BULLETIN OF THE POLISH ACADEMY OF SCIENCES
TECHNICAL SCIENCES
Vol. 56, No. 4, 2008

Applicational possibilities of nonparametric estimation of distribution density for control engineering

P. KULCZYCKI*

Department of Automatic Control, Cracow University of Technology, 24 Warszawska St., 31-155 Cracow, Poland
Systems Research Institute, Polish Academy of Sciences, 6 Newelska St., 01-447 Warsaw, Poland

Abstract. Together with the dynamic development of modern computer systems, the possibilities of applying refined methods of nonparametric estimation to control engineering tasks have grown just as fast. This broad and complex theme is presented in this paper for the case of estimation of density of a random variable distribution. Nonparametric methods allow here the useful characterization of probability distributions without arbitrary assumptions regarding their membership to a fixed class. Following an illustratory description of the fundamental procedures used to this end, results will be generalized and synthetically presented of research on the application of kernel estimators, dominant here, in problems of Bayes parameter estimation with asymmetrical polynomial loss function, as well as for fault detection in dynamical systems as objects of automatic control, in the scope of detection, diagnosis and prognosis of malfunctions. To this aim the basics of data analysis and exploration tasks – recognition of outliers, clustering and classification – solved using uniform mathematical apparatus based on the kernel estimators methodology were also investigated.

Key words: control engineering, nonparametric estimation, density of probability distribution, kernel estimators, data analysis and exploration, Bayes parameter estimation, fault detection, optimal control, robust control.

1. Introduction

In contemporary control engineering, the quality of the control algorithm – although itself a central element responsible for the correct running of an automatic device – most often depends considerably on many other factors which are of both subordinate (e.g. model of an object) and superior (e.g. fault detection system) function, but are always subject to the main goal of this algorithm. Despite it seeming that the development of innovative methods based on knowledge engineering and data exploratory analysis will slowly blur the division between the above factors, these methods are actually only hope for the future rather than for the present. In today's methodology, in all phases and aspects of design and functioning of contemporary automatic control systems, notably important is the correct identification of particular elements (especially an object), and later estimation of parameters and dependencies present there [1–3]. This refers essentially to the stages of preliminary analysis and defining the structure of a control system, synthesis of the control algorithm itself – with additional activities e.g. possible creation of observers and filters as well as prediction subsystems – and also future supervision of correct work in a real-time regime, in the frame of fault detection. A fundamental problem constitutes here the required accuracy – on one hand it should guarantee adequate representation of the modeled reality, and on the other it should not cause difficulty in actual use. In practice this is closely connected to the mathematical apparatus applied.

Generally, the simplest identification methods, closest to intuition and worthy of recommendation wherever possible, are deterministic methods. These however can not always be used, not just because phenomena by their nature are different in character, e.g. uncertain or imprecise, but even deterministic phenomena may have such complex or unknown structures, that artificial introduction of a nondeterministic factor may eventually occur just to describe such phenomena. The most common of these are probabilistic methods [4], well investigated and known, often with clear and suitable possibilities for interpretation.

The primary notion of probabilistic methods is the random variable, followed by its distribution. For simple applications often a sufficient representation of the distribution seems to be given by characteristic parameters (expectation value, variance, median, etc.), though in more complicated cases the frequent use of functional characteristics, e.g. density or distribution function, is necessary. The classic approach here is so-called parametric methods [4]. They are based on making an arbitrary choice, at the beginning, regarding distribution type (e.g. normal or uniform), in practice done with known properties of reality under consideration, the intuition of the researcher or preliminary investigation in this area, at times ratified by hypothesis testing. As a consequence of such a choice, only values of parameters existing in the definition of the assumed type of distribution are estimated – this is why such procedures are referred to as parametric methods. They are simple, easy to understand, widely available in subject literature, and robust to errors and inaccuracies, but their main limited possibilities and the need for preliminary investigations make them less and less acceptable from the point of view of contemporary refined applications. This has led directly to a necessity to find alternative procedures which

*e-mail: kulczycki@pk.edu.pl

do not need any assumptions concerning the type of distribution under research – to underline the difference, they are called nonparametric methods [5, 6]. This has become possible thanks to the rapid development of computer technology, which becomes particularly apparent in the domain of gathering and storing a large amount of data.

The subject of this publication is the presentation of applicational possibilities of nonparametric estimation, in particular based on the near intuitive, in practice often used functional characteristic of a random variable – the density of its distribution. Thus, an illustrative description of fundamental methods for nonparametric estimation of density is presented in Section 2 – their concepts, advantages and disadvantages will be shown here, after which will be quoted subject literature containing detailed aspects. The next two sections, 3 and 4, include a synthetic generalization of results obtained by the author during applications of the kernel estimators methodology – dominant in this type of tasks – to the representative problems of control engineering. Thus, Section 3 presents the use of kernel estimators methodology to calculate optimal – in the Bayes sense – values of parameters of automatic control objects, as an example of a subordinate factor with respect to the control algorithm. Finally, Section 4 describes a fault detection system, after considerations regarding the basic procedures for data analysis and exploration, as an example of a superior – with respect to such an algorithm – factor.

2. Nonparametric estimation of distribution density

This section presents an illustratory comparative analysis of basic methods of nonparametric estimation for the density of probability distribution. Such a characteristic of a random variable is not only convenient for interpretation and – therefore – comprehensive specialist applications, but also enables other characteristics of random variable distribution, both functional and parametric, to be examined. With regard to this far-reaching subject, numerous quotations from subject literature will be given, where one can find more exact aspects of the above tasks.

For simplicity of interpretation and denotation, the first considerations are presented for a one-dimensional random variable. Let then the random variable $X : \Omega \to \mathsf{R}$ be given, with distribution having the density f. Its estimator $\hat{f} : \mathsf{R} \to [0, \infty)$ is calculated on a simple random sample, i.e. the m experimentally obtained independent values x_1, $x_2, ..., x_m$ taken by the variable under investigation.

A trivial representative of nonparametric estimation methods is the histogram (Fig. 1) [7, 8]. Its idea is based on the division of the real numbers set into the bins H_k of identical width h, while the index k is an integer. Over each of these bins, the histogram has a constant value equal to the number of the values of the random sample $x_1, x_2, ..., x_m$ which have fallen into a given bin, divided by mh, therefore

$$\hat{f}(x) = \frac{\#\{x_i \in H_k\}}{mh} \text{ for every } x \in H_k \text{ and } k \text{ integer}, \quad (1)$$

where $\#A$ denotes a size of the set A. Nowadays the histogram is gradually becoming the only effective illustratory tool – even a layman is able to interpret results presented in this form. Unfortunately, however, there is no credible method for selecting the parameter h value or fixing the location of center of the bins, and the histogram's shape seems to be excessively sensitive to these quantities. The derivative of the histogram exists beyond points of contact of the bins, but it is constantly equal to zero there, which significantly hinders even the most basic theoretical analysis. For details see numerous publications, e.g. the textbooks [7, 8].

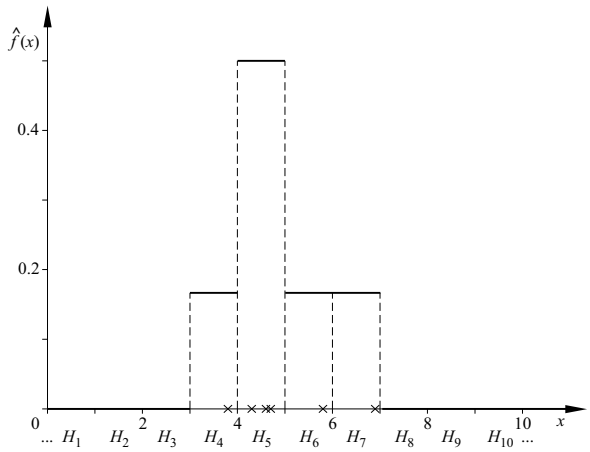

Fig. 1. Histogram

A search for further nonparametric methods provided the next more advanced proposals on different properties and practical usability.

Thus, an unusual idea led to the definition of the nearest neighborhood estimator (Fig. 2) [9, 10]. Its value can be given by the formula

$$\hat{f}(x) = \frac{k-1}{2m\, d_k(x)}, \quad (2)$$

where $d_k(x)$ denotes the distance of the argument x from its k-th nearest neighbor among the elements $x_1, x_2, ..., x_m$; often as the parameter $k \in \mathsf{N}\backslash\{0, 1\}$ the integral part of the number $\sqrt{m}$ is taken. This estimator is therefore a "conjunction" of hyperbolas, while in the places of these "joints" the derivative does not exist. Its graph is therefore irregular and unnatural in shape. What is more, the obvious – concerning the estimator of density of probability measure – condition

$$\int_{-\infty}^{\infty} \hat{f}(x)\,\mathrm{d}x = 1, \quad (3)$$

is not fulfilled, as $\int_{-\infty}^{\tilde{x}_{[1]}} \hat{f}(x)\,\mathrm{d}x$ and $\int_{\tilde{x}_{[m]}}^{\infty} \hat{f}(x)\,\mathrm{d}x$, where $\tilde{x}_{[i]}$ denotes the i-th with respect to size element of the set $x_1, x_2, ..., x_m$, are proportional to $\int_{-\infty}^{\tilde{x}_{[1]}} 1/(\tilde{x}_{[k]} - x)\mathrm{d}x$ and $\int_{\tilde{x}_{[m]}}^{\infty} 1/(x - \tilde{x}_{[m-k+1]})\mathrm{d}x$, that is they equal infinity. Even if

one narrows the considerations to a bounded interval, then calculation of an appropriate constant guaranteeing condition (3) is a difficult task to carry out in practice. For more information see the pioneering work [9] and also the book [10].

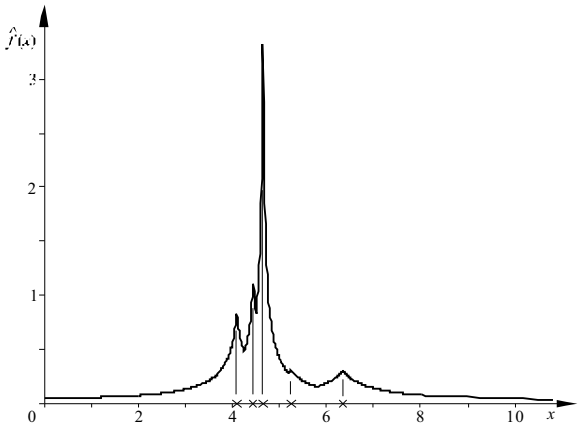

Fig. 2. Nearest neighbourhood estimator

In turn, the concept of the Fourier estimator (Fig. 3) [11] results directly from the general theory of Fourier transformation. Here it is possible to define the estimator only on the bounded interval $\mathsf{D} = [a, b]$. The Fourier estimator is then given by the formula

$$\hat{f}(x) = \frac{a_0}{2} + \sum_{j=1}^{J} [a_j \cos(j\omega x) + b_j \sin(j\omega x)], \tag{4}$$

while J is an appropriately fixed natural number and

$$a_j = \frac{2}{(b-a)m} \sum_{i=1}^{m} \cos(j\omega x_i) \quad \text{for} \quad j = 0, 1, 2, ..., J, \tag{5}$$

$$b_j = \frac{2}{(b-a)m} \sum_{i=1}^{m} \sin(j\omega x_i) \quad \text{for} \quad j = 1, 2, ..., J, \tag{6}$$

$$\omega = \frac{2\pi}{b-a}. \tag{7}$$

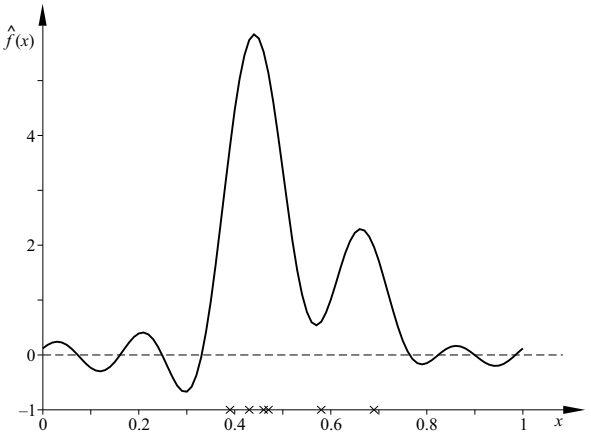

Fig. 3. Fourier estimator

Estimator (4) has a derivative for any order. Moreover, since $a_0 = 2/(b-a)$ and $\int_a^b \cos(j\omega x)\,\mathrm{d}x = \int_a^b \sin(j\omega x)\,\mathrm{d}x = 0$ for $j = 1, 2, ..., J$, then equality (3) is fulfilled. Unfortunately, this does not concern the obvious – concerning the estimator of density of probability measure – condition

$$\hat{f}(x) \geq 0 \quad \text{for every} \quad x \in \mathsf{R}; \tag{8}$$

the Fourier estimator can be negative in some subintervals of the domain D. Generalizing the Fourier estimator leads straight to the concept of orthogonal series estimators [12, 13], also defined in the case $\mathsf{D} = \mathsf{R}$. Maintaining the basic idea of a classic Fourier estimator, various changes in definition (4) are made to the sine/cosine functions, as well as procedures for calculation of coefficient values, arriving at a variety of estimator forms, of different properties and applicational possibilities. Further details are found in the classic work [12] and also the monographs [11, 13].

However a number of further concepts were proposed, from the simple naive estimator [14] to mathematically advanced splines [15], although up to formulating the kernel estimators concept, none of them satisfactorily fulfilled even the most basic theoretical or practical requirements.

Today, the prevalent method of nonparametric estimation is that of the kernel estimators [5–7, 16–20]. The idea of their construction is natural, the interpretations clear, and the form suitable for analysis. They were created at the end of the 1950's independently by Rosenblatt [14] and Parzen [21], and generalized for the multidimensional case by Cacoullos [22], but until the 80's they could be of interest to only a small group of specialists. Widespread research, and above all the application of kernel estimators, is impossible without computers of relatively high calculational capacity and the possibility to display results effectively – at least in the preliminary phase – on the screen.

Returning to the general n-dimensional case, let therefore the n-dimensional random variable $X : \Omega \to \mathsf{R}^n$, with a distribution having the density f, be given. Its kernel estimator $\hat{f} : \mathsf{R}^n \to [0, \infty)$ is defined in its basic form by the formula

$$\hat{f}(x) = \frac{1}{mh^n} \sum_{i=1}^{m} K\left(\frac{x - x_i}{h}\right), \tag{9}$$

where the measurable, symmetrical with respect to zero and having a weak global maximum in this point, function $K : \mathsf{R}^n \to [0, \infty)$ fulfils the condition $\int_{R^n} K(x)\,\mathrm{d}x = 1$ and is called a kernel, whereas the positive coefficient h is referred to as a smoothing parameter. In reference to properties of the estimators presented before, it should be underlined that conditions (3) and (8) are of course fulfilled here.

The interpretation of the above definition is illustrated in Fig. 4 for a one-dimensional random variable. In the case of the single realization x_i, the function K (transposed along the vector x_i and scaled by the coefficient h) represents the approximation of distribution of the random variable X having obtained the value x_i. For m independent realizations x_1, $x_2, ..., x_m$, this approximation takes the form of a sum of these

single approximations. The constant $1/mh^n$ enables the condition $\int_{R^n} \hat{f}(x)\,dx = 1$, required of the density of a probability distribution. For illustration of a more complex, multimodal and multidimensional ($n = 2$) random variable, see Fig. 5.

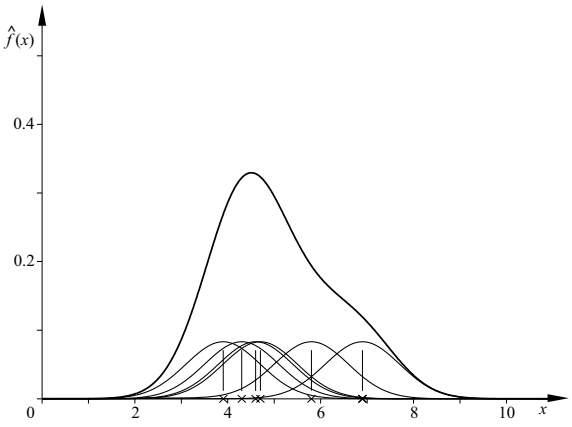

Fig. 4. Kernel estimator for the one-dimensional case

It is worth noting that a kernel estimator allows the modeling of density for practically every distribution, without arbitrary assumptions and most often any preliminary research. Atypical, complex distributions, also multimodal, are regarded here as textbook unimodal. It also allows the recognition of properties of a population described by an investigated random variable, in particular placement of modal values (i.e. local maximums of the density f), symmetries of particular associated components, as well as features of "tails" – properties of the function f for extreme values of the argument x. Furthermore, this information is most often obtained without additional, tiresome and ambiguous test procedures. In the multidimensional case kernel estimators also enable the discovery of total dependences between particular coordinates of the random variable under investigation.

Setting the quantities introduced in definition (9), 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 [5, 6, 19][1]. In particular, the choice of the kernel form has no practical meaning and thanks to this it is possible to take into account firstly properties of the estimator obtained (e.g. its class of regularity, boundary of a support) 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 [5 – Section 3.1; 6 – Section 5.3] and a linear transformation [5 – Section 3.1; 6 – Section 4.2], while for the second, the boundaries of a support [5 – Section 3.1; 6 – Section 2.10]. It is worth mentioning also the possibility of applying data compression and dimensionality reduction procedures – original and useful algorithms can be found e.g. in the book [23 – Sections 2 and 3.4].

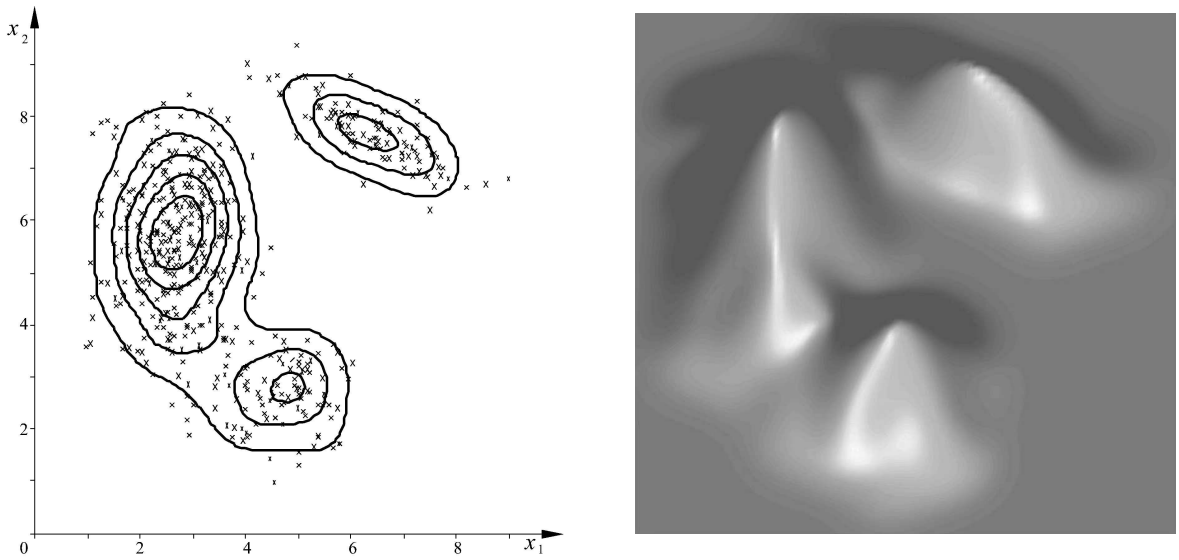

Fig. 5. Kernel estimator for the multidimensional case ($n = 2$)

[1]For calculating a smoothing parameter one can especially recommend the plug-in method in the one-dimensional case [5 – Section 3.1; 19 – Section 3.6], as well as the cross-validation method [5 – Section 3.1; 6, 19 – Section 3.6] in the multidimensional. Comments for the choice of kernel may best be found in [5 – Section 3.1, 19 – Sections 2.7 and 4.5].

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 so chosen that its primitive $I(x) = \int_{-\infty}^{x} K(y)\mathrm{d}y$ may be analytically obtained, then the estimator of the distribution function

$$\hat{F}(x) = \frac{1}{m}\sum_{i=1}^{m} I\left(\frac{x - x_i}{h}\right) \tag{10}$$

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

$$\hat{F}(x) = r \tag{11}$$

constitutes the kernel estimator of quantile of the order $r \in (0, 1)$. For details and proof of strong consistencies see the paper [24].

Polish science has had a sizable input into the progress of applications of nonparametric methods for control engineering and related fields, as well as in the broad range beyond the density estimation task presented earlier. Above all, mention should be made of the team from the Wroclaw University of Technology – Professors Wlodzimierz Greblicki, Zygmunt Hasiewicz, Adam Krzyzak (present of the Concordia University, Canada), Miroslaw Pawlak (present of the University of Manitoba, Canada), Ewaryst Rafajlowicz, with colleagues – and the research groups led by Prof. Jacek Koronacki (the Institute of Computer Sciences of the Polish Academy of Science, Warsaw), Prof. Leszek Rutkowski (the Czestochowa University of Technology), as well as the author of this article in the Cracow University of Technology and the Systems Research Institute of the Polish Academy of Science, Warsaw. Results have been published in many books and papers from renowned publishers and scientific journals. For Polish-speaking readers it is worth mentioning the works [5, 25, 26].

In following parts of this article, the applicational possibilities of nonparametric estimators of distribution density are shown for kernel estimators, as those which appear to posses the greatest universal practical potential. First will be presented results of investigations into the calculation of optimal values for parameters of automatic control object models, and next for synthesis of a statistical fault detection system.

3. Bayes parameter identification with asymmetrical loss function

Besides classic or trivial cases, the creation of an ideal model for an object under automatic control is neither possible, nor even required, as it would be far too complicated for effective use [1–3]. Consequently, absolutely precise determination of the values of parameters contained within is impossible, not only from a metrological point of view, but also due to the fact that such a value does not even exist, while a considered parameter represents an entire range of phenomena impossible to describe in a form of a single number. As identification is in practice always subject to a higher goal (usually conditioned by the control algorithm), then more suitable results can be obtained thanks to the consideration, in the estimation of the parameters' values, of the losses implied through errors encountered here. Often such losses can be described by the function assuming the following asymmetrical and polynomial form:

$$l(\hat{x},x) = \begin{cases} (-1)^k a\,(\hat{x} - x)^k & \text{for} \quad \hat{x} - x \le 0 \\ b\,(\hat{x} - x)^k & \text{for} \quad \hat{x} - x \ge 0 \end{cases}, \tag{12}$$

with $k \in \mathrm{N}\backslash\{0\}$, while the coefficients a and b are positive, and may differ, when x and $\hat{x}$ denote the parameter under investigation and its estimator respectively. Consider therefore the typical situation where one has the m values of the investigated parameter $x_1,\ x_2,\ \ldots\ ,\ x_m$, obtained by independent measuring, and requires the estimator which allows to obtain minimal potential losses. Three basic cases will be investigated in the following: linear (Section 3.1), quadratic (Section 3.2), and higher order polynomial (Section 3.3) – here the cube-case will be described in detail. In every case the final result will be an algorithm for the calculation of values for an optimal estimator, ensuring that its practical implementation does not demand of the user detailed knowledge of the theoretical aspects or laborious research. The results of numerical verification of the procedures investigated here are presented in Section 3.4.

First, however, the basic aspects of the decision theory, in particular in the Bayes approach [27], will be briefly described. Thus, the main aim of this theory is the selection of a concrete decision based only on a representation of measure characterizing the imprecision of states of nature. Let there be given the nonempty set of states of nature $Z = \mathrm{R}$, and the nonempty set of possible decisions $D \subset \mathrm{R}$. Assume that the imprecision of states of nature is of probability type and its distribution is described by the density $f : \mathrm{R} \to [0, \infty)$. Let there be given also the loss function $l{:}D \times Z \to \mathrm{R}$, while its values $l(d, z)$ can be interpreted as losses occurring in a hypothetical case, when the state of nature is z and the decision d is taken. If for every $d \in D$ the integral $\int_{\mathrm{R}} l(d, z)f(z)\mathrm{d}z$ exists, then the Bayes loss function $l_B : D \to \mathrm{R} \cup \{\pm\infty\}$ can be defined as

$$l_B(d) = \int_{\mathrm{R}} l(d,z)\, f(z)\, \mathrm{d}z. \tag{13}$$

Every element $d_B \in D$ such that $l_B(d_B) = \min_{d\in D} l_B(d)$ is called a Bayes decision, and the above procedure – a Bayes decision rule. The Bayes decision minimizes the mean value of losses following the decision d. Further details are found in the book [27].

3.1. Linear case. As an example illustrating the investigations presented in this section, an optimal control system [28, 29] will be considered. Such systems have shown themselves in practice to be sensitive to the inaccuracy of modelling. The control performance index which exists here, however, can also refer to quality of identification allowing the creation of an optimal procedure for the estimation of model parameter values, thereby notably lowering this sensitivity.

Thus, consider the following dynamic system:

$$\begin{bmatrix} \dot{X}_1(t) \\ \dot{X}_2(t) \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} X_1(t) \\ X_2(t) \end{bmatrix} + \begin{bmatrix} 0 \\ \dfrac{1}{M} \end{bmatrix} U(t), \quad (14)$$

where the positive parameter M represents a mass submitted to a force according to Newton's second law of dynamics. Then X_1, X_2 and U denotes position and velocity of the mass, and the force regarded here as a control, respectively. Such a system constitutes a basis for the majority of research in the field of robotics, leading in consequence to much more complex models, specifically suited to the particular problem under investigation. Consider the time-optimal control task, the basic form of which consists of bringing the system's state to the origin, in minimal and finite time, assuming the control values are bounded; for details see the textbook [28 – Section 7]. Fundamental meaning for phenomena existing in the control system lies in proper identification of value of the parameter M. The control is defined in relation to the value of the estimator $\hat{M}$, different in fact from the value of the parameter M in the object. Detailed analysis is found in the publications [30, 31].

Thus, in the purely hypothetical case of $\hat{M} = M$, i.e. when the value of the estimator of this parameter is equal to its true value, the process is regular in character. The system's state reaches the origin in minimal and finite time. However, in the event of underestimation (i.e. for $\hat{M} < M$), overregulations occur in the system – its state oscillates around the origin and reaches it in a finite time, albeit larger than the minimal. Next, in the case of overestimation (i.e. when $\hat{M} > M$), the system's state moves along a sliding trajectory and finally reaches the origin in a finite time, again larger than the minimal. Figure 6 shows the graph of the performance index for values of the estimator $\hat{M}$. One can note that an increase in this index is roughly proportional to the estimation error $|\hat{M} - M|$, although with different coefficients for positive and negative errors. The resulting losses can so be described in the form of an asymmetrical linear loss function, i.e. given by formula (12) with $k = 1$.

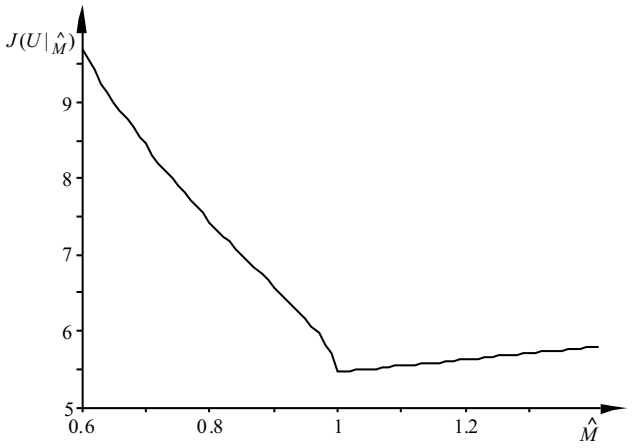

Fig. 6. Value of performance index J obtained for different values of the estimator $\hat{M}$, with $M = 1$

The parameter under investigation, whose value is to be estimated, will be denoted by x. In order to adhere to the principles of decision theory presented earlier at the beginning of Section 3, it will be treated here as the value of a random variable. According to point estimation methodology, it is assumed that the metrologically achieved measurements of the above parameter, i.e. $x_1, x_2, ..., x_m$, are the sum of its "true" (although unknown) value and random disturbances of various origin. The goal of this research is the calculation of the estimator of this parameter (hereinafter denoted by $\hat{x}$), which would approximate the "true" value – the best from the point of view of a practical problem investigated. In order to solve this task, the Bayes decision rule will be used, ensuring a minimum of expectation value of losses. According to the conditions formulated above, the loss function is assumed in asymmetrical linear form:

$$l(\hat{x}, x) = \begin{cases} -a\,(\hat{x} - x) & \text{for} \quad \hat{x} - x \leq 0 \\ b\,(\hat{x} - x) & \text{for} \quad \hat{x} - x \geq 0 \end{cases}, \quad (15)$$

while the coefficients a and b are positive and not necessarily equal to each other. Thus, the Bayes loss function (13) is given by the formula

$$l_B(\hat{x}) = b \int_{-\infty}^{\hat{x}} (\hat{x} - x) f(x)\, \mathrm{d}x - a \int_{\hat{x}}^{\infty} (\hat{x} - x) f(x)\, \mathrm{d}x, \quad (16)$$

where f:R $\rightarrow$ [0,∞) denotes the density of distribution of a random variable representing the uncertainty of states of nature, i.e. the parameter in question. It is readily shown that the function l_B fulfils its minimum for the value being a solution of the following equation with the argument $\hat{x}$:

$$\int_{-\infty}^{\hat{x}} f(x)\, \mathrm{d}x - \frac{a}{a+b} = 0. \quad (17)$$

Since $0 < a/(a+b) < 1$, a solution for the above equation exists, and if the function f has connected support, e.g. it is positive, this solution is unique. Moreover, thanks to equality

$$\frac{a}{a+b} = \frac{\dfrac{a}{b}}{\dfrac{a}{b} + 1}, \quad (18)$$

it is not necessary to identify the parameters a and b separately, rather only their ratio.

The modelling of the density f present in condition (17) will be carried out using statistical kernel estimators, presented in Section 2. Then one should choose a continuous kernel of positive values and so that the function I:R $\rightarrow$ R such that $I(x) = \int_{-\infty}^{x} K(y)\, \mathrm{d}y$ can be expressed by a relatively simple analytical formula. In consequence, this results in a similar property regarding the function U_i:R $\rightarrow$ R for any fixed $i = 1, 2, ..., m$ defined as

$$U_i(x) = \frac{1}{h} \int_{-\infty}^{x} K\left(\frac{y - x_i}{h}\right) \mathrm{d}y. \quad (19)$$

Then criterion (17) can be expressed equivalently in a form of

$$\frac{h}{m}\sum_{i=1}^{m} U_i(\hat{x}) - \frac{a}{(a+b)} = 0. \tag{20}$$

If the left side of the above formula is denoted by $L(\hat{x})$, its derivative is simply

$$L'(\hat{x}) = \hat{f}(\hat{x}), \tag{21}$$

where $\hat{f}$ was given by definition (9). In this situation, the solution of criterion (17) can be calculated numerically on the basis of Newton's algorithm [32] as the limit of the sequence $\{\hat{x}_j\}_{j=0}^{\infty}$ defined by

$$\hat{x}_0 = \frac{1}{m}\sum_{i=1}^{m} x_i, \tag{22}$$

$$\hat{x}_{j+1} = \hat{x}_j - \frac{L(\hat{x}_j)}{L'(\hat{x}_j)} \quad \text{for} \quad j = 0,\, 1,\, \ldots, \tag{23}$$

with the functions L and L' being given by formulas (20)–(21), whereas a stop criterion takes on the form

$$|\hat{x}_j - \hat{x}_{j-1}| \;\leq 0.01\;\hat{\sigma}, \tag{24}$$

where $\hat{\sigma}$ denotes the estimator of the standard deviation obtained from the sample $x_1,\, x_2, ..., x_m$.

3.2. Quadratic case. As an example to illustrate the reason for the case investigated below, consider the problem concerning the classical task of optimal control for a quadratic performance index [28 – Section 9.5] with infinite end time and unit matrix/parameter of the performance index. The object is the dynamic system

$$\begin{bmatrix} \dot{X}_1(t) \\ \dot{X}_2(t) \end{bmatrix} = \begin{bmatrix} \Lambda & 1 \\ 0 & \Lambda \end{bmatrix} \begin{bmatrix} X_1(t) \\ X_2(t) \end{bmatrix} + \begin{bmatrix} 0 \\ \Lambda \end{bmatrix} U(t), \tag{25}$$

where $\Lambda \in \mathsf{R}\backslash\{0\}$. Moreover, let $\hat{\Lambda} \in \mathsf{R}\backslash\{0\}$ represent an estimator of the parameter Λ. An optimal feedback controller is defined on the basis of the value $\hat{\Lambda}$, not necessarily equal to the value of the parameter Λ existing in the object. The values of the performance index obtained for a particular $\hat{\Lambda}$, are shown in Fig. 7. One can see that the resulting graph can be described with great precision by a quadratic function with different coefficients for positive and negative errors, which in fact proves that over- and underestimation of the parameter Λ have other results on the performance index value.

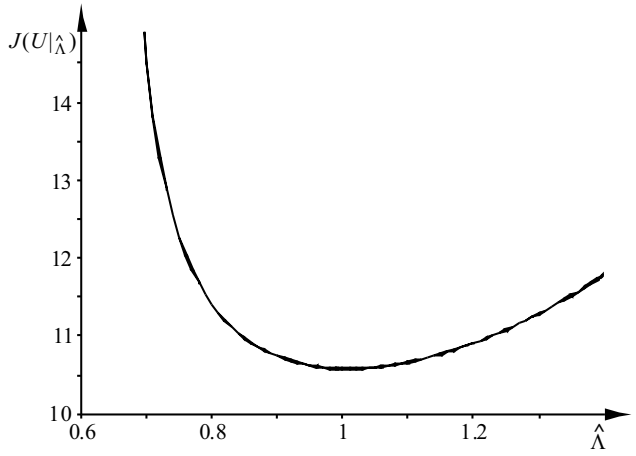

Fig. 7. Value of performance index J obtained for different values of the estimator $\hat{\Lambda}$, with $\Lambda = 1$

To use an analogous methodology to that of the linear case considered in the previous section, the loss function is assumed in quadratic and asymmetrical form defined as

$$l(\hat{x},x) = \begin{cases} a\,(\hat{x}-x)^2 & \text{for} \quad \hat{x}-x \leq 0 \\ b\,(\hat{x}-x)^2 & \text{for} \quad \hat{x}-x \geq 0 \end{cases}, \tag{26}$$

while the coefficients a and b are positive and not necessarily equal to each other. Thus, the Bayes loss function (13) is given by the formula

$$l_B(\hat{x}) = a\int_{\hat{x}}^{\infty} (\hat{x}-x)^2 f(x)\,\mathrm{d}x + b\int_{-\infty}^{\hat{x}} (\hat{x}-x)^2 f(x)\,\mathrm{d}x. \tag{27}$$

One can show that the function l_B fulfils its minimum for the value $\hat{x}$ being a solution of the equation

$$(a-b)\int_{-\infty}^{\hat{x}} (\hat{x}-x) f(x)\,\mathrm{d}x - a\int_{-\infty}^{\infty} (\hat{x}-x) f(x)\,\mathrm{d}x = 0. \tag{28}$$

This solution exists and is unique. As in the linear case, dividing the above equation by b, note that it is necessary to identify only the ratio of the parameters a and b.

Solution of Eq. (28) for a general case is not an easy task. However, if estimation of the density f is reached using statistical kernel estimators, then – thanks to a proper choice of the kernel form – one can design an effective numerical algorithm to this end. Let, therefore, a continuous kernel of positive values, fulfilling the condition

$$\int_{-\infty}^{\infty} xK(x)\,\mathrm{d}x < \infty \tag{29}$$

be given. Besides the functions U_i introduced in Section 3.1, let for any fixed $i = 1, 2, ..., m$ the functions V_i:R $\rightarrow$ R be defined as

$$V_i(x) = \frac{1}{h}\int_{-\infty}^{x} yK\left(\frac{y-x_i}{h}\right)\,\mathrm{d}y. \tag{30}$$

The kernel K should be chosen so the function J:R $\rightarrow$ R such that $J(x) = \int\limits_{-\infty}^{x} y\,K(y)\,\mathrm{d}y$ be expressed by a convenient analytical formula. If an expected value is estimated by the arithmetical mean value of a sample, then criterion (28) can be described equivalently as

$$\sum_{i=1}^{m} [(a-b)(\hat{x}U_i(\hat{x}) - V_i(\hat{x})) + ax_i] - a\hat{x}m = 0. \tag{31}$$

If the left side of the above formula is denoted by $L(\hat{x})$, then one can express the value of its derivative as

$$L'(\hat{x}) = \sum_{i=1}^{m} [(a-b)\,U_i(\hat{x})] - am. \tag{32}$$

In this situation, the solution of criterion (28) can be calculated numerically on the basis of Newton's algorithm (22)–(24).

3.3. Higher order polynomial case. In this section, detailed investigations presented earlier will be supplemented with the polynomial case, that is where the loss function is an asymmetrical polynomial of the order $k \geq 2$ and is therefore given by the following formula:

$$l(\hat{x},x) = \begin{cases} (-1)^k a\,(\hat{x}-x)^k & \text{for} \quad \hat{x}-x \leq 0 \\ b\,(\hat{x}-x)^k & \text{for} \quad \hat{x}-x \geq 0 \end{cases}, \tag{33}$$

while the coefficients a and b are positive, and may differ. Criterion for the optimal estimator $\hat{x}$ is given here in the form

$$\begin{aligned} (-1)^k ak \int\limits_{\hat{x}}^{\infty} (\hat{x}-x)^{k-1} f(x)\,\mathrm{d}x \\ + bk \int\limits_{-\infty}^{\hat{x}} (\hat{x}-x)^{k-1} f(x)\,\mathrm{d}x = 0. \end{aligned} \tag{34}$$

The solution of the above equation exists and is unique.

When the statistical kernel estimators are used with respect to the density f, it is possible again to create an efficient numerical algorithm enabling Eq. (34) to be solved. Let the kernel K be continuous, of positive values and fulfilling the following condition:

$$\int\limits_{-\infty}^{\infty} x^{k-1} K(x)\,\mathrm{d}x < \infty. \tag{35}$$

For clarity of presentation, the case $k = 3$ is presented below. Thus, Eq. (34), after simple transformations, takes on the equivalent form

$$\begin{aligned} (a+b)\left(\hat{x}^2 \int\limits_{-\infty}^{\hat{x}} f(x)\,\mathrm{d}x - 2\hat{x} \int\limits_{-\infty}^{\hat{x}} x f(x)\,\mathrm{d}x + \int\limits_{-\infty}^{\hat{x}} x^2 f(x)\,\mathrm{d}x\right) \\ - a\left(\hat{x}^2 - 2\hat{x} \int\limits_{-\infty}^{\infty} x f(x)\,\mathrm{d}x + \int\limits_{-\infty}^{\infty} x^2 f(x)\,\mathrm{d}x\right) = 0. \end{aligned} \tag{36}$$

Now, with any fixed $i = 1, 2, \ldots, m$, let the functions U_i and V_i defined by dependencies (19) and (30) be given, and furthermore W_i:R $\rightarrow$ R be introduced as

$$W_i(x) = \frac{1}{h} \int\limits_{-\infty}^{x} y^2 K\left(\frac{y-x_i}{h}\right)\,\mathrm{d}y. \tag{37}$$

Making use of the above notations, condition (36) can be expressed in the following form:

$$\begin{aligned} \sum_{i=1}^{m} \Big[(a+b)\left(x^2 U_i(x) - 2xV_i(x) + W_i(x)\right) + 2ax_i x \\ - \lim_{z\to\infty} W_i(z)\Big] - amx^2 = 0. \end{aligned} \tag{38}$$

The solution of the above equation exists and is unique. If its left-hand side is denoted as $L(x)$, then the derivative is

$$L'(x) = \sum_{i=1}^{m} [2(a+b)\,(xU_i(x) - V_i(x)) + 2ax_i] - 2amx. \tag{39}$$

Finally, the desired estimator can be calculated numerically through Newton's algorithm (22)–(24), while the functions L and L' are given by formulas (38)–(39). The above investigations can be analogously transposed to a higher order of asymmetrical polynomial loss function (12), although on account of their extreme nature, they seem to be useful mainly for atypical applicational tasks.

3.4. Numerical simulation results. The operation of the algorithm designed here has been checked in detail using a numerical simulation, also for the optimal control tasks considered as motivation in Sections 3.1 and 3.2. In the case $a = b$, the results were close to medium value, however, when $a \neq b$, the algorithm provided possibilities that cannot be achieved using classical methods, by appropriately shifting the value of the estimator in the direction associated with smaller losses, where intensity of this process was stimulated by the parameter k depending on the nature of the system under research. Many different distributions were examined including also multimodal with asymmetrical modes. In each case, as the size of a random sample m increases, the mean estimation error and its standard deviation tend to zero. From an applicational point of view, these fundamental properties are demanded of estimators used in practice. This above all states that, as the sample size increases, the estimators' values achieved tend to the desired value, and their dispersion decreases. This allows for the obtaining of any required precision, although the proper sample size must be guaranteed. In practice this implies a necessity for compromise between these two quantities. A satisfactory degree of precision was obtained when the size of the sample was between 10 and 200, i.e. for $m \in [10, 200]$; in particular, the bigger values became necessary when the difference between parameters a and b increased.

One may construe that the benefits arising from application of the method presented here are greater the more complex the control system is, and over- and underestimation of

a model's parameters have a more differing influence on performance index, i.e. when asymmetry of the loss function is more distinct.

This section also contains material worked on together with Malgorzata Charytanowicz and Aleksander Mazgaj, included in the common publications [33–36].

4. Fault detection

The task of fault detection and diagnosis has lately become one of the most important challenges in modern control engineering [37–39]. Although it plays a superior role in the hierarchy of layers of a control system, from the perspective of its total utility it has proven most advantageous to adapt the methodology used in this respect to the conditions prevailing in the lower layers, in particular the control algorithm. The result in practice is an enormous, indeed excessive diversity of concepts used in the design of fault detection systems. Among 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.

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

- detection, so discovery of the existence of potential anomalies in the technical state of a supervised device;
- 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 using kernel estimators methodology. First, Section 4.1 presents possible applications of kernel estimators 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 4.2, where the fault detection system designed here is described, will consist mainly of references to earlier material, and integrate them into one coherent idea. Results of numerical verification are described in Section 4.3.

4.1. Kernel estimators for data analysis and exploration procedures. The application of kernel estimators in basic tasks of data analysis and exploration (for an original approach see also [40]) 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 $X : \Omega \to \mathsf{R}^n$ is 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 [41]. The significance test based on the kernel estimators methodology will now be described [42].

Let therefore the random sample $x_1, x_2, \ldots, x_m$ treated as representative, and so 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 $\tilde{x} \in \mathsf{R}^n$ is a typical element will be tested against the hypothesis that it is not, and therefore should be treated as an outlier. The statistic $S : \mathsf{R}^n \to [0, \infty)$, used here, can be defined by

$$S(\tilde{x}) = \hat{f}(\tilde{x}), \tag{40}$$

where $\hat{f}$ denotes a kernel estimator of density obtained for the random sample $x_1, x_2, \ldots, x_m$ mentioned above, while the critical set takes the left-sided form $A = (-\infty, \hat{q}]$ when $\hat{q}$ constitutes the kernel estimator of quantile of the order r (for its description see the end of Section 2) calculated for the sample $\hat{f}(x_1), \hat{f}(x_2), \ldots, \hat{f}(x_m)$, with the assumption that random variable support is bounded (see also Section 2) to nonnegative numbers. Further details can be found in the publication [42].

Secondly, the aim of clustering is the division of a data set – for example given in the form of the random sample x_1, $x_2, \ldots, x_m$ – into subgroups (clusters), with every one including elements "similar" to each other, but with significant differences between particular subgroups [43, 44]. 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 [45, 46] based on kernel estimators, taking advantage of the gradient methods concept [47] will be presented now.

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

$$x_j^0 = x_j \quad \text{for} \quad j = 1, 2, ..., m, \tag{41}$$

$$x_j^{k+1} = x_j^k + b\frac{\nabla \hat{f}(x_j^k)}{\hat{f}(x_j^k)} \quad \text{for} \quad j = 1, 2, ..., m \quad \text{and} \quad k = 0, 1, \ldots, \tag{42}$$

where $b > 0$ and ∇ denotes a gradient. 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

$$|D_k - D_{k-1}| \leq c\, D_0, \tag{43}$$

where $c > 0$, $D_0 = \sum_{i=1}^{m} \sum_{j=i+1}^{m} d(x_i, x_j)$, $D_{k-1} = \sum_{i=1}^{m} \sum_{j=i+1}^{m} d(x_i^{k-1}, x_j^{k-1})$ and $D_k = \sum_{i=1}^{m} \sum_{j=i+1}^{m} d(x_i^{k}, x_j^{k})$, i.e. they are the sums of the distances d between particular elements of the random sample under consideration before the beginning of algorithm (41)–(42) and having performed the $(k-1)$-th and k-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 $x_1^{k^*}, x_2^{k^*}, \ldots, x_m^{k^*}$ (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 enables it to be suited to a true data structure. Moreover, possible changes in intensity of the smoothing parameter modification procedure allow 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 [45, 46].

Thirdly, the application of kernel estimators in a classification task [43, 44] is considered. Let the number $J \in \mathsf{N}\backslash\{0,1\}$ be given. Assume also, that the possessed random sample $x_1, x_2, \ldots, x_m$ has been divided into J nonempty and separate subsets $\{x_1', x_2', ..., x_{m_1}'\}$, $\{x_1'', x_2'', ..., x_{m_2}''\}$, ... , $\{x_1^{''\ldots'}, x_2^{''\ldots'}, ..., x_{m_J}^{''\ldots'}\}$, 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 $\tilde{x} \in \mathsf{R}^n$ should be reckoned.

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 $\hat{f}_1, \hat{f}_2, ..., \hat{f}_J$ denote kernel estimators of density calculated for subsets $\{x_1', x_2', ..., x_{m_1}'\}$, $\{x_1'', x_2'', ..., x_{m_2}''\}$, ... , $\{x_1^{''\ldots'}, x_2^{''\ldots'}, ..., x_{m_J}^{''\ldots'}\}$, respectively, treated here as samples. If sizes $m_1, m_2, \ldots, m_J$ 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

$$m_1 \hat{f}_1(\tilde{x}),\ m_2 \hat{f}_2(\tilde{x}), ..., m_J \hat{f}_J(\tilde{x}) \tag{44}$$

is the greatest. Some additional information can be found in the publication [48].

4.2. Statistical fault detection system. The procedures presented in the previous section, for recognition of atypical elements (outliers), clustering and classification, based on kernel estimators, provides a complete and methodologically consistent mathematical tool to design an effective statistical 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 $x \in \mathsf{R}^n$, 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

$$x_1,\ x_2, ..., x_{m_0}, \tag{45}$$

as well as the set

$$x_1^*,\ x_2^*, ..., x_{m_*}^*, \tag{46}$$

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 (46) into $J \in \mathsf{N}\backslash\{0,1\}$ 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:

$$x_1',\ x_2', ..., x_{m_1}', \tag{47}$$

$$x_1'',\ x_2'', .., x_{m_2}'', \tag{48}$$

$$\vdots$$

$$x_1^{''\ldots'},\ x_2^{''\ldots'}, ..., x_{m_J}^{''\ldots'}, \tag{49}$$

while $\sum_{j=1}^{J} m_j = m_*$. Where there is no such division, one can automatically divide set (46) into subsets (47)–(49) using the clustering algorithm presented in Section 4.1, 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 4.1, can be applied. Assume therefore that the random sample considered there, including elements treated as typical, constitutes set (45) representing the correct functioning conditions for a supervised device, while $\tilde{x}$ denotes its current state. Applying the above-mentioned 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 (47)–(49) characterizing particular types of faults being diagnosed, then after the above-described detection of anomalies, one can – by applying directly the procedure for Bayes classification presented at the end of Section 4.1 – infer which of them is being dealt with.

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 (40) and $m_1\hat{f}_1,\ m_2\hat{f}_2, ..., m_J\hat{f}_J$ to be seen in formula (44), 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 $\hat{f},\ \hat{f}_1,\ \hat{f}_2,\ \ldots\ , \hat{f}_J$ 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 [39 – Chapter 4; 49 – Chapter 3 and additionally Chapter 4].

4.3. Numerical simulation results. The proper operation of the fault detection system investigated in this section 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 whose dynamics were modeled by the differential inclusion

$$\ddot{y}(t) \in H(\dot{y}(t)) + u(t), \tag{50}$$

where y expresses the position of the object, u is a control with values limited to the interval $[-1, 1]$, and the function H represents a multi-valued discontinuous model of resistance to motion. In the event of no resistance to motion, i.e. when $H \equiv 0$, inclusion (50) is reduced to a differential equation 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 (50) 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 is located; for details see the papers [50, 51]. The symptom vector was assumed as a 3-dimensional vector whose coordinates designate the absolute values of control, resistance to motion, and velocity, i.e. $|u(\,\cdot\,)|$, $|H(\,\cdot\,)|$, and $|\dot{y}(\,\cdot\,)|$, 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 above 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 transgressed 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 unfortunate direction of its evolution.

This section also contains material included in the articles [52, 53] as well as worked on together with Malgorzata Charytanowicz, Karina Daniel, Piotr A. Kowalski, and Cyprian Prochot, described in the common publications [43, 46–48, 54].

5. Summary and final comments

In this paper a problem from the area of applications of nonparametric estimation for control engineering tasks was considered. The subject of research was limited to the task of estimation of density of distribution – a convenient and often used basic functional characteristic of a random variable. Firstly an illustratory comparative analysis of most common methods was carried out. Next the results were generalized and synthetically presented for many years of author's research into the application of the most useful – for this purpose – kernel estimators, in particular to the tasks of optimal parameter identification, as well as a fault detection system, after considerations regarding the basic procedures for data analysis and exploration: the recognition of outliers, clustering and classification. The material collected here point to the great possibilities of application of this methodology – convenient for interpretation and implementing – in tasks of widely understood automatic control, as well as experimental research confirmed its suitability.

Kernel estimators were also the subject of applicational investigations in practical problems of general systems analysis, among others the demand-based design of an optimal base-station system of wireless data transmission LMDS [55] and the design of a marketing support strategy for a mobile phone operator [54]. They were also successfully used for statistical inference [56] and data exploration procedures [57]. These cases, however, go beyond the assumed scope of this article.

The main ideas shown above were briefly described as the conference-paper [58]. This work also contains material widely presented in the book [5].

Acknowledgements. The following publication also includes material of research in the field of kernel estimators carried out with my junior colleagues Malgorzata Charytanowicz, Ka-

rina Daniel, Piotr A. Kowalski, Szymon Lukasik, Aleksander Mazgaj, Cyprian Prochot, and Jacek Waglowski.

It is a particular honour to make the above acknowledgments in a publication dedicated to Prof. Henryk Górecki. As an author of this text, I myself was already instructed by his pupils – among others Prof. Wojciech Mitkowski and Prof. Ryszard Tadeusiewicz. My junior colleagues therefore belong to the generation of scientific great-grandchildren of Prof. Henryk Górecki. I write this dedication in the hope that this publication has allowed us to express utmost respects and gratitude to him for giving us a new way of looking at the world of science and research, distinctive for his own Cracovian school of control engineering.

REFERENCES

[1] F. Morrison, *The Art of Modeling Dynamic Systems*, Multiscience Press, New York, 1991.
[2] H.E. Nusse and J.A. Yorke, *Dynamics: Numerical Exploration*, Springer-Verlag, New York, 1997.
[3] T.S. Soderstrom and P. Stoica, *System Identification*, Prentice Hall, Englewood Cliffs, 1994.
[4] R.L. Scheaffer and J.T. McClave, *Probability and Statistics for Engineers*, Duxbury, Boston, 1990.
[5] P. Kulczycki, *Kernel Estimators in Systems Analysis*, WNT, Warsaw, 2005, (in Polish).
[6] B.W. Silverman, *Density Estimation for Statistics and Data Analysis*, Chapman and Hall, London, 1986.
[7] D.W. Scott, *Multivariate Density Estimation*, Wiley, New York, 1992.
[8] J.W. Tukey, *Exploratory Data Analysis*, Addison-Wesley, Reading, 1977.
[9] D.O. Loftsgaarden and C.P. Quesenberry, "A nonparametric estimate of a multivariate density function", *Annals of Mathematical Statistics* 36, 1049–1051 (1965).
[10] A. Pagan and A. Ullah, *Nonparametric Econometrics*, Cambridge University Press, Cambridge, 1999.
[11] D.L. Kreider, R.G. Kuller, D.R. Ostberg, and F.W. Perkins, *Introduction to the Linear Analysis*, Addison-Wesley, Reading, 1966.
[12] N.N. Cencow, "Evaluation of an unknown distribution density from observations", *Soviet Mathematics* 3, 1559–1562 (1962).
[13] W. Greblicki and M. Pawlak, *Nonparametric System Identification*, Cambridge University Press, Cambridge, 2008.
[14] M. Rosenblatt, "Remarks on some nonparametric estimates of a density function", *Annals of Mathematical Statistics* 27, 832–837 (1956).
[15] R.L. Eubank, *Spline Smoothing and Nonparametric Regression*, Dekker, New York, 1988.
[16] L. Devroye and L. Gyorfi, *Nonparametric Density Estimation: The L_1 View*, Wiley, New York, 1985.
[17] E.A. Nadaraya, *Nonparametric Estimation of Probability Density Function and Regression Curves*, Kluver, Dordrecht, 1989.
[18] B.L.S. Prakasa Rao, *Nonparametric Functional Estimation*, Academic Press, Orlando, 1983.
[19] M.P. Wand and M.C. Jones, *Kernel Smoothing*, Chapman and Hall, London, 1994.
[20] L. Wasserman, *All of Nonparametric Statistics*, Springer, New York, 2006.
[21] E. Parzen, "On estimation of a probability density function and mode", *Annals of Mathematical Statistics* 33, 1065–1076 (1962).
[22] T. Cacoullos, "Estimation of a multivariate density", *Annals of the Institute of Statistical* Mathematics 18, 179–189 (1966).
[23] S.K. Pal and P. Mitra, *Pattern Recognition Algorithms for Data Mining*, Chapman and Hall, London, 2004.
[24] P. Kulczycki and A.L. Dawidowicz, "Kernel estimator of quantile", *Universitatis Iagellonicae Acta Mathematica* XXXVII, 325–336 (1999).
[25] W. Greblicki, "Nonparametric systems identification", *Archives of Control Sciences* 36, 277–290 (1991), (in Polish).
[26] Z. Hasiewicz and P. Sliwinski, *Orthogonal Wavelets with Compact Support. Application in Nonparametric Systems Identification*, EXIT, Warsaw, 2005, (in Polish).
[27] J.O. Berger, *Statistical Decision Theory*, Springer-Verlag, New York, 1980.
[28] M. Athans and P.L. Falb, *Optimal Control*, McGraw-Hill, New York, 1966.
[29] H. Gorecki, *Optimization of Dynamic Systems*, PWN, Warsaw, 1993, (in Polish).
[30] P. Kulczycki and R. Wisniewski, "Fuzzy Controller for a System with Uncertain Load", *Fuzzy Sets and Systems* 131, 185–195 (2002).
[31] P. Kulczycki, R. Wisniewski, P.A. Kowalski, and K. Krawiec, "Hard and soft sub-time-optimal controllers for a mechanical system with uncertain mass", *Control and Cybernetics* 33, 573–587 (2004).
[32] D. Kincaid and W. Cheney, *Numerical Analysis*, Brooks/Cole, Pacific Grove, 2002.
[33] P. Kulczycki and M. Charytanowicz, "Asymmetrical conditional Bayes parameter identification for control engineering", *Cybernetics and Systems* 38, 229–243 (2008).
[34] P. Kulczycki and M. Charytanowicz, "Automatic control tasks featuring asymmetrical conditional identification of parameters", *Proc. 13th IEEE/IFAC Int. Conf. on Methods and Models in Automation and Robotics* 0236, CD-ROM (2007).
[35] P. Kulczycki and A. Mazgaj, "Bayes parameter identification with polynomial asymmetrical loss function", *Proc. 17th IFAC World Congress* 12395, CD-ROM (2008).
[36] P. Kulczycki and A. Mazgaj, "Parameter identification for asymmetrical polynomial loss function", (to be published).
[37] J.J. Gertler, *Fault Detection and Diagnosis in Engineering Systems*, Dekker, New York, 1998.
[38] J. Korbicz, "Robust fault detection using analytical and soft computing methods", *Bull. Pol. Ac.: Tech.* 54 (1), 75–88 (2006).
[39] P. Kulczycki, *Fault Detection in Automated Systems Using Statistical Methods*, Alfa, Warsaw, 1998, (in Polish).
[40] R. Tadeusiewicz and R. Ogiela, "Structural approach to medical image understanding", *Bull. Pol. Ac.: Tech.* 52 (2), 131–139 (2004).
[41] V. Barnett and T. Lewis, *Outliers in Statistical Data*, Wiley, Chichester, 1994.
[42] P. Kulczycki and C. Prochot, "Detection of outliers by nonparametric estimation method", in: *Operation and Systems Research: Decision Making – Theoretical Basics and Applications*, pp. 313–328, eds. R. Kulikowski, J. Kacprzyk, and R. Slowinski, EXIT, Warsaw, 2004, (in Polish).
[43] J.H. Hand, H. Mannila, and P. Smyth, *Principles of Data Mining*, MIT Press, Cambridge, 2001.
[44] D.T. Larose, *Discovering Knowledge in Data. An Introduction to Data Mining*, Wiley, New York, 2005.
[45] P. Kulczycki and M. Charytanowicz, "A complete gradient clustering algorithm", in: *Control and Automation: Current*

Problems and Their Solutions, pp. 312–321, eds. K. Malinowski and L. Rutkowski, EXIT, Warsaw, 2008, (in Polish).

[46] P. Kulczycki and M. Charytanowicz, "A complete gradient clustering algorithm formed with kernel estimators", (to be published).

[47] K. Fukunaga and L.D. Hostetler, "The estimation of the gradient of a density function, with applications in Pattern Recognition", *IEEE Trans. on Information Theory* 21, 32–40 (1975).

[48] P. Kulczycki and P.A. Kowalski "Classification of imprecise information of interval type with reduced model samples", in: *Operation and Systems Research: Environment, Space, Optimization*, pp. 305–314, eds. O. Hryniewicz, A. Straszak, and J. Studzinski, IBS PAN, Warsaw, 2008, (in Polish).

[49] B. Abraham and J. Ledolter, *Statistical Methods for Forecasting*, Wiley, New York, 1983.

[50] P. Kulczycki, "A random approach to time-optimal control", *J. Dynamic Systems, Measurement, and Control* 121, 542–543 (1999).

[51] P. Kulczycki, "Fuzzy controller for mechanical systems", *IEEE Transactions on Fuzzy Systems* 8, 645–652 (2000).

[52] P. Kulczycki, "Data anlaysis using kernel estimators for systems diagnosis", in: *Processes and Systems Diagnosis*, pp. 231–238, eds. J. Korbicz, K. Patan, and M. Kowal, EXIT, Warsaw, 2007, (in Polish).

[53] P. Kulczycki, "Statistical kernel estimators for design of fault detection, diagnosis, and prognosis system", *The Open Cybernetics and Systemics Journal* 2, 180–184 (2008).

[54] P. Kulczycki and K. Daniel, "An algorithm to support marketing strategy for a mobile phone operator", in: *Operation and Systems Research 2006: Methods and Techniques*, pp. 245–256, eds. J. Kacprzyk and R. Budzinski, EXIT, Warsaw, 2006.

[55] P. Kulczycki and R. Waglowski, "On the application of statistical kernel estimators for the demand-based design of a wireless data transmission system", *Control and Cybernetics* 34, 1149–1167 (2005).

[56] P. Kulczycki, "A test for comparing distribution functions with strongly unbalanced samples", *Statistica* LXII, 39–49 (2002).

[57] S. Lukasik, P.A. Kowalski, M. Charytanowicz, and P. Kulczycki, "Fuzzy model identification using kernel-density-based clustering", in: *Development in Fuzzy Sets, Intuitionistic Fuzzy Sets, Gemeralized Nets and Related Topics. Applications*, vol. 2, pp. 135–146, eds. K. Atanassov, P. Chountas, J. Kacprzyk, M. Krawczak, P. Melo-Pinto, E. Szmidt, and S. Zadrozny, EXIT, Warsaw, 2008.

[58] P. Kulczycki, "Nonparametric estimation for control engineering", *Proc. 4th WSEAS/IASME Int. Conf. on Dynamical Systems and Control*, 115–121 (2008).