Link Analysis of ETF data

SP21: NETWORK SCIENCE: 10075
project
Pranav Gujarathi (pgujarat@iu.edu)
Vishal Bhalla (vibhalla@iu.edu)

1 Research Question

Often network based relationships in real world are open to interpretation, and the basis for link formation is decided based on direct indicators or interpreted indicators. For instance, in Stock/ETF prices, we know that price of every stock influences the price of every other stock, which in turn controls the price of the stock at the next time interval. Historically, pearson coefficient between the prices expressed as a continuous time series is good indicator of the relationships between different ETFs, and correspondingly for link formation. In our project, we discuss alternative methods for qualitative and quantitative estimation for such said correlation - ie, the using Machine Learning to predict the prices and using the coefficients as values for quantitative weights.

2 Dataset and background

An exchange traded fund (ETF) is a type of security that tracks an index, sector, commodity, or other asset, but which can be purchased or sold on a stock exchange the same as a regular stock. An ETF can be structured to track anything from the price of an individual commodity to a large and diverse collection of securities. ETFs can even be structured to track specific investment strategies. We are using a kaggle dataset[1] for our analysis. The dataset contains individual ETF prices in time series form, however the time ranges vary all the way from the 1970s to current day. For consistent analysis, it is important to have a small set of ETFs from the pool all of which have an intersecting time range, and that period should be continuous. The dataset being large, after thorough computation as well as using University [2] resources, we were able to seperate a set of 156 ETFs with intersecting time range from February 2005 to November 2017 (total of 3200 continuous data points)

3 Research

As part of our analysis we have cleaned up the available data to choose the data for dates where we have data points for all the ETF's in consideration. We used 2 different approaches to analyse our data set.

3.1 Approaches

- ML Techniques to calculate weights and identify correlation followed by community detection
- Applying backbone threshold to complete graph and apply community detection techniques

It is important to note the stark differences amongst both the approaches. While correlation coefficient does measure the relationship between two time series, it is has own limitations. The accuracy of the coefficient value is highly dependent on the time series following a normal or approximately normal distribution. Additionally, the value is an independent measure of correlation, and does not take into consideration bi-variate relationships, or non linear relationships. It also does not take into account external factors that might cause correlation without causation - for instance, if two ETF prices have similar seasonal trend might show a high correlation even if they do not affect each others' prices. These reasons lead us to explore alternative models for understanding relationships between prices, leading us to formulate the Elastic Net based approach.

3.2 Methodology

- Classical comparison(Pearson coefficient) : To have a comparison baseline, we used pearson correlation coefficient as a baseline technique for measure weights/linkages. To have limited number of edges, we chose only those with magnitude above a threshold (0.8) to have links.
- ML Techniques : Here we went ahead with the hypothesis that the cost of one ξ_t at time t is a function of the price of all the ETF's at time *t-1*, as well as the timedelta[3]. We used Elastic Net to perform analysis on our model and derive the correlation. Only non zero correlations were preserved, which due to

[1] https://www.kaggle.com/borismarjanovic/price-volume-data-for-all-us-stocks-etfs

[2] https://kb.iu.edu/d/anrf

[3] Time represented a continuous feature, with the first point being 0 and the rest incremented continuously based on the difference

the normalization involved in Elastic Net Regression. These normalized coefficients then become the edge weights in our network. It is interesting to note that, after thorough experimentation, we realized that predicting the price itself led to poorly fit model, we modelled the difference between the prices, not the price itself as a function of all the prices. Hence if ξ_t^i represents price of ETF *i* at time *i*, we put forth the hypothesis of the following relationship -

$$\triangle\xi_t^i = \xi_t^i - \xi_{t-1}^i = f(\xi_{t-1}^0, \xi_{t-1}^1, ..., \xi_{t-1}^n, t)$$

where *f* is estimated using the Elastic Net Regressor mode *t* is the term representing timedelta

- Backbone Threshold approach : We created a complete graph with all the ETF tickers we had left after cleanup and found the Pearson's correlation between each of the tickers. This correlation was used as the edge weight of the graph. Thereafter we performed the backbone analysis on the graph and eliminated the edges with alpha greater than .25 .

We applied the different community detection techniques like Louvain algorithm, Asynchronous and Synchronous label propagation to the resulting graphs.

3.3 Observations

(Please refer to Appendix 1 for Images)

On application of Louvain we observe that multiple distinctive communities can be identified in both the graphs. The modularity calculated on both the graphs are very different from each other(.03 in case of ML based graph and .4 in case of backbone elimination graph). This is expected and from noticing the graph as well, we know that in reality the ETFs are much more interconnected, as unlike Stocks, these represent community of prices, and hence there is expected to be higher influence amongst ETFs.

On applying the Asynchronous and Synchronous label propagation algorithms we see that for the ML generated graph the modularity is extremely low almost tending to 0 and there is only 1 community for the graph.

However when the same techniques are applied to the graph generated using the Pearson's coefficients there are distinct multiple communities that can be observed. The modularity calculated also is very near to what we see with Louvain.

3.4 Deductions

We see that the two approaches tried are showing similar results when we apply Louvain technique for community detection, similar modularity and communities are identified however when it comes to other methods like label propagation the differences are stark.

We need to continue working on this further with applying other techniques and algorithms(e.g. SBM - Stochastic Block Models) to further analyze this data.