Selecting Representative Constituents in CSI 300 Index

Yikun Hu

March 23, 2020

1 Introduction

1.1 Background

Source distribution plays a key role in world economic. Investing in a corporate means providing funds to target corporate for their growth at a reasonable price in order to obtain income and capital appreciation from its growth of performance. The investment process mainly consists of conducting investment analysis, making investment decisions, taking investment actions and monitoring and review. Selecting a good quality corporate from the beginning of the investment process may significantly reduce the investment costs, e.g. opportunity cost, hedging cost, exit costs.

1.2 Problem

How to select appropriate constituents for tracking an index. There are thousand corporates in the public traded markets. It is hard for fund managers to analyze every constituent, even using simple methods.

Top-down strategies help us to select corporates from industries to specific corporate, but the view on a specific industry will lead to confirmation bias. Before picking a good quality corporate for deep analysis, a flexible framework for screening and segmenting corporate will help fund managers to save much time for the following analysis. The direction of this project is to provide a objective advise for fund managers while a potential target comes into his mind. But there is not any standard for assessing the level of objective. Analyzing a corporate is as looking an object in the sun light. We have to observe it from different directions to fully understand the object.

1.3 This project aims to 1) cluster corporates by mean-variance characteristics and financial ratios in order to combine Markovitz's efficient market theory (even this theory is not realistic, but his view on selecting corporate by statistic method promotes systematic investment analysis in the real world) and fundamental analysis at the beginning stage of screening corporate, 2) pick up representative constituent from each cluster for mimic the Index with financial preference.

2 Data acquisition

2.1 Data collection

For this project we only use the constituents in CSI 300 Index.

We can obtain all data that we need by **JoinQuant API**. The number of queries is limited to one million per day. We can use following code to download the data that we need. We need two datasets for this project: 1) Market closing price of CSI 300 Index constituents, 2) their financial ratios.

2.1.1Collection of Market closing price

The start date is 1st of Jan 2010, and the end date is 1st of Feb 2020. Here we got closing prices of 300 constituents in 2447 trading days. And then, the data has been downloaded and saved into the **Daily_Trading_Data.csv**.

2.1.2Collection of Financial ratios

According to the JoinQuant API Manual, we use the API function **query** to obtain the financial ratios by 2018 year of all the CSI 300 Index Constituents. After this, the financial ratio data has been downloaded and saved into the **Financial_Ratio_Data.csv**.

2.2 Data Preprocessing

2.2.1Market Data

The source data may contain NaN because some constituents' shares were not traded at specific date when declaring important event, some constituents were listed into the CSI 300 Index in the middle of period from 2010 to 2020. For this project we need to calculate the rate of return of each constituent, and then the mean and standard deviation. For calculation the rate of return of each constituent, we will use the function pct_change() with a dataframe without trading date. When we got the dataframe with rates of return, we can use the function **describe()** to get the results of statistical analysis.

2.2.2Financial Ratio Data

First of all, we need to check whether there are NaN in the dataframe.

There are 7 NaN in **operating_profit_to_profit**. For this project we only take account for the companies whose financial ratios are in the database. But we can put the companies whose financial ratios are not fully saved in the database into a list for further researching (Hint: the financial ratios are calculated on the basis of the data in financial reporting. For example, **the operating profit to profit** can be calculated by dividing **operating profit** to **net profit**. Items, **operating profit** and **net profit**, are revealed in **Income Statement** in Financial Reporting.

At last, there are 285 constituents can be used for future analysis.

2.3 Feature Selection

The daily rates of return reflect market participants' views on constituents. For quantitative analysis we can use the mean-variance to measure the risk of constituents in this project.

But market is not always right. The driver of trading may be influenced by emotion, psychological status, market liquidity and the other factors. There are many evidences proof that the market is not always efficient, and market data cannot reflect the real financial status of corporates.

Dealing with this problem, we also conduct qualitative analysis based on financial ratios. For this project, we use 5 financial ratios.

3 Exploratory Data Analysis

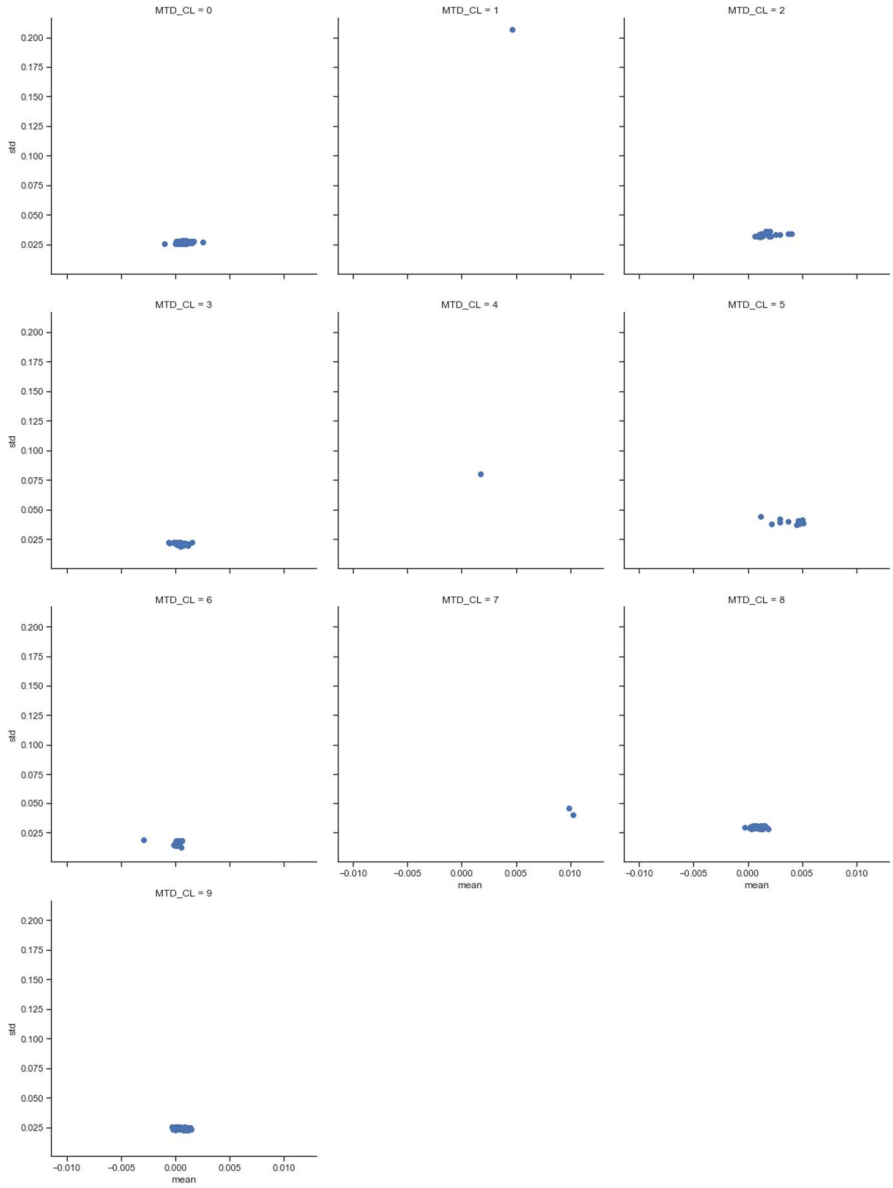
In this project, we use K-mean algorithm to cluster the constituents in CSI 300 Index into 10 groups.

3.1 Clustering constituents by Market Trading Data

The number of constituents in each group is shown below:

No. cluster	Amount
0	70
1	1
2	26
3	49
4	1
5	10
6	22
7	2
8	51
9	68

In the graphic shown below, Cluster 0, 2, 3, 8, 9 contain more constituents. The other clusters contain much less. This means the representatives of Cluster 0, 2, 3, 8, 9 will take higher weights for index tracking in the future.

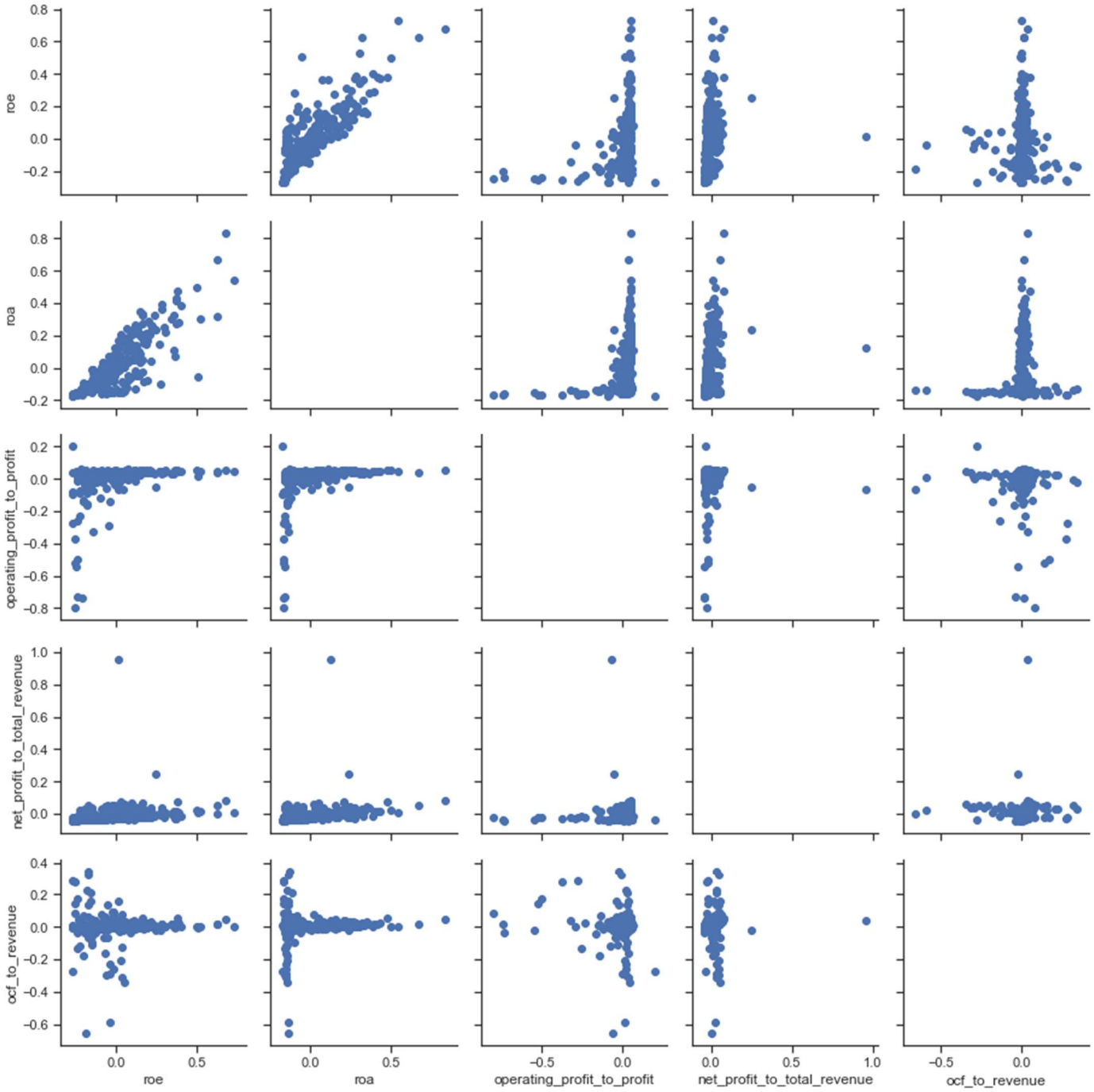


3.2 Clustering constituents by Financial Ratio Data

The number of constituents in each group is shown below:

No. cluster	Amount
0	55
1	3
2	9
3	10
4	8
5	3
6	192
7	1
8	2
9	2

We used 5 financial ratios for clustering. It is impossible to plot a 5-dimension graphic. But we can plot the pairwise relationship graphics to understand the relationship between these financial ratios.



The graphics show that it is possible to use linear regression to mimic **ROA** by **ROE**, and **ROA** (**ROE**) is linear independent on the other financial ratios.

For this project there is only 5 financial ratios have been used. When more financial ratios will be used in the future, plotting pairwise data relationship will help use to quickly find out financial ratios that are linear independent relationship with each other.

3.3 Combining Market Data and Financial Ratio Data

After combining market data and financial ratio data into a dataframe we can find the constituents that allocates in the intersections between Market Clusters and Financial Ratio Clusters. Such

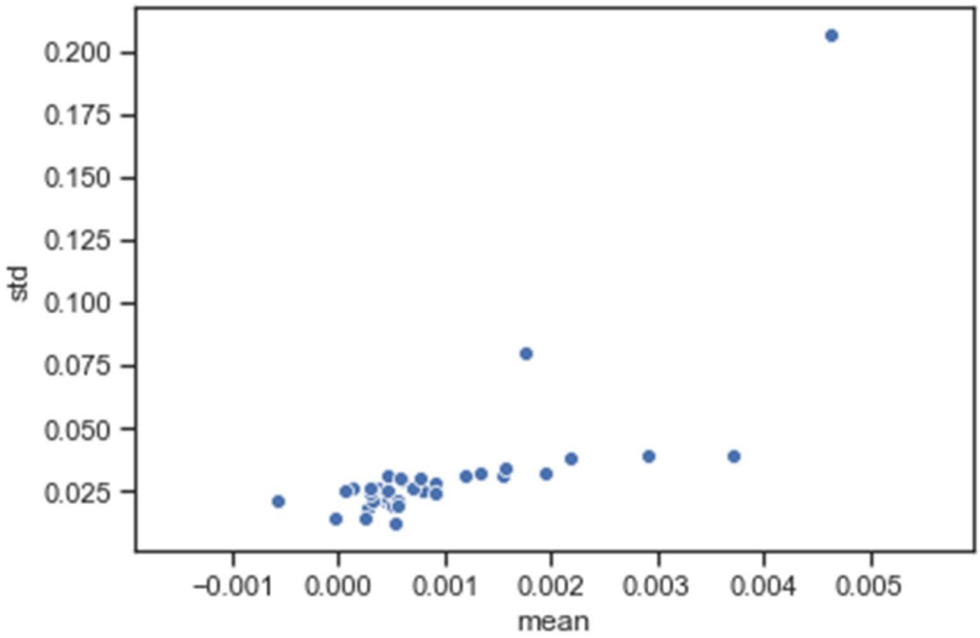
constituents are considered as the representatives of intersection between a market cluster and a financial ratio cluster. In this project these constituents can be picked by the maximum ROE. In the future they can be picked by the other financial ratios that fund managers prefer to.

4 Result

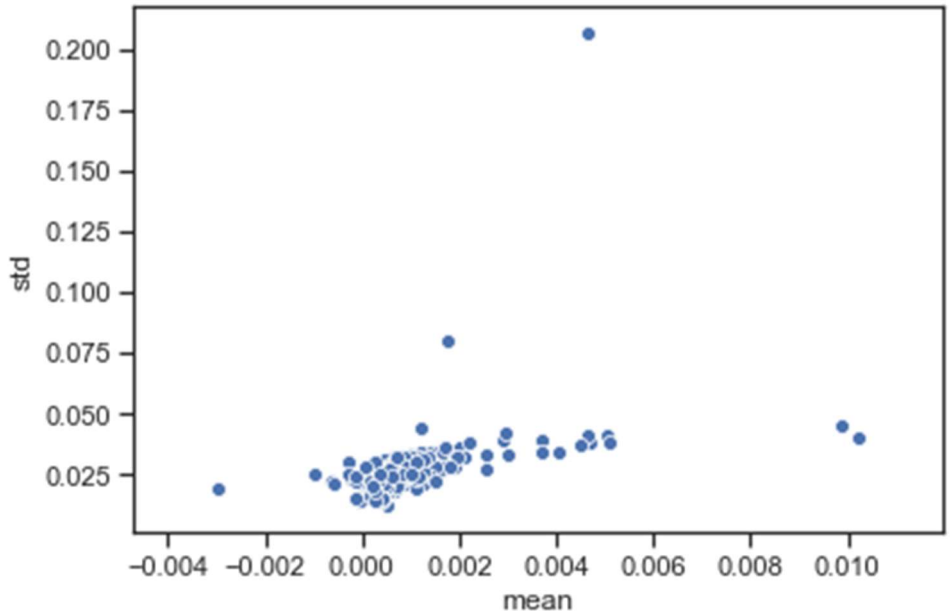
Finally, we got 36 representative constituents as our result.

The stock tickers of there constituents are shown below.

```
['600332.XSHG', '600570.XSHG', '600104.XSHG',  '603019.XSHG', '601229.XSHG', '002555.XSHE', '300142.XSHE', '601555.SHG', '000625.XSHE', '601788.XSHG', '601577.XSHG', '601997.XSHG', '601336.XSHG', '601319.XSHG', '601601.XSHG', '601162.XSHG', '601198.XSHG', '601688.XSHG', '300059.XSHE', '601009.XSHG', '000776.XSHE', '601988.XSHG', '600030.XSHG', '601628.XSHG', '600733.XSHG', '000629.XSHE', '002607.XSHE', '000895.XSHE', '300413.XSHE', '600900.XSHG', '600516.XSHG', '600309.XSHG', '600674.XSHG', '600705.XSHG', '600061.XSHG', '000783.XSHE']
```



Return distribution of representative constituents



Return distribution of constituents in CSI 300 Index

5 Conclusion

The cost of fully tracking an index is very high. In this project, we use K-mean algorithm for clustering constituents in CSI 300 Index by their market data and financial ratios in order to find out the representative in each cluster. The number of constituents for tracking CSI 300 Index has been reduced from 300 (fully) to 36. There 36 representative constituents allow fund managers to tracking CSI 300 Index with lower costs. Additionally, modifying the number of clusters and financial ratios for picking representative constituents provides fund managers with advices while creating a portfolio.

6 Future Direction

Beyond this project we may add more financial ratios and pick up representatives by specified financial ratios (ROE is used in this project). We may use these representatives to build a new index tracking CSI 300 Index. But our index is not just tracking the CSI 300 Index, but also reflects the bias of the representative picker (fund manager). For a conservative fund manager, he may choose the **debt ratio** as a reference to pick the representative. A fund manager focusing on corporate's cash flow from operation will use the **ocf_to_revenue** as a reference. For building a tracking index, constituents' weights should be defined. In this project constituents' weights are not considered. This will be considered in the future.