молодой ученый. 2017. № 25. С. 13–20.

5. Korneeva A., Chernova S., Shishkina A. Nonparametric algorithms for recovery of mutually unbeatted functions on observations // Applied Methods of Statistical Analysis. Nonparametric methods in cybernetics and system analysis – AMSA'2017 (18–22 September). Krasnoyarsk. P. 64–72.

6. Лонер Р. Л., Уилкинсон Г. Н. Устойчивые статистические методы оценки данных : пер. с англ. под ред. Н. Г. Волкова. М. : Машиностроение, 1984. 229 с.

7. Box G. E. P. Non-normality and test on variances // Biometrika. 1953. Vol. 40. P. 318–335.

8. Робастность в статистике. Подход на основе функций влияния / Ф. Хампель [и др.]. М. : Мир, 1989. 512 с.

9. Шуленин В. П. Математическая статистика. Ч. 1. Параметрическая статистика : учебник. Томск : НТЛ, 2012. 540 с.

10. Шуленин В. П. Математическая статистика. Ч. 2. Непараметрическая статистика. Томск : НТЛ, 2012. 388 с.

11. Шуленин В. П. Математическая статистика. Ч. 3. Робастная статистика. Томск : НТЛ, 2012. 520 с.

12. Stigler S. M. Simon Newcomb, Percy Daniel and history of robust estimations // J. Amer. Statist. Assoc. 1973. Vol. 68. P. 872−879.

13. Tukey J. W. A survey of sampling from contaminated distributions // Contributions to Prob. Statist. / Ingram Olkin, ed. Stanford Univ. Press, 1960. P. 448–485.

14. Tukey J. W. Bias and confidence in not-quite large samples (Abstract) // Ann. Math. Statist. 1958. Vol. 29. P. 614.

15. Tukey J. W. Data Analysis, Computation and Mathematics // Quarterly of Applied Mathematics. 1972. Vol. XXX, No. I. Special Issue: Symposium on the Future of Applied Mathematics. P. 51–65.

16. Tukey J. W. Exploratory Data Analysis. Reading, Mass. : Addison Wesley, 1977.

17. Huber P. J. Robust estimation of location parameter // Ann. Math. Statist. 1964. Vol. 35. No. 1. P. 73–101.

18. Китаева А. В. Робастное и непараметрическое оценивание характеристик случайных последовательностей : дис. … д-ра физ.-мат. наук. Томск, 2009. 324 с.

19. Надарая Э. А. Непараметрическое оценивание плотности вероятностей и кривой регрессии. Тбилиси : ТГУ, 1983. 194 с.

20. Медведев А. В. Основы теории адаптивных систем / СибГАУ. Красноярск, 2015. 526 с.