Outlier Detection in Time Series

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 effected 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 prescene 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.

Throughout this report, we will look at four different types of outliers: additive outliers (AO), innovational outliers (IO), 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 in *Joint Estimation of Model Parameters and Outlier Effects in Time Series* by Chung Chen and Lon-Mu 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 Eurpoean countries from 1999 to 2013. After successful detection, we will also begin exploring ways to incoorporate this information into our model building process to obtain reliable parameter estimates and predictions.

Types of outlier in ARIMA model

We say that X_t folows 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

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

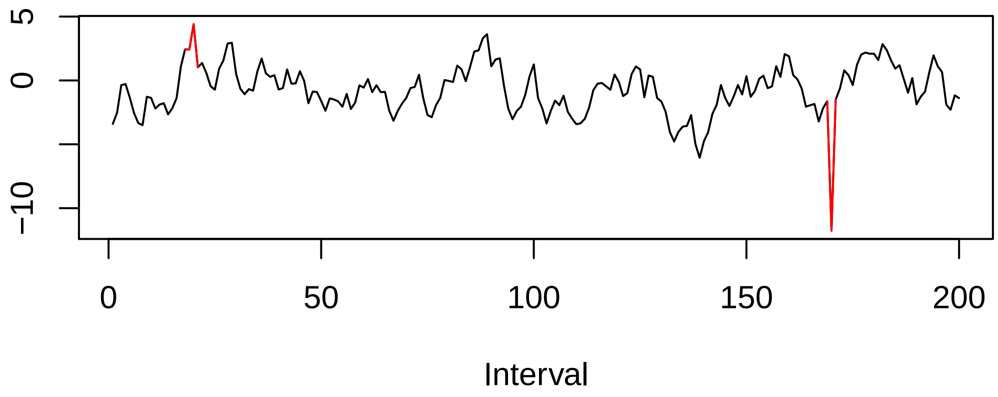
where $I_t(t_1) = 1$ if t = t_1, and 0 otherwise. $I_t(t_1)$ acts as an indicator function for the outlier impact occurence, with t_1 being the location (possibly unknown) of the outlier. $\omega \frac{A(B)}{G(B)H(B)}$ denotes magnitude and pattern of some event. We will assume these are not previously 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).

Additive outliers (AO)

An additive outlier (AO) correspond to an exogenous change of a single observation of the time series. It is associated with inciden like measurement errors or impulse effect due to external effect. A time series $y_1 .... y_T$ affected by AO at t=k is given by

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

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



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 $I_t(t_1) = 0$.

Innovational Outliers (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 be 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

$$X_t^* = (X_t + \omega I_t(t_1))\frac{\phi(B)}{\alpha(B)\omega(B)} \tag{4}$$

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

$$\psi(B) = \frac{\theta(B)}{\alpha(B)\phi(B)} = \psi_0 + \psi_1 B + \psi_2 B^2 + ... + \psi_{max(p,d,q)} B^{max(p,q,d)}, \text{where } \psi_0 = 1 \tag{5}$$

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.\

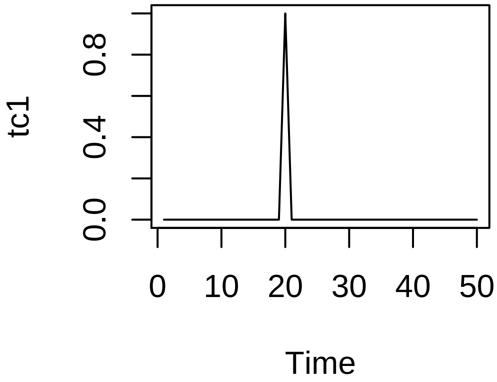
Temporal Change outliers (TC)

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

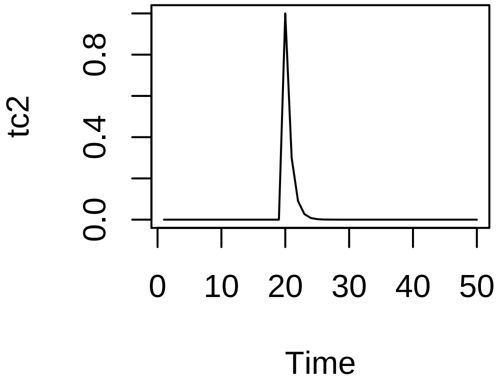
$$X_t^* = X_t + \frac{\omega I_t(t_1)}{(1 - \delta B)} \tag{6}$$

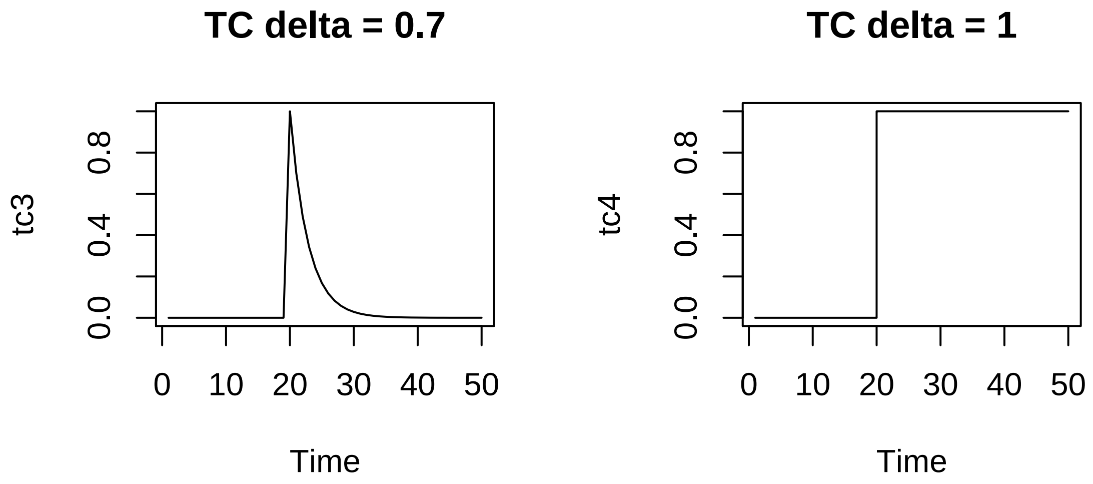
We can see that $\frac{A(B)}{G(B)H(B)} = \frac{1}{(1-\delta B)}$ and that these outliers are **independent** of the model chosen.

TC delta = 0



TC delta = 0.3





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.

Level Shifts Outliers (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

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

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.

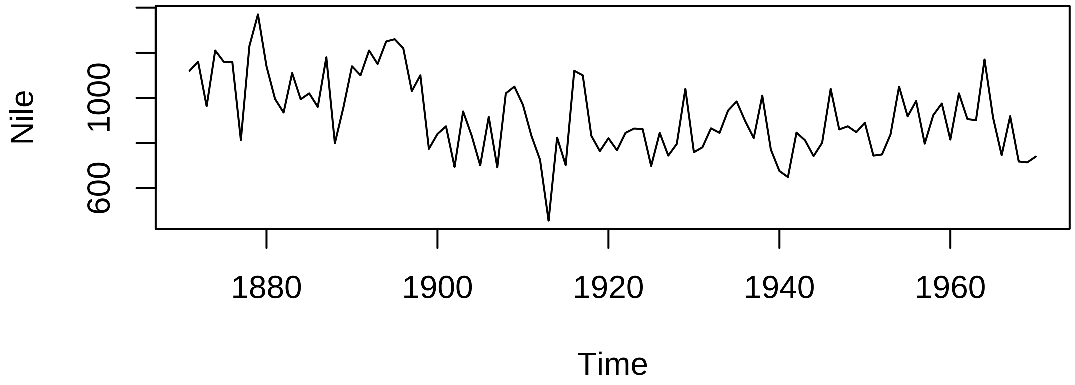
OUTLIER DETECTION AND ESTIMATION PROCEDURE

EXPERIMENTS

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).

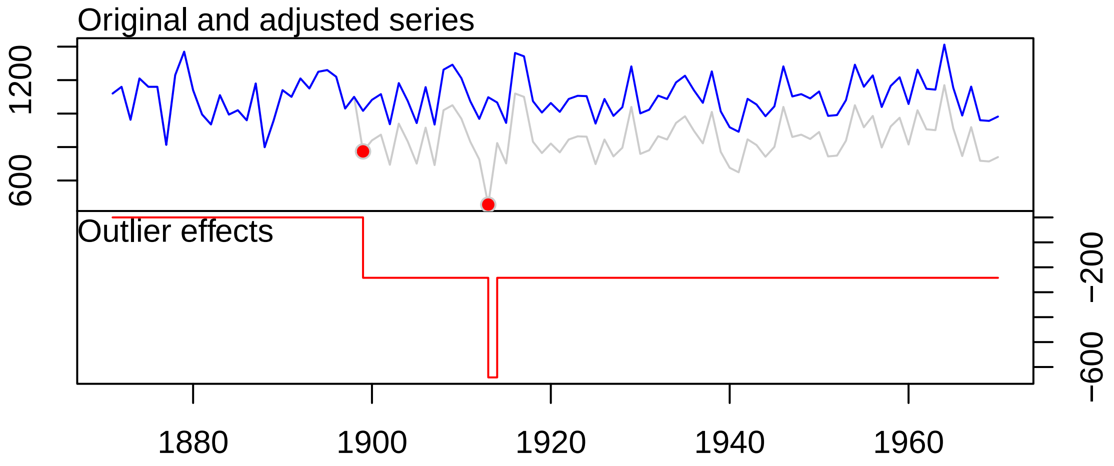
```
library(tsoutliers)
library(fma)

data(Nile)
plot(Nile)
```



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

Applying the two stage outleir detection process using `tsoutliers` function generates.

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



Additionaly the `tso` suggests an ARIMA order (0,0,0)

Example 2. Outlier from yahoo dataset

CONCULSION

APPENDIX

```r
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
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 autlier 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)
```