Modeling Stock Signals Using Principle Component Analysis Association Rules Learning

Ian Schweickart
Department of Mathematics
Harvey Mudd College
Claremont CA, 91711
ischweickart@math.hmc.edu

Abstract

The goal of this paper is to create a method for constructing more accurate eigen-portfolios by incorporating association rule learning with a Principal Component Analysis algorithm. Association rules are learned from the components of the Dow Jones Industrial Average index. These associations allow us to reduce the number of features supplied to the Principal Component Analysis by excluding stocks with less significant associations. Principal Component Analysis is used to reduce the number of features necessary to model the given data. Efficacy of the feature-reduced model is found by plotting our model against the fund that tracks the Dow Jones Index.

1 Introduction

Predicting the stock market has been the relentless quest of countless traders passed, and surely to come. To do so requires either acting fast and making trades before others react to a position, or having superior information to other traders. Such information is hard to come by in this day and age, but with the virtually limitless amounts of data to be gathered from the finance industry, and the continual improvement of computing power, a novel algorithm might just be able to beat the market.

An algorithm like that would be a fantastic discovery, and make for a killer final project, but this is not that project. Instead, this paper seeks to explore one of the many machine learning algorithms employed in today's world of finance. Principle Component Analysis (PCA) is extremely useful in modeling the movement of a stock (or a portfolio of stocks) due to its ability to capture the covariance between difference stocks in a intuitive and manipulable way.

Since PCA is already a common industry practice, my goal is to strengthen its predictive power by incorporating an association rule learning algorithm into the process. Association rule learning algorithms are able to find associations betweeen different elements across large and sparse datasets, and have to ability to be tailored to fit many needs. In the case of this paper, I will need an association rule learning algorithm to look for similar movements in stock behavior.

1.1 Focus

The focus of this project is to explore the applications PCA in a novel way. By applying an association rule learning algorithm to similar stocks, the hope is that I can feed PCA a tailored "index" that allows for a more accurate prediction of the target signal. I hope to show that association analyses can have a significant effect on the applications of PCA, and that PCA can be used to produce an interpreatable predictive model.

29th Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain.

1.2 Background

The motivation to explore this type of algorithm and specifically, its application to stock market data, stemmed from my original thesis topic. I was planning on finding associations between global events and significant fluctuations in a stock's price, then use what I found to create a predictive model. The issue with this approach is that I had no way of knowing whether or not a stock's movement was actually affected by a global event, or if the market was just having a particularly good or bad day. Thankfully I came across a paper by (Avellandeda and Lee, 2008) [1], on the topic of statistical arbitrage in the U.S. equities market. This paper gave me the foundation for finding confident associations between stock price movements and a given event unrelated to the market. This foundation was based on a method of taking the market-driven variance out of a stock signal by utilizing PCA.

The second aspect of this project follows from a technique I became familiar with over the summer. Association rule mining is a technique for finding significant associations between two or more "items" in a "transaction", where items are stock returns by day and transactions refer to each day. An algorithm of this kind will produce "rules" consisting of support, confidence, and lift statistics. For a select stock X, and the remaining stocks in the data set I_i,

$$\text{Support}(X \Rightarrow I_i) = P(X \wedge I_i) = P(X)P(I_i|X) = P(I_i)P(X|I_i)$$
$$\text{Confidence}(X \Rightarrow I_i) = P(I_i|P_X)$$
$$\text{Lift}(X \Rightarrow I_i) = \frac{P(I_i|X)}{P(I_i)} = \frac{P(X \wedge I_i)}{P(X)P(I_i)}.$$

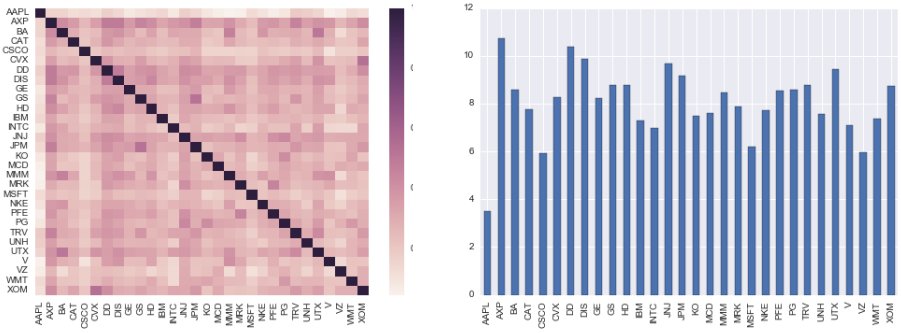
Support measures how often a rule is true in any transaction; a support of 0.1 implies that the rule occurs 10% of the transactions in the data set. Confidence measures how often I_i is present when X is present, so a confidence value of 0.5 implies that when X is present in a transaction, 50% of the time I_i is also present. Finally, lift measures the correlation between X and I_i. If the lift value is 1, then the two are considered independent. A lift value greater than 1 implies positive correlation.

1.3 Data

The data used for this project came from a DVD-ROM purchased from [3] containing 42 gigabytes of stock data for 1239 stock symbols over a 13-year period. The data set contains all component stocks for the Dow Jones Industrial Average, the Dow Jones Transportation Average, the Dow Jones Utility Average, the Nasdaq 100 Index, and the S&P 500 Index, as well as most active Exchange Traded Funds. This data was recorded on a minute by minute basis, between market open (9:30am) and market close (4:00pm), from December 30th, 2002 up until the day it was purchased, which happened to be October 7th, 2016.

I decided to use data from the Dow Jones Industrial Average (DJI) since it seemed to be the least dirty. By least dirty, I mean each stock had no less than 50% of the total minutes that could possibly be recorded per day. I sampled the DJI for the month of December in 2013 because the market during this time seemed to have a strong underlying, linearly increasing trend. An underlying trend is helpful in this this case because this has a higher likelihood to result in a principal component with a larger percentage of explained variance, after applying PCA, more of that in section 3.

To get a sufficient number of data points, I used minute by minute data from the aforementioned time period. The DJI consists of 30 different stock symbols which have been chosen be industry professionals to diversify the index as much as possible. While this could bring rise to problems that are explained later on in this paper, the techniques I have employed allow for my analysis regardless of portfolio diversification.



(a) Heat Map of DJI Components Correlations (b) Sums of Pairwise Stock Correlations

Figure 1: Initial Correlation Statistics

2 Preliminary Investigation

To get the stocks into a form that was usefully interprettable I used the formulas for standardized returns referenced in [1]

$$Y_{ik} = \frac{R_{ik} - \bar{R}_i}{\bar{\sigma}_i},$$

where R_{ik} is the returns of stock i on day k. In this form, the data can easily be made into an empirical correlation matrix

$$\rho_{ij} = \frac{1}{M-1} \sum_{k=1}^{M} Y_{ik} Y_{jk},$$

which is symmetric and non-negative definite. Here M is the number of days over which the returns are being gathered and i, j refer to different stocks. We can see in figure 1a that when $i = j$, $\rho_{ij} = 1$ as expected. This heat map explains how each of the stocks' standardized returns is correlated. The bar chart shown in 1b shows us which stocks are the most associated with the rest of the stocks that make up the DJI. This chart is composed of column-wise sums of the Pearson's correlation coefficients found in figure 1a.

3 Principal Component Analysis

The main source of inspiration for exploring this technique came from [1], which gave an intuitive example of a scenario that benefited from using PCA. Papers similar to [1], as well as in other industry applications, PCA is primarily used as an indicator of risk in a portfolio. The idea is that a portfolio should be maximally diversified in terms of risk, and if the principal component of the portfolio explained too much of the variance, the portfolio's risk is not diversified enough. While this this is a useful application, my interest in PCA is for a different reason. Since PCA has the ability to model an index, or stock, while accounting for less than 100% of the variance of said index/stock, it has the potential to allow me to draw conclusions about movements in price that are not explained by the PCA (which implies the market in this case).

Due to the nature of stock market data, the many dimensions necessary for concrete analysis can be overwhelming, and detract from the interpretability of results. To reduce the dimensionality of this data I applied Principal Component Analysis. While this is helpful in creating more informative, less unwieldy components out of the data set, it also has the benefit of allowing me to control the amount of variance I use in reconstructing the original data. In practice, a model to explain the movements of the stock market is called an eigenportfolio; instead of a portfolio of stocks, it is a portfolio consisting of the eigenvectors produced by PCA.

Running PCA on a portfolio containing every component of the DJI produced the explained variances in figure 2. We can see that none of the components explains a drastically disproportionate percentage of the variance. The principal component, the one that is supposed to carry a significant portion of

Figure 2: Variance Explained by New Components

the variance, does not even explain 18% of the variance of this portfolio. As previously mentioned, this is probably due to the fact that the components in this index are chosen specifically to increase diversity and minimize exposure to risk. Thus finding a disproportionate principal component here could indicate an error in calculation.

4 Association Rule Mining

Employing an association rule mining algorithm to this project was primarily to gain a deeper understanding of the connections between the data. It also allowed for a more descriptive interpretation of the data correlations than just simply doing a correlation matrix. This technique was useful when applied in concert with PCA since it allowed me to tailor the data such that the less similar stocks were pruned from the data set, thus strengthening the predictive power of the principal component.

In order to find associations between the DJI components I ran the algorithm on the components in the DJI along with DIA, an Exchange Traded Fund that tracks the DJI as a whole. The resulting plot is shown in figure 3. We can see that each of these associations rules was present in over a third of the trading days in December and there was a minimum of 75% confidence in the rules, which is extremely promising.

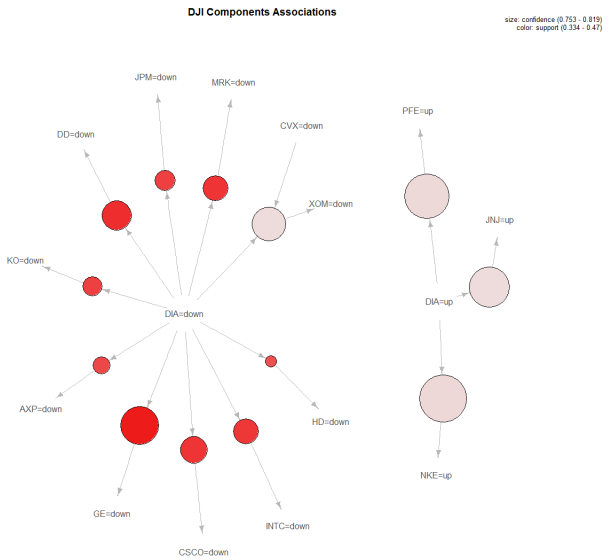


Figure 3: Association Rules Plot of DJI Components

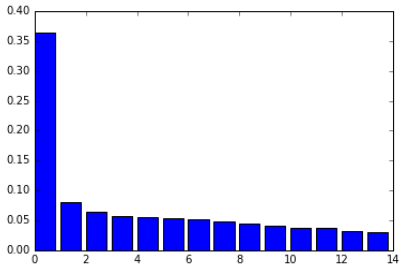


Figure 4: Variance Explained by New Components

With these 14 stocks that had significantly higher associations than the rest I created the previously mentioned "tailor" portfolio. When performing PCA on this new portfolio, there was a marked improvement in variance explained by the principal component, as seen in figure 4.

5 Results

Using the new components produced by running PCA on the new set of stocks, I was able to create a predictive model (orange) of the standardized returns of DIA (blue) as seen in figure 5.

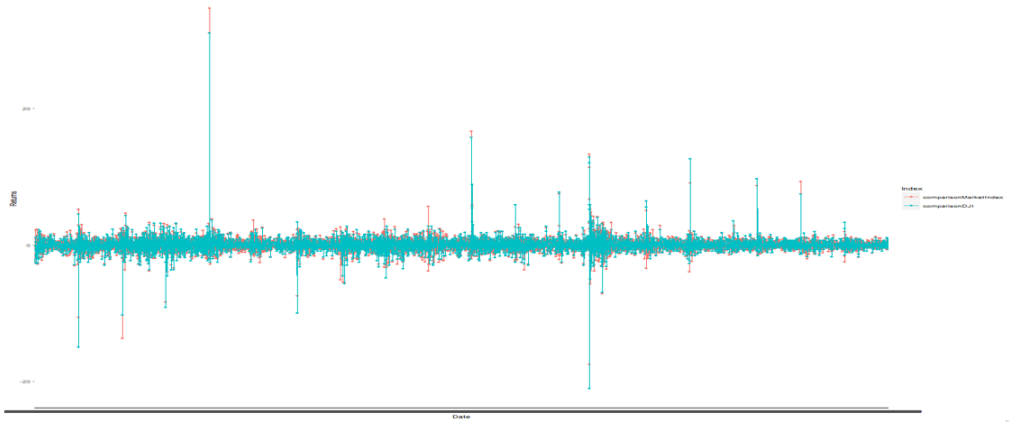


Figure 5: PCA Model of DJI Standardized Returns

6 Discussion

Figure 5 shows that a relatively accurate model can be created using the methods detailed in this paper. What is more promising is that this model appears to be more accurate than the model created by the original portfolio. After differencing the model and DIA, for both the new and old portfolio, the total squared error is improved by over 50%. This combination of techniques follows in the pattern of good machine learning algorithms being made better when combined with other good algorithms, that strengthen some part of the original algorithm.

6.1 Future Work

The first issue that should be addressed in future work on this topic needs to be the quality of the data. Not only does the applicability of the results depend on accurate data, but also the ability to get results. Missing minutes in this data set were the main hindrance to accurate results since these minutes had to be filled with the previous minutes, a reliable solution, but one that introduced its own variance to the problem. Clean data would not only improve the results of this project, but it

would also allow for more flexibility as my choice of data was constrained to the stocks that passed the already low bar of cleanliness.

Future work relating to this project could look into sampling the data at different time scales. This project primarily dealt with minute to minute data, but the issue of dirty data could be potentially avoided by examining larger time scales. This approach could lead to totally different findings as stock prices on the minute scale are much more volatile than they are at larger time scales.

Given clean data and an accurate model, future work relating to this project is wide open. As I plan to do for my thesis, using this method I hope to be able to take the difference between a stock's signal and the signal PCA has created to model that stock. This method theoretically produces a stationary signal that models the variance of the stock with respect to itself. It follows that countless hypotheses can be tested for why that stock, or any stock, varies the way it does, with no external sources of variance. The methods in this paper represent the framework for endless hypothesis testing for research purposes, trading strategies, or whatever else might intrigue someone with interest in the stock market.

References

Note: All of my code can be viewed at `https://github.com/ischweickart/MATH189r-Final-Project`

[1] Avellandea, M, & Lee, J.H. (2008) Statistical Arbitrage in the U.S. Equities Market. *Quantitative Finance*, 10(7), 761–782. Cambridge, MA: MIT Press.

[2] Conway, D, & White, J.M. (2012) Machine Learning for Hackers. 218–227 O'Reilly Media.

[3] Historical Intraday Stock Data. (n.d.). Retrieved October 7, 2016, from http://pitrading.com/historical-stock-data.html