CSIC 5011 Project 1: Finance Data PCA, Parallel Analysis

Meilan WANG[1] and Di LIU[2] {mwangau, dliuah}@connect.ust.hk

[1]: Department of Civil and Environmental Engineering, HKUST [2]: Department of Mechanical and Aerospace Engineering, HKUST

1. Introduction

The SNP’ 500 is an American stock market index based on the market capitalizations of about 500 large companies having common stock listed on the NYSE, NASDAQ, or the Cboe BZX Exchange. Those stocks are representative of the industries in the United States economy.

We carried out PCA and Parallel Analysis over a time series of stock closed prices in SNP’500. Through the analysis, we find out some interesting interpretation of principal components in the real world.

2. Overview

Dataset

The dataset contains 1258-by-452 matrix with closed prices of 452 stocks in SNP’500 for workdays in 4 years

Methodology

PCA, Parallel Analysis

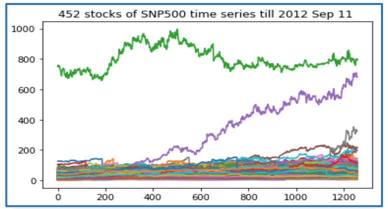
Why?

This is a dataset with hundreds of companies or thousands of days as features. In order to know more about the dataset, or modelling with the dataset, dimension reduction is necessary. With further projection the original company time series on to the principal component constructed space, the sector distribution on the overall SNP’500 stock distribution has explainable meaning.

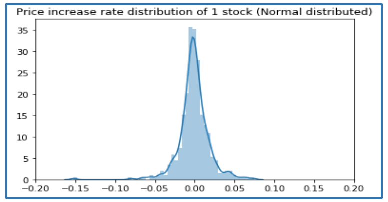
With parallel analysis, we can confirm if the number of eigenvalues we chose is in the reasonable range.

How?

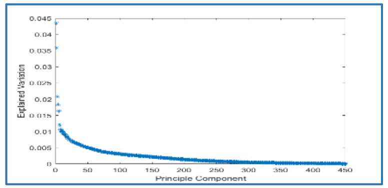
Normalize price → PCA → Project in PC space → Analysis



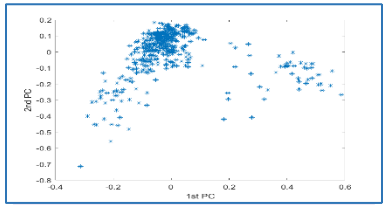
Fg1. 452 Stock Time Series Overview



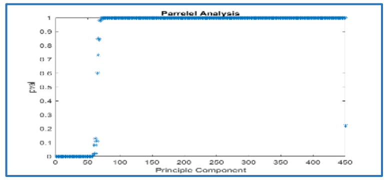
Fg2. Normalize before fit PCA



Fg3. PCA explain variance ratio (except outliers)



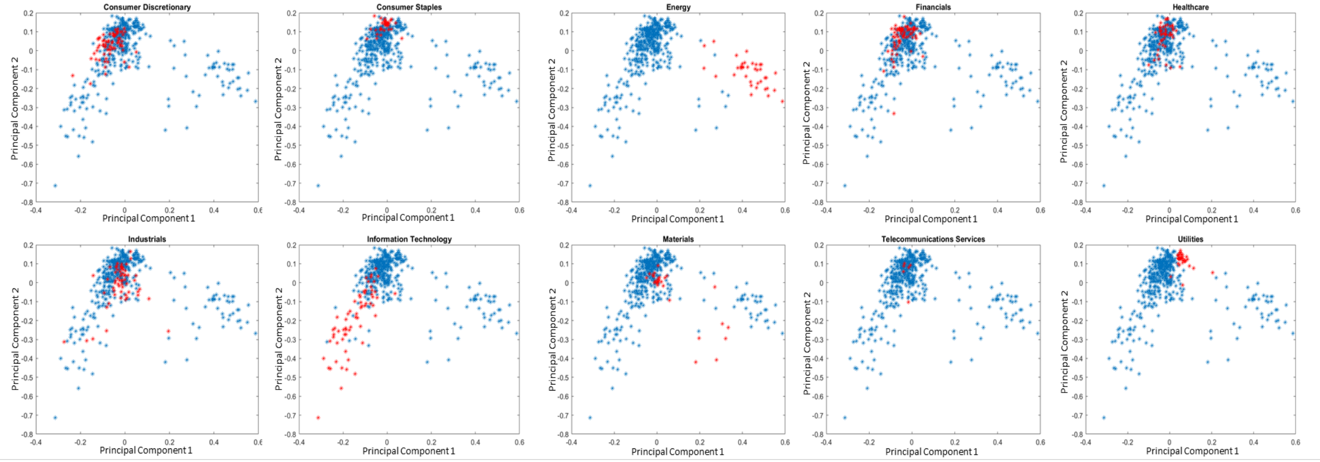
Fg4. Project original data on the PC1 x PC2 space



Fg5. PCA explain variance ratio (except outlier companies)

3. Analysis

By calculating the “log(current day price – previous day price)” from the original dataset, we normalized the original dataset (Fg2). Then we conducted the PCA to find out the components have most significant impact. While first two component contributes over 8.5% of explained variance ratio, we project the original data on the space of PC1 x PC2. By coloring out different industry sectors on the overall SNP’500 companies distribution. The Energy sector and IT sector is more distant than downstream and traditional industry sectors. The principal components are possible to explain with the sector-wise distribution figure.



4.Conclusion and Further Work

1. The Energy and IT sector is more distant than downstream and daily-life related industry sectors.
2. The principal component 1 could be upstream and downstream of business chain, while the principal component 2 could be the distance to daily-life.
3. The upstream and down stream is in different clusters.

If we have exact time information, or more information about the companies, we could conducted PCA using time as feature, or fit PCA within sector companies; With PCA result, we could build a predicting model with less features.

5. Contribution

PCA on subgroup data with Anaconda-Jupyter, poster

- ➢ Meilan WANG

PCA and Parallel Analysis on overall SNP500 with Matlab

- ➢ Di LIU