Autoregressive model order determination

Abstract: Here investigation of some approaches for model order identification in the autoregressive model is presented for univariate time series prediction. The approaches are implemented in a software library used for the sake of financial predictions. The results for some real financial series using the considered alternative approaches are summarized and conclusions are presented for their applicability.

Keywords: Autoregressive model; Model order identification; Time series prediction

I. INTRODUCTION

Time series prediction is important from both theoretical and practical point of view and it is applicable in financial, physical, social and many other fields. The prediction approaches can be separated into two main categories - univariate and multivariate. Each of them has their own characteristics and their usage depends to the specific problem that should be solved. The univariate prediction is the simpler one but in the same time it is not less important especially when there is a lack of information for the involved factors influencing the time series behavior into consideration. In this paper the univariate time series prediction is considered emphasizing to the model identification stage. The considered approaches are analyzed from practical point of view and implemented in a software library for time series prediction.

The main steps in time series prediction according to the classical Box-Jenkins approach [1] are as follows.

First of all the series is plotted and its characteristics are analyzed. If there is trend it should be removed by differencing or other pre-processing. Then the model order identification and parameters estimation are performed.

A. Model identification

In this stage the process supposed to represent the time series is identified as autoregressive or moving average or combination of both. In addition their orders are identified. In this paper only the autoregressive model is considered and thus this stage involves determination of the model order that is the number of the model parameters. Let suppose a discrete time series x is given

$$x_1, x_2, \dots, x_n \tag{1}$$

Having an autoregressive model the i-th prediction can be calculated by:

$$\hat{x}_{i+1} = \phi_i x_i + \phi_{i-1} x_{i-1} + \cdots + \phi_{i-k} x_{i-k} + e_{i+1} \tag{2}$$

where k is the model order, e is an error term with zero mean and ϕ are model parameters.

The model order determination is based on identifying the best value of k according to a given criterion. Using (2) an arbitrary number of predictions can be iteratively generated for the historical period and compared to the available values x. The predictions could not be generated for the first k values of the series because there are not input values for the model.

The model order identification is an issue often avoided in the literature and not deeply investigated and described. In the classical Box-Jenkins methodology the analysis of the autocorrelation function (ACF) and partial autocorrelation function (PACF) are used for this purpose [1] [6]. The goal is to find the point where they have not significant values and that point is considered as the model order. Roughly the significance level can be estimated as

$$\frac{n}{\sqrt{2}} \tag{3}$$

where n is the time series size.

If a fixed approach is used for the next stage of parameters estimation the model order is the most important for the prediction quality. That is why in this paper the model order determination is emphasized investigating some new approaches.

B. Parameters estimation

When the model order k is determined the coefficients

$$\phi_1, \phi_2, \dots, \phi_k \tag{4}$$

should be calculated by one of the known approaches. For example Yule-Walker equations [6] or ordinary least squares approach can be used. The aim is to find the best parameters according to the next stage which checks the chosen model to best fit the available data.

C. Model checking

One possible approach for model checking is to separate some part of the time series for parameters estimation and another part for error calculation. When the model order is determined and the parameters are estimated the predictions can be calculated for some future time horizon m:

$$\hat{x}_{n+1}, \hat{x}_{n+2}, \dots, \hat{x}_{n+m} \tag{5}$$

applying (2) and if there are the real values available

$$x_{n+1}, x_{n+2}, \dots, x_{n+m} \tag{6}$$

the residuals can be calculated as

$$d = \sum_{i=1}^{m} (x_i - \hat{x}_i)^2 \tag{7}$$

Trying the model identification and parameters estimation stages for different model orders k a family of residuals dj are obtained and the smallest one determines the best model order.

This work is supported by Eurorisk Systems Ltd..

Thus given that the autoregressive methodology is used in our investigation the steps above can be separated into two main stages: model identification and prediction. The former is based on historical values analysis and the latter uses the determined model from the first stage in order to generate the predictions by propagating the input values through the model

II. IMPLEMENTED APPROACHES

The approaches presented here are implemented as functionalities in a software library and imported in a real software application for the purpose of prediction of financial time series representing commodities, indices, etc. From practical point of view we found that the simpler models are almost always more useful and in the same time they produce practically usable results. The more complex models are based on a huge number of settings parameters most of which should be determined by an expert and often guessed based on the trial and error approach. The practical applications however should be strongly automated. The steps in which the implemented model works are presented below together with three different methods for model order identification which results are analyzed.

A. Data pre-processing and post-processing

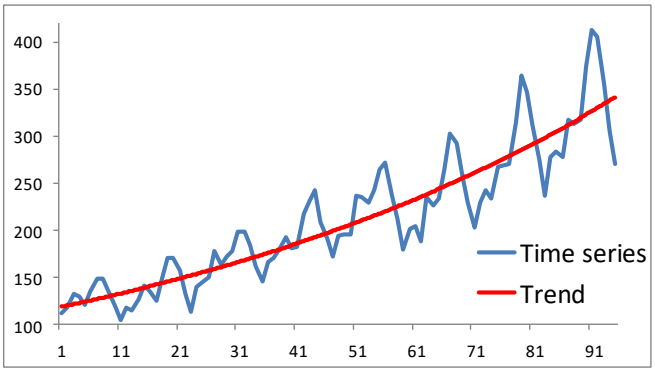
In order to facilitate the prediction making the time series non-depending from its characteristics some pre-processing should be performed before the model building and post-processing after the prediction [11]. Thus the series is transformed to stationary one by removing any additive or multiplicative trend as well as the increasing seasonal effect. In our case first a check is performed are there zero values in the series and if there are such the series is transformed by simple discrete differentiation:

$$y_t = x_t - x_{t-1} \tag{7}$$

Otherwise, if there are not zero values, so called "performances" are calculated:

$$y_t = \frac{x_t - x_{t-1}}{x_{t-1}} \tag{8}$$

Thus the series y is one value shorter than the original series. In fig. 1 the original and pre-processed series are shown with their trends.



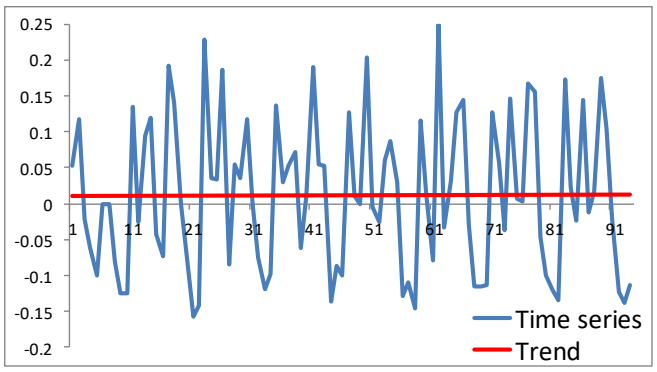


Fig. 1. An original series with its trend shown up and the series after pre-processing down

After the prediction the reverse operation is performed by:

$$x_t = y_t, t = 0 \tag{9}$$

and

$$x_t = y_t + y_{t-1}, t > 0 \tag{10}$$

if differences have been calculated or

$$x_t = (y_t + 1)x_{t-1}, t > 0 \tag{11}$$

if performances have been used.

B. Model identification

Given a time series the autoregressive model is based on the collecting of the model data as shown in fig. 2.
There is a variety of predictive models based on this approach. In addition to the autoregressive based methods like AR, ARMA, ARIMA, SARIMA, ARMAX, SETAR [5] and so on, the neural networks predictive methods are also realized in this way [4] [7] [10] using the sliding window approach thus creating the input-output training vectors.

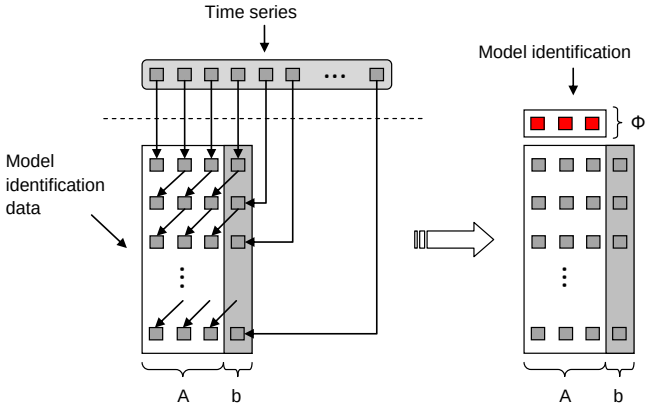


Fig. 2. A time series and building of a simple autoregressive predictive model

The number of the columns of A is determined by the model order identification stage. Given the model order the parameters

Φ are calculated by ordinary least squares method solving the following matrix equation:

$$\Phi = (A^T A)^{-1} A^T b \quad (12)$$

where b, A and Φ are as follows

$$\underbrace{\begin{pmatrix} x_{k+1} \\ x_{k+2} \\ \vdots \\ x_n \end{pmatrix}}_{b} = \underbrace{\begin{pmatrix} x_1 & x_2 & \cdots & x_{k-1} & x_k \\ x_2 & x_3 & \cdots & x_k & x_{k+1} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ x_{n-k} & x_{n-k+1} & \cdots & x_{n-2} & x_{n-1} \end{pmatrix}}_{A} \underbrace{\begin{pmatrix} \phi_1 \\ \phi_2 \\ \vdots \\ \phi_{k-1} \\ \phi_k \end{pmatrix}}_{\Phi} \quad (13)$$

Thus the approach for calculation of the parameters is considered as preliminary chosen and it is not analyzed here. In contrast the model order determination is not chosen to be only one and the following approaches are implemented.

A. Analyzing the autocorrelation partial autocorrelation functions

As in the classical Box-Jenkins procedure these functions are built for the time series and the point, in which they become smaller than a given significance level determines the model order. This approach is traditionally used in the autoregressive prediction.

B. Brute force searching

Using this approach predictions are done subsequently for the historical values by all possible model orders and the error is calculated according to one of the following criteria.

- *Trend brute force searching*

  If there are n historical values $x_1 \ldots x_n$ the linear trend La is built for them and it is extrapolated in the future time horizon m for which predictions will be performed. The future time horizon in our experiments is chosen to be equal to the historical series size n. After that for every possible model order k from min_o to max_o the predictions $x_{n+1} \ldots x_{2n}$ are produced and the linear trend L_b is built for the series $x_1 \ldots x_{2n}$ of the historical and the predicted values. The value k for which the Euclidean distance d_t between the extrapolated L_a and L_b is minimal is chosen to be the model order. For maximal effectiveness in practical solution our investigation shows that min_o and max_o should be chosen to be $\frac{n}{5}$ and $\frac{2n}{5}$ respectively, where n in the series size, because almost all model orders are determined to be in this range.

- *Variation brute force searching.*

  In this approach the variation S_a is calculated for the historical values $x_1 \ldots x_n$. For every possible model order k from min_o to max_o the series is predicted by producing $x_{n+1} \ldots x_{2n}$ and the variation S_b is calculated for $x_{n+1} \ldots x_{2n}$. The value k for which the Euclidean distance $d_v = |S_a - S_f|$ between the historical and predicted values variation is minimal is chosen to be the model order.

- *Trend and variation brute force searching*

  Here the combination between the previous two approaches is used. Both d_t and d_v are calculated together with their sum $d_s = d_t + d_v$ for every model order k from min_o to max_o. The value k for which d_s is minimal is chosen to be the model order. Here different weights can be applied to both d_t and d_v in the calculation of d_s.

The distance in these calculations could also be some other, for example Mahalanobis, other Minkowski measure, R (Pearson correlation) squared or adjusted R squared, etc. [2] [3] [8] [9]

All these approaches are investigated for time series with different characteristics and the results are summarized in table 1 where the average mean squared errors for all series are shown.

TABLE I. MODEL ORDER DETERMINATION APPROACHES

Approach	Average error (%)
Variation	21.25
PACF	24.25
Trend	27.00
Trend and Variation	27.75

The time series are discrete with daily observations of real financial data for 255 business days. The first 205 values are used to apply the presented approaches for model order determination and last 50 values are compared with the predictions generated by the determined model order.

In fig. 3 examples of predictions with confidence levels are shown for series representing aluminum and silver prices in a practical software solution. The series is shown with red line, prediction with green and the upper and lower confidence bounds are shown with yellow and blue color respectively.

The collected real series does not contain all values because they represent different indicators. That is why some of them should be interpolated before the preprocessing stage. In order to preserve their characteristics like mean and variation the interpolation is performed by Brownian bridge [12].

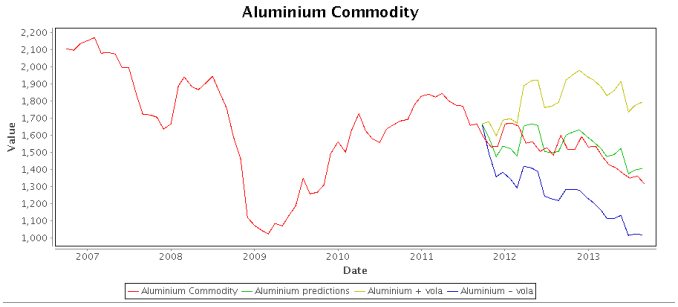


Fig. 3. Example predictions for Aluminium and Silver

III. CONCLUSIONS AND FUTURE WORK

The summary data in table 1 shows that the variation brute force model order searching produces best results. However taking into account that it is a brute force method it works slow executing all modeling stages.

The predictions are shown in red color and their moving average smoothing is shown in blue color. The moving averages together with the confidence lines are the usable data in the practical usage. All predictions shown are by variation brute force searching of the model order. It is obvious that the predicted series also adheres to the main trend. This effect is caused by the pre-processing and post-processing of the time series. The prediction is done to the processed series that preserves its real trend.

According to our investigations unfortunately it is not possible to apply any optimization approach to find the best model order because there is not any dependence of the prediction error by the model order. In fig. 4 – fig.6 examples of prediction of different series types are shown by using the prototype realized in Java.

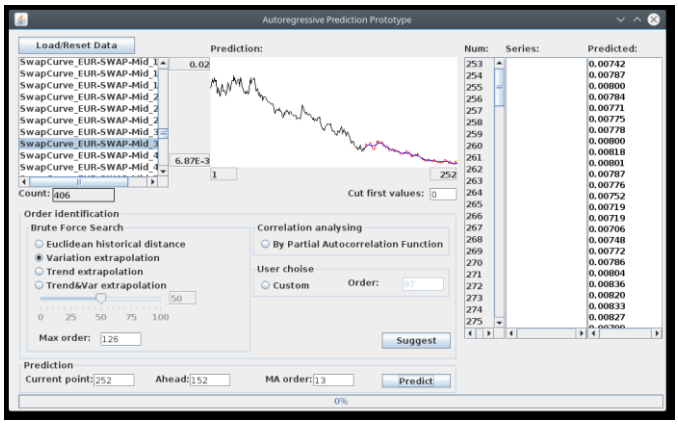


Fig. 4. Prediction of a swap curve

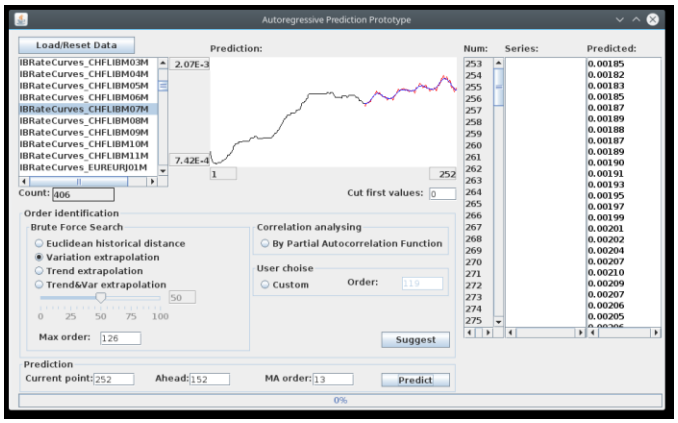


Fig. 5. Prediction of an IB Rate curve

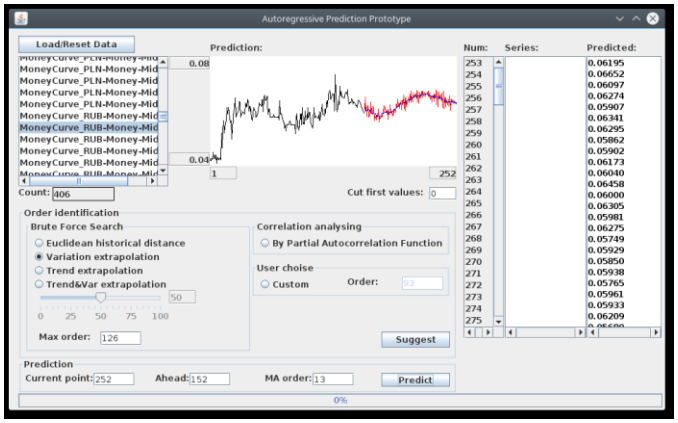


Fig. 6. Prediction of a money market series

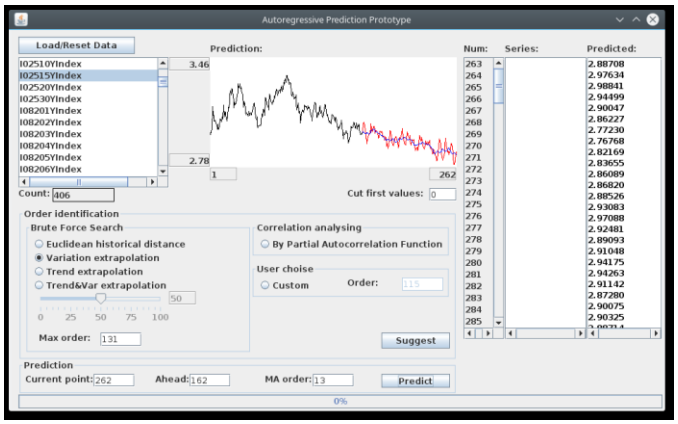


Fig. 7. Prediction of an index series

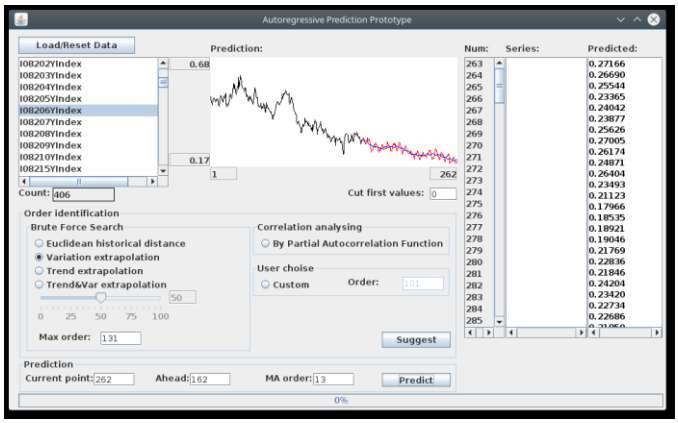


Fig. 8. Prediction of an index

It is also important that the series type is not considered. If the predicted values go below zero correction is not made though this may not matter for the series. For example if a commodity price is predicted to be negative this is not corrected.

REFERENCES

[1] Box G. E. P., G. M. Jenkins. Time Series Analysis: Forecasting and Control, San Francisco, Holden-Day, 1970.

[2] Brocklebank, J. C., D. A. Dickey. SAS for Forecasting Time Series, second edition. SAS Institute Inc., 2003, 398 p.

[3] Chatfield, C. The analysis of time series. An introduction. Fifth edition. Chapman & Hall/CRC, 1996, 304 p

[4] Faraway, J. Time series forecasting with neural networks: a comparative study using the airline data. Appl. Statist., Vol. 47, No. 2, 1998. pp. 231-250.

[5] Fu, Q., H. Fu, Y. Sun. Self-Exciting Threshold Auto-Regressive Model (SETAR) to Forecast the Well Irrigation Rice Water Requirement. Nature and Science, Vol. 2 no. 1, 2004, pp. 36-43.

[6] Hamilton, J. Time Series Analysis. Princeton University Press, ISBN: 0-691-04289-6, 1994.

[7] Huang, W., Y. Nakamori, S. Wang, H. Zhang. Select the Size of Training Set for Financial Forecasting with Neural Networks. Lecture Notes in Computer Science, Vol. 3497/2005, Springer-Verlag, 2005, pp. 879-884.

[8] McNames, J., J. A. K. Suykens, J. Vandewalle. Time Series Prediction. Competition. Internation Journal of Bifurcation and Chaos, Vol. 9, No. 8, 1999, pp. 1485-1500.

[9] Palit, A. K., D. Popovic. Computational intelligence in time series forecasting. Theory and engineering applications. Springer-Verlag, 2005.

[10] Touretzky, D., K. Laskowski. Neural Networls for Time Series Prediction. 15-486/782: Artificial Neural Networks, Lectures, Carnegie Mellon Univeristy, Fall 2006. data. Appl. Statist., Vol. 47, No. 2, 1998. pp. 231-250.

[11] Virili, F., B. Freisleben. Nonstationarity and Data Preprocessing for Neural Network Predictions of an Economic Time Series. International Joint Conference on Neural Networks, Vol.5, 2000, pp.129-134.

[12] http://www.riskmetrics.com: Long Run Technical Document.

ABOUT THE AUTHOR

Ventsislav Nikolov
Eurorisk Systems Ltd.
Varna, Bulgaria
Vnikolov at eurorisksystems dot com