A New Algorithm for Automated Box–Jenkins ARMA Time Series Modeling Using Residual Autocorrelation/Partial Autocorrelation Functions

Article · January 2006

CITATIONS
8

READS
952

2 authors, including:

Qiang Song
Atlanta Forecasting Systems, LLC
23 PUBLICATIONS **3,813** CITATIONS

SEE PROFILE

IEMS Vol. 5, No. 2, pp. 00-00, December 2006.

A New Algorithm for Automated Box-Jenkins ARMA Time Series Modeling Using Residual Autocorrelation/Partial Autocorrelation Functions

Qiang Song
Intelligent Systems and Control Lab
School of Industrial and Systems Engineering
Georgia Institute of Technology
Atlanta, GA 30332
E-mail: qsong3@hotmail.com

Augustine O. Esogbue†
Intelligent Systems and Control Lab
School of Industrial and Systems Engineering
Georgia Institute of Technology
Atlanta, GA 30332
E-mail: aesogbue@isye.gatech.edu

Received Date, November 2005; Accepted Date, June 2005

Abstract. Box-Jenkins time series modeling technique is a powerful tool. Yet, it requires a substantial amount of manual work and statistical skills for the user by inspecting the sample autocorrelation and partial autocorrelation function plots of a time series. This, in our opinion, prevents the technique from further applications in many areas. Therefore, it is highly desirable to automate the Box-Jenkins modeling technique. This paper has two major contributions. First, a new algorithm is proposed where by inspecting the sample autocorrelation function or the partial autocorrelation function of the *model residuals*, which is the key difference from Box-Jenkins', an ARMA(p, q) model can be automatically identified, estimated and diagnosed, without any manual interventions, for a stationary or invertible time series. And second, a new algorithm is proposed to identify the characteristics of correlograms, a fundamental step in the automated modeling processes.

Keywords: Time Series Modeling, Forecasting, Box-Jenkins Technique, ARMA Models, Autocorrelation Function, Identification, Estimation, and Diagnostics.

1. INTRODUCTION

Box-Jenkins time series modeling technique (Box, Jenkins, and Reinsel, 1994) is a powerful tool and has found wide applications in modeling dynamic processes encountered in finance, economics, business, engineering, and many other fields. Particularly, it has been proven to be a powerful tool for forecasting in those fields. However, this powerful technique consists of only three steps. The first step is *identification*. In this step, the sample autocorrelation function (ACF) and the partial autocorrelation function (PACF) are plotted to tentatively identify the most likely candidate for a time series model. The plots of these functions, called correlograms, are then compared with the theoretical autocorrelation functions and partial autocorrelation functions of a standard time series such as moving average MA(q), autoregressive AR(p), or mixed ARMA(p, q) for some small values of p and q. Correlograms of standard MA(q) and AR(p) time series have typical characteristic features that are easy to identify. For example, the theoretical autocorrelation function of an MA(q) time series has a cutoff after q lags, and the partial autocorrelation function of an AR(p) time series has a cutoff after p lags. Thus, any good similarities of the sample correlograms of a time series to either of the correlograms of a stan-

† : Corresponding Author

dard MA(q) or AR(p) time series will recommend a selection of the corresponding model as a candidate for the data. But, it is not so easy to identify an ARMA(p, q) model because the theoretic correlograms of an ARMA (p, q) time series do not have as palpable characteristics as those of an MA(q) or AR(p) times series. The rule of thumb is that if the correlograms do not look like that of either an MA(q) or an AR(p), then an ARMA(p, q) might be a good candidate with p and q chosen properly (Box, Jenkins, and Reinsel, 1994). Once the model is chosen, the next step, called *estimation* step, is to estimate the parameters in the model. Many methods are available to identify the model parameters (Box, Jenkins, and Reinsel, 1994, and Åström and Wittenmark, 1995). The final step is the *diagnostic* step where the model and the parameters are verified against the data, and some metrics and statistics are calculated in this step. A proper model will be used in either forecasting, or other applications, while an improper model may demand some rework by repeating these three steps until a satisfactory model has been identified.

The disadvantage of the Box-Jenkins' time series modeling technique lies in Step 1 where correlograms are inspected visually. This requires the modeler to have special skills and expertise, and demands a significant amount of manual work as well. For this reason, it has become the major obstacle in automating this technique in mass production of time series models and data processing (Talluri and van Ryzin, 2004, p. 450). The goal of this paper is to propose a new algorithm for automating Box-Jenkins' time series modeling technique. Only stationary or invertible time series are studied in this paper.

The remainder of this paper is organized as follows. In section 2, a literature review is presented on the Box-Jenkins modeling technique where various related modeling techniques and methods are discussed. Section 3 is devoted to introducing the Box-Jenkins' modeling technique where the three technical steps are depicted in detail. In section 4, the new algorithm to identify time series models will be developed where empirical illustrations will be used for developing the idea. Numeric examples, seen in section 5, are provided to illustrate the efficacy of the new algorithm. And conclusions and discussions are found in section 6.

2. LITERATURE REVIEW

Automated time series modeling is nothing new in both literature and applications. The simplest method might be to keep a list of candidate models each of which has a pair of p and q values. With the parameters estimated using the data, each model in the list is tested and diagnosed using the data under a given criterion. The best model in the list is recommended for use. The disadvantage of this type of methods is that no recipes are readily available to determine how many candidate models should be kept in the list to have the maximum possibility of obtaining the best model. Variants of such methods can be quite sophisticated. For example, methods such as those of employing information criteria are also used in practice. In these types of methods, a range of ARMA models is estimated by the maximum likelihood method and for each model, the AIC or BIC criterion is evaluated with different values of the decision variables p and q (Wu and Pandit, 1979, and Akaike, 1974). The pair of p and q that yields the minimum criterion value is the preferred model of the data. But, it has been argued that these methods should be viewed as supplementary guidelines to assist in the model identification process, in particular they should not be used as a substitute for careful examination of the characteristics of the estimated autocorrelation and partial autocorrelation functions of a time series, and critical examination of the residuals of the model should always be included as a major aspect of the overall model identification process (Box, Jenkins, and Reinsel, 1994).

A different method can be perceived by assuming that $p=q+1$, *i.e.*, a special structure of ARMA(n, n-1) model can be tested with different n values, and the best model found is used (Wu and Pandit, 1979). But, if $p > q+1$ in the true model then this modeling technique will encounter problems. A variant of this method is to decrease the value of p or q by 1 from some specific values (http://obelia.jde.aca.mmu.ac.uk/resdesgn/arsham/opre330Forecast.htm#rboxjenkm). This is essentially a local search method where the neighborhood is defined around (n+1, n).

The most prominent and successful automated modeling algorithm might be the one implemented in a commercial product introduced in Reilly (1980) and Reilly (1987). This algorithm follows the three steps of Box-Jenkins technique, and automates the entire process. Briefly, this algorithm starts with checking the autocorrelation and partial autocorrelation functions of the time series. Based on the characteristics of the ACF and PACF, the algorithm will determine a p or q value. Then, from the p or q value the algorithm performs a step-down analysis by decreasing the value by 1, and checks if all the parameters in the model are significant. This process continues until all the parameters are significant. After this, the residuals will be examined to see if there is a probabilistic structure still existing. If the ACF of residual process is similar to that of a moving average process, then the model is added an MA part. The algorithm will perform a step-down analysis with respect to the MA part until all its parameters are significant. We refer the interesting readers to Reilly (1980) for details of this algorithm. Compared with all the aforemen-

tioned methods, this method is evidently superior as characteristics, such as the ACF and the PACF, of each individual time series are used as the decision-making criterion. But, it is not clear from the publication how this algorithm handles the situations where the ACF or PACF of the time series does not have a significant characteristic so that a proper p or q is hard or impossible to determine, and it is not clear either in this case how to apply this algorithm. In addition, no details are available on the logic of how the characteristics of the ACF or PACF are extracted automatically by a piece of computer codes.

Time series modeling is not only a science but also an art. It requires subjective judgment and, sometimes, special skills. As a result, different modelers may come up with different models even if the same modeling technique is applied simultaneously by all the different modelers to the same given data set. Technically, there are abundant estimation methods available. Different estimation methods may generate different values of parameters. Therefore, it is not quite uncommon to see that for a time series there might be plethora of proper and competitive models. No matter what methods are applied, and no matter what the modeler's personal preference is, the three steps of the Box-Jenkins modeling technique should be revered and exercised in time series modeling processes. But, this may impose a big burden on the modeling process because the Box-Jenkins technique demands a substantial amount of manual work by visually checking the correlograms of the time series. It is in this step that subjective judgment and preference are incorporated and that it is made extremely difficult to automate the model building process. The difficulty, as will be seen later, can be circumvented by inspecting the correlograms of, instead of the time series itself, *the model residuals*. The rationale is that for an ARMA(p, q) time series, if somehow we can guess the AR(p) part well and apply this model to the data, then the residuals of the model should be an MA(q) time series. Thus, the autocorrelation function of the residuals will have a cutoff after q lags. Similarly, if we can guess the MA(q) part well and apply this model to the data, then the residuals of the model should be an AR(p) time series, and its partial autocorrelation function will have a cutoff after p lags. Well-established statistics exist to study the properties of the residuals (Box, Jenkins, and Reinsel, 1994). Thus, it is possible to design algorithms that utilize the correlograms of the model residuals to automatically build an ARMA(p, q) time series model. This is the goal of this paper with two major contributions: One is to develop an algorithm to identify p or q from the autocorrelation plots, and the other to automate the modeling process by means of inspecting residual correlograms.

3. BOX-JENKINS TIME SERIES MODELING TECHNIQUE

In this section, we briefly review the Box-Jenkins time series modeling technique.

Let $\{x_n\}$ be a time series where $n = 0, 1, 2, \ldots$. Then, an autoregressive model of the time series is described by the following difference equation:

$$x_t = a_1 x_{t-1} + a_2 x_{t-2} + \cdots + a_p x_{t-p} + \xi_t \tag{1}$$

Where p is the order of the model, and ξ_t is an *i.i.d.* noise process. In this model, as is quite evident, the current value of the time series is expressed as a finite linear summation of previous values of the time series and the current value of the noise. Such a model is called an autoregressive (AR) process of order p, denoted as AR(p).

A moving average model is described by the following difference equation:

$$x_t = \xi_t + b_1 \xi_{t-1} + b_2 \xi_{t-2} + \cdots + b_q \xi_{t-q} \tag{2}$$

where q is the order of the model, and ξ_t is an *i.i.d.* noise process. Such a model is called a moving average (MA) process of order q and denoted as MA(q). Evidently, in such a model the current value of the time series is expressed as a linear summation of a finite number of the current and previous values of a noise process.

A mixed model of both (1) and (2) is given by the following difference equation:

$$\begin{aligned} x_t = {} & a_1 x_{t-1} + a_2 x_{t-2} + \cdots + a_p x_{t-p} \\ & + \xi_t + b_1 \xi_{t-1} + b_2 \xi_{t-2} + \cdots + b_q \xi_{t-q} \end{aligned} \tag{3}$$

where p and q are the orders of the model, and such a model is called autoregressive-moving average(ARMA) process and denoted as ARMA(p, q).

The Box-Jenkins time series modeling technique consists of three steps when applied to stationary time series. In the first step, p and q are identified by visually inspecting the plots of the autocorrelation and/or partial autocorrelation functions of the time series. By studying the correlograms of theoretic time series, it has been found that an MA(q) time series has an autocorrelation function that has a cutoff after q lags and the partial autocorrelation function of an AR(p) time series has a cutoff after p lags. Such visual characteristics of the autocorrelation function of an MA(q) time series and the partial autocorrelation function of an AR(p) are extremely helpful in identifying p and q values. But, unfortunately, in general the correlograms of mixed autoregressive-moving average time series do not possess characteristics as obvious as those of MA(q) and AR(p)

time series. For example, a theoretic ARMA(p, q) series may have an autocorrelation function which has an infinite number of damped exponentials and/or damped sine waves after the first q–p lags while the partial autocorrelation function has an infinite number of damped exponentials and/or sine waves after the first p–q lags. Such plots are even hard for experienced modelers to use to glean good information about the values of p and q. For this reason, it has been difficult to automate the Box-Jenkins modeling technique. But, once p and q are known, then in the second step parameters in the model will be estimated. This can be done by using either the least square method, or the maximum likelihood method. Whether or not a model and its parameters are properly chosen will determine the properties of the model residuals. If the residuals behave like a white noise, then the model and parameters are regarded as proper else this modeling process has to be repeated until a satisfactory model and satisfactory parameters are found.

4. MAIN RESULTS

In this section, we will introduce the new algorithm that utilizes only the sample autocorrelation function or partial autocorrelation function of the residual process to identify both p and q. Unlike the classical Box-Jenkins modeling technique where the correlograms of the time series itself are used to identify p and q, and the correlograms of the residuals are used only for the purpose of checking the validity of the model, in the new algorithm *the correlograms of the residuals* are used to identify both p and q, and to check the validity of the model as well. As will be shown later, either the autocorrelation function or the partial autocorrelation function of the residuals alone enables us to identify both p and q simultaneously.

4.1 Overview of the New Algorithm

To proceed, let us rewrite ARMA(p, q) model (3) in a different form as follows:

$$x_t = a_1 x_{t-1} + a_2 x_{t-2} + \cdots + a_p x_{t-p} + \gamma_t \tag{4}$$

where

$$\gamma_t = \xi_t + b_1 \xi_{t-1} + b_2 \xi_{t-2} + \cdots + b_q \xi_{t-q} \tag{5}$$

and $\{\gamma_t\}$ can be regarded as an MA(q) time series. Obviously, (3) is equivalent to (4) and (5). That is, we purposefully decompose an ARMA(p, q) time series into two parts: one is an autoregressive process and the other a moving average process. γ_t can be seen as the model error in (4). But, it is not required here that γ_t be uncorrelated, as implied by (5). Instead, as shown later, γ_t could be correlated and it is the dependence of γ_t that we can draw information from about q based on a chosen value of p.

Our new algorithm can be detailed as follows. Suppose that both (4) and (5) are time series. We pick a value for p of the autoregressive part. Applying an estimation algorithm and the data, we obtain a set of model parameters of $a_1, a_2, \cdots, a_p$. Then, from this AR(p) model we obtain a model residual time series $\{\gamma_t\}$. If this time series is a white noise, then its autocorrelation function has a value of virtually zero for any non-zero lags, and this characteristic is very easy to identify. In this case, we have identified p correctly and we know that the model is an AR(p). If $\{\gamma_t\}$ is correlated, and if its autocorrelation function has a cutoff after q lags, then $\{\gamma_t\}$ is an MA(q) time series and we need to add an MA(q) part to the model. Else, if its autocorrelation function has tailoff, it means that $\{\gamma_t\}$ is either an autoregressive or a mixed time series. In either case, it suggests that in (4) the p value was not chosen properly. If so, we simply increase p by 1, and repeat the above process.

To help illustrate this point, we plot the sample autocorrelation function and partial autocorrelation function of a time series $\{x_t\}$, generated by the following ARMA(2, 4) model $x_t - x_{t-1} + 0.84x_{t-2} = \xi_t + 0.8\xi_{t-1} + 0.57\xi_{t-2} + 0.1\xi_{t-3}$ where $\{\xi_t\}$ is an *i.i.d* N(0, 1) random variable. The correlograms are shown in Figure 1 and Figure 2 respectively. The standard deviation σ of the autocorrelation values and partial autocorrelation values after lag > 2 is about 0.1 if using an approximation (Box, Jenkins, and Reinsel, 1994). From Figure 1 no clear information can be obtained about the MA part of the model as no cutoff can be observed if the 2σ rule is applied. Although Figure 2 indicates that an AR(2) or AR(4) model might be a candidate, as we know it, the true model is an ARMA and evidently the partial autocorrelation function provides information that is not good enough. But, what can we obtain from the autocorrelation function of the residuals of the model if an AR(2) model is assumed? For easy illustration purpose, we calculate the residuals as $\gamma_t = x_t - \hat{x}_t$ where $\hat{x}_t = x_{t-1} - 0.84x_{t-2}$, i.e., we assume the parameters of the AR part are correctly estimated. The corresponding correlograms are shown in Figure 3. Evidently the ACF has a cutoff after 2 lags, which indicates that if an AR(2) model is assumed for the data, an MA(2) component is missing. This tells us at least an AR(2, 2) model should be considered as a starting point. This should be regarded as a significant improvement compared to using the correlograms of the observed data. The same should be obtained if an MA model is assumed for the data and the residual correlograms will indicate an addition of an AR component to the model.

As we inspect the autocorrelation function of $\{\gamma_t\}$ during the identification process, we expect to see only

three possible scenarios: One is that $\{\gamma_t\}$ is a white noise, another is that the autocorrelation function of $\{\gamma_t\}$ has a cutoff after lag q, and the other is that no significant characteristics can be found. In the first scenario, $q = 0$ and we are done, as this time series can be modeled as an AR(p) time series. In the second scenario we have identified both p and q and found that an ARMA(p, q) might be a proper model . But to verify this model, we need to check the model residuals ξ_t, using the estimated parameters and the data. If the resulting ξ_t is a white noise, then the model is proper. Else, if the autocorrelation function of ξ_t has a cutoff after $w \geq 1$ lag, then this means q is still not chosen properly. Hence in this case we need to repeat the above identification process by changing p to p+1. If the autocorrelation function of ξ_t has tail-off, this means that $\{\gamma_t\}$ still contains some autoregressive information of x_t, and we should increase p by 1 and repeat the above process. In summary, the only criterion for a proper model in our algorithm is a white noise residual process.

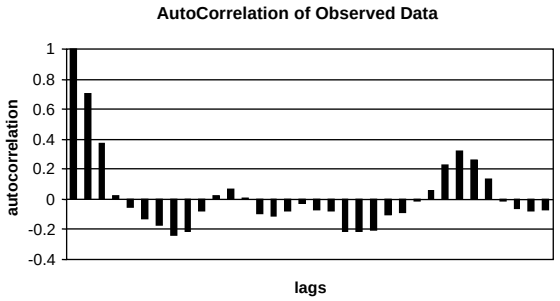

Figure 1. Autocorrelation Function of Observed Data.

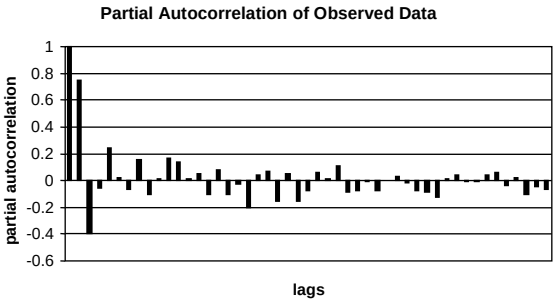

Figure 2. Partial Autocorrelation Function of Observed Data.

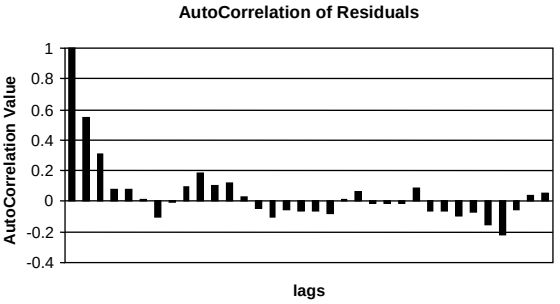

Figure 3. Autocorrelation Function of the Residuals.

4.2 Algorithm for Identifying Characteristics of Correlograms

To automate the modeling process, we need to develop an algorithm to identify whether or not an autocorrelation function of the model residuals is of a white noise, or an MA(q) time series, or neither of the two. This is a fundamental step in the automated time series modeling processes.

To achieve this goal, we will utilize a statistical hypothesis test. As is known, the autocorrelation of a white noise is zero for all non-zero lags. For the estimated autocorrelation function of a white noise, the autocorrelation coefficients can be regarded as a random variable whose variance can be approximated with the following formula (Box, Jenkins, and Reinsel, 1994),

$$\sigma = \frac{1}{\sqrt{N}} \tag{6}$$

where N is the sample size of the data used to calculate the autocorrelation function. Thus, for a given significance level α, if the percentage of autocorrelation coefficients of nonzero lags that are outside the confidence interval ($-k\sigma$, $k\sigma$) is less than 1-α where k is chosen properly, then there is a good reason to believe that the residual process is a white noise. Else, we need to test further whether the residuals are a moving average or an autoregressive time series.

Suppose the residuals are a moving average time series and can be modeled by an MA(q). We know that after q lags the autocorrelation functions are zeros for a theoretical moving average time series. But this may not be true for real time series, and especially this may not be true if we use estimated autocorrelation coefficients from actual data. Similar to a white noise, the variance of the estimated correlation coefficients after q lags can be approximated by (6). Thus, if for a given significance level α, the percentage of correlation coefficients is less than 1-α that are outside the confidence interval ($-k\sigma$, $k\sigma$) after q lags where k is chosen properly, then there is a strong reason to believe that the residuals are a moving average time series with order q. The same idea can be applied to test if the residuals are an autoregressive time series, the difference being that instead of autocorrelation function being used, the partial autocorrelation function of the model residuals is used in the testing process.

Now comes the problem of how to determine q value. For a theoretical moving average time series, it is known that after q lags, its autocorrelation function will become zero. But, for a practical moving average time series for which a proper model is yet to be identified, we do not know its order q in advance, and therefore when we count the number of autocorrelation coeffi-

cients that are outside the confidence interval ($-k\sigma$, $k\sigma$), we do not know where to start to count as the basis, and hence it is seemingly impossible to apply a statistical test approach. To circumvent this difficulty, we propose the following heuristic which should be easy to perceive. Intuitively, dependence between values of a stationary moving average time series should be in general a decreasing function of lags. Thus, we expect that the autocorrelation function is a decreasing function of lags as well. If this is acceptable, then to find where to start to count the autocorrelation coefficients outside the confidence interval, we can use the following algorithm where ρ_l is the autocorrelation coefficient at lag l. Let the input to the algorithm be the correlograms of the residual process.

Algorithm 1

Step 1. Define significance level α, and estimate σ using (6). Set $l = 0$.

Step 2. If ρ_l is outside the confidence interval ($-k\sigma$, $k\sigma$), set $l = l+1$, and repeat Step 2. Else, go to Step 3.

Step 3. Calculate the percentage ϕ of autocorrelation coefficients from lag l to the maximum lag. If ϕ is less than or equal to $1-\alpha$, then let $q = l$ and stop. Else, q is undetermined and stop.

Comments are in order on this algorithm. If $q = 0$ from the above algorithm, then the residual can be said to be a white noise. The same algorithm can be applied using partial autocorrelation functions to determine the p value, and if $p = 0$, then the residual is a white noise. Note that this algorithm implements a statistical hypothesis test and proposes a method to estimate q so that the number of autocorrelation coefficients outside the confidence interval can be counted. How to determine the value k according to the significance level α depends upon the assumption of the error distribution. In this paper the N(0, 1) normal distribution is assumed for the model errors or residuals and k is chosen to be 2. If the input to this algorithm is the correlograms of the time series itself, then this algorithm can identify p or q. Needless to say this algorithm is easy to implement as a fundamental step for automated time series modeling processes. Below, we present the main algorithm of this paper.

4.3 Main Algorithm

The main algorithm, of three steps, can be stated as follows.

Step 1. Use Algorithm 1 to test if the time series is AR or MA. If neither, let $p = 0$, and go to Step 2.

Step 2. Let $p = p+1$. Use the least square method to estimate the parameters of AR(p), and calculate model residuals. Use Algorithm 1 to test if the residuals are an MA time series. If yes, then go to Step 3 with q identified. Else, repeat Step 2.

Step 3. Use the least square method to estimate parameters for the tentative model ARMA(p, q), and calculate model residuals. Then use Algorithm 1 to test if the residuals are a white noise process. If yes, stop. Else, go to Step 2.

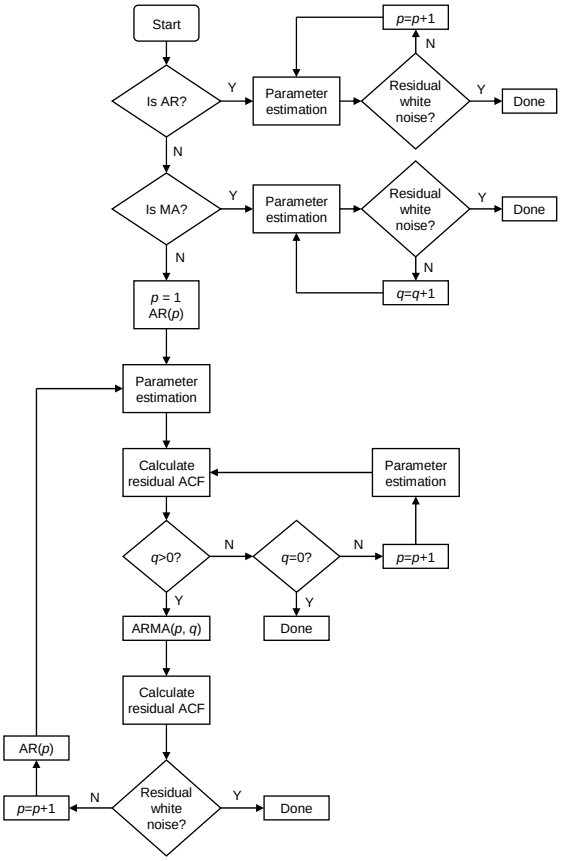

Figure 4. Flow Chart of Main Algorithm for Automated ARMA(p, q) Modeling.

Comments are in order on the main algorithm. First, in the main algorithm, either the least square algorithm or the maximum likelihood algorithm can be used to estimate the parameters of the model. The maximum likelihood algorithm can also provide some information regarding the standard error of the estimate of the parameter. It may not be so easy to get the same information if the least square algorithm is used to estimate the parameter. But, the least square algorithm is simpler and easier to implement. In our implementation of the algorithm, the least square method is used to estimate the model parameters with a gradient-like algorithm for parameter update. The estimation problem is easy to solve for AR or MA models, but is hard for ARMA

models as this involves a highly nonlinear parameter optimization problem. With proper tuning of the optimization algorithm, no convergence problems have been observed so far in the parameter estimation process. Second, in Step 2 the initial model can be an MA(q), and at the beginning of this step q is increased by one. Accordingly in Algorithm 1, then, partial autocorrelation functions should be used. A flowchart of the main algorithm is given in Figure 4 where ACF is used in inspecting the properties of the residuals. Evidently, this flowchart can be modified to use PACF instead without almost any difficulty. And third, in actual implementations of this algorithm, an upper bound on either p or q should be imposed. But how to determine this bound may need human experience and experimentation.

4.4 Convergence of The Main Algorithm

We conjecture that the main algorithm converges if the time series is either invertible or stationary, or both. Although it is beyond the scope of this paper to provide a vigorous proof of the convergence of the algorithm, it suffices to provide a brief argument about the convergence of the algorithm in this section.

The convergence seems to be guaranteed because an invertible ARMA time series can be represented by an AR model of an infinite number of terms and a stationary ARMA time series can be represented by an MA model of an infinite number of terms with arbitrary accuracy (Box, Jenkins, and Reinsel, 1994; Graupe, 1975). In general, if the autocorrelation function or partial autocorrelation function of the observed time series has significant characteristics that the algorithm can recognize, then the algorithm stops with a proper p or q identified. Otherwise, the algorithm will start with an initial p or q, depending on which function is used in the algorithm, ACF or PACF. First, let us assume that the time series is invertible and ACF is used so that an AR model is assumed at the beginning. If the residuals of the model for some p values have a strong characteristic of a moving average process and the algorithm captures this, it stops with a q value identified. If no such a characteristic can be captured, then p will be increased and an iteration process will initiate. As is known, the algorithm increases p by one in each iteration, and this process will continue until when p is large enough so that the AR model can approximate the ARMA time series to an arbitrary precision and the residuals will behave like a white noise. At this time the algorithm will capture this characteristic and terminate. Second, let us consider the case where the time series is stationary and PACF is used so that an MA model is assumed at the start. If the residuals of the model for some q values have strong characteristics of an autoregressive process and the algorithm captures this, it will stop and a proper p value will be detected. Otherwise, if the residuals do not have any strong characteristics that can be captured by the algorithm, the algorithm will increase q by one and a new iteration process will start. As q increases continually, the MA model will approximate the ARMA model as arbitrarily as possible. Thus, when q is large enough the residuals of the model will behave like a white noise and this will be captured by the algorithm. Hence, the algorithm terminates. Finally, if the time series is both stationary and invertible, then the above arguments apply.

5. NUMERICAL EXAMPLES

In this section, we provide numerical results of the algorithm run using some of the time series data available in the literature. What we do is to feed the data as input to the new algorithm and we present the output, i.e., the model parameters and (p, q) values and compare them with the ones reported in the literature.

Example 1. Models for the Sunspots Data

This example has appeared in both Box, Jenkins, and Reinsel (1994), and Brockwell and Davis (1996) as an example of manual modeling of the time series. Table 1 lists the model parameters in the literature and the model obtained using our new algorithm.

As we have about 100 data points, the standard deviation of the autocorrelation coefficients is about 0.1 using Eqn. (6). Thus, using the confidence interval [-0.2, 0.2] with significance level being 0.95 it can be seen from Figure 5 that only one autocorrelation coefficient is outside the confidence interval and the percentage of this is about 1/30=0.033. Thus, we have a strong reason to believe that the residual of the model from the new algorithm is white noise. It appears that the new model is very close to the one obtained in Brockwell and Davis (1996) but quite different from those in Box, Jenkins,

Table 1. Comparison of Models Obtained from Three Sources.

Box, Jenkins, and Reinsel (1994), p.255	Brockwell and Davis(1996), p.97	New Algorithm
$x_t - 1.42x_{t-1} + 0.73x_{t-2} = 14.3 + \xi_t$ $x_t - 1.57x_{t-1} + 1.02x_{t-2} - 0.21x_{t-3} = 11.31 + \xi_t$	$x_t - 1.318x_{t-1} + 0.634x_{t-2} = \xi_t$	$x_t - 1.497x_{t-1} + 0.607x_{t-2} = \xi_t$

and Reinsel (1994). In fact even the two models in Box, Jenkins, and Reinsel (1994) are quite different from each other. This is not surprising because modeling involves some human interference and judgment, and multiple competitive models may exist for the same data set.

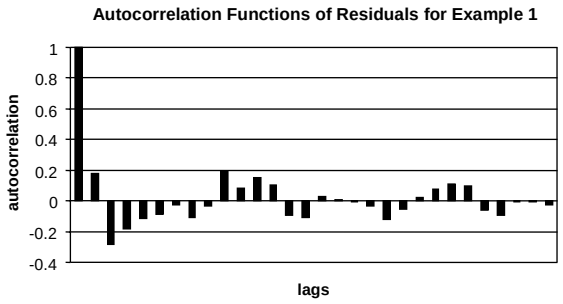

Figure 5. Autocorrelation function of residuals of the model obtained using the new algorithm.

Example 2. Models for the Lake Huron Data

In Brockwell and Davis (1996) a model is set up manually using the Lake Huron data with Box-Jenkins modeling technique. We use the same data set to test the new algorithm and compare the model obtained. Table 2 shows the model found in Brockwell and Davis (1996) and the model identified with the new algorithm.

Table 2. Comparison of Models from Two Sources.

Brockwell and Davis (1996), p.20	New Algorithm
$x_t = 0.791x_{t-1} + \xi_t$	$x_t = 0.996x_{t-1} + \xi_t$

So, both have AR(1) models for the same data set. But, the coefficients are different although quite close. The ACF of the residuals of the new model is shown in Figure 6. The standard deviation of the estimated correlation coefficients is 0.1 using Eqn. (6). From Figure 6, only one out of 40 correlation coefficient data is outside the confidence interval [-0.202, 0.202] and this takes place at lag 3. So, we have a strong belief that the model residual is a white noise.

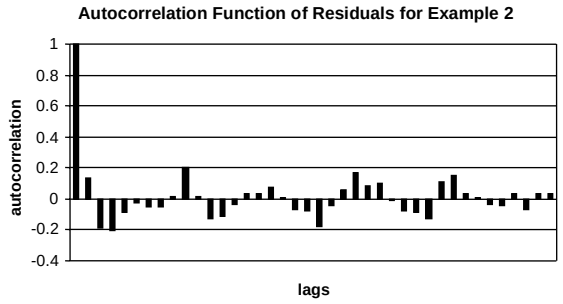

Figure 6. Autocorrelation function of residuals of the model obtained using the new algorithm.

Example 3. Models for the Chemical Process Data

In Box, Jenkins, and Reinsel (1994), two models are developed manually for a chemical process data set. We use the same data set and the new algorithm to create one model. All the models are listed in Table 3. In Box, Jenkins, and Reinsel (1994), two ARMA(1, 1) models are identified whereas with the new algorithm an AR[2] model is identified.

Table 3. Comparison of Models from Two Sources.

Box, Jenkins, and Reinsel (1994), p.255	New Algorithm
$x_t - 0.92x_{t-1} = 1.45 + \xi_t - 0.58\xi_{t-1}$ or $x_t - x_{t-1} = \xi_t - 0.7\xi_{t-1}$	$x_t - 0.5x_{t-1}$ $-0.499x_{t-1} = \xi_t$

The standard deviation of the estimated correlation coefficients is $\sigma = \frac{1}{\sqrt{197}} = 0.0749$ using Eqn. (6). Figure 7 shows the autocorrelation function of the residual of the new model. It can be seen that only two of the autocorrelation coefficient values are outside the confidence interval [-0.15, 0.15] and this takes place at lags 2 and 7 where the values are respectively -0.264 and 0.162407, and the percentage of the data outside the confidence interval is 5%. So, we have a strong belief that the residual process of the new model is a white noise, and hence the model is proper.

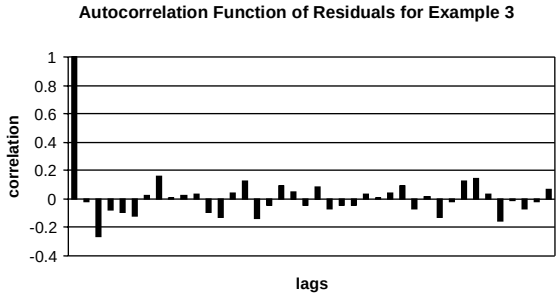

Figure 7. Autocorrelation function of residuals of the model obtained using the new algorithm.

Example 4. An Artificial Time Series Data

In Section 3.1, we use a time series in the illustration of the idea of the new algorithm. Next, assuming the model of the time series is the true model, and generating a set of data using the model, we build a model using the new algorithm and the data generated. For the ease of discussion, we restate the time series model here as $x_t - x_{t-1} + 0.84x_{t-2} = \xi_t + 0.8\xi_{t-1} + 0.57\xi_{t-2} + 0.1\xi_{t-3}$.

To test the new algorithm, 100 data points are generated again, and fed to the algorithm as input. The output of the algorithm is an ARMA(1, 1) model $x_t -$

$0.51x_{t-1} = \xi_t + 0.45\xi_{t-1}$. The model identified is quite different from the "true" model. Figure 8 shows the autocorrelation function of the residual of this model. As 100 data points are used, the standard deviation of the estimated autocorrelation coefficients is $\sigma = 0.1$ using Eqn. (6). From Figure 8, it can be seen that only one out of 40 data is outside the confidence interval [-0.2, 0.2] and this takes place at lag 3. Evidently, we have a strong belief that the residual process is a white noise, and the model is proper albeit it is different from the true model.

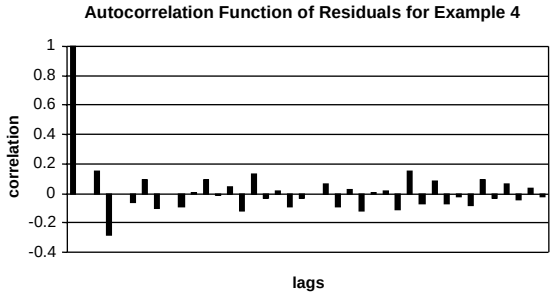

Figure 8. Autocorrelation function of residuals of the model obtained using the new algorithm.

6. CONCLUSIONS AND DISCUSSIONS

In this paper, a new algorithm is proposed for automating time series modeling processes using Box-Jenkins technique, and numerical examples are given. In particular, three examples found in the literature are used to test the algorithm with satisfactory results.

As is known, modeling an ARMA time series is hard using Box-Jenkins technique because for such time series the autocorrelation function or the partial autocorrelation function does not have as significant characteristics as that of AR or MA time series. Thus, it is even harder to determine the values of p and q of an ARMA time series in an automatic mode. In the literature, existing algorithms are available to automate this modeling process, with both pros and cons, as discussed in Section 2. Unlike all other algorithms, statistical hypothesis testing are the only means used in the new algorithm to determine the properties of the residual process and to identify the p or q values of the model in the identification process. This is reflected in both Algorithm 1 and the Main Algorithm. Algorithm 1 is the fundamental step in this new algorithm, and makes automated identification possible using statistical hypothesis.

This is an automatic process. Nevertheless, human experience can play a role. In our testing of the algorithm, it has been found that the significance level to decide whether to reject the hypothesis that the residual process is a white noise is critical in determining the model. Higher rejecting probability leads to a higher order of the model and this is contradictory to the parsimony principle of modeling. Lower rejecting probability tends to generate lower order models. In the algorithm, the percentage of autocorrelation coefficients that are outside the confidence interval is used in decision making. This percentage is directly determined by the maximum number of lags, and this is also determined by the sample size. Thus, how to determine the sample size to maximize the efficacy of the algorithm is an interesting problem. Besides, real data are rarely seen to satisfy the assumptions of normality in the uncertainty. Thus, how to decide the rejecting probability should not be solely determined by a mathematical distribution, such as normal or *t*-distribution. On this aspect it seems reasonable that some extent of tolerance should be permitted. But how much tolerance is allowed should be experimented through real data. Instead, this is where human experience can play a significant role and this is why modeling is an art in addition to being a science. In estimating the standard deviations of autocorrelation coefficients, Eqn. (6) is used. It is needless to say other approximations can be used for this purpose as well. Nevertheless, we leave the question of which approximation might be the best for estimating the standard deviation as a future research problem.

The new algorithm is based on the idea of analyzing the ACF or PACF of the residual of the model, given a chosen model structure and the corresponding estimated parameters. The advantage of this is that either p or q can be selected from the smallest possible values, and p and q are selected separately, not as in some other algorithms where p and q are selected jointly. In addition, as p or q is increased by one at a time and the algorithm terminates the first time it recognizes that the model residual has a strong characteristic of either MA or AR process, the number of parameters so determined in the model should be close to a minimum, which is consistent with the parsimony principle of modeling.

Only stationary time series are considered in this paper. We believe it is the most important step to automate stationary ARMA(p, q) time series modeling as any nonstationary time series can be converted to a stationary one by differencing a few times and by detecting the seasonal cycle, if existing, in the time series. There is a rich literature on this subject, and it is not the interest of the current paper. However, we will discuss this subject in a separate work.

The stopping criterion in this algorithm is the whiteness of the residual process whereas in other algorithms such as that in Reilly (1980) significance of model parameters is the stopping criterion. It will be interesting to know the relationship between the probabilistic properties of parameters and the probabilistic properties of residuals. For example, we are curious to know if all the parameters are significant *the first time*

when the residual process becomes a white noise while p or q increases from, say, 1. This is an interesting research topic in the near future.

It will be interesting to know how this algorithm performs if the maximum likelihood algorithm is used to estimate the parameters whereas the least square algorithm is used in the algorithm implementation. Currently, significance of model parameters is not considered in the algorithm. But, this should be included in the algorithm in the near future.

Both Algorithm 1 and the Main Algorithm can be refined to accommodate human experience. There should be quite a few places in the algorithm where different implementations or refinements are possible. As more studies are carried out, more and better understanding of the new algorithm will be gained, and hence more refinements will be committed.

ACKNOWLEDGEMENT

The authors are grateful to the three anonymous referees for their advice and suggestions on revising the final version of this paper.

REFERENCES

Box, G. E. P., Jenkins, G. M., and Reinsel, G. C. (1994), *Time Series Analysis: Forecasting and Control*, 3rd Ed., Prentice-Hall, Inc., New Jersey.

Brockwell, P. J., and Davis, R. A. (1996), *Introduction to Time Series and Forecasting*, Springer-Verlag, New York.

Reilly, David P. (1980), *Experiences with an automatic Box-Jenkins Modeling Algorithm,* Time Series Analysis, ed. O. D. Anderson, Amsterdam: North-Holland, 493-508.

Reilly, David P. (1987), Experiences with an automatic transfer function algorithm, *Computer Science and Statistics Proceedings of the 19th Symposium on the Interface*, ed. R. M. Heiberger, Alexandria, VI: American Statistical Association, 128-135.

Wu, S. M. and Pandit, S. M. (1979), *Time Series and Systems Analysis*, Modeling and Applications, Wisconsin University.

Akaike, H. (1974), A new look at the statistical model identification, *IEEE Trans. On Automatic Control*, **AC-19**, 716-723.

Schwarz, G. (1978), Estimating the dimension of a model, *Annals of Statistics*, **6**, 461-464.

Talluri, K. T., and van Ryzin, G. J. (2004), *The Theory and Practice of Revenue Management*, Kluwer Academic Publishers, Boston.

Åström, K. J. and Wittenmark, B. (1995), *Adaptive Control*, 2nd Ed., Addison-Wesley.

http://obelia.jde.aca.mmu.ac.uk/resdesgn/arsham/opre330Forecast.htm#rboxjenkm

Graupe, D. (1975), Estimation of Upper Bounds of Errors in Identifying Autoregressive Models, *Int. J. Sys. Sci.*, **6**(1), 1-8.