INFORMATICA, 1992, Vol.3, No.1, 88–97

MODEL ORDER DETERMINATION USING ROBUST HYPOTHESIS-TESTING PROCEDURES

Rimantas PUPEIKIS

Institute of Mathematics and Informatics
Lithuanian Academy of Sciences
2600 Vilnius, Akademijos St.4, Lithuania

Abstract. In the previous paper (Pupeikis, 1990) the problem of model order determination in the presence of outliers in observations has been considered by means of introducing robust analogues of the sample covariance and cross-covariance functions instead of the respective classical function meanings used in the determinant ratio test. The aim of the given paper is the development of statistical hypothesis-testing procedures for determination of the model order of dynamic objects, described by linear difference equations. The results of numerical simulations by computer (Table 1) show the efficiency of the proposed statistical procedures for determining the model order by input-output data in the presence of outliers in observations.

Key words: model order, outlier, robustness, robust inference.

1. Statement of the problem. The larger the number of model parameters, subject to estimating, the more complicated the model itselfs and the greater the amount of calculations required for object identification. On the other hand, it is known (Isermann, 1980) that an increase in the model order not always guarantees the improvement of a mathematical description of the object processes. That is why there appears the problem of a significance of separate parameters or their sets under outliers in observations. Robust inference methods, based on the application of statistical decision rules and hypothesis-testing procedures can be used here.

Consider a single input x_k and single output u_k of a linear discrete-time system, described by the linear difference equation

$$[1+A(z^{-1})]u_k = B(z^{-1})x_k + \xi_k, \tag{1}$$

where

$$A(z^{-1}) = a_1 z^{-1} + \ldots + a_n z^{-n},$$
$$B(z^{-1}) = b_1 z^{-1} + \ldots + b_n z^{-n} \tag{2}$$

are polynomials;

$$a^T = (a_1, \ldots, a_n), \quad b^T = (b_1, \ldots, b_n), \quad c^T = (a^T,\ b^T) \tag{3}$$

are object parameters, subject to the estimation; z^{-1} is a backward shift operator; x_k, $u_k = y_k + \xi_k^*$ are input and output sequences of the mentioned object; y_k is a noiseless sequence of the same object;

$$\xi_k = (1-\gamma_k)v_k + \gamma_k \eta_k \tag{4}$$

is the sequence of independent identically distributed variables with an ε – contaminated distribution of the shape

$$p(\xi_k) = (1-\varepsilon)N(0,\sigma_1^2) + \varepsilon N(0,\sigma_2^2), \tag{5}$$

$p(\xi_k)$ is a probability density distribution of the sequence ξ_k; γ_k is a random variable, taking values 0 or 1 in the probabilities $p(\gamma_k = 1) = \varepsilon$, $p(\gamma_k = 0) = 1-\varepsilon$, v_k, η_k are sequences of independent Gaussian variables with zero means, and variances σ_1^2, σ_2^2, respectively;

$$\xi_k^* = [1+A(z^{-1})]^{-1}\xi_k \tag{6}$$

is the sequence of correlated noise, acting the output of the model.

It is assumed that the roots of $A(z^{-1})$ are outside the unit circle of the z^{-1} – plane. The input signal x_k is a persistent excitation of arbitrary order. The true orders of the polynomials $A(z^{-1})$ and $B(z^{-1})$ are unknown beforehand and are subject to the determination by processing of the input-output data.

2. Model order determination in the absence of outliers in observations. Suppose that $\varepsilon = 0$ in equation (5), therefore $p(\xi_k) = N(0, \sigma_1^2)$. In this case to test if a decrease of the loss function $V(n)$ is significant when the order of model (1) is increased from $n = n_1$ to $n = n_2$ the test quantity

$$F_{n_1,n_2} = \frac{V(n_1) - V(n_2)}{V(n_2)} \frac{s - (2n_2 + 1)}{2(n_2 - n_1)} \tag{7}$$

is used with the asymptotic distribution $F_{2(n_2-n_1),s-(2n_2+1)}$ (Åström and Eykhoff, 1971). In equation (7)

$$\begin{aligned} V(n_1) &= e^T(n_1)e(n_1) = (s - n_1)\sigma_e^2, \\ V(n_2) &= \tilde{e}^T(n_2)\tilde{e}(n_2) = (s - n_2)\sigma_{\tilde{e}}^2 \end{aligned} \tag{8}$$

are loss functions and

$$\begin{aligned} e^T(n_1) &= (e_1, \dots, e_s), \\ \tilde{e}^T(n_2) &= (\tilde{e}_1, \dots, \tilde{e}_s) \end{aligned} \tag{9}$$

are the equation errors or residual sequences of the applied parameter estimation method, using the same input-output data (Isermann, 1980); s is sample size; σ_e^2, $\sigma_{\tilde{e}}^2$ are variances of sequences (9), respectively.

This test is based on the statistical independence of $V(n_2)$ and $V(n_1) - V(n_2)$, which for normal residuals have χ^2 distributions, and $s - (2n_2 + 1)$ and $2(n_2 - n_1)$ degrees of freedom, respectively. Thus, we calculate the test quantity (7) using the loss functions $V(n_1)$ and $V(n_2)$ and at the risk level α we choose $n = n_1$, if $F_{n_1,n_2} < F_\alpha$, where F_α can be taken from the tabulated F distribution.

It is known (Soderström, 1977), that some other order test statistics, based on the loss functions, have been worked out, too.

3. Model order determination only under very large outliers in observations. It was assumed earlier that in equation (5) $\varepsilon = 0$. Now let us consider such a case when this assumption is invalid. Then the statistical decision rule, based on the test quantity (7), for small sample size becomes of little use, since it is not robust even for small departures of $p(\xi_k)$ from the normality. In this case it should be noted that non-normality essentially arises from the tails and if the extreme observations are censored, then there is hardly any difference between a normal sample and the non-normal one, and in that situation whatever is good for normal samples it is automatically good for non-normal ones too (Tiku, Tan and Balakrishnan, 1986). That is why in such a case the hypothesis-testing procedure of the form

$$F^*_{n_1,n_2} = \frac{V^*(n_1) - V^*(n_2)}{V^*(n_2)} \, \frac{s-(2n_2+1)}{2(n_2-n_1)} \tag{10}$$

based on the classical F – test and censored samples

$$e_{(1)}, e_{(2)}, \dots, e_{(s-r_2)}; \quad \tilde{e}_{(1)}, \tilde{e}_{(2)}, \dots, \tilde{e}_{(s-r_2^*)} \tag{11}$$

can be used.

Here

$$\begin{aligned} V^*(n_1) &= e_*^T(n_1) e_*(n_1) = (s-n_1)\sigma^2_{e_*}, \\ V^*(n_2) &= \tilde{e}_*^T(n_2) \tilde{e}_*(n_2) = (s-n_2)\sigma^2_{\tilde{e}_*} \end{aligned} \tag{12}$$

are loss functions and

$$e_*^T(n_1) = (e_{(1)}, \dots, e_{(s-r_2)}), \quad \tilde{e}_*^T(n_2) = (\tilde{e}_{(1)}, \dots, \tilde{e}_{(s-r_2^*)}) \tag{13}$$

are the remaining equation errors of size $s - r_2$ and $s - r_2^*$ after censoring the r_2 and r_2^* numbers of very large outliers in the initial samples $e_1, e_2, \dots, e_s$ and $\tilde{e}_1, \tilde{e}_2, \dots, \tilde{e}_s$, respectively. In addition $e_{(1)} \leqslant e_{(2)} \leqslant \dots \leqslant e_{(s-r_2)}$, $\quad \tilde{e}_{(1)} \leqslant \tilde{e}_{(2)} \leqslant \dots \leqslant \tilde{e}_{(s-r_2^*)}$;

$$\sigma_{e_*} = \frac{B + \sqrt{B^2 + 4AC}}{2\sqrt{A(A-1)}}, \quad \sigma_{\tilde{e}_*} = \frac{B_* + \sqrt{B_*^2 + 4A_*C_*}}{2\sqrt{A_*(A_*-1)}}, \tag{14}$$

$$B = r_2\alpha_2(e_{(s-r_2)} - K), \qquad B_* = r_2^*\alpha_2^*(e_{(s-r_2^*)} - K_*), \tag{15}$$

$$K = \frac{\sum\limits_{i=1}^{s-r_2} e_{(i)} + r_2\beta_2 e_{(s-r_2)}}{m}, \qquad K_* = \frac{\sum\limits_{i=1}^{s-r_2^*} \tilde{e}_{(i)} + r_2\beta_2^* \tilde{e}_{(s-r_2^*)}}{m}, \tag{16}$$

where α_2, α_2^* and β_2, β_2^* are tabulated values, which depend on $q = r_2/s$ and $q^* = r_2^*/s$ (Tiku, Tan and Balakrishnan, 1986);

$$m = s - r_2 + r_2\beta_2, \qquad m_* = s - r_2^* + r_2^*\beta_2^*, \tag{17}$$

$$A = s - r_2, \qquad A_* = s - r_2^* \tag{18}$$

$$\begin{aligned} C &= \sum_{i=1}^{s-r_2} e_{(i)}^2 + r_2\beta_2 e_{(s-r_2)}^2 - mK^2, \\ C_* &= \sum_{i=1}^{s-r_2^*} \tilde{e}_{(i)}^2 + r_2^*\beta_2 e_{(s-r_2^*)}^2 - mK_*^2. \end{aligned} \tag{19}$$

Thus we calculate test quantity (10) using the loss functions $V^*(n_1)$ and $V^*(n_2)$ and at the risk level α we choose $n = n_1$ if $F_{n_1,n_2}^* < F_\alpha$. On the other hand large values of F_{n_1,n_2}^* lead us to the rejection of it.

4. Model order determination in the presence of very large and very small outliers in observations. Let us assume now, that in a random sample there appear observations not only with very large but also with very small meanings. In this case the initial samples $e_1, e_2, \ldots, e_s$ and $\tilde{e}_1, \tilde{e}_2, \ldots, \tilde{e}_s$ of size s ought to be arranged in ascending order of their magnitude in such a way

$$e_{(1)} \leqslant e_{(2)} \leqslant \ldots \leqslant e_{(s)}, \qquad \tilde{e}_{(1)} \leqslant \tilde{e}_{(2)} \leqslant \ldots \leqslant \tilde{e}_{(s)} \tag{20}$$

Afterwards the smallest observations r_1 and r_1^* and the largest ones r_2 and r_2^* in random sequences (20) will be rejected. Then the remaining observations of sizes $s - r_1 - r_2$ and $s - r_1^* - r_2^*$ after censoring the mentioned observations are

$$e_{(r_1+1)}, e_{(r_1+2)}, \ldots, e_{(s-r_2)},$$

and

$$\tilde{e}_{(r_1^*+1)}, \tilde{e}_{(r_1^*+2)}, \ldots, \tilde{e}_{(s-r_2^*)}, \tag{21}$$

respectively.

Therefore, equations (15) – (19) will be transformed into the formulas :

$$\begin{aligned} B &= r_2\alpha_2(e_{(s-r_2)} - K) - r_1\alpha_1(e_{(r_1+1)} - K), \\ B_* &= r_2^*\alpha_2^*(\tilde{e}_{(s-r_2^*)} - K_*) - r_1^*\alpha_1^*(e^*_{(r_1^*+1)} - K_*), \end{aligned} \tag{22}$$

$$\begin{aligned} K &= \frac{\sum\limits_{i=r_1+1}^{s-r_2} e_{(i)} + r_1\beta_1 e_{(r_1+1)} + r_2\beta_2 e_{(s-r_2)}}{m}, \\ K_* &= \frac{\sum\limits_{i=r_1^*+1}^{s-r_2^*} \tilde{e}_{(i)} + r_1^*\beta_1^*\tilde{e}_{(r_1^*+1)} + r_2^*\beta_2^* e_{(s-r_2^*)}}{m_*}, \end{aligned} \tag{23}$$

where α_1, α_2, and α_1^*, α_2^*, β_1, β_2 and β_1^*, β_2^* are tabulated values, depending on r_1, r_2, r_1^* and r_2^*;

$$m = s - r_1 - r_2 + r_1\beta_1 + r_2\beta_2, \quad m_* = s - r_1^* - r_2^* + r_1^*\beta_1^* + r_2^*\beta_2^*, \tag{24}$$

$$A = s - r_1 - r_2, \qquad A_* = s - r_1^* - r_2^*, \tag{25}$$

$$\begin{aligned} C &= \sum_{i=r_1+1}^{s-r_2} e_{(i)}^2 + r_1\beta_1 e_{(r_1+1)}^2 + r_2\beta_2 e_{(s-r_2)}^2 - mK^2; \\ C_* &= \sum_{i=r_1^*+1}^{s-r_2^*} \tilde{e}_{(i)}^2 + r_1^*\beta_1^*\tilde{e}_{(r_1^*+1)}^2 + r_2^*\beta_2^*\tilde{e}_{(s-r_2^*)}^2 - mK_*^2, \end{aligned} \tag{26}$$

It should be mentioned that the meanings of r_1, r_1^* and r_2, r_2^* can be chosen according to Tiku, Tan and Balakrishnan, (1986).

5. Simulation results. The efficiency of the F – test statistics was investigated with the help of numerical simulation by IBM PC/AT. The noiseless sequence y_k was generated by the equation

$$y_k = \frac{z^{-1} + 0.5z^{-2}}{1 - 1.5z^{-1} + 0.7z^{-2}} x_k, \qquad k = \overline{1,100} \tag{27}$$

taken from the paper (Åström and Eykhoff 1971).

The realizations of independent Gaussian variables ζ_k with zero mean and unitary dispersion and the sequences of the first and second order AR models of the form

$$x_k = 0.9x_{k-1} + 0.43\zeta_k, \tag{28}$$

$$x_k = x_{k-1} - 0.5x_{k-2} + \zeta_k \tag{29}$$

were used as the input sequence x_k^*. The realization of the discrete AR process was generated as the additive noise according to equation (6), where

$$A(z^{-1}) = -1.5z^{-1} + 0.7z^{-2}, \tag{30}$$

ξ_k is a sequence of independent identically distributed variables of shape (4) with the ε - contaminated distribution (5) and $\sigma_1^2 = 1$, $\sigma_2^2 = 100$. 10 experiments with different realizations of the noise ξ_k^* were carried out at the noise level $\lambda = \sigma_{\xi^*}^2/\sigma_y^2 = (0.5; 1.0)$. In each i-th experiment five different orders models of the form

$$u_k = \hat{b}_1 x_{k-1} + \hat{a}_1 u_{k-1} + \xi_k, \tag{31}$$

$$u_k = \sum_{j=1}^{2} \hat{b}_j x_{k-j} + \sum_{j=1}^{2} \hat{a}_j u_{k-j} + \xi_k, \tag{32}$$

$$u_k = \sum_{j=1}^{3} \hat{b}_j x_{k-j} + \sum_{j=1}^{3} \hat{a}_j u_{k-j} + \xi_k, \tag{33}$$

$$u_k = \sum_{j=1}^{4} \hat{b}_j x_{k-j} + \sum_{j=1}^{4} \hat{a}_j u_{k-j} + \xi_k, \tag{34}$$

$$u_k = \sum_{j=1}^{5} \hat{b}_j x_{k-j} + \sum_{j=1}^{5} \hat{a}_j u_{k-j} + \xi_k, \tag{35}$$

were used, where the estimates of parameters of the abovementioned models were obtained using the least squares formulas:

$$\hat{c}_s = \left\{ \sum_{k=t}^{s=80} \varphi_k \varphi_k^T \right\}^{-1} \sum_{k=t}^{s=80} \varphi_k u_k. \tag{36}$$

Here $\hat{c}_s^T = (\hat{a}^T, \hat{b}^T)_s$

$$\hat{a}^T = (\hat{a}_1, \dots, \hat{a}_n)_s, \qquad \hat{b}^T = (\hat{b}_1, \dots, \hat{b}_n)_s \tag{37}$$

are model parameters estimates, whereas $n = 1$, $t = 2$; $n = 2$, $t = 3$; $n = 3$, $t = 4$; $n = 4$, $t = 5$, $n = 5$, $t = 6$ for equations (31) - (35), respectively;

$$\varphi_k = (-u_k, \dots, -u_{k+1-n}, x_k, \dots, x_{k+1-n})^T \tag{38}$$

is the vector of s observations of input x_k and output u_k.

Then for $n = \overline{1,5}$ the test quantities (10) were calculated using formulas (22) - (26) and $s = 80$, $r_1 = r_2 = r_1^* = r_2^* = 5$, $\alpha_1 = \alpha_2 = \alpha_1^* = \alpha_2^* = 0.25$, $\beta_1 = \beta_2 = \beta_1^* = \beta_2^* = 0.9$, $\varepsilon = 0.5$.

Table 1 illustrates the averaged by 10 experiments variables (10) according to

$$\overline{F}^*_{n_1,n_2} = \frac{1}{10} \sum_{i=1}^{10} F^{*(i)}_{n_1,n_2} \tag{39}$$

for different n_1, n_2 and their confidence intervals Δ, obtained by the formula

$$\Delta = \pm t_\alpha \frac{\hat{\sigma}_{F^*}}{\sqrt{L}} \tag{40}$$

and calculated for different inputs.

Here $\hat{\sigma}_{F^*}$ is the estimate of the variance σ_{F^*}, $\alpha = 0.05$ is the significance level; $t_\alpha = 2.26$ is the $100(1-\alpha)\%$ point of Studentś t distribution with $\nu = L - 1$ degrees of freedom; $L = 10$ is the number of the experiments.

In this conection the first line of each column for different n_1, n_2 corresponds to the Gaussian process, the second and third lines – to the AR processes (28) and (29) at the input, respectively. It follows from the simulation results, presented here, that for different inputs and the same n_1, n_2 the accuracy of the averaged values (39) will be different. On the other hand, the accuracy of the above-mentioned values for the same input depends on the value of λ. It also follows, that the averaged test quantity $\overline{F}^*_{n_1,n_2}$ for different inputs first decreases significantly and then slightly changes. The

Table 1. Averaged values (39) and their confidence intervals (40) depending on n_1 and n_2 for different λ

F_α	$n_1,\ n_2$	$(\overline{F}^*_{n_1,n_2} \pm \Delta)$ for $\lambda = 0.5$	$(\overline{F}^*_{n_1,n_2} \pm \Delta)$ for $\lambda = 1.0$
2.49	1, 2	29.03 ± 2.99 23.33 ± 2.05 21.25 ± 3.00	24.13 ± 2.29 18.33 ± 2.07 17.71 ± 2.51
	2, 3	0.96 ± 0.52 0.48 ± 0.54 1.04 ± 0.54	0.30 ± 0.39 0.48 ± 0.54 0.44 ± 0.56
	3, 4	0.74 ± 0.34 -0.22 ± 0.16 0.29 ± 0.26	1.06 ± 0.32 0.90 ± 0.37 0.20 ± 0.26
2.53	4, 5	1.48 ± 0.27 -0.08 ± 0.37 0.48 ± 0.41	1.28 ± 0.43 -0.49 ± 0.42 0.71 ± 0.41

estimates of the model order are found $n = 2$ because $\overline{F}^*_{2,3}$ and $\overline{F}^*_{3,4}$, $\overline{F}^*_{4,5}$ do not differ significantly.

6. Conclusions. The results of numerical simulation, carried out by computer, prove the efficiency of robust hypothesis-testing procedures (10) calculated on the basis of the classical test quantity (7) with the asymptotic F distribution and equations (22) - (26), used in the presence of outliers in observations.

REFERENCES

Åström, K.J., and P.Eykhoff (1971). System identification - a survey. *Automatica,* 7(2), 123–162.

Isermann, R. (1980). Practical aspects of process identification. *Ibid.,* **16**(5), 575–587.

Pupeikis, R. (1990). Model order robust determination. *Informatica,* 1(2), 96–

109.

Soderström, T. (1977). On model structure testing in system identification. *Int. J. Control,* **26**, 1–18.

Tiku, M.L., W.Y.Tan and N.Balakrishnan (1986). *Robust Inference.* Marcel Dekker Inc., New York and Basel. 321.pp.

Received February 1992

R. Pupeikis received the Degree of Candidate of Technical Sciences from the Kaunas Polytechnic Institute, Kaunas, Lithuania, 1979. He is a senior research worker of the Department of the Technological Process Control at the Institute of Mathematics and Informatics. His research interests include the classical and robust approaches of dynamic system identification as well technological process control.