A Network Analysis of the Stock Market

Matt McPartlon

January 25, 2018

Introduction

The goal is to build and analyze network for the US stock market based on a distance based function of different stock price time series. Traditional stock market network analysis is usually done on *correlation threshold networks* (defined in next section), however these methods have severe limitations. The distance function proposed deviates from correlation based network construction and is motivated by sequence alignment methods in bioinformatics. Analysis of networks resulting from the distance based network construction will be done, and resulting networks will be used to build portfolios, and identify highly market structure and other important characteristics. A comparison of networks resulting from this approach and those resulting from correlation based threshold networks will also be conducted.

Definitions

For a set of time series $\mathcal{S}$ where each $S \in \mathcal{S}$ is defined at m discrete time steps, let $S(t)$ denote the value of series S at time t. Let $f : \mathcal{S}^2 \to \mathbb{R}$ be any function taking $k-$subsets $\mathcal{S}$ into $\mathbb{R}$, we can define a *threshold graph* $G_\theta = \{V, E, f\}$ on $\mathcal{S}$ where there is a vertex $v(S) \in V$ for each $S \in \mathcal{S}$ and an edge $e = \{v(S_i), v(S_j)\}$ connecting distinct time series in G if and only if $f(S_i, S_j) \geq \theta$. An important example of such a network is the *correlation threshold graph*, where f is the function that computes the correlation between time series S_i and S_j, and $\theta \in [-1, 1]$. For concreteness, say $\theta = 0.5$, then two time series are adjacent is and only if their correlation coefficient is at least 0.5.

Motivation

Correlation threshold networks of stocks are extremely relevant in portfolio management, The investment philosophy for most funds is to "beat the market." In order to achieve that goal, they need to select the stocks that are representative of the market. A common choice would be stocks high market caps which could potentially yield greater impact to the market. If we apply this logic to a stock network, we could try to select stocks (vertices) with high centralities. Furthermore, in a stock market, changes in a stocks price might effect the other stock's prices. If these kinds of interactions can be realized, not only the stock itself, but also the past prices of other stocks interacting with it might be used while predicting its stock price. These techniques could also be used to identify groups of mutually "distant" stocks in order to manage risk, or to detect community structures, where stocks in different communities tend to fluctuate independently.

Although correlation threshold networks have given nice results to the motivations described above, they have several pitfalls. The first is that these networks only detect directional changes in time series. Another is that in periods where the market as a whole performs poorly/strongly, correlation coefficient s may be inflated. The next is that correlation methods operate under the implicit assumption that similar/opposing price movements between two stocks always occur over the same time interval. To be more concrete, it may be the case that $S_1(t+1)$ and $S_2(t)$ are highly correlated at certain times, and $S_1(t+2)$, $S_2(t)$ are highly correlated at other times. Standard correlation based methods could not capture this. This begs the question: Is correlation a true indication of dependence/independence in stocks?

Description of Method

The goal is to analyze the threshold graphs of fixed a set of stock price time series over varying time horizons, and time periods. The threshold function we will use is inspired from sequence alignment methods in Bioinformatics, and involves converting each time series into a sequence of symbols analogous to nucleotides in a DNA sequence. Each symbol x_i will encode some notion of price movement, for example, the sequence of prices $(1, 2, 1, 1, 0, 1)$ could be encoded as (U, D, N, D, U) where U in position i indicates that the price went up from time i to time $+1$, D indicates down, and N indicates no movement. More complicated encodings are possible as well. Encodings that capture the magnitude/scale of movements could be considered, as well as those that capture the micro-structure of price movements - each possible structure could be encoded with its own symbol.

After a sequence of symbols are generated for each stock price time series, we can conduct pairwise sequence alignment techniques to calculate the "distance" between respective stocks corresponding sequences. With these estimates, threshold networks can be built, and examined using standard network analysis techniques. To reassert the generality of this technique, it is not hard to show that the U, D, N encoding above can be used with a high enough gap-penalty set to produce a distance estimate that is proportional to the correlation between two stocks price series.

Previous Work

Several papers [2, 7, 4, 9, 3] have analyzed various market networks where nodes represent some trading entity, and edges reflect the correlation between prices of the entities being traded. Additional papers have tried to generalize this approach. Some authors consider computing cross correlations by including asset returns and volume trading as the main variables to study the financial market. Some of the analysis techniques on correlation threshold graphs include analysis of minimum spanning trees (MST) and planar maximally filtered graphs (PMFG), community detection, clustering coefficients, as well as topological properties such as the average length of shortest paths, cuts, and centrality measures. Some of these methods were used by [3]to empirically confirm an increase in globalization between 2005-2009 by analyzing the dynamics of several international markets. [7] also used these methods to construct diversified portfolios which performed better than several market indexes by selecting stocks with high centrality. Most recently, dependency networks have been used to analyze market structure [5]. These networks rely on the partial correlation between assets - a measure of how the correlation between two variables, e.g., stock returns, is affected by a third variable. The resulting information is used to calculate the dependancy $D(i, j)$ of an asset i on another asset j. Roughly speaking ,$D(i, j)$ is a measure of the average influence of node$j$ on the correlations $C(i, k)$ over all nodes k not equal to j. By modeling financial networks this way, [5] were able to quantitatively show the dependency relationships between the different economic sectors - revealing that the structure of the network is dominated by companies belonging to the financial sector, which are the "hubs" in the dependency network.

Below is a list of some important considerations taken in previous work

- Time period that the data is collected over (Long term vs. short term)
- Time horizon that asset prices are sampled (end of trading day price, hourly prices, weekly returns etc.)
- Time lag effect of stock prices (i.e. rather than consider the correlation between $S_i(t)$ and $S_j(t)$, consider the correlation between $S_i(t + \Delta t)$ and $S_j(t)$ where Δt indicates the amount of "lag").
- The geographic region in which assets are traded (or the exchange that they are traded on)
- The market sector in which assets are traded
- Inclusion of additional information such as trade volume, market indicators, etc

Proposed Analysis

1. Compare distance based threshold networks to correlation threshold networks and dependency networks. (i.e. do they yield the same types of clustering, same degrees, etc)
2. Explore the persistence of these measures over time
3. Explore the effect of different penalties (i.e. the affect of changing the gap penalty, deletion penalty, etc)
4. Explore the affect of different alignment schemes (i.e. local alignment, global alignment, semi-global alignment)
5. Explore the effect of different penalty matrices (i.e. explore the effect of assigning different penalties to aligning different symbols)

6. Explore the distribution of connectedness in the network. Several authors have shown that for correlation threshold networks, the distribution is well approximated by a power law in the tails. Is this the case for the our network? If so, what is the tail-exponent of the degree distribution?

7. Explore how these measures related to realized trading profits and trade correlations (i.e. information diffusion is indeed an important factor, then the centrality should be superior to other measures).

Plan of Attack

- Gather desired market data for highly capitalized US stocks by end of this week (1/20) ✓
- Code Sequence alignment algorithms by end of next week (1/27) ✓
- Generate meaningful symbols for different price movements, and determine a systematic way to determine alignment penalties, then produce some of the graphs described (2/3)
- Explore the effect of different symbol sets, and different penalties, in terms of the properties of the networks produced (2/10)
- Compare to correlation threshold networks, and begin writing up results (2/10-Deadline)

More Recent Work

Showing that Correlation can be captured using sequence alignment methods.

Let $S = s_1, \ldots, s_n$ be a real valued time series indexed by t for $t = 1, \ldots, n$. Let $\bar{S}$, andσ_S denote the mean and standard deviation S respectively. Also, define $\Delta(S) := (\frac{s_t - \bar{S}}{\sigma_S} \; : \; t = 1, \ldots, n)$, the ordered list of so-called "$z - scores$," where $\Delta(S)_t = \frac{s_t - \bar{S}}{\sigma_S}$. For two analogously defined time series X and Y, the correlation of X and Y, denoted by $\rho(X, Y)$, is defined as

$$\begin{aligned} \rho(X,Y) &= \frac{1}{n-1}\sum_{t=1}^{n}\left(\frac{x_t - \bar{x}}{\sigma_X}\right)\left(\frac{y_t - \bar{y}}{\sigma_Y}\right) \\ &= \frac{1}{n-1}\sum_{t=1}^{n}\Delta(X)_t \cdot \Delta(Y)_t \end{aligned}$$

Consider the following encoding of the same two time series X and Y. Begin by defining an alphabet $\mathcal{A} = \{A_\delta : \delta \in \Delta(X) \cup \Delta(Y)\}$. Let $e_X : [n] \to \mathcal{A}^n$ be given by $e_X(t) = A_{\Delta(X)_t}$. Define e_Y analogously. To be concrete, $e_X(t)$ is the letter A_δ such that $\delta = \frac{x_t - \bar{X}}{\sigma_X}$, i.e. the $z-$score of x_t. We now have an encoding of the time series X and Y that captures the same information as their sequence of $z-$scores. We can then create a distance function $d(\cdot, \cdot)$ between two letters A_δ, $A_{\delta'} \in \mathcal{A}$, defined by $d(A_\delta, A_{\delta'}) = \frac{\delta \cdot \delta'}{n-1}$. Using standard pairwise global sequence alignment techniques to align the sequences $\{e_X(t), e_Y(t)\}_{t=1,\ldots n}$ under the described distance function, and setting the gap penalty to positive infinity, we can get the alignment score $s_{X,Y}$ of X and Y. Note that using global alignment with an infinite gap penalty forces $e_X(t)$ to align with $e_Y(t)$, thus

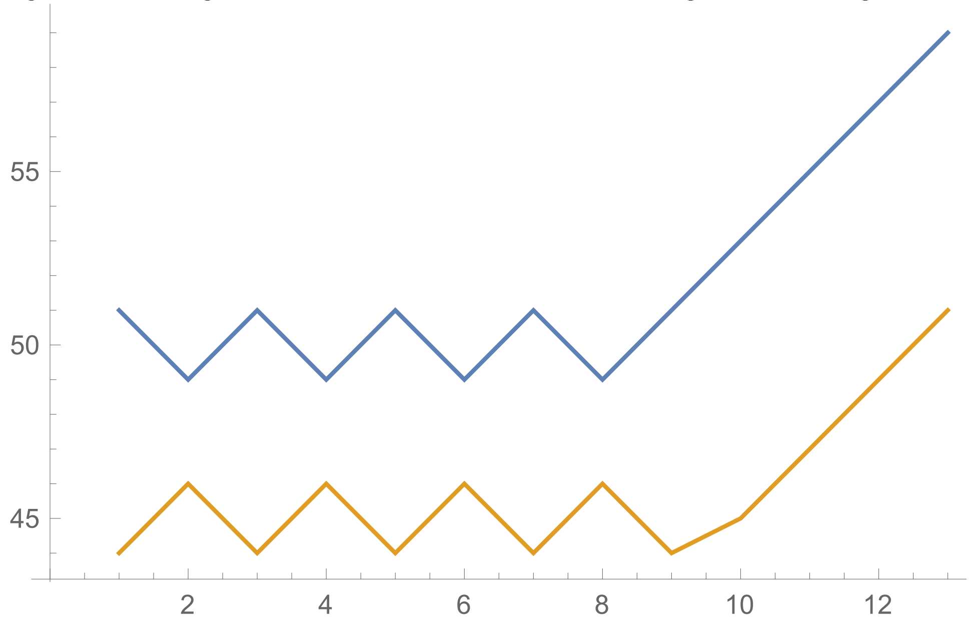
$$\begin{aligned} s_{X,Y} &= \sum_{t=1}^{n} d(e_X(t), e_Y(t)) \\ &= \sum_{t=1}^{n} \frac{\Delta(X)_t \cdot \Delta(Y)_t}{n-1} \\ &= \frac{1}{n-1}\sum_{t=1}^{n}\Delta(X)_t \cdot \Delta(Y)_t = \rho(X,Y) \end{aligned}$$

Although this is somewhat contrived, this result shows that using sequence alignment methods for creating threshold networks is at least as powerful as correlation.

Capturing Abnormal Market Conditions

One of the issues with correlation based methods is that they do not capture abnormal market conditions. Take, for example, two stocks X and Y where x_t and y_t denote the price of X and Y at time t. Suppose That X and Y have strong negative correlation from time $0, 1, \ldots, t_0$, but at time $t_0 + 1$, there is some good news in the market, and both stock values rise, producing a period of positive correlation between the two time series. This is illustrated below.

Figure 1: Two negatively correlated stocks, each responding positively to good news



Here, stock X, (blue) has perfect negative correlation with stock Y, (yellow) from time $1, \ldots, t_0 = 9$, and strong positive correlation from time $t_0 + 1 = 10$ to time $n = 13$. The correlation value between stocks X and Y is 0.75, well above the threshold value used in the construction of many correlation based networks.

It is not uncommon for external events to impact overall market condition. Several papers [8, 6, 1] have been written on this alone, and range from analyzing market swings due to political events to swings caused by sports sentiment. Overall, it is well understood that news of impending events can create large swings in overall market prices for weeks before and after they actually occur.

References

[1] ALEX EDMANS, Jack C. Bao, Nick Barberis, Andrew B. Bernard, Oyvind Bohren, B. Espen Eckbo, Florian Ederer, Tim Johnson, Lisa A. Kramer, Rafael Laporta, Nils Rudi, Petter Rudi, and Rob Stambaugh. Sports sentiment and stock returns. 2006.

[2] B. Ergï¿œer and Z. ï¿œataltepe. Relational stock market network analysis. In *2013 21st Signal Processing and Communications Applications Conference (SIU)*, pages 1–4, April 2013.

[3] M. Eryigit and R. Eryigit. Network structure of cross-correlations among the world market indices. *Physica A: Statistical Mechanics and its Applications*, 388(17):3551–3562, 2009. cited By 43.

[4] G.-G. Ha, J.W. Lee, and A. Nobi. Threshold network of a financial market using the p-value of correlation coefficients. *Journal of the Korean Physical Society*, 66(12):1802–1808, 2015. cited By 3.

[5] Dror Y. Kenett, Michele Tumminello, Asaf Madi, Gitit Gur-Gershgoren, Rosario N. Mantegna, and Eshel Ben-Jacob. Dominating clasp of the financial sector revealed by partial correlation analysis of the stock market. In *PloS one*, 2010.

[6] A. Khalid and G. Rajaguru. The impact of political events on financial market volatility: Evidence using a markov switching process. *Globalisation and Development Centre*, (43), 2010.

[7] Chun-Xiao Nie and Fu-Tie Song. Constructing financial network based on pmfg and threshold method. *Physica A: Statistical Mechanics and its Applications*, 495:104 – 113, 2018.

[8] R Schweitzer. How do stock returns react to special events. *Business Review: The Federal Reserve bank of Philadelphia*, pages 17–29, 1989.

[9] D.-M. Song, M. Tumminello, W.-X. Zhou, and R.N. Mantegna. Evolution of worldwide stock markets, correlation structure, and correlation-based graphs. *Physical Review E - Statistical, Nonlinear, and Soft Matter Physics*, 84(2), 2011. cited By 94.