Jurij GRISZIN

Bialystok University of Technology

doi:10.15199/48.2017.12.04

Adaptive data processing in the presence of outliers with unknown stochastic characteristics

***Abstract**. The paper presents a new adaptive algorithms of data processing when observed data are corrupted by pulse interferences which cause outliers in measurements. It is assumed that stochastic characteristics of the outliers are unknown. The algorithm is based on the nonlinear filtering approach and procedure of on-line calculation of the stochastic characteristics of the outliers described by the Markov chains. The simulation results are presented*

Streszczenie.** W artykule został przedstawiony nowy adaptacyjny algorytm przetwarzania danych zniekształconych zakłóceniami impulsowymi które skutkują anomaliami w pomiarach. Algorytm został opracowany z wykorzystaniem metod oceny stochastycznych parametrów anomalii opisanych łańcuchami Markowa. Przedstawiono wyniki symulacji. (**Adaptacyjne przetwarzanie danych w obecności anomalii o nieznanych charakterystykach stochastycznych).

Keywords: adaptive filtering; robust estimation; pulse interferences; compatibility.
Słowa kluczowe: filtracja adaptacyjna; estymacja krzepka; zakłócenia impulsowe; kompatybilność elektromagnetyczna.

Introduction

The problem of reliable data processing is one of the most important in control theory, telecommunications, telemetric systems, industrial measurement equipment and in other applications. Pulse interferences are the main cause of the outliers in observed data which lead to significant distortions in the results.

The main sources of the outliers in telecommunications systems and industrial measurement devices are [1]: atmospheric radio noise bursts; industrial and vehicular radio frequency pulse interferences; non-controlled phase jumps in transmitter and receiver equipment; distortion of code words in digital communication channels; intentional jamming and some other.

Effective data processing algorithms in the presence of outliers can be developed using proper statistical description of normal and abnormal models of measurement noise and estimated data [2].

Usually the probability density function (pdf) of the normal measurement noise v can be described by the Gaussian pdf $N(v_e, \sigma_v^2)$. The pdf of the outlier can be completely unknown or can be approximated by the pdf with long “tails” (e.g. uniform or Laplace’s pdf). If the statistical characteristics of the data stream x_k and measurement noise v_k are completely known the Bayes’s approach can be used. In a case when the data a priori distributions are uncertain but the measurement noise pdf are known the maximum likelihood (ML) method is widely used for process estimation

For the Laplace’s pdf the ML method results to the median filters. Such estimates are known as the least modulus estimates (LME) [3]. If a priori information concerning statistical characteristics of observed processes is not available or is not reliable one can use so called nonparametric methods of mathematical statistics. They include linear combination of the order statistics (so called L-estimates) and rank statistics (R-estimates) [4]. P. Huber [5] developed the minimax approach to the robust estimation of random values based on the influence functions introduced by F. Hampel [6]. These estimates were called by M-estimates (maximum likelihood estimates under non-standard conditions).

Considerably less attention was paid to the robust estimation of random processes (measurement data) observed in presence of noise and outliers. To cope with this problem we are using nonlinear filtering approach [2]. In this case first of all it is necessary to describe an observation model which depends on a real cause of the pulse interferences and a priori statistical characteristics of the observed processes.

We can present the model of the outliers at the input of the digital estimation filter in the following form [2]:

$$y(k) = H(k)x(k) + \gamma(k)v(k) \tag{1}$$

where γ(k) is the measurement vector, *H(k)*.- the observation matrix, *x(k)* is the data process (the state vector), *v(k)* is the white Gaussian sequences with zero mean and covariance matrixes *R(k)*. The outliers in the measurement equation can be described by the random multiplier γ(k) which can take on values of 1 when the outliers are absent (normal operation) and $\gamma(k) = \sigma >> 1$ when they arise.

A random binary switching function $\gamma(k)$ generally can consist of correlated values with known or unknown a priori statistical characteristics. The models described by the equation (1) is typical for digital data transmission in the presence of transmission noise and pulse interferences.

In practice the probability of switching function γ(k) is unknown and the pulse interferences arise either independently at any one instant of discrete time or can be correlated. Thus the procedure of data processing in such conditions has to be adaptive and includes estimation of unknown stochastic characteristics of the switching function.

In the paper two models of switching function $\gamma(k)$ are studied: 1). γ(k) is independent at any instant of time sequence with unknown probability of arising and 2). γ(k) is the Markov sequence with unknown elements of transition matrix which describes its stochastic properties.

In this paper a new adaptive filtering algorithm robust with respect to the pulse interference with unknown probability of arising and stochastic properties has been developed using the nonlinear filtering approach [2].

Filtering algorithm based on the nonlinear approach

The adaptive filtering algorithm at issue can be designed on the basis of nonlinear suboptimal filter developed in [2] on the assumption that the probabilities of the outliers arising are known.

Let us briefly consider this algorithm. If the outliers are absent the state estimates of the linear dynamic system can be found using the Kalman filter designed on basis of information on system dynamics and observation model [8].

For a linear dynamic system described by the state equation

(2) $$x_{k+1} = \Phi_{k+1,k} x_k + G_{k+1,k} w_k$$

and the observation equation

(3) $$y_k = H_k x_k + \gamma_k v_k$$

where γ_k =1 in the absence of the outliers and $\gamma_k = \sigma$ >>1 when the outliers occur.

Estimation of the state vector at the output of the Kalman filter can be written in the following form [8]:

(4) $$\hat{x}_{k/k} = \hat{x}_{k/k-1} + K_k[y_k - H_k \hat{x}_{k/k-1}])$$

where $\hat{x}_{k/k}$ is the filtering estimate, $\hat{x}_{k/k-1}$ is the prediction estimate:

(5) $$\hat{x}_{k/k-1} = \Phi(k, k-1)\,\hat{x}_{k-1/k-1}$$

Matrix gain

(6) $$K_k = P_{k/k-1} H^T [H P_{k/k-1} H^T + R_k]^{-1}$$

Covariance matrix of predicted errors

(7) $$P_{k/k-1} = \Phi P_{k-1/k-1} \Phi^T + G Q_k G^T$$

Covariance matrix of filtering errors

(8) $$P_{k/k} = [I - K_k H] P_{k/k-1}$$

In a case when the outliers at the input of the filter occur with a priori known probability of arising *q* the estimation of the state vector can be obtained as the weighted sum of the partial estimates $\hat{x}^{(i)}_{k/k}$ corresponding to presence and absence of the outliers in the measurements:

(9) $$\hat{x}_{k/k} = \sum_{i \in 1,\sigma} \hat{x}_{k/k}(\gamma_k = i) P(\gamma_k = i / Y_1^k)$$

The a posterior probability of the measurement channel state $P(\gamma_k = i / Y_1^k)$ depends on the outlier stochastic characteristics. If the outliers are statistically independent the probability can be found as [2, 7]:

(10) $$p_{1/k} = \frac{f(y_k / \gamma_k = 1, Y_1^{k-1})\, p_{1/k-1}}{\sum_{i=1,\sigma} f(y_k / \gamma_k = i, Y_1^{k-1})\, p_{i/k-1}}$$

where $P(\gamma_k = 1 / Y_1^k) = p_{1/k}$ is the a posterior probability of the outliers absence in the observation which can be calculated in real time using current data at the filter input based on the pdf $f(y_k / \gamma_k = i, Y_1^{k-1})$ of predicted estimates.

These probabilities $p_{1/k}$ are used for control of the matrix gain of the filter: should be as:

(11) $$\hat{x}_{k/k} = \hat{x}_{k/k-1} + p_{1/k} K_{1k}\, [y(k) - H_k\, \hat{x}_{k/k-1}]$$

The structure of the filter is presented in Fig.1.

The matrix gain of the filter depends on the current realizations y_k due to the a posterior probability of the outliers absence $p_{1/k}$. It has to be calculated in real time according to the following expression:

(12) $$K_{1k} = P_{k/k}\, H_k^T\, R_k^{-1}$$

where covariance matrix can be presented as:

(13) $$P_{k/k} = P_{k/k-1} - p_{1/k}\, K_{1k}\, H_k\, P_{k/k-1} + p_{1/k}(1 - p_{1/k}) K_{1k}\, S_k\, K_{1k}^T$$

where

(14) $$S_k = [y_k - H_k\, \hat{x}_{k/k-1}][y_k - H_k\, \hat{x}_{k/k-1}]^T$$

is the filter innovation process.

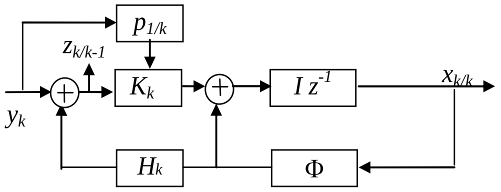

Fig.1. Structure of the robust suboptimal filter (RSF)

Dependence of the filter matrix gain from observations in (11) and (13) makes it nonlinear. It can be shown that the product $p_{1/k} K_{1k}$ in (11) is an equivalent to the Huber "influence function". If the value of the outlier is large then the a posterior probability of the measurement cannel state $p_{1/k}$ converges to zero and general matrix gain is equal to zero as well, the feedback loop in Fig 1. will be disconnected and at the filter output the predicted estimate will appear.

Calculation of $p_{1/k}$ according to (10) depends on a priori information about probabilities of arising of the outliers (pulse interferences). If these probabilities are known and outliers are independent at any one instant of time for calculation of the value $p_{1/k}$ in (10) we have to use $p_{1/k} = q$, where q is known a priori probability of the outliers arising. This case is considered in [2].

However in practice such a situation is not real. Usually probabilities of pulse arising are not known and as a result estimates of the data are not optimal. In this paper a new adaptive filtering algorithm is developed based on the equation (11) and a new procedure of on-line calculation of the probabilities q which determine the value of the a posterior probability $p_{1/k}$. In the paper two cases are considered. The first one deals with independent in time outliers with unknown probability of arising q . The second-studies the situation with correlated in time outliers. In the last case the outliers sequence has been modelled as the Markov chain with unknown elements of the transition matrix. The derivation of the algorithms are presented in the next two sections.

Calculation of probabilities $p_{1/k}$ for independent in time outliers

If probabilities *q* a priori are not known then they can be described as random values with the uniform probability density function on the interval [0, 1].

For a priori known probability *q* the a posterior probability of the observation channel state can be written as:

(15) $$p_{1/k} = \frac{f(y_k / \gamma_k = 1, Y_1^{k-1})\; q}{\sum_{i=1,\sigma} f(y_k / \gamma_k = i, Y_1^{k-1})\; q}$$

where $f(y_k/\gamma_k = i, Y_1^{k-1})$ - conditional PDF for normal ($\gamma_k = 1$) and abnormal ($\gamma_k = \sigma$) measurements.

Using the smoothing properties of the conditional statistical expectation [9] the a posterior probability of the measurement channel state can be found as:

$$p(1/k) = E_q\{P[\gamma(k)=1]/Y_1^k, q/Y_1^k]\} =$$

(16) $$= E_q\{\frac{f_1(k)q}{f_1(k)q + f_\sigma(k)(1-q)}/Y_1^k\} =$$

$$= \int_0^1 \frac{f_1(k)q}{f_1(k)q + f_\sigma(k)(1-q)} f(q/Y_1^k)dq$$

where $f(q/Y_1^k)$ is a posterior pdf of unknown probability *q* which can be presented in a recurrent form using the Bayes' theorem:

(17) $$f(q/Y_1^k) = \frac{f[y(k)/q, Y_1^{k-1}]\, f[q/Y_1^{k-1}]}{\int_0^1 f[y(k)/q, Y_1^{k-1}]\, f[q/Y_1^{k-1}]dq}$$

with initial condition $f[q/y(0)] = 1]$ on the interval [0, 1].

The PDF $f[y(k)/q, Y_1^{k-1}]$ in (17) can be written in the following equivalent form:

(18) $$f[y(k)/q, Y_1^{k-1}] = f_1(k)q + f_\sigma(k)(1-q)]$$

Define $E[q/Y_1^{k-1}]$ by

(19) $$\overline{q(k-1)} = E[q/Y_1^{k-1}] = \int_0^1 q\, f(q/Y_1^{k-1})dq$$

then the expression (17) takes the form

(20) $$f(q/Y_1^k) = \frac{[f_1(k)q + f_\sigma(k)(1-q)]\, f(q/Y_1^k)}{f_1(k)\overline{q(k-1)} + f_\sigma(k)(1-\overline{q(k-1)})}$$

and the a posterior probability of the measurement channel state can be written as the following:

(21) $$p(1/k) = \frac{f_1(k)\overline{q(k-1)}}{f_1(k)\overline{q(k-1)} + f_\sigma(k)(1-\overline{q(k-1)})}$$

Thus the adaptive algorithm for calculation of $p(1/k)$ uses the value $\overline{q(k-1)}$ instead of a priori known probability *q*. The probabilities $\overline{q(k-1)}$ are evaluated recurrently according to equations (17) - (19).

Calculation of probabilities $p_{1/k}$ for outliers described by Markov chain with unknown transition matrix

Correlated in time outlier sequences with unknown stochastic characteristics can be modelled by the simple Markov chain [2]. The elements of the transition matrix are unknown and have to be estimated on-line. Estimated transition matrix is used for calculation of the a posterior probability of the measurement channel state $p_{1/k}$ as is the case in the previous section.

It is assumed that elements of the transition matrix $\mathbf{P}_{ij}$

(22) $$\mathbf{P}_{ij} = \begin{bmatrix} p_{\sigma\sigma} & p_{\sigma 1} \\ p_{1\sigma} & p_{11} \end{bmatrix}$$

do not change on the time observation interval. In fact the matrix is defined by only two numbers- $p_{\sigma\sigma}$ and p_{11}. Suppose that joint pdf of values $p_{\sigma\sigma}$ and p_{11} is the uniform in the square [0,1]×[0,1].

Then the procedure of calculation $p_{\sigma\sigma}$ and p_{11} is similar to that used in the previous section. Omit details of the algorithm derivation. We can present the desired algorithm in the following steps:

1). Calculation of the marginal pdf $f[p_{\sigma\sigma}/Y_1^k]$ and $f[p_{11}/Y_1^k]$ with the initial conditions $f[p_{\sigma\sigma}/y(0)] = f[p_{11}/y(0)] = 1$, p_σ^0, p_1^0:

(23)

$$f(p_{\sigma\sigma}/Y_1^k) = \frac{\sum_{j=1,\sigma} f_j(k)[p_{j\sigma}p(\sigma/k-1)] + \overline{p_{j1}}(k-1)p(1/k-1)]\, f(p_{\sigma\sigma}/Y_1^{k-1})}{\sum_{j=1,\sigma} f_j(k) \sum_{i=1,\sigma} \overline{p_{ji}}(k-1)p(i/k-1)]}$$

(24)

$$f(p_{11}/Y_1^k) = \frac{\sum_{j=1,\sigma} f_j(k)[\overline{p_{j\sigma}}p(\sigma/k-1)] + \overline{p_{j1}}(k-1)p(1/k-1)]\, f(p_{11}/Y_1^{k-1})}{\sum_{j=1,\sigma} f_j(k) \sum_{i=1,\sigma} \overline{p_{ji}}(k-1)p(i/k-1)]}$$

2). Calculation of the expected values $\overline{p_{ii}}(k-1), i = 1, \sigma)$

(25) $$\overline{p_{ii}}(k-1) = E[p_{ii}/Y_1^{k-1}] = \int_0^1 p_{ii}\, f(p_{ii}/Y_1^{k-1})dp_{ii}$$

3). Calculation of the final value of the a posterior probability $p(1/k)$ according to:

(26) $$p(1/k) = \frac{f_1(k)\sum_{i=1,\sigma} \overline{p_{1j}}(k-1)p(i/k-1)]}{\sum_{j=1,\sigma} f_j(k) \sum_{i=1,\sigma} \overline{p_{ji}}(k-1)p(i/k-1)]}$$

with the initial condition

(27) $$p(1/0) = \frac{f_1(0)\, p_1^0}{f_1(0)\, p_1^0 + f_\sigma(0)\, p_\sigma^0}$$

Simulations results

As an example let us consider the data process described by a scalar state equation

$$x_{k+1} = \alpha x_k + w_k$$

where α = 0.9, σ^2_w = 4 10^{-4} and observation equation of the following form:

$$y_k = x_k + \gamma_k v_k$$

where σ^2_v = 25 10^{-4}, γ_k = 1 (normal measurements) with probability *q* = 0.8 (not known to the observer) and γ_k = 10 (outliers) with probability 0.2. It is assumed that the outliers are independent in time.

At first suppose that the probability of the outliers arising are known (*q*=0.8). Then the robust suboptimal filter (RSF) is described by expressions (11)-(15). The variance of the estimation error *P(k/k)* for RSF is showed in Fig.2 (curve 1). For comparison in the same figure the variance of the estimation error *P(k/k)* for the Kalman filter (KF) is presented (curve 2).

As it follows from the schedules calculation of the a posterior probability $p(1/k)$ makes it possible to considerably decrease estimation errors of the SRF in comparison with the KF.

If the a prior probability *q* is not known in this case the adaptive algorithm can be realized using equations (11) - (14) for estimation and equation (21) for calculation of a posterior probability $p(1/k)$. Conditional probability density function $f(q/Y_1^k)$ in (17) can be approximated by discrete function of the following form:

$$f(q/Y_1^k)=\sum_{j=1,N} q_j\delta(q-q_j)$$

where the number of samples *N* determines the accuracy of approximation. In simulations *N* was chosen 50. The results of simulation are shown in Fig.3.

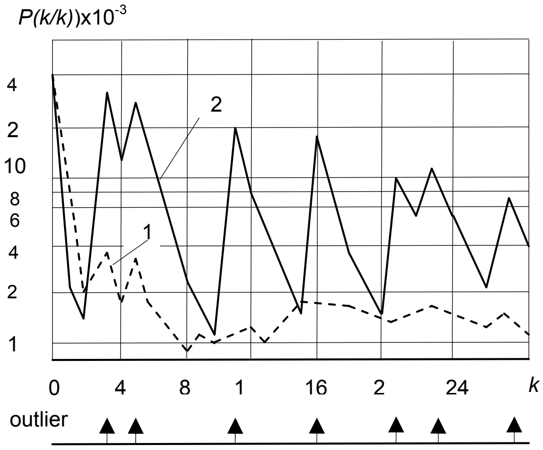

Fig.2. Variances of estimation errors for SRF and KF

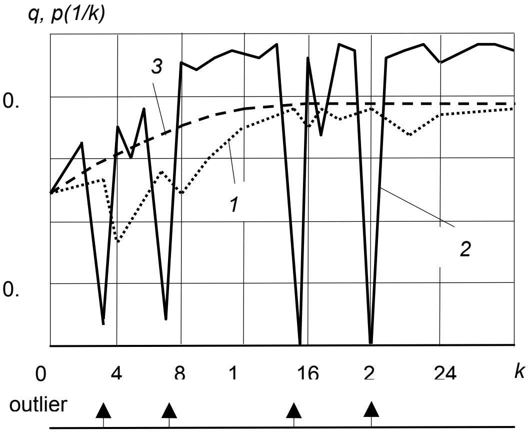

Fig. 3. Probabilities of the measurement channel state: *1* - $\overline{q(k-1)}$, *2* - $p(1/k)$, ***3*** - q_m

A typical realization of $\overline{q(k-1)}$ (curve 1) demonstrates convergence of the adaptation process to the unknown value of pulse interference (*q*=0.8). Dependence 2 shows current values of the a posterior probabilities $p(1/k)$ which control the Kalman filter gain. The average values of $\overline{q(k-1)}$ calculated using 100 realization of the input measurements are presented in Fig.1 by dependence 3.

As the additional simulations results show the accuracy of estimation of measurements with using probabilities calculated on the base of the proposed algorithms (17) – (21) and (23)-(27) are practically the same as in a case of known *q.*

Conclusion

In the paper the problem of developing the adaptive robust algorithms of data processing in industrial measurement devices and telecommunication systems when observed data are corrupted by pulse interferences of unknown probability of arising has been solved. As the base structure was chosen the nonlinear filter designed with using the Gaussian approximation approach. The unknown probability of the interference pulse arising is estimated in real time. The estimated value of the probability is used for changing the nonlinear filter gain. Due to these changes the influence of pulses interferences is eliminated**.**

The proposed algorithm has a recursive structure and can be easily implemented on the basis of the DSP technology. The results of numerical simulations have revealed a high efficiency of the algorithm in measurement devices and telecommunication applications. It can be also used for fault detection and identification in industrial control processes.

Acknowledgment*: The research was conducted within the project S/WE/1/2015, financially supported by Polish Ministry of Science and Higher Education.*

Author*: prof. dr hab. inż. Jurij Griszin, Politechnika Białostocka, Wydział Elektryczny, ul. Wiejska 45 D, 15-351 Białystok, E-mail: j.griszin@pb.edu.pl.*

REFERENCES

[1] J.G. Proakis, M.Salehi, Communication system engineering. Prentice-Hall, Inc. New Jersey, 2002.

[2] J. Griszin, D. Jańczak, A detection-estimation method for systems with random jumps with application to target tracking and fault diagnosis, in “Nonlinear dynamics”/Ed. T. Evans, chapter 15, INTEH, 2010, pp. 343-366.

[3] Patton R., Frank P., Clark R., Fault diagnosis in dynamic systems. Theory and applications. Prentice Hall, N.Y., 1989.

[4] Hajek J., Sidak Z.,Theory of rank tests. Academia, Prague, 1967.

[5] Huber P.J., Robust statistics. John Wiley & Sons, Inc., N.Y.,1981.

[6] Hampel F.R., Ronchetti E.M, Rousseeuw P.J., Stahel W.A., Robust statistics. The approach based on influence functions. John Wiley & Sons, Inc., N.Y.,1986.

[7] J. Griszin, D. Jańczak A robust fixed-lag smoothing algorithm for dynamic systems with correlated sensor malfunctions. *Bulletin of the Polish Academy of Sciences, Technical Sciences,* v.62, NR 3, 2014, pp. 517-523.

[8] Balakrishnan A.V., Kalman filter theory. Optimization Software, Inc., N.Y.,1984.

[9] Feller W., An introduction to probability theory and its applications. John Wiley & Sons, N.Y., 1957