Market Regime Classification Using Correlation Networks

Fan Zhu, Cedric Nam

Advisor: Dr. Michael Aguilar

September 20, 2019

Abstract

Market classification has historically been a very loosely defined standard for financial practitioners. We define a stronger heuristic to classify based on consecutive months of positive or negative returns. Then we take historical data on U.S. industry returns, and create correlation networks to determine if these networks contain intrinsic properties that determine, or are determined by our standard market classification. We find that a machine learning classifier using the x, y, z-coordinates of node position, stress values, and edge weights from 55 different networks can create a model with up to 70% accuracy when classifying new data. If this accuracy remains consistent, this opens the possibility of using these networks to forecast regime shifts by using lagged data, or instead training a model based on data between regimes, i.e., the transition periods.

1 Introduction

Market regime classification has historically been a task with no strong standards. Financial practitioners have a simple "definition," classifying bull markets as those where the market (usually an index such as the S&P500) has risen 20%, and bear markets as those where the market has fallen 20%. But this is still a loose heuristic, and could occur in a matter of days, or months. Furthermore, this method of classification inherently relies on current data. But what if it were possible to know how the market will move, before the market moves? This is closely related to literature based on forecasting metrics such as GDP. Using components of GDP, such as housing prices, unemployment, financial indices, manufacturing, etc., many different groups such as governments, banks, corporations, and individual investors attempt to predict future values of GDP, or at least predict a trend. Similarly, components of the markets, such as different industries, and their relationships with one another should contain intrinsic properties that can produce a classification of the current market regime. There may even exist recurring properties that appear before shifts in regimes. In this way, we may move beyond classification into forecasting, though this paper's focus is solely on classification.

There exists almost no prior work attempting to classify market regimes using networks. While there is similar literature for GDP network creation, and much more on GDP forecasting, our work is dissimilar enough that it does not fall snugly under the umbrella of either group of GDP-related literature. We do not attempt to predict or model exact values of GDP, but rather set up a framework to classify current trends, which in turn leads nicely into future work on forecasting.

Networks were chosen as the classification data source of choice for a variety of reasons. They provide a clear visualization of relationships between different nodes, as can be seen in Section 4. As our chosen method of creating networks created the same consistently given the same data, it was then possible to store geometric coordinates for each point, along with other metrics such as stress values, along with classic metrics such as industry correlations. This essentially tripled the data available with which we could create models for classification.

We use Ken French's 49 industries data to create a series of correlation networks over time, alongside S&P500 data, from 03/01/1978 to 06/28/2019. We create our own heuristic for classifying market regimes, basing our classifications not only on magnitude of returns, but overall duration of majority positive or negative returns. We also introduce a neutral market classification, in which we see both positive and negative returns. Once these periods were identified, correlation networks were formed from the corresponding periods of industry data, which were then analyzed.

In Section 2, we summarize prior literature on related topics, and how we build upon their work. In Section 3, the specific methodology of our paper is detailed, including the data used, our market identification heuristic, and how the data was transformed and classified. In Section 4, we review the results, and in Section 5, we conclude.

2 Literature Review

Given the complexity of the underlying systems of the financial markets, it is natural for researchers to resort to more sophisticated theories and tools in physics to construct market models with better fits of the financial data. The correlation network is one of those physics models that gained popularity among financial economists. The theoretical pavement for this application is that, the price of a financial asset is influenced by both macroeconomic factors and the prices of other assets (Fama and French, 1992). The change of the economic factors and the assets' prices would result in an information flow that eventually affect other assets. Such interconnections between financial assets constitutes the fundamental structures of the market, which can be captured by the correlation network through straightforward visualizations and fitting mathematical models.

Mantegna (1999), among other trailblazers who applied the network theories in financial data, converted the Pearson's correlation coefficients between stocks in the Dow Jones Industrial Average index (DJIA) into Euclidean distances, which were then utilized to construct a minimum spanning tree representing the interactions between the stocks. This hierarchical taxonomy Mantegna identified validated the applications of the network theory in the financial markets. Through implementing the nearest neighbor single linkage cluster algorithm to a dataset of stocks traded at the New York Stock Exchange during a 12-years period, Bonanno and Mantegna (2003) compared the minimum spanning trees formed by the real data and the MSTs formed by the random model and the one-factor CAPM model. The comparisons showed that both the random and the one-factor model failed to capture the real market's hierarchical distribution of significance of the stocks. These results emphasized the complexity of the real market's intrinsic structures to which the simple models such as the random model and the classic CAPM model failed to approximate. Thus, the network model appears to be more attractive to researchers who study the market taxonomy.

There has also been similar work done without correlation networks, to attempt to forecast GDP using financial market data. Kuosmanen and Vataja (2013) found that during different market conditions in Finland, the best indicator variables also changed. Periods of steady growth saw short term interest rates and previous values of output growth as the best economic activity forecasters. During economic turbulence, it was instead the traditional yield spread and stock market returns that best forecast the economy. These results follow older literature on GDP forecasting variables quite consistently, and can be generalized beyond small open economies like Finland. Kuosmanen and Vataja conclude by noting the need for a variety of different forecasting models during different market conditions. They propose, as a solution, the use of an inversion-recession signal, essentially a market regime classifier that determines the best forecasting model to use at the time. Our model could potentially be used in place of their signal, in conjunction with their forecasting models.

3 Methodology

3.1 Data

We used two main sources of data, the 49 industries daily returns from the Kenneth French's data library, and historical S&P500 daily prices data (henceforth market data). To maintain a balanced panel data, we selected the time range from 03/01/1978 to 06/28/2019. The 49 industries returns data were converted into pseudo-prices, so that we could then calculate 20-day log returns for both the 49 industries and market data. This was chosen to imitate monthly returns (20 trading days a month), but provide us with 20 days worth of data per "month." Therefore even a single month of bear market classification could provide up to 20 different networks, or a single network comprised of 20 different data points for each of the 49 industries.

Once different market regime periods were identified (details on this below), we aggregated the data from each period and constructed a single network from each period. In other words, we simplified the different length regimes, such as a four month bull market followed by a three month neutral and then a one month bear, eight data points, into a three data point table consisting of a bull, then neutral, then bear period. The longer a period, the more data we had available for that period's network.

By reducing our data set in this way, we ended with

- 8 bear markets
- 23 bull markets

- 24 neutral markets

which each ranged from one month (20-day period) to over 12 months.

3.2 Market Identification

For the purpose of this paper, we created a new heuristic to classify markets. Expanding upon the classic ”20% up/down” rule-of-thumb, we created a four-month rule to classify as bull, bear, or neutral. We initially classify the market based on the first four months of returns. If three months are positive, it was a bull market. If three were negative, it was a bear. If neither, the market was classified as neutral. From then on, we followed the following rules:

- Four consecutive months of positive returns reclassified as bull.
- Four consecutive months of negative returns reclassified as bear.
- If one month of opposite returns appeared, then two consecutive months of current returns appeared, then the current classification held. E.g., if the market was currently in a bear market, saw a month of positive returns, followed by two months of negative returns, we kept the classification as bear.
- Three months of alternating signs reclassified as neutral. If we were in a bear market, saw a month of positive returns, a month of negative returns, and a month of positive returns, in that order, the classification would be changed from bear to neutral.

- Two or three months of opposite returns, followed by a month of current returns, reclassified as neutral.

- If returns during a single month (a 20-day period) exceeded 20% up or down, we reclassified as bull or bear respectively, regardless of the prior classification or current trends.

3.3 Correlation Network

We chose to use a minimum spanning tree for our correlation networks. MSTs are a well-established method of forming networks from similar data. For our networks, we used Matlab's graph and minimum spanning tree functions to construct them. The specific methodology used to create these trees is as follows (Mantegna and Stanley, 1999):

1. First, we obtain the daily adjusted closing price for each stock and calculate the log returns.

2. Then we compute the correlation coefficient C_{ij} between each pair of stocks and obtain a n-by-n matrix of C_{ij}. The subscripts i, j and k represent different financial assets.

3. We convert the correlation coefficients to Euclidean distances representing the edges in the network. Each edge distance must satisfy three axioms:

    - $d_{ij} = 0$ if and only if i = j
    - $d_{ij} = d_{hi}$

- $d_{ij} \leq \mathrm{d}_{ik} + d_{kj}$

4. However, the direct application of the correlation coefficient does not satisfy these axioms. Thus, we need to transform these correlation coefficients to be qualified for the three axioms. One of the possible transformation functions is

$$d_{ij} = \sqrt{2 \cdot (1 - C_{ij})}$$

5. There are several feasible algorithms to determine the rules of forming links from the Euclidean distance to a correlation network. One of them is the Sammon Projection algorithm, which projects a space of high dimensionality to a low-dimensional space through minimizing the loss of the structural information of the high-dimensional space in its projection of the low dimensional space (Sammon, 1969). The Sammon's error function is shown below, where d_{ij}^* denotes the Euclidean distance between stocks i and j in the original space, and d_{ij} denotes the distance between these two stocks' projections in the low-dimensional space. By the Sammon Projection algorithm, we construct a minimum spanning tree for each market period we classified. Each node in the MST represents an asset and the MST has n-1 edges that minimize the total edge distances in the graph.

$$E = \frac{1}{\sum\limits_{i<j} d_{ij}^*} \sum_{i<j} \frac{(d_{ij}^* - d_{ij})^2}{d_{ij}^*}$$

3.4 Machine Learning for Forecasting

After creating the networks, we use Matlab's Classification Learner App, which creates a number of different machine learning classification models. We will train the network on 90% of the data, and test on the remaining 10%. Specifically, we will be using the x, y, and z-coordinates of each node, edge weights, and stress values from the networks formed in the previous step, and using the classification app to determine if there are any patterns before, during, or after market regime switches, recurring often enough to be used to classify, or even forecast when switches will occur, and what it will switch to.

The most common models used are briefly described below. This is followed by a short explanation on how the models are kept from being overfitted, or being too closely linked to the training data, thus becoming unable to generalize for new data.

3.4.1 Decision Trees

Decision trees are a classification and regression tool that choose features in the provided data to "split" data between, branching downwards like an upside-down tree. They are some of the simplest machine learning outputs to understand and visualize, require little-to-no data prep from the user, and implicitly use feature selection.

3.4.2 Support Vector Machines (SVM)

Support Vector Machines is a discriminative classifier that, given labeled training data, outputs a hyperplane that separates new data into the two classes defined by a subset of the data, called support vectors. If the data is not linearly separable, SVM projects the system to higher and higher dimensions until it can be linearly separated.

SVM models are built by separating data into two subsets, the larger used to identify support vectors, and the smaller one to test accuracy. The main advantage of SVM is its ability to properly identify global, rather than local, minima, which allows SVM to generalize easily to new data

3.4.3 Nearest Neighbors (KNN)

Nearest Neighbors, or K-Nearest Neighbors, is a non-parametric lazy learning algorithm. It makes no assumptions about the distribution of the data, instead determining that information from the data itself, hence its non-parametric characteristic. KNN also has essentially no training period, instead taking new data and comparing it to the ”memorized” test data, and classifying it then, making it a lazy algorithm because it does not formulate a classifying model equation during a traditional testing period. It simply holds all of the input data, and then adds to its collection.

Data in KNN are classified usually by a majority vote of its k-nearest neighbors, joining whatever class it most closely resembles. KNN is a simple to understand, versatile classifier with high accuracy, but suffers from high memory costs, slow predictions, and over-sensitivity to noise and size.

3.4.4 Ensemble Classification

Ensemble Classification is a machine learning classification technique that utilizes a number of different models to create a single, hopefully superior, model. There are a number of different ensemble methods, including bagging (bootstrap aggregating), random forest, boosting, and stacking. Bagging simply creates multiple models using different data samples drawn with replacement (bootstrap sampling), and then averages their results. Random forest is similar to bagging, using bootstrap sampling to create each tree's dataset, except each different tree is given only random subset of all the available features that can split upon, rather than having all features available. Boosting takes weak classifiers, and retrains them on weighted versions of the training data, until they become accurate. Stacking combines multiple classifiers through a meta-classifier. The base models are trained on the complete training data, and successive models are trained on the previous model's outputs.

3.4.5 Overfitting

In a machine learning scheme, a common problem for models is overfitting, when the model is too sensitive to noise or short-term only effects. This means that while the model may be accurate for the training data, it is too accurate, and cannot easily generalize to new data. To avoid this problem, we utilized Matlab's 5-fold cross validation. A classic k-fold cross validation splits the dataset into k subsets, and repeats the training-testing steps k times. Each time, a different subset of the data is used as the testing data, and the remaining k-1 subsets are used as the training data. An example of a 5-fold cross validation is shown in the figure below.

4 Results

4.1 Preliminary Results

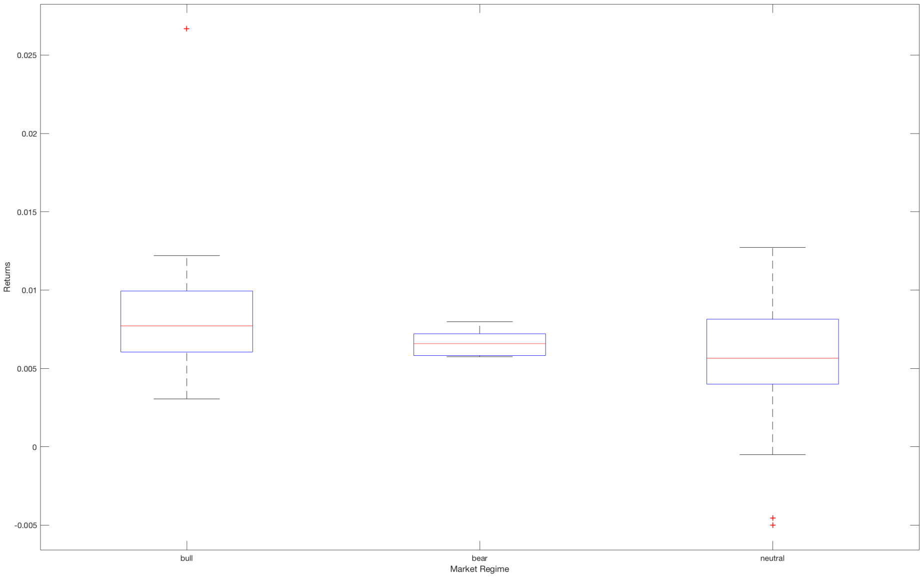
The following are descriptive statistics of our results for each industry's returns during bear (Table 2), bull (Table 3), and neutral markets (Table 4), and the average statistics for returns during bear, bull, and neutral markets (Table 1). Due to the space limit, the sector-wise descriptive statistics tables only showed 5 of the 49 industries: Agriculture, Food, Smoke, Toy, and Household.

Table 1: Average Market for Three Regimes

Regime	Bear	Bull	Neutral
Mean	0.0066	0.0088	0.0054
Standard Deviation	0.0454	0.0441	0.0446
Skewness	-1.0929	-1.0906	-0.9710
Kurtosis	8.5211	9.3000	7.4641
Minimum	-0.3240	-0.2889	-0.2400
25^{th} Percentile	-0.0169	-0.0140	-0.0166
50^{th} Percentile	0.0106	0.0122	0.0105
75^{th} Percentile	0.0339	0.0347	0.0319
Maximum	0.1733	0.1714	0.1672
Range	0.4972	0.4603	0.4072

The Table 1 showed that the bull markets (M = 0.0088, SD = 0.0441) have the highest average returns, while the neutral markets (M = 0.0054, SD = 0.0446) have the lowest mean returns. One possible explanation to this inconsistency with the definition of the bear (M = 0.0066, SD = 0.0454) and neutral markets is that our regime classification heuristics focused on the four-month window of market returns. A longer horizon of alternating positive and negative returns with, nevertheless, a larger drop in its total return, might be classified as a neutral market by the rules. Consistent to the theory, the bear markets exhibited the highest volatility in the returns, while the bull market showed the lowest volatility. Even though the distributions for all three regimes showed negative skewness, the distribution of the bear markets is the most skewed to the left. The bull markets' distribution showed the largest kurtosis, which indicated that it had the fattest tails among the three distributions.

Figure 1: Boxplot for the Returns in Bull, Bear and Neutral Markets



To study whether the average returns of the three regime of markets differed from each other, we performed the t-tests between each pair of the three regimes, without assuming equal variance. The results showed a significant difference between the bull markets and the bear markets, t(25.8928) = 2.1969, p = 0.0372. The neutral markets are also significantly different from the bull markets, t(44.9784) = 2.5407, p = 0.0146. However, the bear markets we classified are not significantly different from the neutral markets, t(27.0283) = 1.2250, p = 0.2311.

Table 2: Five Industries in the Bear Markets

Sector	Agric	Food	Smoke	Toy	Hshld
Mean	0.0106	0.0122	0.0148	0.0079	0.0089
Standard Deviation	0.0619	0.0459	0.0633	0.0751	0.0489
Skewness	-0.7622	-0.6932	-0.4331	-0.9549	-0.9912
Kurtosis	6.5200	6.3432	5.3924	7.7721	7.9251
Minimum	-0.3593	-0.2833	-0.3265	-0.5673	-0.3386
25^{th} Percentile	-0.0206	-0.0113	-0.0202	-0.0337	-0.0178
50^{th} Percentile	0.0143	0.0140	0.0185	0.0116	0.0120
75^{th} Percentile	0.0485	0.0398	0.0537	0.0565	0.0391
Maximum	0.2578	0.2008	0.2801	0.2678	0.1869
Range	0.6171	0.4840	0.6066	0.8352	0.5255

4.2 Correlation Networks

Below are the average networks of each market type. The numbers along the edges represent Euclidean distances, not weights, so a higher number corresponds to a higher distance, and therefore lower correlation. The two nodes with shorter distances indicate stronger correlation between these two sectors.

We found that among Matlab's different classification models, various support vector machine and ensemble models resulted in the best predictions. We achieved a consistent 65% to 70% accuracies on our testing data.

Table 3: Five Industries in the Bull Markets

Sector	Agric	Food	Smoke	Toy	Hshld
Mean	0.0115	0.0146	0.0155	0.0107	0.0116
Standard Deviation	0.0588	0.0442	0.0615	0.0734	0.0478
Skewness	-0.6608	-0.6928	-0.4452	-0.9998	-1.0440
Kurtosis	6.6649	6.4716	5.0420	8.4083	8.3119
Minimum	-0.3218	-0.2514	-0.2894	-0.5019	-0.2996
25^{th} Percentile	-0.0193	-0.0082	-0.0197	-0.0297	-0.0146
50^{th} Percentile	0.0146	0.0165	0.0200	0.0147	0.0149
75^{th} Percentile	0.0471	0.0413	0.0543	0.0576	0.0409
Maximum	0.2371	0.1876	0.2513	0.2630	0.1758
Range	0.5590	0.4390	0.5407	0.7649	0.4754

Table 4: Five Industries in the Neutral Markets

Sector	Agric	Food	Smoke	Toy	Hshld
Mean	0.0097	0.0103	0.0140	0.0070	0.0080
Standard Deviation	0.0614	0.0416	0.0602	0.0707	0.0446
Skewness	-0.6468	-0.8078	-0.3503	-0.7676	-0.8780
Kurtosis	5.9606	6.1501	4.8874	5.9880	6.6407
Minimum	-0.2941	-0.2031	-0.2488	-0.3438	-0.2193
25^{th} Percentile	-0.0217	-0.0106	-0.0199	-0.0310	-0.0155
50^{th} Percentile	0.0128	0.0127	0.0181	0.0125	0.0115
75^{th} Percentile	0.0462	0.0352	0.0506	0.0516	0.0353
Maximum	0.2163	0.1519	0.2297	0.2422	0.1583
Range	0.5104	0.3549	0.4785	0.5860	0.3776

Table 5: T-Test between the Average Bull and Bear Market

Regime	Bear	Bull	Neutral
Mean	0.0066	0.0088	0.0054
Standard Deviation	0.0454	0.0441	0.0446
Skewness	-1.0929	-1.0906	-0.9710
Kurtosis	8.5211	9.3000	7.4641
Minimum	-0.3240	-0.2889	-0.2400
25^{th} Percentile	-0.0169	-0.0140	-0.0166
50^{th} Percentile	0.0106	0.0122	0.0105
75^{th} Percentile	0.0339	0.0347	0.0319
Maximum	0.1733	0.1714	0.1672
Range	0.4972	0.4603	0.4072

Bear Mkt MST.png

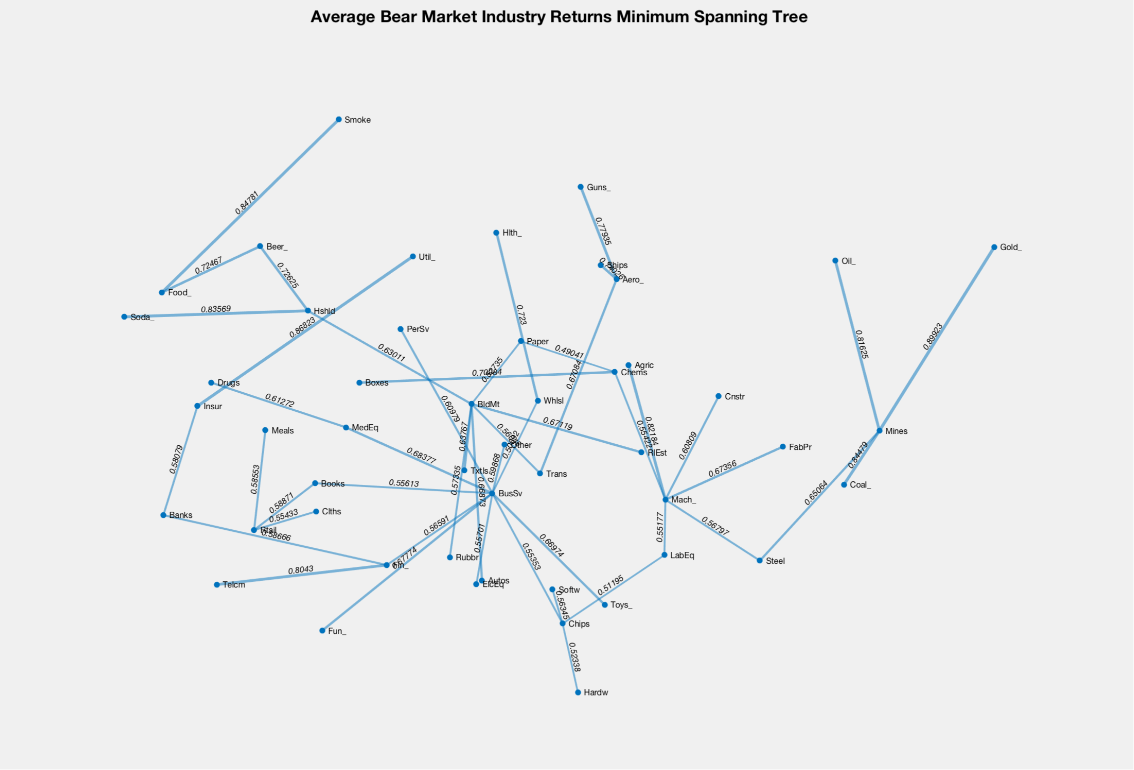


Figure 2: Average Bear Market Network

Bull Mkt MST.png

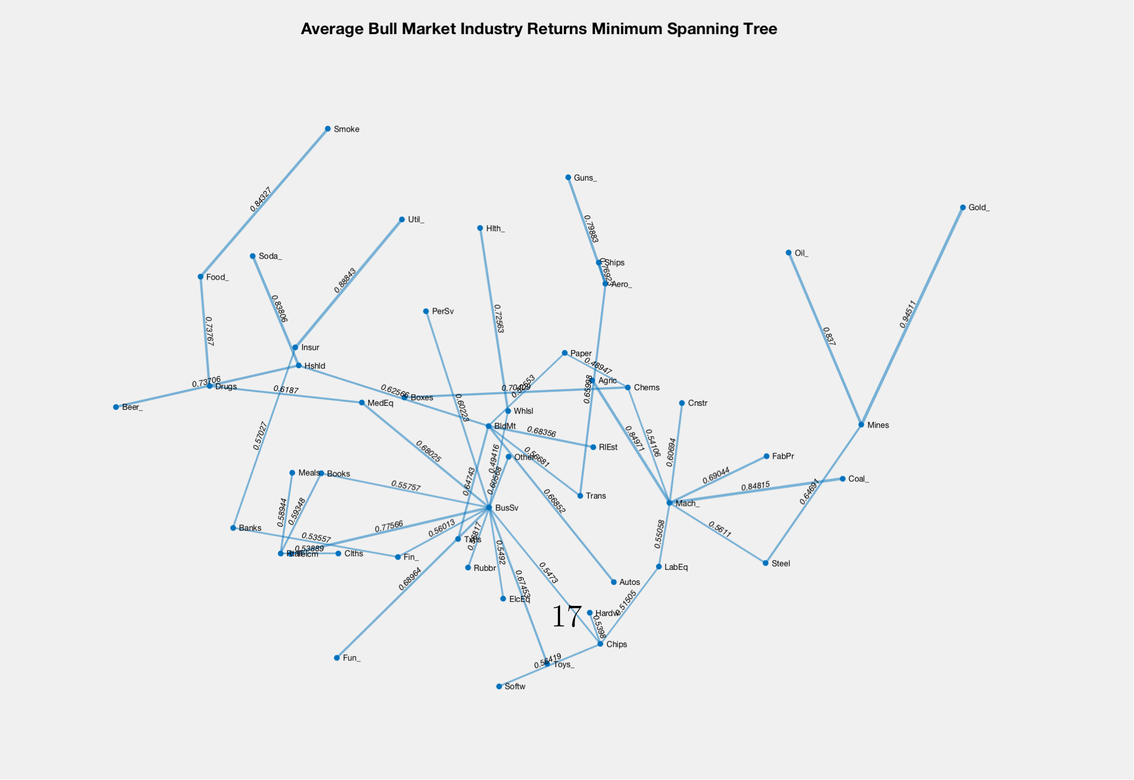


Figure 3: Average Bull Market Network

Neutral Mkt MST.png

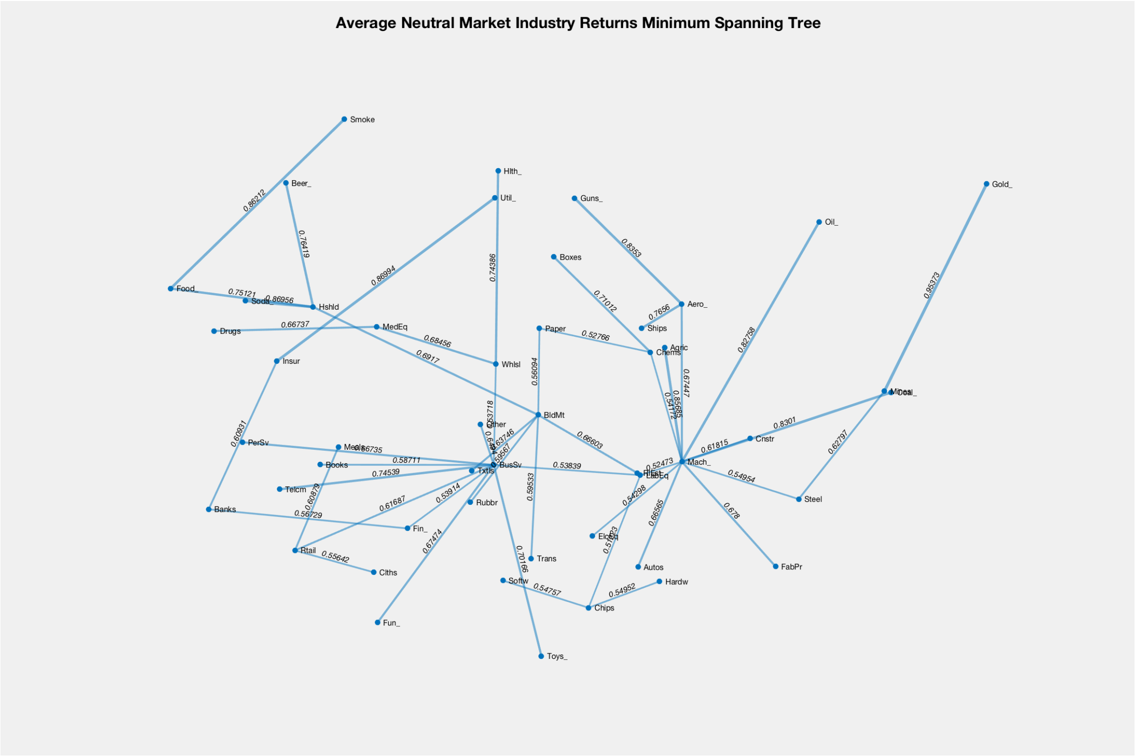
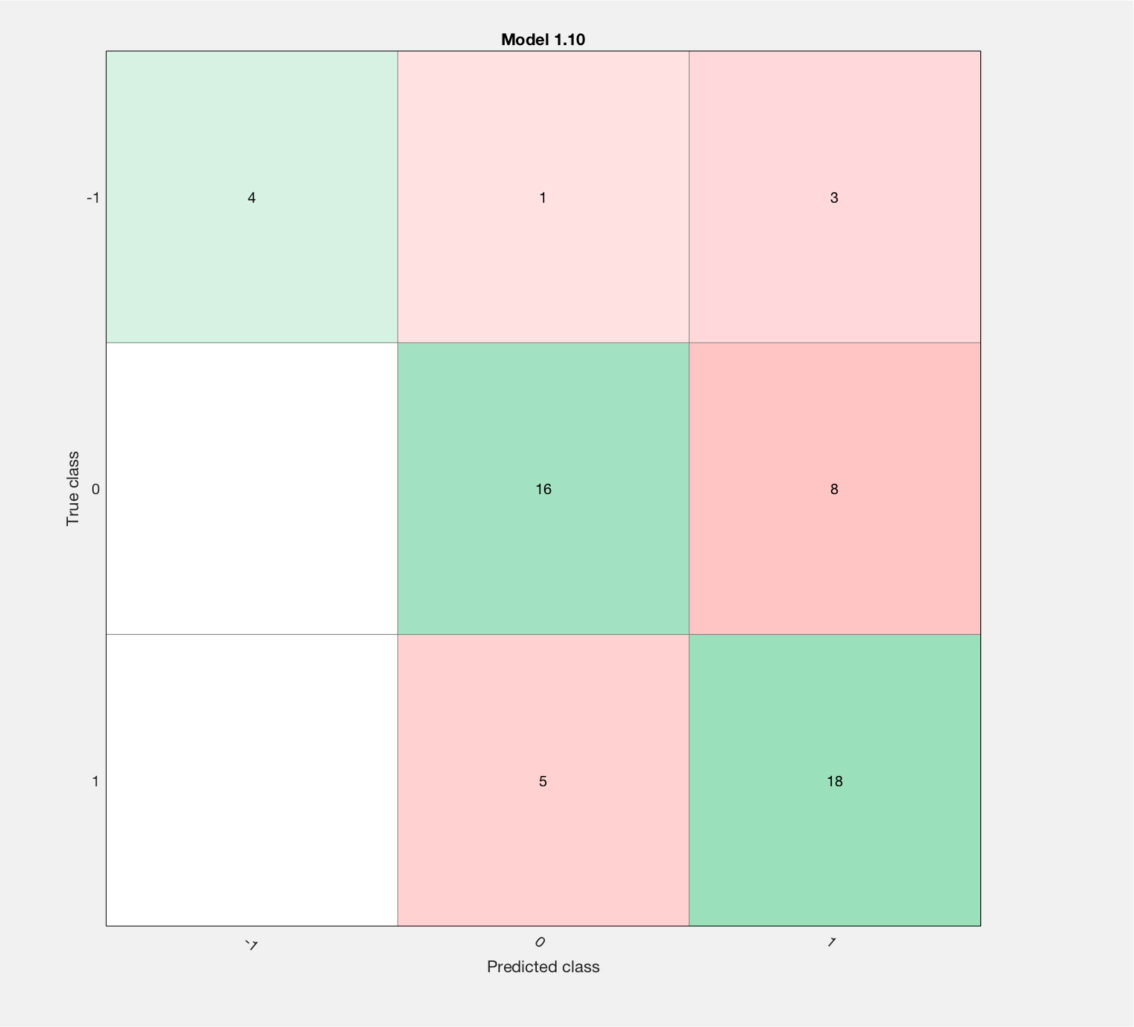


Figure 4: Average Neutral Market Network

Included is the confusion matrix for the Medium Gaussian Support Vector Machine model, consistently one of the highest accuracy models. You can see that the model was especially good at differentiating bear markets from neutral and bull, and was decently good at differentiating between neutral and bull. However, it seems that there may have been a flaw in how our classification heuristic was designed, which then classified periods too positive for bear as bear markets.

Matrix.png



	-1	0	1
-1	4	1	3
0		16	8
1		5	18

Figure 5: Gaussian SVM Confusion Matrix

5 Conclusion

The results we achieved have created a number of possible paths forward. Many improvements can be made upon the methodology of this specific paper, which could then be implemented into a paper focused not on classification, but forecasting regime shifts.

One such improvement would be to create more specific correlation networks. We simplified our results by treating all periods of regimes as equal, though they may have varied wildly in length. Instead, we can extract different portions from each period, such as one month prior to a shift, the first month of a new regime, and then the entire regime like we did. Maybe there exists a pattern of data and network evolution that appears consistently only one month prior to a shift, which we lost by combining with data corresponding to several months before or after.

Another improvement would be to account more strongly for magnitude of changes in returns, rather than just time trends. Our descriptive statistics revealed that during an average bear market, we still saw positive returns, even higher than during an average neutral market. This may have occurred due to a single positive month during the period not causing a reclassification, but being of significantly higher magnitude than the surrounding negative returns. Or possibly, during neutral markets, we saw enough positive returns to classify as neutral, but magnitude-wise, the market was actually overwhelmingly negative and thus deserved a bear classification.

Regarding our results themselves, they appear quite promising. Though 60% to 70% is not stellar, this is actually superior to the personal predictions of many economists. Furthermore, the majority of errors by the model were false positive bulls when it was actually a neutral. The above improvement, accounting more strongly for magnitude of returns, could in turn improve our results, or instead reveal our classifier's accuracy depended on a flawed heuristic.

Our work moves beyond previous work implementing both financial data and correlation networks, by being one of the first to introduce a use for these networks in market classification. We have also created a framework by which a forecasting model for market regimes could be created. This has a number of applications, both theoretical and practical. Theoretically, it opens the door to research on the use of network theory in finance, and how more complex networks and analysis could reveal valuable information. Practically, our work could be used a number of ways. Similar work can be done to show how any subset relate to and influence a larger set, such as industries to the market, national GDPs to international growth, etc. Robust and early classification of market trends provide financial practitioners early warning about hard times ahead, and the all-clear earlier than waiting for a trend to be established in the market itself. Expansion from classifying to forecasting provides similar use, and could also be used by governments and central banks to monitor the future health of their economies and financial markets. This in turn could potentially be linked to GDP/recession-forecasting efforts.

6 Acknowledgement

We would like to thank our advisor, Dr. Michael Aguilar, for his mentorship and guidance. From providing the impetus for this research to aiding in the development

of the econometric model, Dr. Aguilar has provided an immeasurable amount of aid and direction in this process.

References

Bonanno, G., C. G. L. F. and Mantegna, R. N. (2003). Topology of correlation-based minimal spanning trees in real and model markets. *Physical Review. E, Statistical, Nonlinear, and Soft Matter Physics*, 68(4).

Fama, E. and French, K. (1992). The cross-section of expected stock returns. *Journal of Finance*, 47(2):427–465.

Kuosmanen, P. and Vataja, J. (2013). Forecasting gdp growth with financial market data in finland: Revisiting stylized facts in a small open economy during the financial crisis. *Review of Financial Economics*, 23(2):90–97.

Mantegna, R. N. (1999). Hierarchical structure in financial markets. *The European Condensed Matter and Complex Systems*, 11(1):193–197.

Mantegna, R. N. and Stanley, H. E. (1999). *Introduction to Econophysics: Correlations and Complexity in Finance.* Cambridge University Press.

Sammon, J. W. (1969). A nonlinear mapping for data structure analysis. *IEEE Transactions on Computers, C*, 18(5):401–409.