Asian Journal of Scientific Research

ISSN 1992-1454

science alert
http://www.scialert.net

ANSInet
an open access publisher
http://ansinet.com

Asian Journal of Scientific Research 1 (2): 130-137, 2008
ISSN 1992-1454

Detection of Outliers in Time Series Data: A Frequency Domain Approach

O.I. Shittu and D.K. Shangodoyin
Department of Statistics, University of Ibadan, Ibadan, Nigeria

Abstract: We consider the identification and detection of outliers in frequency domain using the spectral method. By assuming both the additive and multiplicative effect of outliers on a series, the parameters of the model were estimated using the maximum likelihood method with a view to measuring the effect of the suspected outlier on the parameter of the series. The occurrence of outliers has led to a shift in the phase and amplitude of the Fourier series thus affected the periodogram estimates. Further more, detection of aberrant observations is more exact in the frequency domain than in the time domain.

Key words: Outliers, fourier transform, periodogram, robust regression

INTRODUCTION

Considerable attention has been devoted to the detection of outliers in discrete univariate time series were developed for univariate samples in time domain. Fox (1972) and Rosner (1975) started study on outlier detection, Haoglin and Iglewicz (1987) worked on resistant rules for labeling outliers. Chang and Tiao (1983) introduced the additive (AO) and Innovative (IO) models, these were further developed by Shangodoyin and Shittu (2000) the Multiplicative (MO) and Convolution (CO) were proposed in using the model identification tools (ACF and PACF). However in almost all the techniques in time domain Tsay (1986). Shangodoyin and Shittu (2003) detected that outliers were found to have some degree of smearing or swamping effects on other regular observations in the series. Also most economic and social data which are no longer linear but continuous in nature just in physics, engineering and medicine are of the continuous type which can be analyzed in frequency domain.

In this research, we determine the occurrence of outliers in time series data that assumes a Gaussian process and has a continuous spectrum using the spectral method of analysis. An algorithm that uses the robust trigonometric regression of Tatum and Hurvich (1993) is proposed. The estimate of the parameters of the model for the contaminated series is obtained by the maximum likelihood method with a view to compare with that obtained by the least squares method by Priestley (1981)and Brillinger (1981). We also assume the additive and multiplicative effect of outliers on the observed process and the measure of impact of outliers on the observed process and the measure of impact of outliers on the observed values shall be estimated as well as the location of the suspected outlier using a proposed algorithm based on the repeated median transform of Siegel (1982).

ESTIMATION OF PARAMETER USING THE MAXIMUM LIKELIHOOD TECHNIQUE

Here, we estimate the parameters of the model using the maximum likelihood technique with a view to comparing them with that obtained by the least Square method in the literature (Priestley, 1981; Brillinger, 1981).

Let X_t be any periodic stochastic process with period 2π with Fourier representation as:

Corresponding Author: O.I. Shittu, Department of Statistics, University of Ibadan, Ibadan, Nigeria

$$\chi_t = A_0 + \sum_{t=1}^{K} R_i \text{Cos}(w_i t + \phi^i) + \varepsilon_t \quad (1)$$

Where:

w_i : The Fourier frequency;
ϕ_i : The phase uniformly distributed on $(0, 2\pi)$
Ri : The amplitude
ε_t : The random error term NID $(0, \sigma)$

Equation 1 can be re-written as:

$$X_t = A_0 + \sum_{i=1}^{K} (A_i \text{Cos } w_i t + B_i \text{Sin. } w_i t) + \varepsilon_t \quad (2)$$

$A_i = R_j \cos \phi_i$ and $B_i = R_j \sin \phi_i$ are parameters to be estimated and ε_t is a purely random process, normal and independently distributed with:

$$E(\varepsilon_t) = 0 \text{ and } (\varepsilon_t^2) = \sigma_\varepsilon^2$$

Where, σ_c^2 is a further unknown parameter and

$$W_i = \frac{2\pi k}{N}, \quad k = 1, 2, 3, \ldots \frac{N}{2}$$

Since $\varepsilon_t \sim$ N I D $(0, \sigma^2)$ the distribution function of ε_t can be given as:

$$g(\xi_t) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2} - \varepsilon_t^2 / \sigma^2} \quad (3)$$

With a corresponding maximum likelihood function:

$$L(\vartheta) = \prod_{i=1}^{n} g(\xi_t) = \frac{1}{(2\pi\sigma^2)^{-n/2}} e^{-\frac{1}{2\sigma^2} \sum_{i=0}^{n} [X_t - A_0 - \sum A_i \text{Cos. } t + B_i \text{Sin. } w_i t]^2}$$

and log- likelihood:

$$L = \text{In}L(\vartheta) = \frac{n}{2}\text{In}(2\pi) - \frac{n}{2}\text{In}\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{K} [X_t - A_0 + \sum_{1}^{k} (A_K \text{Cos } w_i t + B_K \text{Sin } w_i t)]^2 \quad (4)$$

The maximum likelihood estimate of the A_o, A_i and B_i are

$$\hat{A}_o = \sum_{i=1}^{n} X_t / n \quad (5)$$

$$\hat{A}_i = \frac{2}{N}\sum_{t=1}^{n} X_t \cos \omega_i t \qquad i = 1,2,..., \frac{N}{2} \tag{6}$$

and

$$\hat{B} = \frac{2}{N}\sum_{t=1}^{n} X_t \sin \omega_i t \qquad i = 1,2,..., \frac{N}{2} \tag{7}$$

It can be shown that the estimates $\hat{A}_i$ and $\hat{B}$ are unbiased with variances:

$$Var(\hat{A}_i) = \frac{2}{N}\hat{\sigma}_e^2$$

and

$$Var(\hat{B}_i) = \frac{2}{N}\hat{\sigma}_e^2$$

Where:

$$\hat{\sigma}_e^2 = \frac{1}{N\text{-}K}\sum_{t=1}^{N}\left[X_t - \sum_{t=1}^{K}(A_i \text{ Cos } w_i t + B_i \text{ Sin } w_i t)\right]^2$$

is the unbiased estimate of the residual variance.

ESTIMATION OF PARAMETERS OF A CONTAMINATED SERIES

Our focus here is to derive estimates of the parameters of outlier contaminated series.

The Additive Model

Suppose outliers have additive effect on a series, we assume the additive outlier generating model of Tsay (1986).

The additive model is given by:

$$X_t = Z_t + D\xi_t^{(T)} \tag{8}$$

Where, X_t is the observed series; Z_t is the outlier free series; and D is the magnitude of the outlier $\xi_t^{(T)}$ is the time indicator of the outlier such that

$$\xi_t^{(T)} = \begin{cases} 1, t = T \\ 0, t \neq T \end{cases}$$

Using (8) in (4) gives the maximum likelihood function:

$$L(\theta) = \prod_{i=1}^{n} g(x) = \frac{1}{(2\pi\sigma)^{n/2}} e^{\frac{1}{\sigma^2}\sum_{i=1}^{n}[Z_t + D\xi_t^{(T)} - A_o - \sum_{i=1}^{k}(A_1 \cos\omega_1 + B_1 \sin_j)]^2} \quad (9)$$

and the log-likelihood function:

$$L = \ln L(\theta) = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}[Z_t + D\xi_t^{(T)} - A_o - \sum_{i,j=1}^{k}(A_i \cos\omega_i t + B_j \sin\omega_j t)]^2$$

The maximum likelihood estimate of the A_o is:

$$\hat{A}_o = \frac{1}{n}\sum_{i=1}^{n} Z_t + \frac{1}{n}\sum_{i=1}^{n} D\xi_t^{(T)} \quad (10)$$

When $t = ; A_o = \bar{Z}_t + \bar{D}$

the estimate of the magnitude of outlier is:

$$D\xi_t^{(T)} = A_o - \frac{1}{n}\sum_{i=1}^{n} Z_t \quad (11)$$

at $t = T$, $\xi_t^{(T)} = 1$, therefore:

$$\hat{D}_T = \hat{A}_o - \bar{Z}_T \quad (12)$$

and for reasons of orthogonality when there is no outlier:

$$X_T = \hat{A}_o \quad \text{hence,} \quad \hat{D}_T = X_T - \bar{Z}_T \quad (13)$$

The maximum likelihood estimate of the A_i and B_i are:

$$\hat{A}_i = \frac{2}{N}\sum_{i=1}^{n} Z_t \cos\omega_i t \quad (14)$$

and

$$\hat{B}_i = \frac{2}{N}\sum_{i=1}^{n} Z_t \sin\omega_i t \quad (15)$$

It could be observed that from (12), (14) and (15) the occurrence of outlier has influenced only $\hat{A}_0$ (the grand mean) and no influence on $\hat{A}_i$ and $\hat{B}_i$ for (AO) model. However, the influence on $\hat{A}_0$ could be monotone increasing or decreasing depending on whether $\hat{D}$ is positive or negative.

Thus the occurrence of outlier in any series does not affect the periodogram $I(\omega) = \hat{A}_i^2(\omega) + \hat{B}_i^2(\omega)$ for the (AO) model.

The Multiplicative Model (MO)

Suppose outlier have multiplicative effect on a data set, without loss of generality, we consider the multiplicative outlier generating model (MO) (Shangodoyin and Shittu, 2003):

$$X_t = Z_t D\xi_t^T \tag{16}$$

Where, X_t is the observed series; Z_t is the outlier free series; and D is the magnitude of the outlier $\xi_t^{(T)}$ is the time indicator of the outlier such that

$$\xi_t^{(T)} = \begin{cases} 1, t = T \\ 0, t \neq T \end{cases}$$

Using (16) in (4) we have the log-likelihood function:

$$L = InL(\theta) = -\frac{n}{2}In(2\pi) - \frac{n}{2}In\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}[Z_t D\xi_t^{(T)} - A_o - \sum_{i,j=1}^{k}(A_i \cos\omega_i t + B_j \sin\omega_j t)]^2 \tag{17}$$

and the estimate of:

$$\hat{A}_0 = \frac{1}{n}\sum_{i=1}^{n} Z_t D\xi_t^{(T)} \tag{18}$$

$$\text{When } t = T;\ \hat{A}_0 = \bar{Z}_t \bar{D} \tag{19}$$

While the MLE of the magnitude of outlier for the multiplicative model is:

$$D\xi_t^{(T)} = \frac{A_o \sum_{t=1}^{n} Z_t}{\sum_{i=1}^{n} Z^2{}_t} - \left[I^*(\omega_i)\right] / \sum_{i=1}^{n} Z^2{}_t \tag{20}$$

Where:

$$I(\omega_i) = \frac{n}{2}\left(A_i^{\ 2} + \frac{n}{2}B_j^{\ 2}\right)$$

the Normalized Periodogram.

The maximum likelihood estimate of the A_i and B_i are:

$$\hat{A}_i = \frac{2}{N}\sum_{i=1}^{n} Z_t D\xi_t^T \cos\omega_i t \tag{21}$$

and

$$\hat{B}_i = \frac{2}{N}\sum_{i=1}^{n} Z_t D\xi_t^T \sin\omega_i t \tag{22}$$

However when there is no outlier, that is when t # T and $\xi_{\ni t}^{(T)} = 0$, the estimates of $\hat{A}_i$ and $\hat{B}_i$ are

$$\hat{A}_i = \frac{2}{N}\sum_{i=1}^{n} Z_t \cos\omega_i t$$

and

$$\hat{B}_i = \frac{2}{N}\sum_{i=1}^{n} Z_t \sin\omega_i t \text{ ,}$$

respectively.
Which are also unbiased

DETECTION OF OUTLIERS USING THE DERIVED ESTIMATES

The derived estimates shall now be used to diagnose for suspected outliers using the following proposed algorithms.

Algorithm I (Detection of Outlier)

If observations $X_1, X_2, \ldots ; X_N$ can be expresses as a sum of sine and cosine waves as in (2) which can be written as:

$$X_t = \sum_{i=1}^{k} (\alpha_i \text{Cos } \omega_i t + \beta_i \text{ Sin } \omega_i t) + \varepsilon_t$$

Using any of the spread sheet package or Microsoft Excel to

- Obtain the estimate of the Fourier frequencies $\hat{\omega}_i = \frac{2\pi nk}{N}$ for k = 1, 2. . . $^N/_2$ and
- i = 1, 2, . . . k hence the periodogram $I_N(\omega_i)$ for all ω in the range - Π < ω < Π by

$$I_N(\omega) = \left\{\alpha_{(\omega)}^2 + \beta(\omega)^2\right\}$$

Where, $\alpha(\omega)$ and $\beta(\omega)$ are as defined in (9) and (10).

If $\hat{\omega}_i$ is very close to its true value then $\hat{\alpha}(\omega)$ and $\hat{\beta}(\omega)$ will also be close to $\alpha(\omega)$ and $\beta(\omega)$, respectively, hence the squared amplitude will be non-zero. However, if $\hat{\omega}_i$ is substantially far from its expected value the periodogram will be close to zero.

- Determine the value of ω_i; i = 1, 2,. . . . K whose squared amplitude is non-zero.
  Obtain the residual variance and Compute the test statistics:

$$\lambda_i = \frac{X_t}{\hat{\sigma}_\varepsilon^2} \quad \text{for } i = 1, 2, \ldots, N$$

- Determine $\lambda_F = \text{Max}_{(1<F<N)}\ \lambda_i$
- For all $\lambda_{Fi} > C$ where, C is the critical value simulated as 1.00 or 1.10, the observation X_t^s corresponding to λ_{Fi} is declared an outlier for the Fourier frequencies ω_i; i = 1, 2,.... k whose squared amplitude is non-zero.
- Use the Repeated median filter of Siegel (1982) and for all $\omega_i \neq 0$; compute the estimate discrete Fourier transform:

$$X_t^F = \tilde{X} + \sum_{0<t=<N/2}^{N} (\alpha_i \text{Sin}\ \omega_p t + \beta_p \text{Sin}\ \omega_p t)$$

* X_t^F gives the uncontaminated data set whose contamination/outlier has been removed.

DATA ANALYSIS

To show the use of the above algorithm, five different natural and well analysed data were used. They are series A: Zadakat data in a local mosque in Nigeria; series B: Wolfer Sunspot data, a record of activities in the solar system; series C: Batch chemical data; series D: Well analyzed data from Box and Jenkins (1976). Nigerian Consumer Price index data obtained for the Federal Office of Statistics; and series E: Diabetic patient data from the University teaching Hospital, Ibadan, Nigeria.

The algorithm I was used to diagnosed collected data for outliers using the Microsoft Excel package and the results were summarized in the Table 1-3.

Table 1: The timing and magnitude of outliers

Series A: N = 146 (Zadakat data)

Timing (T)	Observed value (O)	Magnitude of outliers (D)
132	135.5	77.27
138	130.0	76.96

Table 2: The timing and magnitude of outliers

Series B: N = 100 (Sunspot data)

Timing (T)	Observed value (O)	Magnitude of outliers (D)
9	154	118.76
10	125	78.17
18	132	101.86
19	131	108.01
20	118	60.82
67	122	98.69
68	138	110.24
79	124	34.01

Table 3: The timing and magnitude of outliers

Series E: N = (UTH data)

Timing (T)	Observed value (O)	Magnitude of outliers (D)
47	58	41.03
63	1	-51.90
83	41	22.30
99	35	17.62

It should be noted that no outlier were detected in series C (N = 48) and series D (N = 70)

CONCLUSIONS

The derived estimates using the Maximum Likelihood Method (MLE) compares favourably with the Least Squares Method (LSM). This confirms the remark made by Priestley (1981) that Maximum likelihood method is a more asymptotically fully efficient method of estimation. It was found out that the contamination has no influence on the estimates of $\hat{A}_i$ and $\hat{B}_i$ for (AO) model. However, the influence on $\hat{A}_0$ could be monotone increasing or decreasing depending on whether $\hat{D}$ is positive or negative, however for the multiplicative model, the influence on the parameter estimates were noticeable under the null hypothesis that there is contamination in the series.

We found that 2, 6 and 4 observations were identified as outliers in series A, B and E, respectively as shown in Table 1-3 while no observation were identified in series C and D. This is not to say that the algorithm can not work for small sample size data (i.e., n<100) as studies have shown that the procedure performs efficiently in any series were contamination is apparent.

It was also observed that using the spectral method of analysis in the frequency domain, the detection of aberrant observations were more exact than in those techniques in discrete domain.

With the Robust repeated median transform, it can also be observed that the issue of swamping or masking effect does not arise as outlying observations can be detected more exactly. The Robust repeated median transform technique is more complex and involves a lot of iterations; it is also more extensive computationally than other techniques.

RECOMMENDATIONS

Because of the fact that the number of outliers present in a set of data can not be determined aprori, it is recommended that every set of data, especially time series data should be diagnosed for outliers; the detected outlier should be treated or accommodated by any known method, before further analysis could be carried out.

Future research should emphasize on the identification and detection of outliers in Multivariate and categorical data as well as the extension of multiple outlier detection technique to the frequency domain.

REFERENCES

Box, G.E.P. and Jenkins, 1976. Time Series Analysis and Control. Holden Day, San Francisco.

Brillinger, 1981. Time Series Data Analysis and Theory. Expanded Edition, McGraw-Hill, Inc., New York.

Chang, I. and G.C. Tiao, 1983. Estimation of time series parameters in the presence of outlier. Technical Report 8, University of Chicago, Statistic and Research Center.

Fox, A.J., 1972. Outlier in time series. Roy. Stat. Soc., 34: 350-363.

Haoglin, D.C. and B. Iglewicz, 1987. Fine-turning some resistant rules for outlier labeling. J. Am. Stat. Ass., 82: 1147-1149.

Priestley, M.B., 1981. Spectral Analysis and Time Series. New York, Academic Press.

Rosner, B., 1975. On detection of many outliers. Technometrics, 17: 221-227.

Shangodoyin, D.K. and O.I. Shittu, 2000. Some recent advances in multiple outlier detection technique. J. Sci. Res., 6: 12-19.

Shangodoyin, D.K. and O.I. Shittu, 2003. Single outlier generating models-new strategies. J. Sci. Eng. Technol., 10: 5166-5177.

Siegel, A.F., 1982. Robust regression using repeated medians. Biometrics, 69: 240-244.

Tatum, L.G. and C.M. Hurvich, 1993. High breakdown methods of time series analysis. J. R. Stat. Soc. B., 55: 881-896.

Tsay, R.S., 1986. Time series model specification in the presence of outlier. J. Am. Stat. Asso., 81: 132-141.