Outliers in Time Series

ST 565 Final Project (Winter 2016)

Amirhosein "Emerson" Azarbakht, Michael Dumelle,
Camden Lopez & Tadesse Zemicheal

Introduction

Data sets often contain observations that are far from other observations. These outliers may be due to measurement error, or they may truly represent anomalous or extreme behavior in the observed phenomenon. Outliers can have a large influence in many statistical analyses, and since we prefer the results of an analysis not to depend on a few unusual observations, it is necessary to develop methods for avoiding or removing outlier effects.

With time series data, because of the time-dependent correlation structure among the observations, outliers present unique problems and have to be dealt with differently from outliers in i.i.d. observations. Suppose we want to model a process as ARIMA(p, d, q). The data may contain measurement errors or may have been affected by some unusual event(s). We would like a way to detect outliers, which we might investigate more closely to determine their origin, and adjust the series so that we can fit a model on what we believe the series would have been without the measurement errors or unusual events.

We will consider four simple models for time series outliers and how they can be detected by estimating their hypothetical effect at each time point using linear regression. Large estimated effect at a given time point suggests an outlier at that point. The estimate of outlier effect also allows us to remove the effect and proceed to model the adjusted series. We will look at a procedure developed by Chen and Liu (1993) to detect and estimate outlier effects while also fitting an ARIMA(p, d, q) model. The R package `tsoutliers` conveniently implements this procedure, and we will see examples of applying it to real time series.

Outlier Models

The following models describe four different ways outliers can enter into an ARIMA(p, d, q) process X_t. For simplicity, we consider only non-seasonal processes, though the models can be extended for seasonal ARIMA processes. X_t is described by

$$X_t = \frac{\theta(B)}{\alpha(B)\phi(B)} Z_t \tag{1}$$

where $\theta(B)$ and $\phi(B)$ have roots strictly outside the unit circle, $\alpha(B) = (1 - B)^d$, and $Z_t \sim_{iid}$ Normal$(0, \sigma^2)$. Now let X_t^* denote the observed series, which is contaminated by outliers. The outlier effect on the series is $X_t^* - X_t$.

Consider two broad types of outlier effects. Either the effect is independent of the underlying X_t process, or the effect gets incorporated into the process and is propagated through the "memory" described by $\theta(B), \phi(B)$ and $\alpha(B)$. Among the first type, the effect can be a simple impulse at one time point—this is referred to as an additive outlier (AO)—or the effect can decay over time—a temporary change (TC)—or remain permanently—a level shift (LS). The second type, where the effect gets propagated by the memory of the process, is basically equivalent to an impulse applied to the "innovation" Z_t. More complicated models are possible, but these four in combination can cover a large variety of cases.

Additive Outlier (AO)

An additive outlier corresponds to an impulse on the series at one point, where the impulse is external to the X_t process. Large measurement errors would clearly be additive outliers; the "impulse" would not represent anything actually added to the X_t process in that case.

The observed series given an AO at time $t = t_1$ is

$$X_t^* = X_t + \omega I_t(t_1) \tag{2}$$

So the outlier effect is simply

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

or ω at $t = t_1$ and zero elsewhere. The effect does not depend on the ARIMA process and has no lingering effects.

Level Shift (LS)

A level shift is a sudden, persisent change in the mean of the time series, or an external impulse that carries on into the future without diminishing. An example might be a change in some performance measure for a company when a new regulation goes into effect.

The observed series when a LS is present can be intuitively expressed as

$$X_t^* = X_t + \omega I_t(t \geq t_1) \tag{4}$$

where $I_t(t \geq t_1) = 1$ for $t \geq t_1$, 0 otherwise. But an alternative expression which will be more convenient later is

$$X_t^* = X_t + \frac{1}{1-B}\omega I_t(t_1) \tag{5}$$

The second form is easier to understand when rearranged to make the outlier effect clear:

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

The application of $(1 - B)$ carries the effect over to all successive points in the series. Note that again the effect does not depend on the ARIMA process.

Temporary Change (TC)

A temporary change is 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.

In this case the observed series is

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

where $0 \leq \delta \leq 1$. Again, this formulation is clearer when rearranged:

$$(1 - \delta B)(X_t^* - X_t) = \omega I_t(t_1) \tag{8}$$

The outlier effect gets carried over but with some decay controlled by δ. Notice that the temporary change outlier is a generalization of the AO and LS types. An AO is equivalent to a TC with $\delta = 0$, and a LS is a TC with $\delta = 1$. The value of δ used to characterize TC outliers in a particular setting must be chosen by the researcher. A suggested value is $\delta = 0.7$.

Innovational Outlier (IO)

The innovation (or "white noise") Z_t can be thought of as combining the effects of a large number of independent, unmeasured variables (plus, possibly, routine measurement error due to limited precision—distinct from the gross error that could result in an AO). Suppose that something causes an impulse of ω to be added to the innovation at time t_1. (An example might be an earthquake creating an impulse in seismic activity.) Then X_t^* becomes

$$X_t^* = \frac{\theta(B)}{\alpha(B)\phi(B)}\left[Z_t + \omega I_{t_1}(t)\right] \tag{9}$$

where $I_{t_1}(t) = 1$ for $t = t_1$ and 0 otherwise. Rearranging and substituting in X_t, we see that the outlier effect is

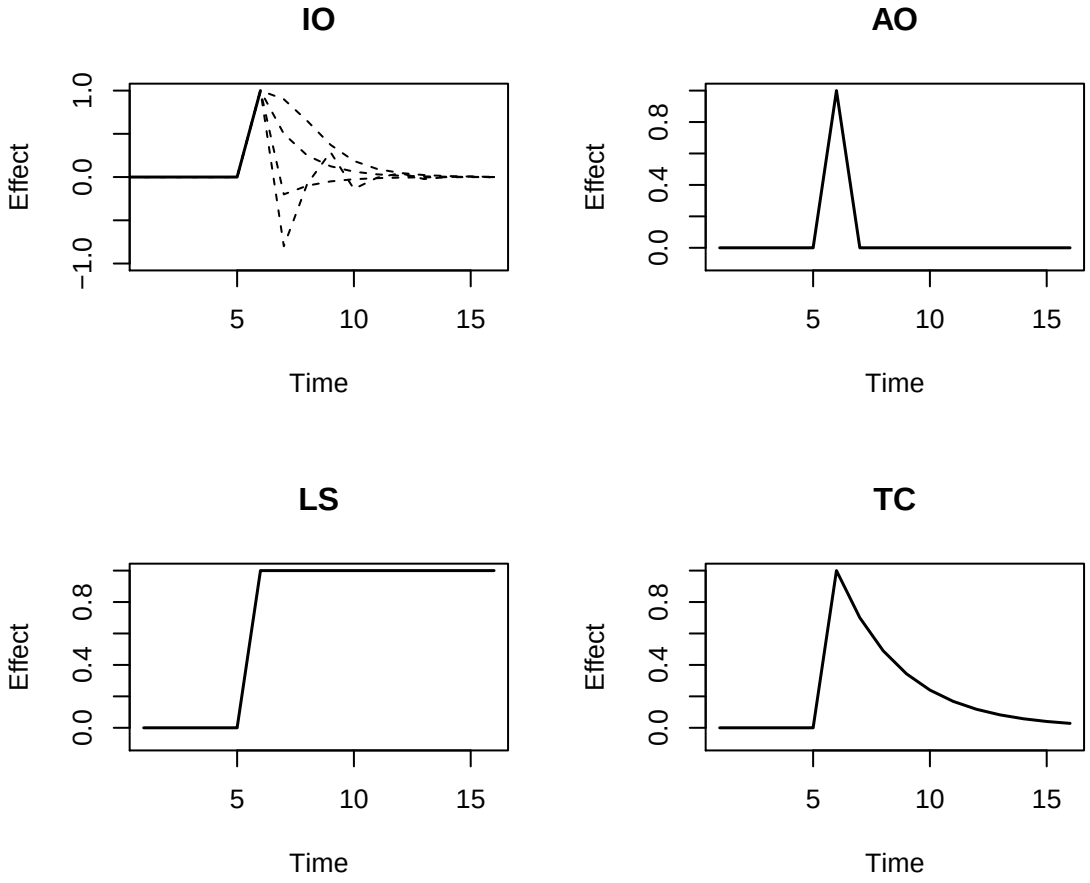
$$X_t^* - X_t = \frac{\theta(B)}{\alpha(B)\phi(B)}\omega I_{t_1}(t) \tag{10}$$

Thus, the outlier effect depends on the ARIMA model. When an IO occurs at $t = t_1$, the effect of this outlier at $t = t_1 + k$, $k \geq 0$, is

$$\begin{aligned} X_{t_1+k}^* - X_{t_1+k} &= \frac{\theta(B)}{\alpha(B)\phi(B)}\omega I_{t_1}(t_1 + k) && (11)\\ &= \psi(B)\omega I_{t_1}(t_1 + k) && (12)\\ &= \psi_k \omega && (13) \end{aligned}$$

where $\psi(B) = \frac{\theta(B)}{\alpha(B)\phi(B)}$ and ψ_k is the k^{th} coefficient of $\psi(B)$, corresponding to the B^k term. If the time series is stationary, the ψ_k coefficients tend to zero as k increases. This implies that the IO effect in a stationary series diminishes and becomes zero given enough time. In a non-stationary series, an IO can cause a permanent level shift.

The various outlier effects are summarized by the following figure. The IO effect is shown with four different stationary ARMA models to emphasize that the effect depends on the model.



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 linear regression Then we point out some practical issues, and the section culminates in a description of the procedure proposed by Chen and Liu to jointly estimate model parameters and outlier effects in a series.

Estimating the Effect of a Single Outlier

Suppose we have an observed time series X_t^* that arose from an ARIMA(p, d, q) process X_t with one outlier at time t_1. For now, assume we know the ARIMA model parameters, t_1, and the type of outlier. We will find an estimate of the outlier magnitude ω. (Later we discuss how this calculation is useful when the locations and types of outliers are unknown.)

$$\pi(B) = \frac{\alpha(B)\phi(B)}{\theta(B)} = 1 - \pi_1 B - \pi_2 B^2 - \pi_3 B^3 - ... \tag{14}$$

Since $X_t = \frac{\theta(B)}{\alpha(B)\phi(B)} Z_t$, applying $\pi(B)$ to X_t would result in the purely random residuals, Z_t. But X_t^* is contaminated by the outlier, so the residuals that result from applying $\pi(B)$ to X_t^* are also contaminated. Let $\hat{e}_t = \pi(B)X_t^*$ be those contaminated residuals. For the different outlier types, we have

$$\text{IO: } \hat{e}_t = \omega I_t(t_1) + Z_t \tag{15}$$

$$\text{AO: } \hat{e}_t = \omega\pi(B) I_t(t_1) + Z_t \tag{16}$$

$$\text{LS: } \hat{e}_t = \omega \frac{\pi(B)}{1 - B} I_t(t_1) + Z_t \tag{17}$$

$$\text{TC: } \hat{e}_t = \omega \frac{\pi(B)}{1 - \delta B} I_t(t_1) + Z_t \tag{18}$$

These equations can be put in the form of a linear regression where the responses are the residuals $\hat{e}_t$ and the predictors are different for each type of outlier. The outlier magnitude ω is the coefficient to be estimated:

$$\hat{e}_t = \omega x_t + Z_t \tag{19}$$

For all outlier types, $x_t = 0$ for $t < t_1$ and $x_t = 1$ for $t = t_1$. For $t = t_1 + k$, $k \geq 1$, the value of x_t depends on the outlier type:

$$\text{IO: } x_{t_1+k} = 0 \tag{20}$$

$$\text{AO: } x_{t_1+k} = -\pi_k \tag{21}$$

$$\text{LS: } x_{t_1+k} = 1 - \sum_{j=1}^{k} \pi_j \tag{22}$$

$$\text{TC: } x_{t_1+k} = \delta^k - \sum_{j=1}^{k-1} \delta^{k-j}\pi_j - \pi_k \tag{23}$$

The least-squares estimate for ω can then be calculated for each type of outlier, using the appropriate x_t values:

$$\hat{\omega} = \frac{\sum_{t=t_1}^{n} \hat{e}_t x_t}{\sum_{t=t_1}^{n} x_t^2} \tag{24}$$

Note that when $t_1 = n$ (the last observation in the time series), it is impossible to distinguish the different types of outliers, as $x_{t_1} = 1$ for all types, and this is the only predictor value available.

Once we have an estimate of ω corresponding to a particular type of outlier at $t = t_1$, we can use equations (2, 7, 10, 12) to remove the outlier effect and obtain and adjusted series which estimates X_t, the uncontaminated process.

Using the Single-Outlier Estimate

For the estimation of ω, we assumed the ARIMA model parameters and the outlier type and location t_1 were known. In practice, they are not. The model parameters must first be estimated, then $\hat{\omega}$ can be calculated for each type of outlier at each $t_1 = 1, \ldots, n$. If $\hat{\omega}$ is large at some t_1, and largest at t_1 for one type of outlier, it might be justified to say there is an outlier of that type at t_1. This raises two issues.

First, estimates for the model parameters will be influenced by outliers. Eventually we want parameter estimates uninfluenced by outliers, but we have to start somehow. Iterating repeatedly through estimation of model parameters, detection of outliers, and adjustment of the series to remove outlier effects seems a reasonable strategy. This is what procedure described below does.

The other issue is that the $\hat{\omega}$ values are calculated using different sets of predictors and responses in the regression, so they cannot be compared immediately. Dividing $\hat{\omega}$ by its standard error results in standardized statistics which are approximately normally distributed:

$$\hat{\tau} = \frac{\hat{\omega}}{\hat{\sigma}/\sqrt{\sum_{t=t_1}^{n} x_t^2}} \tag{25}$$

where $\hat{\sigma}$ is an estimate of $\sqrt{\text{Var}(Z_t)}$ (see comment below). The $\hat{\tau}$ statistics allow for comparison among various outlier types at various times.

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 under consideration. Or the effects of an outlier 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

The $\hat{\tau}$ statistic requires an estimate of σ. Since we are assuming the presence of outliers which could bias the estimate upwards, it is important to use an estimate of σ 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 ARMA model to fit the series after adjusting for outlier effects. The procedure assumes that the order of the ARMA model is known. We will see that implementation of this procedure in `tsoutliers` can also automatically choose the order of the ARMA 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 $\hat{\tau}$ for each type of outlier. 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.
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 a multiple regression based on (24). 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

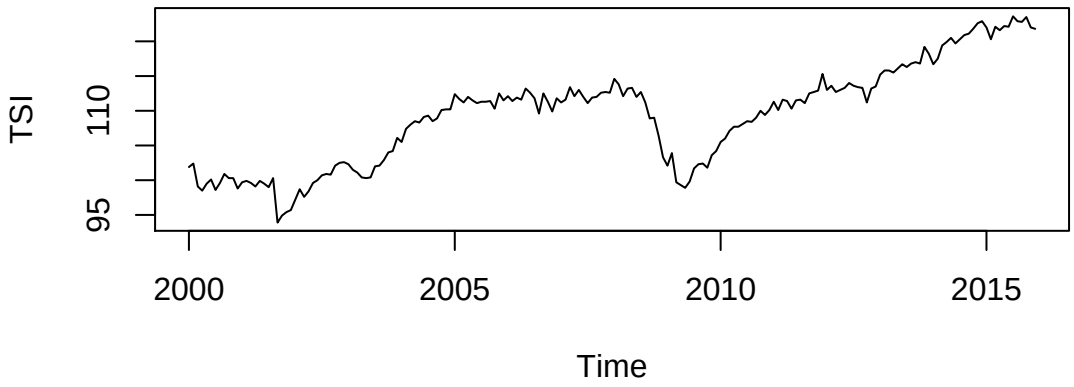
The procedure uses two threshold values, C and ϵ, which are chosen by the user. Based on simulations, Chen and Liu recommend $C = 2.5$ to 2.9 for a shorter series of length less than $n = 100$ or so, and $C = 3.0$ for $n = 100$ to 200. For longer series, $n > 200$, C greater than 3.0 may be appropriate, 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.

Application

Chen and Liu's procedure for outlier detection/estimation is implemented in the R package `tsoutliers`. The function `tso` is the main interface for the procedure. It takes as input a time series object and fits a seasonal ARIMA (or simpler) model while detecting and estimating the effects of outliers. It can look for AO, LS, TC, and IO outliers, plus seasonal level shifts (SLS). The `tso` arguments allow one to specify parameters, including C and δ, and the types of outliers to consider. Another argument can specify the method for fitting a (seasonal) ARIMA model. As mentioned earlier, one option is to have the model automatically selected using, for example, `auto.arima`.

Transportation Services Index

Now we apply the `tso` function to data from the US Bureau of Transportation Statistics. The time series is the Transportation Services Index (TSI), a monthly measure of the volume of services provided by the for-hire transportation sector, including both freight and passenger carriers.



When `tso` is applied with `tsmethod = "auto.arima"`, the suggested model has a seasonal AR(1) component, but the coefficient is insignificant. So we force `tso` to fit only the non-seasonal ARIMA(1, 1, 0) model using the `tsmethod` and `args.tsmethod` arguments.

```
library(tsoutliers)
result <- tso(TSI, types = c("AO", "LS", "TC", "IO"), tsmethod = "arima",
              args.tsmethod = list(order = c(1, 1, 0)))
result$outliers
```

```
##   type ind    time   coefhat     tstat
## 1   TC  21 2001:09 -5.889364 -5.928143
## 2   LS 108 2008:12 -3.884195 -3.633127
```

Two outliers are detected, a temporary change starting in September 2011, and a level shift in December 2008. These outliers have clear interpretations (which is not often the case, we have found): the terrorist attacks on 9/11 and the economic downturn of 2008 both severely impacted transportation services.

Comparing the `tso` model fit to what the fit would have been without adjusting for outliers, we find that accounting for outliers results in a larger (absolute value) AR(1) coefficient and, understandably, a smaller estimated σ^2:

```
fit <- arima(TSI, order = c(1, 1, 0))
rbind(result$fit$coef, c(fit$coef, NA, NA))
```

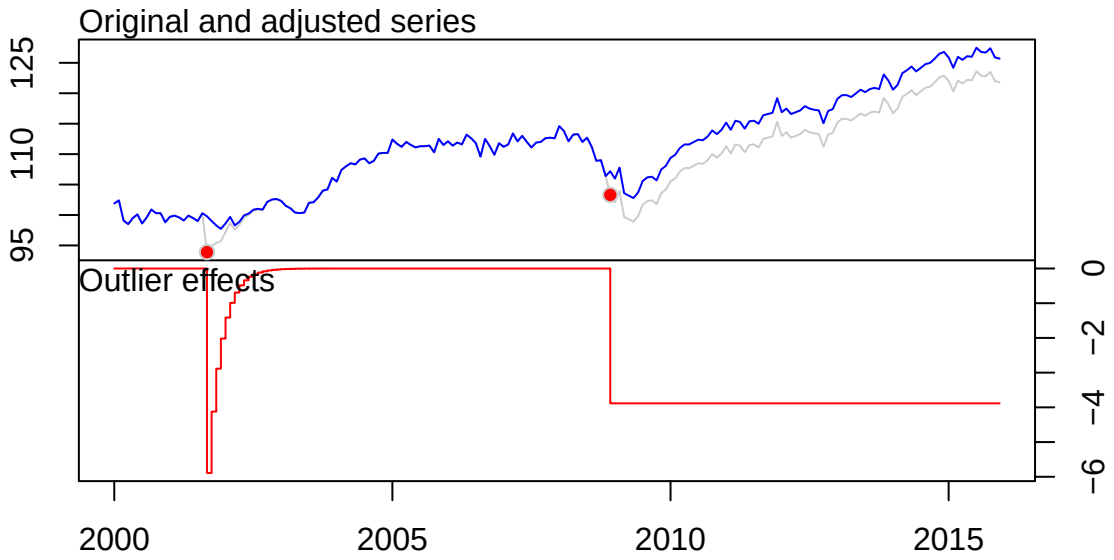
```
##             ar1      TC21     LS108
## [1,] -0.2159957 -5.889364 -3.884195
## [2,] -0.1659520        NA        NA
```

```
rbind(result$fit$sigma2, fit$sigma2)
```

```
##           [,1]
## [1,] 1.136264
## [2,] 1.425164
```

Finally, one can plot the original and adjusted series together with the outlier effects:

```
plot(result)
```



One can obtain the outlier effects at each time point from the `effects` component of the `tso` output. From there, you can obtain the outlier-adjusted time series by subtracting these effects from the original time series.

```
adjusted <- TSI - result$effects
```

Conclusion

By using the methods described above, we can identify and estimate each of the four types of outlier effects. This enables us to build a time series model free of outliers and construct accurate outlier-free forecasts. The `tsoutliers` R package allows us to easily obtain these outlier effect estimates and produce clear graphs displaying time series models contaminated with outliers versus outlier-free models.

References

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.

Galeano, Pena (2013), "Finding Outliers in Linear and Nonlinear Time Series", *Robustness and Complex Data Structures*, Springer, 243–260.

Lopez-de-Lacalle, Javier (2015), "Package 'tsoutliers' ", CRAN documentation, https://cran.r-project.org/web/packages/tsoutliers/tsoutliers.pdf.

Wei, William S. (1990), "Intervention Analysis and Outlier Detection", *Time Series Analysis*, Addison-Wesley, 184–203.

Appendix

```r
# Calculate effect of IO given AR and MA coefficients, out to given lag
innov_outlier_effect <- function (ar, ma, lag) {
  ar <- ifelse(rep(is.null(ar), lag), rep(0, lag),
               c(ar, rep(0, lag - length(ar))))
  ma <- ifelse(rep(is.null(ma), lag), rep(0, lag),
               c(ma, rep(0, lag - length(ma))))
  effect <- numeric(lag)
  effect[1] <- 1
  for (k in 1:(lag - 1)) {
    effect[1 + k] <- ma[k]
    for (j in 1:k) {
      effect[1 + k] <- effect[1 + k] + ar[j] * effect[1 + k - j]
    }
  }
  effect
}
# Calculate and plot effects of each outlier type
par(mfrow = c(2, 2))
n <- 16
t <- 1:n
t1 <- 6 # Outlier here
# IO
ar <- c(-0.4, -0.2, 0.048)
ma <- c(-0.4, -0.2, 0.048)
io1 <- c(rep(0, t1 - 1), innov_outlier_effect(ar, ma, n - t1 + 1))
plot(io1, type = 'l', ylim = c(-1, 1), lty = 2,
     xlab = "Time", ylab = "Effect", main = "IO")
ar <- 0.5
```

```r
ma <- c(0.4, 0.2, 0.048)
io2 <- c(rep(0, t1 - 1), innov_outlier_effect(ar, ma, n - t1 + 1))
lines(io2, lty = 2)
ma <- -0.7
io3 <- c(rep(0, t1 - 1), innov_outlier_effect(ar, ma, n - t1 + 1))
lines(io3, lty = 2)
ma <- NULL
io4 <- c(rep(0, t1 - 1), innov_outlier_effect(ar, ma, n - t1 + 1))
lines(io4, lty = 2)
lines(c(0, t1 - 1, t1), c(0, 0, 1), lwd = 1.5)
# AO
ao <- rep(0, n)
ao[t1] <- 1
plot(ao, type = 'l', ylim = c(-0.1, 1), lwd = 1.5,
     xlab = "Time", ylab = "Effect", main = "AO")
# LS
ls <- rep(0, n)
ls[t1:n] <- 1
plot(ls, type = 'l', ylim = c(-0.1, 1), lwd = 1.5,
     xlab = "Time", ylab = "Effect", main = "LS")
# TC
tc <- rep(0, n)
tc[t1] <- 1
delta <- 0.7
for (k in (t1 + 1):n) {
  tc[k] <- delta * tc[k - 1]
}
plot(tc, type = 'l', ylim = c(-0.1, 1), lwd = 1.5,
     xlab = "Time", ylab = "Effect", main = "TC")

# Application of tsoutliers::tso to Transportation Services Index
TSI <- read.csv("../Data/transport.csv", skip = 10)
TSI <- ts(TSI$TSITTL, start = c(2000, 1), frequency = 12)
plot(TSI)
library(tsoutliers)
result <- tso(TSI, types = c("AO", "LS", "TC", "IO"), tsmethod = "arima",
              args.tsmethod = list(order = c(1, 1, 0)))
result$outliers
fit <- arima(TSI, order = c(1, 1, 0))
rbind(result$fit$coef, c(fit$coef, NA, NA))
rbind(result$fit$sigma2, fit$sigma2)
plot(result)
adjusted <- TSI - result$effects
```