Time Series Modelling towards Netflix Stock Price
Project Report, STA457: Time Series Analysis

Department of Statistics, University of Toronto
Huifeng Wu∗

December 17, 2020

1. Data Description

In presence of the COVID-19 epidemic, stock investors have witnessed tremendous fluctuations in the market. Here I collected adjusted closed stock price of Netflix Inc. (ticker: NFLX) over the past 243 trading days, namely year-to-start data from Jan 1, 2020 to Dec 16, 2020. This report proposes a couple of statistical models to interpret the time series, after which a final model is derived through comprehensive model diagnostics. Eventually, model prediction and spectral analysis are preformed. The dataset is publicly available from Yahoo Finance:
https://ca.finance.yahoo.com/quote/NFLX/history?period1=1576555791&period2=1608178191&interval=1d&filter=history&frequency=1d&includeAdjustedClose=true

2. Data Visualization and Transformation

Figure 1 (a) illustrates that the original series y_t has a strong upward trend, despite a large dip around 50th trading days (March). NFLX seemed to recover from the recession quickly, which might be a result of quantitative easing policies by the government. Nevertheless, it is still challenging to investigate other variability within data, as demonstrated by the slow decay pattern of its sample ACF in Figure 1(b).

It is empirically found that logarithm transformation, followed by two times of differencing, stabilize the data. In fact, the transformed data, say $\nabla^2 log(y_t)$, resembles the variation of the stock growth rate. Both Figure 1(c) and (d) confirm the transformed data as a stationary process with constant mean, and its ACF and PACF decay with higher time lags.

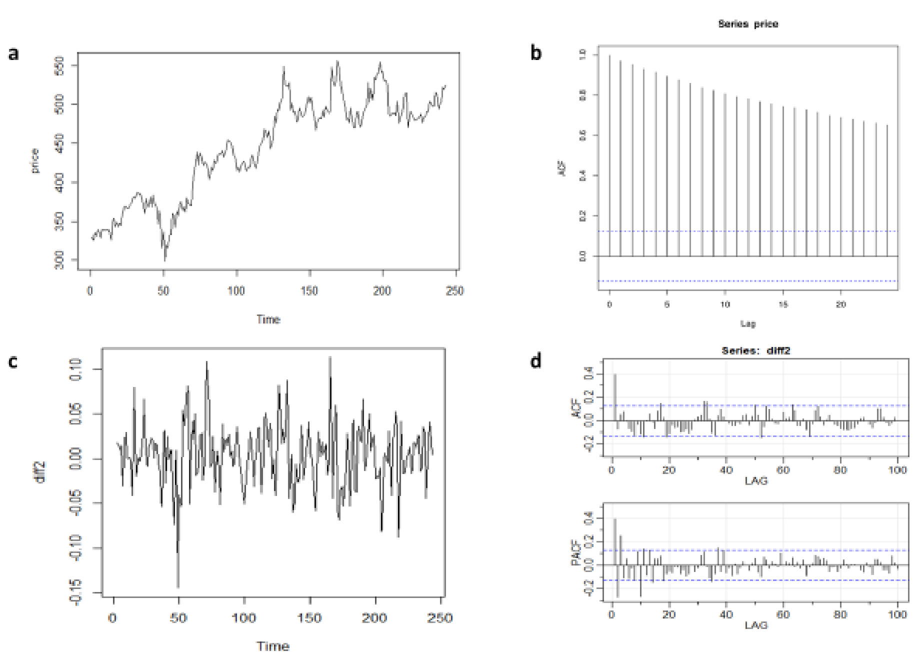

Figure 1. Data plotting: (a)-(b) original data y_t and its sample ACF (c)-(d) transformed data $\nabla^2 log(y_t)$ and its sample ACF and PACF

3. Proposed Models and Dependence Orders

Prior to modelling, it is crucial to identify the dependence orders based on the ACF and PACF plots. Particularly, ACF values cut off at lag 1, whereas the PACF is susceptible of cut-off at lags 3 and 10. For simplicity, two different models are suggested; the models are ARIMA(10,2,1) and ARIMA(3,2,1). Note the input is $log(y_t)$ where second difference would be made.

4. Parameter Estimation

By sarima() function in R, the estimation of parameters is presented in Table 1.

Model	Coefficient Estimates
ARIMA(10,2,1)	**ar1**: -0.2400 **ar2**: 0.0350 **ar3**: -0.0094 **ar4**: 0.1057 **ar5**: 0.0106 **ar6**: -0.0671 **ar7**: -0.0167 **ar8**: -0.1591 **ar9:** 0.0206 **Ar10**: -0.0133 **ma1**: -1.0000
ARIMA(3,2,1)	**ar1**: -0.2524 **ar2**: 0.0467 **ar3**: -0.0269 **ma1**: -1.0000

Table 1. Model parameter estimation

5. Model Diagnostics and Final Model Choice

For model ARIMA(10,2,1) evaluated in Figure 2(left), the standardized residuals oscillate around zero like white noise, where no pattern can be observed. Since there are a few outliers among the data on the 50^{th} trade day, short volatile fluctuations can be found at the corresponding period. The ACF of residuals indicates no significant correlation. Moreover, I examine the normality of the standardized residuals by using QQ plot. It is concluded that they follow a normal distribution that residual points align well with the straight line, although only several end points are out of the significant regions. Finally, I present p-values for Ljun-Box statistics, by which all values are below the significance level. It exhibits great correlation between data, and thus lack of fit for the model.

For model ARIMA(3,2,1) diagnosed in Figure 2(right), the trend is fairly random with mean 0; the pattern is similar to that of the former model, yet a bit smoother. The ACF of the standardized residuals indicates no significant departure from the significance level. Furthermore, the standardized residuals fall within the confident intervals decently, with a few outliers at tail portions in the QQ plot. The normality of standardized residuals is still satisfied here. Last, we can see the p-values for Ljun-Box statistics are quite different from ARIMA(10,2,1). Nearly a quarter of them are insignificant and the rest are in line with the significance level, which is tolerable.

In addition to above tests, Table 2 shows the size of penalty terms for AIC, BIC and AICc, which help us obtain a final model. All three metrics support the second model. I then select ARIMA(3,2,1) as the final model, since it also fits the data better by the pervious diagnostic procedures.

Models	AIC	AICc	BIC
ARIMA(10,2,1)	-4.204731	-4.199948	-4.031214
ARIMA(3,2,1)	-4.218987	-4.218284	-4.146688

Table 2. AIC, BIC and AICc for model candidates

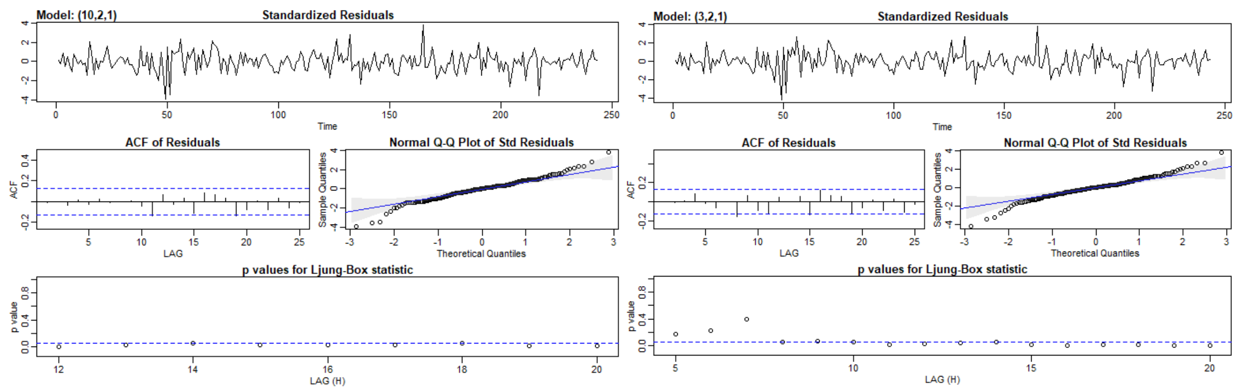

Figure 2. ARIMA(10,2,1) model (left) and ARIMA(3,2,1) model (right) diagnostics

6. Prediction

Using the final model, I forecasted $log(y_t)$ on the next four trade days. Figure 3 shows an increasing trend, which implies the bullish sentiment towards the future price of NFLX as well. The predicted values and 95% confidence intervals for the forecasted values are shown in Table 3.

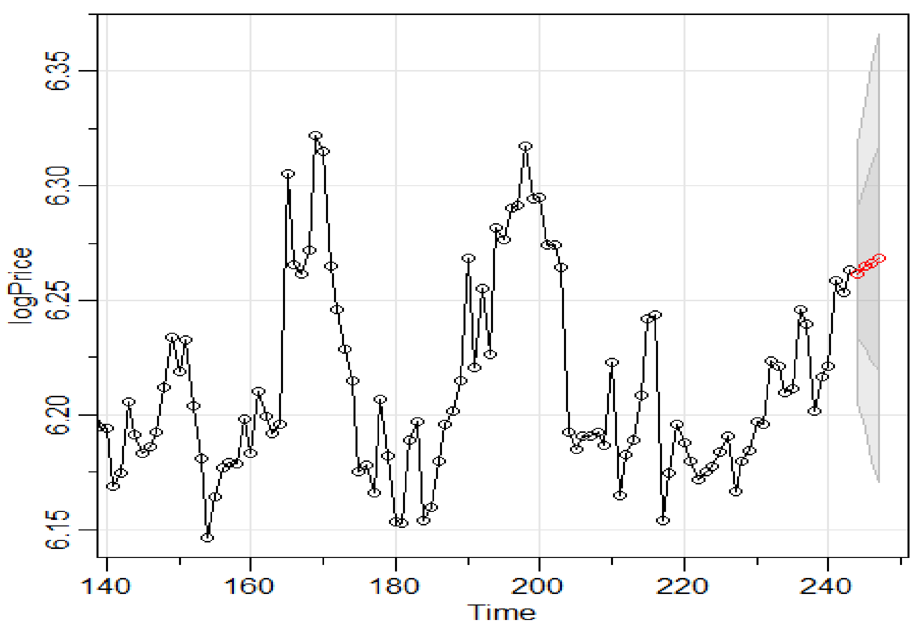

Figure 3. Forecasting using ARIMA(3,2,1)

Time	Estimate	Lower	Upper
t+1	6.261763	6.205997	6.317529
t+2	6.265054	6.195290	6.334817
t+3	6.266274	6.181452	6.351095
t+4	6.268526	6.172633	6.364420

Table 3. Prediction on next 4 trade days from Dec 17, 2020 (inclusive) and 95% confidence intervals

7. Spectral Analysis

Figure 4 plots the periodogram based on $\nabla^2 log(y_t)$, in which the dominant frequencies occurred at ω = 0.1811. The predominant period is then 5.5227, accordingly. All first three dominant frequencies are given in Table 4, alongside their 95% confidence intervals. However, the confidence intervals are too narrow and small that cannot be used to evaluate significance.

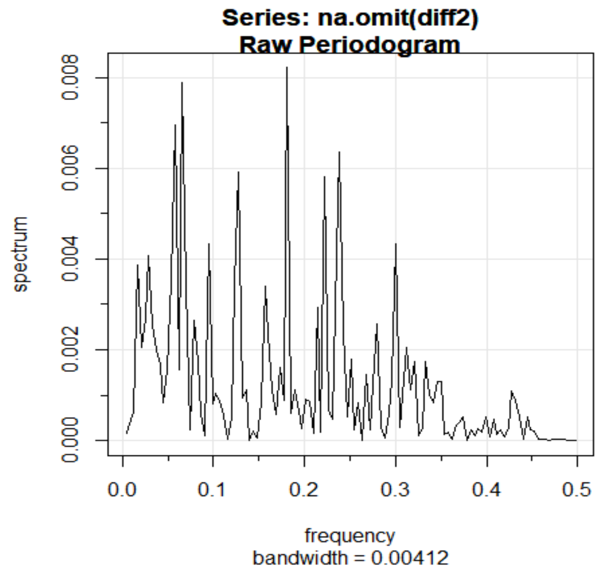

Figure 4. Forecasting using ARIMA(3,2,1)

Dominant Freq	Spectrum	Lower	Upper
0.1811	0.0082	0.002222897	0.3238827
0.0658	0.0079	0.002141572	0.3120333
0.0576	0.0069	0.001870487	0.2725354

Table 4. First three dominant frequencies (from top to bot), spectrums and 95% confidence intervals

8. Discussion

To summarize, this report initially suggested two model ARIMA(10,2,1) and ARIMA(3,2,1) to interpret NFLX stock price in log scale. It is interesting that data on stock price typically have strong correlation, as shown by its ACF plot. Transformation and differencing are usually required.

For parameter estimation, ARIMA(10,2,1) naturally has more number of parameters than ARIMA(3,2,1), which has a risk of overfitting. To obtain the final model, necessary diagnostics were performed. It is particularly worth noting that ARIMA(3,2,1) shows less significant Q-statistics for given lags, even though other tests are equivalently satisfied by both of the models. Eventually, ARIMA(3,2,1) was chosen as the final model. Meanwhile, the criterion such as BIC and AIC also confirmed such a decision.

Using the final model, I forecasted $log(y_t)$ on the next four trade days. From the previous output figure, the current bullish trend will continue, yet slowly. All the values are approximately 6.26. Although spectral analysis was conducted and we computed some dominant frequencies, the confidence intervals were not reasonable enough to offer useful insight for their respective spectrums.

So far ARIMA(3,2,1) seems to be the best model to describe the underlying data, but it still indicates some lack of fit, especially if we narrow down our focus on its Q-statistics of standardized residuals. It is of importance to explore other models and transformation techniques with more collected data.