KYBERNETIKA — VOLUME *38* (2002), NUMBER 2, PAGES 141 – 168

STATISTICAL INFERENCE FOR FAULT DETECTION: A COMPLETE ALGORITHM BASED ON KERNEL ESTIMATORS

Piotr Kulczycki

This article presents a new concept for a statistical fault detection system, including the detection, diagnosis, and prediction of faults. Theoretical material has been collected to provide a complete algorithm making possible the design of a usable system for statistical inference on the basis of the current value of a symptom vector. The use of elements of artificial intelligence enables self-correction and adaptation to changing conditions. The mathematical apparatus is founded on the methodology of testing statistical hypotheses, and on kernel estimators; the theoretical aspects have been documented by mathematical theorems. The work is oriented towards the problem of fault detection in dynamic systems under automatic control, but the basic formula is of a universal nature and can be used in a broad range of applications, including those outside the scope of engineering.

1. INTRODUCTION

The increasing capabilities and universality of the computer installations used in contemporary technical devices have created conditions conducive to the rapid development of methods to detect faults appearing in systems working in the real-time regime. Indeed, it can now safely be stated that the construction of more and more robust systems, with an increased level of security against the consequences of any breakdowns that may occur, is the third stage – following the classic feedback control technique and optimal control – in the development of automatic control engineering. Although fault detection plays a superordinate role in the hierarchy of the individual layers of control, from the perspective of total system utility it has proven most advantageous to adapt the methodology used in this respect to the conditions prevailing in the lower layers. The result in practice is an enormous, indeed excessive heterogeneity of concepts used in the design of fault detection systems [2, 5, 9, 14, 18, 24]. Among the most universal are statistical methods. These very often consist in generating a certain group of variables that characterize the state of technical performance of the device, and then making a statistical inference, on the basis of their current values, as to whether or not the device is working properly, and in the event of a negative response, what is the nature of the malfunction.

The present paper provides all the material needed to design a universal statistical inference system, including:

1. detection of faults, i. e. the discovery that a malfunction has occurred in the system under supervision;
2. diagnosis of faults, which means the recognition of the malfunction;
3. prediction of faults, referring to the anticipation of the risk that a malfunction will occur in the immediate future (along with its presumed classification).

The idea for the system to be proposed here is based on the procedures of mathematical statistics, with particular emphasis on the methodology of kernel estimators. Elements of artificial intelligence are also used, based on the neural networks technique, for purposes of self-correction and adaptation. The following sections provide a full set of formulas defining the structure and the detailed form of the functions and parameter values, which consequently enable the efficient design of a usable fault detection system, without the need for laborious particularized research.

2. FORMULATION OF THE PROBLEM

In the statistical inference system here investigated, one assumes the successive availability for measurement of the so-called symptom vector Z, that is, a finite number of variables whose current values and/or relations among its individual coordinates would be dependent on the technical state of the system under supervision. It is a prerequisite, then, that both proper operating conditions and any type of diagnosed fault be associated with the most highly diversified value sets and/or relations among the coordinates of this vector. (The detailed form of these sets and relations need not be known *a priori*: its identification is an integral part of the procedure here proposed.) The particular coordinates of the symptom vector Z may be coordinates of control, state and response, the current values of the measurable parameters, and a range of other quantities that are characteristic for the given device (e. g. its output capacity, temperature, fuel consumption, etc.), as well as their functions (e. g. differences, powers, etc.).

The current measurements of the values of the symptom vector Z will be the basis for statistical inference, conducted by testing the hypotheses:

$$H_0 \text{ -- proper system operation} \tag{1}$$

in the event of detection, as well as

$$H_k \text{ -- the occurrence of the } k\text{th type of fault} \tag{2}$$

for the diagnosis of a finite number ($k = 1, 2, \ldots, d$) of its differentiable types. It should be emphasized that the range of malfunctions that can be discovered during detection is not limited to the set of faults subject to diagnosis. The following assumptions are accepted:

(A) Z denotes an n-dimensional discrete stochastic process defined on the probability space (Ω, Σ, P) and the set $\mathbb{N} \setminus \{0\}$;

(B) in either case, whether under proper operating conditions or when any type of fault to be diagnosed is occurring, the process Z has stationary one-dimensional distributions (though they may be different for each case);

(C) the bounded mapping $f_0 : \mathbb{R}^n \to [0, \infty)$ of the class $\mathcal{C}^2$, with bounded second derivative, constitutes the density function of the distribution of the random variables $Z(\cdot, j)$ for $j \in \mathbb{N} \setminus \{0\}$, when the system is operating properly;

(D) the bounded mapping $f_k : \mathbb{R}^n \to [0, \infty)$ $(k = 1, 2, \ldots, d)$ of the class $\mathcal{C}^2$, with bounded second derivatives, represents the density function of the distribution of the random variables $Z(\cdot, j)$ for $j \in \mathbb{N} \setminus \{0\}$, whenever the kth type of fault to be diagnosed occurs.

Assumption (B) implies that for every $j \in \mathbb{N} \setminus \{0\}$ the random variables $Z(\cdot, j)$ have identical distributions. Moreover, if $f : \mathbb{R}^n \to \mathbb{R}$ denotes a Borelian function, then this fact also concerns the stochastic process $Y \equiv f \circ Z$, i. e. for any $j \in \mathbb{N} \setminus \{0\}$ the distributions of the random variables $Y(\cdot, j)$ are the same.

3. A REVIEW OF THE ESTIMATORS APPLIED

3.1. Kernel estimator of density function

Let the n-dimensional random variable X be given, whose distribution has the density function f. Its kernel estimator $\hat{f} : \mathbb{R}^n \to [0, \infty)$ is calculated on the basis of the m-element simple random sample $x_1, x_2, \ldots, x_m$, acquired experimentally from the variable X, and is defined in its basic form by the formula

$$\hat{f}(x) = \frac{1}{mh^n} \sum_{i=1}^{m} K\left(\frac{x - x_i}{h}\right), \tag{3}$$

where the function $K : \mathbb{R}^n \to [0, \infty)$, which is Borelian, radially symmetrical relative to zero, and has a weak global maximum at this point, fulfills the condition $\int_{\mathbb{R}^n} K(x)\,\mathrm{d}x = 1$ and is called the kernel, whereas the positive coefficient h is known as the smoothing parameter. The form of the kernel K and the value of the smoothing parameter h is selected most often on the basis of the criterion of the minimum mean square error. In that case one assumes additionally the condition $f \in \mathcal{C}^2$, and the boundedness of the functions f and f''.

It turns out that the form of the function K has no essential importance from the statistical point of view, and for that reason it becomes possible, in selecting this function, to take into account primarily the desired properties of the kernel obtained, e. g. the class of regularity, the boundedness or unboundedness of the support, or other features essential in the case of a particular problem. In practice, the normal kernel

$$K(x) = (2\pi)^{-n/2} e^{-\|x\|^2/2} \tag{4}$$

is in general use. The estimator obtained by its application is of the class $\mathcal{C}^\infty$ and furthermore takes on positive values.

Fixing the value of the smoothing parameter $h$is of vital importance for the quality of the estimator obtained. In practice one uses a criterion whose implementation can be reduced to the formula

$$h = \left(c \frac{1}{m} \right)^{1/(n+4)} , \tag{5}$$

whereas for normal kernel (4):

$$c = \frac{4}{2n+1} . \tag{6}$$

In particular applications the linear transformation $X \equiv RY$ is used, while most often the diagonal version of the matrix R is sufficient:

$$R = [r_{i,j}] = \begin{cases} \sqrt{Var(X_i)} & \text{when} \quad i = j \\ 0 & \text{when} \quad i \neq j, \end{cases} \tag{7}$$

where $Var(X_i)$ denotes the variance of the ith coordinate of the random variable X. In a similar manner, positive results can be gained from the so-called modification of the smoothing parameter, which is performed as follows:

(A) the kernel estimator $\hat{f}$ is specified in accordance with the scheme presented earlier;

(B) the modifying parameters $s_i > 0$ $(i = 1, 2, \dots, m)$ of the form

$$s_i = \left(\frac{\hat{f}(x_i)}{s^{\sim}} \right)^p \tag{8}$$

are calculated, where most often $p = -1/2$, while $s^{\sim}$ is the geometric mean of the numbers $\hat{f}(x_1)$, $\hat{f}(x_2), \dots, \hat{f}(x_m)$, given by the logarithmic equation

$$\log(s^{\sim}) = m^{-1} \sum_{i=1}^{m} \log(\hat{f}(x_i)); \tag{9}$$

(C) the kernel estimator is defined, which, after taking into account linear transformation as well, ultimately assumes the form

$$\hat{f}(x) = \frac{1}{mh^n \det(R)} \sum_{i=1}^{m} \frac{1}{s_i^n} K \left(R^{-1} \frac{x - x_i}{hs_i} \right) . \tag{10}$$

Definition (3) is a particular case of formula (10), where R is a unit matrix and $p = 0$, which implies $s_i \equiv 1$. A primary advantage of the modification procedure, in addition to the considerable improvement of the statistical quality of the estimator, is its greatly reduced sensitivity to strongly conditioned – in the case of kernel estimators – fixing of the value of the parameter h. Thus the approximate formula (5) most often proves in practice to be entirely sufficient.

Finally, the last parameter that needs to be determined is the size of the sample m, and in particular its dependency on the dimension of the tested random variable

n. Table 1 shows the minimum sizes of the sample m_* needed to assure 10 % precision at point zero for the normal standard distribution. These values should be treated as an absolute minimum; however, thanks to the capabilities of current computer systems and the automation of metrological processes, the rapidly rising minimum sample size need not constitute an essential barrier in contemporary applications, even when the dimension of the tested random variable approaches 10.

Table 1.

n	m_*
1	4
2	19
3	67
4	223
5	768
6	2790
7	10700
8	43700
9	187000
10	842000

If the value of the parameter h is made dependent on the size of the sample m, in such a way that

$$\lim_{m\to\infty} h = 0 \tag{11}$$

$$\lim_{m\to\infty} mh^n = \infty, \tag{12}$$

the estimator thus obtained is strongly consistent at every point of continuity of the function f, which means that at these points the value of the estimator $\hat{f}(x)$ is convergent with probability 1 to the estimated value $f(x)$[1]. From the statistical point of view, then, the largest possible sample size is desirable, though in practice a certain compromise is necessary, taking into account the calculational aspects, especially time limitations.

Ultimately formulas (3) – (10) provide a full set of rules enabling the specification of the kernel estimator of the density function of the n-dimensional random variable. A broader discussion of the issues presented in the foregoing section, including also the general forms of dependencies (5) and (7), can be found in [6, 20, 23, 25].

3.2. Kernel estimator of distribution function

The mapping $\hat{F} : \mathbb{R} \to [0,1]$ given as $\hat{F}(x) = \int_{-\infty}^{x} \hat{f}(y)\,\mathrm{d}y$ constitutes the natural kernel estimator of the distribution function of the real random variable X. When the kernel K has the primitive function I, that is $I(x) = \int_{-\infty}^{x} K(y)\,\mathrm{d}y$, then after the application of the linear transformation and the modification of the smoothing parameter, this estimator takes on the form

$$\hat{F}(x) = \frac{1}{m}\sum_{i=1}^{m} I\left(\frac{x - x_i}{Rhs_i}\right). \tag{13}$$

[1] In the case of kernel estimators, properties of an asymptotic nature are of fundamental importance, since these estimators are typically used when the random sample is of large size.

If condition (11) is fulfilled, then this estimator is strongly consistent. In the case of the estimator of the distribution function, the exponential kernel

$$K(x) = \frac{e^{-x}}{(1+e^{-x})^2} \tag{14}$$

can be recommended, since its primitive function has the convenient form

$$I(x) = \frac{1}{1+e^{-x}} \tag{15}$$

and fulfills all the conditions formulated here. The parameter c required in formula (5) is here

$$c = \frac{4}{\pi^{7/2}}. \tag{16}$$

A more detailed discussion of estimator (13), along with a proof of its strong consistency, can be found in [16].

3.3. Kernel estimator of quantile

The term quantile of the rth order of the real random variable X, provided that $0 < r < 1$, is used for any number $q \in \mathbb{R}$ fulfilling the equation $F(q) = r$, where F designates the distribution function. The quantile divides the space of the values of the random variable X into two subsets, $(-\infty, q]$ and $[q, \infty)$, such that their probabilities are r and $1 - r$, respectively. On the basis of dependence (13), the kernel estimator of the quantile of the rth order, denoted by $\hat{q}$, can be given as a solution of the equation

$$\frac{1}{m} \sum_{i=1}^{m} I\left(\frac{\hat{q} - x_i}{Rhs_i}\right) = r. \tag{17}$$

When the kernel K is positive, and condition (11) is fulfilled, then this estimator is strongly consistent. Here also, exponential kernel (14) is recommended. These issues, along with the relevant proofs, were introduced in [16]; the statistical properties are described in [11]. A review of differing concepts can be found in [19, 22].

3.4. Statistical forecasting (trend estimation)

The quantity whose future values are the object of statistical prognostic investigation is treated as a discrete real stochastic process Y. If at the moment j one has at hand a sequence of values of the process Y empirically obtained for $1, 2, \ldots, j$, known as "observations" and denoted by $y_1, y_2, \ldots, y_j$, then by making use of the available statistical methodology it is possible to calculate the forecast $\hat{y}_j^s$, i. e., the estimator of the value of the stochastic process Y for the moment $j + s$, while the parameter $s \in \mathbb{N} \setminus \{0\}$ is called the anticipation of the forecast. In a case where the object of interest is not the strict calculation of the future values of the process Y, but rather only the identification of the trend of the changes, then the classic linear regression method [1] is the basic mathematical tool. In such case the forecast $\hat{y}_j^s$ is obtained from the formula

$$\hat{y}_j^s = C_j^T \begin{bmatrix} 1 \\ -s \end{bmatrix}, \tag{18}$$

where

$$
\begin{aligned}
C_j &= D_j^{-1} d_j && (19)\\
D_j &= \sum_{k=0}^{j-1} w^k \begin{bmatrix} 1 & -k \\ -k & k^2 \end{bmatrix} && (20)\\
d_j &= \sum_{k=0}^{j-1} w^k \, y_{j-k} \begin{bmatrix} 1 \\ -k \end{bmatrix}, && (21)
\end{aligned}
$$

while the constant $w \in (0,1]$ is known as the deactualization parameter. After the next observation y_{j+1} has been obtained at the moment $j+1$, the forecast can be updated by means of the formulas

$$
\begin{aligned}
\hat{y}^s_{j+1} &= C^T_{j+1} \begin{bmatrix} 1 \\ -s \end{bmatrix} && (22)\\
C_{j+1} &= D^{-1}_{j+1} d_{j+1}, && (23)
\end{aligned}
$$

while the matrices D_j and d_j are changed in accordance with the dependencies

$$
\begin{aligned}
D_{j+1} &= D_j + w^j \begin{bmatrix} 1 & -j \\ -j & j^2 \end{bmatrix} && (24)\\
d_{j+1} &= w \begin{bmatrix} 1 & 0 \\ -1 & 1 \end{bmatrix} d_j + \begin{bmatrix} 1 \\ 0 \end{bmatrix} y_{j+1}. && (25)
\end{aligned}
$$

The specification of a concrete value for the anticipation parameter s essentially results from the application conditions; it should be recalled, however, that as its value increases, the forecast obtained becomes less precise. In practice the deactualization parameter w is assumed in such a way that $w \in [0,8; 0,99]$, and there is a particular preference for $w = 0,95$. A decrease in this value improves the likelihood that the model will adapt to changes taking place, but it also increases its sensitivity to interference; the opposite would result from increasing the value of the parameter w. The number of available observations should be no less than 15.

A review of statistical forecasting methodology can be found in [1, 26].

4. DESIGNING A STATISTICAL INFERENCE SYSTEM

4.1. Diagnosis of faults

Diagnosis is accomplished by successively testing a finite number of the hypotheses $H_1, H_2, \ldots, H_d$, stating the occurrence of the assumed types of faults, as against the same alternative hypothesis H_0, representing proper operating conditions; in other words, diagnosis will consist in the independent testing of the truth of d pairs of hypotheses, of the form H_k versus H_0. Accordingly, for any fixed $k = 1, 2, \ldots, d$, let the following random samples be given, composed from the experimentally obtained values of the symptom vector Z:

$z_1^k, z_2^k, \ldots, z_{m_k}^k$ – in the event of the occurrence of the kth type of fault to be diagnosed (26)

$$z_1^0, z_2^0, \ldots, z_{m_0}^0 \text{ – for proper operating conditions} \tag{27}$$

and

$$Z(\omega, j) \text{ – the current value (at the moment } j\text{) of the symptom vector.} \tag{28}$$

The procedure for accepting the hypothesis H_k or H_0 will now be presented. The statistic $S_k : \Omega \times \mathbb{N} \setminus \{0\} \to [0, \infty)$ will be considered, as defined by the formula

$$S_k(\omega, j) = f_k(Z(\omega, j)), \tag{29}$$

where the Borelian mapping $f_k : \mathbb{R}^n \to [0, \infty)$ denotes the density function of the one-dimensional distributions of the symptom vector Z assuming the occurrence of the kth type of fault to be diagnosed. (A detailed description of the specification procedure of the kernel estimator of this function is presented in Section 3.1.) The form of the statistic when defined in this way makes it possible to identify possible changes involving not only the values of the individual coordinates of the symptom vector, but also, and especially, the complex relations that occur among them. For any fixed value j, representing the corresponding moment in time, the value of the statistic S_k can be referred to the probability that the current value of the symptom vector will occur under the condition that the kth type of fault to be diagnosed has appeared. Thus low values for this statistic are an indication in favor of accepting the hypothesis H_0, i. e. this should happen along with the relation

$$S_k(\omega, j) \in A_k, \tag{30}$$

where

$$A_k = (-\infty, a_k]. \tag{31}$$

In the opposite case, if

$$S_k(\omega, j) \in B_k \tag{32}$$

for

$$B_k = \mathbb{R} \setminus A_k, \tag{33}$$

then the hypothesis H_k should be accepted. In order to calculate the critical value a_k the basic formula for statistical decision theory, the Bayes rule, has been used [3]. In Appendix A it is shown that if $\mathsf{a}_k > 0$ and $\mathsf{b}_k > 0$ denote the losses resulting from non-detection of the kth type of fault and from the corresponding false alarm, respectively, then the optimal critical value a_k in the sense of this rule can be obtained from the criterion

$$F_{f_k \circ Z|_0}(a_k) + \frac{\mathsf{a}_k}{\mathsf{b}_k} F_{f_k \circ Z|_k}(a_k) = 1, \tag{34}$$

where $F_{f_k \circ Z|_0}$, $F_{f_k \circ Z|_k}$ denote the distribution functions of the random variables $f_k \circ Z$ for proper operating condition and the occurrence of the kth type of fault to be diagnosed, respectively. In practice, the functions f_k and $F_{f_k \circ Z|_0}$, $F_{f_k \circ Z|_k}$ are available only in the form of the corresponding estimators $\hat{f}_k$, $\hat{F}_{\hat{f}_k \circ Z|_0}$, $\hat{F}_{\hat{f}_k \circ Z|_k}$, and therefore dependence (34) takes on the form of the following equation:

$$\hat{F}_{\hat{f}_k \circ Z|_0}(\hat{a}_k) + \frac{\mathsf{a}_k}{\mathsf{b}_k} \hat{F}_{\hat{f}_k \circ Z|_k}(\hat{a}_k) = 1, \tag{35}$$

while its argument $\hat{a}_k$ constitutes an estimator of the quantity a_k. Appendix B contains a theorem stating that if the values of the smoothing parameters are made dependent on the samples sizes in such a way that the conditions (11) – (12) are fulfilled, then this estimator is strongly consistent, which indicates the formal correctness of the fault diagnosis procedure proposed here. If the left side of dependence (35) is treated as the function $g_k : I\!R \to I\!R$ of the variable $\hat{a}_k$, then applying criterion (35) one should find the solution of the equation $g_k(\hat{a}_k) = 1$. In that case, the following formulas are true:

$$\lim_{\hat{a}_k \to -\infty} g_k(\hat{a}_k) = 0 \tag{36}$$

$$\lim_{\hat{a}_k \to \infty} g_k(\hat{a}_k) = 1 + \frac{\mathsf{a}_k}{\mathsf{b}_k} > 1. \tag{37}$$

The function g_k is continuous, which results from the continuity of the kernel estimator of the distribution function. Taken together with conditions (36) – (37), this states the existence of a solution for equation (36). If the kernel K with positive values is used in estimating the distribution function, then the function g_k is furthermore strictly increasing, which implies the uniqueness of this solution, and makes it possible to apply effective numerical procedures; in particular, this solution can be calculated using the Newton method [7] as the limit of the sequence $\{\hat{a}_k^l\}_{l=1}^{\infty}$ defined by the formulas

$$\hat{a}_k^1 = 0 \tag{38}$$

$$\hat{a}_k^{l+1} = \hat{a}_k^l + \frac{1 - \hat{F}_{\hat{f}_k \circ Z|_0}(\hat{a}_k^l) - \frac{\mathsf{a}_k}{\mathsf{b}_k} \hat{F}_{\hat{f}_k \circ Z|_k}(\hat{a}_k^l)}{\hat{f}_{\hat{f}_k \circ Z|_0}(\hat{a}_k^l) + \frac{\mathsf{a}_k}{\mathsf{b}_k} \hat{f}_{\hat{f}_k \circ Z|_k}(\hat{a}_k^l)} \quad \text{for } l = 1, 2, \ldots, \tag{39}$$

where the denotations $\hat{f}_{\hat{f}_k \circ Z|_0}$, $\hat{f}_{\hat{f}_k \circ Z|_k}$ additionally introduced here constitute the density functions of the random variables $\hat{f}_k \circ Z$ for proper operating conditions and the occurrence of the kth type of fault to be diagnosed, respectively (note that the estimator of the density function constitutes a derivative of the estimator of the distribution function).

Since the random variables $\hat{f}_k \circ Z$ take on only positive values, then the primitive function I can be slightly modified to I_* given as

$$I_*(x) = \begin{cases} 0 & \text{for} \quad x < 0 \\ \frac{I(x) - I(0)}{1 - I(0)} & \text{for} \quad x \geq 0 \end{cases}, \tag{40}$$

in which case the kernel estimators of the distribution function have the support of the form $[0, \infty)$, adequate to this situation. The derivative of this kernel is then

$$K_*(x) = \begin{cases} 0 & \text{for} \quad x < 0 \\ \frac{K(x)}{1 - I(0)} & \text{for} \quad x \geq 0 \end{cases} \tag{41}$$

and the kernel modified in this way should be used to calculate the estimators of the density functions required in algorithm (38) – (39).

The procedure leading to the acceptance of the hypothesis H_k or H_0 thus consists of three phases. To begin with, random samples (26) – (27) should be obtained experimentally, while in the first case it may become necessary to simulate the conditions of the occurrence of the given type of diagnosed fault. The second step involves calculating, on the basis of these samples, the elements of the decision-making process: the functions f_k, and therefore the statistic S_k and the critical value a_k. Finally, the third phase – the only one performed in real time – leading to the acceptance of the hypothesis H_0 or H_k consists merely in inserting the current value of symptom vector (28) to formula (29) and then determining whether relation (30) or (32) is true.

4.2. Detection of faults

The detection of faults will depend on using the current values of the symptom vector to test successively the hypothesis stating that the system under supervision is operating properly. In view of the assumed full spectrum of malfunctions to be detected, implying significant non-specificity of the possible alternative hypothesis and the probability of an error of the second kind, a test of significance has been designed for the needs of detection, according to the principles of the Neyman-Pearson theory [8]. Once again, statistical random samples will be used, as defined for the needs of diagnosis by formulas (27) and (28). If the distribution functions of the random variables from which these samples originate are denoted by F_0 and G, then the hypothesis H_0 stating their identity is to be tested:

$$H_0 : G \equiv F_0. \tag{42}$$

It has been assumed that from the practical point of view this formal form is equivalent to expression (1), which describes the applicational aspects. Thus the detection procedure presented below constitutes a test of consistency in the situation when one of the samples is one-element; the statistical properties of such a test have been presented in [13].

Just as in the case of diagnosis, the verification procedure will be divided here into three phases: experimental acquisition of random sample (27), calculation on that basis of the statistic and the critical set, and finally, testing the hypothesis H_0 while the system is in operation, based on the current value of symptom vector (28). The statistic $S_0 : \Omega \times I\!\!N \setminus \{0\} \to [0, \infty)$ applied for this purpose is defined by the formula

$$S_0(\omega, j) = f_0(Z(\omega, j)), \tag{43}$$

while the Borelian mapping $f_0 : I\!\!R^n \to [0, \infty)$ denotes the density function of the one-dimensional distributions of the symptom vector Z in the case of proper operation of the system under supervision (for the specification of its kernel estimator, see Section 3.1). The assumed form of the statistic makes it possible to recognize the changes taking place both in the individual coordinates of the symptom vector and in their mutual relations. The value of this statistic is associated with the probability of the occurrence of the current symptom vector under proper operating conditions, which implies the left-sided form of the critical set

$$A_0 = (-\infty, a_0], \tag{44}$$

where the critical value a_0 is calculated in such as way as to fulfill the condition

$$F_{f_0 \circ Z|_0}(a_0) = \alpha_0 \tag{45}$$

for the fixed level of significance $\alpha_0 \in (0,1)$ representing the probability of an error of the first kind, which in this case means a false alarm. The critical value a_0 can thus be estimated by using the quantile of the order α_0 of the distribution of the random variables $f_0 \circ Z(\cdot, j)$ under proper operating conditions. Ultimately, if

$$S_0(\omega, j) \in A_0, \tag{46}$$

then it should be inferred that the hypothesis H_0 stating the proper operation of the system is false, whereas when

$$S_0(\omega, j) \notin A_0 \tag{47}$$

there is no basis to reject this hypothesis.

In statistical practice the density function f_0 and the distribution function $F_{f_0 \circ Z|_0}$ are replaced by their estimators $\hat{f}_0$ and $\hat{F}_{\hat{f}_0 \circ Z|_0}$; thus dependence (45) can be written in the form of the equation

$$\hat{F}_{\hat{f}_0 \circ Z|_0}(\hat{a}_0) = \alpha_0, \tag{48}$$

while its solution $\hat{a}_0$ constitutes an estimator of the quantity a_0 from criterion (45). In Appendix C it is shown that if the values of the smoothing parameters are made dependent on the samples sizes in such a way as to guarantee the fulfillment of conditions (11) – (12), then this estimator is strongly consistent. This proves the formal correctness of the fault detection procedure presented here.

A useful concept for calculating the value of the quantile estimator, and thus the quantity $\hat{a}_0$, was presented in Section 3.3. In practice, if one uses a positive kernel K, this value may be calculated by means of the Newton method [7] as the limit of the sequence $\left\{\hat{a}_0^l\right\}_{l=1}^{\infty}$ defined by the formulas:

$$\hat{a}_0^1 = 0 \tag{49}$$

$$\hat{a}_0^{l+1} = \hat{a}_0^l + \frac{r - \hat{F}_{\hat{f}_0 \circ Z|_0}(\hat{a}_0^l)}{\hat{f}_{\hat{f}_0 \circ Z|_0}(\hat{a}_0^l)} \quad \text{for } l = 1, 2, \ldots. \tag{50}$$

Here, also, in view of the fact that the random variable $\hat{f}_0 \circ Z$ takes on positive values, one should use a modified version of the kernel and its primitive funtion defined by formulas (41) and (40).

Thanks to the similarity of the mathematical apparatus used for detection and diagnosis, specialized procedures can be used in the latter case in order to reinforce the detection test, which is universal and thus less conveniently conditioned. Namely, the procedures for diagnosis will be much more effective in discovering the type of fault assigned to them than the detection test, which does not make use of detailed characterizations of the conditions for the occurrence of these faults. It will be advantageous, then, to introduce for the needs of detection the modified hypothesis H_0^* of the form

$$H_0^* = H_0 \wedge \sim H_1 \wedge \sim H_2 \wedge \ldots \wedge \sim H_k, \tag{51}$$

which in practice means that the proper operation of the supervised system would be confirmed by a positive result from the verification of the hypothesis H_0 (previously the detection concept) and negative results from diagnosis tests. From the theoretical point of view, this change does not introduce any new elements, since the rejection of the hypotheses H_k tested in the course of diagnosis would be equivalent to accepting the alternative hypothesis H_0, and so $\sim H_k = H_0$. In practice, on the other hand, this modification increases the effectiveness of the detection system.

For purposes of detection a test of significance has been designed. The result of such a test can only be the rejection of the hypothesis H_0, but no decision is made whether to accept it. From the perspective of application this is of no essential importance, since both accepting and not rejecting the hypothesis H_0 produce in practice the same result: taking no action. There is, however, a certain drawback in the absence of the possibility to fix the level of significance α_0 by comparing the economic consequences of errors of the first and second kind. However, the hypothesis H_0, verified in the course of detection, is likewise an alternative hypothesis for the diagnosis tests; thus the value of the parameter α_0 may be compared with the constants β_k – the probability of an error of the second kind for the kth test of diagnosis. Ultimately, as a preliminary estimate it is proposed to assume

$$\alpha_0 = \min_{k=1,\ldots,d} \beta_k, \tag{52}$$

due to the previously mentioned lesser effectiveness of the universal detection test. Any possible increase in value (52) improves the sensitivity of the detection system, but at the cost of a greater number of false alarms; a reduction produces the opposite effect. The parameters β_k can easily be calculated in the course of solving equations (35), with due regard for the fact that the value of the first factor on its left side is equal to $1 - \beta_k$.

4.3. Prediction of faults

The idea of the prediction system involves successively analyzing the evolution of the symptom vector and making inferences on the basis as to the possibility that improper operating conditions will occur in the future. Both the mere appearance of a fault (the scope of detection) and the particular types subject to diagnosis are predicted. The methodology used combines elements of the theory of testing hypotheses applied in Sections 4.1 and 4.2 for purposes of detection and diagnosis with statistical forecasting (Section 3.4). The object of verification is the supposition that the values of the statistics S_0 and $S_1, S_2, \ldots\ , S_d$ defined by dependencies (43) and (29) will belong in the future to the sets A_0 and $A_1, A_2, \ldots, A_d$ or $B_1, B_2, \ldots, B_d$, given by formulas (44) and (31) or (33).

Accordingly, let $s_0 \in I\!N\backslash\{0\}$ represent the anticipation with which the appearance of a system fault (in the detection sense) is predicted. Treating at the moment $j \in I\!N \setminus \{0\}$ the previous values of the statistic S_0:

$$S_0(\omega, 1), S_0(\omega, 2), \dots, S_0(\omega, j) \tag{53}$$

as the observations $y_1, y_2, \dots, y_j$ (see Section 3.4), one can calculate the forecast $\hat{y}_j^{s_0}$, which represents the estimator of the statistic S_0 for the moment $j + s_0$. Therefore, if there occurs the relation

$$\hat{y}_j^{s_0} \in A_0, \tag{54}$$

then it should be supposed that in the future the hypothesis H_0 will be false, or more precisely, it can be inferred that in s_0 time units a fault will appear. In the opposite case, when

$$\hat{y}_j^{s_0} \notin A_0, \tag{55}$$

there is no basis to reject the judgment that the supervised device will be operating properly in s_0 time units.

Similarly, for each $k = 1, 2, \dots, d$ let the parameter $s_k \in I\!N \setminus \{0\}$ represent the anticipation with which the occurrence of the kth type of diagnosed fault is predicted. At any moment $j \in I\!N \setminus \{0\}$ the past values of the statistic S_k:

$$S_k(\omega, 1), S_k(\omega, 2), \dots, S_k(\omega, j), \tag{56}$$

are treated in the forecasting process as the observations $y_1, y_2, \dots, y_j$. Calculating the forecast $\hat{y}_j^{s_k}$ for the anticipation s_k , one obtains the estimator of the statistic S_k for the moment $j + s_k$. Therefore, if the relation

$$\hat{y}_j^{s_k} \in B_k \tag{57}$$

appears, then the future truth of the hypothesis H_k is expected, and therefore one infers that in s_k time units the kth type of fault to be diagnosed will occur. In the opposite case, i. e. when

$$\hat{y}_j^{s_k} \in A_k, \tag{58}$$

that supposition should be rejected.

The concept of the modified hypothesis H_0^*, as described by formula (51), can be transposed to the problem of fault prediction in a natural way.

The realization of the idea of prediction presented above ultimately requires, in the general case, separate predicting of the value $d + 1$ of the stochastic processes: one regarding the appearance of a system fault (detection) and d associated with the particular types to be diagnosed. The choice of a particular forecasting method is determined by the specific factors conditioning the problem of fault prediction, calculating speed, and the possibility of updating data within the procedure used. On the other hand, it is not so much the actual precision of the forecast that most matters for fault prediction, but only correct trend identification. Under such conditions, the classic regression method, even in its basic form, turns out to be particularly well suited to the problem of fault prediction. In Section 3.4 formulas were presented enabling the calculation of the forecasts $\hat{y}_j^{s_0}$ and $\hat{y}_j^{s_1}, \hat{y}_j^{s_2}, \ldots, \hat{y}_j^{s_d}$ for this method, whereas what follows will discuss the specific aspects of the choice of anticipation and deactualization parameters, and of the minimum number of available observations.

The choice of values for the anticipation parameters s_k results from the technological conditions of the system under supervision. In practice this means choosing the minimum time needed to stop the device, or at least to change the working regime in a matter appropriate to the type of fault forecasted. The postulate of minimization is formulated with regard to the precision of the forecast. If the modified hypothesis (51) is used, then it is advisable to introduce the boundary

$$s_k \leq s_0 \quad \text{for } k = 1, 2, \ldots, d. \tag{59}$$

The value of the deactualization parameter can be chosen according to the universal rules presented in Section 3.4. The natural conditions surrounding the fault prediction process, where the essence of the phenomena being analyzed often consists in dramatic changes taking place in the supervised system, while the technological conditions require relatively large anticipation, may nevertheless require difficult compromises.

Finally, as mentioned in Section 3.4, the minimum number of available observations should be no less than 15. The system proposed here presents no particular difficulties in meeting that requirement. It can easily be rendered that the number of available observations was equal to the number of values of the symptom vector Z obtained for proper operating conditions (27). Since in practice the dimension of this vector is always greater than 1, according to the data shown in Table 1 the necessary number of values to be obtained exceeds even 15 by several orders of magnitude.

5. SELF–CORRECTION

The kernel estimators technique, on the basis of which the methodology for the statistical inference system here designed has been developed, also enables the introduction of effective self-correction, which makes it possible to significantly improve the quality of this system. The concept of self-correction is based on elements of artificial intelligence, specifically neural networks. The procedure will consist of two phases: one performed off-line before the system begins operation, and one on-line, in reaction to erroneous indications during the supervision processes, which are unavoidable in practice.

The kernel estimators of the density function (10) and distribution function (13) can be generalized to the forms

$$\hat{f}(x) = \frac{1}{mh^n \det(R)} \sum_{i=1}^{m} \frac{w_i}{s_i^n} K\left(R^{-1} \frac{x - x_i}{hs_i}\right) \tag{60}$$

$$\hat{F}(x) = \frac{1}{m} \sum_{i=1}^{m} w_i I\left(\frac{x - x_i}{Rhs_i}\right), \tag{61}$$

when the non-negative coefficients w_i $(i = 1, 2, \ldots, m)$ fulfill the condition

$$\sum_{i=1}^{m} w_i = m. \tag{62}$$

In making the appropriate choice of their values, one should give preference to those kernels which have the greatest impact on proper system indications, while the significance of those which prevailed in erroneous decision-making can be gradually eliminated. This corresponds to the learning process in a neural network. A comparison of basic forms (10) and (13) with generalized versions (60) and (61) indicates that $w_i \equiv 1$ should be accepted as the initial values.

The first phase is carried out off-line, on the basis of data obtained in the form of random samples (26) and (27).

In the case of the diagnosis of the kth type of fault, the combination of formulas (26) – (33) implies that the fulfillment of the condition

$$\hat{f}_k(z_j^k) \leq a_k \quad \text{for any } j = 1, 2, \ldots, m_k, \tag{63}$$

is an indication of an error of the first kind, i. e. neglecting this type of fault, whereas the relation

$$\hat{f}_k(z_j^0) > a_k \quad \text{for any } j = 1, 2, \ldots, m_0, \tag{64}$$

constitutes an error of the second kind, that is, a false alarm. In both cases one can introduce a change in the coefficients w_i, denoted for the kth type of diagnosed fault as w_i^k, altering the significance of particular kernels depending on their impact on the appearance of an error. Thus for each index j at which condition (63) is fulfilled, the auxiliary values $\tilde{w}_i^k$ are defined as

$$\tilde{w}_i^k = w_i^k \left(1 + \frac{\frac{w_i^k}{\left(s_i^k\right)^n} K_k \left(R_k^{-1} \frac{z_j^k - z_i^k}{h_k s_i^k} \right)}{\sum\limits_{l=1}^{m_k} \frac{w_i^k}{\left(s_l^k\right)^n} K_k \left(R_k^{-1} \frac{z_j^k - z_l^k}{h_k s_l^k} \right)} \right)^{p_k} \quad \text{for every } i = 1, 2, \ldots, m_k, \tag{65}$$

whereas in the case of every j at which (64) is true:

$$\tilde{w}_i^k = w_i^k \left(1 - \frac{\frac{w_i^k}{\left(s_i^k\right)^n} K_k \left(R_k^{-1} \frac{z_j^k - z_i^k}{h_k s_i^k} \right)}{\sum\limits_{l=1}^{m_k} \frac{w_i^k}{\left(s_l^k\right)^n} K_k \left(R_k^{-1} \frac{z_j^k - z_l^k}{h_k s_l^k} \right)} \right)^{p_k} \quad \text{for every } i = 1, 2, \ldots, m_k, \tag{66}$$

where K_k, h_k, R_k, s_l^k or s_i^k denote the kernel, the smoothing parameter, the transformation matrix, and the modifying parameters, respectively, used in the construction of the estimators associated with the kth type of fault, while the positive exponent p_k states the intensity of the changes. As a preliminary value it is proposed to accept

$$p_k = \log m_k; \tag{67}$$

reducing or enlarging this value decreases or increases the intensity of the changes, respectively. The coefficients $\tilde{w}_i^k$ must be normalized in order to assure condition (64). Finally, one obtains the altered coefficients w_i^k according to the formula

$$w_i^k = \frac{m\tilde{w}_i^k}{\sum_{l=1}^{m_k} \tilde{w}_l^k} \quad \text{for every } i = 1, 2, \ldots, m_k. \tag{68}$$

After these have been specified, one should recalculate the critical values a_k on the basis of criterion (35), using kernel estimators in generalized forms (60) – (61). The above procedure can be repeated until the sum of errors of the first and second kind for the elements of random samples (26) and (27) has stabilized.

In the case of detection, the fulfillment of the relation

$$\hat{f}_0(z_i^0) \le a_0 \quad \text{for any } i = 1, 2, \ldots, m_0 \tag{69}$$

indicates an error of the first kind. As above, it is possible to alter the coefficients w_i, denoted in the case of detection as w_i^0. For every index j at which condition (69) is fulfilled, one defines

$$\tilde{w}_i^0 = w_i^0 \left(1 + \frac{\frac{w_i^0}{\left(s_i^0\right)^n} K_0\left(R_0^{-1} \frac{z_j^0 - z_i^0}{h_0 s_i^0}\right)}{\sum\limits_{l=1}^{m_k} \frac{w_i^0}{\left(s_l^0\right)^n} K_0\left(R_0^{-1} \frac{z_j^0 - z_l^0}{h_0 s_l^0}\right)} \right)^{p_0} \quad \text{for every } i = 1, 2, \ldots, m_0, \tag{70}$$

where K_0, h_0, R_0, s_l^0 or s_i^0 mean the kernel, the smoothing parameter, the transformation matrix, and the modifying parameters, respectively, used in the construction of the estimators for detection; and after normalization

$$w_i^0 = \frac{m\tilde{w}_i^0}{\sum_{l=1}^{m_k} \tilde{w}_l^0} \quad \text{for every } i = 1, 2, \ldots, m_0, \tag{71}$$

while, as before, it is proposed to accept initially

$$p_0 = \log m_0. \tag{72}$$

The number of repetitions of the above procedure for detection should be the maximum among the number of repetitions postulated previously for the individual types of faults to be diagnosed. After each repetition the altered value should be calculated for the level of significance a_0.

The second phase of the self-correction is performed on-line, on the basis of the current value of the symptom vector. This is done according to the scheme presented above, while in the event that an error occurs, the current value of the symptom vector (28) is inserted into the appropriate place in the elements of random samples (26) and (27) in order to check conditions (63) – (64) and (69). In view of the likely incidental nature of such an event, it is not necessary to update the levels of significance. For this same reason one may recommend at least doubling exponents (67) and (72). If the deactualization parameters used in forecasts are other than one, the prediction algorithm on its own accord adjusts itself to the on-line self-correction procedure.

The concept presented above for the self-correction of the system designed in Sections 3 and 4 makes it possible to eliminate off line the non-representative elements obtained in random samples (26) – (27), and adapt on line to changing conditions.

6. EXPERIMENTAL VERIFICATION

The proper operation of the statistical inference system worked out in this paper has been verified experimentally. The supervised object was a mechanical system, subjected to a robust time-optimal control [10,12,15,17], whose dynamics are described by the differential inclusion

$$\ddot{y}(t) \in H(\dot{y}(t), y(t), t) + u(t), \tag{73}$$

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 form

$$H(\dot{y}(t), y(t), t) = v(\dot{y}(t), y(t), t)\, G(\dot{y}(t)), \tag{74}$$

while v denotes a continuous mapping, and G is a piecewise continuous function, additionally multivalued at the points of discontinuity. In the event that resistance to motion is omitted, i. e. when $H \equiv 0$, inclusion (73) is reducible to a differential equation expressing Newton's second law of dynamics. This is a problem of fundamental importance in the control of industrial manipulators and robots. The random time-optimal control takes on the values $+1$ or -1, depending on where among the distinguished sets the system state is located; for details see [10,12,15,17]. The symptom vector was assumed in the following form:

$$Z(\cdot) \equiv \begin{bmatrix} |u(\cdot)| \\ |H(\cdot)| \\ |\dot{y}(\cdot)| \end{bmatrix}, \tag{75}$$

and therefore its coordinates designate successively the absolute values of control, resistance to motion, and velocity. Diagnosis consisted in recognizing two assumed types of faults. The first was assumed to be the reduction of the maximum absolute value of the admissible control by the value $\Delta u \in [0, 1]$, which in practice indicates a fault in the drive system. The second type of diagnosed fault was taken to be an increase in resistance to motion (its 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. In order to calculate the values of the detection and diagnosis parameters, it was assumed that $\mathsf{a}_1/\mathsf{b}_1 = \mathsf{a}_2/\mathsf{b}_2 = 50$, whereas in the case of prediction the anticipation parameters came to 100, and the deactualization parameters were 0,98 (prediction of detection) and 0,95 (prediction of diagnosis).

The remaining quantities were generated in accordance with the suggestions made in Sections 3 – 5.

The results of these experiments verified the concept presented in this paper and confirmed the proper functioning of the statistical inference system here designed. In cases where the manifesting symptoms were abrupt, the malfunction of the device was promptly discovered and correctly recognized within the scope of detection and diagnosis. Figure 1 illustrates the values of the detection statistic S_0 and diagnosis statistic S_1 obtained in such a case and their forecasts. If, on the other hand, the fault was accompanied by a slow progression of symptoms, it was forecasted with a correct indication of the type of fault about to occur (prediction), and at the appropriate moment it was discovered and recognized within the scope of detection and diagnosis (see Figure 2). Self-correction also operated properly, preliminarily eliminating nonrepresentative elements of the random samples and later adapting the system to variable operating conditions.

The experience gathered from these experiments has made it possible to formulate a number of conclusions pertaining to application.

During the application of the system presented here, it may prove advantageous for the purpose of prediction to limit the values of the statistic S_0. In such case, the following quantities should be treated as observations:

$$\min(S_0(\omega, 1), b_0), \min(S_0(\omega, 2), b_0), \ldots, \min(S_0(\omega, j), b_0), \tag{76}$$

given that $b_0 > 0$; this means the upper boundary of the statistic S_0 to the number b_0. The purpose of this conception is to eliminate erroneous indications of prediction resulting from the mere shifting of the symptom vector from areas assigned to exceptionally large values of the statistic S_0 to regions with values which, though indeed smaller, do not in fact give grounds to presume that a fault has occurred. A multiple of the critical value a_0 can be proposed as the constant b_0, specifically $b_0 = 10a_0$. (In Figures 1 and 2 one can see the impact of the boundary $b_0 = 2$ on the forecast of the statistics S_0.)

From the practical point of view, it may also prove advantageous to bound the dimension of the symptom vector Z, for the purposes of individual diagnosis tests, to only those coordinates which are of essential significance for the recognition of the given type of fault. This means that for every $k = 1, 2, \ldots, d$ the definition of statistic (29) should be generalized to the form

$$S_k(\omega, j) = f_k(g_k(Z(\omega, j))), \tag{77}$$

where $g_k : I\!\!R^n \to I\!\!R^{n_k}$ ($n_k \in \{1, 2, \ldots, n-1\}$) is a mapping of the spatial projection $I\!\!R^n$ onto the subspace $I\!\!R^{n_k}$ composed of the previously mentioned coordinates of the symptom vector that are essential for the given type of fault. Then $f_k : I\!\!R^{n_k} \to I\!\!R$ and – in accordance with Table 1 – the sizes of the random samples (26) are subject to reduction.

For those types of diagnosed faults which are not preceded by clear-cut symptoms, prediction can be omitted.

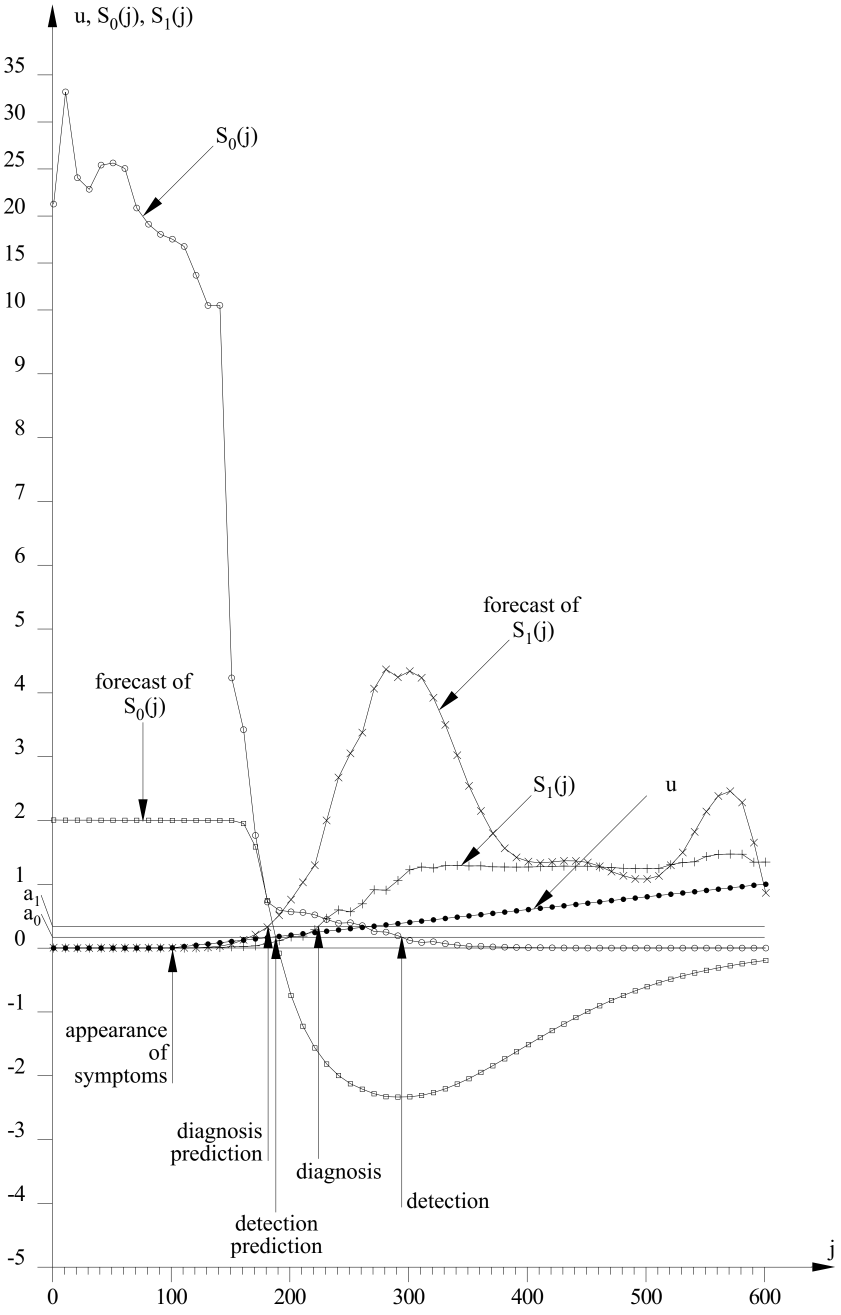

Fig. 1. The fault detection process with abrupt changes.

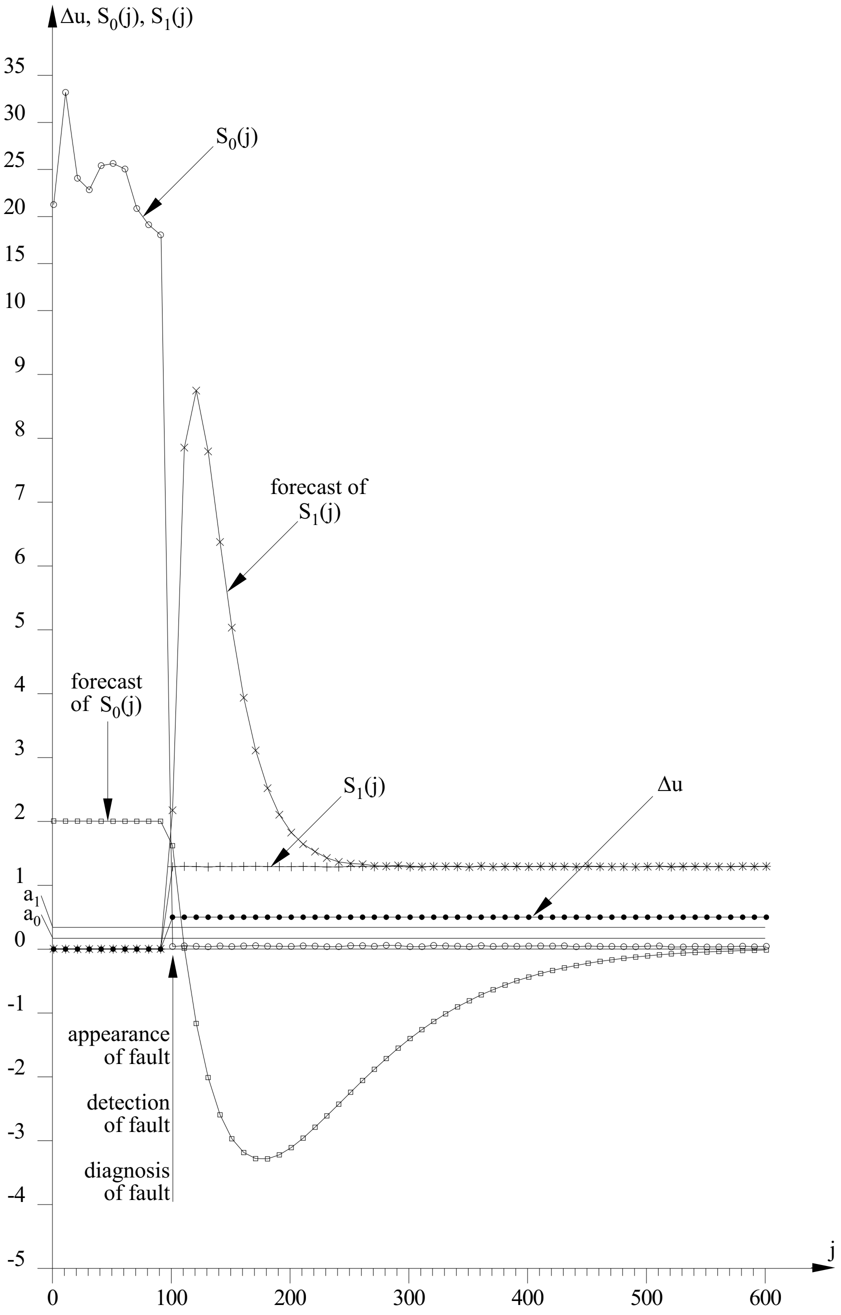

Fig. 2. The fault detection process in the case of slowly progressive symptoms.

7. CONCLUSIONS

In the present paper the concept of a statistical fault detection system has been presented for detection, diagnosis, and prediction associated with these two functions. It is intended for application in real time, while the device under supervision is performing its usual technological operations. In the statistical inference system here designed, the successive availability of the symptom vector has been assumed, that is, of a finite number of variables whose current values and/or the relations between them are dependent on the technical state of the device being supervised. The exact form of these relations need not be given *a priori* – its identification constitutes an integral part of the procedure here proposed. No limitations have been introduced on the form of the statistical patterns characterizing proper working conditions and the types of faults to be diagnosed; in particular, allowance is made for the existence of local extremes. Detailed familiarity with the model of the dynamic system being supervised is likewise not required. Allowance is also made for the simultaneous occurrence of several different types of faults to be diagnosed. The system here presented has the capacity to detect and recognize changes in the values of particular coordinates of the symptom vector, and especially – due to the application of kernel estimators – the complex qualitative and quantitative relations existing among them. These changes may be abrupt, or – thanks to prediction – slowly progressive. A procedure has also been introduced to eliminate less credible data and adapt the system to variable working conditions. Finally, the demands on the automatic control system executing the algorithm here designed do not exceed the capacities of contemporary devices sufficiently advanced to apply fault detection procedures. The possibilities of modern systems in practice limit the dimension of the symptom vector to $7-9$, while due to the global probability of error the number of types of faults to be diagnosed should not exceed $3-5$. The material presented above is documented by mathematical theorems given in the appendices.

The present paper provides a complete algorithm enabling the construction of a usable fault detection system in respect to statistical inference regarding the current value of the symptom vector. This places the following demands on the designer:

(A) defining the symptom vector Z on the basis of the technological conditions and the available methodology [2, 5, 9, 14, 18, 24];

(B) distinguishing d types of faults foreseen for diagnosis;

(C) specifying the quotients $\mathsf{a}_1/\mathsf{b}_1$, $\mathsf{a}_2/\mathsf{b}_2, \ldots, \mathsf{a}_d/\mathsf{b}_d$ representing the ratio of economic losses resulting from neglecting particular types of faults to be diagnosed against the corresponding false alarms;

(D) fixing the anticipation of the forecasts $s_0, s_1, \ldots, s_d$ on the basis of technological requirements; it is suggested that relation (59) be fulfilled;

(E) obtaining experimentally the sequence of values of the symptom vector when the system is operating properly (27) and when the particular types of faults to be diagnosed are occurring (26).

At that point – according to the design here presented – the following steps should be taken:

(F) using the foregoing sequences, specify the kernel estimators of the density functions $f_0, f_1, \ldots, f_d$ (following the instructions given in Section 3.1), and therefore the forms of the statistics $S_0, S_1, \ldots, S_d$;

(G) on the basis of the results obtained at items (E) and (F), calculate the sequences of values of the random variables $f_0 \circ Z$, $f_1 \circ Z$,..., $f_d \circ Z$ assuming the proper operation of the system, and the variables $f_1 \circ Z$, $f_2 \circ Z$,..., $f_d \circ Z$ when respectively the 1st, 2nd,..., dth type of fault to be diagnosed is occurring[2];

(H) using these sequences, define the kernel estimators of the density functions (Section 3.1) and the distribution functions (Section 3.2) of the variables $f_1 \circ Z$, $f_2 \circ Z, \ldots,$ $f_d \circ Z$ when the system is operating properly, and when respectively the 1st, 2nd,...,dth type of fault to be diagnosed is occurring;

(I) applying algorithm (38) – (39), calculate the critical values $a_1, a_2, \ldots, a_d$;

(J) by means of the sequences obtained at item (G), specify – applying algorithm (49) – (50) – the kernel estimator of the quantile of the random variable $f_0 \circ Z$ when the system is operating properly, and therefore the critical value a_0; the order of the quantile can be assumed on the basis of formula (52);

(K) as described in Section 5, establish the procedure for self-correction and perform the operations in the off-line phase;

(L) based on the sequences of values of the random variables $f_0 \circ Z$, $f_1 \circ Z$,..., $f_d \circ Z$ for the case of proper system operation, as calculated at item (G), fix the parameters of the prediction model (Sections 3.4 and 4.3) in the detection sense and for the individual types of faults to be diagnosed.

When the fault detection system is working in real time, after obtaining the successive values of the symptom vector, one proceeds in turn to:

(M) obtain the current values of the statistics $S_0, S_1, \ldots, S_d$, and then verify detection conditions (46) or (47), and diagnosis conditions (30) or (32);

(N) calculate the forecasts for these statistics and check the prediction conditions in the detection sense (54) or (55) and diagnosis sense (57) or (58);

(O) update the prediction models (following the instructions in Sections 3.4 and 4.3);

(P) in the event of erroneous indication, improve the procedure for on-line self-correction (Section 5).

The calculation algorithms for items (F) – (P) have been fully presented in Sections 3 – 5 of this article. References have also been given to the literature, enabling the future creation of more sophisticated individualized versions.

Finally, the material presented here can also be designed using kernel estimators in their conditional version [21]. In the event that particular coordinates of the symptom vector should prove to be significantly dependent on other factors of a variable nature (e. g. environmental temperature), this may lead to a major improvement in the system's practical properties. This problem will be investigated in future research projects.

[2] If – for the sake of example – $z_0^1, z_0^2, \ldots, z_0^{m_0}$ means an m_0-element sequence of the symptom vector Z for proper operating conditions, then the sequence of values of the random variable e. g. $f_k \circ Z$ assuming such conditions should be understood to have the values $f_k(z_0^1), f_k(z_0^2), \ldots, f_k(z_0^{m_0})$.

APPENDICES

Appendix A: Proof of the optimality of the critical value for the diagnosis tests

In this appendix it will be shown that the optimal – in the Bayes sense – critical value for the kth diagnosis test is given as the solution of equation (34).

The basic task of statistical decision theory [3] is the optimal selection of one element from among all possible decisions on the sole basis of probabilistic information about the state of nature (reality), especially when its actual state is unknown. Let the following be given: $\mathcal{N}$ – a non-empty set of possible states of nature, $\mathcal{D}$ – a non-empty set of possible decisions, and the loss function $\ell : I\!\!N \times \mathcal{D} \to I\!\!R \cup \{\pm\infty\}$, in which its value $\ell(\nu, \delta)$ is interpreted as losses resulting from making the decision δ while in reality the state ν is occurring. If the probability space $(\mathcal{N}, \mathcal{S}, \mathcal{P})$ is defined on the set $\mathcal{N}$, and for every $\delta \in \mathcal{D}$ the integral $\int_{\mathcal{N}} \ell(\nu, \delta)\, \mathrm{d}\mathcal{P}(\nu)$ exists, then the mapping $\ell_b : \mathcal{D} \to I\!\!R \cup \{\pm\infty\}$ given as

$$\ell_b(\delta) = \int_{\mathcal{N}} \ell(\nu, \delta)\, \mathrm{d}\mathcal{P}(\nu) \tag{78}$$

is called the Bayes loss function. Every element $\delta_b \in \mathcal{D}$ such that

$$\ell_b(\delta_b) = \inf_{\delta \in \mathcal{D}} \ell_b(\delta) \tag{79}$$

is known as a Bayes decision, while the above procedure is known as the Bayes rule. Its underlying purpose is therefore to minimize the expected value of losses.

In the diagnosis problem here under consideration, for any fixed index k, it is therefore assumed that the set of states of nature $\mathcal{N}_k$ is two-element: ν_k – the occurrence of the kth type of fault to be diagnosed, and ν_0 – proper system operation. Similarly, the set of decisions that can possibly be made $\mathcal{D}_k$ takes on the form: δ_k – accepting the hypothesis H_k stating that the kth type of fault to be diagnosed has occurred, and δ_0 – accepting the hypothesis H_0 representing proper operation. Therefore, if $\mathsf{a}_k > 0$ and $\mathsf{b}_k > 0$ mean the losses incurred by neglecting the kth type of fault and the corresponding false alarm respectively, then the loss function ℓ assumes the form

$$\ell(\nu, \delta) = \begin{cases} 0 & \text{when the state } \nu_k \text{ occurs and the decision } \delta_k \text{ is made} \\ 0 & \text{when the state } \nu_0 \text{ occurs and the decision } \delta_0 \text{ is made} \\ \mathsf{a}_k & \text{when the state } \nu_k \text{ occurs and the decision } \delta_0 \text{ is made} \\ \mathsf{b}_k & \text{when the state } \nu_0 \text{ occurs and the decision } \delta_k \text{ is made.} \end{cases} \tag{80}$$

If the decision δ_k is taken, then the value of Bayes loss function (78) is

$$\ell_b(\delta_k) = \mathsf{a}_k \alpha_k, \tag{81}$$

whereas in the case of the decision δ_0:

$$\ell_b(\delta_0) = \mathsf{b}_k \beta_k, \tag{82}$$

while

$$\alpha_k = P_k(\{\omega \in \Omega : S_k(\omega, j) \in A_k\}) = P_k(\{\omega \in \Omega : f_k(Z(\omega, j)) \leq a_k\}) \quad (83)$$

$$\begin{aligned}\beta_k &= P_0(\{\omega \in \Omega : S_k(\omega, j) \in B_k\}) = P_0(\{\omega \in \Omega : f_k(Z(\omega, j)) > a_k\}) \\ &= 1 - P_0(\{\omega \in \Omega : f_k(Z(\omega, j)) \leq a_k\}), \end{aligned} \quad (84)$$

where P_k and P_0 denote in succession the probabilities in the case of the occurrence of the kth type of fault to be diagnosed and for proper working conditions. According to the principle of the Bayes rule (79), the decision d_k should be made if $\mathsf{a}_k \alpha_k \leq \mathsf{b}_k \beta_k$, whereas the decision δ_0 whenever $\mathsf{a}_k \alpha_k \geq \mathsf{b}_k \beta_k$. The critical value a_k is thus to be specified in such a way as to fulfill the condition

$$\mathsf{a}_k \alpha_k = \mathsf{b}_k \beta_k, \quad (85)$$

i. e. ultimately, after taking into account dependencies (83) – (84):

$$P_0(\{\omega \in \Omega : f_k(Z(\omega, j)) \leq a_k\}) + \frac{\mathsf{a}_k}{\mathsf{b}_k} P_k(\{\omega \in \Omega : f_k(Z(\omega, j)) \leq a_k\}) = 1. \quad (86)$$

The foregoing condition is equivalent to equation (34), whose verification was the purpose of this section.

Appendix B: Proof of the formal correctness of the procedure for fault diagnosis

Theorem 1. Let:

(A) $c > 0$;

(B) X, Y_0, Y_1 represent n-dimensional random variables, defined on the same probability space; their distributions have density functions;

(C) $f_X$denote a density function of the distribution of the random variable X, while $\hat{f}_X$ is its strongly consistent kernel estimator, calculated on the basis of an m_X-element random sample, with the application of a kernel such that the inverse image of any real number is a zero-measure set;

(D) the mappings f_X and $\hat{f}_X$ be Borelian;

(E) $a \in I\!R$ constitute a unique solution of the equation

$$F_{f_X \circ Y_0}(a) + c F_{f_X \circ Y_1}(a) = 1, \quad (87)$$

while $F_{f_X \circ Y_0}$ and $F_{f_X \circ Y_1}$ denote distribution functions of the random variables $f_X \circ Y_0$ and $f_X \circ Y_1$, respectively;

(F) $\hat{a} \in I\!R$ be a solution of the equation

$$\hat{F}_{\hat{f}_X \circ Y_0}(\hat{a}) + c \hat{F}_{\hat{f}_X \circ Y_1}(\hat{a}) = 1, \quad (88)$$

where $\hat{F}_{\hat{f}_X \circ Y_0}$ and $\hat{F}_{\hat{f}_X \circ Y_1}$ represent kernel estimators of the distribution functions of the variables $\hat{f}_X \circ Y_0$ and $\hat{f}_X \circ Y_1$, calculated on the basis of random samples with the sizes m_0 and m_1, while in both cases the values of the smoothing parameters are dependent on their sizes in accordance with conditions (11) – (12).

Then, with probability 1:

$$\lim_{m_X, m_0, m_1 \to \infty} \hat{a} = a; \tag{89}$$

therefore, $\hat{a}$ is a strongly consistent estimator of the quantity a.

Proof. From the strong consistency of the estimator $\hat{f}_X$ it results that with probability 1:

$$\hat{f}_X \circ Y_0 \overset{m_X \to \infty}{\longrightarrow} f_X \circ Y_0. \tag{90}$$

This implies the weak convergence, and consequently the convergence of the distribution functions

$$F_{\hat{f}_X \circ Y_0}(t) - F_{f_X \circ Y_0}(t) \overset{m_X \to \infty}{\longrightarrow} 0 \tag{91}$$

at the points of continuity of the mapping $F_{f_X \circ Y_0}$.

The following dependence, in turn, is true:

$$\begin{aligned} \left|\hat{F}_{\hat{f}_X \circ Y_0}(t) - F_{\hat{f}_X \circ Y_0}(t)\right| &= \left|\int_{-\infty}^{t} \hat{f}_{\hat{f}_X \circ Y_0}(s)\,\mathrm{d}s - \int_{-\infty}^{t} f_{\hat{f}_X \circ Y_0}(s)\,\mathrm{d}s\right| \\ \leq \int_{-\infty}^{t} \left|\hat{f}_{\hat{f}_X \circ Y_0}(s) - f_{\hat{f}_X \circ Y_0}(s)\right| \mathrm{d}s &\leq \int_{-\infty}^{\infty} \left|\hat{f}_{\hat{f}_X \circ Y_0}(s) - f_{\hat{f}_X \circ Y_0}(s)\right| \mathrm{d}s, \end{aligned} \tag{92}$$

where $f_{\hat{f}_X \circ Y_0}$ and $\hat{f}_{\hat{f}_X \circ Y_0}$ denote the density function of the random variable $\hat{f}_X \circ Y_0$ and its kernel estimator, respectively. (The existence of this function results from Assumptions (B) – (C) on the basis of the Radon–Nikodym Theorem [4].) If conditions (11) – (12) are fulfilled, then the right-hand side of inequality (92) is convergent to zero with probability 1, thanks to the strong consistency of the kernel estimators of the density functions in the norm L_1 [6]. This entails the convergence of the left side as well, i. e.

$$\hat{F}_{\hat{f}_X \circ Y_0}(t) - F_{\hat{f}_X \circ Y_0}(t) \overset{m_0 \to \infty}{\longrightarrow} 0 \tag{93}$$

with probability 1.

Now:

$$\hat{F}_{\hat{f}_X \circ Y_0}(t) - F_{f_X \circ Y_0}(t) = \hat{F}_{\hat{f}_X \circ Y_0}(t) - F_{\hat{f}_X \circ Y_0}(t) + F_{\hat{f}_X \circ Y_0}(t) - F_{f_X \circ Y_0}(t); \tag{94}$$

thus, thanks to formulas (91) and (93), it results that with probability 1

$$\hat{F}_{\hat{f}_X \circ Y_0}(t) - F_{f_X \circ Y_0}(t) =^{m_X, m_0 \to \infty}{\longrightarrow} 0 \tag{95}$$

at the points of continuity of the mapping $F_{f_X \circ Y_0}$.

Analogously, the following dependence is true with probability 1:

$$\hat{F}_{\hat{f}_X \circ Y_1}(t) - F_{f_X \circ Y_1}(t) \overset{m_X, m_1 \to \infty}{\longrightarrow} 0 \tag{96}$$

at the points of continuity of the mapping $F_{f_X \circ Y_1}$.

In order to prove thesis (89) it is necessary to show that with any fixed $\varepsilon > 0$ and for sufficiently large m_X, m_0 and m_1

$$\hat{a} \in (a - \varepsilon, a + \varepsilon) \tag{97}$$

with probability 1. Since a distribution function of a probability measure can have at most a countable number of discontinuities, there exist the numbers $t^{\sim}, t^{\approx} \in R$, in which the mappings $F_{f_X \circ Y_0}$ and $F_{f_X \circ Y_1}$ are continuous, and the following condition is fulfilled:

$$a - \varepsilon < t^{\sim} < a < t^{\approx} < a + \varepsilon. \tag{98}$$

The distribution function is also an increasing function, and so, due to the assumed uniqueness of the solution a, the following inequalities are true

$$F_{f_X \circ Y_0}(t^{\sim}) + cF_{f_X \circ Y_1}(t^{\sim}) < 1 \tag{99}$$

$$F_{f_X \circ Y_0}(t^{\approx}) + cF_{f_X \circ Y_1}(t^{\approx}) > 1, \tag{100}$$

i. e., thanks to formulas (95) and (96), for sufficiently large m_X, m_0 and m_1, with probability 1 there also occurs

$$\hat{F}_{\hat{f}_X \circ Y_0}(t^{\sim}) + c\hat{F}_{\hat{f}_X \circ Y_1}(t^{\sim}) < 1 \tag{101}$$

$$\hat{F}_{\hat{f}_X \circ Y_0}(t^{\approx}) + c\hat{F}_{\hat{f}_X \circ Y_1}(t^{\approx}) > 1, \tag{102}$$

which – taking into account the form of equation (88) and the continuity of the kernel estimator of the distribution function – directly establishes the truth of condition (97), and consequently of the thesis to be proved. □

It results from the foregoing theorem – by means of obvious replacements – that according to the statement made in Section 4.1, if a_k $(k = 1, 2, \ldots, d)$ constitutes a solution of equation (34), then $\hat{a}_k$, as a solution of dependence (35), is its strongly consistent kernel estimator.

Appendix C: Proof of the formal correctness of the procedure for fault detection

Theorem 2. Let:

(A) $c \in (0, 1)$;

(B) X, Y represent n-dimensional random variables, defined on the same probability space; their distributions have density functions;

(C) f_X denote a density function of the distribution of the random variable X, while $\hat{f}_X$ is its strongly consistent kernel estimator, calculated on the basis of an m_X-element random sample, with the application of a kernel such that the inverse image of any real number is a zero-measure set;

(D) the mappings f_X and $\hat{f}_X$ be Borelian;

(E) $a \in \mathbb{R}$ constitute a unique solution of the equation

$$F_{f_X \circ Y}(a) = c, \tag{103}$$

while $F_{f_X \circ Y}$ denotes a distribution function of the random variable $f_X \circ Y$;

(F) $\hat{a} \in I\!R$ be a solution of the equation

$$\hat{F}_{\hat{f}_X \circ Y}(\hat{a}) = c, \tag{104}$$

where $\hat{F}_{\hat{f}_X \circ Y}$ represents a kernel estimator of the distribution function of the variable $\hat{f}_X \circ Y$, calculated on the basis of an m-element random sample, while the value of the smoothing parameter is dependent on its size in accordance with conditions (11) – (12).

Then, with probability 1:

$$\lim_{\substack{m_X \to \infty \\ m \to \infty}} \hat{a} = a; \tag{105}$$

therefore, $\hat{a}$ is a strongly consistent estimator of the quantity a.

Proof. Just as in the case of dependence (95), it can be shown that with probability 1:

$$\hat{F}_{\hat{f}_X \circ Y}(t) - F_{f_X \circ Y}(t) \overset{m_X, m \to \infty}{\longrightarrow} 0 \tag{106}$$

at the points of continuity of the mapping $F_{f_X \circ Y}$. Analogously to formula (97), one can prove that with any fixed $\varepsilon > 0$ and for sufficiently large m_X and m, with probability 1 there also occurs

$$\hat{a} \in (a - \varepsilon, a + \varepsilon), \tag{107}$$

which ultimately states the truth of the thesis of the present theorem. □

Thus it can be inferred that if a_0 represents the solution of equation (45), then $\hat{a}_0$, being a solution of dependence (48), is its strongly consistent estimator. Reference was made to this fact in Section 4.2.

(Received April 23, 2001.)

REFERENCES

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

[2] M. Basseville and I. V. Nikiforov: Detection of Abrupt Changes – Theory and Applications. Prentice–Hall, Englewood Cliffs, N.J. 1993.

[3] J. O. Berger: Statistical Decision Theory. Springer–Verlag, New York 1980.

[4] P. Billingsley: Probability and Measure. Wiley, New York 1979.

[5] J. Chen and R. J. Patton: Robust Model–Based Fault Diagnosis for Dynamic Systems. Kluwer, Boston 1999.

[6] L. Devroe and L. Györfi: Nonparametric Density Estimation: the L_1 View. Wiley, New York 1985.

[7] M. L. Dertouzos, M. Athans, R. N. Spann, and S. J. Mason: Systems, Networks, and Computation. McGraw–Hill, New York 1972.

[8] M. Fisz: Probability Theory and Mathematical Statistics. Wiley, New York 1963.

[9] J. J. Gertler: Fault Detection and Diagnosis in Engineering Systems. Dekker, New York 1998.

[10] P. Kulczycki: Almost certain time-optimal positional control. IMA J. Math. Control Inform. *13* (1996), 63–77.

[11] P. Kulczycki: An algorithm for Bayes parameter identification. Journal of Dynamic Systems, Measurement, and Control, Special Issue on the Identification of Mechanical Systems *123* (2001), 611–614.

[12] P. Kulczycki: A random approach to time-optimal control. Journal of Dynamic Systems, Measurement, and Control *121* (1999), 542–543.

[13] P. Kulczycki: A test for comparing distribution functions with strongly unbalanced samples, to appear.

[14] P. Kulczycki: Fault Detection in Automated Systems by Statistical Methods. Alfa, Warsaw 1998.

[15] P. Kulczycki: Fuzzy controller for mechanical systems. IEEE Trans. on Fuzzy Systems *8* (2000), 645–652.

[16] P. Kulczycki and A. L. Dawidowicz: Kernel estimator of quantile. Univ. Iagel, Acta Math. *37* (1999), 325–336.

[17] P. Kulczycki and R. Wisniewski: Fuzzy controller for a system with uncertain load, to appear.

[18] R. S. Mangoubi: Robust Estimation and Failure Detection. Springer–Verlag, London 1998.

[19] R. S. Parrish: Comparison of quantile estimators in normal sampling. Biometrics *46* (1990), 247–257.

[20] B. L. S. Prakasa Rao: Nonparametric Functional Estimation. Academic Press, Orlando 1983.

[21] H. Schiølerand and P. Kulczycki: Neural network for estimating conditional distributions. IEEE Trans. Neural Networks *8* (1997), 1015–1025.

[22] S. J. Sheather and J. S. Marron: Kernel quantile estimators. J. Amer. Statist. Assoc. *85* (1990), 410–416.

[23] B. W. Silverman: Density Estimation for Statistics and Data Analysis. Chapman and Hall, London 1986.

[24] B. Sohlberg: Supervision and Control for Industrial Processes. Springer–Verlag, London 1998.

[25] M. P. Wand and M. C. Jones: Kernel Smoothing. Chapman and Hall, London 1994.

[26] M. West and J. Harrison: Bayesian Forecasting and Dynamic Models. Springer–Verlag, New York 1989.

Doc. Piotr Kulczycki, D.Sc., Ph.D., Systems Research Institute, Polish Academy of Sciences, ul. Newelska 6, PL-01-447 Warsaw. Poland.
e-mail: kulczycki@ibspan.waw.pl