Robust Akaike's Criterion for Model Order Selection

Vladimir Stojanovic[1*], Vojislav Filipovic[1]
[1]Faculty of Mechanical and Civil Engineering in Kraljevo, Department of Energetics and Automatic Control,
University of Kragujevac (Serbia)

The paper considers the model order selection (Output Error model) of the system with constant parameters. Ad hoc selection of model order leads to overparametrization or parsimony problem. To avoid these problems, different selection criterions of the model are used: AIC (Akaike Information Criterion), BIC (Bayesian Information Criterion) and FPE (Final Prediction Error Criterion). In this paper, Akaike's criterion is used, which is obtained by minimization of the Kullback-Leibler information distance. The criterion is basically a generalization of the maximum likelihood method. It is assumed that the stochastic disturbance in the model belongs to the class of ε-contaminated distributions. In such conditions the originally proposed AIC criterion cannot be applied. By determining the least favourable probability density for a given class of probability distribution represents a base for design of the robust version of AIC criterion. Simulations illustrate the behavior of the proposed criterion.

Keywords: model order selection, output error model, ε-contaminated distributions, robust Akaike's criterion

1. INTRODUCTION

Obtaining system models based on the fundamental laws of physics is a difficult problem. In order to facilitate the controller design, for the obtained model, different simplifications of the model are performed. Most often it is a procedure of linearization around the equilibrium point. However, there are the systems that cannot be linearized around an equilibrium point, because there is no equilibrium point. If a linear approximation is found, the resulting model will be valid only for a small region around the linearization point. As an alternative approach the design of controllers is largely based on the use of mathematical models that are obtained during the process of system identification [1,2].

Most identification algorithm assume that the model structure is a priori known. As is well known, a fundamental difficulty in statistical analysis is the choice of an appropriate model and determining the order of a model. In recent years, the necessity of introducing the concept of model has been recognized and the problem is posed how to choose the "best approximating" model among a class of competing models with different numbers of parameters by a suitable model selection criterion given a data set. Also, there is presently a great deal of interest in simple criteria represented by parsimony of parameters for choosing one of a set of competing models to describe a given data set. Therefore, the best model is the one with least complexity, or equivalently the highest information. For example, parameter parsimony requires that the smallest number of factors is chosen, such that the corresponding model fits the data. The selection of a parsimonious model, in general, is a nontrivial problem without the aid of model selection criteria.

Several information theoretic criterion have been proposed for structure selection in linear dynamic input output models. The model which minimizes the criterion is then chosen as the best model from the available set. Examples of the classical criterions are the Final Prediction Error (FPE), Akaike's Information Criterion (AIC) and Bayesian Information Criterion (BIC). These techniques find a tradeoff between goodness of fit and model complexity. The performance of an order-selection criterion is optimal if the model of the selected order is the most accurate model in the considered set of estimated models. Note that this is not necessarily the true model order. If the true process is, e.g. tenth-order, where the last six parameters are insignificant, the estimated fourth-order model will be the most accurate.

Used way for deriving model selection criteria is based on the quantification of "how close are" the probability density of the generating model and the probability density of the fitted approximating model. Several coefficients or "measures" have been introduced in the literature for this quantification. The Kullback-Leibler information distance is the most frequently used information theoretic coefficient for measuring divergence or separation between two probability densities [3]. The Akaike's information criterion (AIC) is a commonly used tool for choosing between alternative models [4].

Here, those results are extended on the case when the measurement noise is a non-Gaussian. Justification of this approach was confirmed in practice [5,6]. Namely, in measurements there are rare, inconsistent observations with the largest part of population of observations (outliers). The presence of outliers can considerably degrade the performance of linearly recursive algorithms based on the assumptions that measurements have a Gaussian distribution.

The synthesis of robust algorithms is of primary interest. The synthesis is based on Huber's theory of robust statistics [6]. As a generator of a recursive algorithm, according Huber's theory, it is defined the functional based on the least favourable probability distribution for a given class of probability distribution. Robust recursive algorithms in the identification of

[*] Corresponding author: stojanovic.v@mfkv.kg.ac.rs

dynamical systems are discussed in [7] while in an area of adaptive control are discussed in [8].

This paper considers the model order selection using robust Akaike's criterion. The recursive algorithm for the OE (output error) model with time invariant parameters have been also discussed. Robustness of the used robust OE parameter estimation algorithm is accomplished by introducing the nonlinear transformation of prediction error (Huber’s function).

The performances of the algorithm are described through simulation results that demonstrate the superiority of the proposed algorithm in relation to the linear algorithm (derived under the assumption that the stochastic noise has a Gaussian distribution).

2. ROBUST PARAMETER ESTIMATION ALGORITHM FOR OE MODEL

The general form of the OE model is

$$y(k)=\frac{B(q^{-1})}{A(q^{-1})}u(k)+e(k) \tag{1}$$

where $u(k)$, $y(k)$ and $e(k)$ are input, output and stochastic noise, respectively. Polynomials $A(q^{-1})$ and $B(q^{-1})$ have the form:

$$\begin{aligned}A(q^{-1})&=1+a_1q^{-1}+\ldots+a_nq^{-n}\\ B(q^{-1})&=b_1q^{-1}+\ldots+b_mq^{-m}\end{aligned} \tag{2}$$

Practical and theoretical studies have shown that in a stochastic model of the system there are some observations that are inconsistent with the largest part of the population (outliers) [5], and that is why the disturbance (measurement noise) $e(k)$ in the model (1) is a non-Gaussian. Hence, the probability density function of the disturbance belongs to approximately normal distribution class:

$$\mathcal{P}_\varepsilon=\left\{p(e):p(e)=(1-\varepsilon)p_1(e)+\varepsilon p_2(e)\right\} \tag{3}$$

in which

$p_1(e)$ □ $\mathcal{N}(0,\sigma_1^2)$, $p_2(e)$ □ $\mathcal{N}(0,\sigma_2^2)$, σ_2^2 □ σ_1^2 .

In other words, the probability density function $p(e)$ represents a mixture of normal (Gaussian) distributions where σ_1^2 and σ_2^2 denote variances. The parameter $0\le\varepsilon<1$ is called the degree of contamination.

Let us introduce an auxiliary model

$$y_M(k)=\frac{B(q^{-1})}{F(q^{-1})}u(k), \tag{4}$$

or in the following form:

$$\begin{aligned}&y_M(k)=-f_1y_M(k-1)-\ldots-f_ny_M(k-n)+\\&+b_1u(k-1)+\ldots+b_mu(k-m)\end{aligned} \tag{5}$$

Since the parameters $a_i(i=1,\ldots,n)$ and $b_i(i=1,\ldots,m)$ are unknown, their estimates are used, so the output of the auxiliary model is calculated as:

$$\begin{aligned}&\hat{y}_M(k)=-\hat{f}_1\hat{y}_M(k-1)-\ldots-\hat{f}_n\hat{y}_M(k-n)+\\&+\hat{b}_1u(k-1)+\ldots+\hat{b}_mu(k-m)\end{aligned} \tag{6}$$

Let $\hat{\theta}$ is the estimated vector of OE parameters, and $\varphi(k)$ is the observation vector of OE parameters:

$$\begin{aligned}&\hat{\theta}=[\hat{f}_1,\ldots,\hat{f}_n,\hat{b}_1,\ldots,\hat{b}_m]^T,\\&\varphi(k)=[-\hat{y}_M(k-1)\ldots-\hat{y}_M(k-n),u(k-1)\ldots u(k-m)]^T\end{aligned} \tag{7}$$

At the moment k , before the estimate $\hat{\theta}(k)$ is known, the prediction of the model is [9]:

$$\hat{y}_M(k)=\hat{\theta}^T(k-1)\varphi(k)\ .$$

The natural definition of the prediction error (residual) is

$$\varepsilon(k)=y(k)-\hat{y}_M(k)\ . \tag{9}$$

The identification criterion (a generator of recursive parameter estimation procedure) is based, according to OE methodology, on the prediction error and has a mathematical form, for systems with constant parameters:

$$\mathcal{J}(\theta)=E\left\{\Phi(\varepsilon(\mathrm{k}))\right\} \tag{10}$$

in which

$$\Phi(\cdot)=-\log p^*(\cdot) \tag{11}$$

In the last relation, $p^*(\cdot)$ represents the least favourable distribution of probability for a given class of probability distribution (3).
This distribution is obtained by using the mathematical machinery of robust statistics [6].

An analytical description of the least favorable probability density $p^*(\cdot)$ is given as follows:

$$p^*(e(\mathrm{k}))=\begin{cases}\dfrac{1-\varepsilon}{2\pi\sigma_1}\exp\left\{-\dfrac{e^2(\mathrm{k})}{2\sigma_1^2}\right\} & |e(\mathrm{k})|\le k_\varepsilon\\[2ex] \dfrac{1-\varepsilon}{2\pi\sigma_1}\exp\left\{-\dfrac{k_\varepsilon}{\sigma_1^2}\left(|e(\mathrm{k})|-\dfrac{k_\varepsilon}{2}\right)\right\} & |e(\mathrm{k})|>k_\varepsilon\end{cases} \tag{12}$$

where k_ε is the Huber function parameter.

The empirical functional for systems with time-invariant parameters has the form (obtained from the relation (10) for sufficiently large *k*):

$$\mathrm{J}_k(\theta)=\frac{1}{k}\sum_{t=1}^{k}\left\{\Phi(\varepsilon(i))\right\} \tag{13}$$

Expanding $\mathcal{J}_k(\theta)$ in the vicinity of the preceding estimate $\hat{\theta}(\mathrm{k}-1)$ in Taylor series, one obtains:

$$\begin{aligned}&J_k(\theta)=J_k\left(\hat{\theta}(\mathrm{k}-1)\right)+\nabla_\theta J_k\left(\hat{\theta}(\mathrm{k}-1)\right)\left[\theta-\hat{\theta}(\mathrm{k}-1)\right]+\\&+O\left(\left\|\theta-\hat{\theta}(\mathrm{k}-1)\right\|^2\right)\end{aligned} \tag{14}$$

where

$$\lim_{\|x\|\to\infty}\frac{O\left(\|x\|\right)}{\|x\|}=0 \tag{15}$$

and $\|\cdot\|$ denotes the Euclidean norm. The desired value $\hat{\theta}(\mathrm{k})$ can be obtained by solving the equation:

$$\nabla_\theta J_k\left(\hat{\theta}(\mathrm{k})\right)=0 \tag{16}$$

from which one can obtain:

$$\hat{\theta}(\mathrm{k})=\hat{\theta}(\mathrm{k-1})-\left[k\nabla_\theta^2 J_k\left(\hat{\theta}(\mathrm{k-1})\right)\right]^{-1}\left[k\nabla_\theta J_k\left(\hat{\theta}(\mathrm{k-1})\right)\right]+ \\ +O\left(\left\|\theta-\hat{\theta}(\mathrm{k-1})\right\|\right) \tag{17}$$

Based on the relation (13) it is obtained:

$$J_k(\theta)=\frac{1}{k}\left[\frac{k-1}{k-1}\sum_{i=1}^{k-1}\Phi(\varepsilon(i))+\Phi(\varepsilon(k))\right]= \\ \frac{1}{k}\left[(k-1)\frac{1}{k-1}\sum_{i=1}^{k-1}\Phi(\varepsilon(i))+\Phi(\varepsilon(k))\right] \tag{18}$$

or in the form:

$$kJ_k(\theta)=(k-1)J_{k-1}(\theta)+\Phi(\varepsilon(k)) \tag{19}$$

By differentiating the last relation twice one can obtain:

$$k\nabla_\theta^2 J_k(\theta)=(k-1)\nabla_\theta^2 J_{k-1}(\theta)+\Psi^{'}(\varepsilon(k))\varphi(k)\varphi^T(k) \tag{20}$$

where $\Psi(\cdot)=\Phi^{'}(\cdot)$.

Let us assume further that the following assumptions are satisfied:

a) The estimate $\hat{\theta}(\mathrm{k})$ is in the vicinity of the estimate $\hat{\theta}(\mathrm{k-1})$

b) The estimate $\hat{\theta}(\mathrm{k-1})$ is optimal at the instant k-1.

Taking $\theta=\hat{\theta}(\mathrm{k-1})$ in the relation(20), one can obtain:

$$k\nabla_\theta^2 J_k(\hat{\theta}(\mathrm{k-1}))=(k-1)\nabla_\theta^2 J_{k-1}(\hat{\theta}(\mathrm{k-1}))+ \\ +\Psi^{'}(\varepsilon(k))\varphi(k)\varphi^T(k) \tag{21}$$

From the assumption a) follows

$$\nabla_\theta^2 J_k(\hat{\theta}(\mathrm{k}))\cong\nabla_\theta^2 J_k(\hat{\theta}(\mathrm{k-1})) \tag{22}$$

Based on this, the relation (21) takes the form

$$k\nabla_\theta^2 J_k(\hat{\theta}(\mathrm{k-1}))=(k-1)\nabla_\theta^2 J_{k-1}(\hat{\theta}(\mathrm{k-2}))+ \\ +\Psi^{'}(\varepsilon(k))\varphi(k)\varphi^T(k) \tag{23}$$

Based on the assumption a) it also follows

$$O\left(\left\|\hat{\theta}(\mathrm{k})-\hat{\theta}(\mathrm{k-1})\right\|\right)=0 \tag{24}$$

By introducing the notation $\overline{R}(\mathrm{k})=k\nabla_\theta^2 J_k(\hat{\theta}(\mathrm{k-1}))$ from relations (17) and (23) one can obtain:

$$\hat{\theta}(\mathrm{k})=\hat{\theta}(\mathrm{k-1})-\overline{R}^{-1}(\mathrm{k})\left[k\nabla_\theta J_k\left(\hat{\theta}(\mathrm{k-1})\right)\right] \tag{25}$$

$$\overline{R}(\mathrm{k})=\overline{R}(\mathrm{k-1})+\Psi^{'}(\varepsilon(\mathrm{k}))\varphi(k)\varphi^T(k) \tag{26}$$

From the assumption b) it follows $\nabla_\theta J_{k-1}(\hat{\theta}(\mathrm{k-1}))=0$. Based on this condition, and if $\theta=\hat{\theta}(\mathrm{k-1})$ is put in the relation (25), one obtains:

$$k\nabla_\theta J_k(\hat{\theta}(\mathrm{k-1}))=-\Psi(\varepsilon(k))\varphi(k) \tag{27}$$

Finally, based on relations (25) - (27) a recursive algorithm is obtained:

$$\hat{\theta}(\mathrm{k})=\hat{\theta}(\mathrm{k-1})+\overline{R}^{-1}(\mathrm{k})\varphi(k)\Psi(\varepsilon(k)) \tag{28}$$

$$\overline{R}(\mathrm{k})=\overline{R}(\mathrm{k-1})+\Psi^{'}(\varepsilon(\mathrm{k}))\varphi(k)\varphi^T(k) \tag{29}$$

The algorithm (28) - (29) includes the inverse matrix $\overline{R}^{-1}(\mathrm{k})$. To avoid this let us introduce the matrix $P(\mathrm{k})=\overline{R}^{-1}(\mathrm{k})$. Using this notation and applying the matrix inversion lemma [1], from (28) and (29), one can obtain the definitive form of a recursive algorithm for identification of dynamic systems with time-invariant parameters:

$$\hat{\theta}(\mathrm{k})=\hat{\theta}(\mathrm{k-1})+P(\mathrm{k})\varphi(k)\Psi(\varepsilon(k)) \tag{30}$$

$$\mathrm{P(k)}=P(\mathrm{k-1})-\frac{P(\mathrm{k-1})\varphi(k)\varphi^T(k)P(\mathrm{k-1})}{\left[\Psi^{'}(\varepsilon(\mathrm{k}))\right]^{-1}+\varphi^T(k)P(\mathrm{k-1})\varphi(k)} \tag{31}$$

$$\varepsilon(\mathrm{k})=y(\mathrm{k})-\hat{\theta}^T(\mathrm{k-1})\varphi(k) \tag{32}$$

$$\Psi(x)=\min\left\{|x|,k_\varepsilon\right\}\mathrm{sgn}(x) \tag{33}$$

$$\Psi^{'}(x)=\begin{cases}1 & |x|<k_\varepsilon \\ 0 & otherwise\end{cases} \tag{34}$$

The function defined by the relation (33) is the Huber function [6]. It is derived for a class of distributions(3). It is shown on the following figure.

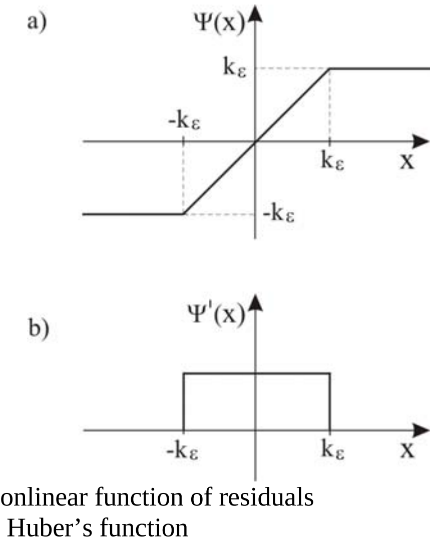

Fig. 1 Nonlinear function of residuals

a) Huber's function

b) Derivative of Huber's function

3. ROBUST AKAIKE'S CRITERION

In a general case, a model of system can be described by an assumed probability density function of measurements. This probability density is put in correspondence with the exact probability density measurements. The consistency between two probability densities describes the Kullback - Leibler information distance. By minimization of the information distance it is obtained the criterion for determining the model order [1]. For the given model order this criterion is identical to the maximum likelihood criterion. If it is assumed that the model (1) has constant parameters and a stochastic noise e(k) has a Gaussian distribution, Akaike's criterion has the form:

$$W_A(k) = \sum_{i=1}^{k} \varepsilon^2(i) + p, \quad p = n + m \tag{35}$$

in which k represents a number of measurements and p is a number of parameters. In this paper, it is necessary to define Akaike’s criterion for a general case:

a) The system parameters are time-invariant
b) The stochastic noise has a non-Gaussian distribution described by the relation (3)

Based on relations (11) and (12) it is obtained:

$$\Phi\left(\varepsilon(\mathrm{k})\right) = \begin{cases} \dfrac{\varepsilon^2(\mathrm{k})}{2\sigma_1^2} + \ln\dfrac{\sqrt{2\pi}\sigma_1}{1-\varepsilon} & \left|\varepsilon(\mathrm{k})\right| \le k_\varepsilon \\ \dfrac{k_\varepsilon}{\sigma_1^2}\left(\left|\varepsilon(\mathrm{k})\right| - \dfrac{k_\varepsilon}{2}\right) + \ln\dfrac{\sqrt{2\pi}\sigma_1}{1-\varepsilon} & \left|\varepsilon(\mathrm{k})\right| > k_\varepsilon \end{cases} \tag{36}$$

Since in the paper estimation algorithm is based on robust statistics [2], the criterion for the selection of the model structure will be called robust Akaike's criterion. Taking into account conditions a) and b) this criterion has the form:

$$W_{RA}(k) = \sum_{i=1}^{k} \Phi(\varepsilon(k)) + p, \quad p = n + m \tag{37}$$

Based on the point of criterion minimum(37), polynomial orders $A(\cdot,\cdot)$ and $B(\cdot,\cdot)$ are determined.

Remark 1: The criterion (37) determine models collection because when p is determined from minimum of the criterion there are multiple combinations of polynomial orders m and n which satisfy the condition. Because, it is adopted:

$$n = m, \quad p = 2n \tag{38}$$

4. SIMULATION RESULTS

The proposed robust Akaike’s criterion has been tested on the following OE model:

$$y(k) = \frac{0.5q^{-1} + 0.3q^{-2}}{1 - 0.7q^{-1} + 0.5q^{-2}} u(k) + e(k) \tag{39}$$

The system identification example, is based on measured 1000 input-output data points obtained during the experiments.

During the simulations, it is assumed that measured noise has non-Gaussian distribution:

$$\mathcal{P}_\varepsilon = \left\{ p(e) = (1-\varepsilon)\cdot\mathcal{N}(0;0.1) + \varepsilon\cdot\mathcal{N}(0;10) \right\}. \tag{40}$$

PRBS signal is used for input signal. Figs. 2 to 4 show noise signal, system input and corresponding system output, respectively.

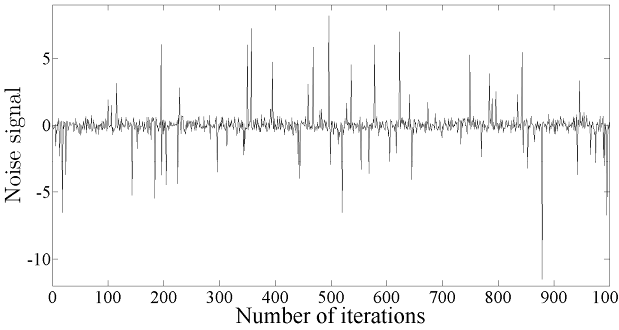

Fig.2 A non-Gaussian noise sequence, $\varepsilon = 0.1$

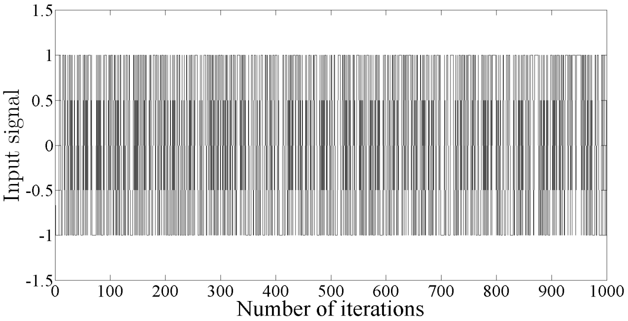

Fig.3. PRBS excitation signal

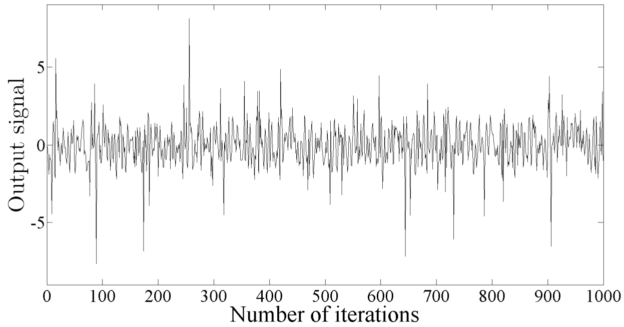

Fig.4. Measured output signal of the system with contamination $\varepsilon = 0.1$

Based on the point of criterion minimum (37), for nine different model orders, it is shown that the observed system can be best described by a second order model, see Fig 5.

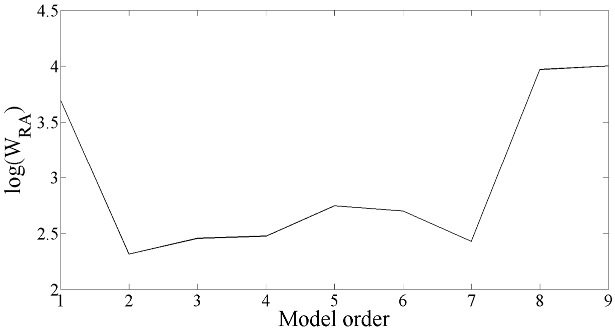

Fig. 5 RAIC criterion for selection of model order

To demonstrate the superiority of the proposed robust OE identification algorithm, a comparison with linear OE identification algorithm [9], when input signal is PRBS signal, is made.

The simulation results are compared in terms of mean square error (MSE), defined by

$$MSE = \ln\left(\mathrm{E}\left\|\hat{\theta}(k) - \theta(k)\right\|^2\right) \quad (41)$$

Figs. 6 to 8 show parameter estimates, and mean square errors.

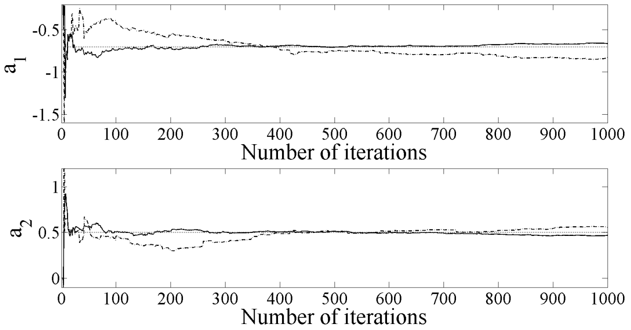

Fig.6. Estimates of parameters a_1 and a_2 obtained in nongaussian noise environment with contamination $\varepsilon = 0.1$ (solid line: Parameter estimates Robust OE, dash-dot: Parameter estimates using linear OE algorithm, dotted line: True parameter values)

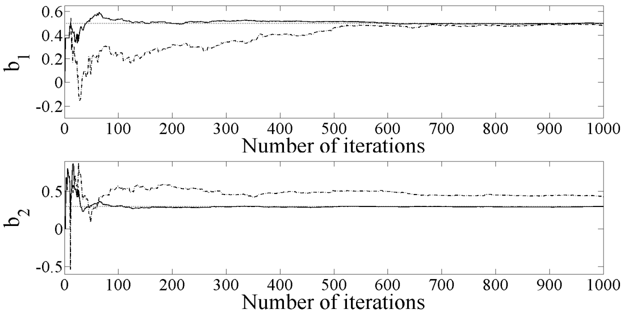

Fig.7. Estimates of parameters b_1 and b_2 obtained in nongaussian noise environment with contamination $\varepsilon = 0.1$ (solid line: Parameter estimates Robust OE, dash-dot: Parameter estimates using linear OE algorithm, dotted line: True parameter values)

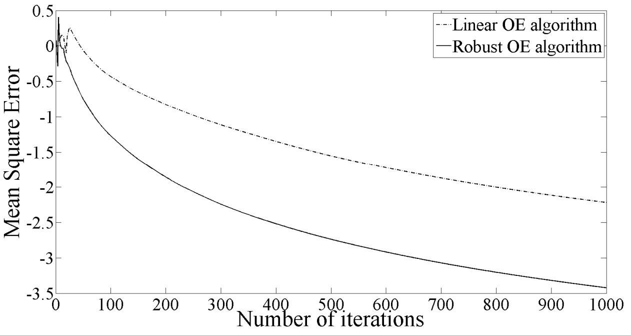

Fig.8. Mean square error, obtained in nongaussian noise environment with contamination $\varepsilon = 0.1$

Figs. 9 and 10 show noise signal and system output respectively, the contamination $\varepsilon = 0.2$.

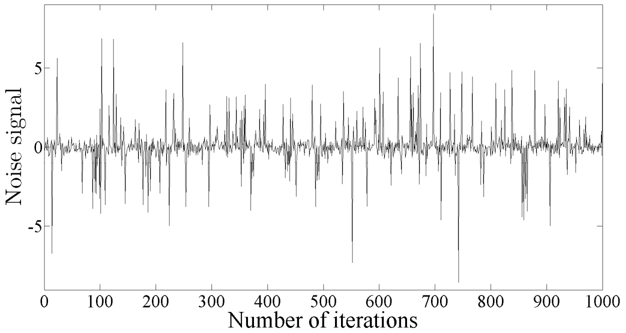

Fig.9 A non-Gaussian noise sequence, $\varepsilon = 0.2$

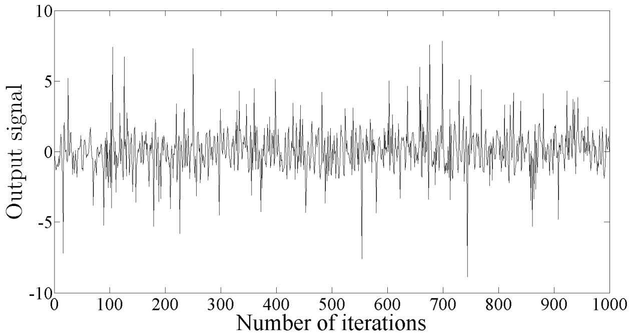

Fig.10. Measured output signal of the system with contamination $\varepsilon = 0.2$

Figs. 11 to 13 show parameter estimates, and mean square errors.

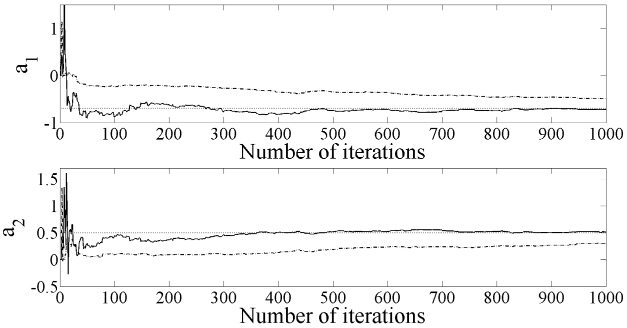

Fig.11. Estimates of parameters a_1 and a_2 obtained in nongaussian noise environment with contamination $\varepsilon = 0.2$ (solid line: Parameter estimates Robust OE, dash-dot: Parameter estimates using linear OE algorithm, dotted line: True parameter values)

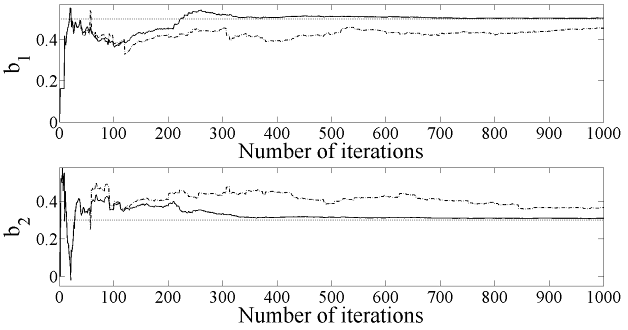

Fig.12. Estimates of parameters b_1 and b_2 obtained in nongaussian noise environment with contamination $\varepsilon = 0.2$ (solid line: Parameter estimates Robust OE, dash-

dot: Parameter estimates using linear OE algorithm, dotted line: True parameter values)

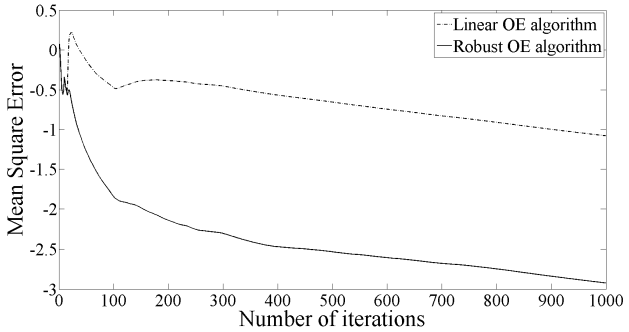

Fig.13. Mean square error, obtained in nongaussian noise environment with contamination $\varepsilon = 0.2$

Comparing Figs. 8 and 13, it can be clearly seen that the superiority of the proposed robust OE algorithm is greater in higher degrees of contamination.

5. CONCLUSION

The basic objective of this paper is to consider how the proposed robust Akaike’s criterion copes with the problem of the robust parameter estimation of the time invariant OE model. It assumed that the output measurement of plant is disturbed by Non-Gaussian noise.

The good behavior of proposed robust Akaike’s criterion as well as robust identification procedure for OE model is illustrated on the simulation example of the second order model.

ACKNOWLEDGEMENTS

The authors would like to express their gratitude to the Serbian Ministry of Education, Science and Technological Development for supporting this paper through projects TR33026 and TR33027.

REFERENCES

[1] L. Ljung, “System Identification: Theory for the user”, 2nd ed., Prentice – Hall, New Jersey (USA) (1999)

[2] T. Söderström, P. Stoica, “System Identification”, Prentice – Hall, London (England) (1989)

[3] S. Kullback, R. A. Leibler, “On information and sufficiency,” The Annals of Mathematics Statistics, Vol. 22(1), pp. 76–86, (1951)

[4] H. Akaike, “A new look at the statistical model identification”, Automatic Control, IEEE Transactions on, Vol. 19(6), pp. 716 – 723, (1974)

[5] V. Barnett, T. Lewis, “Outliers in Statistical Data”, 3rd ed. Wiley-Blackwell, New York (USA) (1994)

[6] P.J. Huber, E.M. Ronchetti, “Robust Statistics”, 2nd ed., Wiley, New Jersey (USA), (2009)

[7] V. Filipovic, B. Kovacevic, “On robust AML identification algorithm”, Automatica, Vol. 30(11), pp. 1775 – 1778, (1994)

[8] V. Filipovic, B. Kovacevic, “On robustified adaptive minimum-variance controller”, International Journal of Control, Vol.65(1), pp.117-129, (1996)

[9] I.D. Landau, R. Lozano, M. M'Saad, “Adaptive Control”, 1st ed. Springer, Berlin (Germany) (1998)