〈研究論文〉

Journal of the Korean Institute of Industrial Engineers
Vol. 10, No. 1, June 1984

An Approach for the Automatic Box-Jenkins Modelling

Park, Sung Joo*
Hong, Chang Soo**
Jeon, Tae Joon*

Abstract

The use of Box-Jenkins technique is still very limited due to the high level of knowledge required in comprehending the technique and the cumbersome iterative procedure which requires a large amount of cost and time.

This paper proposes a method of automating the univariate Box-Jekins modelling to overcome the limitations of subjective identification in iterative procedure by using Variate Difference method, D-statistic and Pattern Recognition algorithm combined with Akaike's Information Criterion.

The results of the application to real data show that the average performance of automatic modelling procedure is better or not worse, at least, than those of the existing models which have been manually set up and reported in the literature.

1. Introduction

The Box-Jenkins (B-J) forecasting technique is one of the most powerful and accurate forecasting techniques known today. Granger and Newbold [6] have shown the superiority of Box-Jenkins methodology over other Baysian methodologies empirically, while Anderson [3] have also pointed out the B-J forecasting model is one of the most accurate models in the short-term forecasting.

Despite these advantages of the B-J technique, the use of the B-J technique is still very limited due to the high level of knowledge required in comprehending the technique and the cumbersome iterative procedure which requires a large amount of cost, time and effort to implement to the real data.

* Korea Advanced Institute of Science and Technology
**Dae Woo Heavy Ind., Inc.

To overcome these limitations of the B-J technique, Hill and Woodworth (H-W) [8] have proposed an approach to automate Box-Jenkins modelling. They combined a "Pattern Recognition" algorithm to identify an initial set of models, followed by the Final Prediction Error (FPE) criterion to choose among the models. The H-W model is one of the earliest attempts to automate the B-J method. However, there are some limitations in this approach.

1) The order selection criterion FPE is applied during the model estimation procedure. But FPE criterion must be applied at the end of the estimation procedure.
2) In diagnostic checking, only the over-parameterization is considered. Hence, the final model is not approoapriate if the model is under-parameterized.
3) The FPE criterion exihibits limitations due to the high dependency on the number of observations.

The primary purpose of this paper is to develop a method of automating the univariate B-J modelling procedure by using D-statistic and Pattern recognition algorithm combined with Akaike's Information Criterion to overcome the limitations of the H-W model.

2. Automatic Modelling

The general class of models that needs to be defined is the ARIMA (p, d, q), that is, autoregressive integrated moving average model.

$$\phi_p(B)\nabla^d(Z_t - u) = \phi_o + \phi_q(B)a_t \quad \cdots \quad (1)$$

where $\phi_p(B) = 1 - \phi_1 B - \ldots - \phi_p B_p$

$\phi_q(B) = 1 - \phi_1 B - \ldots - \phi_q B_q$

$\nabla^d = (1 - B)^d$

$u = E(Z_t)$

a_t = independently distributed white noise with mean 0 and variance $\sigma_a{}^2$.

In order to set up ARIMA model automatically, the identification and diagnostic checking stages of the B-J iterative modelling procedure need to be automated.

First, the degree of the differencing d must be selected in order to obtain the stationary series and the orders of p and q of the tentative ARMA (p, q) models must be identified,

second, parameters for all tentative models are estimated, and one model is selected among tentative models, and

finally, the selected model is diagnostically checked. If the selected model is adequate, it is used in forecasting. Otherwise, the selected model is modified using the above procedure again.

1) Difference Operater Selection

In order to set up the B-J model, the difference operater must be selected to transfer the non-stionary series into the stationary series. According to Box and Jenkins [4], the degree of differencing is normally 0, 1, 2. Hence the degree is restricted as follows.

$0 \leqslant d \leqslant 2$ for non-seasonal series

$d \geqslant 0,\ D \leqslant 1$ for seasonal series

To select the differencing operator, the "Variate Difference" method is used which select the degree d and D which minimize the variance of the series W_t where $W_t = (1 - B)^d(1 - B^s)^D Z_t$.

2. Tentative Model Selection

Most of the empirical models can generally be represented with a combination of orders within the range.

$$0 \leqslant p, q, P, Q \leqslant 2$$

(1) Nonseasonal Series

Gray, Kelly and McIntire (GKM) [7] have proposed an alternative to Box and Jenkins approach of ARMA (p, q) model identification based on the D (p, q) statistic which can be used to identify orders p and q automatically.

However, the pure moving average process can not be idenfied by the D statistic. To overcome this limitation, Pattern Recognition Algorithm is used in this paper to determine the order q of ARMA (O, q) model.

(2) Seasonal Series

For seasonal series, the efficiency of the use of GKM's D (p, q) statistic is restricted by the number of observations of a series, since the calculation of the D (p, q) in a seasonal series requires a relatively large number of the autocorrelations. Also, inappropriate orders can be identified because autocorrelations at a large lag may not have any statistical meanings. Even though appropriate orders are identified, the resulting model can be over-parameterized because of the non-multiplicative model form.

The Pattern Recognition algorithm using the autocorrelation at lag S, 2S and 3S is applied to identify the orders P and Q of the SARIMA (Seasonal ARIMA) model. Then the orders p and q of ARMA (p, q) model are identified by the method for the non-seasonal series using the autocorrelation function of residual series generated from the estimated SARIMA model, which results in a multiplicative model of the order $(p, d, q) \times (P, D, Q)_S$.

3. Initial Values of Parameters

The calculation of the initial estimates of an ARMA (p, q) process is based on the first p+q+1 autocovariances of $W_t = \nabla^d Z_t$.

1) The autoregressive parameters $\phi_1, \phi_2, \ldots, \phi_p$ are estimated from the autocovariances $c_{q-p+1}, \ldots, c_{q+1}, c_{q+2}, \ldots, c_{q+p}$ by the Yule-Walker equation.
2) Using the esimates ϕ obtained in 1), the first q+1 autocovariances c_j' (j = 0, 1, ..., q) of the derived series

$$W_t' = W_t - \phi_1 W_{t-1} - \cdots - \phi_p W_{t-p} \quad \ldots\ldots\ldots\ldots \quad (2)$$

are calculated.

3) Finally, the autocovaraiances $c_0', c_1', \ldots, c_q'$ are used in linearly convergent process to compute initial estimates of the moving average parameters $\theta_1, \theta_2, \ldots, \theta_q$ and of the residual variances σ_a^2.

4. Final Model Selection

To select a model from the tentative models, a criterion is needed. In addition to the residual mean square, the number of parameters should be considered in the selection criterion based on the Principle of Parsimony.

H. Akaike propoed the Final Prediction Error Criterion and the Akaike's Information Criterion which are based on the principle of entropy maximizations.

(1) Final Prediction Error Criterion (FPE)

The FPE for the model with p parameters is defined as,

$$FPE(P) = [(N+P)/(N-P)]*S(P) \qquad (3)$$

where N = Number of observations in the series
P = Number of parameters in the model
S(P) = The estimate of the variance of the white noise from the model with P parameters.

The FPE criterion is to choose the model whose FPE value is minimum. One apparent drawback of the FPE is that it is difficult to choose the correct model order of the model with high degree of consistency when N is large. This follows from the ratio $[(N+P)/(N-P)]$ which is insensitive to the changes in P when $N \gg P$ as was shown by Gersch and Sharp [5]. For economic and business time series which are very often in the range of 40 to 80 data points, the ratio $[(N+P)/(N-P)]$ will be very sensitive to N.

(2) Akaike:s Information Criterrion (AIC)

The AIC is defined as,

$$AIC(P) = N*\log S(P) + 2P \qquad (4)$$

The constant 2 is a penalty parameter, with which we can prevent the final selected model from being over-parameterized. The AIC is choose the model whose AIC value is minimum.

The relationship between FPE and AIC can be shown as follows.

$$\begin{aligned} \log FPE(P) &= \log [(N+P)/(N-P)*S(P) \\ &= \log S(P) + 2P/N \end{aligned}$$

as N goes to infinity.

Therefore,

$$AIC(P) = N*\log FPE(P) \qquad (5)$$

Thus, for a series in which the number of observations are sufficiently large, the above two criteria give us the same orders.

5. Diagnostic Checking

The inadequacy of the selected model can be classified into two cases, that is, over-parameterization and under-parameterization.

(1) Over-parameterization

This case implies that there exists unnecessary parameter in the selected model. If a certain confidence interval of each parameter contins 0, the parameter can be considered to be unnecessary. The unnecessary parameter can be deleted from the model and the model is re-estimted with the remaining parameters, using estimated value of the remaining parameters as initial values.

If the re-estimated model is better than the original model, the original model is over-parameterized, and the re-estimated model is selected as final model. Otherwise, original model is not over-parameterized.

(2) Under-parameterization

If the selected model is under-parameterized, the residual series from the selected model is autocorrelated which can be modified and added by the Pattern Recognition algorithm.

To check the overall appropriateness of the fitted model, the Portmanteau Lack of Fit test is used which considers the estimated autocorrelations collectively.

Suppose that we have the first K autocorrelations $r_k(a)$ from any ARIMA (p, d, q) process and if the fitted model is appropriate, then the statistics,

$$Q = \Sigma_{k=1}^{K} \; r_k^2 \, (a) \quad \ldots\ldots (6)$$

is appropriately distributed as a Chi-square distribution with degree of freedom K-p-q. By comparing the observed value of Q to a table of the percentage points of Chi-square distribution, a test of the hypothesis of the model adequacy can be made.

3. Application

The automatic Box-Jenkins modelling procedure is applied to actual time series, of which appropriate models have already been set up and reported by specialists in literatures or books. The sample series consists of 8 non-seasonal series and 5 seasonal series as given in Table 1.

The results are summarized in Table 2 and Table 3. The results show that the automatic modelling procedure selects the same difference operator as that of the literature models except Series F. It can also be observed that AIC and FPE criterion are identical in selecting the model from the tentative model.

Table 1. Sample Time Series

No.	Description	Reference
Series A	Chemical Process Concentration Readings.	[4]
Series B	IBM Common Stock Closing Prices.	[4]
Series C	Daily Drybulb Temperatures at noon on Ben Nevis.	[3]

Series D	Yields from Batch Chemical Process.	[4]
Series E	Dow-Jones Utilities Index.	[3]
Series F	Simulated Series from $Z_t = 0.9 Z_{t-1} + a_t$.	[3]
Series G	Chemical Process Viscosity Readings.	[4]
Series H	Chemical Process Temperature Readings.	[4]
Series I	Woman Unemployment in U.K.	[3]
Series J	Mean Monthly Air Temperature at Nottingham Castle.	[3]
Series K	Sales of Company X.	[3]
Series L	International Airline Passengers.	[4]
Series M	Passenger Miles Flown ol Domestic Services.	[3]

Table 2. Summary of Results for Nonseasonal Series

Series (# of points)	Tentative Models ARIMA (p, d, q)	Selected Model ARIMA (p, d, q)	Literature Model ARIMA (p, d, q)
Series. A (197)	(0, 1, 0) (2, 1, 0) (0, 1, 1)	(0, 1, 1)	(1, 0, 1) or (0, 1, 1)
Series. B (369)	(0, 1, 0) (3, 1, 3) (1, 1, 3) (0, 1, 1)	(0, 1, 1)	(0, 1, 1)
Series. C (200)	(0, 1, 0) (2, 1, 2) (0, 1, 2)	(0, 1, 2)	(0, 1, 2)
Series. D (70)	(0, 0, 0) (1, 0, 0) (3, 0, 2) (0, 0, 2)	(0, 0, 2)	(2, 0, 0)
Series. E (78)	(0, 1, 0) (1, 1, 1) (0, 1, 2) (1, 1, 0)	(1, 1, 0)	(1, 1, 1)
Series. F (100)	(0, 1, 0) (2, 1, 1) (1, 1, 1)	(0, 1, 0)	(1, 0, 0)
Series. G (310)	(0, 1, 0) (2, 1, 1) (1, 1, 1)	(0, 1, 0)	(1, 0, 0)
Series. H (226)	(0, 2, 0) (3, 2, 1) (1, 2, 1)	(1, 2, 1)	(0, 2, 2)

Table 3. Summary of Results for Seasonal Series.

Series (# of Points)	Tentative Models (p, d, q) (P, D, Q) s	Selected Model (p, d, q) (P, D, Q)s	Literature Model (p, d, q) (P, D, Q)s
Series. I	(0, 1, 0) (0, 1, 0)		
(67)	(0, 1, 0) (0, 1, 1)		
	(2, 1, 0) (0, 1, 0)		
	(1, 1, 1) (0, 1, 0)		
	(0, 1, 1) (0, 1, 0)		
	(1, 1, 0) (0, 1, 0)	(1, 1, 0) (0, 1, 0)	(1, 1, 0) (0, 1, 0)
Series J.	(0, 0, 0) (1, 1, 0)		(0, 0, 0) (1, 1, 0)
(240)	(0, 0, 0) (0, 1, 2)		
	(2, 0, 0) (0, 1, 2)		
	(2, 0, 2) (0, 1, 2)		
	(0, 0, 1) (0, 1, 2)	(0, 0, 1) (0, 1, 2)	
	(1, 0, 0) (0, 1, 2)		
Series. K	(0, 0, 0) (0, 1, 0)		
(77)	(0, 0, 0) (0, 1, 1)		
	(2, 0, 0) (0, 1, 0)	(2, 0, 0) (0, 1, 0)	
	(1, 0, 1) (0, 1, 0)		
	(0, 0, 2) (0, 1, 0)		(0, 0, 2) (0, 1, 1)
Series L.	(0, 1, 0) (0, 1, 0)		
(144)	(0, 1, 0) (0, 1, 1)		
	(2, 1, 1) (0, 1, 1)		
	(0, 1, 1) (0, 1, 1)	(0, 1, 1) (0, 1, 1)	(0, 1, 1) (0, 1, 1)
Series. M	(0, 1, 0)(0, 1, 2)		
(119)	(0, 1, 0) (0, 1, 2)		
	(2, 1, 0) (0, 1, 2)		(0, 1, 3) (0, 1, 2)
	(2, 1, 0) (0, 1, 2)	(2, 1, 1) (0, 1, 2)	

The selected models correspond to the literature models for series A, B, C, G, I ald L. But for series E, J, and K, the selected model is better than the literature model since the literature model is contained in the set of the tentative models. For series D, F, H, and M, the selected model is different from the literature model and the literature model is not contained in the tentative models, therefore, it is necessary to analyze each model in detail.

For the series D, F, and H, the literature model is better than the selected model in terms of the forecasting accuracy, but for the series M, the selected model is better than the literature model. For the series F, it can be said that the selected model is equal to the literature model since the order of p is approximately 1. The detailed analysis is summarized in Table 4.

Table 4. Summary of Detailed Analysis

Series		Selected Model	Literature Model
Series. D	a.	$Z_t - 51.173 =$ $(1 - 0.323B + 0.310B^2)a_t$	$(1 + 0.339B + 0.19B^2)(Z_t - 51.095)$ $= a_t$
	b.	119.57	119.35
	c.	338.873	338.744
Series. F	a.	$(1 - B) Z_t = a_t$	$(1 - 0.914B)(Z_t - 0.427) = a_t$
	b.	1.104	1.078
	c.	9.910	9.520
Series. H	a.	$(1 - 0.665B)(1 - B) Z_t =$ $(1 - 0.881B) a_t$	$(1 - B) Z_t =$ $(1 - 0.253B - 0.248B^2) a_t$
	b.	0.018	0.020
	c.	-897.322	-884.142
Series. M	a.	$(1 - 0.439B - 0.075B^2)(1 - B)$ $(1 - B^{12}) Z_t = (1 - 0.877B)$ $(1 - 0.47B^{12} - 0.137B^{24}) a_t$	$(1 - B)(1 - B^{12}) Z_t =$ $(1 - 0.419B - 0.082B^2 - 0.23B^3)$ $(1 - 0.447B^{12} - 0.167B^{24}) a_t$
	b.	24.457	23.830
	c.	390.433	387.342

Note: a. Detailed Model
b. Residual Mean Square
c. Value of AIC

4. Conclusion

A method of automating the univariate Box-Jenking modelling procedure by using the D-statistic and pattern recognition technique combined with Akaike's Information Criterion is suggested in this paper.

The procedure has been fully programmed and implemented in the CYBER 174 and applied to 13 sample time series data whose models have already been set up and reported in the literatures. The results showed that the average performance of the automatic modelling procedure is better than or as good as the performance of the manual modelling.

These results suggest the practical applicability of the automatic procedure for the univariate time series which can remove the major limitation of the Box-Jenkins technique.

Automation of transfer function modelling can be the area for the further study.

References

(1) Akaike, H., "A New Look at the Statistical Model Identification.", *IEEE Trans.* AC-19, PP.716-723, 1974.

(2) Akaike, H., "A Bayesian Extension of the Minimum AIC Procedure of Autoregressive Model Fitting", *Biometrica,* 66, PP.239-242, 1979.

(3) Anderson, O.D., *Time Series Analysis and Forecasting the Box-Jenkins Approach,* Butterworths, London, 1976.

(4) Box, G.E.P., and Jenkins, G.M., *Time Series Forecasting and Control,* Holden Day, New York, 1971.

(5) Gersch, W., and Sharp, D.R., "Estimation of Power Spectrum with Finite Autoregressive Models", *IEEE Trans.* AC-18., PP.367-369, 1973.

(6) Granger, C.W.J., and Newbold J.P., "Experience with Forecasting Univariate and Combination of Forecast", *J. R. Stastist. Soc.,* A137, PP.131-165, 1974.

(7) Gray, H.L., Kelly, G.D., and McIntrie, D.D., "A New Appraoch to ARMA Modelling", *Commun. Statis.-Simula. Computa.,* B7, PP.1-77, 1978.

(8) Hill, G.W., and Woodworth, D., "Automatic Box-Jenkins Forecasting", *J. Opl. Res. Soc.,* 31. PP. 413-422, 1980.