Spring 2019

Optimizing Investment Portfolio Allocation

Nathan Boone
Carroll College

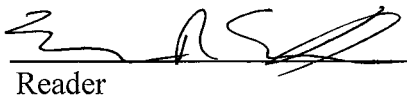
SIGNATURE PAGE

This thesis for honors recognition has been approved for the

Department of Mathematics

Director — 12-10-18 Date

Ted Wendt
Print Name



Reader — 12/10/18 Date

Eric Sullivan
Print Name

Reader — 12/10/18 Date

Peter Larsen
Print Name

Optimizing Investment Portfolio Allocation

Nathan Boone

December 14, 2018

Abstract

This analysis presents a model that could be used to inform a portfolio manager in which sectors of the stock market to invest in for the equity portion of an investment portfolio. The model incorporates a linear program to suggest the optimal combination of sectors to invest in with respect to various constraints. We use a linear projection model to estimate the upcoming month's return of each sector, with monthly changes in economic indicators as inputs. The time frame for the historical returns data that we use to build the model runs from $8/1/2008 - 9/1/2018$. Additionally, we apply a Monte Carlo simulation to the projected returns over one hundred thousand iterations. The model suggests which combination of sectors may offer the highest return on investment with consideration to the potential volatility of each sector.

Contents

1 Introduction

Optimizing an investment portfolio has long been a highly complicated task for money managers and individual investors. While there are multiple standard techniques and approaches that have evolved over the past few decades, there still remains the inevitable nature of black swan events, stock market crashes, and economic downturns that make portfolio management a very difficult task. After all, if it were easy, we would all be rich and have no need to refine our methods and approaches. One of the accepted techniques used by rational investors is to diversify their holdings to minimize exposure to risk, thus increasing the probability of an acceptable return. To achieve this concept of diversification, portfolio managers seek to spread their investments across a variety of investment types: stocks, bonds, real estate, commodities, foreign securities, index funds, government issues treasury bills, precious metals, and derivatives to name a few. The types of investments a portfolio might include, the specific choices within those investments, and the percentage of a portfolio to invest in each area is a task that is much greater than the scope of this analysis.

This project seeks to build a model that optimizes the equity portion of an investment strategy of a portfolio. We formulate a linear program that is designed to help inform a decision on the balance of investment into the 11 major stock market sectors, as well as a miscellaneous sector. The model uses historical returns over the last ten years to project the monthly returns of each sector, with the idea that each quarter a portfolio manager might re-balance the holdings. The model focuses on a strategy of allocating equity investments across the major sectors, and does not suggest any other forms of investment such as bonds, real estate, or treasury notes. It is assumed that 100% of the funds available can be used for the decision, and suggests sector in which to invest, as opposed to specific stocks within each sector. We used a multiple regression model with a 10-fold cross validation to inform the projected returns of each sector. We implemented a "best subset" variable selection technique on the each of the sectors, and tested that against the monthly returns of four economic indicators: Unemployment rate (UMENP), gross domestic product (GDP), consumer price index (CPI), and the Daily Federal Funds rate (DFF). In addition to selecting variables against the monthly returns, we tested against a three month, and six month moving average of changes in each economic indicators. Thus, the model for each sector may contain any combination of the four variables, with respect to the monthly change, three month moving average, or six month moving average of change.

In Section 2, we will highlight the background and motivation behind this analysis. In Section 3, we will discuss the data sources that we use for historical stock returns, and how we went about calculating the sectors monthly returns. Additionally, we will explain which economic indicators will be used in the projection models. Section 4 discusses the type of projection model we used to calculate expected returns of each sector, and explain the variable selection technique we used to chose appropriate predictors. Next, in Section 5, we will present a unique approach to classifying sectors as neighbors, and how those classifications will be used as a constraint to the linear program. Section 6 discusses the formulation of the linear program, and how we built the objective function and constraint matrix. In Section 7 and Section 8 we present the results of the linear program, and discuss how to interpret our findings. Finally, Section 8 highlights some of the short comings associated with this model.

2 Background and Motivation

A common theme among the financial world is to be diversified with your investments. However, sometimes this term is used as a "be all end all", and lacks a specific direction as to what diversification actually looks like. In this project, we seek to offer a unique, data driven approach to help inform and optimize investment decisions. We will limit our analysis to focus on how to allocate investments into 12 sections- consisting of the 11 major sectors of the stock market, and a miscellaneous sector. While it is important to recognize that a necessary part of any investment strategy is to have a mix of securities, fixed income bonds, treasury bills and other types of investments, we will limit our analysis to domestic securities. Additionally, we will not seek to suggest which specific stocks are suitable for an investment decision, but rather focus on the different sectors of the stock market as a whole.

The goal of this project is to create a model that can inform the decision making of an individual, and relies on a few general assumptions. This model does not attempt to inform the individual how frequently different investment decisions should be made, and does not account for the investment horizon of a portfolio. The model focuses on monthly return projections, with the assumption that each quarter or so, a portfolio manager might look to rebalance the holdings based on information from the model. The model also does not account for the level of risk each individual may have.

3 Data Sources and Relevant Software

3.1 Data Sources

The entirety of this analysis was completed using R. To retrieve the stock prices and returns data we used the "TTR" package and the "stocksymbols()" command within that package, to download stock data of symbols traded on the major U.S. markets: AMEX, NASDAQ and the NYSE. This package offers various statistics for each stock, and we utilized the stock symbols, along with respective market capitalization values and sector classification. The package only had access to current market cap values and the sectors included the 11 standard sectors of the market, a miscellaneous sector, and any stock that did not have a value for either of these categories was omitted from this analysis.

This package also offers a function, "getSymbols", used to retrieve historical closing prices for stocks traded on the major public markets. In conjunction to this function, "monthlyReturn" offers the capability to calculate monthly returns for stocks across the defined time interval. This package can calculate various intervals of returns over time, for this analysis we used monthly returns as the standard. Finally, we used various commands from the "tidyverse" package to aggregate the stocks by their respective sector with monthly return data. Using the aggregated monthly return values, we calculated the weighted average for each month to represent the sector's monthly return.

3.2 Market Capitalization Classification

Typically, sector averages are weighted by each stock's market capitalization for the respective period. However, we did not have access to continuous market cap values, so instead used a

Classification	Cutoff Value	Weight
Mega	\$100 B	2,000
Large	\$10 B	200
Mid	\$2 B	40
Small	\$300 M	6
Micro	\$50 M	1
Nano	< \$50M	1

Table 1: Cap size classifications and cutoff values. The stock is classified by the if it has a market cap greater than the cutoff value, and less than the class above it. Nano cap is any value less than \$50 million. Weight is the relative multiple of the smallest cutoff value to all other values.

different approach to estimate each stock's appropriate market cap weight. Rather, we used the current market cap of each stock to assign a market cap classification, using industry standard cap sizes and cutoff values, as seen in Table 1.

Since a stock's market valuation is a direct reflection of its price per share, it is not reasonable to assume that its has maintained its current market cap value over the past ten years. However, the assumption is that most stocks will have remained in a similar classification over the time interval. For example, today Apple stock is classified as a mega cap stock, with almost \$1 trillion in market cap, and in 2008, Apple was valued at roughly \$150 billion, which is still classified as a mega cap stock. Although the relative weight has changed, we assume that most stocks have remained in the same market class. The weight assigned to each class is the relative quotient of each cutoff value to the smallest cutoff value, which is \$50 million. However, since the minimum value for a nano class is technically 0, it is weighted equally with micro. We believe this is a reasonable assumption, because it has a negligible impact on the sector's final return value. With each of the stocks assigned an appropriate weight and sector, we used the historical monthly returns to estimate a weighted average sector return by month.

3.3 Economic Data

Next, we downloaded historical data for the four economic indicators that will be used in the projection models. The indicators include unemployment rate (UNEMP), Gross Domestic Product (GDP),Consumer Price Index (CPI, or Inflation), and the effective federal funds daily rate (DFF). We chose to use these indicators as a basis for this analysis, as they pertain to various aspects of the economic landscape. Since the stock returns are calculated on a monthly basis, we needed to analyze the economic indicators on a monthly basis as well. Two of the indicators, UMEMP and CPI, are reported monthly, and the monthly percent change is available. DFF is reported daily, but we only considered the monthly percent change; change from the first day to the last day of each month. Additionally, GDP is reported quarterly, so we used the quarterly percent change and a simple linear interpolation to fill in the two months between each quarterly change. Finally, we aggregated the economic indicator and stock returns by month, with all units representing percent change from the month prior.

4 Projecting Expected Returns

The goal of the projection model is to predict the sector return of the proceeding month using inputs from trends in the economic indicators. We then used the projected returns for each sector

Sector	Return Timeframe	Variable	Adj R^2
Finance	Monthly	CPI, GDP	0.186
Basic Industrials	Monthly	CPI, UNEMP, GDP	0.247
Capital Goods	Monthly	CPI, GDP	0.182
Technology	Monthly	CPI	0.155
Energy	6 month MA	UNEM, GDP , CPI	0.065
HealthCare	Monthly	CPI, DFF	0.143
Public Utilites	Monthly	CPI, GDP	0.137
Consumer Services	Monthly	CPI	0.185
Miscellaneous	Monthly	CPI, DFF	0.123
Consumer Non Durables	Monthly	CPI, GDP	0.207
Consumer Durables	N/A	N/A	0.010
Transportation	Monthly	CPI, UNEMP	0.057

Table 2: Results of the variable section process that returns most appropriate predictive variable(s). Each sector was tested against the monthly returns, and both the three month and six month moving average of returns. Note that Consumer Durables returned results of an adjusted R^2 value less than 0.01; so we are excluding a projection model for that sector.

to inform the objective weights of the linear program. We wanted to offer these models enough flexibility to account for the fact that different sectors, or industries, of the market are impacted in different ways by trends in the economy. Since each sector is has unique characteristics, it was important to address the fact that using every economic indicator might be inappropriate, and that some sectors may respond to changes in the economy with delay. Therefore, we implemented a "best subset"" variable selection process on each of the sectors, considering three ways of presenting the economic indicators: monthly change, a three month moving average and a six month moving average of changes in the indicators. This meant that each sector would consist of three models, and as a training set we excluded the returns from the current year- 2018.

The results of the variable selection are presented in Table 2. Each sector suggested that a unique combination of indicators and time frame were best suited to predict the returns of that sector. Our criteria for selecting the combination of best variables was to maximize the adjusted R^2 for each possible combination of variables and monthly return. As shown in Table 2, the adjusted R^2 values generally fall into the range of 0.12-0.25; with the exception of three sectors. The results of the Consumer Durables sector suggested that no combination of the variables and time frame would be an appropriate predictor of the sector's return, as the highest adjusted R^2 value was less than 0.01. Therefore, we did not build a projection model for that sector.

Using the results from the variable selection process, we built a predictive model for each sector (excluding Consumer Durables). The models used a 10 fold cross validation technique over the entire data set, and predicted using a linear regression model. The cross validation method is useful for minimizing the potential over-fitting of a model. The results of this model inform the weights of the objective function of the linear program by projecting the next month's return for each sector, based on the current values of the economic indicators. However, since each sector should have a degree of error associated with the return projection, we used the historical standard error over the ten year period to estimated the error of the projection. We will discuss how we applied the error to a random normal distribution further, in Section 6. For the Consumer Durables sector, we assume the expected return will be the average monthly return over the entire time frame, and assume the historical standard error for the projection. The 10 fold cross validated model does not produce coefficient values for each of the models. Therefore,

we built a general linear model over the entire data set in order to produce an equation for each sector that could be used outside of this data set.

Additionally, we tested each of the models to see if any of the variables demonstrated multi-collinearity; our results showed that the variable inflation factor was negligible for each model. The coefficients of the variables for each model, along with various model statistics are available in the Appendix, in Section C.

5 Sector Neighborhoods

Diversification has proven to be a key component of any investment strategy. However, we wanted to see if any of the sectors have a tendency to trade together. Thus, the purpose of this section of our analysis is to identify which sectors could be grouped into "neighborhoods". Investing in too many sectors within the same neighborhood could lead to an over-diversification of a portfolio, increasing exposure to risk.

To develop a more comprehensive model, we analyzed the trends of the monthly returns of sectors relative to each other. This offers an alternative approach to the direct relationship between the economic indicators and sector returns. This technique seeks to identify trends in the returns of sectors, and highlight how often any two sectors move together over time. Our approach to analyzing these relationships identifies both direction and magnitude of the movements in sectors by showing the cross over of two techniques-which we refer to as the Distance Technique, and the Direction Technique.

5.1 Distance Technique

With the first technique, we examined the movements of the sectors showing the "distance" between each sector's returns over the 10 years of data. While we will use the term distance to describe the correlation between a pair of sectors, we really are measuring the angle between two sectors by using a vector projection. To do this, we began with the monthly return data for each sector, and normalized each element in the matrix, call this matrix R (returns). Then, we multiplied $R^T R$, which produced a 12×12 matrix, D (distance). This method represents a series of vector projections of each sector-vector pair. By approaching this process as a vector projection, we are able to quantify the correlation of each pair of sectors. Each element in D represents how each sector relates to each other with a value indicating the sum product of the normalized movement relationship between two sectors for each month. This value represents a correlation between the two vectors, where the inverse cosine of this value would be the angle between the two vectors. In other words, the values ranged from 0-1, where 0 signals a perfectly negatively correlated stock, and 1 signals a perfect positively correlated stock. However, to actually be able to model which sectors might trade as neighbors, we arbitrarily chose a cut off correlation value of 0.70; any two sectors with a correlation greater than this value are considered a neighbor to each other. This produced a binary table, where a 1 represents a neighbor, as presented in Table 8 of Appendix B. Table 3 shows the results of this method in list form. Four of the sectors do not have any neighbors, while the rest of the sectors generally seem to be correlated to each other.

Sector Neighbors - Distance Technique.

Sector #	Sector Name	Neighbors
1	Finance	2,3,4,6,10
2	Basic Industrials	1,3,4,6,8,9,10
3	Capital Goods	1,2,4,6,9,10
4	Technology	1,2,3,6,9
5	Energy	None
6	HealthCare	1,2,3,4,9,10
7	Public Utilites	None
8	Consumer Services	2,10
9	Miscellaneous	2,3,4,6,10
10	Consumer Non Durables	1,2,3,6,8,9,
11	Consumer Durables	None
12	Transportation	None

Table 3: Results of sector neighbors using the distance technique with a cutoff value of 0.70.

Sector Neighbors- Direction Technique

Sector #	Sector Name	Neighbors
1	Finance	3,4,6,11
2	Basic Industrials	5
3	Capital Goods	1,4,6,7,9,11
4	Technology	1,3,6,7,9,11
5	Energy	2
6	HealthCare	1,3,4,7,9,11
7	Public Utilites	3,4,6,9,11
8	Consumer Services	None
9	Miscellaneous	3,4,6,7,11
10	Consumer Non Durables	None
11	Consumer Durables	1,3,4,6,7,9
12	Transportation	None

Table 4: Neighbor matrix for direction technique using a cutoff value of 70.

5.2 Direction Technique

For the second technique, we counted the number of months in which any two sectors had the same direction of positive or negative returns, respectively, regardless of magnitude of those values. This means that any two sectors could trade at most 120 times (120 month time frame) together in the same direction, or -120 times which means that every month they had opposite signs of returns. This model does not account for magnitude, which helps reduce the skew that the first model might have for sectors with large swings relative to more stable sectors, while focusing on the directional movements of two sectors. The first step was to reduce the monthly return value to -1 or 1 if the return was negative or positive, respectively. Then, by using a similar approach as discussed in Section 5.1, we multiplied $R^T R$, and produced a correlation matrix. Again, we chose an arbitrary cutoff value of 70 to classify each pair of sectors as neighbors; a value greater than 70 represents a neighbor. The direction matrix is presented in Table 10, in Appendix B. Table 4 shows a list form of the results of this method. The results resemble the pattern of the distance method; a few of the sectors have zero or one neighbors, while the rest show a tendency to all be grouped into the same neighborhood.

Final Sector Neighbors

Sector #	Sector Name	Neighbors
1	Finance	3,4,6
2	Basic Industrials	None
3	Capital Goods	1,4,6,9
4	Technology	1,3,6,9
5	Energy	None
6	HealthCare	1,3,4,9
7	Public Utilites	None
8	Consumer Services	None
9	Miscellaneous	3,4,6
10	Consumer Non Durables	None
11	Consumer Durables	None
12	Transportation	None

Table 5: Neighbor results after combining both techniques.

5.3 Combining the Two Techniques

Finally, the combination of these two models helps offer a more complete picture of identifying the trends in similar sectors. For two sectors to be considered neighbors, both models would need to suggest that they are in the same neighborhood. Since the output of neighborhoods are binary after an arbitrary cutoff value, the strength of correlation between each neighbor is not consider. Simply, if both models suggest a relationship, then the two sectors are considered to be neighbors. The result of combination of both techniques is presented in Table 5.

This model contributed to an aspect of the constraints of the linear program. For example, one of the constraints might suggest that you should not invest in more than three neighbors, since these sectors tend to trade together in the long run, which suggests a false sense of diversification. More tables that show the correlation matrices for each of the techniques is available in Appendix B.

6 The Linear Program

A linear program is an optimization technique that seeks to minimize or maximize an objective function, relative to various constraints and bounds. For this model, we are seeking to maximize the projected monthly return on an investment portfolio by investing across the major sectors of the stock market. We applied various constraints to this model including:

- A single sector can only account for up to 25% of the total investment.
- We cannot invest in more than three "neighbors".
- Less than 100% of the available capital can be used.

We chose to limit maximum percentage of an individual sector to 25% arbitrarily, but believe that it is reasonable to invest 100% of capital in at least four sectors. Additionally, we chose three neighbors as the cutoff value to prevent a possible over-diversification of investment. Finally, we allow the model flexibility to allocate less than 100% of capital across the sectors, in the event that only a few sectors have positive expected returns.

To inform the weights of the objective function, we used the most recent month of economic data as the input to the individual projection models for each sector. As previously discussed, we did not build a model for the Consumer Durables sector. Instead, for this sector, we used the historical mean of returns over the time interval. One of the great challenges of predicting returns on equities is that there are a variety of outside forces that cannot be totally accounted for. To address the volatility of each sector, we incorporated a Monte Carlo simulation to our projections. A Monte Carlo simulation is a useful technique when modeling the uncertainty of the projected returns on the sectors. To implement this simulation, we assume a normal distribution for the monthly returns each sector, with the expected return representing the mean and the standard error of the sector's historical monthly returns as one standard deviation. For each sector simultaneously, the Monte Carlo simulation draws a randomly distributed value from each sector, and uses that result as the representative return of each sector for one iteration of the linear program. Then, we simulated this process over 100,000 iterations while holding the constraints constant. In other words, the Monte Carlo method allows the model to optimize 100,000 scenarios of the returns within a random distribution of each sector, helping to identity which sectors might be more stable and unstable. A discussion of how we tested for normal distribution of the monthly sector returns is available in Appendix D.

Additionally, we used the results of the neighborhood analysis to inform one of the constraints of the LP. We set this constraint to limit investment to three neighbors. The logic behind this constraint is that if sectors tend to trade together as neighbors, then investing in the same neighborhood would increase exposure to risk. We arbitrarily chose the cutoff of three neighbors, and when we tested the LP by only varying this value, there was little to no change in the outcome; suggesting that this is not a binding constraint.

Another constraint is that you cannot invest more than 25% available capital in a single sector. Also, you do not have to invest all 100% of available capital. This would only apply if many of the returns were projected to be negative, and after maxing out the positive sectors you had remaining capital, the model would suggest just holding some of that capital out of the investment.

7 Results

As previously discussed, we simulated the Monte Carlo and LP process over 100,000 iterations. The output of the results is simply the average percentage of capital invested into each sector for each iteration. Table 6 presents the results of the analysis. The average value reflects the average optimal percentage of the portfolio to invest into each sector. As reference, the maximum value we could expect for the average selection percentage is 25%, due to the constraint that limit investment into a single sector at 25% of the available capital. This means that transportation was selected as the most optimal sector in a majority of the simulation, suggesting in theory that transportation offers the best value of risk-reward. The average expected return over the simulation is 3.75% for the proceeding month.

There are two ways in which we could practically apply the results of this model to inform an investment decision. One way would be allocate investment by the average selection value across all the sectors. In theory, this would violate the neighbor constraints, since we would be investing in all of the sectors (the sum of the average selection is 100%). However, by following

these distributions exactly, the potential risk of over-diversification would have been account for in to some degree over the course of the simulation. The second way we could interpret these results is to chose four of the most optimal sectors based on the simulation. For example, we could equally invest a quarter of our portfolio into each of the first four sectors presented in Table 6.

	Names	Avg selection
1	Transportation	22.875
2	Technology	17.875
3	Miscellaneous	12.875
4	ConsumerDurables	12.25
5	BasicIndustries	11
6	ConsumerServices	9.125
7	CapitalGoods	5.50
8	HealthCare	5.25
9	Finance	2.875
10	Energy	0.25
11	ConsumerNonDurables	0.125
12	PublicUtilities	0
-	Expected Return	3.75

Table 6: Aggregated results of linear program over 100,000 iterations of the Monte Carlo simulation.

8 Discussion and Analysis

The results of this analysis should be used as tool to help inform a portfolio manager to which sectors may offer a positive return on investment for the coming month and quarter. This model is not intended to suggest that any specific approach to interpreting and applying these results. Additionally, there are a variety of sector specific investment decisions that an individual would need to make before investing in a sector. It is also recommended that the results of this model be used with the same data set that we used for the analysis. There are many different names and ways to classify the sector or industry that a company belongs to, and many companies have diverse operations that fit into multiple sector definitions. See Appendix A for more precise explanation of definitions and examples of stocks that are in each defined sector based on this database.

Model Challenges and Future Work

The model offers a unique data driven perspective into how to diversify an equity portfolio, however does have a few challenges that are worth noting. First, our analysis of the returns and classification of stocks relies on a single data set from the "TTR" package in R. While we are confident that the historical stock prices are accurate, the sector and industry classifications do have overlap and discrepancies with other financial analyzing sources. For this reason, the results of this model should be used in conjunction with the same data set, in order to properly identify which sector each stock is classified. Additionally, the sector titles used in this data set

are different than those by defined by the Exchange Trade Fund database[1] source that we used to define each sector. The sectors defined in the dataset that we used differed from those in the ETFdb source, so we had to try and match them despite clear overlap between descriptions and companies.

Another challenge of this model is that it focuses on monthly returns of stocks and economic indicators to predict the returns of sectors and optimize the investment decision. This model could offer a more robust analysis if we repeated the process using quarterly, semi annual, or annual returns of sectors and economic indicators in conjunction with the monthly returns model. Since our model only projects out the proceeding month, this model may imply a short term view on investing, which generally is a less ideal approach to having a disciplined and diverse investment strategy.

As discussed in Section 4 and Table 2, the sectors showed a relatively weaker Adj R^2 value than we anticipated, and some of the variable used in the models did not demonstrate statistical significance. The analysis could be improved by analyzing more economic indicators, and other factors that might better explain the returns of sectors across various time frames. We also highlight, in Section C, that some of the sectors have high leverage points, which have a disproportionate effects on the model relative to the other data points. It would be good to removed these high leverage points and reconstruct a model without them.

9 Sources and Acknowledgements

We would like to express our gratitude to Dr. Ted Wendt, for his guidance and support of this project. We would also like to thank Dr. Eric Sullivan, and Dr. Peter Larsen, for serving a final readers for this report.

[1] "Select The 11 Sectors Of the Stock Market" https://etfdb.com/etf-education/the-10-sectors-of-the-stock-market/

A Sector Descriptions

- Finance: The financial sector consists of banks, investment funds, insurance companies and real estate firms, among others. In general, the majority of the revenue generated by the sector comes from mortgages and loans that gain value as interest rates rise. Examples of the largest companies in this sector include major banks such as JP Morgan Chase (JPM), and Bank of America (BAC).

- Basic Industrials: The basic industrials sector consists of mining, refining, chemical, forestry and related companies that are focused on discovering and developing raw materials. Since these companies are at the beginning of the supply chain, they are vulnerable to changes in the business cycle. Also referred to as "Materials" or "Consumer Defensive". The largest companies in this space include: Procter and Gamble (PG), Unilever (UN) and BHP Group (BHP).

- Capital Goods: The industrial sector consists of aerospace, defense, machinery, construction, fabrication and manufacturing companies. In general, the industrys growth is driven by demand for building construction and manufactured products like agricultural equipment. Also referred to as "Industrial". The largest companies in this space include Boeing Co. (BA) and Toyota Motor Co. (TOM).

- Technology: The technology sector consists of electronics manufacturers, software developers and information technology firms. In general, these businesses are driven by upgrade cycles and the general health of the economy, although growth has been robust over the years. Leaders in this industry include the likes of Apple (AAPL), Microsoft (MSFT) and Google or Alphabet Inc. (GOOG).

- Energy: The energy sector consists of oil and gas exploration and production companies, as well as integrated power firms, refineries and other operations. In general, these companies generate revenue thats tied to the price of crude oil, natural gas and other commodities. Leaders in this sector include Exxon Mobile (XOM), Total S.A (TOT), and Chevron Corp. (CVX).

- Consumer Services The consumer services sector consists of retailers, media companies, consumer service providers, apparel companies and consumer durables. In general, these companies benefit from an improving economy when consumer spending accelerates. Also known as the "Consumer Discretionary" sector. Leaders in this group include Amazon Inc. (AMZN), Walmart Inc. (WMT), The Home Depot (HD, and Comcast Corp. (CMCSA).

- Health Care: The health care sector consists of biotechnology companies, hospital management firms, medical device manufacturers and many others. In general, the sector is considered to be both a growth opportunity and defensive play since people will always require medical aid. The leaders of this sector include Johnson Johnson (JNJ), United Health Group (UNH), and Pfizer Inc. (PFE).

- Public Utilities: The utilities sector consists of electric, gas and water companies as well as integrated providers. In general, the sector generates consistent recurring income by charging consumers and businesses that provide higher-than-average dividend yields. This

data set also includes telecommunications equipment companies in this sector, such as Verizon Communications (VZ) and ATT Inc (T). However, more traditional leaders of this sector include NextEra Energy Inc. (NEE) and Duke Energy Corp (DUK).

- Consumer Non Durables: The consumer non durables sector consists of food and beverage companies as well as companies that create products consumers are unwilling to cut from their budgets. In general, these companies are defensive plays capable of withstanding an economic downturn. Also referred to as "Consumer Staples". Leaders in this sector include The Coca-Cola Company (KO), PepsiCo (PEP), and Nike Inc. (NKE).

- Consumer Durables: The consumer durables sector consists of companies specializing in building products, container and packaging, household and personal products, and diversified industrials. Companies in this sector may contain overlap in their operations with companies in the capital goods or industrial sector. Leaders in this sector of the data set include AMETEK Inc (AME), Kimberly-Clark Corp. (KMB), Colgate-Palmolive Co. (CL) and ABB Ltd (ABB).

- Transportation: The transportation sector consists of companies whose operations focus on delivery services of goods via trucking, air, marine and railroad freight. Top companies in this sector include: Union Pacific Corp. (UNP), United Parcel Service (UPS), and FeDex Copr (FDX). This sector may contain overlap in definition with the Capital Goods sector.

- Miscellaneous This sector includes a variety of companies whose operations did not fit into any one of the sectors because of overlap. Many of these companies include Financial Services and Technology, which might overlap Technology and Financials. Leaders of this sector, as defined by this data set, include Alibaba Group Holding Limited (BABA), Visa Inc (V), PayPal Holdings Inc. (PYPL).

B Sector Neighborhood Matrices

Index of sector numbers

Sector Index	Sector Name
1	Finance
2	BasicIndustries
3	CapitalGoods
4	Technology
5	Energy
6	ConsumerServices
7	HealthCare
8	PublicUtilities
9	Miscellaneous
10	ConsumerNonDurables
11	ConsumerDurables
12	Transportation

Table 7: To be used as reference for neighbor matrices indexing.

Sector distance matrix

	1	2	3	4	5	6	7	8	9	10	11	12
1	0.00	0.82	0.87	0.75	0.16	0.82	0.44	0.58	0.65	0.72	0.27	0.37
2	0.82	0.00	0.86	0.73	0.24	0.81	0.46	0.73	0.70	0.77	0.29	0.37
3	0.87	0.86	0.00	0.85	0.19	0.90	0.56	0.63	0.79	0.75	0.31	0.36
4	0.75	0.73	0.85	0.00	0.21	0.82	0.49	0.58	0.80	0.70	0.32	0.22
5	0.16	0.24	0.19	0.21	0.00	0.21	0.08	0.16	0.11	0.19	0.06	0.13
6	0.82	0.81	0.90	0.82	0.21	0.00	0.52	0.67	0.79	0.82	0.33	0.28
7	0.44	0.46	0.56	0.49	0.08	0.52	0.00	0.42	0.45	0.47	0.17	0.17
8	0.58	0.73	0.63	0.58	0.16	0.67	0.42	0.00	0.58	0.79	0.25	0.19
9	0.65	0.70	0.79	0.80	0.11	0.79	0.45	0.58	0.00	0.71	0.29	0.21
10	0.72	0.77	0.75	0.70	0.19	0.82	0.47	0.79	0.71	0.00	0.29	0.13
11	0.27	0.29	0.31	0.32	0.06	0.33	0.17	0.25	0.29	0.29	0.00	0.06
12	0.37	0.37	0.36	0.22	0.13	0.28	0.17	0.19	0.21	0.13	0.06	0.00

Table 8: Sector distance matrix which should be read by column. The higher value represents a strong correlation of similar movement between two sectors. The default value that we chose is 0.70 to classify a pair of sectors as neighbors. For example, in column 1, sector 4 would be classified as a neighbor to sector 1, but sector 9 would not.

Sector distance rank matrix

	1	2	3	4	5	6	7	8	9	10	11	12
1	0	2	2	4	7	4	7	6	6	5	7	1
2	2	0	3	5	1	5	5	2	5	3	4	2
3	1	1	0	1	5	1	1	4	2	4	3	3
4	4	5	4	0	3	2	3	5	1	7	2	5
5	11	11	11	11	0	11	11	11	11	10	10	10
6	3	3	1	2	2	0	2	3	3	1	1	4
7	8	8	8	8	10	8	0	8	8	8	9	8
8	7	6	7	7	6	7	8	0	7	2	8	7
9	6	7	5	3	9	6	6	7	0	6	6	6
10	5	4	6	6	4	3	4	1	4	0	5	9
11	10	10	10	9	11	9	9	9	9	9	0	11
12	9	9	9	10	8	10	10	10	10	11	11	0

Table 9: Sector distance rank matrix which should be read by column. The lower value of rank implies the sector (by row) which is the closest neighbor to the respective sector (by column). Note that this matrix should not be symmetric about the diagonal. For example, sector 3 is the closet neighbor of sector 1; but sector 1 is the second closest neighbor to sector 3.

C Projection Model Statistics

Table 12 presents the coefficient values for each of the variables that were used in the projection models. Table 13 presents a few values that represent error and fit of the models, as well as the variables that demonstrate statistical significance for each model. The MSE value is a result of the average error using the 10 fold cross validation approach, while the adjusted R^2 and significance values are from the general linear model.

Additionally, Figure 1 shows three plots of the residuals for their respective models. We only included these three plots for the sake of space and readability. The first plot, Finance, is a good representation of what all other models look like (with the exception of Energy and Transportation). We do not see any correlation with the residuals, as reinforced by the flat

Sector direction matrix

	1	2	3	4	5	6	7	8	9	10	11	12
1	0	56	84	80	50	74	62	38	64	56	78	52
2	56	0	68	56	70	62	54	58	52	56	62	44
3	84	68	0	88	58	94	78	38	76	56	94	52
4	80	56	88	0	50	86	78	46	76	52	82	44
5	50	70	58	50	0	48	36	44	38	42	52	54
6	74	62	94	86	48	0	76	56	70	66	88	42
7	62	54	78	78	36	76	0	48	70	58	76	34
8	38	58	38	46	44	56	48	0	54	66	48	26
9	64	52	76	76	38	70	70	54	0	60	78	44
10	56	56	56	52	42	66	58	66	60	0	54	24
11	78	62	94	82	52	88	76	48	78	54	0	42
12	52	44	52	44	54	42	34	26	44	24	42	0

Table 10: Values indicate sum of correlated and non-correlated directional movements of monthly returns between two sectors. This means in two sectors both saw positive returns in a month, regardless of magnitude, then that month is added to value, if they do not have the same direction of returns that month, that month is subtracted. Since we analyzed 120 months of returns, a value of 120 would signify perfect positive correlation, while -120 would signify perfect negative correlation. The cutoff value for this technique is 70. This matrix is symmetric about the diagonal.

Direction Neighbor Matrix

	1	2	3	4	5	6	7	8	9	10	11	12
1	0	0	1	1	0	1	0	0	0	0	1	0
2	0	0	0	0	1	0	0	0	0	0	0	0
3	1	0	0	1	0	1	1	0	1	0	1	0
4	1	0	1	0	0	1	1	0	1	0	1	0
5	0	1	0	0	0	0	0	0	0	0	0	0
6	1	0	1	1	0	0	1	0	1	0	1	0
7	0	0	1	1	0	1	0	0	1	0	1	0
8	0	0	0	0	0	0	0	0	0	0	0	0
9	0	0	1	1	0	1	1	0	0	0	1	0
10	0	0	0	0	0	0	0	0	0	0	0	0
11	1	0	1	1	0	1	1	0	1	0	0	0
12	0	0	0	0	0	0	0	0	0	0	0	0

Table 11: Binary matrix in which a 1 represents a neighbor relationship between a pair of sectors. These neighbors are a result of the direction technique, presented in Table 10. This matrix is symmetric about the diagonal.

trend line. However, in both the Energy and Transportation plots, we can see that they have a few outliers. These points are high leverage points, and may have an disproportionate effect on our model relative to the other data points. However, despite the high leverage points, the flat trend lines suggest there is no correlation in the residuals. We should note that the adjusted R^2 values for the Energy and Transportation sectors are an order of magnitude higher than the other models. We believe that this is partially due to a few high leverage points in their historical returns. Removing these high leverage points may be a valuable aspect of future work.

Coefficients of Variables from Projection Models

Sector model	intercept	CPI	UNEMP	GDP	DFF
Finance	-0.015	9.017	-	1.364	-
Basic Industrials	-0.011	10.098	0.284	1.089	-
Capital Goods	-0.002	7.611		0.686	-
Technology	0.007	7.480	-	-	-
Energy	-0.015	-	1.232	5.594	0.020
HealthCare	0.011	6.908	-		-0.013
Public Utilites	-0.003	3.684	-	0.438	
Consumer Services	0.007	6.876	-	-	-
Miscellaneous	0.013	5.942	-	-	-0.012
Consumer Non Durables	-0.001	5.744	-	0.392	-
Consumer Durables	-	-	-	-	-
Transportation	0.009	10.861	-	-	-

Table 12: Coefficients of models for each sector. Each model uses a different combination of predictive variables. For the Consumer Durables sector, non of the variables seemed to be appropriate predictors, based on the best subset selection process outline in Section 4, and therefore does not have a projection model.

Model Statistics and Variable Significance

Sector model	MSE	Adj R^2	CPI	UNEMP	GDP	DFF
Finance	0.004	0.171	***	-	-	x
Basic Industrials	0.003	0.242	***	x	x	-
Capital Goods	0.003	0.170	***	-	x	-
Technology	0.003	0.133	***	-	-	-
Energy	0.036	0.007		x	∘	-
HealthCare	0.004	0.081	***	-	-	x
Public Utilites	0.001	0.121	***	-	x	-
Consumer Services	0.002	0.170	***	-	-	-
Miscellaneous	0.002	0.114	***	-	-	x
Consumer Non Durables	0.001	0.178	***	-	x	-
Consumer Durables	-	-	-	-	-	-
Transportation	0.024	0.029	*	-	-	-

Table 13: This table presents various error statistics and variable significance for each of the predictive models. The Mean Squared Error (MSE) is the average error of the model using the 10 fold cross validation approach. The Adjusted R^2 value is from the general linear model. The statistical significance values are also from the general linear model. See Table 14 for the significance code values. The "x" represents variables that are used in the respective model, but do not demonstrate statistical significance.

Significance Codes

P-value range	Symbol
0 - 0.001	***
0.001 - 0.01	**
0.01 - 0.05	*
0.05 - 0.1	∘
> 0.1	x

Table 14: Statistical significance codes used in Table 13. Note that an "x" means that variable is not significant, but still used in the model because of the variable selection results.

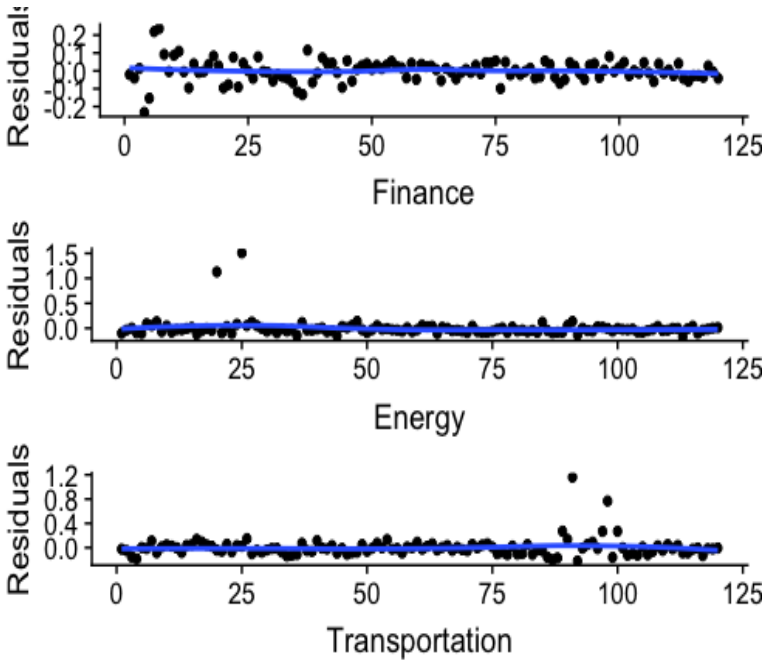


Figure 1: Plots of the residuals of three models with a trend line. The Finance plot is representative of the rest of the models that are not included in the figure. The Energy and Transportation models also do not show an obvious correlation, but do have a few high leverage points.

D Normality Tests

The following Tables, 2-5, show the pairwise test that we used to check for normality for each of the sectors. In the Monte Carlo simulation, we assume normal distribution of the historical monthly returns of each sector. For each sector, the figure on the left shows a histogram of the monthly returns. The figure on the right is the Q-Q plot of the monthly returns of the respective sector. For all cases, the monthly returns appear to be normally distributed in the histograms. Likewise, the Q-Q plots do not suggest that assuming normal distribution would be inappropriate. The figures are broken into four figures due to the size and readability of the graphs.

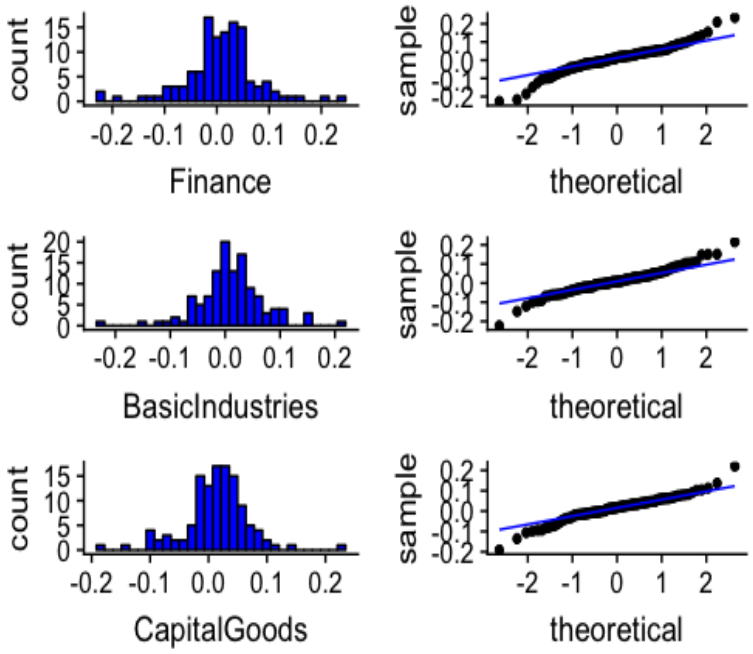


Figure 2: Histogram and Q-Q plots of the historical monthly returns for three sectors.

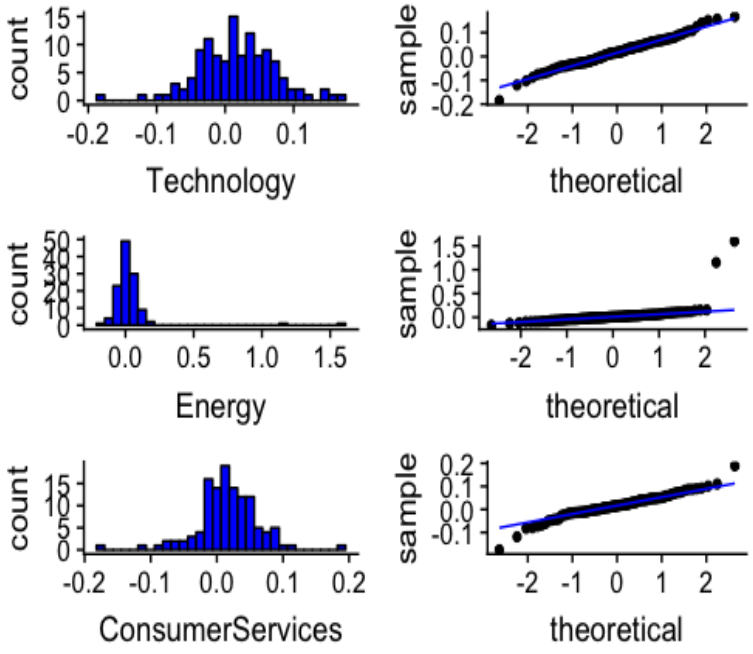


Figure 3: Histogram and Q-Q plots of the historical monthly returns for three sectors.

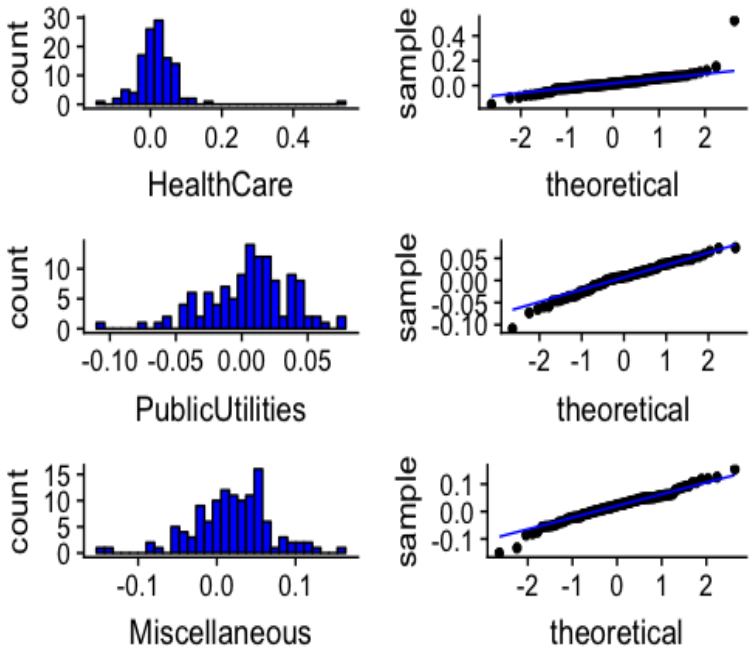


Figure 4: Histogram and Q-Q plots of the historical monthly returns for three sectors.

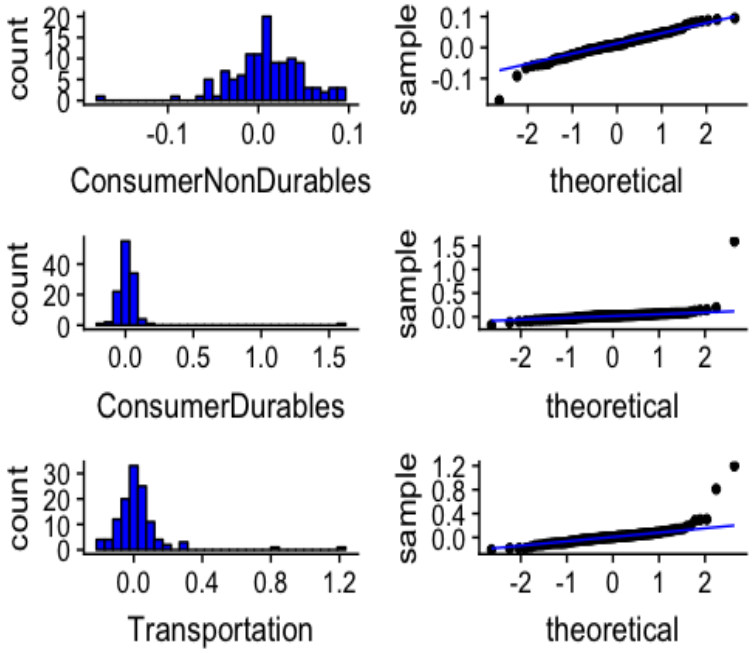


Figure 5: Histogram and Q-Q plots of the historical monthly returns for three sectors.