

Statistical Kernel Estimators for Design of a Fault Detection, Diagnosis, and Prognosis System

Piotr Kulczycki*

Cracow University of Technology, Department of Automatic Control, Cracow, Poland

Systems Research Institute, Polish Academy of Sciences, Warsaw, Poland

Abstract: This paper presents the concept of a fault detection system covering detection, diagnosis, and prognosis associated with them. To this aim procedures of data analysis and exploration, based on the nonparametric method of kernel estimators were applied. This method allows the useful characterization of probability distributions without arbitrary assumptions regarding their membership to a fixed class. The investigated system was proved to be effective with respect to both abrupt as well as slowly progressing symptoms of arising anomalies. This work has been aimed at the problem of fault detection in dynamical systems as objects of automatic control, although the general formula is universal and can be used for a wide range of tasks, also outside of engineering.

Keywords: Fault detection system, data analysis and exploration, nonparametric estimation, kernel estimators, control engineering.

1. INTRODUCTION

The problem of fault detection and diagnosis has lately become one of the most important challenges in modern control engineering. Early discovery of anomalies appearing in a controlled system, from an industrial robot to a nuclear reactor, most often allows serious incidents and even catastrophes to be avoided, which could save material damage, or even loss of human life. Secondly, identification of kind and location of these anomalies is of fundamental relevance, especially when supervising large systems like complex chemical installations, as well as modern ships and airplanes. The importance of the above actions is multiplied by a psychological factor expressed by an increased feeling of safety, as well as – for the producer – prestige and commercial reputation. Finally, economic reasons often translate into a significant decrease in running costs, above all by ensuring proper technological conditions as well as rationalizing overhauls and reducing repairs.

Among the many different procedures used with this aim, the most universal are statistical methods. These very often consist of generating a certain group of variables that characterize the technical state of the device (i.e. its working condition), and then making a statistical inference based on their current values, as to whether or not the device functions correctly, and in the event of a negative response, as to the nature of the anomaly appearing. For general concepts and useful details see the books [1-3].

This paper presents the concept of a fault detection system covering:

- detection, so discovery of the existence of potential anomalies in the technical state of a device under supervision;
- diagnosis, that is identification of these anomalies;
- prognosis, i.e. warning of the threat of their occurrence in the near future, together with anticipated classification.

The mathematical apparatus will be based on statistical inference, mainly kernel estimators methodology. Kernel estimators, whose development is directly related to the current expansion of computer technology, make up the leading method of nonparametric estimation for applicational purposes. The main goal here constitutes the calculation of various characteristics of probability distributions without arbitrary assumptions concerning their membership to a fixed class. First, Section 2.1 describes the basics of the methodology of statistical kernel estimators, and next Section 2.2 presents possible applications to fundamental problems of data analysis and exploration. In the concept dealt with here, kernel estimators will be applied to tasks of recognition of atypical elements (outliers), clustering and classification. It is worth noting that use of a single methodology for all investigated tasks significantly simplifies the process of synthesis of a fault detection system being worked upon. Consequently, Section 3, where the fault detection system designed here is described, will consist mainly of references to earlier material, and integrate them into one coherent idea. And finally, Section 4 presents results of experimental verification of the investigated concept and final suggestions.

The following text also contains material of research in the field of kernel estimators carried out together with Mal-

*Address correspondence to this author at the Cracow University of Technology, Department of Automatic Control, ul. Warszawska 24, PL-31-155 Cracow, Poland; Email: kulczycki@pk.edu.pl or Systems Research Institute, Polish Academy of Sciences, ul. Newelska 6, PL-01-447 Warsaw, Poland; E-mail: kulczycki@ibspan.waw.pl

gorzata Charytanowicz, Karina Daniel, Piotr A. Kowalski, Szymon Lukasik, Aleksander Mazgaj, Cyprian Prochot, and Jacek Waglowski, in particular contained in the publication [4]. Many applicational aspects can also be found in the works [5, 6]. The preliminary version of this paper was presented as [7].

2. KERNEL ESTIMATORS

2.1. Methodology

Let the n -dimensional random variable $X : \Omega \rightarrow \mathbb{R}^n$, with a distribution having the density f , be given. Its kernel estimator $\hat{f} : \mathbb{R}^n \rightarrow [0, \infty)$ is calculated on the basis of the m -elements simple random sample x_1, x_2, \dots, x_m , experimentally obtained from the variable X , and is defined in its basic form by

$$\hat{f}(x) = \frac{1}{mh^n} \sum_{i=1}^m K\left(\frac{x-x_i}{h}\right), \quad (1)$$

where the measurable function $K : \mathbb{R}^n \rightarrow [0, \infty)$, symmetrical with respect to zero and having a weak global maximum in this point, fulfils the condition $\int_{\mathbb{R}^n} K(x) dx = 1$ and is called a kernel, whereas the positive coefficient h is referred to as a smoothing parameter. It is worth noting that a kernel estimator allows the identification of density for practically every distribution, without any assumptions concerning its membership to a fixed class. Atypical, complex distributions, also multimodal, are regarded here as textbook unimodal. In the multidimensional case this enables, among others, the discovery of total dependences between particular coordinates of the random variable under investigation.

Setting the quantities introduced in definition (1), i.e. choice of the form of the kernel K as well as calculation of the value for the smoothing parameter h , is most often carried out according to the criterion of minimum of an integrated mean-square error. Broader discussion and practical algorithms are found in the books [8-10]¹. In particular, the choice of the kernel form has no practical meaning and thanks to this it is possible to first take into account properties of the estimator obtained (e.g. its class of regularity, boundary of a support, etc.) or aspects of calculations, advantageous from the point of view of the applicational problem under consideration. Practical applications may also use additional procedures, some generally improving the quality of the estimator, and others – optional – possibly fitting the model to an existing reality. For the first group one should recommend the modification of the smoothing parameter [8 – Section 3.1.6; 9 – Section 5.3.1] and a linear transformation [8 – Section 3.1.4; 9 – Section 4.2.1], while for the second, the boundaries of a support [8 – Section 3.1.8; 9 – Section 2.10]. It is worth mentioning also the possibility of applying data condensation and dimensionality reduction pro-

cedures – original and useful algorithms can be found e.g. in the book [11 – Sections 2.5 and 3.4].

Kernel estimators allow modeling of the distribution density – a basic functional characteristic of random variables. Consequently this is fundamental in obtaining other functional characteristics and parameters. For example, if in a one-dimensional case the kernel K is such chosen that its primitive function $I(x) = \int_{-\infty}^x K(y) dy$ may be analytically obtained, then the estimator of the distribution function

$$\hat{F}(x) = \frac{1}{m} \sum_{i=1}^m I\left(\frac{x-x_i}{h}\right) \quad (2)$$

can be easily calculated. Next, if the kernel K has positive values, the solution for the equation

$$\hat{F}(x) = r \quad (3)$$

constitutes the kernel estimator of quantile of the order $r \in (0, 1)$, denoted below as \hat{q} . For details and proofs of strong consistencies see the publications [8 – Section 3.2.2; 12].

2.2. Use for Data Analysis and Exploration

The application of kernel estimators for basic tasks of data analysis and exploration will be considered in this section, as subsequently will the recognition of atypical elements (outliers), clustering and classification. In all three cases the n -dimensional random variable $X : \Omega \rightarrow \mathbb{R}^n$ will be considered.

First, in many problems of data analysis the task of recognizing atypical elements (outliers) – those which differ greatly from the general population – arises. This enables the elimination of such elements from the available set of data, which increases its homogeneity (uniformity), and facilitates analysis, especially in complex and unusual cases. In practice, the recognition process for outliers is most often carried out using procedures of statistical hypotheses testing [13]. The significance test based on the kernel estimators methodology will now be described [14].

Let therefore the random sample x_1, x_2, \dots, x_m treated as representative, therefore including a set of elements as typical as possible, be given. Furthermore, let $r \in (0, 1)$ denote an assumed significance level. The hypothesis that $\tilde{x} \in \mathbb{R}^n$ is a typical element will be tested against the hypothesis that it is not, and therefore should be treated as an outlier. The statistic $S : \mathbb{R}^n \rightarrow [0, \infty)$, used here, can be defined by

$$S(\tilde{x}) = \hat{f}(\tilde{x}), \quad (4)$$

where \hat{f} denotes a kernel estimator of density, obtained for the random sample x_1, x_2, \dots, x_m mentioned above, while the critical set $A \subset \mathbb{R}$ takes the left-sided form

$$A = (-\infty, \hat{q}], \quad (5)$$

¹ For calculating a smoothing parameter one can especially recommend the plug-in method in the one-dimensional case [8 – Section 3.1.5; 10 – Section 3.6.1], as well as the cross-validation method [8 – Section 3.1.5; 9 – Section 3.4.3] in the multidimensional. Comments for the choice of a kernel may best be found in [8 – Section 3.1.3; 10 – Sections 2.7 and 4.5].

when \hat{q} constitutes the kernel estimator of quantile of the order r , calculated for the sample $\hat{f}(x_1), \hat{f}(x_2), \dots, \hat{f}(x_m)$, with the assumption that a random variable support is bounded to nonnegative numbers.

Secondly, the aim of clustering is the division of a data set – for example given in the form of the random sample x_1, x_2, \dots, x_m – into subgroups (clusters), with every one including elements “similar” to each other, but with significant differences between particular subgroups [15, 16]. In practice this often allows the decomposition of a large data set with differing characteristics of elements into subsets containing elements of similar properties, which considerably facilitates further analysis, or even makes it possible at all. The following clustering procedure based on kernel estimators, taking advantage of the gradient methods concept [17] will be presented now.

Here the natural assumption is made that clusters are associated with modes – local maximums of the density kernel estimator \hat{f} , calculated for the considered random sample x_1, x_2, \dots, x_m . Within this procedure, particular elements are moved in a direction defined by a gradient, according to the following iterative algorithm:

$$x_j^0 = x_j \text{ for } j = 1, 2, \dots, m \quad (6)$$

$$x_j^{k+1} = x_j^k + b \frac{\nabla \hat{f}(x_j^k)}{\hat{f}(x_j^k)} \text{ for } j = 1, 2, \dots, m \text{ and } k = 0, 1, \dots, \quad (7)$$

where $b > 0$ and ∇ denotes a gradient. Thanks to the proper choice of form of the kernel K , a suitable analytical formula for the gradient ∇ becomes possible. In practice the value $b = h^2/(n+2)$ may be used.

As a result of the following iterative steps, the elements of the random sample move successively, focusing more and more clearly on a certain number of clusters. They can be defined after completing the k^* -th step, where k^* means the smallest number k such that

$$|D_k - D_{k-1}| \leq c D_0, \quad (8)$$

where $c > 0$, and $D_0 = \sum_{i=1}^m \sum_{j=i+1}^m d(x_i, x_j)$ as well as

$$D_k = \sum_{i=1}^m \sum_{j=i+1}^m d(x_i^k, x_j^k) \text{ and analogously } D_{k-1}, \text{ i.e. they}$$

are the sums of the distances between particular elements of the random sample under consideration before the beginning of algorithm (6)-(7) and having performed the k -th and $(k-1)$ -th step, respectively. For practical purposes $c = 0.001$ may be used. Thus, after the k^* -th step, one should calculate the kernel estimator for mutual distances of the elements $x_1^{k^*}, x_2^{k^*}, \dots, x_m^{k^*}$ (under the assumption of nonnegative support of the random variable), and next, the value can be found where this estimator takes on the local minimum for the smallest value of its argument, omitting a possible minimum in zero. Finally, particular clusters are

assigned those elements, whose distance to at least one of the others is not greater than the above value.

Thanks to the possibility of change in the smoothing parameter value, it becomes possible to affect the range of a number of obtained clusters, albeit without arbitrary assumptions concerning the strict value of this number, which allows it to be suited to a true data structure. Moreover, possible changes in intensity of the smoothing parameter modification procedure enable influence on the proportion of clusters located in dense areas of random sample elements to the number of clusters on the “tails” of the distribution. For a detailed description of the above procedure see the publications [18, 19].

Thirdly, the application of kernel estimators in a classification task is considered. Let the number $J \in \mathbb{N} \setminus \{0, 1\}$ be given. Assume also, that the available random sample x_1, x_2, \dots, x_m has been divided into $J \in \mathbb{N} \setminus \{0, 1\}$ nonempty and separate subsets $\{x_1, x_2, \dots, x_{m_1}\}, \{x_1, x_2, \dots, x_{m_2}\}, \dots, \{x_1, x_2, \dots, x_{m_j}\}$, while $\sum_{j=1}^J m_j = m$, representing classes with features as mutually different as possible. The classification task requires deciding into which of them the given element $\tilde{x} \in \mathbb{R}^n$ should be reckoned [15, 16].

The kernel estimators methodology provides a natural mathematical tool for solving the above problem in the optimal – in the sense of minimum for expectation of losses – Bayes approach. Let thus $\hat{f}_1, \hat{f}_2, \dots, \hat{f}_J$ denote kernel estimators of density calculated for the subsets $\{x_1, x_2, \dots, x_{m_1}\}, \{x_1, x_2, \dots, x_{m_2}\}, \dots, \{x_1, x_2, \dots, x_{m_j}\}$, respectively, treated here as samples. If the sizes m_1, m_2, \dots, m_j are proportional to the “frequency” of appearance of elements from particular classes, the considered element \tilde{x} should be reckoned into the class for which the value

$$m_1 \hat{f}_1(\tilde{x}), m_2 \hat{f}_2(\tilde{x}), \dots, m_j \hat{f}_j(\tilde{x}) \quad (9)$$

is the greatest.

3. FAULT DETECTION SYSTEM

The procedures presented in the previous section, for recognition of atypical elements (outliers), clustering and classification, based on kernel estimators, provide a complete and methodologically consistent mathematical tool to design an effective fault detection system for dynamical systems, covering detection, diagnosis, and also prognosis associated with them.

Assume that the technical state of a device under supervision may be characterized by a finite number of quantities measurable in real-time. These will be denoted in the form of the vector $x \in \mathbb{R}^n$, called a symptom vector. One can interpret this name noting that symptoms of any occurring anomalies should find the appropriate reflection in the features of a such-defined vector. More strictly, it is required that both correct functioning conditions and any type of diagnosed fault are connected with the most different sets of

values and/or dissimilar relations between coordinates of the above vector as possible.

Assume also the availability of a fixed set of values of the symptom vector, representative for correct functioning conditions of a supervised device

$$x_1, x_2, \dots, x_{m_0}, \tag{10}$$

as well as the set

$$x_1, x_2, \dots, x_M, \tag{11}$$

characteristic in the case of occurrence of anomalies. From the point of view of transparency of the designed fault detection system, in particular its function of diagnosis, it is worth dividing set (11) into $J \in \mathbb{N} \setminus \{0,1\}$ the most possibly different – in the sense of the values of particular coordinates of the symptom vector and/or relations between them – subsets assigned to the previously assumed types of diagnosed faults:

$$x_1, x_2, \dots, x_{m_1} \tag{12}$$

$$x_1, x_2, \dots, x_{m_2} \tag{13}$$

⋮

$$x_1, x_2, \dots, x_{m_j}, \tag{14}$$

while $\sum_{j=1}^J m_j = M$. Where there is no such division, one can automatically divide set (11) into subsets (12)-(14) using the clustering algorithm presented in Section 2.2, although this then often requires laborious interpretation concerning each of them.

Fault detection will first be considered. With this aim the procedure for the recognition of atypical elements, described at the beginning of Section 2.2, can be applied. Assume therefore that the random sample x_1, x_2, \dots, x_m considered there, including elements treated as typical, constitutes set (10) representing the correct functioning conditions for a supervised device, while \tilde{x} denotes its current state. Applying the above procedure for the recognition of atypical elements, one can confirm if the present conditions should be regarded as typical or rather not, thus showing the appearance of anomalies.

For fault diagnosis, if one already is in possession of samples (12)-(14) characterizing particular types of faults being diagnosed, then after the above described detection of anomalies, one can – applying directly the procedure for Bayes classification presented at the end of Section 2.2 – infer which of them is being dealt with. Note that the range of faults which can be discovered by detection, may significantly exceed all types of faults assumed to be diagnosed.

Finally, if subsequent values of the symptom vector, obtained successively during the supervising process, are available, then it is possible to realize fault prognosis. It can be carried out by separate forecasts of values of the function \hat{f} given by dependence (4) and $m_1\hat{f}_1, m_2\hat{f}_2, \dots, m_j\hat{f}_j$ to be

seen in formula (9), and inferences based on these forecasts for detection and diagnosis, according to guidelines presented in the previous two paragraphs. To calculate the values of forecasts of the functions $\hat{f}, \hat{f}_1, \hat{f}_2, \dots, \hat{f}_j$ it is recommended to use the classical linear regression method separately, though in a version enabling easy updating of a model during successive collection of subsequent current values of the symptom vector. Appropriate formulas are found in the books [20 – Chapter 3 and additionally Chapter 4; 3 – Chapter 4].

4. EXPERIMENTAL VERIFICATION AND FINAL SUGGESTIONS

The proper operation of the fault detection system investigated in this paper has been verified experimentally, on the basis of an example which is simple yet useful in illustrative interpretation. The supervised object was a mechanical system with dynamics modeled by the differential inclusion

$$\ddot{y}(t) \in H(\dot{y}(t)) + u(t), \tag{15}$$

where y expresses the position of the object, u is a control with values limited to the interval $[-1,1]$, and the function H , characterizing resistance to motion, is piecewise continuous and additionally multivalued at the points of discontinuity (particularly at zero it can represent phenomena connected with static friction). In the event of no resistance to motion, i.e. when $H \equiv 0$, inclusion (15) is reduced to a differential equation $\ddot{y}(t) = u(t)$ expressing Newton's second law of dynamics. The above task constitutes therefore a problem of fundamental importance in the control of industrial manipulators and robots. Object (15) was subjected to a robust time-optimal control, which took on the values $+1$ or -1 , depending on where among the distinguished sets the system state was located; for details see the papers [21, 22].

The symptom vector was assumed in the following form of the 3-dimensional vector, which for any fixed moment t can be described in the following form:

$$x(t) = \begin{bmatrix} |u(t)| \\ |H(\dot{y}(t))| \\ |\dot{y}(t)| \end{bmatrix}, \tag{16}$$

and therefore its coordinates designate the absolute values of control, resistance to motion, and velocity, respectively. Diagnosis consisted of recognizing two types of faults. The first was assumed to be the reduction of maximum absolute value of the admissible control, which in practice indicates anomalies in the drive system. The second type of diagnosed fault was taken to be an increase in resistance to motion (whose values are strongly dependent on velocity) – in practice this would indicate that the displacement mechanisms are malfunctioning. Thus the first type of fault to be diagnosed entailed recognizing changes in the value of a single coordinate of the symptom vector, while the second involved the relations among particular coordinates.

The results of these experiments positively verified the concept presented in this paper and confirmed the proper functioning of the statistical inference system designed here.

In cases where the symptoms appeared abruptly, the anomalies of the device were promptly discovered and correctly recognized within the scope of detection and diagnosis. If, on the other hand, the fault was accompanied by a slow progression of symptoms, it was forecast with a correct indication of the type of fault about to occur (scope of prognosis), and later it was also discovered and identified in detection and diagnosis. One should underline that fault prognosis, still rare in practical applications, proved to be highly effective in the case of slowly progressing symptoms, discovering and identifying anomalies before the object's characteristics moved beyond the range for correct conditions for a system's functioning, thanks to the proper recognition of the change in the trend of values of the symptom vector, which indicates an adverse direction of its evolution.

The concept of the fault detection system presented in this paper is universal and complete from a methodological point of view. However, the specifics of each problem allow a designed fault detection system to be especially suited to its conditioning. As an example, for purposes of fault diagnosis proposed above, one can suggest to limit the dimension of the symptom vector, for the purposes of individual diagnosed types, to only those coordinates which are of essential significance for the recognition of the given type of fault – in the above task to $|u(t)|$ for examination of the drive system (the first type of diagnosed faults), and to $|H(\dot{y}(t))|$, $|\dot{y}(t)|$ for displacement mechanisms (the second one).

The above research applied the normal kernel

$$K(x) = \frac{1}{(2\pi)^{n/2}} e^{-x^T x/2}, \quad (17)$$

as this is differentiable and assumes positive values in its whole domain. Unfortunately it does not possess a primitive function and that is why, to calculate the kernel estimator of quantile, the one-dimensional Cauchy kernel

$$K(x) = \frac{2}{\pi(x^2 + 1)^2} \quad (18)$$

was used; its primitive function can be given in the analytical form

$$I(x) = \frac{1}{\pi} \left[\frac{x}{x^2 + 1} + \arctg(x) + \frac{\pi}{2} \right]. \quad (19)$$

Kernels (17)-(18) can be recommended for the majority of practical applications.

REFERENCES

- [1] J. J. Gertler, *Fault Detection and Diagnosis in Engineering Systems*. New York: Dekker, 1998.
- [2] R. Isermann, *Fault-Diagnosis System: An Introduction from Fault Detection to Fault Tolerance*. Berlin: Springer, 2005.
- [3] P. Kulczycki, *Wykrywanie uszkodzen w systemach zautomatyzowanych metodami statystycznymi*. Warsaw: Alfa, 1998.
- [4] P. Kulczycki and K. Daniel, "Algorytm wspomagania strategii marketingowej operatora telefonii komorkowej," in *Badania operacyjne i systemowe 2006: metody i techniki*, J. Kacprzyk and R. Budzinski, Eds. Warsaw: EXIT, 2006, pp. 245-256.
- [5] P. Kulczycki, "Estymatory jadowe w badaniach systemowych," in: *Techniki informacyjne w badaniach systemowych*, P. Kulczycki, O. Hryniewicz, and J. Kacprzyk, Eds. Warsaw: WNT, 2007, pp. 79-105.
- [6] P. Kulczycki, "Kernel Estimators in Industrial Applications," in: *Soft Computing Applications in Industry*, B. Prasad, Ed. Berlin: Springer-Verlag, 2008, pp. 69-91.
- [7] P. Kulczycki, "Analiza danych z uzyciem estymatorow jadowych w zastosowaniu do diagnostyki systemow," in: *Diagnostyka procesow i systemow*, J. Korbicz, K. Patan, and M. Kowal, Eds. Warsaw: EXIT, 2007, pp. 231-238.
- [8] P. Kulczycki, *Estymatory jadowe w analizie systemowej*. Warsaw: WNT, 2005.
- [9] B. W. Silverman, *Density Estimation for Statistics and Data Analysis*. London: Chapman and Hall, 1986.
- [10] M. P. Wand and M. C. Jones, *Kernel Smoothing*. London: Chapman and Hall, 1994.
- [11] S. K. Pal and P. Mitra, *Pattern Recognition Algorithms for Data Mining*. London: Chapman and Hall, 2004.
- [12] P. Kulczycki and A. L. Dawidowicz, "Kernel estimator of quantile," *Universitatis Iagellonicae Acta Mathematica*, vol. XXXVII, pp. 325-336, 1999.
- [13] V. Barnett and T. Lewis, *Outliers in Statistical Data*. Chichester: Wiley, 1994.
- [14] P. Kulczycki and C. Prochot, "Wykrywanie elementow odosobnionych za pomoca metod estymacji nieparametrycznej," in: *Badania operacyjne i systemowe: podejmowanie decyzji – podstawy teoretyczne i zastosowania*, R. Kulikowski, J. Kacprzyk, and R. Slowinski, Eds. Warsaw: EXIT, 2004, pp. 313-328.
- [15] J. H. Hand, H. Mannila, and P. Smyth, *Principles of Data Mining*. Cambridge: MIT Press, 2001.
- [16] D. T. Larose, *Discovering Knowledge in Data. An Introduction to Data Mining*. New York: Wiley, 2005.
- [17] K. Fukunaga and L. D. Hostetler, "The estimation of the gradient of a density function, with applications in Pattern Recognition," *IEEE Trans. Inf. Theory*, vol. 21, pp. 32-40, 1975.
- [18] P. Kulczycki and M. Charytanowicz, "Kompletny algorytm gradientowej klasteryzacji," in: *Sterowanie i automatyzacja: aktualne problemy i ich rozwiazania*, K. Malinowski and L. Rutkowski, Eds. Warsaw: EXIT, 2008, pp. 312-321.
- [19] P. Kulczycki and M. Charytanowicz, "A Complete Gradient Clustering Algorithm Formed with Kernel Estimators," *in press*.
- [20] B. Abraham and J. Ledolter, *Statistical Methods for Forecasting*. New York: Wiley, 1983.
- [21] P. Kulczycki, "A Random Approach to Time-Optimal Control," *J. Dynamic Syst., Measurement, Cont.*, vol. 121, pp. 542-543, 1999.
- [22] P. Kulczycki, "Fuzzy Controller for Mechanical Systems," *IEEE Trans. Fuzzy Syst.*, vol. 8, pp. 645-652, 2000.