Outliers in Time Series

ST 565 Final Project (Winter 2016)

Amir Azarbakht, Michael Dumelle,
Camden Lopez & Tadesse Zemicheal

Introduction

Throughout the quarter, we have used various methods to model time series data. However, we have not encountered any data that has been affected by the presence of one or more outliers. In various statistical settings, outliers can have a profound impact on the way data is analyzed. If outliers are not properly adjusted for, inference on parameters and prediction of future observations will be less reliable. In some extreme cases, it may be meaningless. Consider a time series subject to the presence of outliers. How do we quantify an outlier in the time series setting? Are the outliers seemingly random, or is there some pattern among them? How do we properly account for the effect of these outliers? These are all questions we intend to answer throughout this report, and understanding them is crucial to correctly handling time series data.

[*Goal is to model "normal" or "baseline" behavior without model fit being influenced by anomalies (or measurement errors)*]

Throughout this report, we will look at four different types of outliers: innovational outliers (IO), additive outliers (AO), temporary change outliers (TC), and level shift outliers (LS). We intend to examine the effects that these outliers have on standard ARIMA models. For simplicity, we will only consider outliers in non-seasonal models. This problem becomes increasingly complex when a seasonality component is added. We intend to use an iterative algorithm described by Chen and Liu (1993) to detect these four outlier types. This algorithm is implemented in the R package `tsoutliers`. We will apply this algorithm to the `ipi` data set in `tsoutliers` and use graphical tools to aid our analysis. This data set contains economic indices from several European countries from 1999 to 2013. After successful detection, we will also begin exploring ways to incorporate this information into our model building process to obtain reliable parameter estimates and predictions.

Outlier Models

We say that X_t follows an ARIMA (p,d,q) model if it has the form

$$X_t = \frac{\theta(B)}{\phi(B)(1-B)^d} Z_t \tag{1}$$

where B is the standard backshift operator, Z_t is white noise process (identically and independently distributed as $N(0, \sigma^2)$), $\phi(B) = 1 - \alpha_1 B - \alpha_2 B^2 ... - \alpha_p B^p$, $\theta(B) = 1 - \beta_1 B - \beta_2 B^2 - ... - \beta_p B^q$, d is the order of differencing, and t = 1,. . . ,n, where n is the number of observations in the time series. It is further assumed the roots of $\phi(B)$ and $\theta(B)$ are outside unit circle (this gives us stationarity and invertibility) and have no common factors. Now, to describe a time series which is influenced by a nonrepetitive event, we will consider the following model

$$X_t^* = X_t + \omega \frac{A(B)}{G(B)H(B)} I_1(t_1) \tag{2}$$

where $I_t(t_1) = 1$ if t = t_1, and 0 otherwise. $I_t(t_1)$ acts as an indicator function for the outlier impact occurence, with t_1 being the location (possibly unknown) of the outlier. $\omega \frac{A(B)}{G(B)H(B)}$ denotes magnitude and pattern of some event. We will assume these are not previously known. Outliers are then classified by imposing a structure on $\frac{A(B)}{G(B)H(B)}$ and estimating a value of ω. In this report, we will focus on four main types of outliers: Additive (AO), Innovational (IO), Level Shift (LS), and Temporary Change (TC).

Innovational Outlier (IO)

An innovational outlier is regarded as an initial impact with effects lingering over future observations. For example, consider a time series recording seismic wave activity in a specific area. Now, assume an earthquake occurs. Seismic wave activity will increase dramatically, and the effects of this earthquake will be present long time thereafter. More formally, we define as innovational outlier as follows.

An innovational outlier at time t, is of the form

$$X_t^* = (X_t + \omega I_t(t_1)) \frac{\phi(B)}{\alpha(B)\omega(B)} \tag{3}$$

One can see that IO's are obtained by letting $\frac{A(B)}{G(B)H(B)} = \frac{\phi(B)}{\alpha(B)\omega(B)}$

It is important to note that these outliers are **not independent** of the model, as they rely on $\theta(B)$, $\alpha(B)$, and $\phi(B)$ terms. Thus, future values of the time series are affected by the IO. When an IO outlier occurs at time $t = t_1$, the effect of this outlier on $Y_{t_1+k} = \omega\psi_k$, where $k \geq 0$ and ψ_k is the kth coefficient of the polynomial ψ_k is defined as

$$\psi(B) = \frac{\theta(B)}{\alpha(B)\phi(B)} = \psi_0 + \psi_1 B + \psi_2 B^2 + ... + \psi_{max(p,d,q)} B^{max(p,q,d)}, \text{where } \psi_0 = 1 \tag{4}$$

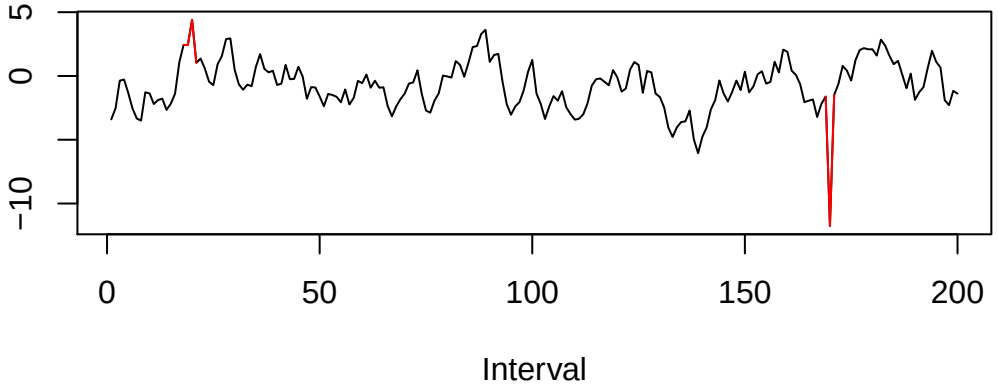
Because the time series is stationary and invertible, we have that the ψ's tend to 0. This implies that the IO we will deal with produce a temporary effect (consider the sales example

above) In other words, the IO at time t_1 does not effect the value in the time series at time $t_1 + k$, for large enough k. In other cases, it is also possible that the IO produces an initial effect and a subsequent permanent level shift.

Additive Outlier (AO)

An additive outlier (AO) corresponds to an exogenous change of a single observation of the time series. It is associated with incident like measurement errors or impulse effect due to external forces. For an example of an additive outlier, consider a time series recording highway traffic. If rainfall is abnormally larger than expected, traffic would likely be much higher than normal. More formally, an additive outlier at time t is defined as

$$X_t^* = X_t + \omega I_t(t_1) \tag{5}$$



which is obtained by setting $\frac{A(B)}{G(B)H(B)} = 1$. The effect of these outliers on the response at time t is **independent** of the ARIMA model initially chosen, and does not depend on any function of B. We see that an additive outlier simply acts as a shift in the value of the response when $t = t_1$. The rest of the model remains unchanged when $t \neq t_1$, as $I_t(t_1) = 0$.

Level Shift (LS)

A level shift outlier (LS) produces a sudden, permanent change in the time series values. Consider a time series of stock market prices for a given company. Now, assume there is a new regulation the company must adhere to, greatly changing how the company operates. Stock prices would likely jump up or down and stay there, depending on the regulation. More formally, a level shift outlier is defined as follows

$$X_t^* = X_t + \frac{\omega I_t(t_1)}{(1-B)} \tag{6}$$

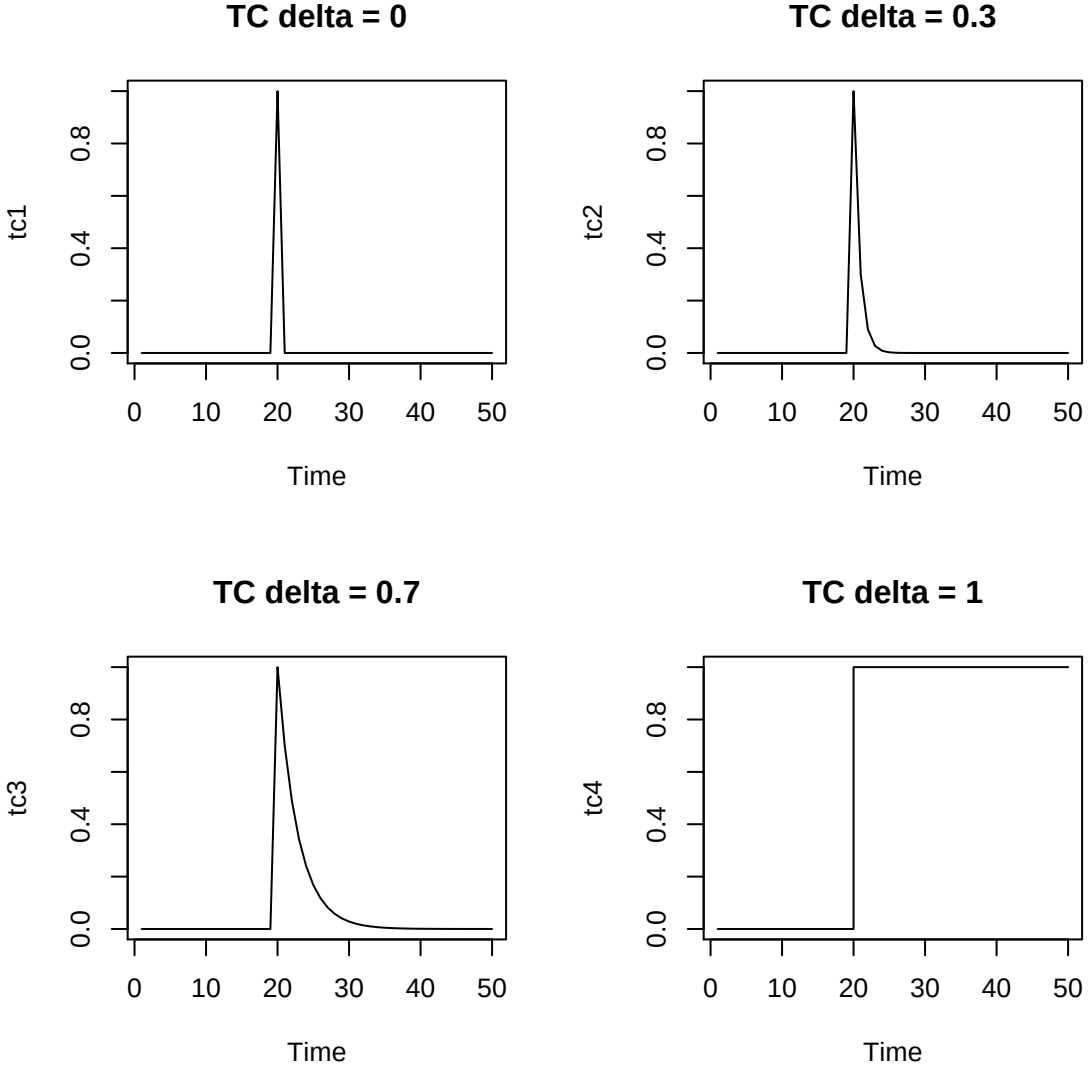
which is obtained by setting $\frac{A(B)}{G(B)H(B)} = \frac{1}{(1-B)}$. Moreover, we see that these outliers are **independent** of the original model, as they simply represent a shift in the response.

Temporary Change (TC)

A temporary change (TC) outlier produces an abrupt change in the time series which decays over time. For example, consider a time series recording daily profit for a retail store. Now suppose there is some month long sporting sale at the store which draws large crowds. The restaurant experiences inflated sales due to the influx of people. It is likely that at the beginning of the sale, a lot of people will visit the store. As the month progresses the total sales per day should decrease, because many people have already gotten the good deals. After the month ends, sales will go back to normal. This illustrates the effect of the temporary change outlier. We rigorously define the temporary change outlier as

$$X_t^* = X_t + \frac{\omega I_t(t_1)}{(1 - \delta B)} \tag{7}$$

We can see that $\frac{A(B)}{G(B)H(B)} = \frac{1}{(1-\delta B)}$ and that these outliers are **independent** of the model chosen.



Notice that the temporary change outlier is a generalization of the additive and level shift outliers ($\delta = 0$ and $\delta = 1$, respectively). In the special level shift case, the time series actually does not decrease over time, as it is a permanent change. δ is used to model the pace of the decaying effect, and is often specified according to the specific needs of the researcher.

Outlier Detection and Estimation

In this section we consider the problem of detecting the presence of an unknown number of outliers occuring at unknown times in a time series. The outliers may be any of the four types described above. First, we describe how the effects of a single outlier can be estimated using multiple regression. Then we point out some practical issues, and the section culminates in a description of the procedure proposed by Chen and Liu (1993) to jointly estimate model parameters and outlier effects in a series.

Estimating the Effect of a Single Outlier

When examining the effect of outliers, we must first assume the time series parameters are known. Next, define the polynomial

$$\pi(B) = \phi(B)\alpha(B)/\theta(B) = 1 - \pi_1 B - \pi_2 B^2 - \pi_3 B^3 - ... \tag{8}$$

where the π_j weights for large j become essentially 0, as the roots of $\theta(B)$ all have absolute value greater than 1 (implying the infinite series converges). We can estimate residuals, $\hat{e}_t$, with the expression

$$\hat{e}_t = \pi(B)Y_t^*, \text{ for } t = 1, 2, ... \tag{9}$$

It is important to note that these residuals are subject to outlier effects. We can express the residual for the four types of outliers by the formulas below.

$$\hat{e}_t = \omega\pi(B)I_t(t_1) + a_t \qquad \text{Additive Outlier} \tag{10}$$

$$\hat{e}_t = \omega I_t(t_1) + a_t \qquad \text{Innovational Outlier} \tag{11}$$

$$\hat{e}_t = \frac{\omega\pi(B)I_t(t_1)}{1 - B} + a_t \qquad \text{Level Shift} \tag{12}$$

$$\hat{e}_t = \frac{\omega\pi(B)I_t(t_1)}{1 - \delta B} + a_t \qquad \text{Temporary Change} \tag{13}$$

We can express the above equations alternatively as $\hat{e}_t = \omega x_{it} + a_t$, for $t = t_1, t_1+1, t_1+2, ..., n$ and $i = 1, 2, 3, 4$, where

$$x_{it} = 0 \text{ for all } i \text{ and } t < t_1 \tag{14}$$

$$x_{it_1} = 1 \text{ for all } i \text{ and } k \geq 1 \tag{15}$$

$$x_{1(t_1+k)} = 0 \tag{16}$$

$$x_{2(t_1+k)} = -\pi_k \tag{17}$$

$$x_{3(t_1+k)} = 1 - \sum_{j=1}^{k} \pi_j \tag{18}$$

$$x_{4(t_1+k)} = \delta^{k-j} - \sum_{j=1}^{k-1} \delta^{k-j}\pi_j - \pi_k \tag{19}$$

Using the above formulas allows us obtain the least squares estimate for the effect of an outlier, ω, at $t = t_1$. These are provided below.

$$\hat{\omega}_{\text{AO}}(t_1) = \frac{\sum_{t=t_1}^{n} \hat{e}_t x_{2t}}{\sum_{t=t_1}^{n} x_{2t}^2} \tag{20}$$

$$\hat{\omega}_{\text{IO}}(t_1) = \hat{e}_{t_1} \tag{21}$$

$$\hat{\omega}_{\text{LS}}(t_1) = \frac{\sum_{t=t_1}^{n} \hat{e}_t x_{3t}}{\sum_{t=t_1}^{n} x_{3t}^2} \tag{22}$$

$$\hat{\omega}_{\text{TC}}(t_1) = \frac{\sum_{t=t_1}^{n} \hat{e}_t x_{4t}}{\sum_{t=t_1}^{n} x_{4t}^2} \tag{23}$$

$$\tag{24}$$

It is interesting to note that when $t_1 = n$ (last observation in a time series) it is impossible to distinguish what type the outlier is, as $x_{it_1} = 1$ for all i, and this is the only value in each sum term above. Next, one can standardize these estimated values to obtain asymptotically normal test statistics of the outlier effects.

$$\hat{\tau}_{\text{AO}}(t_1) = \frac{\hat{\omega}_{\text{AO}}(t_1)\left(\sum_{t=t_1}^{n} x_{2t}^2\right)^{1/2}}{\hat{\sigma}_a} \tag{25}$$

$$\hat{\tau}_{\text{IO}}(t_1) = \frac{\hat{\omega}_{\text{IO}}(t_1)}{\hat{\sigma}_a} \tag{26}$$

$$\hat{\tau}_{\text{LS}}(t_1) = \frac{\hat{\omega}_{\text{LS}}(t_1)\left(\sum_{t=t_1}^{n} x_{3t}^2\right)^{1/2}}{\hat{\sigma}_a} \tag{27}$$

$$\hat{\tau}_{\text{TC}}(t_1) = \frac{\hat{\omega}_{\text{TC}}(t_1)\left(\sum_{t=t_1}^{n} x_{4t}^2\right)^{1/2}}{\hat{\sigma}_a} \tag{28}$$

$$\tag{29}$$

where $\hat{\sigma}_a$ is an estimate of the standard deviation. We can compare these test statistics to a standard normal distribution to find appropriate p-values when testing $H_0 : \omega = 0$ vs $H_1 : \omega \neq 0$.

These results help us decide whether an unusual observation is actually an outlier or is due to natural, random variation. Once we have deemed an observation an outlier, it is possible to adjust a time series model to account for this information and improve future predictions.

Effects of Multiple Outliers

Estimation of outlier effects and calculation of statistics as described above becomes complicated when multiple outliers are present. When it is assumed (as above) that only one outlier is present, but in fact there are many, the presence of the other outliers may bias the estimate of the effect of one at the time t under consideration. Or the effects of an outlier at t may be masked by the effects of an earlier outlier.

The problem of bias and masking suggests that the effects of all of the outliers should be estimated jointly. But doing so when the number and time locations of the outliers are unknown would require much more complex calculations. The procedure described below detects outliers one at a time but attempts to solve the bias and masking problem by repeatedly adjusting and re-examining both the time series and the set of identified outliers.

Estimating σ_a

The statistics described above require an estimate of σ_a. Since we are assuming the presence of outliers which could bias the estimate upwards, it is important to use an estimate of σ_a that is robust to outliers. Three options are the median absolute deviation (MAD), a trimmed standard deviation, and the usual standard deviation of residuals calculated with the residual at time t_1 omitted.

Detection and Estimation Procedure

Chen and Liu (1993) proposed an iterative procedure for detecting outliers, estimating their effects, and estimating the parameters of an ARIMA model to fit the series after adjusting for outlier effects. The procedure assumes that the order of the ARIMA model is known. We will see that implementation of this procedure in `tsoutliers` can also automatically choose the order of the ARIMA model.

The procedure has three stages. In Stage I, outliers are accumulated one-by-one in order of descending magnitude of $\hat{\tau}$ statistic. In Stage II, any outliers that can no longer be considered outliers after taking into account the effects of all the others are dropped. In Stage III, the model estimates are fixed at their final values and one last round of outlier accumulation-elimination results in the final set of outliers and estimates of their effects.

Stage I

1. Compute the maximum likelihood estimates of the model parameters.
2. For $t = 1, \ldots, n$ calculate $\hat{\tau}_{IO}(t), \hat{\tau}_{AO}(t), \hat{\tau}_{LS}(t)$, and $\hat{\tau}_{TC}(t)$. Let $\hat{\tau}_{max}$ be the maximum of the absolute values of the $\hat{\tau}$ statistics. If $\hat{\tau}_{max} > C$, where C is a constant chosen beforehand, there is potentially an outlier at the time and of the type for which $\hat{\tau}_{max}$ occurred. Remove the effect of the outlier from the series using equations above.
3. Repeat (2) until no more $\hat{\tau}$ statistics indicate potential outliers.
4. If any outliers have been found in (2–3), go back to (1), computing the parameter estimates on the series that has now been adjusted for outlier effects, and iterate through (2–3) again. If no outliers were found, proceed to Stage II.

Stage II

5. Jointly estimate the outlier effects using the same multiple regression model above. Calculate the $\hat{\tau}$ statistics. If any of the statistics no longer exceed the threshold C, drop the corresponding outlier.
6. Repeat (5) until no more outliers are dropped.
7. Obtain jointly estimated outlier effects using only the remaining outliers and adjust the original series by removing their effects.
8. Compute the maximum likelihood estimates of the model parameters using the adjusted series. If the change in residual standard error from the previous estimate exceeds ϵ, another threshold constant chosen beforehand, go back to (5), using the new parameter estimates. Otherwise, proceed to Stage III.

Stage III

9. The parameter estimates found in (8) are the final estimates for the model. Using only these model parameters, iterate through (2–3) and (5–6) to obtain a final set of outliers and estimates of their effects.

Comments

The procedure uses two threshold values, C and ϵ, which are chosen by the user. Based on simulations, Chen and Liu (1993) recommend $C = 2.5$ to 2.9 for a shorter series of length less than $n = 100$ or so, and $C = 3.0$ for $n = 100$ to 200. For longer series, $n > 200$, C greater than 3.0 may be appropriate, but in any case multiple values should be tried to assess how sensitive the results are to C. The value of ϵ controls the accuracy of the parameter estimates, and a suggested value is 0.001.

Examples

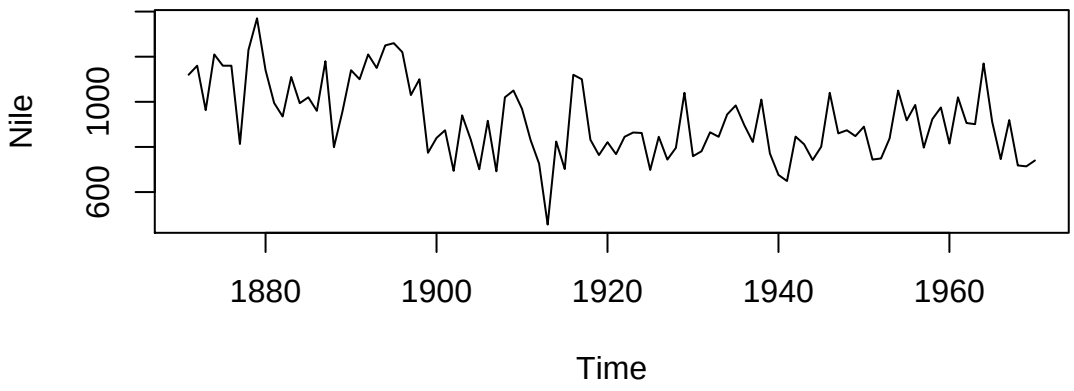
[*Possible questions to address, in addition to showing one or two examples:*]

- What functions does the `tsoutliers` package offer? What exactly do they do? In particular, how exactly is `auto.arima` used within `tso`?
- What arguments need to be specified? What are the defaults? When should one provide arguments other than the defaults?
- How does one decide which types of outliers to look for?
- How does one choose C and δ?
- What is the output of the R function(s) and how does one interpret it?

Outlier in measurement of the annual flow of the river Nile at Ashwan from 1871-1970.

Fitting auto.arima model gives an ARIMA(1,1,1).

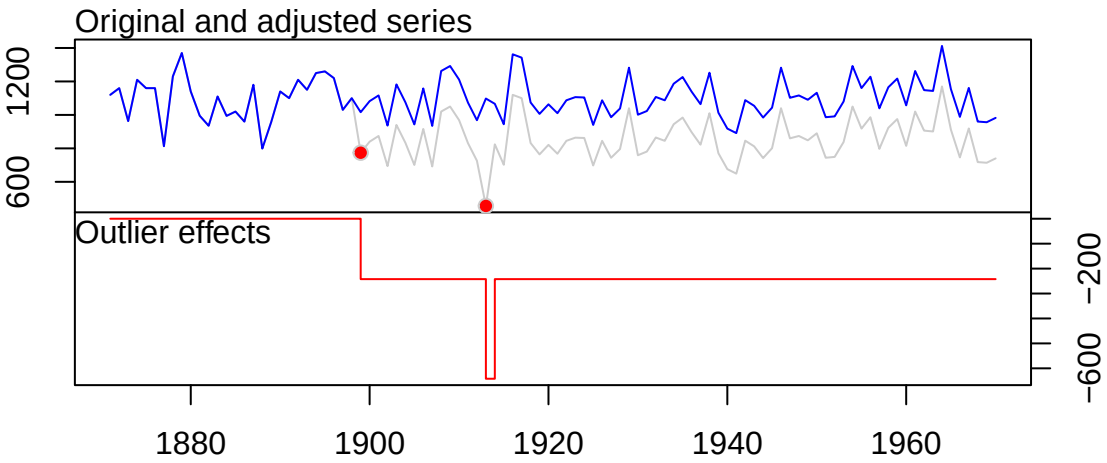
```
library(tsoutliers)
library(fma)

data(Nile)
plot(Nile)
```




```
fit<- auto.arima(Nile)
#fits with ARIMA(1,1,1)
print(fit)
```

Applying the two stage outlier detection process using `tsoutliers` function generates

```
nile.outliers <- tso(Nile,types=c("AO","LS","TC"))
nile.outliers
plot(nile.outliers)
```



Additionally the `tso` suggests an ARIMA order (0,0,0).

Example 2. Outlier from yahoo dataset

Conclusion

References

Chang, I., Tiao, G. C., and Chen, C. (1988), "Estimation of Time Series Parameters in the Presence of Outliers," *Technometrics*, 30, 193–204.

Chen, C. and Liu, Lon-Mu (1993), "Joint Estimation of Model Parameters and Outlier Effects in Time Series," *Journal of the American Statistical Association*, 88, 284–297.

Chen, C., and Tiao, G. C. (1990), "Random Level Shift Time Series Models, ARIMA Approximation, and Level Shift Detection," *Journal of Business and Economic Statistics*, 8, 170–186.

Fox, A. J. (1972), "Outliers in Time Series," *Journal of the Royal Statistical Society*, Ser. B, 34, 350–363.

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

Appendix

```
knitr::opts_chunk$set(echo = FALSE, message = FALSE, warning = FALSE,
           results = "hide", fig.height = 3, fig.width = 6)
library('tsoutliers')
library('ggplot2')
#simulation time series data
```

```r
set.seed(9)
# Arma model ARMA(1,1)
wn <- arima.sim(model=list(ar=c(0.5,0.3),ma=c(0.5,0.0),sd=1),200)
#add additive outlier with weight of 10 times to current value at two section 20 and 1
wn[20]<- 5*wn[20]
wn[170] <- 6*wn[170]
plot(1:200,wn,type='l',main="",xlab="Interval",ylab="")
lines(18:21,wn[18:21],col='red')
lines(169:171,wn[169:171],col='red')
#Example of temporal shift outliers
#source http://stats.stackexchange.com/users/48766/javlacalle
tc <- rep(0, 50)
tc[20] <- 1
tc1 <- filter(tc, filter = 0, method = "recursive")
tc2 <- filter(tc, filter = 0.3, method = "recursive")
tc3 <- filter(tc, filter = 0.7, method = "recursive")
tc4 <- filter(tc, filter = 1, method = "recursive")
par(mfrow = c(2,2))
plot(tc1, main = "TC delta = 0")
plot(tc2, main = "TC delta = 0.3")
plot(tc3, main = "TC delta = 0.7")
plot(tc4, main = "TC delta = 1", type = "s")
#dev.off()
library(tsoutliers)
library(fma)

data(Nile)
plot(Nile)
fit<- auto.arima(Nile)
#fits with ARIMA(1,1,1)
print(fit)
nile.outliers <- tso(Nile,types=c("AO","LS","TC"))
nile.outliers
plot(nile.outliers)
```