Time Series Analysis: Project 2

Due on May 18, 2021

Professor Tianwei Yu

Peng Deng

Contents

1 Problem statement

Using the `getSymbols()` command in the R package `"quantmod"`, you can retrieve the stock price of Alibaba at New York Stock Exchange (NYSE) between 2015-01-01 and 2019-12-31. Please use the daily closing price to conduct the following analyses.

(1) Explore the data. Comment on data characteristics. Is any transformation necessary?

(2) Try to fit an arima model to the data.

(3) Using the residual of the model you select from step 2, examine if there is change in variance over time. If the answer is yes, fit a model to describe the behavior of the residuals.

(4) Using the model from step 2, forecast the next month's data, and compare the forecast results to real data of 2020 .

(5) Using the differential data of the closing stock price, fit a Hidden Markov Model with 2 hidden states. Does there appear to be two distinct states?

(6) If you first take log of the price, observe the data and then repeat the analysis in (5), does your conclusion change? Why? If there is any difference, which analysis do you think is more meaningful?

(7) Write a report to document your analysis and summarize your findings. Attach your code at the end of the report.

2 Data overview

Firstly, we can have a quick look at the dataset, we plot the data as Figure 1. As we can see from Figure 1, there is not seasonal variation in the time series data, and the main trend of the data is increasing. More specifically, we can see that the trend in the first three years is increasing but not linear, and the data in the last two years changes dramatically. Thus, we can take **log transformation** to make the trend in the first three years more linear and to reduce the variation in the last two years.

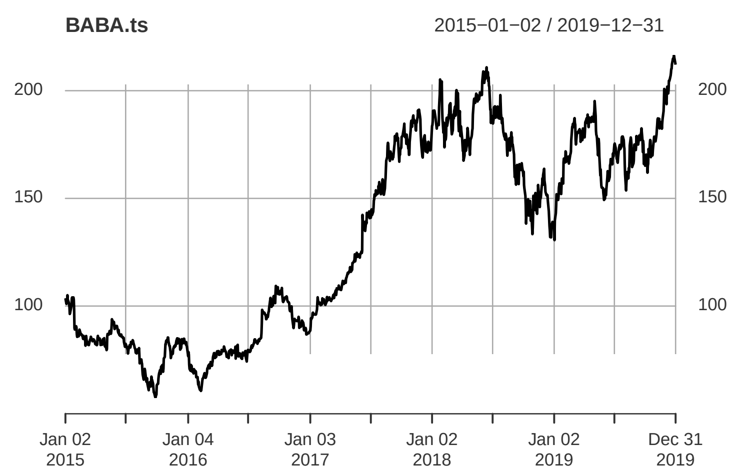


Figure 1 The stock price of Alibaba between 2015-01-01 and 2019-12-31

3 Fit the ARIMA model and do forecast

In this section, we would like to fit an ARIMA model to the stock price of Alibaba and do forecast.

3.1 Fit the ARIMA model

In order to determine the parameters in ARIMA, we call "auto.arima" in R as follow

```
BABA.arima = auto.arima(BABA.ts)
```

Then, we get the parameters of ARIMA is: ARIMA(2,1,2). The fitted figures are as Figure 2.

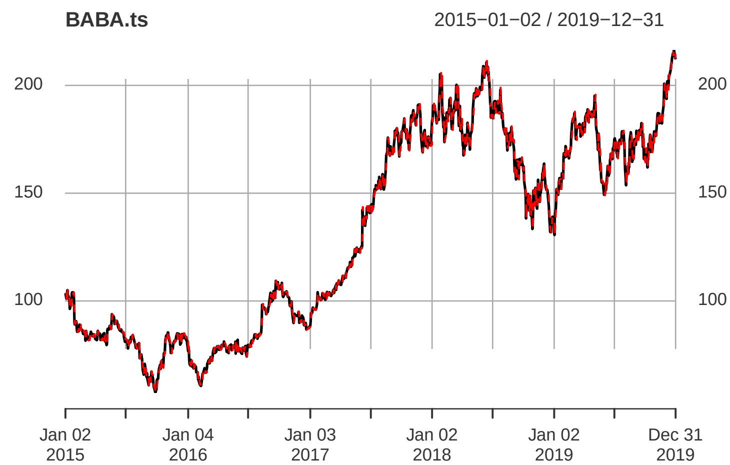


Figure 2 ARIMA model fitted figure

Then, we plot the residuals and the ACF figure of the residuals according to the ARIMA model result as Figure 3, which means we are going to check residuals. As we can see from Figure 3, there is significant acf, so that the model is not very good. However, we can do the Ljung-Box test of the residuals to see if the residuals exhibit serial correlation, the result is as follow:

```
        Ljung-Box test

data:  Residuals from ARIMA(2,1,2)
Q* = 6.0748, df = 6, p-value = 0.4149

Model df: 4.   Total lags used: 10
```

As we can see from the Ljung-Box test, the p-value equals 0.4149, which is larger than 0.05, so that we should not reject the null hypothesis, which means the residuals does not exhibit serial correlation. Thus, we can believe that the ARIMA model is suitable.

3.2 Do the forecast

Then, we would like to forecast the next month's data using the ARIMA model we fitted. As we can see from Figure 4, the forecast is not good, and the value of the forecast is almost constant and it is far away from the true value. We zoom in the Figure 4(a) as Figure 4(b) in order to see clearly and do comparison. Thus, maybe the ARIMA model is not good enough to do the forecast in this problem.

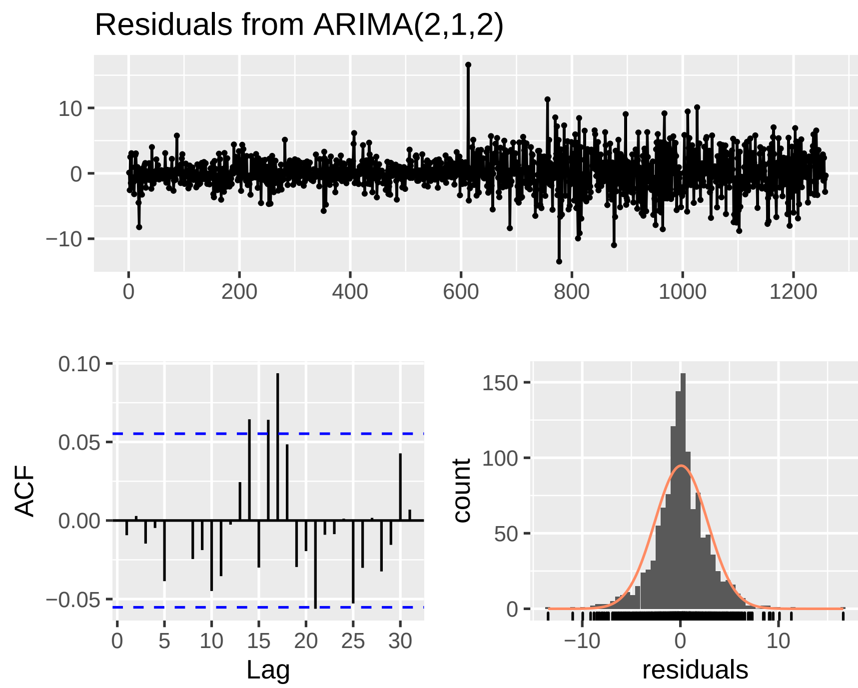
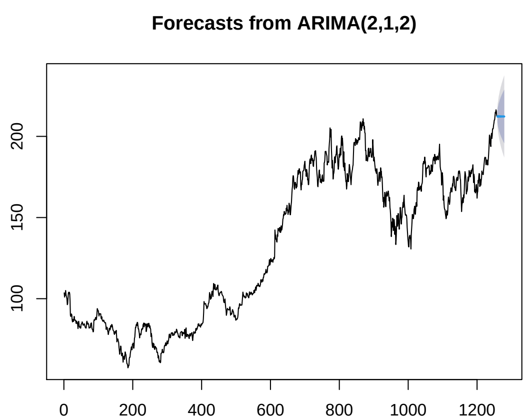
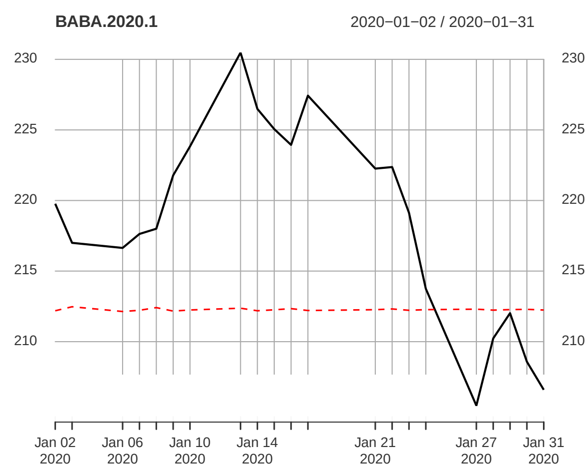


Figure 3 Check residuals of ARIMA model



(a) The forecast



(b) The comparison of forecast and true data

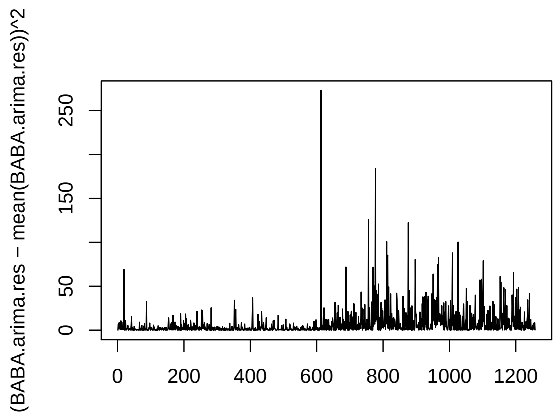
Figure 4 The forecast of the next month

4 Check the residuals

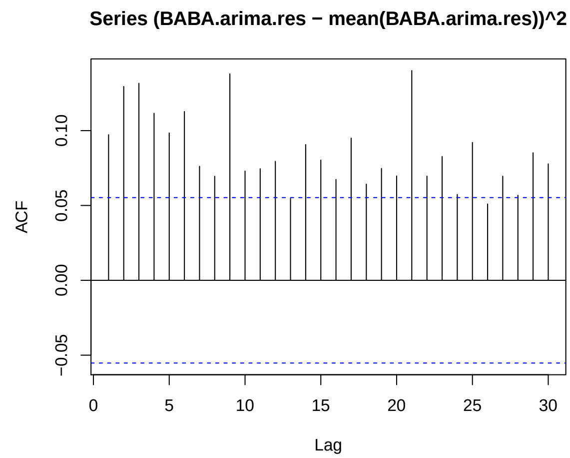
In this section, we would like to deal with the residuals of the ARIMA model in Section 3.

4.1 Examine the change in varaince

As we can see from Figure 3, the variance of residuals seems not constant over time. Then, we can plot the correlogram of the squared values to detect the volatility as Figure 5. As we can see, the variance is correlated in time, since the series exhibits volatility which can be derived from Figure 5(b). Thus, we can see that there is change in variance over time.



(a) The squared residuals



(b) The acf of the squared residuals

Figure 5 The squared value of residuals time series

4.2 Fit the residuals

We fit the residuals with GARCH(1,1) model, the series with fitteds standard deviation is as Figure 6. And the time sequence plot of the estimated conditional variance is as Figure 7.

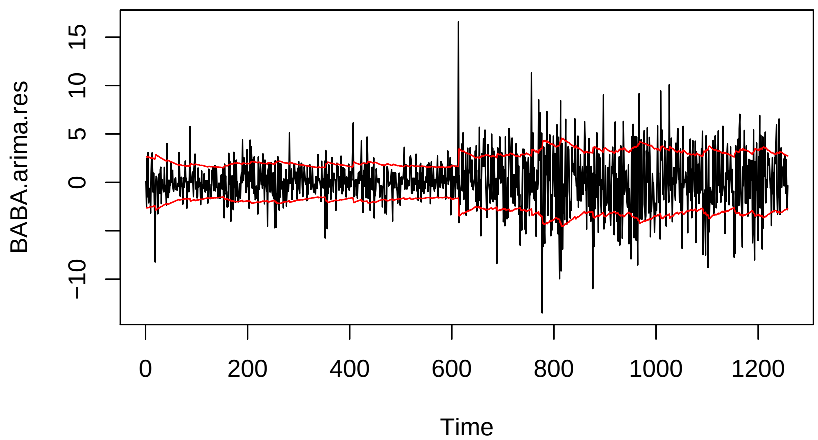


Figure 6 The residuals with GARCH(1,1) model fitted standard deviation

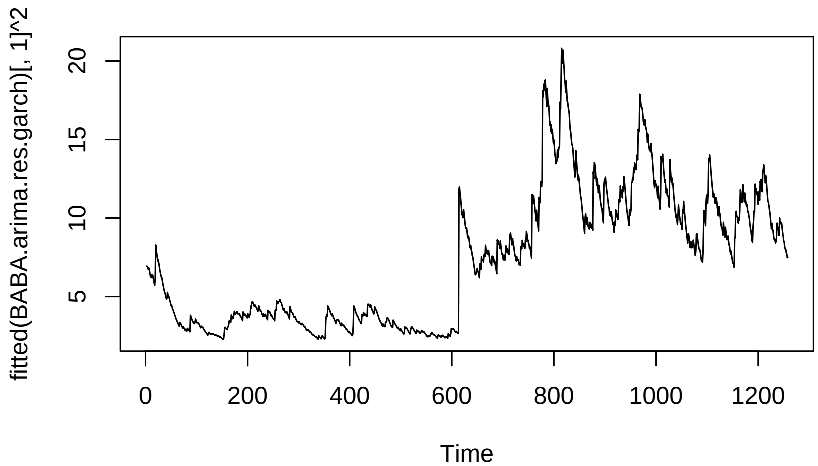


Figure 7 The time sequence plot of the estimated conditional variance

Then, we would like to check if the model have addressed the conditional heteroskedasticity. As we know, if all estimated conditional variance captures the conditional variance, the residuals will be white noise with mean 0 and variance 1.

Thus, we would like to check the residuals of GARCH(1,1) model. The test result is as follow and the test figures are as Figure 8.

```
Diagnostic Tests:
        Jarque Bera Test

data:  Residuals
X-squared = 4475.9, df = 2, p-value < 2.2e-16

        Box-Ljung test

data:  Squared.Residuals
X-squared = 0.11738, df = 1, p-value = 0.7319
```

As we can see from the Jarque Bera Test result, the p-value $< 2.2e^{-16}$, which means the distribution of residuals is not normal. On the other hand, as we can see from the Box-Ljung test result, the p-value equals 0.7319, which means there is no serial autocorrelation in residuals. With the ACF figure of residuals in Figure 8, we ca derive that the residuals can be treated as white noise, which means the GARCH(1,1) model has addressed the conditional heteroskedasticity, although the distribution of residuals is not normal.

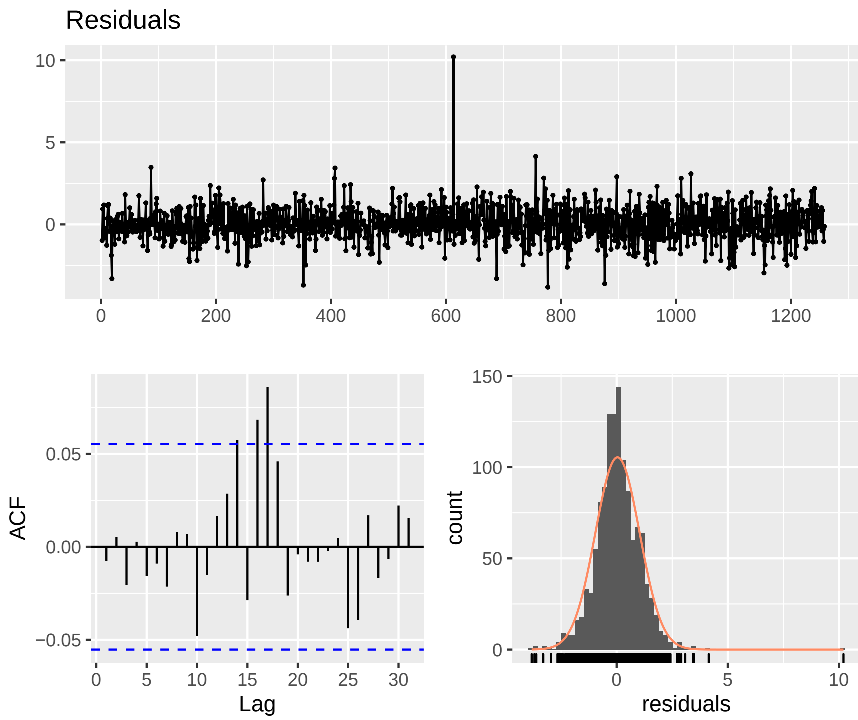


Figure 8 The time sequence plot of the estimated conditional variance

5 HMM model

5.1 Do not take log to the original data

Firstly, we would like to take differential to the closing stock price, the differential data of the closing stock price is showed as Figure 9. As we can see from Figure 9, it seems that there are two hidden states in the series.

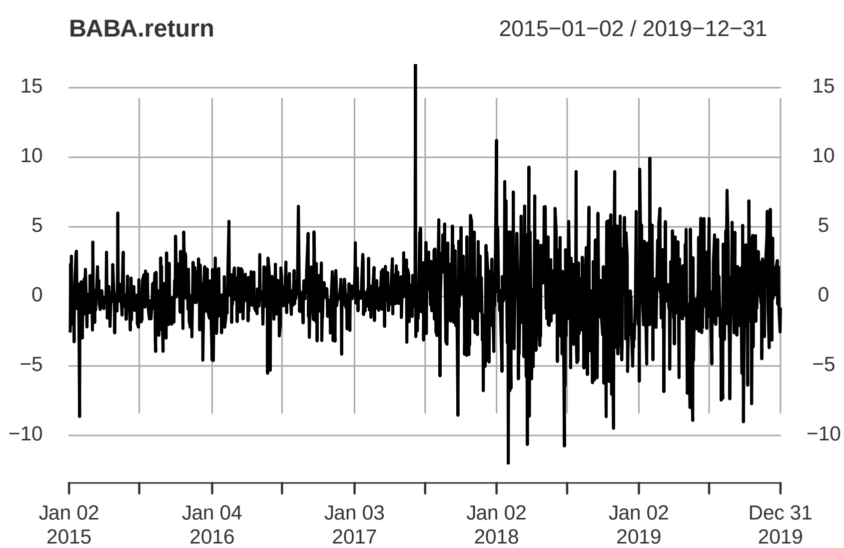
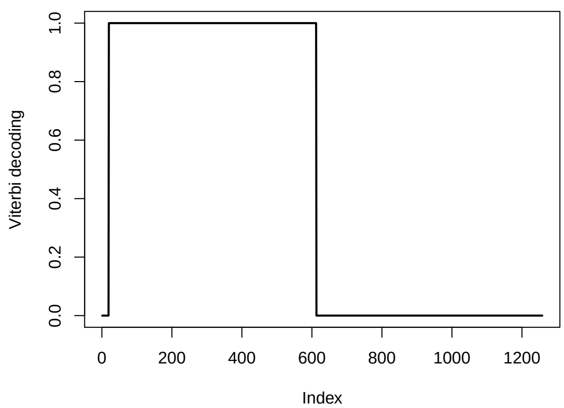
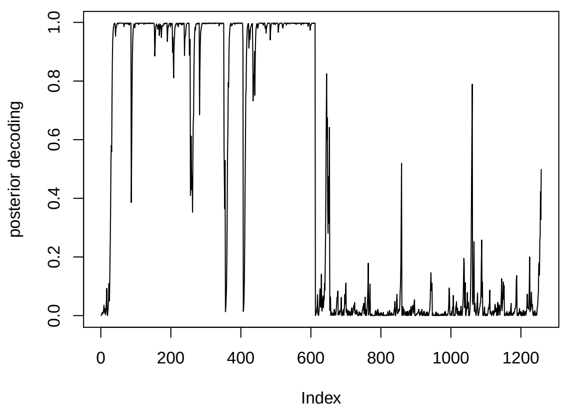


Figure 9 The differential data of the closing stock price

Then, we fit a Hidden Markov Model with 2 hidden states. The result is as Figure 10.



(a) The viterbi decoding



(b) The posterior decoding

Figure 10 The HMM model result

As we can see from Figure 10(a), the Viterbi decoding tells us there appear to be two distinct states, which meets our expectation as we mentioned before.

5.2 Take log to the original data

Firstly, we take log to the original price data, then the log price is as Figure 11.

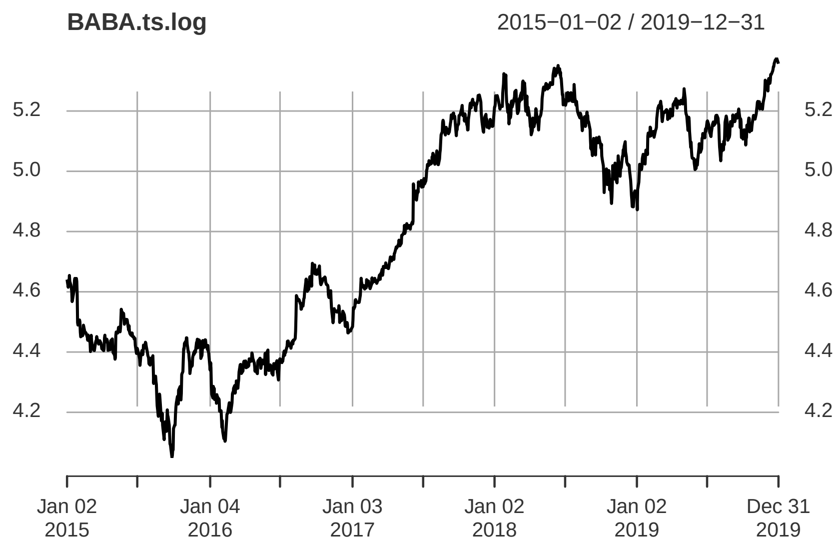


Figure 11 The log data of the closing stock price

Then, we again take differential to the log closing stock price, the differential data of the log closing stock price is showed as Figure 12. As we can see from Figure 12, it seems that there are not distinct hidden states in the series.

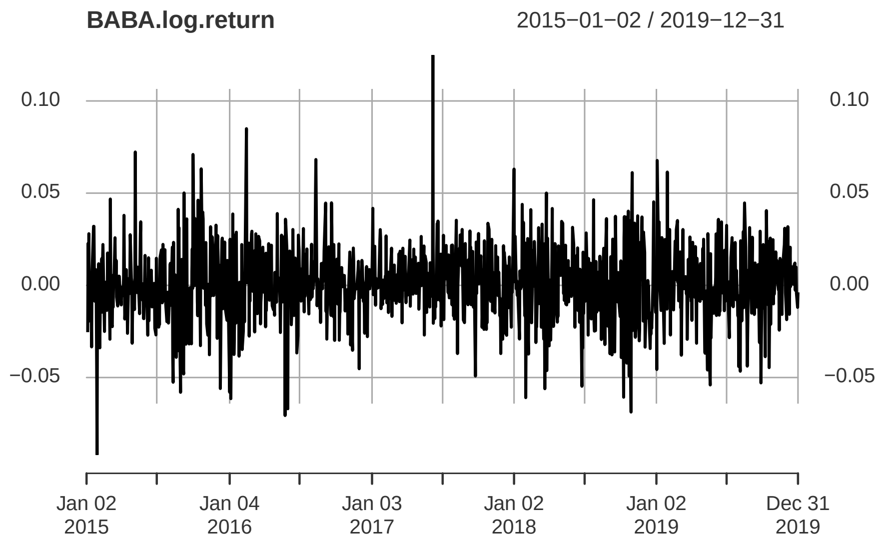
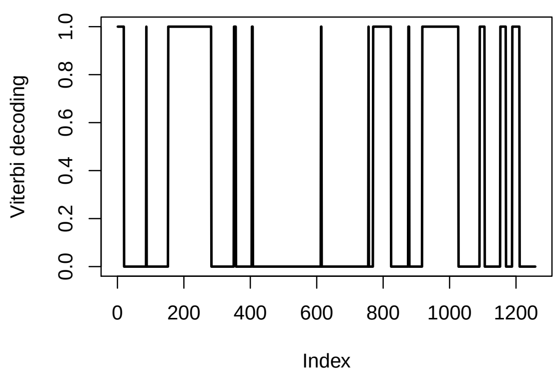
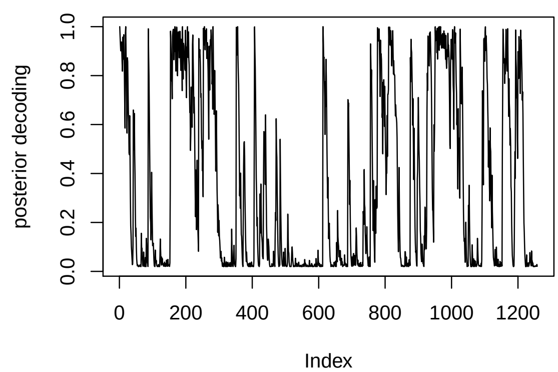


Figure 12 The differential data of the log closing stock price

Then, we fit a Hidden Markov Model with 2 hidden states. The result is as Figure 13.



(a) The Viterbi decoding



(b) The posterior decoding

Figure 13 The HMM model result

As we can see from Figure 13(a), there does not appear to have two distinct states because the viterbi decoding varies very frequently.

Thus, we can derive that after taking log to the original price data, the conclusion changes. There are no longer two distinct state is the return series. I think the analysis "Do not take log to the original data" is more meaningful, because in this analysis we can find two distinct states, which means the method is good to capture the feature of the original data.