DataH195a Project

Zehao Zhao

August 31, 2021

1 Introduction and Motivation of the Project

This project will primarily be focus on the relationship between Chinese stock markets, and the relationship with its counterpart in Hong Kong and US stock market. Usually, when a policy being make, or a news on the company has been published, there will be ups and downs on the consumer's confidence toward one company. However, according to my past observation, groups of stockholders from China mainland, Hong Kong, and the US share different thought process sometimes toward certain news. I want to use data to check the relationship between them. The motivation of doing this project is bacause I started my own investment during covid, and it works out great, my position almost doubled in less than one year, and I want to explore about investments. One of the technique I used is to observe chinese and hong kong company and then make a decision on whether to short or long the investment in its counterpart in the US stock market. I found it can be a potential direction to research and find the relationships for systematical investment.

2 Dataset Description and Exploratory Data Analysis

2.1 Dataset Description and Dataset Gathering

The dataset comes from Tushare API, which is a company that provides API for daily stock market prices. It has provided databases on the stock market around the world. I have already written codes to incorporate the dataset into my code. Once I have provided it with the token number of the stock, it will generate information such as open price, highest price, lowest price, low price, the volume of transaction, etc.

2.2 Dataset Pre-processing

I implemented data pre-processing on python jupyter notebook, please check out my github repository for more information about codes. First, I gather data for one particular Chinese stock Jiangshan Oupai (just an example of my investment) I used the build in function to get the daily trading data, using regex to change the string to timestamp format, and then join with another table that contains the full calendar. The reason of doing so is because not everyday is trading day. The market will close during weekend and holiday. I treat weekend and holiday as no trading, and therefore the price level off. I also add which day of the week from 1-7 and new format of date for drawing in the dataframe.

2.3 Exploratory Data Analysis

We then use the periodogram to explore the frequency of oscillation more in details. From the left panel of Figure 1, we can notice three dominant frequency. The first one is the Three-week frequency that we have already discussed about in the overall trend, The second is the weekly (5 day) frequency, and third one is

Figure 1: Jiangshan Stock Price Over Time

the 3-week frequency. Moreover, figure 1 shows the effects of different days of the week. Furthermore, it is clear to see that as the mean increases in the time series, the variance also increases, so it is reasonable to use log VST to stabilize the data. From this point, we will be working with log(open) as we proceed with our modelings.

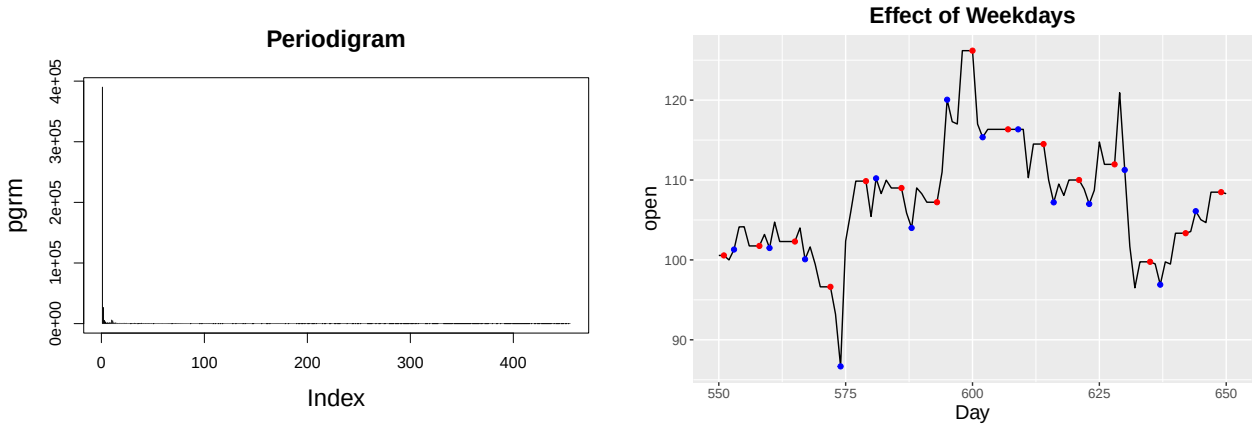

Figure 2: The left panel is the periodigram of open prices. The right panel shows the comparison of new price on Saturnday and Monday. The red dots and blue dots are correspond to Saturnday and Monday. respectively

3 Methodology Description

To model the natural signal in this data, both a parametric model and a differencing approach are used.Both of these models of the signal will be complimented with ARMA models for the remaining noise.

3.1 Differencing

First, as we previously addressed, there are 3 different dominant frequency as we look at the periodigram, 1, 5, and 20. Since it does not make sense to do a differencing with lag of 1 (the frequency is too small), so we will only use lag 5 and 20 in our model. To pursue stationary, we first take a difference with lag of 5 to

get rid of weekly oscillation, then we take a lag of 20 to get ride of three-week oscillation, and by taking the two seasonal difference we can also get rid of quadratic trend. Finally, we can take a first order difference in order to get rid of potential cubic trend, as the rate of price increase or decrease.

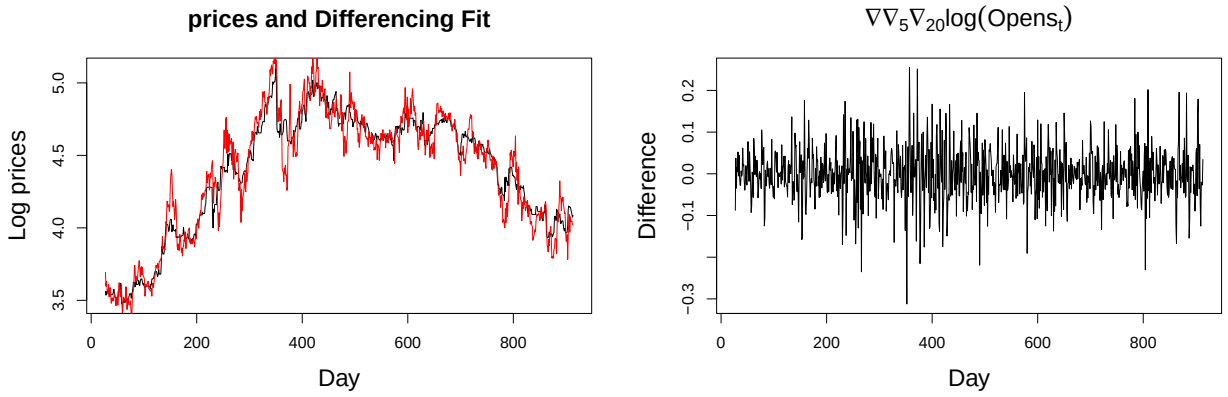

Figure 3: Diagnostics for differencing signal model. The left panel shows the fit of differencing model(red). The plot shows the difference value, to be assessed for trend and seasonality

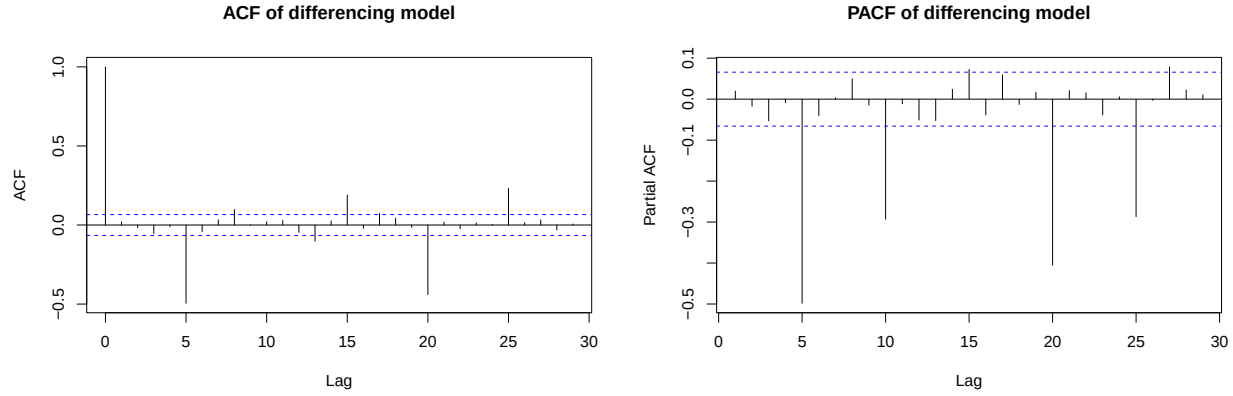

Figure 4: Autocorrelation function (ACF) and partial autocorrelation function (PACF) values for the differencing model.

3.1.1 Differencing with ARMA(0,8)

First, there are no clear cutoff in the PACF plot, it is more like an exponential decaying trend, so it is clear that q does not equal to 0. As we observe the ACF plot, there is a clear cutoff at lag 8, after which most ACF values are within the white noise C.I. Therefore, an ARMA(0,8) model would be used. Most p-values are above 0.05 according to the Ljung-Box plot in figure 4, so it can be considered to be a good fit.

3.1.2 Differencing with ARMA(0,1)

First, there are no clear cutoff in the PACF plot, it is more like an exponential decaying trend, so it is clear that q does not equal to 0. As we can observe from the ACF plot, there is high magnitude of auto-correlation at lag 1, then follows with insignificant values. Although there are some high auto-correlation magnitudes

at lag 5, 7,8, but they might just happen by chance. Therefore, an ARMA(0,1) model would be used. The first few p-values are above 0.05 according to the Ljung-Box plot in figure 4, it is not the most ideal fit, but it is relatively a good fit compare to other SARIMA models in terms of p-value.

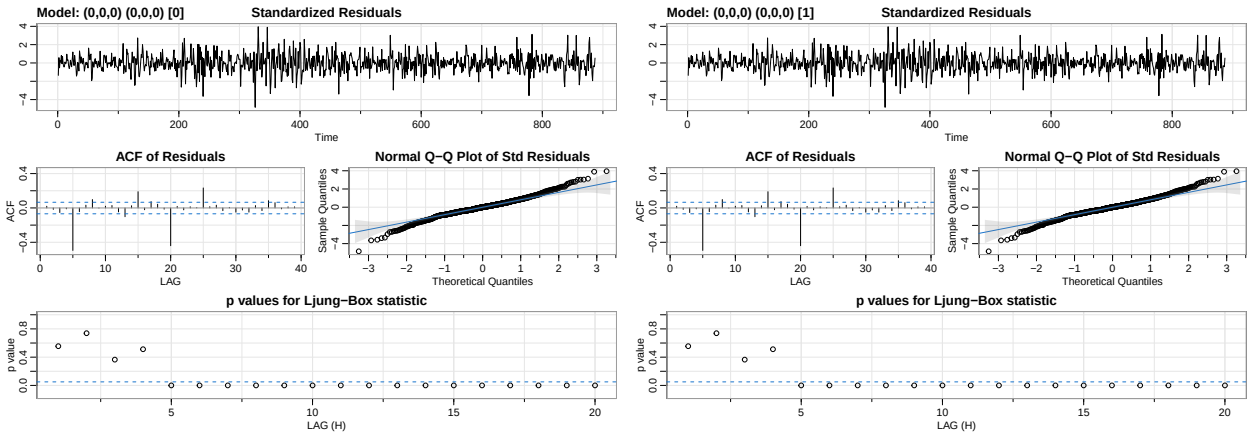

Figure 5: The left panel shows the sarima() function output for the fit of ARMA(0,8) model for differencing model's residuals. The right panel shows the sarima() function output for the fit of ARMA(0,1) model for differencing model's residuals. We will use the Ljung-Box plot to exam the fitness of our ARMA model.

3.2 Parametric Signal Model

For Parametric model, we want to create sinusoids and indicator variables to capture the seasonality, also a parametric curve to capture the trend. First, we still apply the log VST to stabilize the variance. In regard to trend, I am using a fourth degree polynomial equation to capture the feature that price curve has the trend of increasing to decreasing to increasing and then to decreasing again. To capture the weekly seasonality, I use indicator variables for each day of th week. To capture the semi-weekly and Three-week seasonality, sinusoids will be employed.

Figure 5 presents the fit as well as the residuals, which appear reasonably stationary.

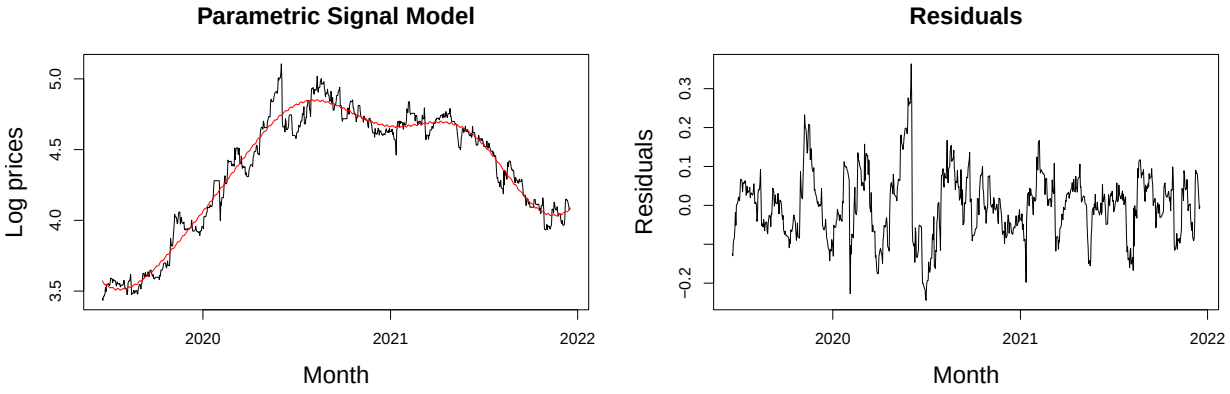

Figure 6: The parametric signal model. The left panel shows this model's fitted values in red, plotted new open prices data in black. The right panel shows the residuals of this model.

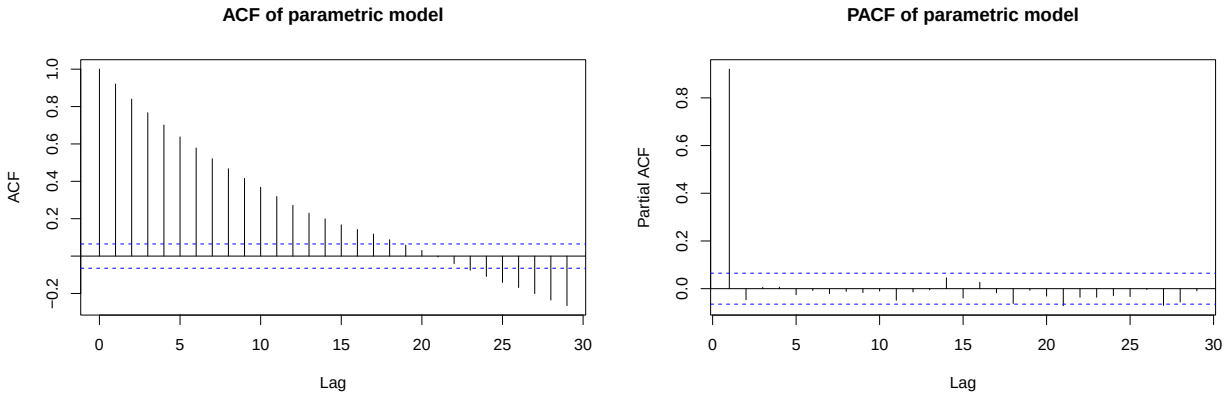

Figure 7: Autocorrelation function (ACF) and partial autocorrelation function (PACF) values for the parametric model.

3.2.1 Parametric Signal with ARMA(1,0)

The ACF plot has a exponential decaying trend which suggests non-zero p value, and PACF plot suggests a clear cut-off at lag 1, which indicates a zero q value, and a p value of 1. Hence, an ARMA(1,0) model would be a reasonable fit for the noise. According to the Ljung-Box plot in figure 7, all p-values above 0.05, so it can be considered as a good fit.

3.2.2 Parametric Signal with ARMA(2,1)

This second noise model will be chosen with the R function, auto.sarima(), which generates a result of ARMA(2,1). According to the Ljung-Box plot in figure 7, all p-values above 0.05, so it can be considered as a good fit.

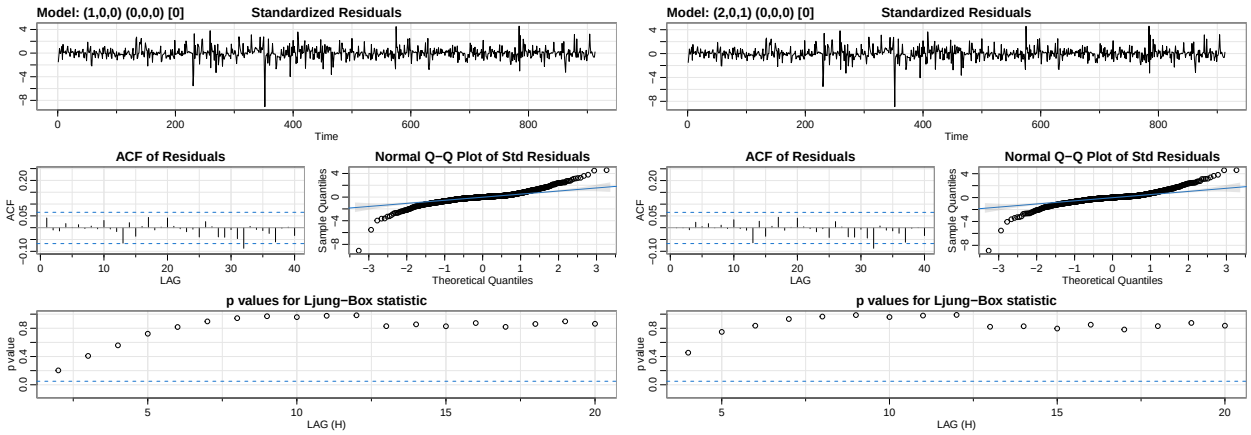

Figure 8: The left panel shows the sarima() function output for the fit of ARMA(1,0) model for parametric model's residuals. The rightanel shows the sarima() function output for the fit of ARMA(2,1) model for parametric model's residuals. We will use the Ljung-Box plot to examine the fitness of our ARMA model.

4 Model Selection and Proposed outcomes

These four model options are compared through time series cross validation. The non-over lapping testing sets roll through the data until 12/17/2021, in 10 day segments. The training sets consist of all data that occur before the appropriate testing set. The models' forecasting performances will be compared through root-mean-square prediction error (RMSPE). Although the RMSPE is in term of log prices, but it still provides the best selection. The model with the lowest RMSPE will be chosen as the model for prediction of the stock price for the next 10 days (weekends included). Table 1 shows that the differencing model with ARMA(0,1) has the lowest cross-validated forecasterror.Thus the differencing model with ARMA(0,1) is the chosen forecasting model.

Table 1: Cross-validated out-of-sample root mean squared prediction error for the four models under consideration.

	RMSPE
Parametric Model + ARMA(1,0)	0.3766187
Parametric Model + ARMA(2,1)	0.3767884
Three-week Differencing + Weekly Differencing + First order Differecing ARMA(0,1)	0.0580801
Three-week Differencing + Weekly Differencing + First order Differecing ARMA(0,8)	0.0629983

5 Results and Relevance of the Project

5.1 Results

A Three-week differencing + Weekly differencing + First order differencing model with ARMA(0,1) noise will be used to forecast. Let Y_t represents the new prices at time t with addictive noise term X_t. X_t is a additive noise term, which is a stationary process defined by ARMA(0,1). W_t is defined as white noise with variance σ_W^2. In the end, we need to apply exp() function to transform $\log(Y_t)$ into Y_t which is the forecasting that we want.

$$\begin{aligned}\log(Y_t) = &\log(Y_{t-5}) + \log(Y_{t-20}) - \log(Y_{t-25}) \\ &+ \log(Y_{t-1}) - \log(Y_{t-6}) - \log(Y_{t-22}) + \log(Y_{t-26}) \\ &+ X_t\end{aligned} \tag{1}$$

$$X_t = W_t + \theta W_{t-1} \tag{2}$$

$$Y_t = \exp(\log(Y_t)) \tag{3}$$

5.2 Prediction

Figure 10 shows the forecast value for stock price 12/18/2021 to 12/28/2021.

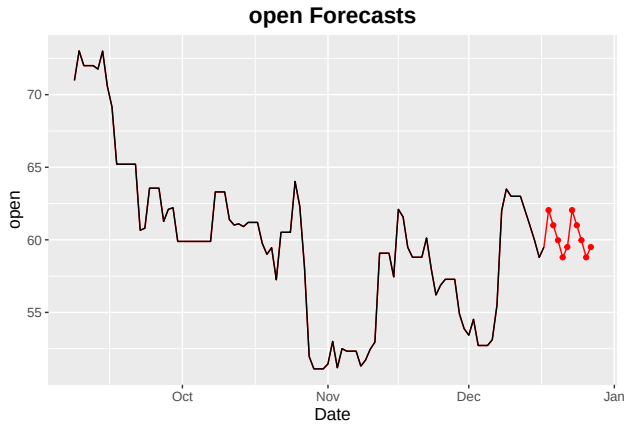

Figure 9: Forecasts of open prices. The x-axis is time in months. The black line is the recent historical new open prices data. The red points are the forecasts for 12/18/2021 to 12/28/2021.

5.3 Relevance

In this project, we successfully give a prediction on the next ten days stock price change. The relevance is that, after I got familar with time series algorithms, I can then employed it on multiple stock markets and analyze their difference in prices change within a time frame. In addition, the fourier transformation and signal processing can also be applied on other markets, such as, crypto and ETFs.

6 Blind spots and Ethical Issues

There are lots of blind spots. Since companies are different if they have they are registered in different areas, we cannot assume both branches are the same. In addition, there are lots of companies that are duel-listed; I need to traverse lots of companies to give a statistically significant conclusion. What's more, analyzing markets only in China, Hong Kong, and the United States can be a biased sample, I need to be very careful on drawing the association, correlation, causal inference, and conclusion.

7 Acknowledgement and Reference

I acknowledge Professor Eric Van Dusen's help on guiding me through completing the project step by step from data pre-processing to the final model and ethical analysis. I also gained example code on time series analysis from stat 153 class projects and textbooks.