Malaysian Journal of Science **27** (2): 107 – 113 (2008)

Improvement on Additive Outlier Detection Procedure in Bilinear Model

Mohd. Isfahani Ismail[1], Ibrahim Mohamed[1*] and Mohd. Sahar Yahya[2]

[1] Institute of Mathematical Sciences, Faculty of Science, University of Malaya, 50603 Kuala Lumpur, Malaysia
* imohamed@um.edu.my (corresponding author)
[2] Centre for Foundation of Studies in Science, University of Malaya, 50603 Kuala Lumpur, Malaysia

Received 21st February 2008, accepted in revised form 2nd June 2008.

ABSTRACT This paper considers the problem of outlier detection in bilinear time series data; with special focus on two most basic models BL(1,0,1,1) and BL(1,1,1,1). The formulation of effect of additive outlier on the observations and residuals has been developed and the least squares estimator of the outlier effect has been derived. Consequently, an outlier detection procedure employing bootstrapping method to estimate the variance of the estimator has been proposed. In this paper, we propose to use the mean absolute deviance and trimmed mean methods to improve the performances of the procedure. Using simulation works, we show that trimmed method has successfully improved the performance. Subsequently the procedure is applied to a real data set.

ABSTRAK Kertas kerja ini mempertimbangkan masalah pengesanan data terpencil di dalam data siri masa bilinear; fokus diberikan kepada dua model asas BL(1,0,1,1) dan BL(1,1,1,1). Zaharim *et. al.* [8] telah mencadangkan formulasi kesan data terpencil tertambah ke atas cerapan dan ralat dan seterusnya menerbitkan penganggar kuasa dua ralat terkecil. Kaedah *bootstrapping* telah digunakan untuk menganggar varians bagi penganggar tersebut. Di dalam kertas kerja ini, kami menggunakan *mean absolute deviance* dan *trimmed mean* untuk memperbaiki pencapaian kaedah tersebut. Dengan menggunakan kerja simulasi, kami telah menunjukkan bahawa kaedah berdasarkan *trimmed mean* telah berjaya memperbaiki pencapaian prosedur. Seterusnya, prosedur ini telah digunakan ke atas data nyata.

(Bilinear, additive outlier, least squares method, bootstrapping, rainfall data)

INTRODUCTION

Fox [1] was among the first to study the problem of outliers in time series. He developed likelihood ratio tests to detect additive outlier and innovational cutlier in non-seasonal autoregressive models of order *p*. Others followed, for instance, [2, 3, 4, 5]. Fewer studies are found on the detection of additive outlier in bilinear models; Chen [6] used Gibbs sampling method for general bilinear model while Ismail *et al.* [7] and Zaharim *et al.* [8] used least squares method for two most simplest order of bilinear model. In this paper, we attempt to improve the performance of the least squares procedure. Instead of using the standard formula of variance, we utilize the mean absolute deviance and trimmed mean methods to estimate the variance of the estimators. We show that the performance is better when trimmed mean method is used.

BILINEAR MODEL

Granger and Andersen [9] had formally introduced the general bilinear model, denoted by BL(p,q,r,s), which is given by

$$Y_t = \sum_{i=1}^{p} a_i Y_{t-i} + \sum_{j=1}^{q} c_j e_{t-j} + \sum_{k=1}^{r} \sum_{\ell=1}^{s} b_{k\ell} Y_{t-k} e_{t-\ell} + e_t \qquad \textbf{(1)}$$

where a_i, c_j and $b_{k\ell}$ are any real numbers satisfying the stationary condition of the model whereas Y_t and e_t are the observation and residual respectively, t = 1, 2, 3, ... The e_t's are assumed to follow normal distribution with mean zero and precision τ, $\tau > 0$. The model is a simplified case of nonlinear Volterra series expansions and extension of general linear autoregressive moving average model of orders p and q.

Various methods of estimating the parameters of bilinear models are available. In this paper, the nonlinear least squares estimation method as proposed by Priestley [10] is used. The method is recursive in nature and the estimates are obtained when the convergence property is satisfied.

THE OUTLIER DETECTION PROCEDURE

The procedure for detecting outliers as proposed by Zaharim *et. al.* [8] is described here. The procedure is meant to detect additive outlier in data generated from BL(1,1,1,1) model, which is given by

$$Y_t = a\,Y_{t-1} + c\,e_{t-1} + b\,Y_{t-1}\,e_{t-1} + e_t \qquad \textbf{(2)}$$

The results also holds for BL(1,0,1,1) models by taking c = 0 in the preceding formulae.

Let Y_t^* be the observed values from BL(1,1,1,1) process with an additive outlier occurs at time point $t = d$ with magnitude ω and e_t^* be the resulting residual when BL(1,1,1,1) is fitted on the contaminated data, $t = 1,2,\ldots,n$. Further, let Y_t and e_t be the residuals 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 $Y_t^* = Y_t$ and $e_t^* = e_t$. For $t \geq d$ and $k \geq 0$, the formulations for Y_t^* and e_t^* is described below.

Let an additive outlier occurs in BL(1,1,1,1) model at time t = d. It has been mentioned in many papers, including [6], that contaminated observation due to additive outlier will differ from the original observations according to the following rule:

$$Y_t^* = \begin{cases} Y_t & t \neq d \\ Y_t + \omega & t = d \end{cases} \qquad \textbf{(3)}$$

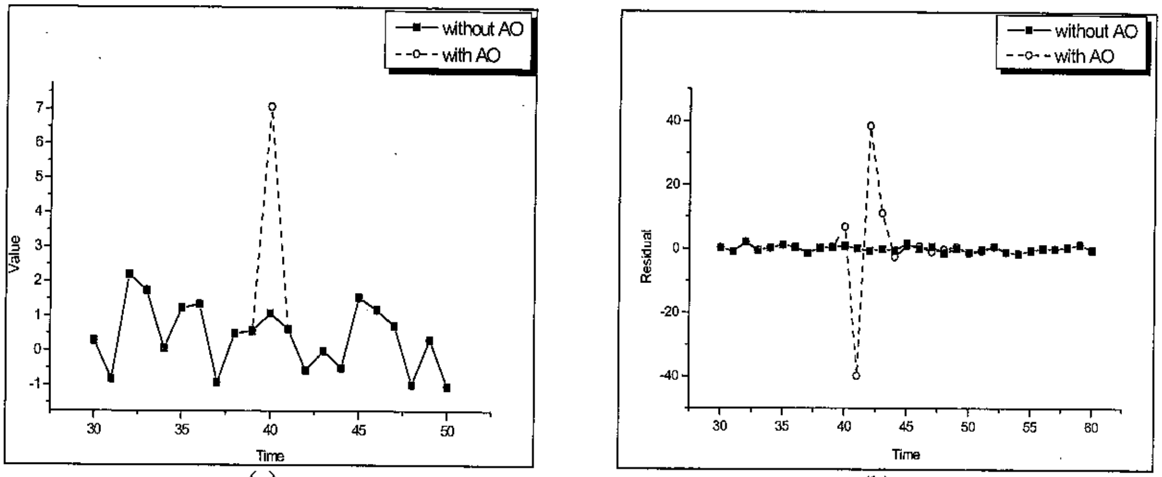
The rule suggests that the shock caused by an additive outlier affects the original observation at $t = d$ only with a magnitude ω and the rest remain unaffected as illustrated in Figure 1(a). Consequently, the residuals will be affected and differ from the original residuals. Zaharim *et. al.* [8] had shown that the effect on residuals can be described by the following formulation:

$$e_{d+k}^* = e_{d+k} + (-1)^k f_{d+k} \qquad \textbf{(4)}$$

where

$$f_{d+k} = \begin{cases} \omega & k=0 \\ \omega(a + b e_d) + \left(bY_d^* + c\right) f_d & k=1 \\ \left(bY_{d+(k-1)}^* + c\right) f_{d+(k-1)} & k=2,3,\ldots \end{cases}$$

Several residuals after $t = d$ were disturbed as illustrated in Figure 1(b).



(a)

(b)

Figure 1. The effect of AO on (a) observations (b) residuals

The statistics to measure the magnitude of outlier effects for additive outlier can now be obtained using the least squares method by minimizing the equation:

$$S = \sum_{t=1}^{n} e_t^2$$

$$= \sum_{t=1}^{d-1} e_t^2 + \sum_{k=0}^{n-d}\left(e^*_{d+k} - \{-1\}^k f_{d+k}(\omega)\right)^2 \qquad \textbf{(5)}$$

Equation (5) is then minimized with respect to ω, yielding the following measure of outlier effect for additive outlier:

$$\hat{\omega} = \frac{\sum_{k=0}^{n-d}\left[\{-1\}^k e_{d+k} A_k\right]}{\sum_{k=0}^{n-d} A_k^2} \qquad \textbf{(6)}$$

where

$$A_k = \begin{cases} 1 & k = 0 \\ (a + be_d) + (bY_d + c) & k = 1 \\ (bY_{d+(k-1)} + c)A_{k-1} & k \geq 2 \end{cases}$$

Zaharim *et. al.* [8] further used the bootstrap method to obtain the estimates of the standard deviation of $\hat{\omega}$. The importance of bootstrap method has been highlighted in many applications on time series data [11]. It is carried out through the process of drawing random samples with replacement from the residuals as described below:

(a) Let $(e_1, e_2, ..., e_n)$ be the original residuals. Sampling with replacement is carried out from the original residuals giving a bootstrap sample of size n, say, $e^{*(1)} = (e_1^*, e_2^*, ..., e_n^*)$. This is repeated a large number of times, say B times, giving B sets of bootstrap samples $e^{*(1)}, e^{*(2)}, ..., e^{*(B)}$.

(b) For each bootstrap sample $e^{*(M)}$, M = 1, 2, ..., B, we calculate $\tilde{\omega}_M$.

(c) The sample standard deviation of $\tilde{\omega}$ is given by

$$\tilde{\sigma}_{BS} = \left\{\frac{\sum_{M=1}^{B}\left(\tilde{\omega}_M - \bar{\tilde{\omega}}_{BS}\right)^2}{(B-1)}\right\}^{1/2} \qquad \textbf{(7)}$$

where

$$\bar{\tilde{\omega}}_{BS} = B^{-1} \sum_{M=1}^{B} \tilde{\omega}_M .$$

Efron and Tibshirani [11] showed that as B → ∞, $\tilde{\sigma}_{BS,t}$ approaches $\hat{\sigma}$, the bootstrap estimate of the standard deviation.

Let H_0 denote the hypothesis that $\omega = 0$ in the bilinear model considered and H_1 denotes the situations $\omega \neq 0$ in bilinear model with additive outlier respectively at time t. The following test statistics can be used to test the hypothesis:

Table 1. The performance of three procedures for BL(1,0,1,1) models

BL(1,0,1,1)							
CO-EFFICIENTS	**MAGNITUDE OF OUTLIER**	**METHODS**	**PROPORTION OF CORRECT DETECTION**				**MIS-DETECTION**
			2.5	**3.0**	**3.5**	**4.0**	
a=0.1 b=0.1	3	Standard	0.56	0.42	0.20	0.14	0.43
		Trimmed mean	0.43	0.42	0.35	0.27	0.57
		MAD	0.44	0.38	0.26	0.23	0.54
	4	Standard	0.75	0.68	0.56	0.39	0.23
		Trimmed mean	0.72	0.71	0.67	0.61	0.28
		MAD	0.70	0.66	0.56	0.41	0.30
	5	Standard	0.91	0.87	0.77	0.68	0.09
		Trimmed mean	0.90	0.90	0.90	0.87	0.10
		MAD	0.92	0.91	0.87	0.76	0.08
a=-0.2 b=0.4	3	Standard	0.33	0.26	0.11	0.04	0.64
		Trimmed mean	0.27	0.25	0.19	0.13	0.70
		MAD	0.26	0.18	0.08	0.06	0.73
	4	Standard	0.36	0.30	0.17	0.14	0.64
		Trimmed mean	0.26	0.26	0.22	0.21	0.74
		MAD	0.23	0.22	0.18	0.11	0.65
	5	Standard	0.53	0.50	0.45	0.38	0.47
		Trimmed mean	0.43	0.43	0.43	0.39	0.57
		MAD	0.48	0.44	0.40	0.35	0.52
a=0.3 b=-0.2	3	Standard	0.51	0.34	0.25	0.15	0.46
		Trimmed mean	0.48	0.45	0.32	0.26	0.51
		MAD	0.44	0.36	0.25	0.18	0.55
	4	Standard	0.76	0.66	0.55	0.35	0.24
		Trimmed mean	0.68	0.68	0.63	0.55	0.32
		MAD	0.70	0.63	0.52	0.37	0.29
	5	Standard	0.77	0.72	0.61	0.50	0.22
		Trimmed mean	0.78	0.77	0.75	0.69	0.22
		MAD	0.76	0.71	0.58	0.56	0.24

$$\hat{\tau}_t = \frac{\left(\hat{\omega}_t - \overline{\tilde{\omega}}_{BS,t}\right)}{\tilde{\sigma}_{BS,t}} \qquad \textbf{(8)}$$

where $\overline{\tilde{\omega}}_{BS,t}$ is bootstrap mean and $\tilde{\sigma}_{BS,t}$ is the bootstrap standard deviation of statistics of interest at time *t*. The following procedure can now be used to detect the occurrence of additive outlier at time *t*:

(1) Compute the least squares parameter estimates of model based on the original data. Hence, obtain the residuals.

(2) Compute $\hat{\tau}_t$ for each *t*, *t* = 1, 2, ..., *n* using the residuals obtained in Stage 1.

(3) Let $\eta_t = \max_{t=1,2,\ldots,n}\left\{\left|\hat{\tau}_t\right|\right\}$. Given a pre-determined critical value C, if $\eta_t > C$, then there is a possibility of an additive outlier occurring at time *t*.

THE IMPROVED VERSION OF OUTLIER DETECTION PROCEDURE

In this paper, we attempt to improve the procedure presented in the previous section.

(a) the mean absolute deviance method
Instead of using equation (7) to calculate the standard deviation of $\hat{\omega}$, we propose to use the method suggested by Hampel *et al.* [12] in which the standard deviation is computed using the following relationship

$$\hat{\sigma}_{MAD} = 1.483 \times \text{median}\left\{\left|\hat{\omega}_t - \tilde{\omega}\right|\right\}$$

where $\tilde{\omega}$ is the median of the bootstrap estimates, $\tilde{\omega}_M$.

(b) the 10% trimmed mean method
The calculation of standard deviation used the trimmed sample such that smallest and largest 10% of $\tilde{\omega}_M$ are removed from the calculation. Equation (7) is then used to give the standard deviation, $\hat{\sigma}_{TM}$.

These methods are expected to be able to overcome the problem of overestimation in the computation of standard deviation.

SIMULATION

The outlier detection procedure is now applied to cases characterized by a combination of the following factors:

(a) two underlying BL(1,0,1,1) and BL(1,1,1,1) models but with different combinations of coefficients.
(b) a single additive outlier at $t = 40$ in samples of size 100.
(c) three different values of outlier effect; $\omega = 3$, 4 and 5.
(d) critical value; 2.5, 3, 3.5 and 4.

Table 2. The performance of three procedures for BL(1,1,1,1) model

BL(1,1,1,1)							
CO-EFFICIENTS	**MAGNITUDE OF OUTLIER**	**METHODS**	**PROPORTION OF CORRECT DETECTION**				**MIS-DETECTION**
			2.5	**3.0**	**3.5**	**4.0**	
a1=0.1 a2=-0.1 b=0.1	3	Standard	0.85	0.67	0.42	0.12	0.15
		Trimmed mean	0.67	0.64	0.57	0.40	0.33
		MAD	0.71	0.57	0.34	0.14	0.29
	4	Standard	0.74	0.61	0.43	0.34	0.21
		Trimmed mean	0.66	0.64	0.60	0.50	0.34
		MAD	0.67	0.60	0.44	0.31	0.33
	5	Standard	0.81	0.74	0.67	0.52	0.14
		Trimmed mean	0.79	0.79	0.79	0.71	0.21
		MAD	0.83	0.74	0.67	0.50	0.17
a1=-0.4 a2=0.1 b=0.2	3	Standard	0.55	0.38	0.24	0.15	0.38
		Trimmed mean	0.45	0.42	0.31	0.25	0.55
		MAD	0.41	0.38	0.29	0.14	0.53
	4	Standard	0.67	0.57	0.43	0.29	0.28
		Trimmed mean	0.62	0.59	0.56	0.48	0.38
		MAD	0.63	0.59	0.41	0.31	0.36
	5	Standard	0.89	0.83	0.67	0.58	0.08
		Trimmed mean	0.85	0.84	0.79	0.74	0.15
		MAD	0.84	0.78	0.70	0.61	0.16
a1=0.3 a2=-0.3 b=0.2	3	Standard	0.66	0.49	0.43	0.29	0.29
		Trimmed mean	0.58	0.53	0.44	0.36	0.42
		MAD	0.52	0.36	0.29	0.26	0.43
	4	Standard	0.81	0.81	0.58	0.42	0.19
		Trimmed mean	0.66	0.66	0.62	0.59	0.34
		MAD	0.71	0.64	0.46	0.39	0.29
	5	Standard	0.77	0.73	0.69	0.65	0.23
		Trimmed mean	0.81	0.81	0.78	0.74	0.19
		MAD	0.77	0.73	0.73	0.73	0.23

Six different series were generated to contain a single additive outlier. For each model, 500 series of length 100 were generated using the *rnorm* procedure in S-Plus. Summary of the performance of the procedures is given in Table 1 and Table 2. In each table, the values in columns 4-7 represent relative frequency or proportion of correctly detecting additive outlier with correct location at $t = 40$ for critical values equal 2.5, 3, 3.5 and 4 respectively for different methods and magnitude of outlier. On the other hand, values in column 8 give the relative frequency of misdetecting the additive outlier at time point different from $t = 40$.

Two main results are observed. Firstly, all three procedures perform quite well. As expected, the performance of the procedures improves when larger value of ω are used. Also, as larger critical values are used, the proportions of detection decrease. However, the performance is reduced when larger coefficient values are used. It is known that when larger coefficient values are used, there tends to be more spikes appearing in the data generated from bilinear process. Consequently, it is expected to be harder to detect the outlier especially for small values of ω. Secondly, in general, the procedure based on trimmed mean has improved the detection of additive outlier compared to the standard procedure. However, the performance of the procedure based on MAD does not differ much from the standard procedure.

APPLICATION: LOCAL RAINFALL DATA

The rainfall data was collected from Kampung Aring weather station, Kelantan, Malaysia for the period of August 1995 till July 2002. The plot of monthly average in millimeter is given in Figure 2. It can be observed that the data is generally stationary in mean and variance except at time points 41 and 77, where heavy rainfalls were heavy.

Nonlinearity test has been widely used to determine whether a given data set is linear or nonlinear [13, 14]. Two such tests are Keenan's test and F-test. When applied on our data, the tests give p-values of 0.0293 and 0.2777 respectively. The Keenan's test strongly suggests that the data is nonlinear and should be fitted using nonlinear time series model.

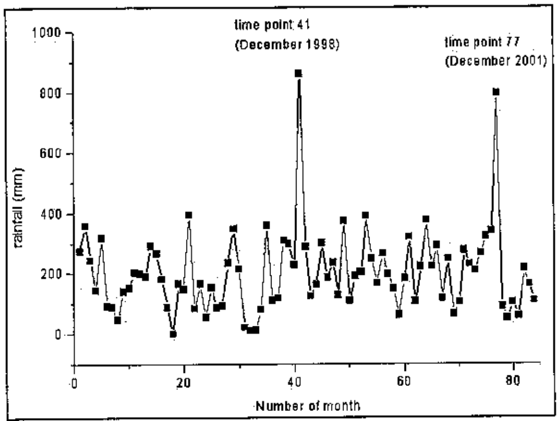


Figure 2. Plot of the Kampung Aring rainfall data

We apply both the BL(1,0,1,1) and BL(1,1,1,1) models on the data. The results are tabulated in Table 3. In can be seen that the values of σ_e^2, AIC, BIC and SBIC for BL(1,0,1,1) model are lower than that of the BL(1,1,1,1) model. Further, diagnostic check-up on the resulting residuals of BL(1,0,1,1) model suggests that the model is pmodel is preferable.

The detection procedure based on BL(1,0,1,1) model is then applied on the data. Results are given in Table 4. Note that, if lower critical point is used, say 2.5, then all three procedures are able to detect the additive outlier at time point 0 as the values of test statistics are greater than 2.5. However, if we choose cut point of 3.5, then only the procedure based on trimmed mean method will detect the outlier. That means that the new procedure based on the trimmed mean will still detect observation 40 as outlier when critical value as high as 4.0 is used.

CONCLUSION

The outlier detection procedure for BL(1,0,1,1) and BL(1,1,1,1) to detect additive outlier that occurs at a particular time point t by using the values of the improved test statistics is proposed in this paper. Simulation study showed that, in general, the three procedures work well in detecting additive outlier with the procedure based on trimmed mean method shows better results compare to the others. The proportion of correct detection is higher when the magnitude of outlier effect is large. The detection procedure is applied on a local rainfall data set and it is able to detect an additive outlier in the data set.

Table 3. Results for selected models for the Kampung Aring Data

MODEL	BL(1,0,1,1)	BL(1,1,1,1)
PARAMETER ESTIMATES	**a = 0.364, se(a) = 0.151** **b = –0.001, se(b) = 0.0004**	**a = 0.624, se(a) = 0.387** ***c* = –0.289, se(c) = 0.402** ***b* = –0.001, se(b) = 0.0004**
Variance of residuals, σ_e^2	18998.02	60116.00
Akaike's Information Criteria (AIC)	1068.66	1075.71
Akaike's Bayesian Information Criteria (BIC)	837.14	847.62
Schwarz's Criterion (SBIC)	835.14	844.62

Table 4. The test statistic value of additive outlier detection procedure on the Kampung Aring rainfall data based on BL(1,0,1,1) model

	METHODS		
POINTS	**STANDARD**	**MAD**	**TRIMMED MEAN**
40	2.9604	3.1308	4.010

ACKNOWLEDGEMENTS

The authors wish to thank the referees and Professor. C. H. Sim for valuable comments and suggestions.

REFERENCES

1. Fox, A. J. (1972). Outliers in time series. *Journal of the Royal Statistical Society* **B 34**: 350 - 363.
2. Tsay, R. S. (1986). Time Series Model Specification in the Presence of Outliers. *Journal of the American Statistical Association* **81**: 132 - 141.
3. Chang, I., Tiao, G. C. and Chen, C. (1988). Estimation of time series parameters in the presence of outliers. *Technometrics* **30**: 193 - 204.
4. Abraham, B. and Chuang, A., (1989) Outlier detection and time series modeling. *Technometrics* **31**: 241 - 248.
5. 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.
6. Chen, C. W. S. (1997). Detection of additive outliers in bilinear time series. *Computational Statistics and Data Analysis* **24**: 283 - 294.
7. Ismail, M. I, Mohamed, I. B. and Yahya, M. S. (2006). The Derivation of Measure of an Additive Outlier Effect in BL(1,0,1,1) Process. *Institute of Mathematical Sciences Technical Report* **12**.
8. Zaharim, A., Mohamed, I.B., Ahmad, I., Abdullah, S. and 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.
9. Granger, C. W .J. and Andersen, A. P. (1978). *Introduction to Bilinear Time Series Models*. Gottinge: Vandenhoeck and Ruprecht.
10. Priestley, M. B. (1991). *Non-linear and Non-stationary Time Series Analysis*. San Diego: Academic Press.
11. Efron, B. and Tibshirani, R. (1986). Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. *Statistical Science* **1**: 54 - 77.
12. Hampel, F. R., Ronchetti, E. O., Rousseeuw, P. J. and Stahel, W. A., (1986). *Robust statistics: The Approach based on Influence Functions*. Toronto: John Wiley.
13. Tsay, R. S. (1986). Nonlinearity test for time series. *Biometrika* **73**: 461 - 466.
14. Keenan, D. M. (1985). A Tukey non-additivity type test for time series nonlinearity. *Biometrika* **72**: 39 - 44.