A New Clipping Approach for Robust ACF Estimation

Samuel Flimmel [1] | *University of Economics in Prague, Prague, Czech Republic*
Ivana Malá | *University of Economics in Prague, Prague, Czech Republic*
Jiří Procházka | *University of Economics in Prague, Prague, Czech Republic*
Jan Fojtík | *University of Economics in Prague, Prague, Czech Republic*

Abstract

The importance of working with sufficiently robust methods has been rising in recent years. This growth is related to the extensive usage of highly frequent data, which we currently encounter in many fields including finance. Since with an increasing number of observations, the probability of outlier presence also rises. Moreover, as it is known, standard methods are not able to work correctly with outliers and, consequently, standard estimates are often biased. We focus on estimators of autocorrelation function for univariate time series, for which we propose a method based on clipping an original time series and working with a binary time series instead. The clipping helps to deal with outliers and the estimation is not affected as much as with standard methods. We also derive an asymptotical distribution of the estimator, what gives our method a major advantage in comparison with other robust methods, which are often presented without this. Furthermore, knowing the distribution of the estimator allows us to perform statistical inference.

Keywords

ACF, robust estimation, clipping, confidence interval, time series

JEL code

C10, C22

INTRODUCTION

The autocorrelation function (ACF) expresses the correlation among observations of the time series. It plays an important role in time series theory because it partially describes the relationship among the observations of the series. Furthermore, it gives us an overview of the time series, and we can use it to investigate or to model the time series.

Estimation of the ACF can be negatively affected by many factors. One of them is an outlier presence, very relevant nowadays, when we face many problems related to the extensive usage of big data. There are several robust methods for ACF estimation designed to be able to take account of outliers. These methods should, naturally, be less sensitive to outliers and should lead to better results in general (Chan and Wei, 1992).

There has been a plethora of approaches proposed by many authors. Chakhchoukh (2010) presented a median approach, where he suggested to use the median instead of the mean in the standard estimator. Ma and Genton (2000) proposed a Gnanadesikan-Kettenring approach based on the special relationship

[1] Department of Statistics and Probability, Faculty of Informatics and Statistics, University of Economics in Prague, Nám W. Churchilla 4, Prague 3, Czech Republic. Corresponding author: e-mail: samuel.flimmel@vse.cz.

for an autocovariance function (Gnanadesikan and Kettenring, 1972), while Maronna et al. (2006) proposed a robust filtering approach. Dürre et al. (2015) presented a useful overview of robust ACF estimators, including those we do not mention specifically. However, the above mentioned estimators are almost always presented without specifying the distribution. Therefore, we are not able to test the significance of the ACF order, or to test hypotheses related to the ACF.

In this paper, we present a new approach for ACF estimation. This approach is based on clipping, i.e. the process, when we replace original observations by zeros when they are below a given threshold, resp. by ones when they are above. We do not only construct the ACF estimator, but we also derive the asymptotical distribution of the ACF estimator. Using the distribution of the estimator, we suggest an analogous approximation to Bartlett's approximation (Bartlett, 1946), which is used to determine the ACF order significance.

We apply the proposed clipping approach in a simulation study in order to compare its results with a standard sample estimator. Finally, both approaches are used in a study with real world data, where we investigate the behavior of the 1-year (1Y) historical volatility of Bitcoin logarithm of daily returns.

The methodology of the clipping approach is presented in Section 1. The distribution of the estimator is derived in Section 2. The simulation study is presented in Section 3. The real data study is presented in Section 4. The last section includes conclusions of our study.

1 METHODOLOGY

Let $\{Z_n, n \in N_0\}$ be a stationary time series. Then we can define an autocovariance function of the lag k, $k \in Z$, $R(k)$ as

$$R(k) = E(Z_k - \mu)(Z_0 - \mu), \tag{1}$$

where μ is the expected value of the process.

We define an autocorrelation function (ACF) of the lag k, $k \in Z$, $\rho(k)$ of the stationary process $\{Z_n, n \in N_0\}$ as

$$\rho(k) = \frac{R(k)}{\sigma^2}, \tag{2}$$

where σ^2 is the variance of the time series.

We define a sample autocorrelation function of the lag k, $k \in Z$, $\hat{\rho}(k)$, of the data $Z_0, ..., Z_m$ as

$$\hat{\rho}(k) = \frac{\sum_{n=k}^{m} (Z_n - \bar{Z})(Z_{n-k} - \bar{Z})}{\sum_{n=0}^{m} (Z_n - \bar{Z})^2}, \tag{3}$$

where $\bar{Z}$ is the mean of the data.

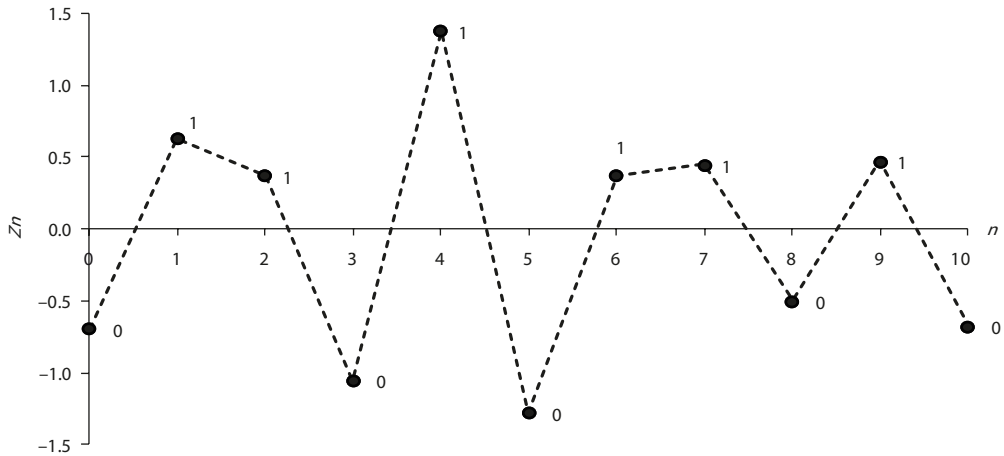
Let $\{Z_n, n \in N_0\}$ be a strictly stationary time series with autocorrelation function $\rho_Z(k)$. For a fixed h, $h \in R$, a so called threshold, we define $\{X_n, n \in N_0\}$ followingly:

$$X_{n,h} = \begin{cases} 1, Z_n \geq h, \\ 0, Z_n < h. \end{cases} \tag{4}$$

A time series $\{Z_n, n \in N_0\}$ is called original and $\{X_{n,h}, n \in N_0\}$ is known as clipped or hard-limited. Let us denote an autocorrelation function of this clipped time series by $\rho_X(k)$.

In Figure 1, we can see an illustrative example of clipping the original time series $\{Z_n, n \in N_0\}$ at the threshold $h = 0$.

Figure 1 Illustrative example of clipping at the threshold h = 0



Source: Own construction

In Kedem (1980) it was proved, under the assumption of a zero threshold ($h = 0$) and a zero mean ($\mu = 0$) strictly stationary Gaussian original time series $\{Z_n, n \in N_0\}$, that

$$\rho_X(k) = \frac{2}{\pi} \arcsin \rho_Z(k), k \in Z. \tag{5}$$

Easily, it can be rewritten into

$$\rho_Z(k) = \sin\left(\frac{\pi}{2} \rho_X(k)\right), k \in Z. \tag{6}$$

Using (6) we can construct a new robust ACF estimator $\tilde{\boldsymbol{\rho}}_{\boldsymbol{Z}}(\boldsymbol{k})$. The construction can be divided into the following steps:

1. Derive a clipped ($\mathbf{h = 0}$) time series $\{\boldsymbol{X}_{n,0}, \boldsymbol{n} \in \boldsymbol{N}_0\}$ from the original time series $\{\boldsymbol{Z}_n, \boldsymbol{n} \in \boldsymbol{N}_0\}$ which is of our interest.
2. Calculate a sample $\hat{\boldsymbol{\rho}}_{\boldsymbol{X}}(\boldsymbol{k})$ from the clipped time series $\{\boldsymbol{X}_{n,0}, \boldsymbol{n} \in \boldsymbol{N}_0\}$.
3. Calculate an estimation $\tilde{\boldsymbol{\rho}}_{\boldsymbol{Z}}(\boldsymbol{k})$ from the $\hat{\boldsymbol{\rho}}_{\boldsymbol{X}}(\boldsymbol{k})$ using Formula (6).

The clipping in step 1 helps to face outliers. It is similar to widely used trimming methods, however, the loss of information has a different nature.

2 DISTRIBUTION OF THE ESTIMATOR $\tilde{\boldsymbol{\rho}}_{\boldsymbol{Z}}(\boldsymbol{k})$

Bartlett's approximation (Bartlett, 1946) is frequently used for determination of the ACF order significance.

It can be formulated as

$$\hat{\rho}_Z(k) \overset{as.}{\sim} N\left(\rho_Z(k), \frac{v_k^Z}{m}\right), \tag{7}$$

where:

$$v_k^Z = \sum_{i=1}^{\infty} \left(\rho_Z(i+k) + \rho_Z(i-k) - 2\rho_Z(i)\rho_Z(k) \right)^2 \tag{8}$$

and m is the number of observations.

Analogously, we can formulate a similar approximation of our estimator

$$\tilde{\rho}_Z(k) \overset{as.}{\sim} N\left(\rho_Z(k), \frac{\pi^2}{4m} \cos^2\left(\frac{\pi}{2} \rho_X(k) \right) v_k^X \right), \tag{9}$$

where:

$$v_k^X = \sum_{i=1}^{\infty} \left(\rho_X(i+k) + \rho_X(i-k) - 2\rho_X(i)\rho_X(k) \right)^2 \tag{10}$$

and m is the number of observations.

We can prove Formula (9) using Bartlett's approximation and Delta method (Greene, 2003). Bartlett's approximation for the clipped time series $\{X_{n,0}, n \in N_0\}$ yields the following:

$$\hat{\rho}_X(k) \overset{as.}{\sim} N\left(\rho_X(k), \frac{v_k^X}{m} \right). \tag{11}$$

Delta method is a result concerning the asymptotic distribution of the transformed random variable in a specific situation. If there is a time series $\{Z_n, n \in N_0\}$ satisfying:

$$\sqrt{n}(Z_n - \theta) \xrightarrow{D} N(0, \sigma^2), \tag{12}$$

where θ and σ^2 are finite valued constants and $\xrightarrow{D}$ denotes convergence in distribution, then

$$\sqrt{n}\left(g(Z_n) - g(\theta) \right) \xrightarrow{D} N\left(0, \sigma^2 \left(\frac{d}{dx} g(\theta) \right)^2 \right), \tag{13}$$

for any function $g(x)$ satisfying that $\frac{d}{dx} g(\theta)$ exists and is non-zero valued.

Finally, if we set $g(x) = \sin\left(\frac{\pi}{2} x \right)$, then Delta method gives us approximation (9), because

$$\frac{d}{dx} g(x) = \frac{d}{dx} \sin\left(\frac{\pi}{2} x \right) = \frac{\pi}{2} \cos\left(\frac{\pi}{2} x \right), \tag{14}$$

$$\left(\frac{d}{dx} g\left(\rho_X(k) \right) \right)^2 = \frac{\pi^2}{4} \cos^2\left(\frac{\pi}{2} \rho_X(k) \right). \tag{15}$$

Easily, we can show equivalence (16), which we use later to prove statement (18),

$$\rho_X(k) = 0 \Leftrightarrow \rho_Z(k) = 0. \tag{16}$$

Since $\rho_Z(k) = 0$ only if $\rho_Z(k) = 0$, and $\sin\left(\frac{\pi}{2}\rho_X(k)\right) = 0$ only if . So it comes from Formulas (5) and (6).

Similarly, we use a special case of equation (10) for $\rho_X(k) = 0$ for $k > k_0$. We have:

$$v_k^X = \sum_{i=1}^{\infty}\left(\rho_X(i+k)+\rho_X(i-k)-2\rho_X(i)\rho_X(k)\right)^2 = \sum_{i=1}^{\infty}{\rho_X}^2(i-k) = \sum_{i=k-k_0}^{k+k_0}{\rho_X}^2(i-k) = \sum_{j=-k_0}^{k_0}{\rho_X}^2(j). \tag{17}$$

To determine the significant order of an ACF by our estimator, we use the following statement, which is proved by Formulas (16) and (17). If $\rho_Z(k) = 0$ for $k > k_0$, then

$$\tilde{\rho}_Z(k) \overset{as.}{\sim} N\left(0, \frac{\pi^2}{4m}\sum_{i=-k_0}^{k_0}{\rho_X}^2(i)\right), k > k_0. \tag{18}$$

So, we would look for k_0 that holds

$$\left|\tilde{\rho}_Z(k)\right| \geq u_{1-\alpha}\sqrt{\frac{\pi^2}{4m}\sum_{i=-k_0}^{k_0}\rho_X(i)}, k > k_0, \tag{19}$$

where $u_{1-\alpha}$ is a $(1 - \alpha)$% quantile of the standard Gaussian distribution and α is a significance level.

3 SIMULATION STUDY

In the presented simulation study, we compare our estimator with a standard sample estimator. We use MA(q) time series $\{Z_n, n \in N_0\}$:

$$Z_n = \varepsilon_n + \theta_1\varepsilon_{n-1} + \theta_2\varepsilon_{n-2} + \cdots + \theta_q\varepsilon_{n-q}, \tag{20}$$

where $\{\varepsilon_n, n \in N_0\}$ and $\theta_1, \theta_2, \ldots, \theta_q$ are parameters of the time series.

The simulation study was designed in the R software (R Core Team, 2015).

We run 10 000 simulations with 1 000 observations, which we contaminate with additive outliers (Fox, 1972).

In the additive outlier (AO) model, we assume that we do not observe the process of interest $\{Z_n, n \in N_0\}$ but, actually, we observe a process $\{Y_n, n \in N_0\}$ defined as

$$Y_n = Z_n + O_n, \tag{21}$$

where processes $\{Z_n, n \in N_0\}$ and $\{O_n, n \in N_0\}$ are assumed to be independent of one another.

Let $\{O_n, n \in N_0\}$ be a process with independent and identically distributed (i.i.d.) random variables that have a normal mixture distribution with a degenerate central component:

$$O_n \sim (1-\varepsilon)\delta_0 + \varepsilon N\left(\mu_O, \sigma_O^2\right), \tag{22}$$

where δ_0 is the point mass distribution located at zero, and we assume that the normal component $N\left(\mu_O, \sigma_O^2\right)$ has a variance significantly higher than the process $\{Z_n, n \in N_0\}$, $\sigma_O^2 \gg \sigma_Z^2$.

The probability of outlier occurrence is represented by ε, which is usually small. Consequently, the probability of occurrence of two outliers in a row is a much smaller ε^2, which means that the AO model generates mostly isolated outliers.

The percentage of outliers present in a single simulation is chosen randomly with a uniform distribution, i.e. $\varepsilon \sim U\left([0.00, 0.05]\right)$. We use the outlier standard deviation $\sigma_O = 10$.

The absolute values of the parameters of the MA(q) are generated randomly with a uniform distribution, i.e. $\theta_i \sim U([0.2,1.0])$, $i = 1,2, \ldots, q$. Values of θ_i being close to zero are not taken into account because they are difficult to observe. The sign of the parameters is generated randomly with Bernoulli's distribution with the probability of a success $p = 0.5$.

We divide our simulation study into 2 parts. In the first part, we work with MA(1), MA(2) and MA(3) times series, where we estimate ACF of the series and compare the methods using mean average error (MAE) criterion.

In the second part, we work with MA(3), which have theoretically significant ACF to 3rd order and we estimate the highest significant order according to the methods. In both parts we use $h = 0$ for the clipping method.

For the first part, we run 10 000 simulations for every model, so together 30 000 simulations. We obtain results summarized in Table 1.

Table 1 Comparison of the standard sample estimator and the clipping approach estimator in point estimation in the simulation study

MA(q)	ρ(k)	Standard approach	Clipping approach
MA(1)	$\rho(1)$	0.2258	0.0361
MA(2)	$\rho(1)$	0.1674	0.0368
	$\rho(2)$	0.1679	0.0396
MA(3)	$\rho(1)$	0.1468	0.0373
	$\rho(2)$	0.1365	0.0399
	$\rho(3)$	0.1277	0.0414

Source: Own construction

In Table 1, we can see that our clipping method gives better results for every model and every order of autocorrelation function. It is caused by the bias of the standard method (Maronna et al., 2006).

For the second part we use a significant level $\alpha = 0.05$ and we obtain results summarized in Table 2.

Table 2 Comparison of the standard sample estimator and the clipping approach estimator in significant order in the simulation study.

Significant order	Standard approach	Clipping approach
1	2.94%	0.22%
2	10.01%	5.48%
3	70.26%	77.48%
4	3.83%	3.65%
5	3.98%	4.11%
6	4.38%	4.13%
7 and more	4.60%	4.93%

Source: Own construction

In Table 2, we can see how many simulations have the highest significant order of ACF at presented values 1, 2, ..., 6, 7 and more . We see that the standard approach tends to underestimate the significant order more than the robust estimator based on clipping. It is caused by outliers which weaken correlation between neighbor observations (Maronna et al., 2006), i.e. autocorrelation function.

4 REAL DATA APPLICATION

We investigate a daily time series of Bitcoin (the most used crypto currency) log returns. We have data (close prices) from *Yahoo Finance* from the period from 16/07/2010 to 18/08/2018 (2 956 observations). Usually, log returns are investigated instead of the original rates, because of a non-stationarity and a high autocorrelation, which could be misleading. Firstly, we define a log return:

$$r_n = \log\left(\frac{p_n}{p_{n-1}}\right), n > 1, \tag{23}$$

where p_n is a price for n-th observation.

Then we calculate the 1-year historical volatility from the log returns:

$$\sigma_n^{1Y} = \sqrt{\frac{1}{364}\sum_{i=n-364}^{n}\left(r_i - \overline{r_n}\right)^2}, n > 365, \tag{24}$$

where:

$$\overline{r_n} = \frac{1}{365}\sum_{i=n-365}^{n} r_i, n > 365 \tag{25}$$

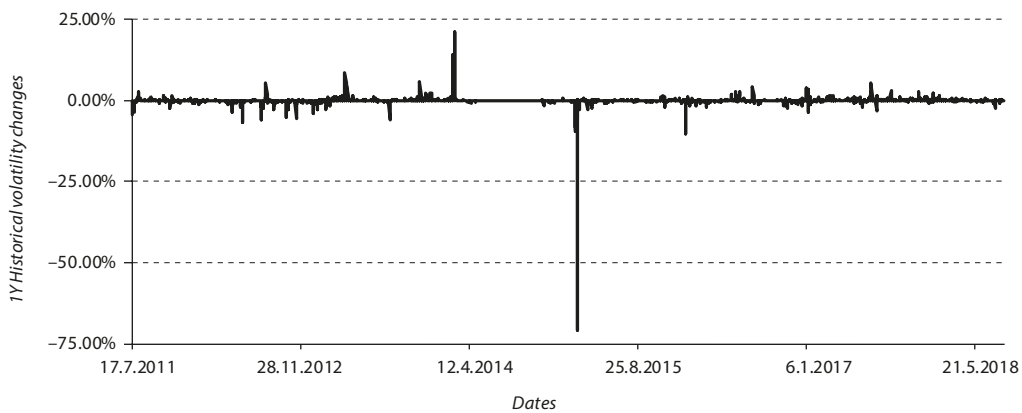
and 365 is the number of days within a year.

The point of our interest is the change (delta) of volatility, so we have to define the logarithmic change of the 1Y historical volatility:

$$\Delta_n^{1Y} = \log\left(\frac{\sigma_n^{1Y}}{\sigma_{n-1}^{1Y}}\right), n > 366. \tag{26}$$

Our logarithmic changes of the 1Y historical volatility are displayed in Figure 2.

Figure 2 Logarithmic changes of the yearly historical volatility of Bitcoin log returns

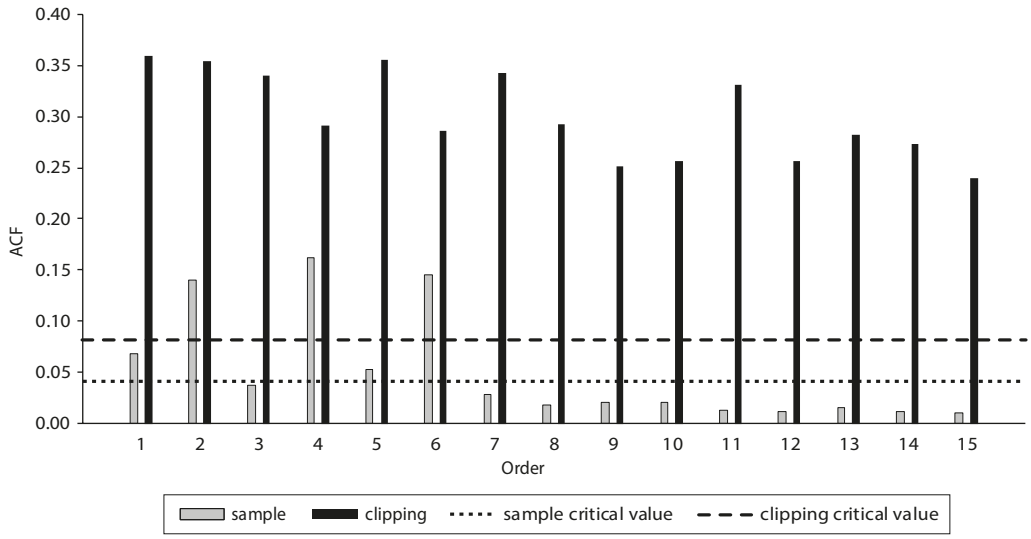


Source: Own construction

We can see a few isolated outliers in Figure 2. The biggest one is from 26/02/2015 and it is caused by losing a high log return from 26/02/2014, which was caused by price change from $135.78 to $593.14. The presence of these outliers could lead to a weaker autocorrelation.

Autocorrelation functions for standard sample estimator and clipping approach estimator are shown in Figure 3. It contains also critical values for a significant level $\alpha = 0.05$ for both estimators.

Figure 3 Comparison of the standard sample ACF estimation and the clipping approach ACF estimation with critical values



Source: Own construction

We can see the significance to 6th order of ACF for standard sample method. On the other hand, the clipping approach shows much higher correlations (with maximum 0.35 and average 0.3 over first 15 orders) and it is significant to the last presented order. The standard sample method could mislead us to a lower order of ACF and a weaker autocorrelation (with maximum 0.15 and average 0.05 over first 15 orders), but in reality, we should consider higher lags, or, alternatively, we could try to model the time series using AR(p) and have satisfying results.

DISCUSSION AND CONCLUSION

We have constructed a new robust ACF estimator based on clipping. Furthermore, we presented the asymptotical distribution of the estimator. We consider the knowledge of the distribution as a major advantage in comparison with other robust estimators, since it allows us to investigate the significance of the ACF orders or to test relevant hypotheses that occur when solving particular problems.

In Section 3, we have designed a simulation study, where we have compared the clipping approach estimator with the standard sample estimator using data contaminated with additive outliers. Firstly, we have compared the point estimates of the methods and our proposed clipping approach method has given better results for all cases. Secondly, we have exploited the knowledge of the clipping approach estimator and the standard sample estimator, the distribution of which is well known. The simulation study has shown us the underestimation of the standard approach in comparison with the clipping approach. This has confirmed our expectations, since additive outliers should weaken the relationship between neighboring observations.

Finally, we have compared both approaches in a study with real world data, where we have investigated the logarithmic change of 1Y historical volatility of Bitcoin log returns. The standard approach has suggested a weaker autocorrelation. We tend to trust more the clipping approach, since it has shown a stronger autocorrelation, which could lead us even to AR(p) time series.

In conclusion, we suggest to use robust methods in the case of additive outliers. On the other hand, for innovative outliers, it may be better to use the standard approach (Flimmel et al., 2017). If there is a need to know the distribution of the estimator, we definitely recommend the clipping approach estimator.

ACKNOWLEDGMENTS

This study was supported by the grant F4/17/2017 (Robustnosť v úmyselnom useknutí časového radu), which has been provided by the Internal Grant Agency of the University of Economics in Prague.

References

BARTLETT, M. S. On the Theoretical Specification and Sampling Properties of Autocorrelated Time-Series. *Supplement to the Journal of the Royal Statistical Society*, 1946, 8(1), pp. 27–41.

DÜRRE, A., FRIED, R., LIBOSCHIK, T. Robust estimation of (partial) autocorrelation. *Wiley Interdisciplinary Reviews: Computational Statistics*, 2015, 7(3), pp. 205–222.

FLIMMEL, S., ČAMAJ, M., MALÁ, I., PROCHÁZKA, J. Comparison of Robust Methods for ARMA Order Estimation. In: *Applications of Mathematics and Statistics in Economics (AMSE 2017)*, 2017, pp. 133–144.

FOX, A. J. Outliers in time series. *Journal of the Royal Society*, 1972, 34(3), pp. 350–363.

GNANADESIKAN, R. AND KETTENRING, J. R. Robust estimates, residuals, and outlier detection with multiresponse data. *Biometrics*, 1972, 28, pp. 81–124.

GREENE, W. H. *Econometric Analysis, 5th Edition*. Prentice-Hall, New Jersey, 2003.

CHAKHCHOUKH, Y. A new robust estimation method for ARMA models. *IEEE Transactions on Signal Processing*, 2010, 58(7), pp. 3512–3522.

CHAN, W.-S. AND WEI, W. W. A comparison of some estimators of time series autocorrelations. *Computational statistics & data analysis*, 1992, 14(2), pp. 149–163.

KEDEM, B. *Binary Time Series. Lecture Notes in Pure and Applied Mathematics. Vol. 52.* New York and Basel: Marcel Dekker, 1980.

MA, Y. AND GENTON, M. Highly robust estimation of the autocovariance function, *Journal of time series analysis*, 2000, 21(6), pp. 663–684.

MARONNA, R., MARTIN, D., YOHAI, V. *Robust Statistics: Theory and Methods*. Chichester: John Wiley & Sons, 2006.

R CORE TEAM. *R: A Language and Environment for Statistical Computing* [online]. R Foundation for Statistical Computing, Vienna, Austria, 2015. <https://www.r-project.org>.