MATH 6380J Mini Project 1

The application of PCA on financial data

YE Yushi (12207896)

2017.03.13

MATH 6380J Mini Project 1

The application of PCA on financial data

Problem Description

In this report, we want to apply principle component analysis (PCA) on the financial data provided by Professor Yao in his website, which contains close price for 452 stocks from SNP'500 in 1252 trading days.

Our objective is to find regularities hidden in the data, and based on that, forms some strategies, which, hopefully, can gain extra profit compared to market average.

To do so, firstly, the whole dataset will be divided into two parts: the first 1000 days will be used as training set, where we do many kinds of analysis and try to form trading strategies based on that; the last 252 days will be used as testing set, where we test our strategies' performance.

Why PCA?

There are two main reasons why we apply PCA on this financial data:

1. Financial data are high-dimensional. By using PCA, one can summarize the original data into a much lower-dimensional one (i.e. the first few principle components), which can also explain most variation of the original data. The variation or volatility of stocks is what people care very much, for purpose of either gaining extra return or managing portfolio risk.
2. For this dataset, besides closing price, we also have information about stocks' industry classification. The 452 stocks are divided into 10 industries: "Consumer Discretionary", "Consumer Staples", "Energy", "Financials", "Health Care", "Industrials", "Information Technology", "Materials", "Telecommunications Services" and "Utilities". We are curious about how different each industry section behaves. Obviously it is not smart to compare stocks from two different industries one by one. Instead, we can use PCA to summary the variation information of each industry by a few principle components and use them as representatives for industry comparison.

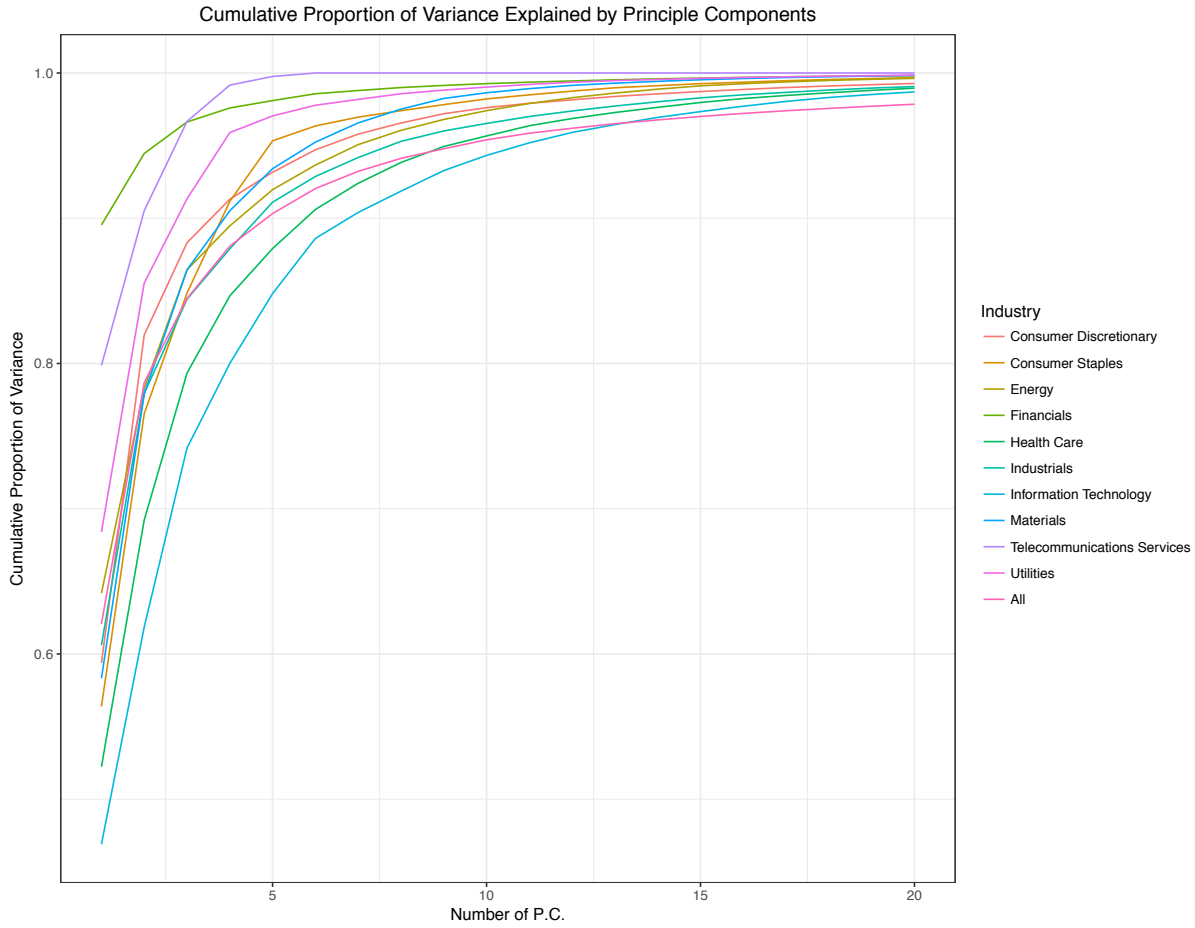
Exploratory Analysis

There are 10 industries in our dataset. Each one contains dozens of stocks. Firstly, we want to extract a few principle components for each industry, which, hopefully, can explain most variance for this industry. Then we can focus on such principle components' behavior instead of all individual stocks. To do so, we apply PCA on all industries and calculate the importance matrix. As an example, the following plot shows the importance matrix for industry "Consumer Discretionary".

	PC1	PC2	PC3	PC4	PC5	PC6
Standard deviation	90.330	55.716	29.551	20.294	15.978	14.558
Proportion of Variance	0.594	0.226	0.064	0.030	0.019	0.015
Cumulative Proportion	0.594	0.820	0.883	0.913	0.932	0.947

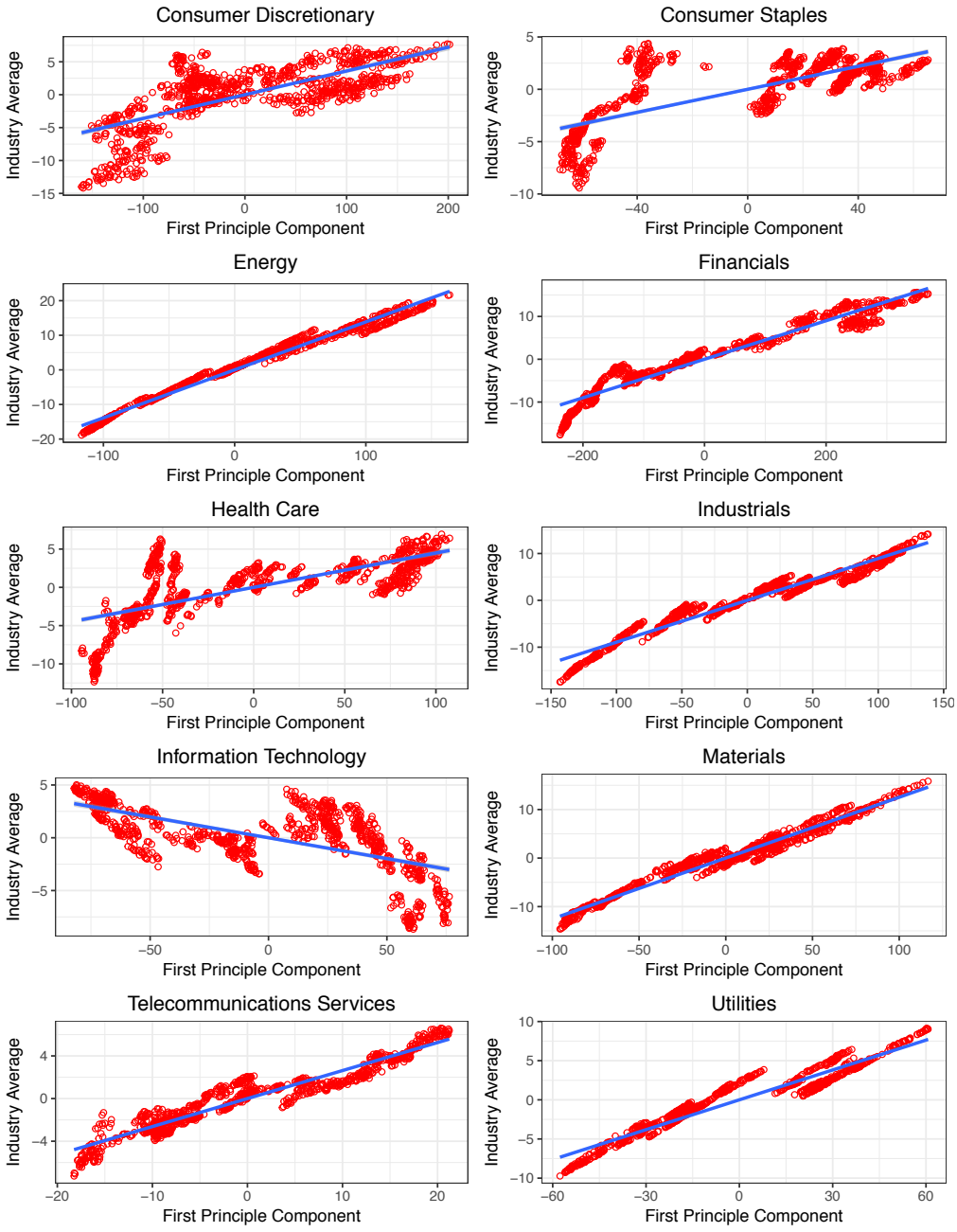
In this table, the first row represents the standard deviations of the principal components (i.e., the square roots of the eigenvalues of the covariance matrix), Since principal components are mutually **orthogonal** variables, the summation of their variance will be exactly the variance of the original data. The second row shows the proportion of variance explained by each principle component; and the third row shows the cumulative summation of such proportion.

Then, we draw a plot which summarizes such information for all 10 industries and the whole SNP'500 training set (labelled as "All"). From this plot, we found that different industry sections do behave different. For example, for industry "Financial", the first principle component can explain nearly 90% of the total variance; but for industry "Information Technology", if we want to explain 90% of the total variance, we must consider the first 7 or 8 principle components. This sounds to be reasonable: on one hand, we know that financial companies' main business are quite similar, and their stocks' performance will be affected heavily by economical events such as interest rate adjustment; on the other hand, companies belong to information technology industry may develop totally different business. For example, Apple's main product is iPhone while Google's revenue relies on its search engine very much.



It's a common sense that market average affects all stocks in the market very much. Then we may raise a question that is the first principle component, which explain the most variance in the original dataset, has a strong relation with market average (or industry average)? In the following, we draw scatter plots between the first principle component and industry average for all the 10 industries in our dataset, along with a regression line fitted by these data.

From these plots, we found that for most industries, the first component and industry average has a strong correlation. But for some industries whose constituent stocks are not connected with each other very closely, this kind of relationship may not be so obvious.



Trading Strategy

Based on the previous analysis, here we develop a simple strategy. The procedures are listed below:

① Estimate the rotation matrix using the training set;

② Apply the rotation matrix on the testing set to get the first principle component.

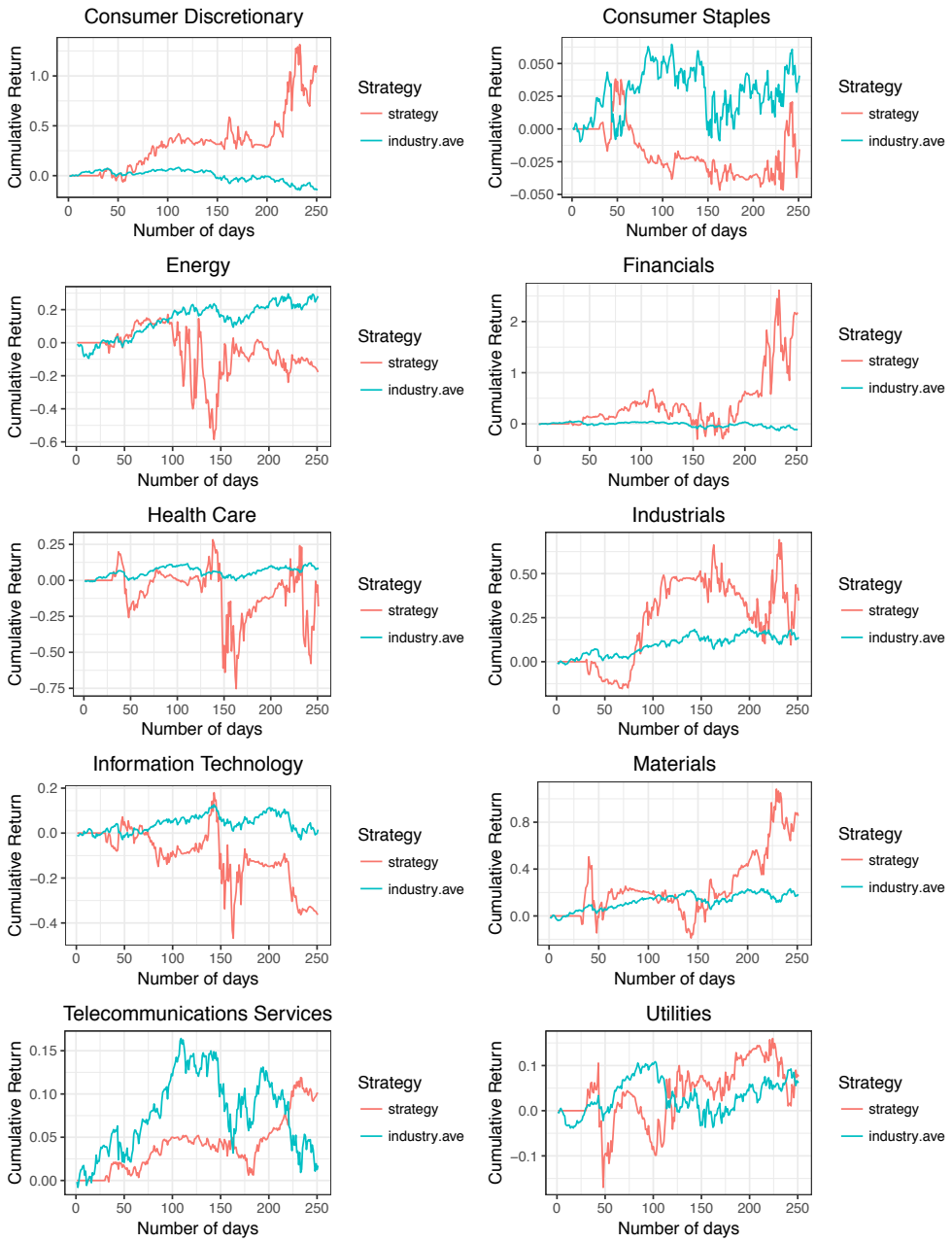
③ On each day, using the data from the previous 30 days and do a univariate linear regression between the stock price against the first component for each stock.

④ Collect the residuals of the above linear regression and calculate z-scores based on that. The higher positive Z-score implies stock price runs more "aggressively" than the first principle component; the lower negative Z-score implies stock price runs more "poorly" than the first principle. Based on the mean-reverting principle, we set up a portfolio with weights proportional to the negative of Z-scores for each stock.

⑤ After we long / short a stock (at today's closing price), we will only hold it for one day and sell it tomorrow at closing price.

Performance

For 10 industries, the strategy's performance are showed in the following plot.



From the above results, we found that this simple strategy only works on some special industries whose first principle component can explain large proportion of total variance, such as Financials and Materials. This enlighten us that we may only focus on these special industries and build a pair-trading strategy based on the first several principle components. However, since deadline is near, we may put this work outside the report. But from the previous analysis, we do found PCA has significant use on financial datasets.