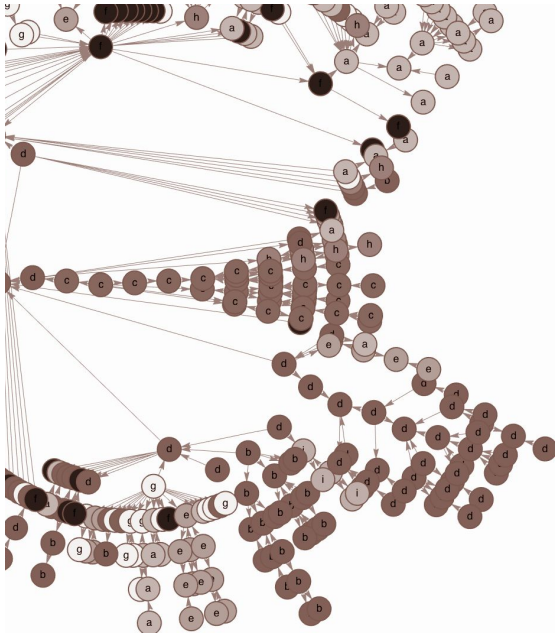


Optimizing Portfolio based on Stock Market Latent Structure

Ganyu Lian
Qiuxuan Lin
Aleena Polansky

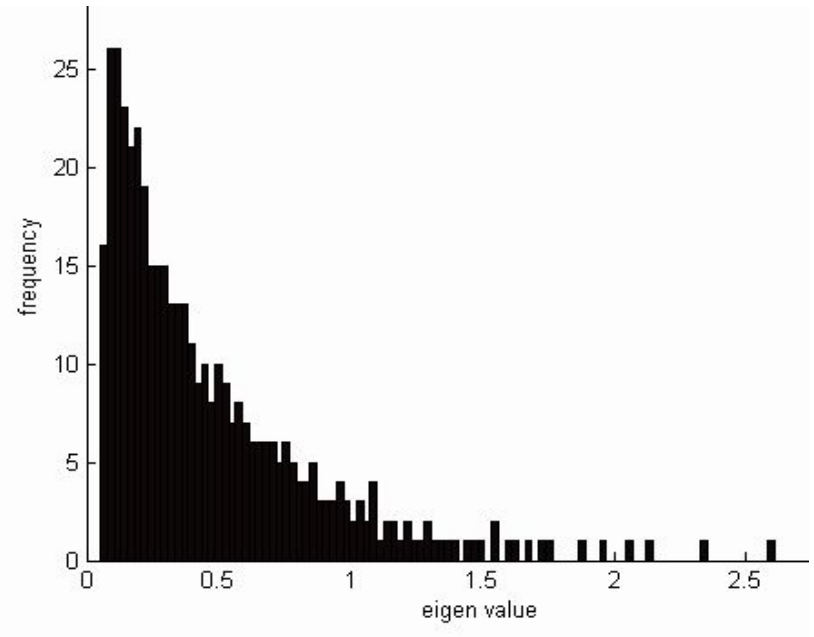
Department of Electrical and Computer Engineering
Boston University

Outline

- Inspiration - Our starting point
- Data Acquisition
- Machine Learning Algorithms
- Application in Portfolio Optimization

Random Matrix Theory Inspiration

1. The largest eigenvalue represents the trend or influence of the whole market.
2. The next few largest eigenvalues show the real interrelation between stocks and can help us with clustering.
3. The very smallest eigenvalues corresponding to eigenvectors that are highly localized, i.e., only a few stocks contribute to them.
4. Most of the eigenvalues(about 95%) in the bulk represent market randomness.



Limits

1. Can only cluster about 10% of all the stocks.

2. Some stocks are clustered into many classes.

Plerou, Vasiliki, et al. "Random matrix approach to cross correlations in financial data."

Dataset Spans 437 stocks and 1538 days

➔ **Preprocessed data: Logreturn**

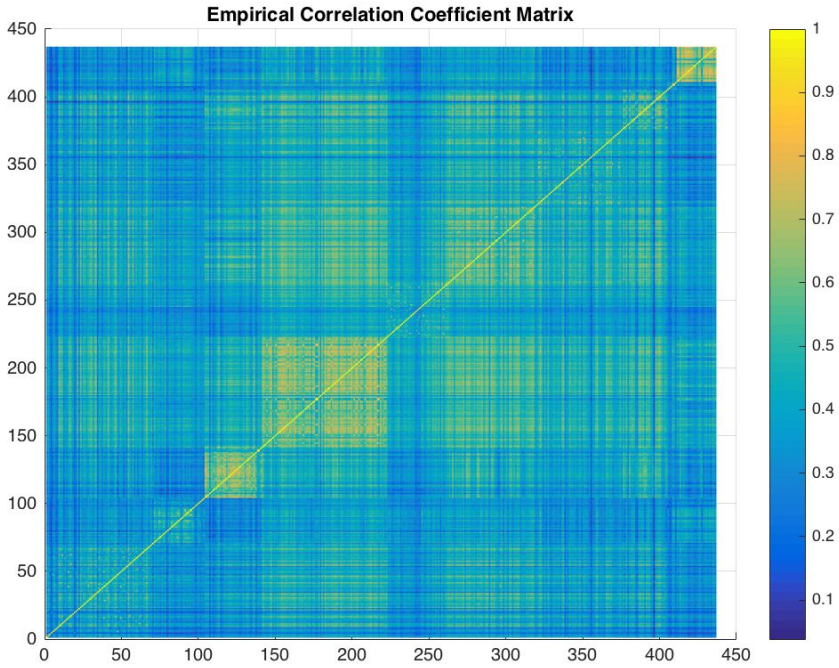
$$R = \log(P_t) - \log(P_{t-1})$$

➔ **Label: Industrial categories**

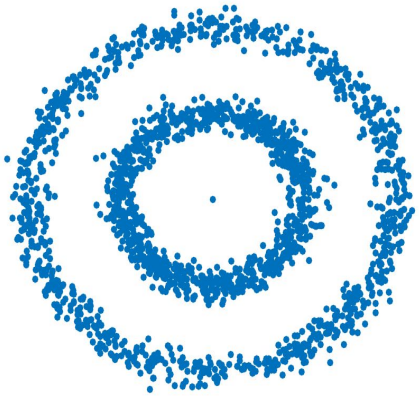
Source: 2010.01 to 2016.03

Yahoo Finance S&P 500http://finance.yahoo.com/

GICS sectors Wikipedia



Machine Learning

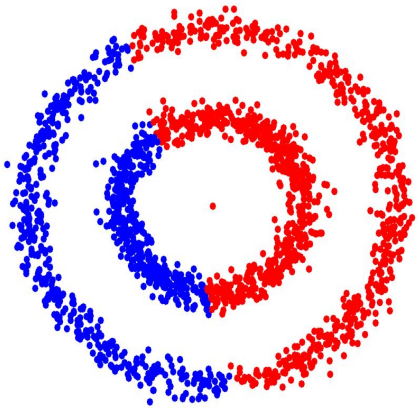


- ➔ **Clustering**
- ➔ Algorithms perform differently on different datasets

  e.g. large or small scale, text, speech, image

- ➔ Evaluation problem

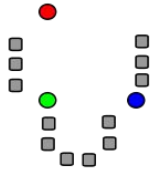
  As much as we'd like to discover some hidden pattern *and to try to ignore any given categorical information*

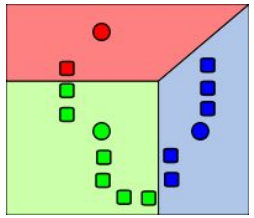


K-means Clustering

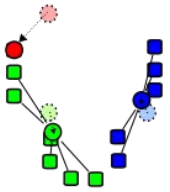
Optimization problem:

$$\arg\min_{\mathbf{S}} \sum_{i=1}^{k} \sum_{\mathbf{x} \in S_i} \|\mathbf{x} - \boldsymbol{\mu}_i\|^2$$





Directly apply K-means on logreturn?



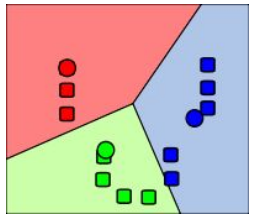


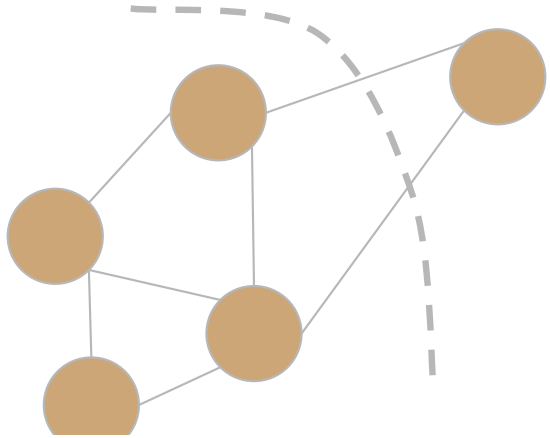
Image credit: Wikipedia K-means Clustering https://en.wikipedia.org/wiki/K-means_clustering

Spectral Clustering

Dimensionality Reduction

A new optimization problem from the graph theory point of view

Minimizes 'cut' : $\sum_{i=1}^{k} \text{cut}(A_i, \overline{A}_i)$



Von Luxburg, Ulrike. "A tutorial on spectral clustering." *Statistics and computing* 17.4 (2007): 395-416.

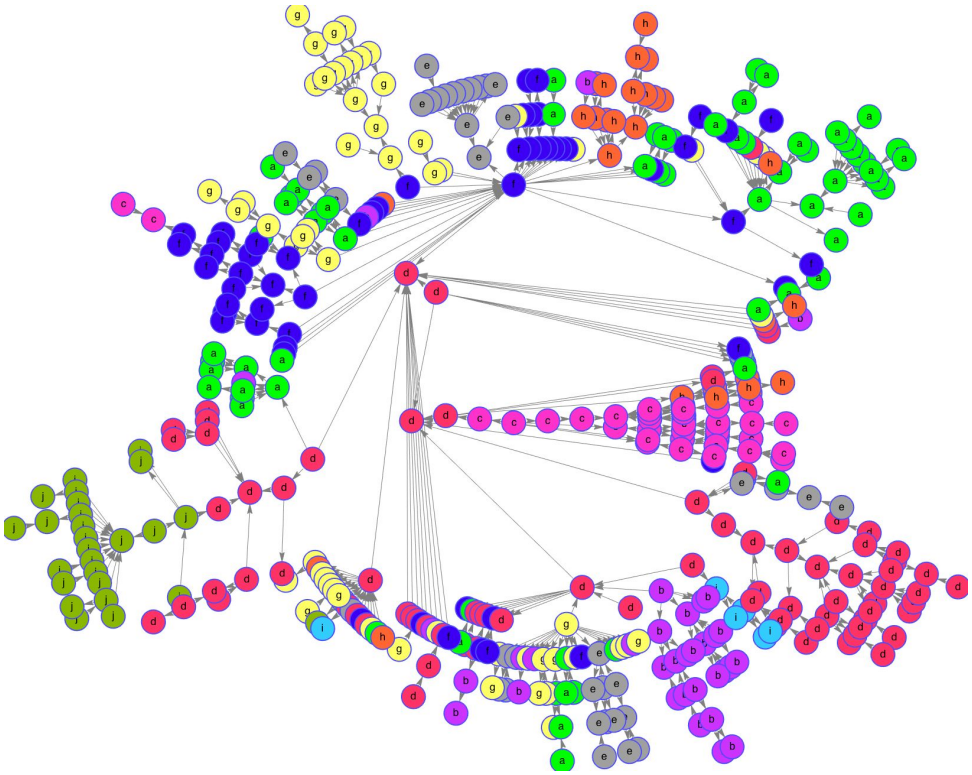
Evaluation at a glance

➔ Maximum Spanning Tree [1]

$$purity = \frac{1}{n}\sum_{y}\max_{c} n_{c,y}$$

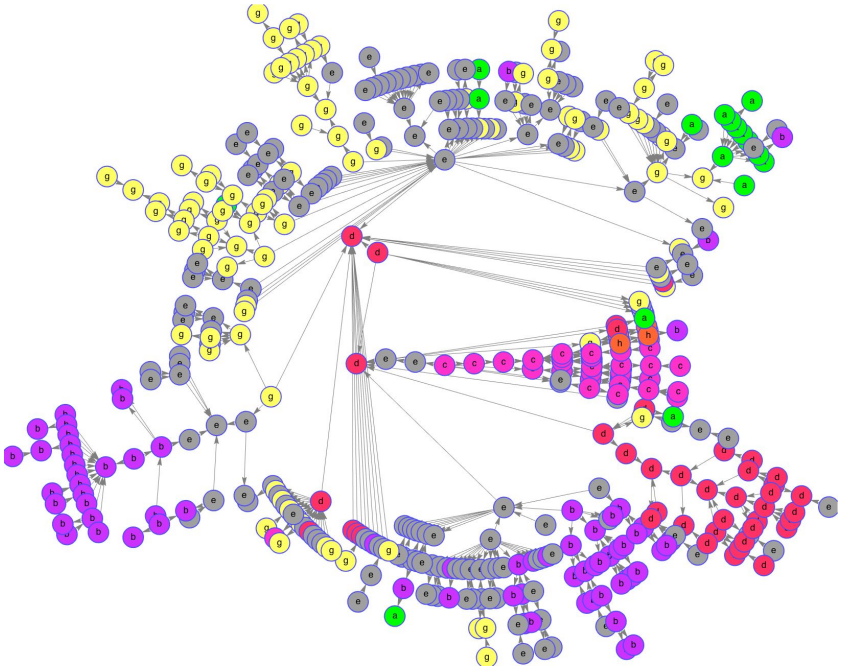
➔ Other performance metrics [2]

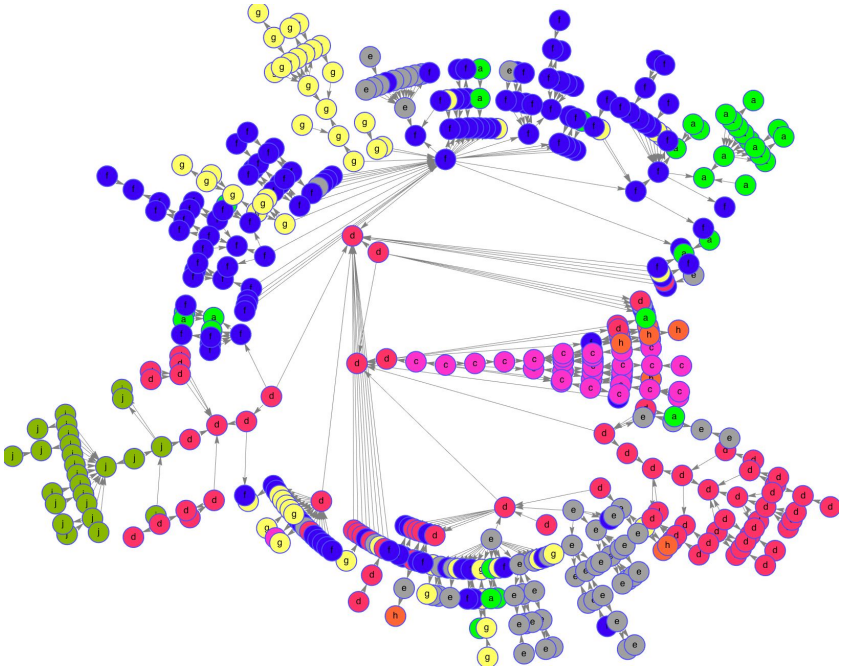
e.g. Entropy,NMI



1. Heimo, Tapio, Kimmo Kaski, and Jari Saramäki. "Maximal spanning trees, asset graphs and random matrix denoising in the analysis of dynamics of financial networks." *Physica A: Statistical Mechanics and its Applications* 388.2 (2009): 145-156.
2. "Evaluation of Clustering." *Evaluation of Clustering*. Web. 20 Apr. 2016.<http://nlp.stanford.edu/IR-book/html/htmledition/evaluation-of-clustering-1.html>

K-means Clustering vs. Spectral Clustering





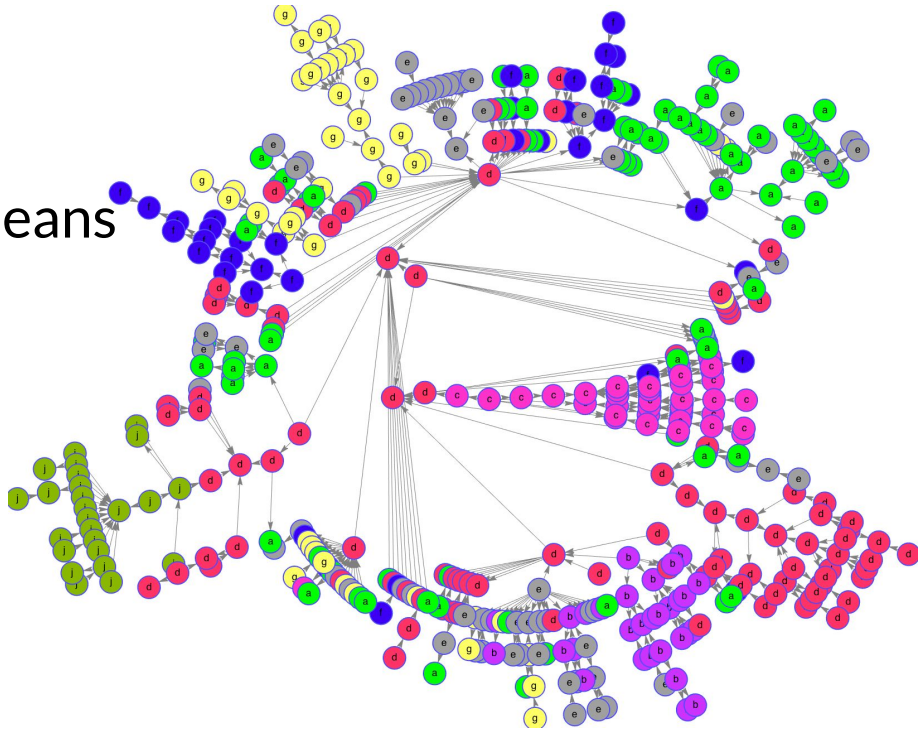
RMT + K-means

Just like Spectral Clustering, we

- Introduce RMT conclusions to K-means
- Eliminate market influence
- Create a new time-serie dataset:

  ‘Residual’

➔ Result



Plerou, Vasiliki, et al. "Random matrix approach to cross correlations in financial data."

Observations

Unbalanced data

6 stocks in Telecom, 29 Material and 27 Utility

Strong and weak correlations matter

e.g. why such labels as Utility would cluster better than others

➔ **MAKE USE OF CLUSTERING RESULTS**

GICS Sectors

a - Consumer Discretionary
b -Consumer Staples
c -Energy
d - Financials
e - Health Care
f - Industrials
g - Information Technology
h - Materials
i- Telecommunications Services
j - Utilities

Application - Optimizing Portfolio

New data set:

Smaller data set by choosing 5 stocks from each of 8 industry categories---total 40 stocks

Stocks price of 2010 for training and constructing a portfolio

Stocks price of 2011 to test how good our portfolio is

Reason: balanced data set; computational problem.

Applied cluster method: k-means; RMT+k-means; Spectral clustering

Portfolio Composition

Markowitz Portfolio Theory

(Mean-Variance Analysis)

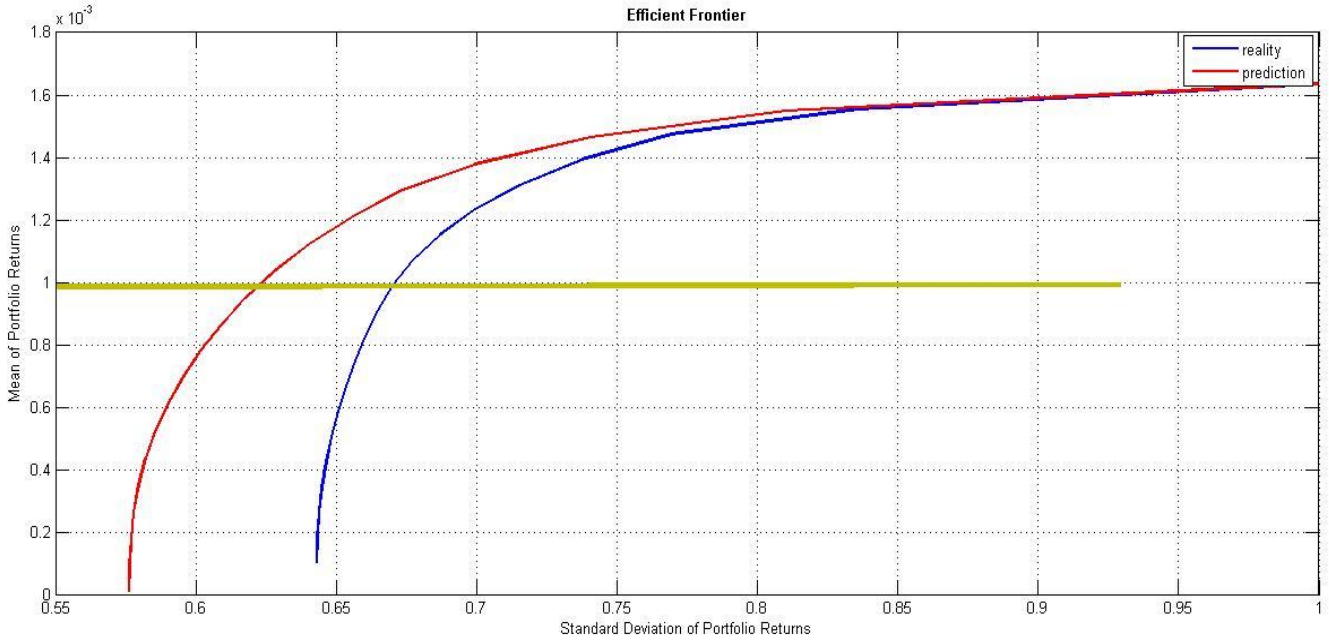
---1952 Nobel Prize in economics

min $$\sigma_p^2 = \sum_i \sum_j w_i w_j \sigma_i \sigma_j \rho_{ij}$$

s.t $$\mathrm{E}(R_p) = \sum_i w_i \, \mathrm{E}(R_i)$$

$$\sum_i w_i = 1$$

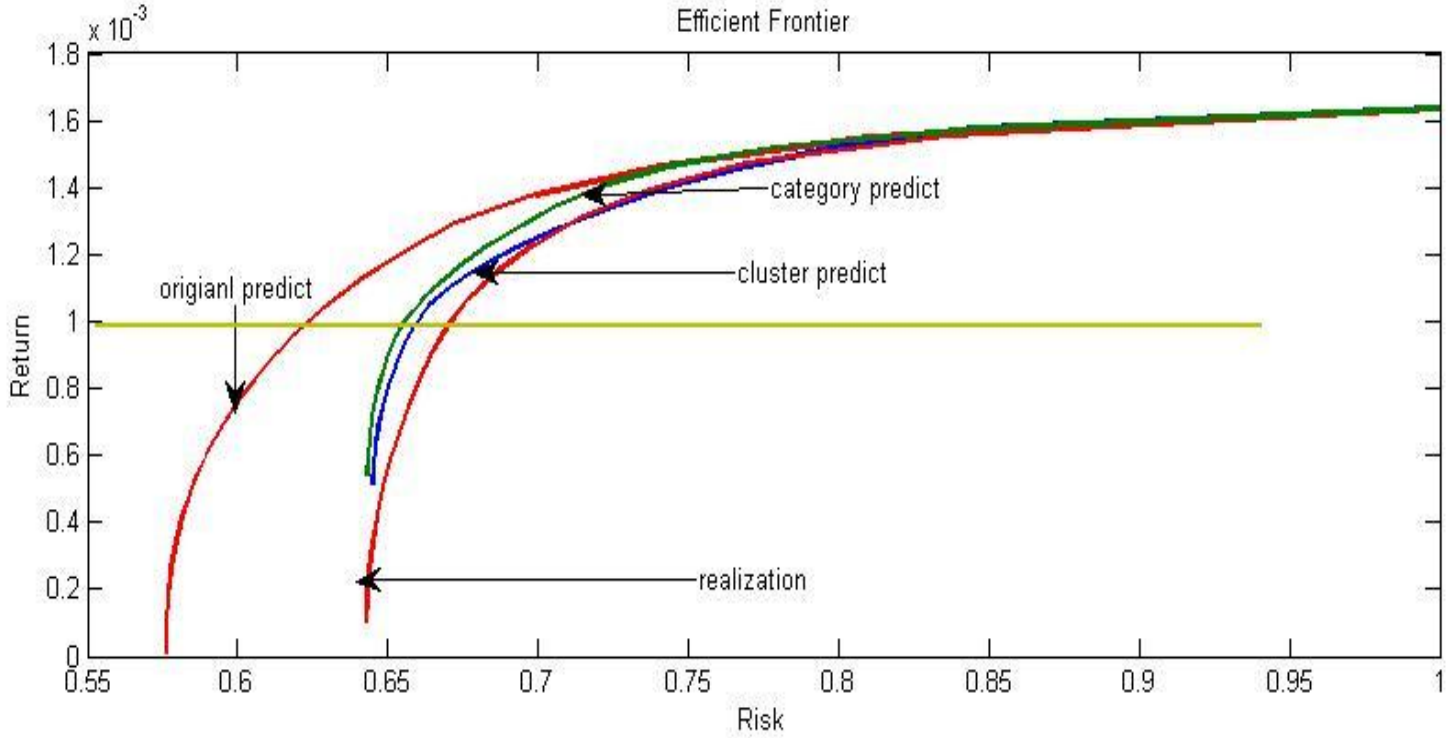
Efficient Frontier Plot



Using information only in 2010 to predict best portfolio for 2011 will underestimate the actual risk!!

What to do? Choosing just one stock from each cluster, and form a portfolio of these 8 stocks.

Efficient Frontier Plot



1.Using clusters to predict optimal portfolio beats the whole market strategy, even though with much fewer choice of portfolio composition.

2. It' s better than just using Industry category as well!!!

Reference

1. Heimo, Tapio, et al. "Spectral and network methods in the analysis of correlation matrices of stock returns." Physica A: Statistical Mechanics and its Applications 383.1 (2007): 147-151.

2. Plerou, Vasiliki, et al. "Random matrix approach to cross correlations in financial data." Physical Review E 65.6 (2002): 066126.

3. Sharpe, M. "Lognormal model for stock prices." (2004).

4. Langone, Rocco, Raghvendra Mall, and Johan AK Suykens. "Clustering data over time using kernel spectral clustering with memory." Computational Intelligence and Data Mining (CIDM), 2014 IEEE Symposium on. IEEE, 2014.

5. Onnela, Jukka-Pekka, et al. "Asset trees and asset graphs in financial markets." Physica Scripta 2003.T106 (2003): 48.

6. Edelman, Alan, and N. Raj Rao. "Random matrix theory." Acta Numerica 14 (2005): 233-297.