IMPORTANCE OF ROBUST METHODS FOR ARMA ORDER ESTIMATING

Samuel Flimmel – Matej Čamaj – Ivana Malá – Jiří Procházka

Abstract

The current time is specific with loads of information that are stored about weather, sport or financial products. In financial market are used automatized algorithms that trade automatically based on various market information. These trades are made on second basis that implies possibility of price changes in a very short time frame, so financial markets produce loads of data. Consequently, there is higher probability to face an outlier and standard methods are not able to process them correctly. ARMA processes are well known and widely used in the financial sector. One of the important steps is to estimate an order of the process and outliers often cause bias in the order estimation. In time series theory there exist several outlier models such as additive outliers (AO), replacement outliers (RO) or innovations outliers (IO). We show importance of robust methods for ARMA order estimating via simulation study. Several robust methods are compared with standard methods. In addition we provide also real data study.

Key words: ARMA order estimating, robust methods, ARMA process

JEL Code: C02, C22, G10

Introduction

ARMA process is a well known and widely used to explain residue of randomness in the random process. ARMA order estimating is a very important step in time series analysis. After solving seasonality and stationarity of the process we need to estimate ARMA orders and then we can estimate the parameters.

Currently, when we are facing the big data problems, the importance of using robust methods is growing. Robust methods are usually more insensitive to outliers and they give better estimation in case of outlier presence. It was already shown by (Chan, 1992).

(Dürre, 2015) made a nice overview of the most important robust methods. For more detailed description of the method you can see (Maronna, 2006), (Ma, 2000) or others.

In Section 1, we establish some notation that we work with in this paper. In Section 2, we briefly introduce 2 robust method and the standard method that we use in our comparison. In Section 3, we show results from our simulation study, by which we compare the methods.

1 Definitions and notation

Let us define Gaussian white noise, which is a zero mean mutually uncorrelated time series $\{\varepsilon_n, n \in N_0\}$ with unknown constant variance $\sigma_\varepsilon^2 > 0$.

We define an autoregressive process AR(p) by equation

$$X_n = \varphi_1 X_{n-1} + \varphi_2 X_{n-2} + \cdots + \varphi_p X_{n-p} + \varepsilon_n, \tag{1}$$

where $\varphi_1, \varphi_2, \ldots, \varphi_p \in R$ are parameters, $\{\varepsilon_n, n \in N_0\}$ is a white noise and $\varphi_p \neq 0.$

We define a moving-average process MA(q) by equation

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

where $\theta_1, \theta_2, \ldots, \theta_q \in R$ are parameters, $\{\varepsilon_n, n \in N_0\}$ is a white noise and $\theta_q \neq 0.$

Finally, we define an autoregressive–moving-average process ARMA(p,q) by equation

$$X_n = \varphi_1 X_{n-1} + \varphi_2 X_{n-2} + \cdots + \varphi_p X_{n-p} + \varepsilon_n + \theta_1 \varepsilon_{n-1} + \theta_2 \varepsilon_{n-2} + \cdots + \theta_q \varepsilon_{n-q}, \tag{3}$$

where $\varphi_1, \varphi_2, \ldots, \varphi_p, \theta_1, \theta_2, \ldots, \theta_q \in R$ are parameters, $\{\varepsilon_n, n \in N_0\}$ is a white noise and $\varphi_p, \theta_q \neq 0.$

We define an autocovariance function of lag k $R(k)$of stationary process $\{X_n, n \in N_0\}$ as

$$R(k) = E(X_k - \mu)(X_0 - \mu), \tag{4}$$

where μ is an expected value of the process.

Let us define autocorrelation function (ACF) of lag k $\rho(k)$ of stationary process $\{X_n, n \in N_0\}$as

$$\rho(k) = \frac{R(k)}{\sigma_X^2}, \tag{5}$$

where σ_X^2 is an variance of the process.

Let us define partial autocorrelation function (PACF) of lag k $\alpha(k)$ of stationary process $\{X_n, n \in N_0\}$ as

$$\begin{aligned} &\alpha(1) = \rho(1) \\ &\alpha(k) = corr\left(X_k - \tilde{X}_k, X_0 - \tilde{X}_0\right), k > 1 \end{aligned} \tag{6}$$

where *corr* denotes correlation and $\tilde{X}_0$ ($\tilde{X}_k$) is projection of X_0 (X_k) onto the Hilbert's space spanned by $X_1, X_2, X_3, \ldots, X_{k-1}$.

2 Estimation methods

We introduce briefly all methods that we use in a simulation study. Firstly, we need to estimate an autocorrelation function of the process. We have $m+1$ observations $X_0, X_1, \ldots, X_m$, from which we estimate the ACF.

Let us start with a standard method, e.g. (Hamilton, 1994)

$$\hat{\rho}_S(k) = \frac{\sum_{i=0}^{m-k}\left(X_{i+k} - \bar{X}\right)\left(X_i - \bar{X}\right)}{\sum_{i=0}^{m}\left(X_i - \bar{X}\right)^2}, \tag{7}$$

where $\bar{X}$ is an average of $X_0, X_1, \ldots, X_m$.

We introduce two robust methods: method based on the Gnanadesikan-Kettenring approach and method based on the robust filtering.

The method based on the Gnanadesikan-Kettenring approach, which was introduced by (Gnanadesikan and Kettenring, 1972), is defined as

$$\hat{\rho}_{GK}(k) = \frac{Q_{m-k}^2(u+v) - Q_{m-k}^2(u-v)}{Q_{m-k}^2(u+v) + Q_{m-k}^2(u-v)}, \tag{8}$$

where *u* is the vector $(X_{m-k}, X_{m-k+1}, \ldots, X_m)$, *v* is the vector $(X_0, X_1, \ldots, X_k)$ and Q_m is robust estimator of the scale. It was proposed by (Croux, 1992) and it is defined as:

$$Q_m = c\left[\left|X_i - X_j\right|, i < j\right]_l, \tag{9}$$

where $[\cdot]_l$ is *l*th order statistic and *l* is defined as

$$l = \left\lfloor \frac{\binom{m}{2} + 2}{4} \right\rfloor + 1, \tag{10}$$

where $\lfloor \cdot \rfloor$ denotes the floor function. The factor *c* is for consistency, for the Gaussian distribution *c* = 2.2191. The method of this robust ACF estimator was presented by (Ma, 2000).

The robust filtering approach takes the time series structure into account. The idea is to have robust filtered values instead of the original observation and calculate ACF from these filtered values. Practically we replace outliers by some reasonable values.

Firstly, we estimate “long” AR process, which we use for robust filtering. Consequently, we obtain fitted values using the robustly filtered τ-scale estimate and finally calculate autocorrelation function. The method of this robust ACF estimator was presented by (Maronna, 2006).

When we already have the estimation of ACF, we are able to estimate PACF too. There exist the theorem (e.g. (Yafee, 2000)) describing a relation between ACF and PACF

$$\alpha(k)=\frac{\begin{vmatrix} 1 & \rho(1) & \cdots & \rho(k-2) & \rho(1) \\ \rho(1) & 1 & \cdots & \rho(k-3) & \rho(2) \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \rho(k-1) & \rho(k-2) & \cdots & \rho(1) & \rho(k) \end{vmatrix}}{\begin{vmatrix} 1 & \rho(1) & \cdots & \rho(k-1) \\ \rho(1) & 1 & \cdots & \rho(k-2) \\ \vdots & \vdots & \ddots & \vdots \\ \rho(k-1) & \rho(k-2) & \cdots & 1 \end{vmatrix}}, k>1, \tag{11}$$

where $|\cdot|$ represents determinant. Having the ACF and the PACF of the process we can estimate orders of the ARMA(p,q) process.

3 Simulation study

The simulation study was designed in software *R* and we use *R* package *robts*. But the package is still not approved by CRAN, so a few functions were coded by authors of this paper to validate correctness of the package. After validation we use functions from package to obtain estimations in the simulation study.

We use Bartlett’s approximation (Bartlett, 1946) for determination significant order k of ACF

$$\hat{\rho}(k) \sim N\left(0, \frac{1+2\sum_{i=1}^{k_0}\hat{\rho}(i)^2}{m}\right), k>q_0, \tag{12}$$

if $\rho(k)=0$ for $k>q_0$. So we search the q_0 that holds

$$|\hat{\rho}(k)| > 2\sqrt{\frac{1+2\sum_{i=1}^{k_0}\hat{\rho}(i)^2}{m}}, k>q_0. \tag{13}$$

Similarly, for partial autocorrelation function we use Quenouille’s approximation (Quinouille, 1949):

$$|\hat{\alpha}(k)| > 2\sqrt{\frac{1}{m}}, k>p_0. \tag{14}$$

ARMA process is known as a process without p_0 and q_0 in (13) and (14). Of course there always exist some p_0 and q_0 (for stationary ARMA process) that will hold both inequalities (13) and (14), but long orders are not preferable from the practical point of view. Maximum value of p_0, respecitvely q_0, was being chosen to be 6. If there is no $p_0 \leq 6$, respectively $q_0 \leq 6$, we assume there exist no p_0, respectively q_0, at all.

We use additive outlier model and innovative outlier model (see e.g. (Maronna, 2006)) in the simulation study.

For every case we run 5000 simulations with 1000 observations. For probability (ε) of outliers being present in one simulation we choose 3 cases: ε = 0%, ε = 1% and ε = 5%.

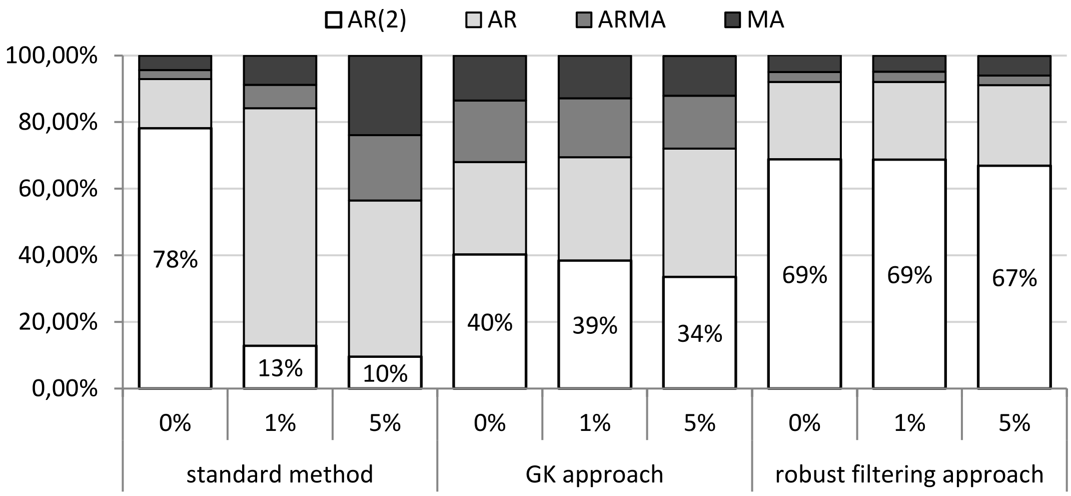
Every simulation is evaluated simultaneously with all 3 described methods as AR with order p between 1 and 6, MA with order q between 1 and 6 or general ARMA. In case of simulation for AR process we collect MA processes of all orders into one category and analogously we do the same for MA process. The simulations are evaluated according the rules we mentioned above.

3.1 Autoregressive process AR(2)

Absolute value of the parameters of the AR(2) process are generated randomly with uniform distribution, i.e $\varphi_i \sim U((0.2,1.0))$.. Values 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 probability of success π= 0.5. Subsequently, we check whether these parameters give a stationary process and we repeat the procedure until it is necessary.

Results for AO model ($\sigma_A = 10$) can be seen in Figure 1.

Fig. 1: Process AR(2) with contaminated data by AO model.

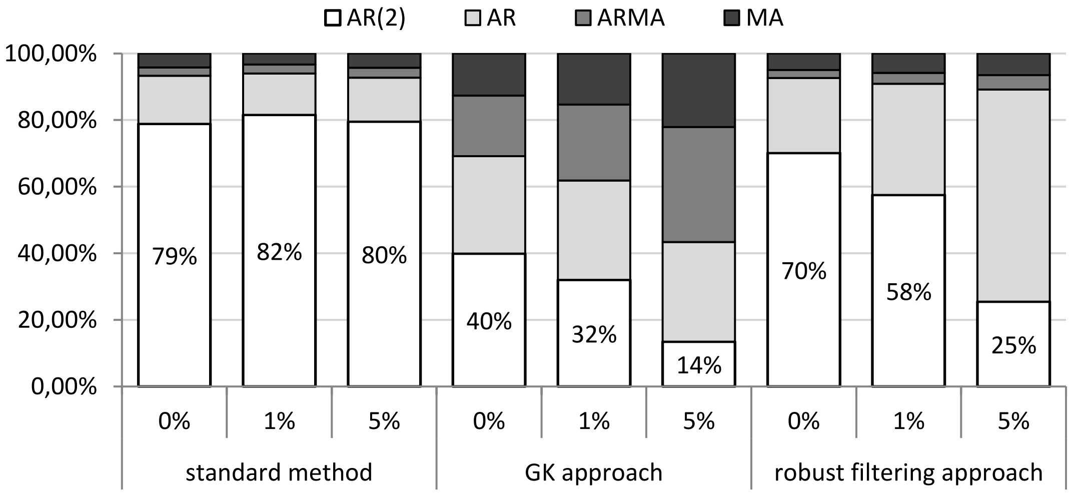


Source: Authors' own calculations

We can see, that standard method is not able to process the outliers. By increasing probability of outlier presence, which means more outliers present in the observations, accurancy of correct evaluation of ARMA orders is decreasing drastically. Naturally, the standard method gives the best result in case of no outliers in the observations. But the difference between the standard method and the approach based on robust filtering is quite small. As we expect, both of robust methods give similar result for every probability of outlier presence. If we compare only robust methods, approach based on robust filtering seems to be better choice.

Results for IO model ($\sigma_I = 10$) are given in Figure 2.

Fig. 2: Process AR(2) with contaminated data by IO model.



Source: Authors' own calculations

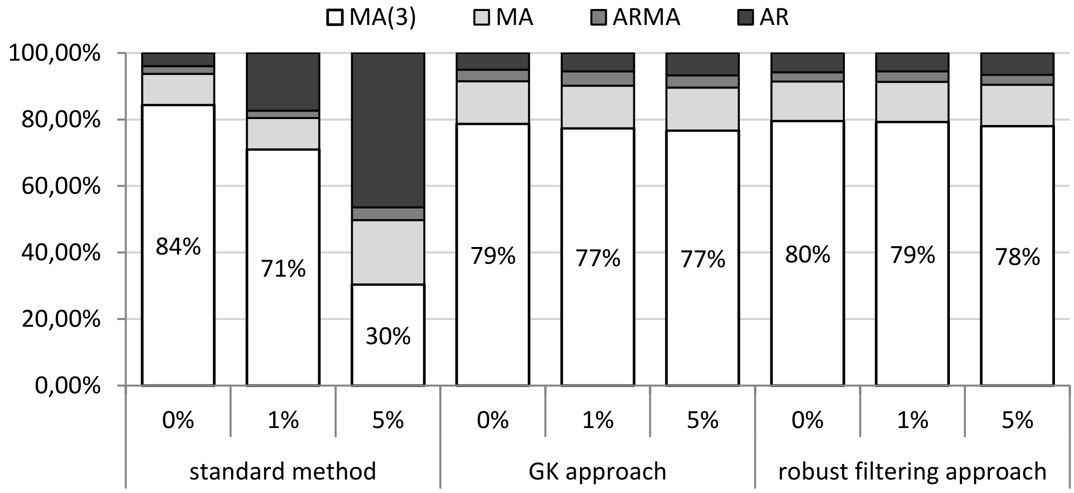
We can see, that innovative outliers have no impact to the standard method. On the contrary, presented robust methods are more sensitive to IOs. Again if we compare only these two robust methods, approach based on robust filtering gives better resutls.

3.2 Moving-average process MA(3)

Similarly as for the AR(2), absolute value of the parameters of the MA(3) process are generated randomly with uniform distribution, i.e. $\theta_i \sim U((0.2,1.0))$. Values close to zero are not taken into, because they are difficult to observe. The sign of the parameters is generated randomly with Bernoulli's distribution with probability of success $\pi = 0.5$.

Results for AO model ($\sigma_A = 10$) can be seen in Figure 3.

Fig. 3: Process MA(3) with contaminated data by AO model.

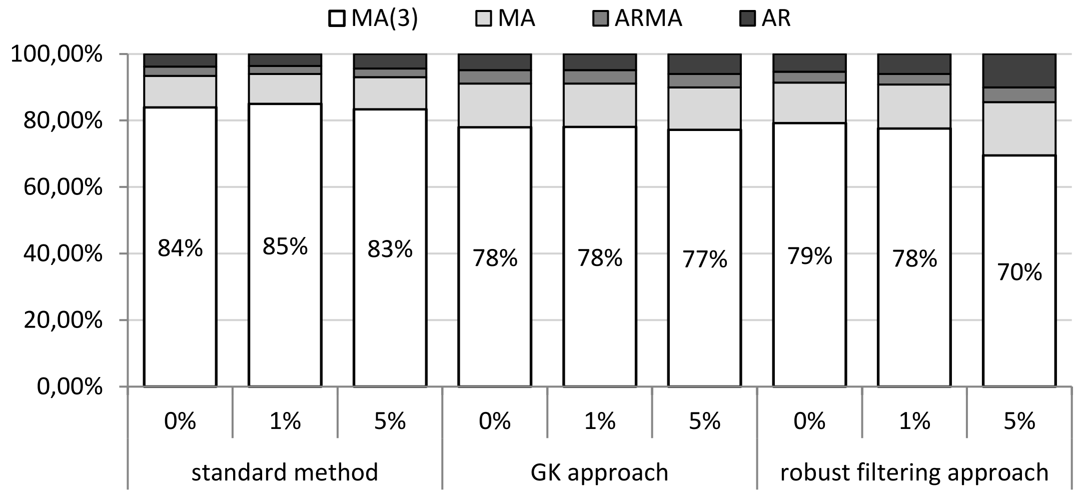


Source: Authors' own calculations

The results are much better in comparison with process AR(2). Similarly as for process AR(2) with AOs, we can see decreasing trend in accuracy of ARMA order estimating for the standard method. Higher probability of outlier presence affects the standard method significantly more than both robust methods. Both robust methods seems to process AOs very well, the accuracy of ARMA order estimating is almost the same regardless of the probability of outlier presence.

Results for IO model ($\sigma_I = 10$) can be seen in Figure 4.

Fig. 4: Process MA(3) with contaminated data by IO model.



Source: Authors' own calculations

We can see the similar figure as before that implies robustness of standard method for IOs. In this case also robust methods give good results. For ε = 5% we can see slightly better results for GK approach in comparison with robust filtering approach.

Conclusion

We briefly introduced the standard method and two robust methods for ACF estimating. Using ACF estimates we estimated also PACF and then we introduced the algorithm for ARMA order estimating.

We provided simulation study and compare the methods. We saw, that the AR process is affected by outliers much more in comparison with MA process.

The AOs have strong impact on the standard method and we should not use the method in these situations. Both robust methods gave better results than the standard method. The method based on robust filtering looked even slightly better than the method based on GK approach.

On the other hand, the IOs have no impact on the standard method, but the robust methods are affected by them.

We would recommend to check the outlier presence at first. Then we should try to detect the nature of outliers. If we detect the innovative outliers, we should use the standard method. But if we detect the additive outliers, we should definitely use one of the robust method. Otherwise we risk to estimate ARMA orders incorrectly.

Acknowledgment

This paper is 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. (1946). On the Theoretical Specification and Sampling Properties of Autocorrelated Time-Series. *Supplement to the Journal of the Royal Statistical Society, 8*(1), 27. doi:10.2307/2983611

Chan, W. (1992). A note on time series model specification in the presence of outliers. *Journal of Applied Statistics*, 19(1), 117-124. doi:10.1080/02664769200000010

Croux, C., & Rousseeuw, P. J. (1992). *Explicit scale estimators with high breakdown point.* Berkeley, CA: Math. Sciences Research Inst.

Dürre, A., Fried, R., & Liboschik, T. (2015). Robust estimation of (partial) autocorrelation. *Wiley Interdisciplinary Reviews: Computational Statistics, 7*(3), 205-222. doi:10.1002/wics.1351

Gnanadesikan, R., & Kettenring, J. R. (1972). Robust Estimates, Residuals, and Outlier Detection with Multiresponse Data. *Biometrics*, *28*(1), 81. doi:10.2307/2528963

Hamilton, J. D. (1994). *Time series analysis.* Princeton, NJ: Princeton University Press.

Ma, Y., & Genton, M. G. (2000). Highly Robust Estimation of the Autocovariance Function. *Journal of Time Series Analysis*, *21*(6), 663-684. doi:10.1111/1467-9892.00203

Maronna, R. A., Martin, R. D., & Yohai, V. J. (2006). *Robust statistics: theory and methods.* Chichester: Wiley.

Quenouille, M. H. (1949). Approximate tests of correlation in time-series. *Mathematical Proceedings of the Cambridge Philosophical Society, 45*(03), 483. doi:10.1017/s0305004100025123

Yaffee, R. A., & McGee, M. (2009). *Introduction to time series analysis and forecasting: with applications of SAS and SPSS*. San Diego: Academic Press.

Contact

Samuel Flimmel
University of Economics, Prague
W. Churchill Sq. 1938/4, Prague, Czech Republic
samuel.flimmel@vse.cz

Matej Čamaj
University of Economics, Prague
W. Churchill Sq. 1938/4, Prague, Czech Republic
matej.camaj@vse.cz

Ivana Malá
University of Economics, Prague
W. Churchill Sq. 1938/4, Prague, Czech Republic
malai@vse.cz

Jiří Procházka
University of Economics, Prague
W. Churchill Sq. 1938/4, Prague, Czech Republic
xproj16@vse.cz