

## Selection of the order of autoregressive models for spectral analysis of noise corrupted signals



Jakub Jeřábek \*

Department of Electrical Engineering, Faculty of Electrical Engineering and Informatics, Univerzita Pardubice, Pardubice, Czech Republic

### ARTICLE INFO

#### Article history:

Received 23 February 2017

Received in revised form

23 September 2017

Accepted 24 September 2017

#### Keywords:

ARMA

AR model

Order estimation

System identification

ACF

### ABSTRACT

This paper presents the theoretical basis of autoregressive (AR) modelling in spectral analysis. Autoregressive modelling includes a model identification procedure based on an autocorrelation function (ACF) of the incoming signal and its statistical evaluation. This is necessary to choose the best order of an AR model that best describes the given set of data. Spectral analysis gives information about the frequency content of a signal. The AR method is an alternative to discrete Fourier transform (DFT) in the computing of a power spectrum density function of a signal, but provides a smoother power spectral density than DFT.

© 2017 The Authors. Published by IASE. This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

### 1. Introduction

Autoregressive moving average model (ARMA) describes a behavior of time series. The time series assumes a discrete set of time ordered data obtained from observation of an incoming signal over time, in equally spaced time intervals. The primary goal analysis of a time series is to understand the basic mechanism that generates evaluated data, and based on this, forecast future values of the series.

In telecommunication, radio signals are represented by a stochastic time series. Only a stochastic signal can carry some information. For evaluation of incoming signals we are interested in estimation of the power spectrum density. Normally this is performed by applying the discrete Fourier transform (DFT) directly to the signal, but this is not satisfactory when applied to random signals.

A possible solution is to organize the samples of a signal in segments of equal length, and compute DFT for each segment, and finally average all DFTs. Thus, averaging reduces randomness. Methods based on this approach are called non parametric method. One variety of nonparametric methods is a periodogram and its modifications. For short data lengths, the periodogram technique suffers from poor statistical stability and poor spectral resolution.

The autoregressive model includes a procedure for model identification, which is important for

choosing the order of the autoregressive process that best describes the given finite record of the signal. It provides a smoother power spectrum than DFT, but has the disadvantage of difficult model identification (Rejfeek et al., 2015).

The objective of this paper is to explain the theoretical background of AR methods and estimate the best order of an AR model, based on statistical evaluation of the autocorrelation function (ACF).

### 2. Arma model

Normally we do not have enough information about a basic mechanism that generates evaluated signals for creating a precise model. So we assume that the signal is created as output of the linear time invariant (LTI) system, with transfer function of  $H(z)$  White noise is the input with spectral density  $\sigma^2$ .

The LTI system is described by differential equations:

$$x(k) + a_1x(k-1) + a_2x(k-2) + \dots + a_px(k-p) = b_0v(k) + \dots + b_qv(k-q) \quad (1)$$

where  $\{a_n, b_m\}$  are corresponding AR and MA parameters;  $\{v_n, x_m\}$  Are inputs and output sequence. System transfer function is given by:

$$H(z) = \frac{B(z)}{A(z)} = \frac{\sum_{m=0}^q b_m z^{-m}}{1 + \sum_{n=1}^p a_n z^{-n}} \quad (2)$$

If we substitute  $z = e^{j\omega t}$ , we get frequency response. Spectral density of time series  $x(k)$  is given by:

\* Corresponding Author.

Email Address: [Jakub.Jerabek1@student.upce.cz](mailto:Jakub.Jerabek1@student.upce.cz)<https://doi.org/10.21833/ijaas.2017.012.016>

2313-626X/© 2017 The Authors. Published by IASE.

This is an open access article under the CC BY-NC-ND license

(<http://creativecommons.org/licenses/by-nc-nd/4.0/>)

$$S_p(\omega) = |H(e^{j\omega t})|^2 S_n(\omega) \tag{3}$$

Therefore the signal  $n(k)$  is white Gaussian noise with a variance  $\sigma^2$ .and can be written:

$$S_p(\omega) = |H(e^{j\omega t})|^2 \sigma^2 \tag{4}$$

Each stationary series that is generated by a stochastic process can be assigned to one of the following models.

The first is the Moving average model (MA), which is a linear combination of past error, and its realization is by a FIR filter. The second is the Autoregressive model (AR), which is realized by a IIR filter (Schlindwein and Evans, 1992).

### 3. AR model

The AR model assumes that the current values of the series can be described by finite linear aggregate previous values of the series and current value of white noise. An AR model process of zero mean and order  $p$  can be defined as (Schwarz, 2009):

$$x(k) = n(k) - a_1x(k - 1) - a_2x(k - 2) - \dots - a_px(k - p). \tag{5}$$

We can rewrite equation into form:

$$n(k) = x(k) + \sum_{l=1}^p a_l x(k - l) \tag{6}$$

Representation of this equation is the IIR filter of  $p$ -order, and contains  $p + 1$  parameters which have to be estimated from the data (Fig. 1).

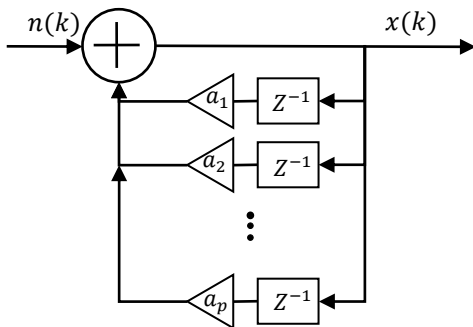


Fig. 1: Realization of AR model using IIR filter

The predictor coefficients are usually estimated using the least-square minimization technique.

A possible solution is through the Yule-Walker equation, which connects an autocorrelation function of time series and its parameters. Deriving of the Y-W equation comes from multiplying Eq. 5 by  $x(k - m)$  and then averaging (Rejfeek et al., 2015).

$$E[x(k)x(k - m)] = -\sum_{l=1}^p a_l E[x(k - l)x(k - m)] + E[n(k)n(k - m)] \tag{7}$$

We suppose that the last value of noise is not correlated with previous values of the series. Therefore is

$$E[n(k)n(k - m)] = 0$$

and we can write:

$$R_{xx}(m) = -a_1R_{xx}(m - 1) - a_2R_{xx}(m - 2) - \dots - a_pR_{xx}(m - p) \tag{8}$$

If we use Eq. 8 to write the expression for  $1 < m < p$ , we get set of linear equation (Porat, 1997).

$$\begin{bmatrix} R_{xx}(1) \\ R_{xx}(2) \\ \vdots \\ R_{xx}(p) \end{bmatrix} = \begin{bmatrix} R_{xx}(0) & R_{xx}(-1) & \dots & R_{xx}(-p+1) \\ R_{xx}(1) & R_{xx}(0) & \dots & R_{xx}(-p+2) \\ \vdots & \vdots & \ddots & \vdots \\ R_{xx}(p-1) & R_{xx}(p-2) & \dots & R_{xx}(0) \end{bmatrix} \begin{bmatrix} -a_1 \\ -a_2 \\ \vdots \\ -a_p \end{bmatrix}$$

The matrix form of the Y-W equation and solution is:

$$a = R^{-1}r \tag{9}$$

and for  $m = 0$ , we get the equation,

$$R_{xx}(0) = -\sum_{l=1}^p a_l R_{xx}(-l) + E[n(k)n(k - m)] \tag{10}$$

therefore  $E[n(k)n(k - m)] \neq 0$  is necessary to provide derivation. The Eq. 5 is multiplied by  $n(k)$  and then averaged.

$$E[n(k)n(k - m)] = -\sum_{l=1}^p a_l E[x(k - l)x(k - m)] + E[n(k)n(k)] \tag{11}$$

where

$$E[n(k)n(k)] = \sigma^2$$

now we can rewrite the Eq. 10 for variance of noise.

$$\sigma^2 = \sum_{l=1}^p a_l R_{xx}(-l). \tag{12}$$

In summary, we get set of  $p$  equations for  $p$  coefficients  $\{a_n\}$  and an equation for variance of noise. Once the parameters  $\{a_n\}$  and  $\sigma^2$ , of the process are known, the power spectrum density of the AR model is given by:

$$S_p(\omega) = \frac{\sigma^2}{|1 + \sum_{l=1}^p a_l e^{-j\omega t}|^2} \tag{13}$$

Identification: The basic assumption is that the time series is stationary. To achieve stationary, it is necessary to provide any of following procedures. The first procedure is to plot data and provide visual inspection (FIG.). Here some seasonality and nonstationary effect trends can be seen. If data is nonstationary, it is necessary to provide transformation. Several ways of transforming the data into a form suitable for modeling, includes filtering, differencing, moving average, and Box-Cox transformation.

As shown in Fig. 2, noise corrupted signal  $s(k)$  and spectrum of the signal (FFT).

Statistical evaluation of series: In Fig. 3, a histogram, probability plot for normal distribution and boxplot are presented. Based on the information from Fig. 3 we can suppose that data comes from Gaussian probability distribution. Parameters are shown in Table 1.

Estimation of the order: The Yule-Walker equation can be used to estimate the parameters of the AR (p) model for given p. The order p can be determined by result of AFC. Before estimation of parameters we must first know the model. Is necessary to choose a preliminary model and afterwards, estimate parameters and do diagnostics to confirm that the model is appropriate. The most telling properties of the time series is the correlation function. Often we don't know the correlation function. The first step is to estimate the correlation function from obtained data.

$$\overline{R_{xx}}(m) = \frac{1}{N} \sum_{i=1}^{N-m-1} x(i)x(i-m), \text{ pro } 1 \leq m \leq p \quad (14)$$

In Fig. 4, ACF is presented. We can see that the correlation function falls within the bounds  $1.96/\sqrt{N}$

for  $k > 15$ . Therefore we may suspect that the data can be modeled by an AR (15) process.

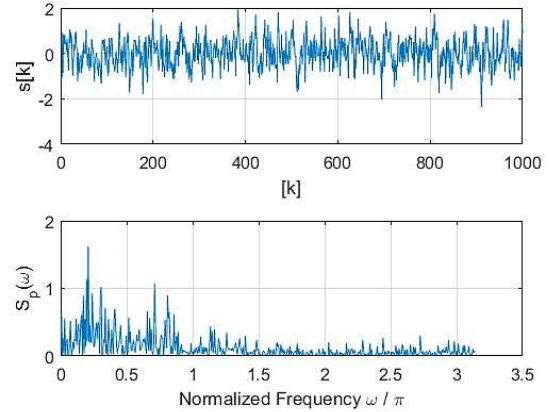


Fig. 2: Noise corrupted signal s(k) and spectrum of the signal (FFT)

**Table 1: Statistical parameters**

Statistical parameters of time series		
Mean	Variance	Standard deviation
0.006	0.4004	0.6327

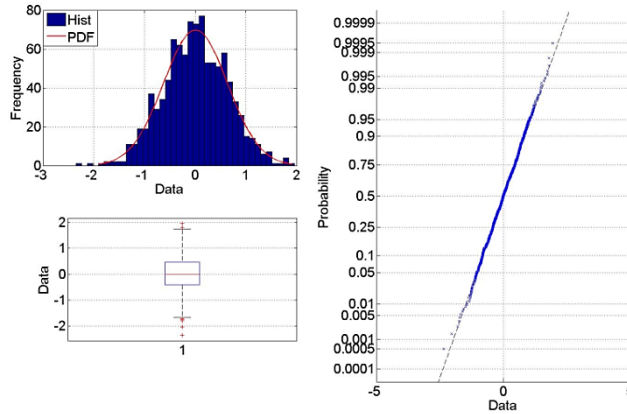


Fig. 3: Statistical evaluation of time series

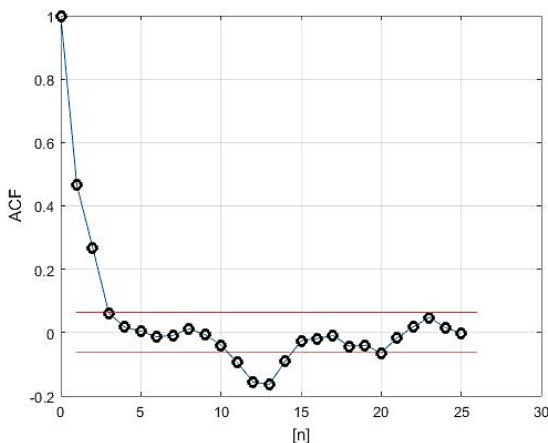


Fig. 4: Autocorrelation function of the time series

Parameters estimation: The next step after predicting of order of the AR model is the necessity to estimate the parameters of the model. This can be solved by using the Yule-Walker equation as is shown in previous section.

Disadvantages of using the Yule-Walker equation is a great number of operations needed to obtain

estimated parameters. There are two basic problems. First is, that mostly we don't know the accurate order of the AR model. So, it is necessary to test the model for different order, and then verify each order by statistical methods. Another problem is that the parameters of the time series can change in time and recalculation of parameters is required. In Fig. 5, the estimated power spectral density of the best order of the AR model is shown. There are three significant frequencies. For higher order of the AR model, non-existent frequencies began to occur and for the lower order, the significant frequencies were not displayed.

Verification: After fitting a model to a given time series, it is necessary to provide a check whether the model fits to the data. There are several ways of checking if the model is satisfactory. The commonly used approach to checking is to examine the residuals. Due to the behavior of residuals, the fitted model is used to test adequacy. The first step in residual analysis is calculating residuals given the fitted model and observed data.

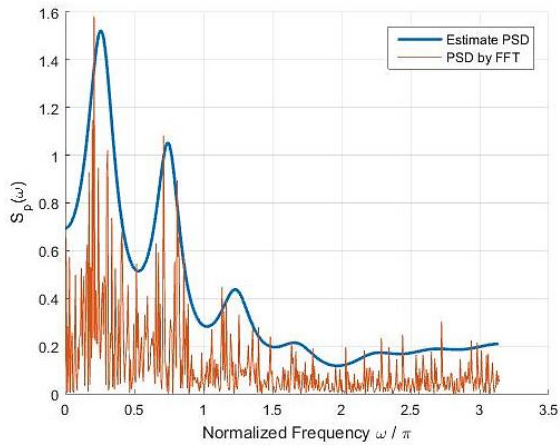


Fig. 5: Estimate AR power spectral density for the best order of AR model

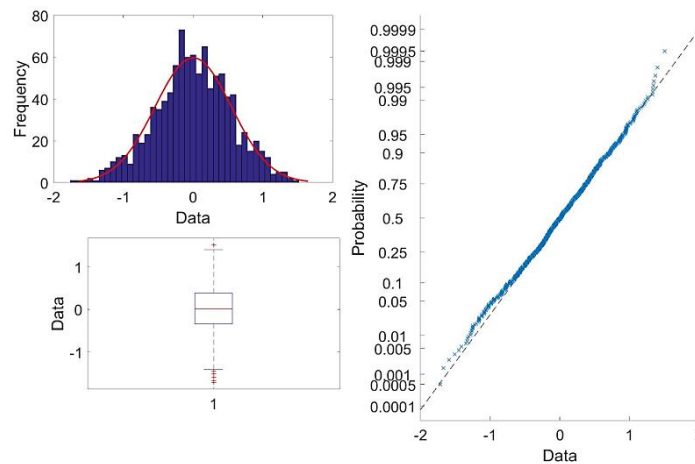


Fig. 6: Statistical evaluation of residuals

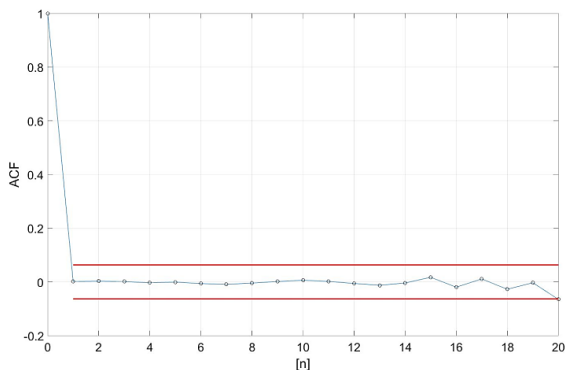


Fig. 7: Autocorrelation function of the residuals

#### 4. Conclusion

In this article the procedure of evaluation of a time series is shown. The first step in evaluation proved the basic assumption, that the time series is stationary. From plotting data we can provide a visual inspection of stationary of the time series. The second step was determining the order of the AR model. The autocorrelation function was the significant instrument for this, which gives important information about attributes of a time series. The third step was estimating the parameters of the selected order of the AR model. Estimation was made through the Yule-Walker equation. After

Residuals are defined as noise calculated from the equation.

$$n(k) = x(k) - \hat{x}(k) = x(k) + \sum_{l=1}^p a_l x(k-l), a_0 = 1 \quad (15)$$

Statistical evaluation of residuals: Statistical evaluation of residuals and autocorrelation function is used to confirm the suitability of the selected model. In Fig. 6, histogram, probability plot for normal distribution, and boxplot are presented. Based on information from Fig. 6 and Fig. 7, we can suppose that residuals come from the Gaussian probability distribution with zero mean and constant variance. In Fig. 7 the ACF function of residuals is presented. The values of the ACF are equal to zero for every  $n \neq 0$ .

estimating parameters, the power spectral density from the AR model can be determined, and is more easily interpreted than the PSD obtained by FFT. Significant frequency and their amplitudes can be easily determined from AR PSD.

#### Acknowledgment

This paper was supported by internal grant University of Pardubice SGS FEI\_2016\_22 and by the grant GAČR 15-24688S.

#### References

- Porat B (1997). Course in digital signal processing. John Wiley and Sons, New York, USA.
- Rejček L, Buresova D, Fiser O, and Brazda V (2015). Comparison of parametric methods for radar signal processing. In the 25<sup>th</sup> International Conference on Radioelektronika (RADIOELEKTRONIKA), IEEE, Pardubice, Czech Republic: 141-144. <https://doi.org/10.1109/RADIOELEK.2015.7128986>
- Schlindwein FS and Evans DH (1992). Autoregressive spectral analysis as an alternative to fast Fourier transform analysis of Doppler ultrasound signals. Diagnostic Vascular Ultrasound, 8: 74-84.
- Schwarz D (2009). Lineární systémy a modely časových řad. Institute of Biostatistics and Analyses, Brno, Czech Republic. Available online at: [www.iba.muni.cz](http://www.iba.muni.cz)