Emily Mai
Brian Tran
SMART Program
07 June 2021

Time Series Analysis with Causal Discovery

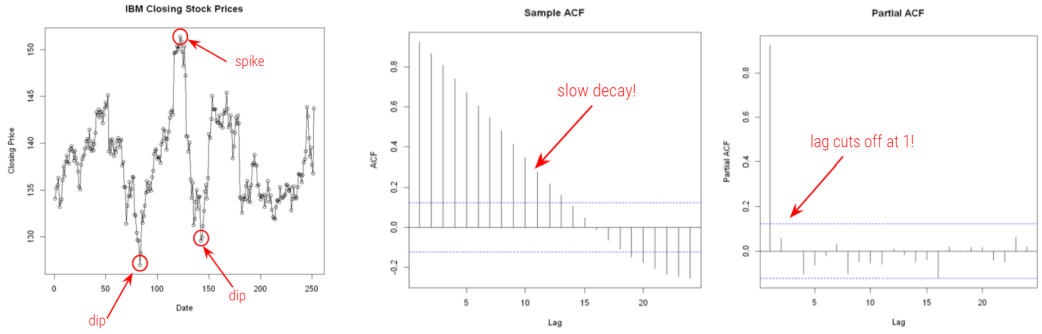
Introduction

International Business Machines Corporation (IBM) is an American multinational technology company based in New York and founded in 1911, with operations in over 170 countries. In addition to providing integrated solutions and services, IBM also produces and sells computer hardware, middleware, data platform software.

Data - The present study uses the stock market dataset obtained through Kaggle where each stock contains common features such as: date, opening, high, low, close, adjusted close, and volume. For the purpose of the study, the IBM stock was used, specifically the closing prices from 2019 to 2020.

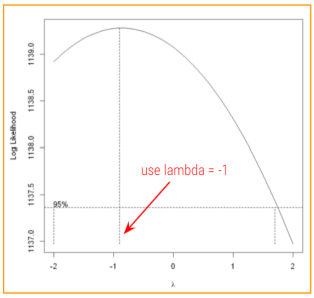
Problem Statement - The goals of the study are 1) identify a possible model for IBM dataset and 2) forecast future IBM stock prices.

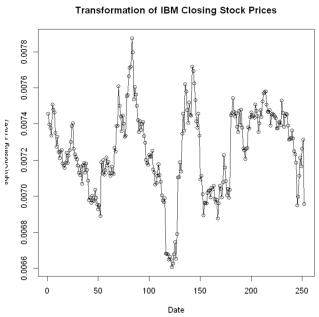
Model Specification



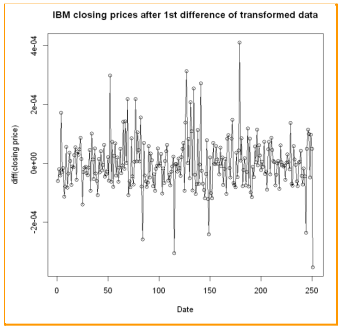
The original scatterplot, shown above, of Y_t = value of variable Y (closing price) at time t (daily dates), suggests that the process does not seem to be stationary indicated by a few dips at certain time intervals. Moreover, the sample ACF, which usually determines the component of the Moving Average (MA) process, does not provide much information besides nonstationarity due to slow decay among time lags. The partial ACF suggests an Autoregressive (AR) process since the lag spike cuts off at 1. Nonetheless, our speculation about nonstationarity is further confirmed through the Augmented Dicky-Fuller (ADF) test (p-value=0.207). The issue of nonstationarity may also suggest a deterministic trend and hence, the present study will select a "differencing" approach. Obtaining stationarity will allow for the use of the theory of stationary processes for modeling analysis and forecasting.

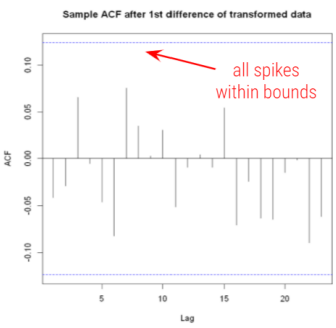
Box-Cox Transformation - Before a difference is applied, a transformation must be taken to ensure constant variance. In order to determine an adequate transformation, the Box-Cox criterion needs to be used to determine the value of the transformation parameter, λ. An approximate 95% confidence interval for λincludes λ = -1 which suggest an inverse transformation, $T(Y_t) = Y_t^{-1}$.

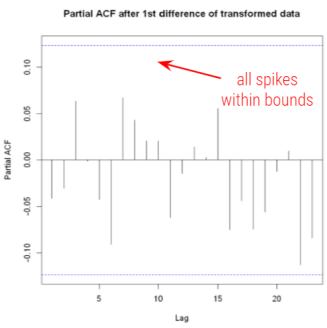




Taking a Difference - The scatterplot, sample ACF, partial ACF of the inverted transformed data shows no difference when compared to the initial plot of the IBM closing and, therefore, the same conclusion of nonstationarity previously made still holds. Furthermore, a quantitative test, the ADF test, for stationarity supports the speculation that the observed time series is not stationary (p-value = 0.2123). The observed result of the ADF test recommends a difference of the inverted transformed data. Upon taking a difference, the time series data appears to resemble a constant mean process and the ADF test (p-value = 0.01) confirms that there is sufficient evidence that the series is stationary at the $\alpha = 0.05$level. The plots of the inverted differenced data further supports that the series is now stationary with all spikes within the bounds for both the sample ACF and partial ACF.







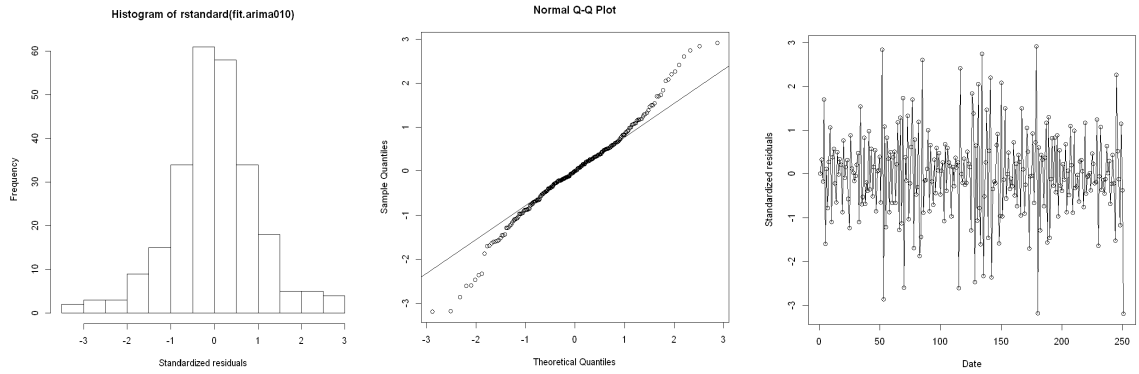
Model Specifications - Due to a difference being taken, the models chosen will be ARIMA (p,d,q) models where “p” refers the the order of the AR component, “d” the number of differences needed to arrive at stationary ARMA(p,q) process, and “q” the order of the MA component. The sample ACF, PACF and EACF can aid in selecting the model. From the wedge of zeros in the EACF, the suggested principal model is: ARIMA (0,1,0).

Model Fitting and Diagnostics - *Model Fitting* - An ARIMA(p,d,q) process with p = 0, d = 1, and q = 0 is called ARIMA(0,1,0) process and can be expressed as:

$$\Delta Y_t^{-1} = (Y_t - Y_{t-1})^{-1} = e^t$$

To estimate the unknown parameters, the method of maximum likelihood estimator (MLE) was used.

Model Diagnostics - To determine if ARIMA(0,1,0) is an adequate model, normality and independence of the standardized residuals needs to be assessed visually using QQplot, histogram and plot of the standardized residuals and through statistical tests such as the Shapiro-Wilks test and the Runs test.

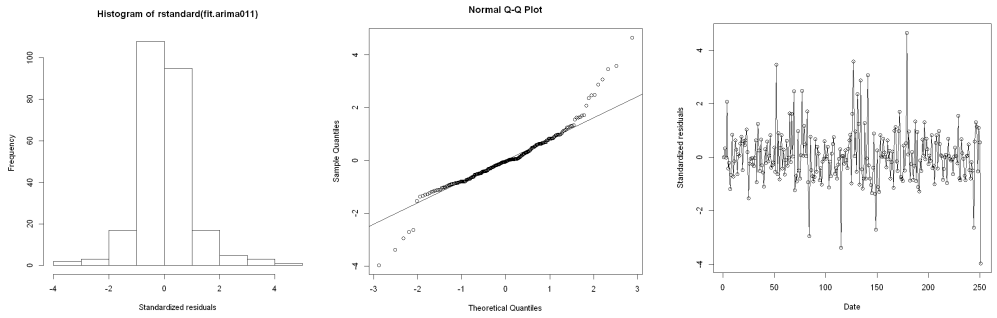


Normality Assumption - Visually, the QQ plot displays obvious deviations and outliers in the tails, However, the histogram appears to be normally distributed and the standardized residuals exhibit no obvious pattern and appears to be random. However, further model diagnostics through quantitative tests must be performed to truly determine the adequacy of the model. From the output of the Shapiro-Wilks test for H_0: "standardized residuals are normally distributed" with observed p-value of 0.0009831, we rejected the null hypothesis at $\alpha = 0.05$ level. Therefore, there is sufficient evidence to indicate that the normality assumption is not met.

Independence Assumption - Visually, the time series of the plot of the standardized residuals displays no discernable patterns and looks to be random. However the runs test does not support the visual assessment: In testing H_0: "standardized residuals are independent", with p-value of 3.28e-07, we rejected the null hypothesis $\alpha = 0.05$level. Therefore, there is sufficient evidence to indicate that the independence assumption is not met.

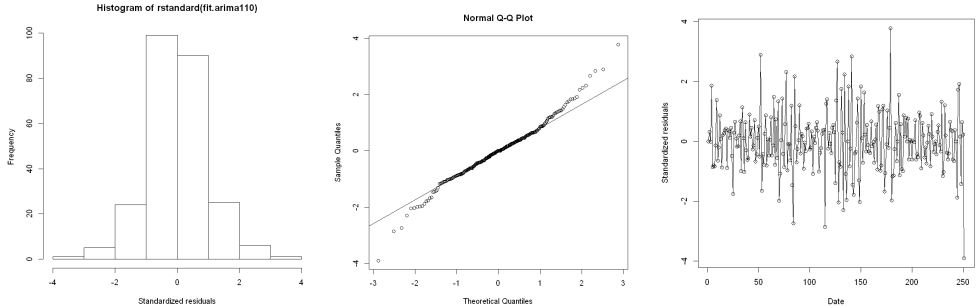
Overfitting - The principal model specifications were not met in terms of necessary assumptions. Hence, the study will attempt to overfit the ARIMA(0,1,0) with the ARI(1,1) and IMA(1,1). When overfitting any model, before any model diagnostics is performed, the significance of the parameter θis assessed by checking whether or not 0 is in the 95% confidence interval.

An ARIMA(1,1,0) ↔ ARI (1,1) can be expressed as: $Y_t^{-1} = (1 + \Phi)Y_{t-1}^{-1} - \Phi Y_{t-2} + e^t$



From Output FIXME, it can be concluded that θis significantly different from 0 and further model diagnostics shows that the normality assumption is violated - only the independence assumption is met.

An ARIMA(0,1,1) ↔ IMA(1,1) can be expressed as: $Y_t^{-1} = Y_{t-1}^{-1} + e^t - \theta e_{t-1}$

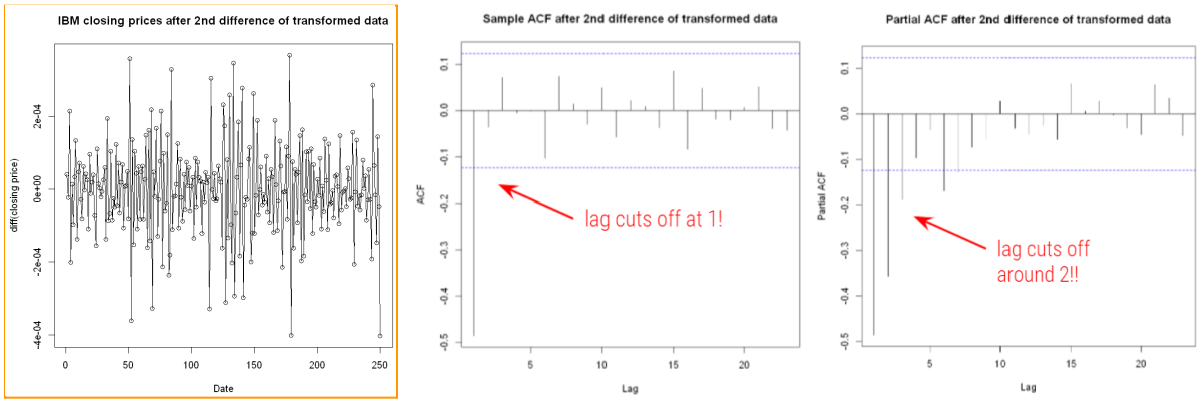


From Output FIXME, it can be concluded that θis significantly different from 0 and further model diagnostics shows that the normality assumption is violated - only the independence assumption is met.

Because all models, both principal and overfitted, failed to meet all the necessary assumptions, a 2nd difference will be applied followed by the necessary model specification, fitting, and diagnostics.

Model Specification After 2nd Difference

Taking 2nd Difference - After taking a 2nd difference of the inverted transformed data, the time series appears to resemble a constant mean process. The sample ACF and PACF also suggest the process is stationary since the time lags cut off in the early stages. The ADF test ($p < 0.01$) further confirms that there is sufficient evidence that the series is stationary at the $\alpha = 0.05$ level.



Model Specification - The sample ACF suggests an ARIMA(0,2,1) since the lag cuts off at 1 and the PACF suggests an ARIMA(2,2,0) since the lag cuts off at around 2. The EACF further verifies the model selection made through the wedge of zeros. Although the ARIMA(3,2,1) would normally be considered a false positive, the model will be considered and tested to investigate all possible model candidates.

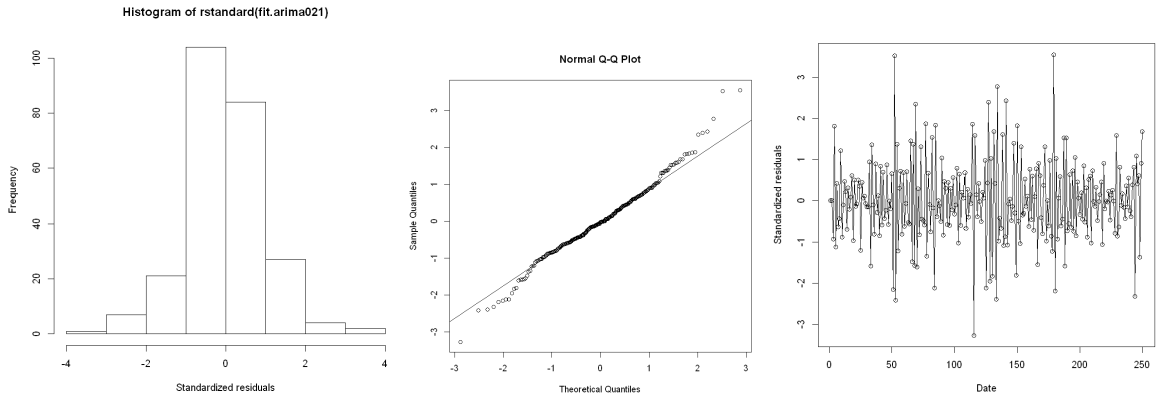
Model Fitting and Diagnostics After 2nd Difference

Model Fitting - ARIMA(0,2,1): An ARIMA(p,d,q) process with p = 0, d = 2, and q = 1 is called ARIMA(0,2,1) process and can be expressed as:

$$\Delta Y_t^{-1} = 2Y_{t-1}^{-1} - Y_{t-2}^{-1} - \theta_1 e_{t-1}$$

To estimate the unknown parameters, the method of maximum likelihood estimator (MLE) was used. Output FIXME, using ML method, the approximate large-sample confidence intervals for each MA component does not include zero, indicated that θis significantly different from zero. Based on this result, no more complex MA-type model needs to be examined.

Model Diagnostics - Similar to the previous model diagnostics process, to determine if ARIMA(0,2,1) is an adequate model, normality and independence of the standardized residuals needs to be assessed visually and through statistical tests.



Normality Assumption - From Output FIXME, when assessing normality visually and through statistical tests, we can conclude there is sufficient evidence to indicate that the normality assumption is not met.

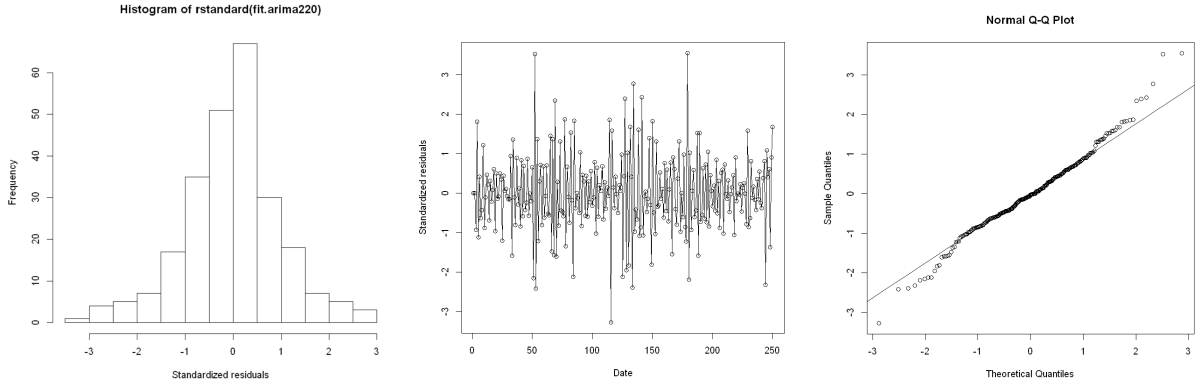
Independence Assumption - From Output FIXME, when assessing independence visually and through statistical tests, we can conclude there is sufficient evidence to indicate that the independence assumption is not met.

Model Fitting - ARIMA(2,2,0): An ARIMA(p,d,q) process with p = 2, d = 2, and q = 0 is called ARIMA(2,2,0) process and can be expressed as:

$$\Delta Y_t^{-2} = (Y_t - \Phi_1 2Y_{t-1} + \Phi_2 Y_{t-2})^{-1} + e_t$$

To estimate the unknown parameters, the method of maximum likelihood estimator (MLE) was used. From Output FIXME, using ML method, the approximate large-sample confidence intervals for each MA component does not include zero, indicated that θis significantly different from zero. Based on this result, no more complex AR-type model needs to be examined.

Model Diagnostics - Similar to the previous model diagnostics process, to determine if ARIMA(2,2,0) is an adequate model, normality and independence of the standardized residuals needs to be assessed visually and through statistical tests.



Normality Assumption - From Output FIXME, when assessing normality visually and through statistical tests, we can conclude there is sufficient evidence to indicate that the normality assumption is not met.

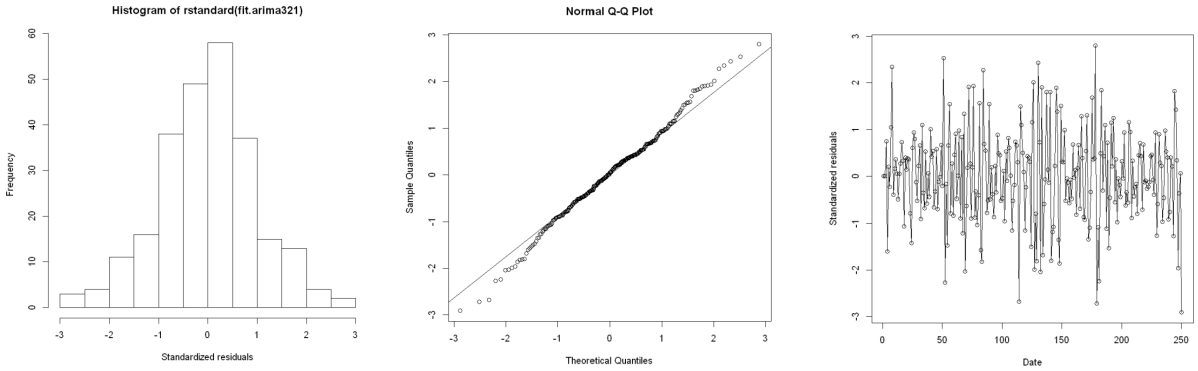
Independence Assumption - From Output FIXME, when assessing independence visually and through statistical tests, we can conclude there is sufficient evidence to indicate that the independence assumption is not met.

Model Fitting - ARIMA(3,2,1): An ARIMA(p,d,q) process with p = 3, d = 2, and q = 1 is called ARIMA(3,2,1) process and can be expressed as:

$$\Delta Y_t^{-2} = (Y_t - \Phi_1 2Y_{t-1} + \Phi_2 Y_{t-2} + \Phi_3 Y_{t-3})^{-1} + e_t - \theta_1 e_{t-1}$$

To estimate the unknown parameters, the method of maximum likelihood estimator (MLE) was used. Output FIXME, using ML method, the approximate large-sample confidence intervals for each MA component and AR component does not include zero, indicated that θis significantly different from zero. Based on this result, no more complex ARMA-type model needs to be examined.

Model Diagnostics - Similar to the previous model diagnostics process, to determine if ARIMA(3,2,1) is an adequate model, normality and independence of the standardized residuals needs to be assessed visually and through statistical tests.



Normality Assumption - From Output FIXME, when assessing normality visually and through statistical tests, we can conclude there is sufficient evidence to indicate that the normality assumption is met.

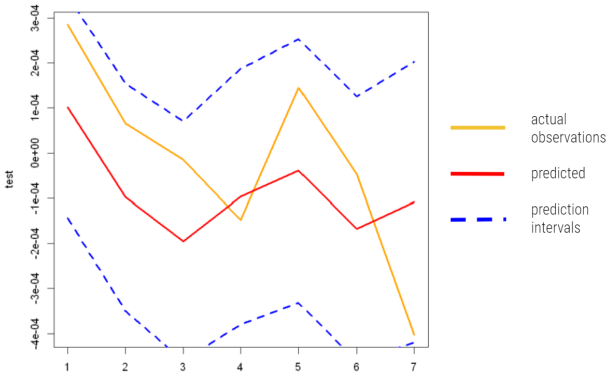
Independence Assumption - From Output FIXME, when assessing independence visually and through statistical tests, we can conclude there is sufficient evidence to indicate that the independence assumption is met.

Since ARIMA(3,2,1) meets all assumptions, we will use this model for forecasting.

Forecasting

In order to perform cross-validation on the forecasted results and original observations, the data must be subsetted. The data will be split into training and testing data where training data is all the observed values excluding the last 7 observations and testing data only contains the last 7 observations. After fitting the ARIMA(3,2,1) model and forecasting 7 days ahead, we get the following table which includes the prediction intervals, predicted values, and actual observations:

date	test	arima321.predict.pred	LP	UP
<date>	<dbl>	<ts>	<ts>	<ts>
2020-01-23	2.853204e-04	1.019709e-04	-0.0001431939	3.471356e-04
2020-01-24	6.541239e-05	-9.761678e-05	-0.0003499929	1.547594e-04
2020-01-27	-1.546254e-05	-1.949258e-04	-0.0004598012	6.994971e-05
2020-01-28	-1.476431e-04	-9.658137e-05	-0.0003800871	1.869243e-04
2020-01-29	1.448775e-04	-3.956217e-05	-0.0003314694	2.523451e-04
2020-01-30	-4.794795e-05	-1.677917e-04	-0.0004611359	1.255524e-04
2020-01-31	-4.029080e-04	-1.083898e-04	-0.0004192035	2.024240e-04



When examining the graph of the predicted versus original data, it can be seen that the predictions obtained using the false positive ARIMA(3,2,1) model is not perfect but reflects the overall pattern of the original last 7 observations. In addition, the predicted values are relatively close to the actual observations.

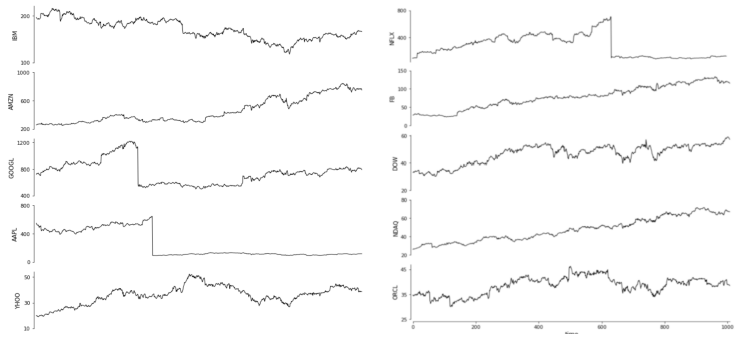
Conclusion

It can be concluded that the predicted values using the ARIMA model is not perfect which, intuitively, makes sense because how can we use one single stock's historical data to predict its future prices. If this were the case and if predicting stock prices were so simple, we would all be rich. It may be important to consider that the price of a stock may possibly be affected by another stock throughout time and casual discovery allows us to do so.

Application in Causal Discovery

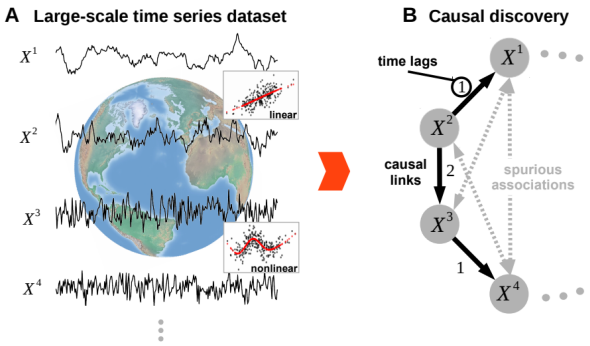
Introduction

Data - For the purpose of the application, stocks of various big-tech corporations are analyzed and used such as IBM, Amazon, Google, Apple, Yahoo, Netflix, Facebook, Oracle, DOW, and NASDAQ.



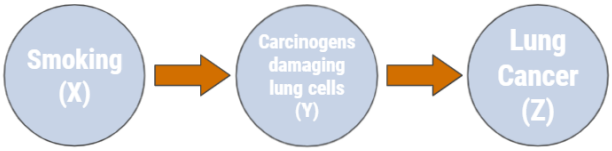
Problem Statement - The goal of the application is to determine whether or not stocks can affect one another through time.

Background - Causal discovery aims to reconstruct the underlying causal dependencies and/or relationships from large-scale time series datasets, accounting for both linear and nonlinear relationships, along with different time lags. Traditional pairwise correlations that yield spurious associations where two or more events or variables are associated but not causally related which is not very informative. Causal discovery, on the other hand, emphasizes in assessing the significance of causal links and discovering possible linkages between variables.



Important Terms

Part of understanding causal discovery is understanding the next three terms: Causal effect, Mediation, and Susceptibility. Given a scenario that depicts a relationship between variables X, Y, and Z shown in the figure below. Causal effects measure how influential a certain variable is and in this case X and Y would be the causal effects because carcinogens in smoking causes carcinogen damaged lung cells and carcinogen damaged lung cells is one cause for lung cancer. Y is the mediator because smoking doesn't directly cause lung cancer. It is the carcinogen damaged lung cells from smoking that causes lung cancer. Y and Z are the susceptible variables and essentially receptors. In this example, smoking makes your lungs more vulnerable to damage and similar carcinogen damaged lung cells makes you more vulnerable to developing lung cancer.



Methods

PCMCI - For the purpose of this application, PCMCI is the major method used to find the causal linkages. Peter-Clark Momentary Conditional Independence (PCMCI) can be essentially split into two components: PC selection algorithm and Momentary Conditional Independence (MCI) test. PC algorithm aims to estimate the parents of a variable, i.e. a stock, at time t for all other variables. Whereas, MCI tries to test whether a variable at time t-tau affects another variable at time t. PCMI is chosen as the main method as opposed to the other two methods introduced in

the research paper (Detecting and quantifying causal associations in large nonlinear time series datasets) due to its high linkage detection power.

Results

Linear Partial Correlation on Stock Dataset - The linear partial correlation is represented by a linkage graph (Figure a). It is important to note that there are two different lines: one without arrows and one with arrows. A line with no arrows simply represents a relationship and a line with an arrow represents a possible causal relationship or linkage. In this case, Netflix has an arrow to Google with lag 2. This essentially means that if Netflix's stock prices were to change, two days later we should see the price of Google's stock change as well. Another presentation of such an example is through a time series graph (Figure b). The same conclusion in Figure a can also be made in Figure b. The only difference is Figure b represents the temporal dependence structure of the linear partial correlation.

Figure A)

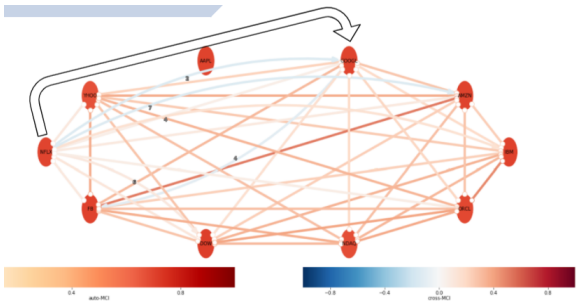
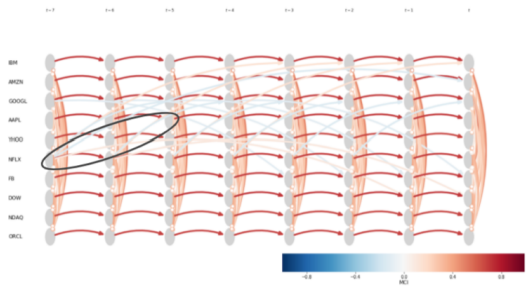
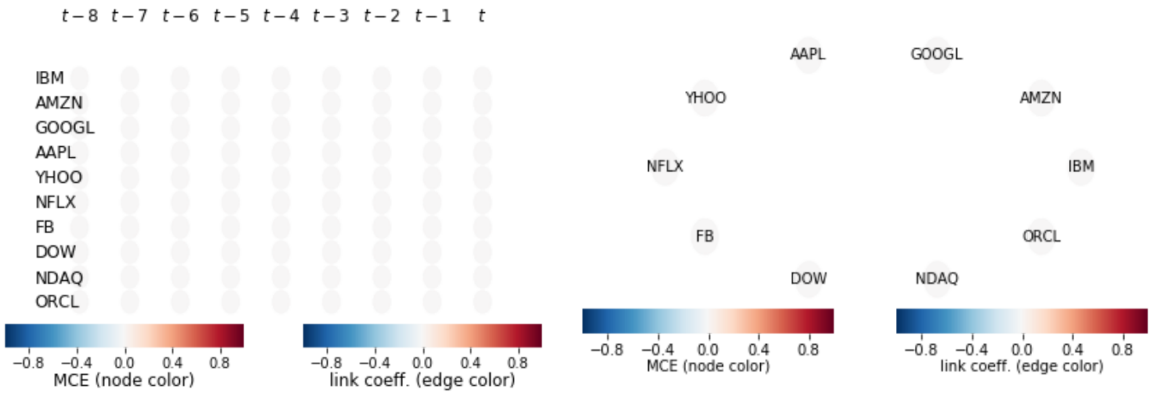


Figure B)



Linear Mediation on Stock Dataset - From the figure below, it is obvious that there are no linear or arrows in both the time series graph and the linkage graph. A possible reason for why we don't see any arrows or line is that the mediation strength among stocks is very weak and meaning most causal relationships are a direct cause-and-effect.



Summary Bar Charts - From the bar chart for average causal effect of stocks, we can see that the stock with the highest average causal effect is Netflix. This means that Netflix is the most influential on all other stocks and any change in Netflix stock prices can potentially cause other stocks to rise or fall. From the bar chart for average casual susceptibility of stocks the highest value corresponds to Google followed by IBM and Facebook. It can be concluded that Google is the most susceptible to the effect of other stocks such as Netflix (Netflix is the most influential, previously discussed). Lastly the only stock with a mediation effect is Google. Although Google is a mediator, it is extremely weak which explains why in the above graphs we do not see any lines or arrows in linkage graphs and time series graphs.

Conclusion

From the application, it is evident that stocks can, in fact, be affected by other stocks through time. Using the summary bar charts, we can apply what we discovered through this analysis toward a real life situation. Given that you are a shareholder or if you have stocks in one of the above tech corporations. By knowing that Netflix is the most influential stock and Google is the most susceptible to the effect of Netflix, meaning change whether rise or fall is most likely going effect Google the most, if you are planning to invest or if you have stock in Google you should be more aware of the stock market for Netflix.

Appendix

All output and code attached as a pdf.

Works Cited

Cryer, J. D., & Chan, K. (2011). Time series analysis: With applications in R. New York: Springer.

Forecasting: Principles and PRACTICE (3rd ed). (n.d.). Retrieved May 31, 2021, from https://otexts.com/fpp3/.

Runge, J., Nowack, P., Kretschmer, M., Flaxman, S., & Sejdinovic, D. (2019). Detecting and quantifying causal associations in large nonlinear time series datasets. Science Advances, 5(11).

Onyshchak, O. (2020, April 2). *Stock Market Dataset*. Kaggle. https://www.kaggle.com/jacksoncrow/stock-market-dataset.