Names: Steven Petersen, Drew McClelland, Curtis Miller

Principle Component Analysis of S&P 500 and Fortune 500 companies

Many financial securities in the markets move similarly for various reasons. Sometimes these stocks are in the same industry; shocks to a common resource in the industry will affect every member of the industry and thus one can reasonably expect similar stock movements in response to the shock. On the other hand, companies may respond differently to shocks. For example, a sudden drop in the price of oil (as seen recently) may have very adverse effects for oil companies but help airlines, and one would expect their stocks to move in opposite directions.

The examples seem obvious, but these relationships quickly become tangled, especially when sophisticated financial instruments become involved and create unexpected relationships. For example, the Russian and Argentinian economies may not have much in common or any great inherent interdependence, but when securities issued by these two countries are grouped into the "developing economy" category (and mutual funds begin placing them in the same portfolios), their securities start to become correlated. Hedging strategies may also prompt securities to become correlated (either positively or negatively), as signals relevant to one security's value prompt the buying or selling of another (seemingly unrelated) security.

Navigating this complex web of interrelationships both natural and brought about by advanced and complex portfolio strategies becomes extremely difficult.

Our objective is to use data mining techniques to try and identify similar stocks, based exclusively on their price movements. The idea of stocks being "similar" is still preserved, but we would like similarity to be drawn from the data using some dimensionality reduction technique so that one can see, visually, which stock behave similarly, which oppositely, and which orthogonally (in the sense that they don't move in the same or opposite direction).

Approach

We used principal component analysis (PCA) to reduce the dimensionality of closing stock prices taken over a length of time. A matrix was constructed with each stock corresponding to a specific row in the matrix. Each column of the matrix had a different representation based on which technique we were currently employing. The first approach used put the closing price of each stock on that particular day (column 1: day 1, column 2: day 2…). This approach was extremely vulnerable to stock price scaling (in other words, some stocks are priced in units much higher than others).

A log-difference approach was implemented to combat this effect. This was done by taking each column to be the log-difference between one day and the day before (column 1: log(day 1), column 2: log(day 2) - log(day 1), column 3: log(day 3) - log(day 2), etc.). This approach closely approximates the percentage change between days and yielded less erratic results. An approach using the actual percentage change (column 1: 0, column 2: (day 2 - day 1) / day 2, column 3: (day 3 - day 2) / day 3, …) was implemented and yielded similar results.

We implemented a final augmentation in an effort to remove the total trend of the stock market from our data set. This was done by calculating the log difference matrix for the parent index and subtracting it from each row in the stock matrix. This technique had a significant effect on scaling the data points but did not help to change the shape of the dataset.

Data Set

We originally used stocks composing the S&P 500 index in this experiment to provide a diverse set of stocks to test the approach on. However, after some analysis it became clear that these stocks were intentionally chosen to be anti-correlated in order to produce a more stable portfolio, which was preventing our analyses from yielding interesting results. In order to produce more significant results we chose to rerun our experiment using stocks in the current list of Fortune 500 companies. This became our dataset.

Using this list of stocks we performed PCA on the 500 companies over a varying period of time from 25 days to 2500 days. Increasing the number of days used in our analysis appeared to spread our data out even as the significance of each component was reduced. Our original approach was also focused on using date ranges after the stock market crash in 2008 in order to avoid the turbulence caused by this event. However, the graphs produced by our technique were most interesting when we included prices during the 2008 financial crisis.

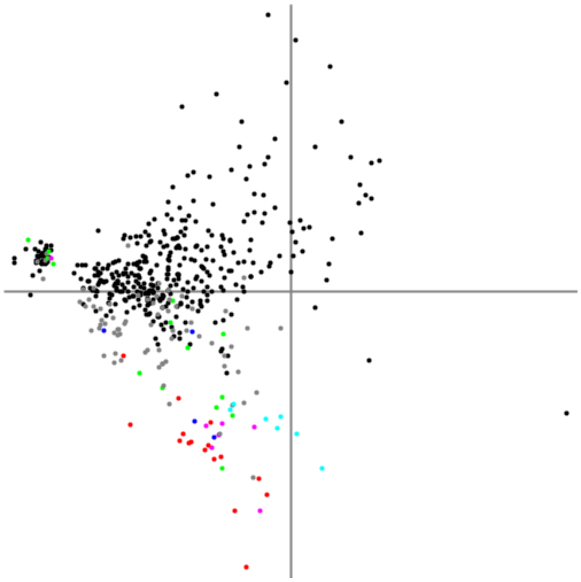
Another interesting result from our experiment was the effects caused by our two components. Surprisingly, our second component seemed to have a much more significant effect on separating industries. The first component, on the other hand, seemed to increase the separation between stocks within a specific industry.

The percentage change approach was used in producing our final results. We used a start date of July 12, 2008 and took the prices over 150 days to provide an analysis over a time period including the stock market crash. In addition, we used a start date of August 12, 2010 and took the prices over 1000 days to use as a comparison to these results.

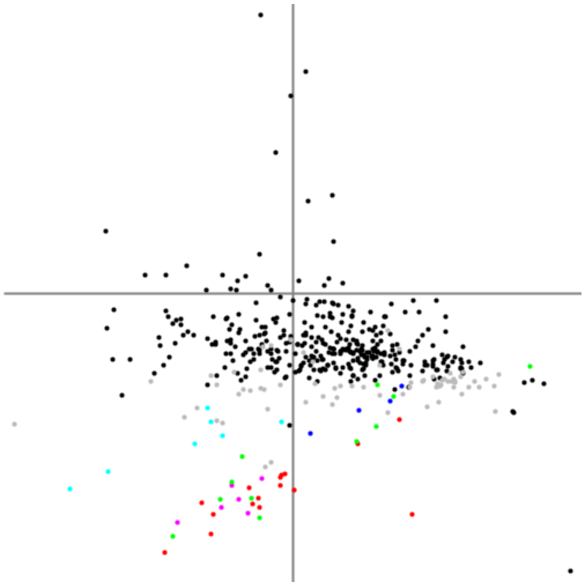
Analysis

We did manage to discover groups of stocks that perform similarly. After running PCA with two components, we graphed all the stocks in two dimensions, each dimension representing each stock's score on a principal component, with the x axis being the first principal component score and the y axis the second principal component score. Unfortunately, our data didn't seem to cluster very well. We came across some clusters in some analyses over certain intervals of time, but they were poor clusters. Because of this issue, we decided to compare several industries. Many industries clustered around some area, but other areas are more spread out.

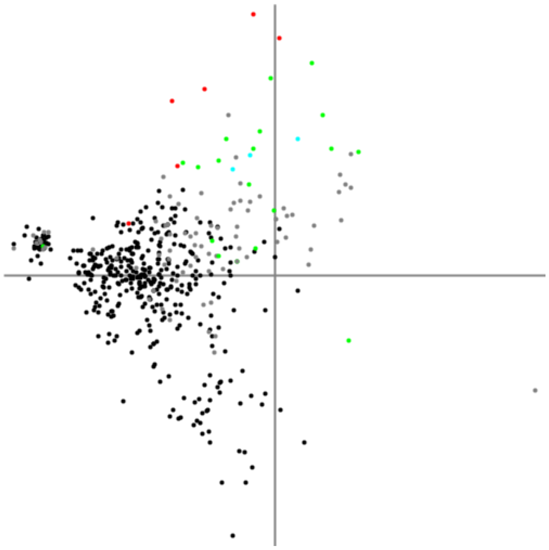
Using stock data taken during the stock market crash (July 12, 2008: 150 days) we were able to find some clustering related to industries. As seen below, Energy (blue), Metals (cyan), Crude Oil Production/Mining (red), Oil and Gas Equipment/Services (magenta), and Petroleum Refining (green) all perform similarly. Some industries performed similarly but were clustered closer to the center, like Chemicals, Construction and Farm Machinery, Construction/Engineering, Food Production, Industrial Machinery, Pipelines, and Gas/Electric Utilities (all gray).



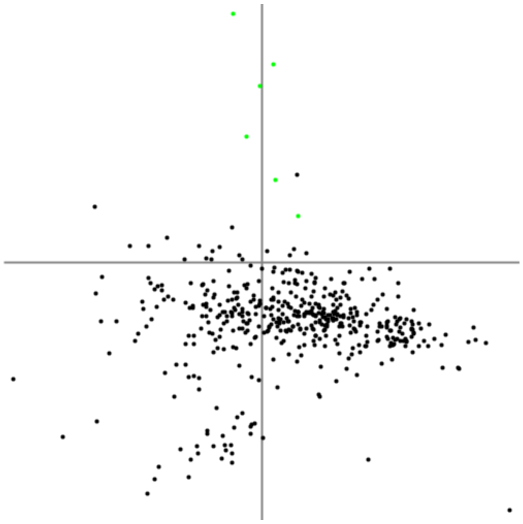
Over a different period of time, (August 12, 2010: 1000 days) these stocks performed similarly (see below).



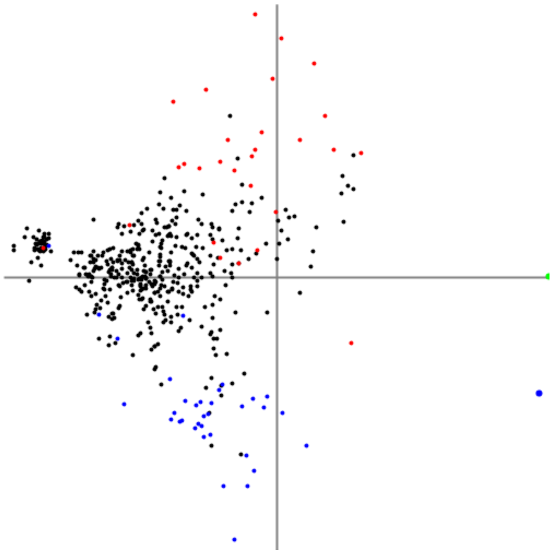
As seen below, Airlines (red), Commercial Banks (green), and Homebuilders (cyan) all performed similarly during the stock market crash (July 12, 2008: 150 days). Some industries performed similarly but were clustered closer to the center, like Automotive Retailing/Services, Entertainment, General Merchandisers, Home Equipment/Furnishings, Life/Health Insurance, Property/Casualty Insurance, Motor Vehicles and Parts, and Real Estate (all gray).



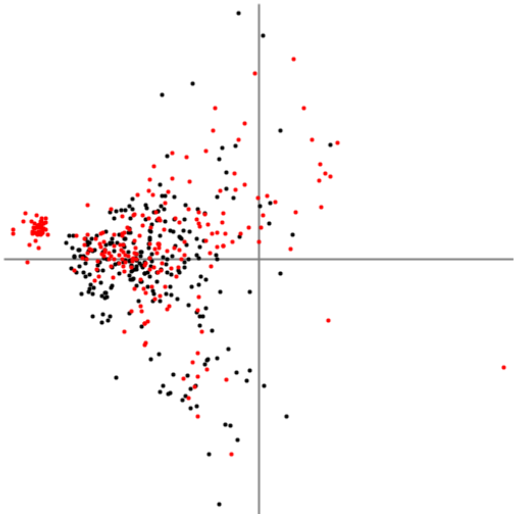
These industries had little correlation over another time period (August 12, 2010: 1000 days); Airlines (green) topped the graph and everything else mostly blended in with the main cluster.



The inverse relationship between Airlines, Homebuilders and Commercial Banks (Red) versus Metals, Energy, Crude-Oil Production, and Oil & Gas Equipment (Blue) is shown in the graph below (July 12, 2008: 150 days).



The cluster on the left has little correlation to a particular industry, and each industry corresponds to companies all over the graph, as seen below. (It has companies from 27 of the 72 industries. Those 27 industries represent about 50% of the companies.) The cluster did contain a relatively significant amount of manufacturing and petroleum related industries. The Industrial Machinery, Motor Vehicles, Metals, and Oil related industries made up 13% of the whole data set (63/491), but made up 29% (11/38) of the cluster.



Summary

From our graphs, we found an upper cluster of similarly performing industries and a lower cluster of similarly performing industries.

In the lower cluster, Energy, Metals, Crude Oil Production/Mining, Oil and Gas Equipment/Services, and Petroleum Refining all perform more similarly than to the other industries. This makes sense, since these companies engage primarily in mining and processing natural resources. Industries like Chemicals, Construction and Farm Machinery, Construction/Engineering, Food Production, Industrial Machinery, Pipelines, and Gas/Electric Utilities all perform similarly to those companies, though not as strongly. These are mostly related to mining and processing natural resources.

In the upper cluster, Airlines, Commercial Banks, and Homebuilders all perform more similarly relative to other industries, with a weaker similarity to Automotive Retailing/Services, Entertainment, General Merchandisers, Home Equipment/Furnishings, Life/Health Insurance, Property/Casualty Insurance, Motor Vehicles and Parts, and Real Estate. On the surface, these industries do not appear to have much in common, yet their stocks appear to move similarly.

Thus, we have identified with the help of principal component analysis two groups of stocks that do not appear to move similarly. After some thought this clustering does make sense. Consider, for example, the recent collapse in oil prices. For oil companies, the oil shock is a major threat to the companies’ profitability and stability, while for airlines (in another cluster) the shock greatly reduced one of these companies major expenses, which should help these stocks’ return.

In conclusion, PCA does seem to facilitate investigating stock similarity. Our analysis seemed to produce placement of stocks that seem intuitive. We are particularly encouraged by the fact that stocks in the same industry seem to appear nearby. Stocks that are distant from one another also seem intuitively "dissimilar". Thus we believe that principal component analysis of stock price movements is useful, and one could consider using it for stock analysis in order to see through complex interrelationships.

Contributions

Group Member	Contributions
Steven Petersen	Collected our data using a Chrome Extension that downloads all the relevant stocks in bulk. (Downloading 500 spreadsheets individually would be really painful, each file needed to be renamed.) Wrote the Chrome Extension that produced all our graphs. It takes processed data from our Python program, then graphs it on a canvas. It has some tools to help find similarly performing stocks and highlight stocks or industries a given color. (Mouse over a part of the graph to find the nearest point and the stock/industry it represents.)
Drew McClelland	Wrote a Python program to convert raw text data into a Numpy matrix for each stock from a start date over a range of days. Wrote methods that performed PCA over a Numpy matrix to reduce dimensionality of the data using one or two (or more) components. Wrote a method to output PCA results to a text file for easy analysis. Analyzed data and refined approaches.
Curtis Miller	Produced and refined approaches and modelling used in data analysis. Verified and determined significance and interpretation of program results. Provided analysis of PCA results. Assisted in technical writing and editing of reports.