Outlier evaluation for the bilinear time series model

Mohamed, I. B.[1] Ismail, M. I.[2]

[1]Institute of Mathematical Sciences, University of Malaya, 50603 Kuala Lumpur, MALAYSIA.
E-mail: imohamed@um.edu.my
[2]Institute of Mathematical Sciences, University of Malaya, 50603 Kuala Lumpur, MALAYSIA.
E-mail: mohdisfahani@perdana.um.edu.my

Keywords: Bilinear, Outlier, Least squares method, Bootstrapping, Rainfall data.

Abstract

The problem of detecting an outlier and then identifying its type for bilinear time series data is studied. The effects of temporary change type of outlier on the observations and residuals for general bilinear processes are considered and the corresponding least-squares measure of the decision threshold is proposed. Due to the complexity of the statistics, we use a bootstrapping method to estimate the mean and standard deviation of the threshold statistics. We look at the ability of the proposed procedure to correctly detect temporary change type of outlier when compared to additive outlier and innovational outlier procedures developed in previous studies. The performances of three bootstrap-based procedures are investigated through simulation studies and shown to be good.

1 Introduction

One of the problems encountered in statistical analysis is the existence of outliers in data including time series data. Previous studies have shown that outliers affect the performance of standard statistical methodology in modeling, forecasting and diagnostic purposes. Chen and Liu (1993) had proposed an iterative procedure for detecting and identifying four different type of outliers; additive outlier (AO), innovational outlier (IO), temporary change (TC) and level change (LC) in integrated autoregressive moving average (ARIMA) models. The approach has been extended to several nonlinear time series models. Charles and Darne(2005) considered detecting AO and IO in GARCH(1,1) models while Battaglia and Orfei (2005) used the Taylor's expansion to approximate the original observations and residuals for AO and consequently detecting AO and IO for general nonlinear model. On the other hand, Zaharim et. al (2006) considered detecting all four types of outlier for bilinear model with order (1,1,1,1) where a new definition of IO was proposed. Later, Ismail et al. (2008) studied on improving the detection of AO in data generated from generalized bilinear process of order (p, q, r, s). In this paper, we extend the work for detecting TC and consequently identifying TC from AO and IO using three bootstrap-based procedures.

This paper is organized as follow: Section 2 discusses the general theory of bilinear model. Section 3 presents the definition and formulation of effects of AO, IO and TC on observations and residuals. In section 4, the statistics to measure the magnitude of outlier effects for AO and IO in BL(p, q, r, s) process are presented while the corresponding statistic for TC is derived. Sections 5 and 6 contain description of three bootstrap-based procedures and their performance based on simulation studies respectively.

2 Bilinear models

The idea of bilinear models was first initiated with the applications on control theory and a real in-depth statistical study was started later by Granger and Andersen (1978). The general bilinear model, denoted by BL(p, q, r, s), is given by

$$Y_t = \sum_{i=1}^{p} a_i Y_{t-i} + \sum_{j=1}^{q} b_j Y_{t-j} + \sum_{k=1}^{r}\sum_{\ell=1}^{s} b_{k\ell} Y_{t-k} e_{t-\ell} + e_t \tag{1}$$

where a_i, c_j and $b_{k\ell}$ are any real numbers satisfying the stationary condition of the model (1) whereas Y_t and e_t are the observation and residual respectively, $t = 1, 2, 3, \ldots$ The e_t's are assumed to follow normal distribution with mean zero and precision τ ($\tau > 0$). The first two components on the right-hand side of (1) are the autoregressive moving average model with parameters p and q, ARMA(p, q). The second last component is the nonlinear term which are used to explain the nonlinearity characteristic of the data being modeled. Thus, ARMA (p, q) is a special case of the BL(p, q, r, s) when $r = s = 0$.

Various methods of estimating the parameters of bilinear models are available. In this paper, the nonlinear least squares estimation method is used to estimate the parameters as suggested by Priestley (1991). The method is recursive in nature and the estimates are obtained when the convergence property is satisfied.

3 Outlier specification in general bilinear model

We used the statistics for measuring AO effect proposed by Ismail et al. (2008) and IO effect based on Battaglia and Orfei (2005). The formulation of outlier effect on observations and residuals are described here.

Let $Y^*_{t,TP}$ be the observed values from BL(p, q, r, s) process such that an outlier of type TP, where TP can be AO, IO or TC, occurs at time point $t = d$ with magnitude ω and let $e^*_{t,TP}$ be the resulting residual when BL(p, q, r, s) is fitted on the contaminated data, $t = 1, 2 \ldots, n$. Further, let Y_t and e_t be the observations and residuals at time t that would have been obtained if there were no outliers in the data and will be referred herewith as 'original observation' and 'original residual' respectively. For $t < d$, clearly $Y^*_{t,TP} = Y_t$ and $e^*_{t,TP} = e_t$. For $t \geq d$ and $k \geq 0$, we consider the definition of AO used by many authors including Chen and Liu (1993) as follows

$$Y^*_{d+k,AO} = \begin{cases} Y_{d+k} & k > 0 \\ Y_{d+k} + \omega & k = 0. \end{cases} \tag{2}$$

Ismail et al. (2008) showed that the residuals are changed according to the following formulation

$$e^*_{d+k,AO} = e_{d+k} + \omega A_{k,AO} \tag{3}$$

where

$$A_{k,AO} = \begin{cases} -1 & \text{for } k = 0 \\ (a_k + \sum_{j=1}^{s} b_{kj} e_{d+k-j}) & \\ -\sum_{j=1}^{k} (\sum_{i=1}^{r} b_{ij} Y^*_{d+k-i,AO} + c_j) A_{k-j,AO} & \text{for } k \geq 1 \end{cases}$$

Equation (2) suggests that the shock caused by an AO affects the original observation at $t = d$ only with a magnitude ω and the rest remain unaffected. Consequently, several original residuals from $t = d$ onward are affected as described in equation (3).

Battaglia and Orfei (2005) used similar definition of IO by Chen and Liu (1993) for ARIMA models such that the original disturbance at time $t = d$ in the residuals only change the residuals at that particular point as below

$$e^*_{d+k,IO} = \begin{cases} e_{d+k} & k > 0 \\ e_{d+k} + \omega & k = 0 \end{cases} \tag{4}$$

Note that this definition is different from Zaharim et. al (2006). Consequently, we formulate the effect of IO on observations. It is given by the following formulation

$$Y^*_{d+k,IO} = Y_{d+k} + f_k \tag{5}$$

where

$$f_k = \begin{cases} \omega & \text{for } k = 0 \\ c_1\omega + b_{11}\omega^2 + \sum_{i=1}^{r} b_{i1}\omega Y_{d+1-i} + (a_1 + \sum_{j=1}^{s} b_{1j} e_{d+1-j}) & \text{for } k = 1 \\ c_k\omega + \sum_{i=1}^{k} (a_i + \sum_{j=1}^{s} b_{ij} e_{d+k-j}) f_{k-i} & \text{for } \geq 2 \end{cases}$$

It can be seen that IO will change the original observation not only at time $t = d$ but also several subsequent observations.

We use similar definition of TC given by Chen and Liu (1993), that is,

$$Y^*_{d+k,TC} = Y_{d+k} + \delta^k \omega \tag{6}$$

where δ is dampening factor of the TC effects. The effects of TC on residuals can be shown to follow the formulation below

$$e^*_{d+k,TC} = e_{d+k} + \omega A_{k,TC} \tag{7}$$

where

$$A_{k,TC} = \begin{cases} 1 & \text{for } k = 0 \\ \delta^k - \sum_{i=1}^{k}(a_i + \sum_{j=1}^{s} b_{ij} e_{d+k-j})\delta^{k-i} & \\ \quad + \sum_{j=1}^{k}(\sum_{i=1}^{r} b_{ij} Y^*_{d+k-i,TC} + c_j) A_{k-j,TC} & \text{for } k \geq 1 \end{cases}$$

Notice that, for TC case, more than one observations and residuals are affected.

4 Statistics to measure the outlier effect

Ismail et al. (2008) derived the least squares statistics for AO as follows:

$$\hat{\omega}_{AO,d} = \frac{\sum_{k=0}^{n-d} e^*_{d+k,AO} A_{k,AO}}{\sum_{k=0}^{n-d} A^2_{k,AO}} \tag{8}$$

where $A_{k,AO}$ is as given in equation (3). On the other hand, Battaglia and Orfei (2005) gave the statistics for IO case as

$$\hat{\omega}_{IO,d} = e^*_{d,IO} \tag{9}$$

The statistics to measure the magnitude of outlier effects for TC is now derived. It is obtained using the least squares method by minimizing $S = \sum_{t=1}^{n} e_t^2$. Thus, we have

$$S = \sum_{t=1}^{d-1} e_t^2 + \sum_{k=0}^{n-d} (e^*_{d+k,TC} - \omega A_{k,TC})^2 \tag{10}$$

Equation (10) was then minimized with respect to ω yielding the following measures of outliers effects for TC:

$$\hat{\omega}_{TC,d} = \frac{\sum_{k=0}^{n-d} e^*_{d+k,TC} A_{k,TC}}{\sum_{k=0}^{n-d} A^2_{k,TC}} \tag{11}$$

In linear ARMA cases, an exact expression of Var($\widehat{\omega}_{AO}$) and Var($\widehat{\omega}_{TC}$) could be derived. However, in the current bilinear cases, the complexity of the formulae make the determination of an algebraic expression are insurmountable. As an alternative, the bootstrap method was used to obtain the estimates of the variances. The method has emerged as powerful tools for constructing inferential procedures in modern statistical analysis. It is carried out through the process of drawing random samples with replacement from the observed residuals. The method has been applied on time series, for example, by Chen and Romano(1999), Pascual et al. (2004) and Zaharim et. al (2006).

5 A general single detection procedure to identify type of outlier

Following Tsay (1986) and Chang et al.(1988), the test statistics for AO or TC, denoted by TP, used are as follows:

$$\widehat{\tau}_{TP,t} = \frac{(\widehat{\omega}_{TP,t} - \bar{\widetilde{\omega}}_{TP,BS,t})}{\widetilde{\sigma}_{TP,BS,t}} \tag{12}$$

where $\bar{\widetilde{\omega}}_{BS} = B^{-1} \sum_{M=1}^{B} \widetilde{\omega}_M$ is the bootstrap mean of statistics of interest, $\widetilde{\sigma}^2_{BS} = \frac{\sum_{M=1}^{B} (\widetilde{\omega}_M - \bar{\widetilde{\omega}}_{BS})^2}{B-1}$ is the bootstrap standard deviation at time t, $\widetilde{\omega}_M$ is the values of statistics from M-th bootstrap sample, M = 1,2,..., B, and B is the number of bootstrap samples considered.

On the other hand, Battaglia and Orfei (2005) used the following statistics for IO case:

$$\hat{\tau}_{IO,d} = \frac{\hat{\omega}_{IO,d}}{\sigma_{I\hat{O},d}} \tag{13}$$

where

$$\sigma_{I\hat{O},d} = \frac{e^{*2}_{r+1} + ... + e^{*2}_{d-1} + e^{*2}_{d+1} + ... + e^{*2}_{n}}{n - r}$$

Ismail et al. (2008) proposed improvement of their procedure for AO case. Thus, we use the 10% trimmed mean method and median absolute deviance given by $\widetilde{\sigma}_{MAD} = 1.483 \times \text{median}\{|\widetilde{\omega} - \widetilde{\omega}_{MED}|\}$ to estimate the variances of the estimators in equations (8),(9)and (11) and investigate whether the performance of procedures improve.

In general, the time point where an outlier occurs is unknown. Hence, a procedure for identifying the type of outlier at a particular point t in BL(p, q, r, s) model is needed. It begins with modeling the original time series $\{Y_t\}$ by assuming that there is no outlier in the data. The maximum values of the test statistics are examined. The procedure is described below:

1. Compute the least squares estimates of BL(p, q, r, s) model based on the original data. Hence, obtain the residuals.
2. Compute $\widehat{\tau}_{AO,t}$, $\widehat{\tau}_{IO,t}$ and $\widehat{\tau}_{TC,t}$ for each t, $t = 1, 2, \ldots, n$, using the residuals obtained in Stage 1.
3. Let $\eta_t = \max\limits_{t=1,2,\ldots,n} \{|\widehat{\tau}_{AO,t}|, |\widehat{\tau}_{IO,t}, |\widehat{\tau}_{TC,t}||\}$. Given a pre-determined critical value C, if $\eta_t > C$ then there is a possibility of an AO, IO or TC occurring at time t.

Through the suggested procedure, the occurrence of AO, IO or TC can be detected at any time t.

6 Simulation

In this section, simulation studies are carried out to investigate the sampling behaviour of $\hat{\tau}_{TC}$ and performance of the outlier detection procedure in detecting outlier and consequently identifying TC in data sets.

6.1 Sampling behaviour of test statistics

The outlier detection procedure is developed based on the maxima of the test statistics measuring the effects of AO, IO and TC. The simulation study in this section is carried out in order to investigate the sampling properties of the maxima of the outlier test statistics. For each model, three cases of sample size are examined, n = 60, 100 and 200. The random errors are assumed to follow standard normal distribution. For each model and each sample size, 500 series are generated. The focus is to examine the sampling behaviour of $\eta_{TP} = \max\limits_{t=1,2,\ldots,n} \{|\widehat{\tau}_{TP,t}|\}$, for TP = TC only. In particular, the percentiles of the test statistics at the 1%, 5% and 10% levels are estimated when no outlier is present in the series.

The results are tabulated in Table 1. Columns 3-5, 6-8 and 9-11 give percentile values for the standard bootstrap method (SB), trimmed mean method (TM) and mean absolute median method (MAD) respectively. Overall, it can be seen that sample sizes ranging from 60 to 100 only slightly change the critical values for all models considered. However, the values increase when $n = 200$ is considered especially for large parameter values. The values are also generally higher for TM method. Based on the results, critical values of 3.0 to 4.0 seem to be appropriate for a series with length of 60 to 200. In practice, more than one critical value is recommended for the analysis.

Table 1: Critical values for TC.

Model	**Sample Size**	**SB**			**TM**			**MAD**		
(a_1, b_{11})		P90	P95	P99	P90	P95	P99	P90	P95	P99
(0.1,0.1)	60	3.01	3.07	3.22	3.85	4.00	4.16	3.56	3.77	3.92
(0.1,0.1)	100	3.30	3.56	3.81	4.15	4.31	4.61	3.70	3.90	4.44
(0.1,0.1)	200	3.49	3.67	4.19	4.44	4.47	5.06	3.94	4.24	4.51
(-0.2, -0.2)	60	3.28	3.36	3.48	4.14	4.27	4.46	3.95	4.09	4.43
(-0.2, -0.2)	100	3.20	3.34	3.67	4.02	4.24	4.71	3.54	3.70	4.16
(-0.2, -0.2)	200	3.24	3.40	3.52	4.10	4.30	4.59	3.71	3.84	4.10
(-0.1,0.3)	60	3.31	3.50	4.38	4.22	4.43	5.57	3.54	3.64	5.00
(-0.1,0.3)	100	3.49	3.61	3.88	4.32	4.48	5.10	3.81	4.04	4.37
(-0.1,0.3)	200	4.01	4.15	5.53	5.00	5.27	7.51	4.40	4.62	6.86
(0.5,0.1)	60	3.41	3.98	4.09	4.58	4.80	5.41	4.08	4.24	4.88
(0.5,0.1)	100	3.52	3.99	4.61	4.36	5.10	6.05	4.01	4.66	5.50
(0.5,0.1)	200	3.89	4.16	4.30	5.10	5.29	5.77	4.10	4.42	4.61

6.2 Performance of the detection procedure

We now investigate the performance of the outlier detection procedure to identify the type of outlier. The procedure is applied to cases characterized by a combination of the following factors:

1. one underlying BL(1,0,1,1) models with different combinations of coefficients.
2. a single TC at $t = 40$ in samples of size $n = 100$.
3. three different values of outlier effect; $\omega = 3$, 5 and 7.
4. critical values; $C = 3$, 3.5 and 4.0.

Table 2: Performance of procedure for identifying TC.

Model	ω	**C=3.0**			**C=3.5**			**C=4.0**		
(a_1, b_{11})		BS	TM	MAD	BS	TM	MAD	BS	TM	MAD
(0.1,0.1)	3	0.38	0.34	0.36	0.26	0.24	0.24	0.10	0.10	0.10
(0.1,0.1)	5	0.84	0.72	0.64	0.82	0.70	0.62	0.72	0.60	0.52
(0.1,0.1)	7	0.79	0.61	0.54	0.79	0.61	0.54	0.79	0.61	0.54
(-0.2,-0.2)	3	0.39	0.39	0.27	0.27	0.27	0.21	0.08	0.08	0.06
(-0.2,-0.2)	5	0.88	0.83	0.79	0.88	0.83	0.79	0.88	0.83	0.79
(-0.2,-0.2)	7	1.00	0.86	0.86	1.00	0.86	0.86	1.00	0.86	0.86
(0.2,-0.2)	3	0.24	0.21	0.27	0.16	0.14	0.21	0.04	0.02	0.04
(0.2,-0.2)	5	0.78	0.72	0.60	0.72	0.70	0.58	0.64	0.66	0.56
(0.2,-0.2)	7	0.93	0.93	0.76	0.93	0.93	0.76	0.93	0.93	0.76
(-0.3,-0.3)	3	0.06	0.06	0.06	0.02	0.02	0.02	0.00	0.00	0.00
(-0.3,-0.3)	5	0.46	0.46	0.37	0.41	0.41	0.33	0.28	0.28	0.24
(-0.3,-0.3)	7	0.80	0.92	0.76	0.80	0.92	0.76	0.68	0.80	0.64

Series are generated to contain a TC. The standard deviation of the noise process for each model is set to be unity. For a given model, 500 series of length 100 are generated using the *rnorm* procedure in S-Plus. Summary of the performance of procedure for identifying TC are given in Table 2. The values represent relative frequency or proportion of correctly detecting TC at t = 40. For instance, the proportions for model (0.1,0.1) with $\omega = 3$ is 0.38 for BS method. That means, 38% of values is greater than 3 giving the percentages of correctly detecting AO at time 40 out of 500 simulations.

It can be seen that the performances of the three procedures improves when larger value of ω is considered. Further, the performance of BS method are generally better than the other two methods.

7 Conclusion

The outlier detection procedure for BL(p,q,r,s) to identify TC from possible type of outliers - AO, IO and TC - that occurs at a particular time point t was proposed in this paper. Simulation study showed that, in general, the procedure works well in detecting and identifying TC. The proportion of correct detection is higher when the magnitude of outlier effect is large.

References

Battaglia, F. and Orfei, L. (2005). Outlier Detection and Estimation in Nonlinear Time Series. *Journal of Time Series Analysis*, 26: 107-121, Blackwell.

Chang, I., Tiao, G. C. and Chen, C. (1988). Estimation of time series parameters in the presence of outliers. *Technometrics*, 30: 193-204, American Statistical Association.

Charles, A. and Darne, O. (2005). Outliers and GARCH models in financial data. *Economics Letters*, 86: 347-352, Elsevier B.V.

Chen, C. W. S. (1997). Detection of additive outliers in bilinear time series. *Computational Statistics and Data Analysis*, 24: 283-294, Elsevier B.V.

Chen, C. and Liu, L. M. (1993). Joint estimation of model parameters and outlier effects in time series. *Journal of American Statistical Society*, 88: 284-297, American Statistical Association.

Chen, H. and Romano, J. P., (1999). Bootstrap-assisted goodness-of-fit tests in the frequency domain. *Journal of Time Series Analysis*, 20 (6): 619-654, Blackwell.

Granger, C. W. J. and Andersen, A. P. (1978). *Introduction to Bilinear Time Series Models*. Gottinge, Vandenhoeck and Ruprecht.

Ismail, M. I, Mohamed, I. B, Yahya, M. S. (2008). Improvement on Additive Outlier Detection Procedure in Bilinear Model. *Malaysian Journal of Science*, In press, University of Malaya Press.

Pascual, L. ,Romo, J. and Ruiz, E. (2004). Bootstrap predictive inference for ARIMA processes. *Journal of Time Series Analysis*, 25 (4): 449-465, Blackwell.

Priestley, M. B. (1991). *Non-linear and Non-stationary Time Series Analysis*. San Diego, Academic Press.

Tsay, R. S. (1986). Time Series Model Specification in the Presence of Outliers. *Journal of the American Statistical Association*, 81: 132-141, American Statistical Association.

Zaharim, A., Mohamed, I. B., Ahmad, I., Abdullah, S., Omar, M. Z. (2006). Performances Test Statistics for Single Outlier Detection in Bilinear (1,1,1,1) models. *WSEAS TRANSACTIONS ON MATHEMATICS*, 5(12): 1359-1364, WSEAS.