Scientific Research Publishing

Open Journal of Statistics, 2019, 9, 543-554
https://www.scirp.org/journal/ojs
ISSN Online: 2161-7198
ISSN Print: 2161-718X

Application of the Improved Generalized Autoregressive Conditional Heteroskedast Model Based on the Autoregressive Integrated Moving Average Model in Data Analysis

Qi Yang, Yishu Wang

Qingdao University, Qingdao, Shandong, China
Email: 1131142973@qq.com, yishu6661@126.com

How to cite this paper: Yang, Q. and Wang, Y.S. (2019) Application of the Improved Generalized Autoregressive Conditional Heteroskedast Model Based on the Autoregressive Integrated Moving Average Model in Data Analysis. *Open Journal of Statistics*, **9**, 543-554.
https://doi.org/10.4236/ojs.2019.95036

Received: August 14, 2019
Accepted: September 3, 2019
Published: September 6, 2019

Abstract

This study firstly improved the Generalized Autoregressive Conditional Heteroskedast model for the issue that financial product sales data have singular information when applying this model, and the improved outlier detection method was used to detect the location of outliers, which were processed by the iterative method. Secondly, in order to describe the peak and fat tail of the financial time series, as well as the leverage effect, this work used the skewed-t Asymmetric Power Autoregressive Conditional Heteroskedasticity model based on the Autoregressive Integrated Moving Average Model to analyze the sales data. Empirical analysis showed that the model considering the skewed distribution is effective.

Keywords

Forecasting, Outliers, Improved GARCH Model, Partial T-APARCH Model Based on ARIMA Model

1. Introduction

Time series models play very important roles in many business decisions. In the current big data era, all walks of life are faced with the problem of modeling and time sequence prediction. For example, the e-commerce platform needs to predict the future sales of all commodities; in the pre-sales industry, both online and offline pre-sales require significant time series forecasting. These data are non-linearly correlated in time series, with most of them affected by product

promotion, inventory situation and market competition among other, which may lead to outliers in time series. Meanwhile, sales during the holiday promotion period are relatively volatile and flat, resulting in an asymmetry of yield fluctuation and the rate of return usually does not follow the normal distribution, showing skewness and peak thick tail. Therefore, how could such data be modeled and predicted remains an open question.

Let's start with classic time series models, such as the Autoregressive Integrated Moving Average model-Generalized Autoregressive Conditional Heteroskedast (ARIMA-GARCH) model [1] and normal Asymmetric Power Autoregressive Conditional Heteroskedasticity (APARCH) model [2] based on the Autoregressive Integrated Moving Average (ARIMA) model [3]. On the one hand, although the model can solve the heteroscedastic effect in the residual sequence, it provides no solution for singular information when data are applied to the Generalized Autoregressive Conditional Heteroskedast (GARCH) model [4]. On the other hand, the classical time series model in parameter estimation is usually based on the assumption of normal distribution, which does not fit well the distribution of volatility in practical applications. The explosive growth of new algorithm development makes this issue even more worthy of attention.

Therefore, in this study, an improved GARCH model, termed the Pro-GARCH model, was proposed to solve the problem of singular information in the data. The improved method is described below. First, in the GARCH (p, q) model, the rank of the Hessian matrix H is defined as $R_1 = p+q+1$. We performed QR decomposition on the Hessian matrix H, *i.e.* $H = QR$; R (QR) represents the rank of the matrix after QR decomposition. Application of data to the model ($R_1 \neq R(QR)$) leads to the generation of singular information, and the rank of the matrix after QR decomposition is now increased by one, *i.e.* $R_2 = R(QR)+1$, so that $R_2 = R_1$, which allows to solve the problem of singular information. Since the rank of the matrix is changed after QR decomposition, the estimated value of a given parameter is not affected. Secondly, because the matrix is singular, the inverse matrix of the Hessian matrix was obtained by determining the generalized inverse of the matrix, yielding the Pro-GARCH model. Furthermore, a skewed-t APARCH model [5] based on the ARIMA model [3] was proposed. After assessing JD sales data, the results showed that the model was superior to the classical time series model in the accuracy of parameter estimation and prediction, and could more accurately describe the skewness problem in the sequence.

The remainder of the article is as follows. In the second part, the definitions of Pro-GARCH model and skewed-t APARCH model based on the ARIMA model will be provided. In the third part, the modeling process of skewed-t APARCH model based on the ARIMA model will be described. In the fourth part, we applied the novel and traditional classical time series models to JD sales data, respectively, and compared the results. Empirical analysis showed that the model is better than other models.

2. Model

2.1. The Pro-GARCH Model

The Pro-GARCH (p, q) model we proposed is:

$$\begin{cases} x_t = f\left(t, x_{t-1}, x_{t-2}, \cdots\right) + \varepsilon_t \\ \varepsilon_t = \eta_t \sqrt{h_t} \\ \varepsilon_t \mid \varepsilon_{t-1}, \varepsilon_{t-2}, \cdots \sim N\left(0, \sqrt{h_t}\right) \\ h_t = \alpha_0 + \sum_{i=1}^{q} \alpha_i \varepsilon_{t-i}^2 + \sum_{i=1}^{p} \beta_i h_{t-i} \end{cases} \tag{1}$$

where, $\{x_t\}$ is the deterministic information fitting, η_t is independent and follows the standard normal distribution, and $\alpha_0 > 0$, $\alpha_i > 0$, $\beta_i > 0$ and $\alpha_i + \beta_i < 1$. The improved GARCH (p, q) model can also be rewritten as the ARMA (p, q) model for ε_t^2, *i.e.*

$$\varepsilon_t^2 = \alpha_0 + \sum_{i=1}^{q} \left(\alpha_i + \beta_i\right) \varepsilon_{t-i}^2 + v_t - \sum_{i=1}^{p} \beta_i v_{t-i}, \ i = 1, \cdots, T \tag{2}$$

where, $v_t = \varepsilon_t^2 - h_t$ [4].

With an outlier in the data, the actual sequence is not ε_t, but an observation sequence e_t, defined as

$$e_t^2 = \varepsilon_t^2 + \omega_t \xi\left(B\right) I_t\left(T\right) \tag{3}$$

where $I_t\left(T\right)$ is the indicator function, ω_t and $\xi\left(B\right)$ denote the magnitude and dynamic model of the outlier effect, respectively.

2.2. The Partial T-APARCH Model Based on the ARIMA Model

The skewed-t APARCH model [5] based on the ARIMA model [3] can be defined as:

$$\begin{cases} \Phi\left(B\right) \nabla^d x_t = \Theta\left(B\right) \varepsilon_t \\ E\left(\varepsilon_t\right) = 0, Var\left(\varepsilon_t\right) = \sigma_\varepsilon^2, E\left(\varepsilon_t \varepsilon_s\right) = 0, s \neq t \\ E\left(x_s \varepsilon_t\right) = 0, \forall s < t \\ \varepsilon_t = \mu_t + a_t, a_t = \sigma_t \eta_t, \eta_t \sim D\left(0,1\right) \\ \sigma_t^\delta = \omega + \sum_{i=1}^{q} \alpha_i \left(\left|\varepsilon_{t-i}\right| + \gamma_i \varepsilon_{t-i}\right)^\delta + \sum_{j=1}^{p} \beta_j \sigma_{t-j}^\delta \end{cases} \tag{4}$$

where, $\nabla^d = \left(1-B\right)^d$, $\Phi\left(B\right) = 1 - \phi_1 B - \cdots - \phi_p B^p$ is an autoregressive coefficient polynomial for a stationary reversible ARIMA (p, q) model; $\Theta\left(B\right) = 1 - \theta_1 B - \cdots - \theta_p B^p$ is the moving smoothing coefficient polynomial for the stationary reversible ARIMA (p, q) model; μ_t is the conditional mean, $D\left(0,1\right)$ represents a distribution with mean and variance of 0 and 1, respectively; $\omega > 0, \delta \geq 0, \beta_j \geq 0 \left(j = 1, \cdots, p\right), \alpha_i \geq 0, -1 < \gamma_i < 1$; η_t is independent, identically distributed and follows the skewed $t\left(0,1,v,\xi\right)$ distribution. The purpose of the power function in Equation (4) is to improve the transformation of model fitting.

3. Processing

1) We built the ARIMA model [6] as follows. a) Data preprocessing. First, the Pro-GARCH model and the improved outlier detection method [7] were used to detect IO type outliers of the data. Then, the iterative method [8] was used to process the outliers and generate new data; Secondly, the stability and pure randomness of the new time series data were tested. b) Model identification. After calculating the sample autocorrelation coefficient and partial correlation coefficient, the appropriate ARIMA model was selected to fit the observation sequence. c) Model prediction and diagnosis. The established model was used to predict future trends of time series values and evaluate the model by analyzing whether parameter estimates are significant and the residual is a white noise sequence.

2) Autocorrelation and heteroscedasticity test for the residual sequence of the ARIMA model were performed by a statistical method using the Portmanteau Q test and the LM test [1].

3) Model identification. After constructing the ARIMA model, the residual sequence was modeled by APARCH (1, 1), selecting skewed-t distribution.

4) Model diagnosis. Whether the residual sequence was a white noise sequence and the parameter estimation significant was assessed.

5) Model prediction. The final fitted model was obtained and evaluated.

4. Results

In this section, the construction process of the skewed-t Asymmetric Power Auto-regressive conditional heteroskedasticity (APARCH) model based on the ARIMA model is introduced in detail. Compared with the ARIMA-GARCH and normal APARCH models, respectively, based on the ARIMA model, the validity of the skewed-t APARCH model based on the ARIMA model was demonstrated. All simulations in this paper were performed in R.

4.1. Data Preprocessing

We analyzed the sales data of Jingdong. **Figure 1** shows the presence of outliers in the data. Therefore, we first used the Pro-GARCH model and the improved outlier detection method to process outliers and generated new data. This result was satisfactory. Among them, the data obtained after processing the abnormal values are shown in **Figure 2** and were recorded as $\{train\_x\}$.

4.2. Model Establishment

The pure randomness test results showed that the P value of the LB test statistic was very low under the first-order to sixth-order delay (see **Table 1**). Therefore, we determined that the sequence belonged to a non-white noise sequence and could model the data.

The stability of the sequence $\{train\_x\}$ was verified by the timing diagram method. As shown in **Figure 2**, the sequence was non-stationary. After using the

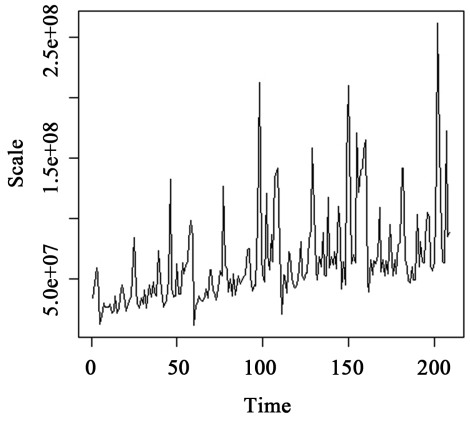


Figure 1. Data with outliers.

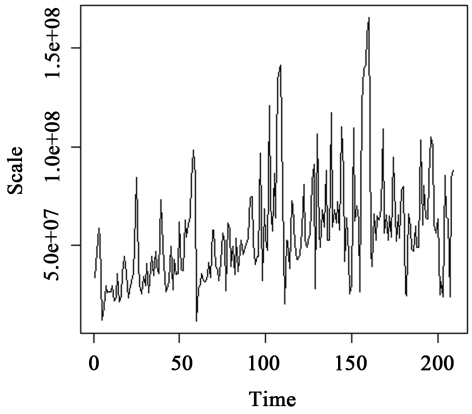


Figure 2. Data after outlier processing.

Table 1. P value of LB test statistics.

Delay order	P value	Delay order	P value
1	2.38e - 13	4	<2.2e - 16
2	<2.2e - 16	5	<2.2e - 16
3	<2.2e - 16	6	<2.2e - 16

ARIMA (p, d, q) model to fit the sequence $\{train\_x\}$, the first-order difference of the sequence $\{train\_x\}$ is shown in **Figure 3**. As shown in **Figure 3**, the sequence $\{train\_x_1\}$ after the first-order difference was stationary, so in the ARIMA (p, d, q) model, the order of the difference was 1, *i.e.* $d = 1$.

Figure 4 shows the autocorrelation (ACF) and partial correlation (PACF) plots of the sequence $\{train\_x_1\}$. The system's automatic ordering was compared with the ACF and PACF graphs. With $p = 1$ and $q = 1$, the model fitting was most reasonable; therefore, the ARIMA (p, d, q) model most suitable for this sequence was the ARIMA (1, 1, 1) model. Regarding the heteroscedasticity test, since the P values of the LM and Portmanteau Q tests were low (see **Table 2**), the residuals were heteroscedastic. Due to space constraints, only the first six P values were reported in this paper.

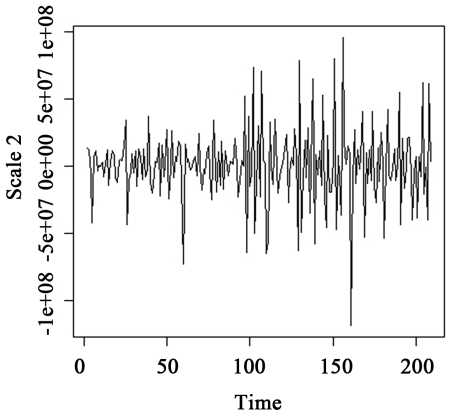


Figure 3. Data $\{train\_x_1\}$ after first-order difference.

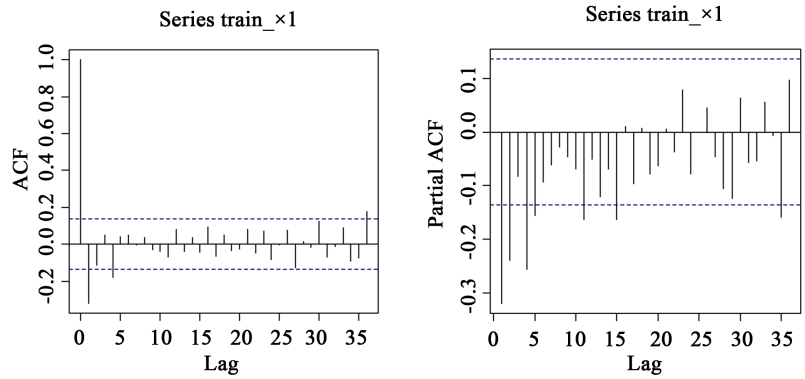


Figure 4. Autocorrelation and partial correlation plots for $\{train\_x_1\}$.

Table 2. P values for LM and portmanteau Q tests.

Delay order	P value of the LM test	P value of the Portmanteau Q test
1	2.081e - 05	2.025e - 05
2	3.748e - 05	2.479e - 06
3	0.0001276	2.002e - 06
4	0.0003327	1.911e - 06
5	1.669e - 06	8.294e - 10
6	1.788e - 06	1.874e - 09

Therefore, this study selected the ARIMA (1, 1, 1)-GARCH (1, 1) and APARCH (1, 1) models based on the ARIMA (1, 1, 1) model for modeling the sequence under normal and partial t distributions. **Table 3** provides the parameter estimation results of the new model.

4.3. Evaluation Criteria

In order to evaluate the accuracy of the model, mean error (*MSE*), mean absolute error (*MAE*), root mean square error (*RMSE*) and mean absolute percentage error (*MAPE*) were used. The smaller the variance of each loss function, the

Table 3. Model parameter estimation results.

ARIMA-GARCH		APARCH (1, 1) - N	APARCH (1, 1) - skewed t
		Based on the ARIMA model	Based on the ARIMA model
ar1	−0.375	0.367***	0.367***
ma1	0.431	−0.951***	−0.951***
omega	0.012***	0.022***	0.006*
α_1	0.161***	0.174***	0.138***
β_1	0.794***	0.797***	0.883***
γ_1	-	0.08	0.122
δ	-	1.385***	1.302***
Skew	-	-	0.925***
Shape	-	-	4.196***

Note: "***", "**", "*", "." indicate that the parameters are significant.

smaller the prediction error, and the more accurate the prediction. The calculation formula was as follows:

$$MSE = (1/N) * \sum_{t=1}^{N} \left(\tilde{\sigma}_t^2 - \hat{\sigma}_t^2 \right)^2 \tag{5}$$

$$MAE = (1/N) * \sum_{t=1}^{N} \left| \tilde{\sigma}_t^2 - \hat{\sigma}_t^2 \right|^2 \tag{6}$$

$$RMSE = \left[(1/N) * \sum_{t=1}^{N} \left(\tilde{\sigma}_t^2 - \hat{\sigma}_t^2 \right)^2 \right]^{1/2} \tag{7}$$

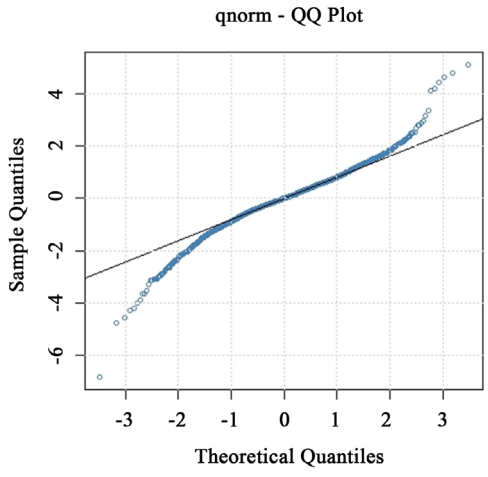
$$MAPE = (1/N) * \sum_{t=1}^{N} \left| 1 - \tilde{\sigma}_t \hat{\sigma}_t \right| \tag{8}$$

where, $\tilde{\sigma}_t^2$ is the realized volatility, estimated using high frequency data, and $\hat{\sigma}_t^2$ is the predicted volatility at time t; N is the number of the performance evaluation data.

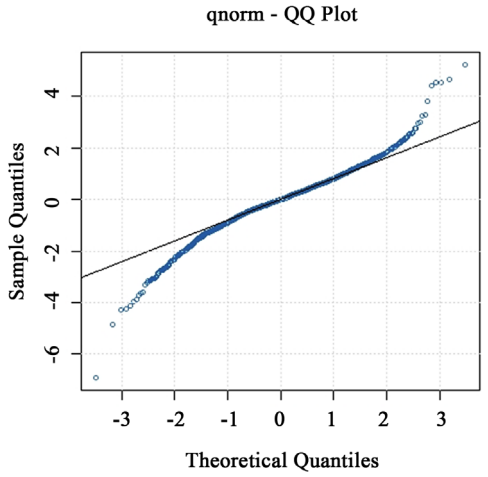
4.4. Prediction Results

Tables 4-6 provide the standardized residual test of the ARIMA (1, 1, 1)-GARCH (1, 1) and APARCH (1, 1) models based on the ARIMA (1, 1, 1) model under the assumption of normal and partial t distributions, respectively. Table 7 shows the prediction effect of the 20 steps of the model. Figure 5 provides the standardized residual QQ diagrams of ARIMA (1, 1, 1)-GARCH (1, 1) and APARCH (1, 1) models based on the ARIMA (1, 1, 1) model in normal and partial t distributions, respectively.

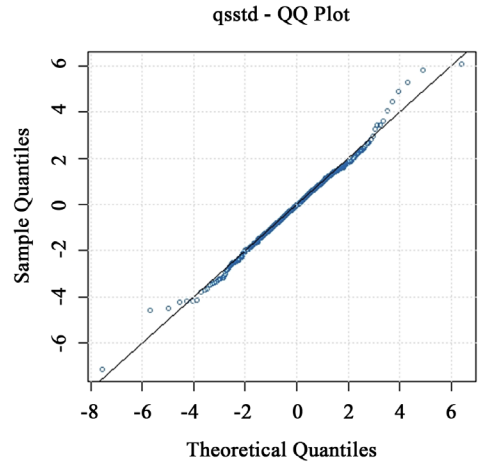
Four evaluation indicators were assessed, including parameter estimation saliency, standardized residual test, evaluation criteria and standardized residual QQ map. 1) parameter estimation of the skewed-t APARCH (1, 1) model based on the ARIMA model was more significant. 2) Considering the standardized



(a)



(b)



(c)

Figure 5. (a) The standardized residual QQ plot of ARIMA (1, 1, 1)-GARCH (1, 1) model; (b) The standardized residual QQ diagrams of APARCH (1, 1) models based on the ARIMA (1, 1, 1) model in normal distributions; (c) the standardized residual QQ diagrams of APARCH (1, 1) models based on the ARIMA (1, 1, 1) model in partial t distributions.

Table 4. Standardized residual test of the ARIMA (1, 1, 1)-GARCH (1, 1) model.

			Statistic	P-Value
Ljung-Box Test	R	Q (10)	4.569693	0.9180108
Ljung-Box Test	R^2	Q (10)	8.465982	0.5834181
LM Arch Test	R	TR^2	9.115951	0.6929961

Table 5. Standardized residual test of the normal APARCH model based on the ARIMA model.

			Statistic	P-Value
Ljung-Box Test	R	Q (10)	10.1409	0.4282198
Ljung-Box Test	R^2	Q (10)	9.741699	0.4634398
LM Arch Test	R	TR^2	10.24594	0.5943943

Table 6. Standardized residual test of the skewed-t APARCH model based on the ARIMA model.

			Statistic	P-Value
Ljung-Box Test	R	Q (10)	10.46054	0.4010589
Ljung-Box Test	R^2	Q (10)	16.42381	0.08812606
LM Arch Test	R	TR^2	17.94269	0.1174418

Table 7. Analysis of the prediction effect of the model.

	MAPE	MSE	MAE	RMSE
ARIMA-GARCH model	0.402	0.097	0.286	0.311
ARIMA-normal APARCH	0.414	0.108	0.303	0.329
ARIMA-skewed-t APARCH	0.399	0.1	0.284	0.317

residuals test, all three models completely eliminated the ARCH effect and the correlation between sequences, so they could not be rejected. 3) The skewed-t APARCH (1, 1) model based on the ARIMA model was slightly higher in accuracy compared with the other two models. 4) The skewed-t distribution had a better fitting effect in the standardized residual map, indicating that the influence of introducing bias on model fitting was significant. Therefore, the skewed-t APARCH (1, 1) model based on the ARIMA model had a better prediction ability. Therefore, the final prediction model was as follows:

$$x_t = 1.367x_{t-1} + 0.367x_{t-2} + \varepsilon_t + 0.9511\varepsilon_{t-2} \tag{9}$$

$$\sigma_t^{1.302} = 0.006 + 0.138\left(\left|\varepsilon_{t-1}\right| + 0.122\varepsilon_{t-1}\right)^{1.302} + 0.883\sigma_{t-1}^{1.302} \tag{10}$$

where, the skewness is 0.925 [9], the model coefficient is greater than 0 and satisfies $-1 < \gamma < 1$. Equations (9) and (10) are the mean and variance equations, respectively.

Further, the residual, predicted confidence interval and 95% confidence in-

terval of the partial-t APARCH (1, 1) model based on the ARIMA (1, 1, 1) model are depicted in Figure 6. The residual was almost completely within the confidence interval, indicating that model prediction was more accurate; the volatility of the model is presented in Figure 7.

5. Discussion

First, the Pro-GARCH model solves the singular information problem in data. Secondly, as shown in Figure 5, the skewed-t APARCH model based on the ARIMA model could better capture the peak thick tail, skewness and leverage effect in the sequence. Finally, Table 3 and Table 7 show that the model is superior to the ARIMA-GARCH and APARCH models based on the ARIMA model under the assumption of normal distribution in the significance of parameter estimation and accuracy of prediction, respectively.

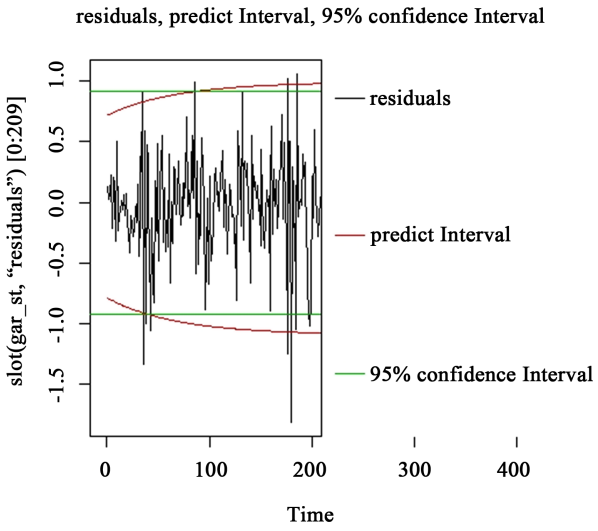


Figure 6. Residual, predictive confidence interval and 95% confidence interval plots.

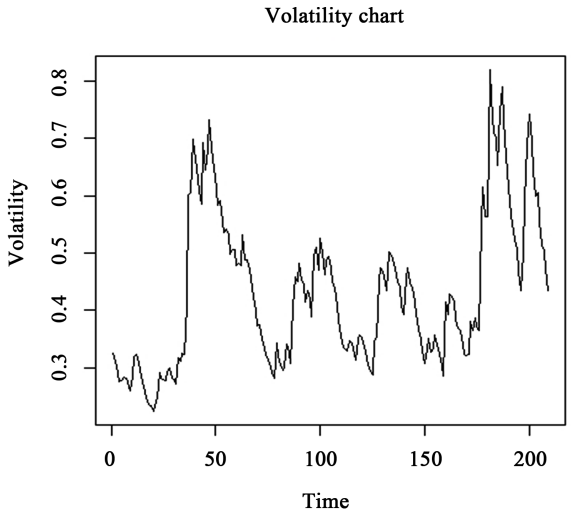


Figure 7. Model volatility.

6. Conclusions

In this study, the Pro-GARCH model and the improved outlier detection method were used, and the iterative method was used to process outliers in the time series to obtain a new time series. The ARIMA-GARCH and normal APARCH models based on the ARIMA model, and the skewed-t APARCH model based on the ARIMA model were compared. Some concluding observations can be summarized as follows:

1) Using the Pro-GARCH model and the improved outlier detection method to process data and selecting absolute deviation of the median (MAD) as a robust estimation of the standard deviation of the model, the location of outliers could be found most accurately;

2) Compared with the ARIMA-GARCH and normal APARCH models based on the ARIMA model, the skewed-t APARCH model based on the ARIMA model could better capture the spikes and thick tails, skewness and leverage effects, and the model had elevated prediction ability;

3) No prediction method could stand out in any time series. Although the skewed-t APARCH model based on the ARIMA model showed good predictive power, it did not achieve the expected results, and there were certain losses; this model is not flexible and cannot be applied to multiple products simultaneously. This is a huge challenge for time series modelers and requires further research.

Acknowledgements

This work is supported by the National Natural Science Foundation of China (11801294).

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] Wang, Y. (2015) Time Series Analysis with R.

[2] Ding, Z., Granger, C.W.J. and Engle, R.F. (1993) A Long Memory Property of Stock Market Returns and a New Model. *Journal of Empirical Finance*, **1**, 83-106. https://doi.org/10.1016/0927-5398(93)90006-D

[3] Hipel, K.W. and Mcleod, A.I. (1978) Preservation of the Rescaled Adjusted Range: 2. Simulation Studies Using Box-Jenkins Models. *Water Resources Research*, **14**, 509-516. https://doi.org/10.1029/WR014i003p00509

[4] Bollerslev, T. (1986) Generalized Autoregressive Conditional Heteroskedasticity. *Journal of Econometrics*, **31**, 307-327. https://doi.org/10.1016/0304-4076(86)90063-1

[5] Carolyn, O., Betuel, C. and Pitos, B. (2018) Modeling Exchange Rate Volatility Using APARCH Models. *Journal of the Institute of Engineering*, **14**, 96-106. https://doi.org/10.3126/jie.v14i1.20072

[6] Jonathan, D.C. (2011) Time Series Analysis with Applications in R.

[7] Wang, Z.J. and Wang, B.H. (2014) An Improved Time Series IO Type Outlier Detection Method. *Statistics & Decision*, **22**, 4-6.

[8] Charles, A. and Darné, O. (2005) Outliers and GARCH Models in Financial Data. *Economics Letters*, **86**, 347-352. https://doi.org/10.1016/j.econlet.2004.07.019

[9] Wang, X.M. (2008) Misunderstanding of the Concepts of Skewness and Kurtosis. *Statistics and Decision*, **12**, 145-146.