Robust Identification of AR Models Based on Empirical Risk Minimization

Vojislav Filipović

Abstract: **In this paper the problem of parameters estimation of AR process in the presence of noise uncertainty is considered. Off-line estimation method is based on empirical risk minimization (method important in theory of learning and generalization). It is supposed that properties of stohastic process is not known exatcly. More realistic assumptions is that we have apriori information about the class of distributions to which stochastic process belongs. In such situation philosophy of robust statistics is used. For [1]estimates consistency and asymptotic normality is proved.**

Kewords: **Empirical risk, robust statistics, consistency, asymptotic normality**

I. INTRODUCTION

The representation of a stochastic process by model dates back to an idea that was originated by Yule [1]. Next step was the observation that stationary stochastic process is decomposed into the sum of a deterministic and purely random component [2]. The deterministic component is perfectly predictable from the infinite past and has the form of finite combination of siuusoids. The purely random component can be represented as the output of linear system driven by white noise. Very large class of a stochastic processes can be represented by autoregressive models (AR). That kind of models is related to fundamental theorem in the decomposition of time series [2] and parameters of model can be computed by solving Yule-Walker equations (system of linear equations) [3]. The AR models have applications in: geophisics, speech processing, radars, weather prediction and in many others areas [4], [5]. It is extremely important prediction that theory of stochastic processes and recursive procedure for estimation will play in close future very important role in the development of quantum computers [6].

In this paper we will consider autoregresive parameter estimation under the different assumptions in comparison with the standard approach. Namely, the most commonly used assumption is that the stochastic disturbance has Gaussian model. Beliefs in existence a kind of continuity principle according to which the results of inference would change only a small amount if the actual model deviated only a small amount from the assumed model are unjustified [7]. Estimation algorithms, based on the Gaussian model, have been found to be espicially ineficient when the real distribution belongs to the heavy tailed variety, giving rise to the occasionally very large outliers [8]. According to outliers type which can occur in practice in this paper we will consider disturbance uncertainty in the form of innovation outliers [9].

In this paper we will consider identification of AR models when stochastic process is non-Gaussian. We use the iterative procedure (off-line identification). Tne minimized functional is nonlinear and includes a priori information about the probability distribution of observations. Namely, using game theory it is possible to find least favourable pdf within a prespecified pdf class P to which the real noise pdf belongs [7]. For such pdf asymptotic estimation error covariance matrix has a saddle point. In theoretical investigation of iterative algorithm we extend the results for scalar parameters [10], [11] on the case of AR models. It is proved consistency and asymptotic normality for estimated parameters.

II. PROBLEM FORMULATION

Let the system under consideration be described by a linear single input-single output AR model [12]

$$A\left(q^{-1}\right)y(i)=e(i) \tag{1}$$

where

$$A\left(q^{-1}\right)=1+a_1q^{-1}+...a_nq^{-n} \quad , \quad n\geq 0 \tag{2}$$

characteristic polynomial in backwards shift operator q^{-1} with unknown coeficients

$$\theta^T=[a_1...a_n]$$

but with known orders n. Here $y(i)\in R^1$ and $e(i)\in R^1$ and $e(i)$ is stochastic process. The stochastic process $\{e(i)\}$ has the properties

$$\mathrm{E}\{e(i)\}=0 \quad , \quad \mathrm{E}\left\{e^2(i)\right\}<\infty$$

where $\mathrm{E}(\cdot)$ is mathematical expectation operator.

Autoregressive model (1) can be rewritten in the next form

$$y(i)=Z^T(i)\theta+e(i) \tag{4}$$

where $Z^T(i)=[-y(i-1)...-y(i-n)]$ is vector of measurement. The main goal is parameters estimation based on minimization of next functional

$$J(\theta)=E\left\{H\left(y(i)-Z^T(i)\theta\right)\right\} \tag{5}$$

where $H:R^1\rightarrow R^1$ is a function which depends from distribution of stochastic process $e(i)$. When exact

V. Filipovic is with the RCT, PO Box 126, 15300 Loznica, Serbia and Montenegro, (e-mail: vfilip@eunet,yu)

knowladge about stochastic process is absent functional (5) can be replaced with the empirical functional

$$J_i(\theta)=\frac{1}{i}\sum_{k=1}^{i}H\left(y(k)-Z^T(i)\theta\right) \tag{6}$$

For parameters estimatin we use next procedure

$$\hat{\theta}_i^{(k)}=\arg\min J_i(\theta) \tag{7}$$

This estimate can be interpreted as a solution of next equations

$$\sum_{k=1}^{i}Z^T(k)\psi\left(y(k)-Z^T(k)\theta_i^{(k)}\right) \tag{8}$$

Where $\psi(\cdot)=H'(\cdot)$. Described approach is known as a empirical functional method. As a special cases appear the maximum likelihood method and a least squares method. In this paper we want to find the conditions under which

$$\lim_{i\to\infty}J_i(\theta)=J(\theta) \tag{9}$$

Direct consequence of relation (9) is consistent estimates, i.e.

$$\lim_{i\to\infty}\hat{\theta}_i^{(k)}=\theta_0 \quad \text{, w.p. 1} \tag{10}$$

Remark 1. In the literature one can find different assumptions about the properties of $e(i)$. Process $e(i)$ can be considered as a uniform bounded quantity [6], [13], i.e.

$$|e(i)|\le k \quad , \quad k\in[0,\infty)$$

The general description of $e(i)$ can be presented in the form [14]

$$e(i)=e_1(i)+e_2(i)$$

where $e_1(i)$ is Gaussian process and $e_2(i)$ is uniformly bounded.

In many cases the main assumption is that $e(i)$ has Gaussian distribution. As explained in the introduction of this paper such assumption often is unjustified and more realistic assumption is that we have apriori information about the class of distributions to which the real process $e(i)$ belongs. Such approach will be considered in this paper. Two important classes of distributions are

a) The gross error model

$$F_{1\varepsilon^*}=\left\{P:P=(1-\varepsilon^*)\Phi+\varepsilon^* G,\ G\ \textit{is symmetric}\right\},$$

b) The Kolmogorov model

$$F_{2\varepsilon^*}\left\{P:P\ \textit{is symmetric and}\ \sup|P(x)-\Phi(x)|<\varepsilon^*\right\}$$

In both models Φ is Gaussian distribution and ε^* is known number in (0,1].

Remark 2. Uncertainty in the model (1) can be in the form of unmodeled dynamic which is dominated by

$$|\eta(i)|<\varepsilon\sum_{k=0}^{i=1}a^{i-k}\left(|y(k)+1|\right) \quad , \quad a\in(0,1) \quad ,$$

$\exists\varepsilon>0$

Two models are important

a) AR models with structural uncertainty

$$\left[1+\mu_1 q^{-1}H_1\left(q^{-1}\right)\right]A\left(q^{-1}\right)=e(i)+\xi(y,\mu)$$

where $H_1\left(q^{-1}\right)$ is polynomial with unknown coefficients and orders and μ_1 is constant.

$$|\xi(y,\mu)|\le\mu_1\sum_{k=0}^{j=1}a_1^{i-k}\left(|y(k)|+1\right) \quad , \quad \mu_1\ge 0 \quad ,$$

$a_1\in(0,1)$

b) AR model with slowly varying parameters

$$y(t)+a(1,i-1)y(i-1)+\ldots+a(n,i-1)y(i-n)=e(i)$$

with $\|\theta(i-1)-\theta\|\le\mu_2$, $\mu_2\ge 0$ and

$$\theta^T(i-1)=[a(1,i-1),\ldots a(n,i-1)]$$

In the [16] is considered problem of estimation in the presence of noise uncertainty and unmodeled dynamic.

Remark 3. The identification in H_∞ in the presence of non-Gaussian noise is considered in [17].

III. CONSISTENT ESTIMATION

The very important question is consistency of estimates. For estimates given by relations (6) and (7) we will prove next theorem. The proof of theorem is based on verification of conditions of Theorem 1 from reference [10] where scalar case is considered.

Theorem 1. Let us suppose that

1° $A\left(q^{-1}\right)$ is stable polynomial

2° Random variables $e(i)$ is uniformly bounded, independent, equaly distributed and belongs to apriori known a class of distributions

3° Function $H:R^1\to R^1\cup\{+\infty\}$ is nonnegative semicontinuous and finite except on set Lebesque measure zero and on set $D=\left\{u\in R^1,H(u)<\infty\right\}$

4° Set of parameters $C_\theta\subset R^n$ is compact

5° for $\forall u\in R^1$ is

$$l(v)=\int_{R^1}[H(u+v)-H(u)]d\phi(u)>0$$

and for $\forall\theta\in R_\theta$, different from θ_0 probability measure μ satisfies condition

$$\mu\left\{Z,Z^T(\theta-\theta_0)\ne 0\right\}>0$$

6° A is closed interval, Q is open set and S is countable set

Then

$$P\left\{\lim_{i\to\infty}\hat{\theta}_i^{(k)}=\theta_0\right\}=1 \quad \blacksquare$$

Proof: In the frame of proof we will verifay conditions of Theorem 1 which is proved for scalar case in [10].

A1 Let us introduce

$$H\left(e+Z^T(\theta-\theta_0)\right)=f(y,\theta) \tag{11}$$

and let us define sets

$$\{y: f(y,\theta)\in A\ ,\ \forall\theta\in Q\} \tag{12}$$

$$\{y: f(y,\theta)\in A\ ,\ \forall\theta\in Q\cap S\} \tag{13}$$

Suppose that exists y_1 such that $f(y_1,\theta)\in A$ for $\theta\in Q\cap S$ and $f(y_1,\theta)\notin A$ for $\exists\theta^1\in Q$. One can find sequence $\theta_1,\theta_2,...,\theta_n$ which belongs to set S such that $\lim_{n\to\infty}\theta_n=\theta^1$.

From condition 4° of theorem follows

$$\lim_{n\to\infty}f(y_1,\theta_n)=f\left(y_1,\theta^1\right) \tag{14}$$

According with assumption $f(y_1,\theta_n)\in A$. The A is closed set (condition 6 of Theorem) and because $f\left(y_1,\theta^1\right)\in A$. That is in contradiction with the initial assumption from where follows that sets which are defined with relations (12) and (13) are identical.

A2 Using condition 3° of theorem we have

$$\inf_{\theta'\in Q}H\left(e+Z^T(\theta-\theta')\right)\to F\left(e+Z^T(\theta_0-\theta)\right)\ \text{wp1} \tag{15}$$

if neighbourhood of Q, which contains θ, is degenerated in $\{\theta\}$

A3 From condition 3° of theorem follows

$$E\left\{H\left(e+Z^T(\theta_0-\theta)\right)-H(e)\right\}^-<\infty\ ,\ \forall\theta\in C_\theta \tag{16}$$

$$E\left\{H\left(e+Z^T(\theta_0-\theta)\right)-H(e)\right\}^+<\infty\ ,\ \forall\ \theta\ \in\ C_\theta \tag{17}$$

where $u^+=\max(u,0)$ and $u^-=\min(u,0)$

A4 From condition 5° of theorem we have

$$E\left\{H\left(e+Z^T(\theta_0-\theta)\right)-H(e)\right\}=$$

$$\int\int\left[H\left(e+Z^T(\theta_0-\theta)\right)-H(\theta)\right]dP(\theta)\,d\mu(Z)=$$

$$\int l\left(Z^T(\theta_0-\theta)\right)\mu(dZ)>0 \tag{18}$$

A5 Using assumption 4° of Theorem it is enough to prove condition A5 of Theorem 1 from [10]

$$\inf_{\theta'\in C_\theta}\left[H\left(e+Z^T(\theta_0-\theta)\right)-H(e)\right]\ge h(e,Z)=$$

$$-\sup_{\theta\in C_\theta}\left|H\left(e+Z^T(\theta_0-\theta)\right)-H(e)\right|I_D\cdot$$

$$\cdot\left\{Z,e: e+Z^T(\theta_0-\theta)\in D\ ,\ \forall\theta\in C_\theta\right\} \tag{19}$$

where $I_A(\omega)$ is a set indicator funcrtion

$$I_A(\omega)=\begin{cases}1 & \omega\in A\\ 0 & \omega\notin A\end{cases} \tag{20}$$

Using condition 1° and 2° of theorem Lipschitz condition for $H(u)$ for $u\in D$ is

$$\mathrm{E}\left|h(e,Z)\right|=\int\sup_{\theta\in C_\theta}\left|e+Z^T(\theta_0-\theta)-H(e)\right|\mu(dZ)\le$$

$$\le L\int\sup_{\theta\in C_\theta}\left|e+Z^T(\theta_0-\theta)\right|\mu(dZ)<\infty$$

where $<\in(0,\infty)$ is Lipschitz constant.

Theorem is proved.

Theorem 1 is possible to prove using ergodic theory of stochastic processes [18].

IV. Asymptotic Normality of Estimates

The next important step in the analysis of stochastic algorithms is a speed of convergence. That fact can be established by convergence of estimates in distribution. Results of such kind will be presented in the theorem in this section.

We first will present two definitions.

Definition 1 [19]. For large class of sets $U\subset X$ and $D_1\subset\Theta$ exists contable set S of points $\theta_j\in\Theta$ such that for $\exists A\in U$ and $\exists Q\in D_1$ set of points

$$\{\omega: g(\theta,\omega)\in A\ ,\ \forall\theta\in Q\}$$

differ from set

$$\left\{\omega: g(\theta_j,\omega)\in A\ ,\ \forall\theta_j\in S\cap Q\right\}$$

onliy for subset of fixed set N which has P-measure zero and is independent from sets A and Q. That is separable stochastic process.

Definition 2 [19]. Let us suppose that S is subset of set Θ. Mean squere variation of stochastic process $X(t)$ on set S is supremum of next sums

$$\sum_j\left\{\left\|X\left(t_{j+1}\right)-X\left(t_j\right)^2\right\|\right\}^{\frac{1}{2}}$$

for finite set of points $\{t_j\}$ such that $t_j<t_{j+1}$ and $t_j\in S$ for $\forall_j$.

We also, will formulate next result from [20]

Result [20}. Suppose that S is closed subset on R^1 and $x(t),t\in S$ is separable stochastic process continuous in mean square sense. Also, suppose that set S has inf $t_0\in S$. Then, for $\forall B>0$

$$P\left\{\sup_{t\in S}|x(t)-x(t_0)|>B\right\}\le\frac{24L^2}{B}$$

where L is variation of process on set S.

Let us introduce next

$$q(y(i),\theta)=Z(i)\psi(\nu(i,\theta)) \tag{21}$$

$$\nu(i,\theta)=y(i)-Z^T(i)\theta \tag{22}$$

Theorem 2. Suppose that next assumptions is valid

1° Zeros of polynomial $A(q^{-1})$ are insaide of unit circle

2° $\{e(i)\}$ is independent and equaly distributed stochastic process and $E\{e(i)\}=0$, and $E\{e^2(i)\}=\sigma^2<\infty$

3° Vector function $\lambda(\theta)=E\{q(y(i),\theta)\}$ is continuously differentiable in point θ_0 and matrix $\nabla\lambda(\theta_0)$ is nonsingular

4° Matrix $E\{q(y(i),\theta_0)q^T(y(i),\theta_0)\}$ is nonsingular and finite

5° Function $\psi(\cdot)$ is continuous with respect to θ, i.e. $\lim_{\theta'\to\theta}|\psi(\nu(i,\theta'))-\psi(\nu(i,\theta))|=0$ wp1

6° Exists $(\varepsilon,\delta)>0$ for which

$$P\left\{\frac{1}{\sqrt{i}}\sum_{k=1}^{i}q(y(k),\hat\theta^{(i)})\ge\varepsilon\right\}\le\frac{\delta}{3}$$

7° Estimates of unknown parameters are consistent , i.e

$$\lim_{i\to\infty}\hat\theta^{(i)}=\theta_0$$

Then

$$\sqrt{i}(\hat\theta^{(i)}-\theta_0)\xrightarrow{D}N(0,\Gamma)$$

where $N(0,\Gamma)$ is normal distribution with mean value equal to zero and covariance matrix

$$\Gamma=\frac{E\{\psi^2(\nu(i,\theta))\}}{\{E\psi'(\nu(i,\theta))\}^2}\left[E\{Z(i)Z^T(i)\}\right]^{-1}$$

Proof: Owing the problem with space the proof is omited▪

V. Experimental Results

In this part of paper we will use the Monte-Carlo simulation for experimental analysis of robust off-line estimation. Model of AR process has the form

$$(1-1.5q^{-1}+0.7q^{-1})y(i)=e(i) \tag{23}$$

Stochastic process $e(i)$ has a distribution

$$e(i):(1-\varepsilon)N(0,1)+\varepsilon N(0,100) \tag{24}$$

where $N(0,\Sigma)$ is normal distriburion with zero-mean and varianca Σ. Here we use $\varepsilon=0.01$ where ε is degree of contamination.

As a nonlinear function (nonlinear transformation of prediction error) we use a Huber function

$$\psi_{HU}(x)=\min\{b,\max\{x,-b\}\}=x\min\left(1,\frac{b}{|x|}\right) \tag{25}$$

In this paper we will use $b=3$.

For experiments a population of 100 observation is used.

For parameters estimation we combine two iterative preocedures. On first ten iterations we use gradient algorithm and after that, for increase speed of convergence, we use Gauss-Newton algorithm.

H	a_1	a_2
0	0.000	0.000
5	-0.959	0.176
10	-1.060	0.261
15	-1.110	0.310
20	-1.140	0.337
25	-1.170	0.365
30	-1.180	0.383
35	-1.510	0.713
40	-1.510	0.713

Table 1. Parameters estimation when linear estimation algorithm is used $(\psi(x)=x)$

H	a_1	a_2
0	0.000	0.000
5	-0.790	0.519
10	-1.400	0.364
15	-1.135	0.664
20	-1.510	0.633
25	-1.490	0.727
30	-1.530	0.723
35	-1.500	0.706
40	-1.500	0.706

Table 2. Parameters estimation when nonlinear algorithm is used $\psi(x)=\psi_{HU}(x)$

Experimental results is presented in the next two tabelas. In the Table 1 are results when for non-Gaussian process $e(i)$ is used linear estimation algorithm $(\psi(x)=x)$ and in the Table 2 for the same case used nonlinear (robust) algorithm where $\psi(x)=\psi_{HU}(x)$. For the second case we have better results.

VI. Conclusion

In this paper the problem of off-line parameters estimation of autoregresive model, in the presence of noise uncertainty, is considered. It is supposed that a priori is known only class of distribution to which

stochastic process belongs. In such situation the methodology of robust statistics is used. Consistency and asymptotic normality for estimated parameters is proved.

REFERENCES

[1] G.U. Yule, "On a method of investigating periodicites in distributed series, with special reference to Woferis sunspot numbers", *Phil. Trans. Royal Soc. (London),* vol. A226, pp 267-298, 1927

[2] H. Wold, *A Stady in the Analysis of Stationary Time Series*. Aluqist and Wicksells, Uppsala, Sweden, 1938

[3] G.E.P. Box and G.M. Jenkins, *Time series Analysis: Forecasting and Control*. Prentice-Hall, New Jersey, 1994

[4] R.H. Shumwey and D.S Stoffer, *Time Series analysis and It's Applications*. Springer Verlag, Berlin, 2000

[5] G. Jenkins and D. Watts, *Spectral analysis and It's Applications*.Emerson Adams, New York, 2001.

[6] O.N. Granichin, *Introduction to Methods of Stochastic Optimization and Estimation* (in Russian). S, Peterburg University, S. Peterburg, 2003

[7] P.J. Huber, *Robust Statistics*. John Wiley and Sons, New York, 1981

[8] V.Barnet and T. Lewis, *Outliers in Statistical Data.* John Wiley and Sons, New York,1978

[9] R.D. Martin and V.J.Yohai, "Influence functional for time series", *The Annals of Statistics,* vol. 14, pp. 173-220, 1986

[10] P.J. Huber, "The bahaviour of maximum likehood estimates under nonstandard conditions", *Proc. of the 5th Berkley Symposium on Mathematical Statistics and Probability*, vol. 1, pp. 221-233, 1967

[11] A.B. Tsybakov, "About the minimization method of empirical risk in identification problems", *Automatika i Telemehanika,* No 9, pp. 77-85, 1981

[12] L. Ljung, *System Identification*. Theory for the User. Prentice-Hall, New Jersey, 1999

[13] J. Chen and G. Gu, *Control-Oriented System Identification. An H_∞ Approach*. John Wiley and Sons, New York, 2000

[14] V.Z. Filipovic, "Identification in H_∞ in the presence of deterministic and stochastic noise", *In Proc. of Third Europen Control Conference ECC 95*, Roma, Italy, 1995

[15] H.F. Chen and L. Guo, "Robustness analysis of identification and adaptive control for stochastic systems", *System and Control Letters*, No 9, pp. 131-140, 1987

[16] V.Z. Filipovic, "Robust estimation in the presence of noise uncertainty and unmodeled dynamics", *In Proc. Adaptive Systems in Control and Signal Processing*, Budapest, Hungary, 1995

[17] V.Z. Filipovic, "Identification in H_∞ in the presence of non-Gaussian noise", In Proc. *System Identification*, Santa Barbara, USA, 2000

[18] O.H. Bustos and V.J. Yohai, "Robust estimation for ARMA models" *American Statistical Association*, vol. 81, pp. 155-168, 1986

[19] I.I. Gihman and A. V. Skorohod, *Introduction to Theory of Stochatic Processes*. (in Russian), Nauka, Moskva, 1977

[20] Le Cam, "On the assumptions used to prove asymptotic normality of maximum linelihood estimates", *Annals of Mathematical Statistics*, vol. 41, pp. 802-821, 1970

[21] Y. S. Chou and H. Teicher, *Probability Theory: Independence, Interchangeability, Martingales,* Springer Verlag, Berlin, 1988