# N-Nearest Portfolio Allocation

An application of Financial Time Series Clustering

Chun-Hui,Wu

2017.12.18

# Outlines

1. Introduction

2. Data pre-processing

3. Cluster method
   - DTW - HAC
   - Cepstral - HAC
   - Cepstral - Kmeans

4. Result

5. Conclusion

# 1.Introduction

- What is clustering?
  - An unsupervised method to learn “specific structure” in data via Algorithm
  - Hard-clustering : K-means , Hierarchical Clustering
  - Soft-Clustering: Gausian Mixture Model

- Compare
  - time series plot.
  - mean-std plot.

# 2. Data pre-processing

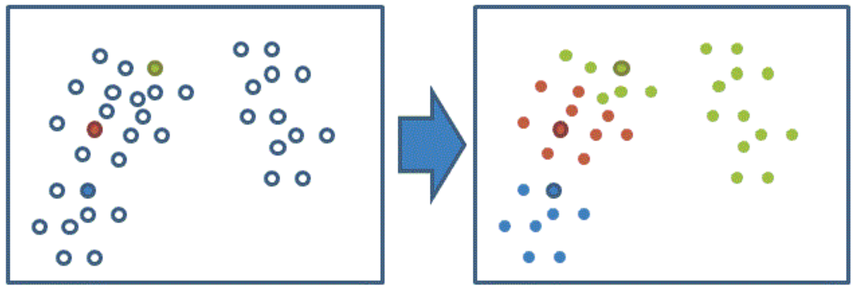
- The data set consists of 77 mutual fund price process with different time length .
- Choose the price data in every series which dates are between 2014/10/8 and 2017/10/16 .
- Since we want to build time series model , and use the Markwitz Frame work to analysis ,we have to smooth missing values in series.
- All series are normalized to 2014/10/8 ,
  by the following formula $P_t = \frac{p_t - p_1}{p_1}$

# 3. Clustering Method

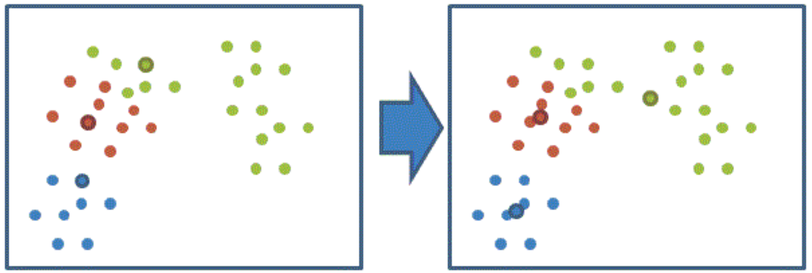
- K-means clustering
- Hierarchichal clustering
- DTW distance
- Cepstral method

# Kmeans Clustering

- In K-means method , we need the coordinate and the definition of distance between two points.



- Once we have the coordinate of the data , we repeat the following two step until the mean converge.



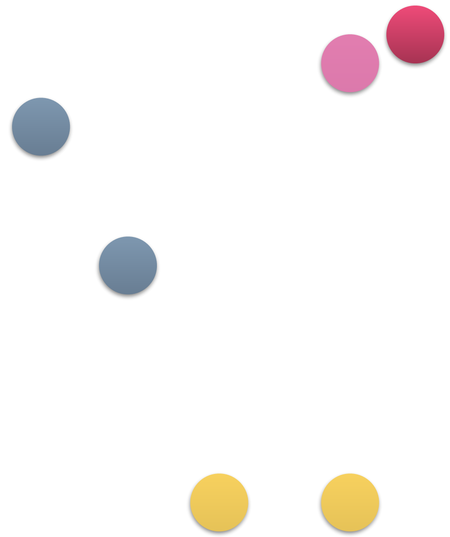
# Hierachichal Clustering

- In hierarchical clustering, we need to define distance between object , and distance between cluster . (Do not need coordinate!)

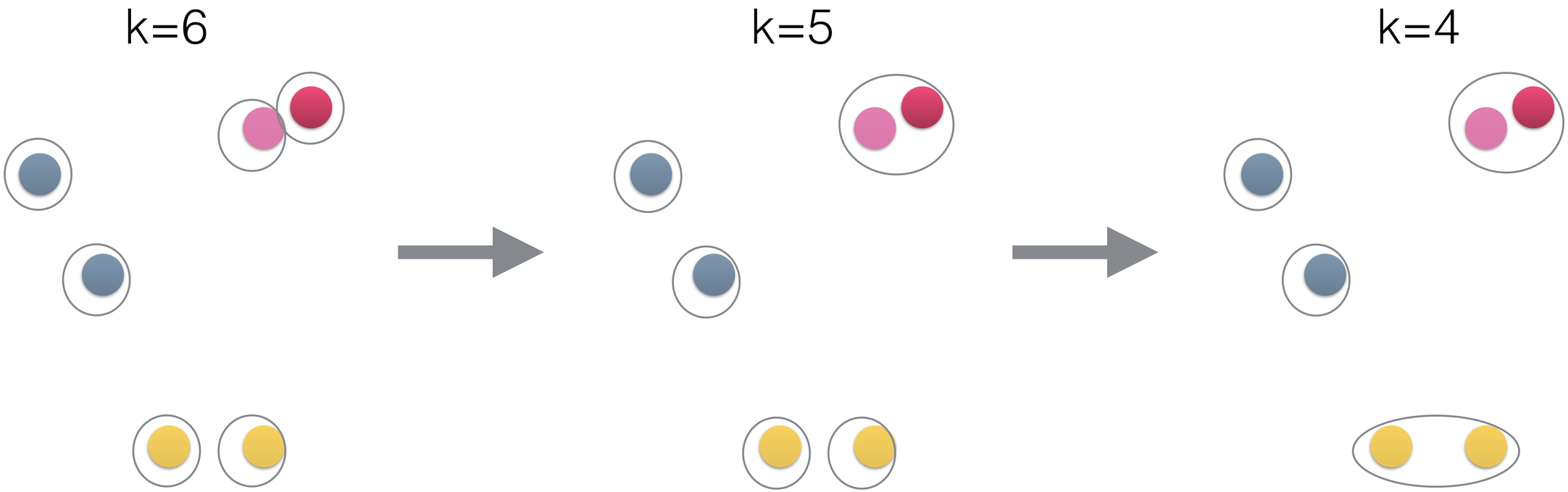
- There are several way to define distance between cluster
  - complete linkage

$$D(Clust_i, Clust_j) \equiv \max_{s,k} d(s,k), s \in Clust_i, k \in Clust_j$$

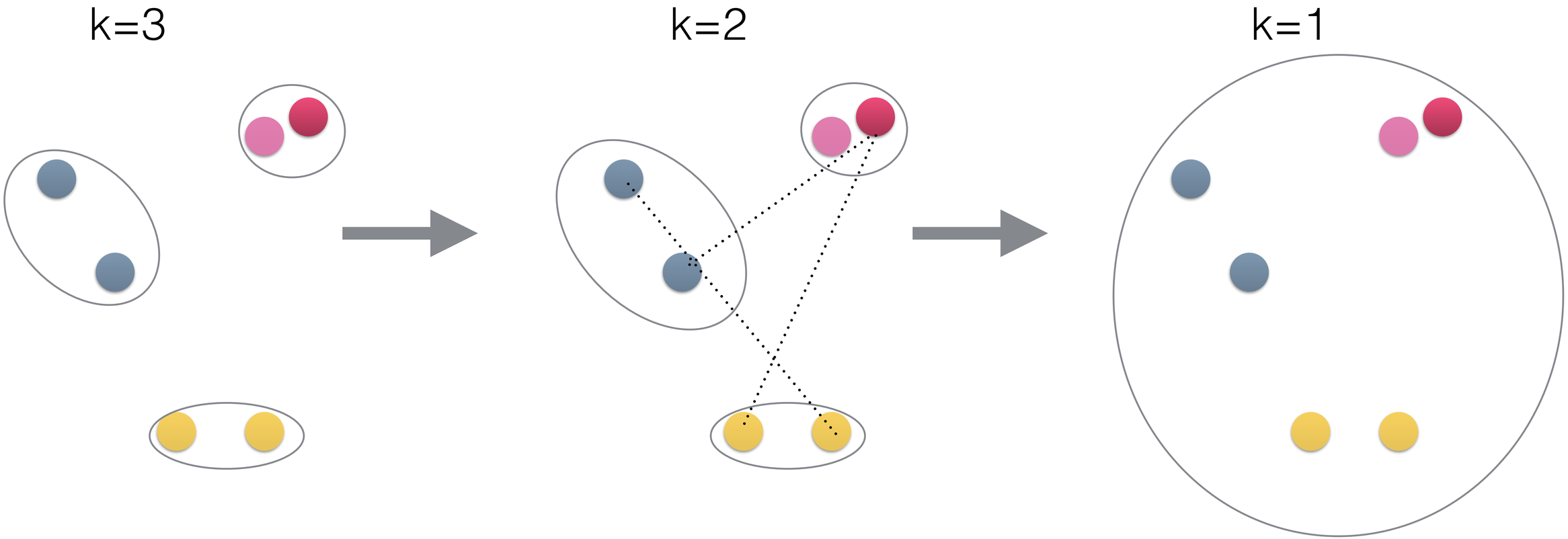
- Agglomerative clustering :
  Data points starts in its own cluster, and repeat the step — merge the closest two clusters .



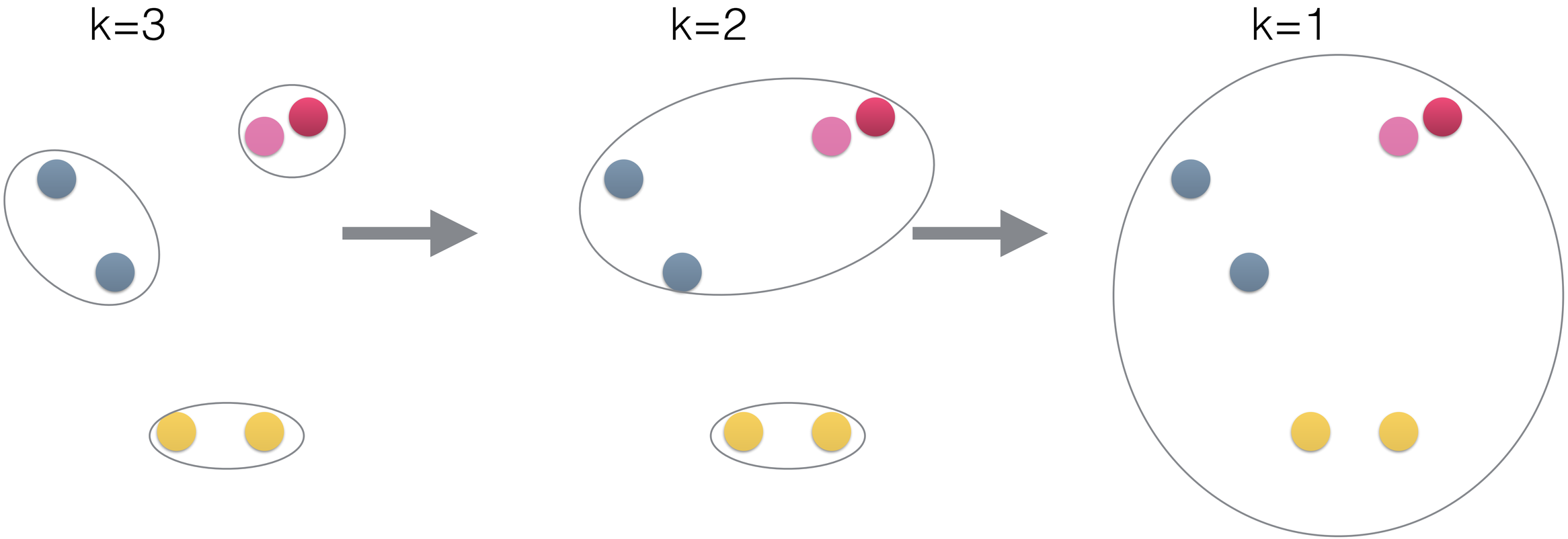
# Agglomerative with Complete linkage distance



# Agglomerative with Complete linkage distance



# Agglomerative with Complete linkage distance



# Raw based method

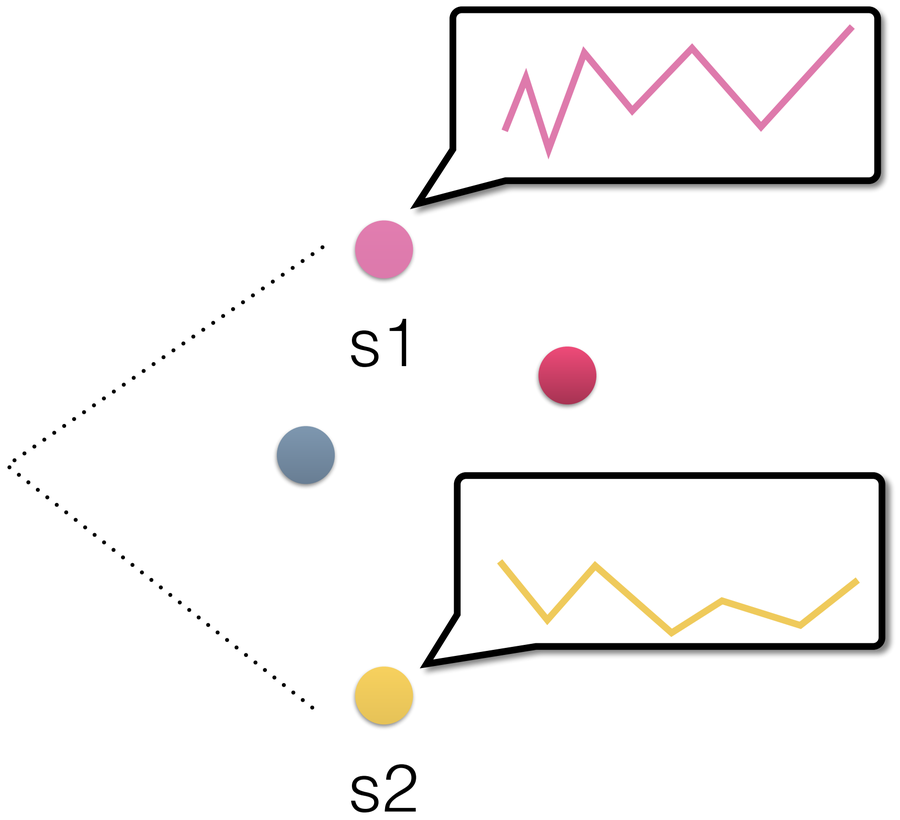
- Define distance between objects

  -> apply clustering method

- Distance Measure is more important than clustering method

- General for almost every domain

Define
d(s1,s2)



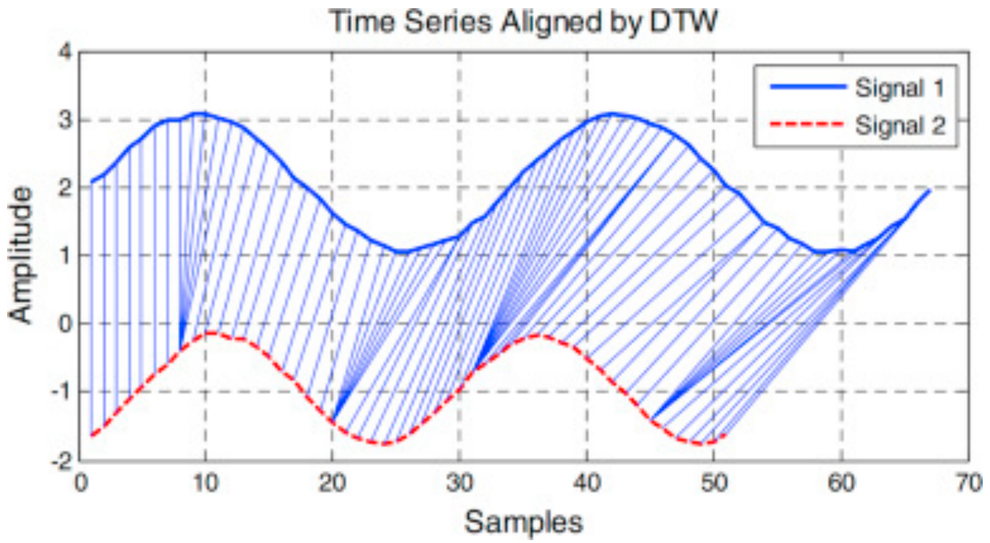
# DTW Distance

- Given two time series
  s1 (length m ), s2 (length n) ,
  the DTW distance is defined by:

$$D(i,j) = |s1(i) - s2(j)| + min(D(i,j-1), D(i-1,j-1), D(i-1,j))$$

- One should recursively solve the formula above to find DTW distance

- DTW can compare the stretched or compressed time series with different time length

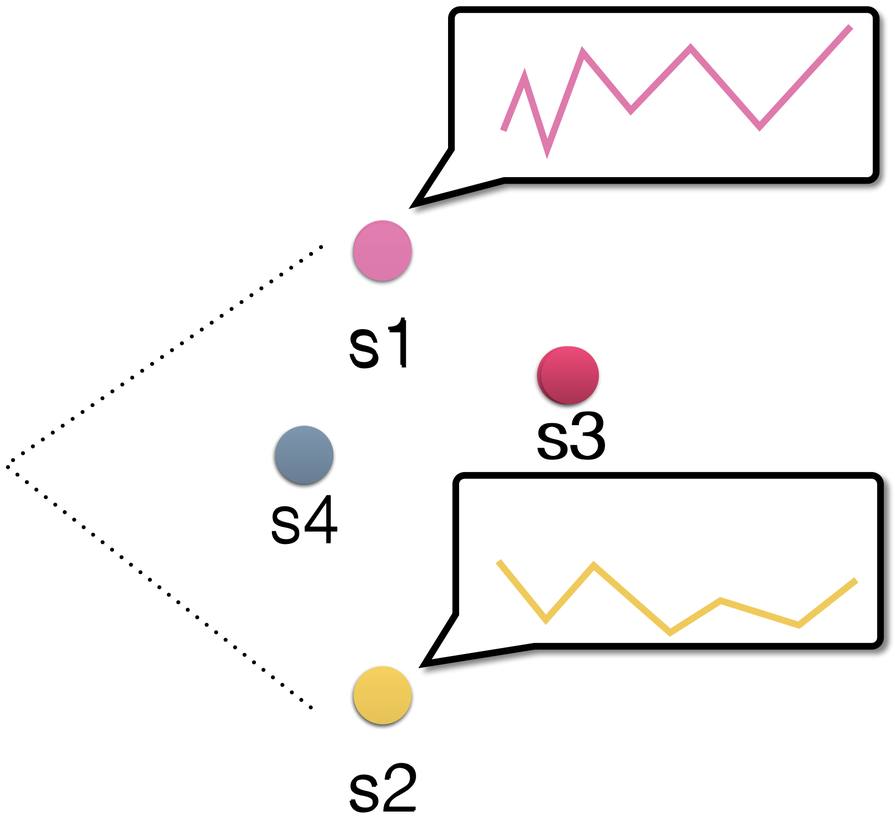


# Calculate Dissimilarity Matrix

Define d(si,sj) s.t.

(1) d(si,sj)=d(sj,si)

(2) d(si,si)=0



# Calculate Dissimilarity Matrix

| | s1 | s2 | s3 | s4 |
|---|---|---|---|---|
| s1 | 0 | d12 | d13 | d14 |
| s2 | d12 | 0 | d23 | d24 |
| s3 | d13 | d23 | 0 | d34 |
| s4 | d14 | d24 | d34 | 0 |

# Cepstral Method

- Domain dependent
- Different models for different applications
- Assume all series follow ARIMA(15,1,0)
- Calculate the cepstral coefficient to define the “distance” of series
- Cepstral method can apply K-means or HAC

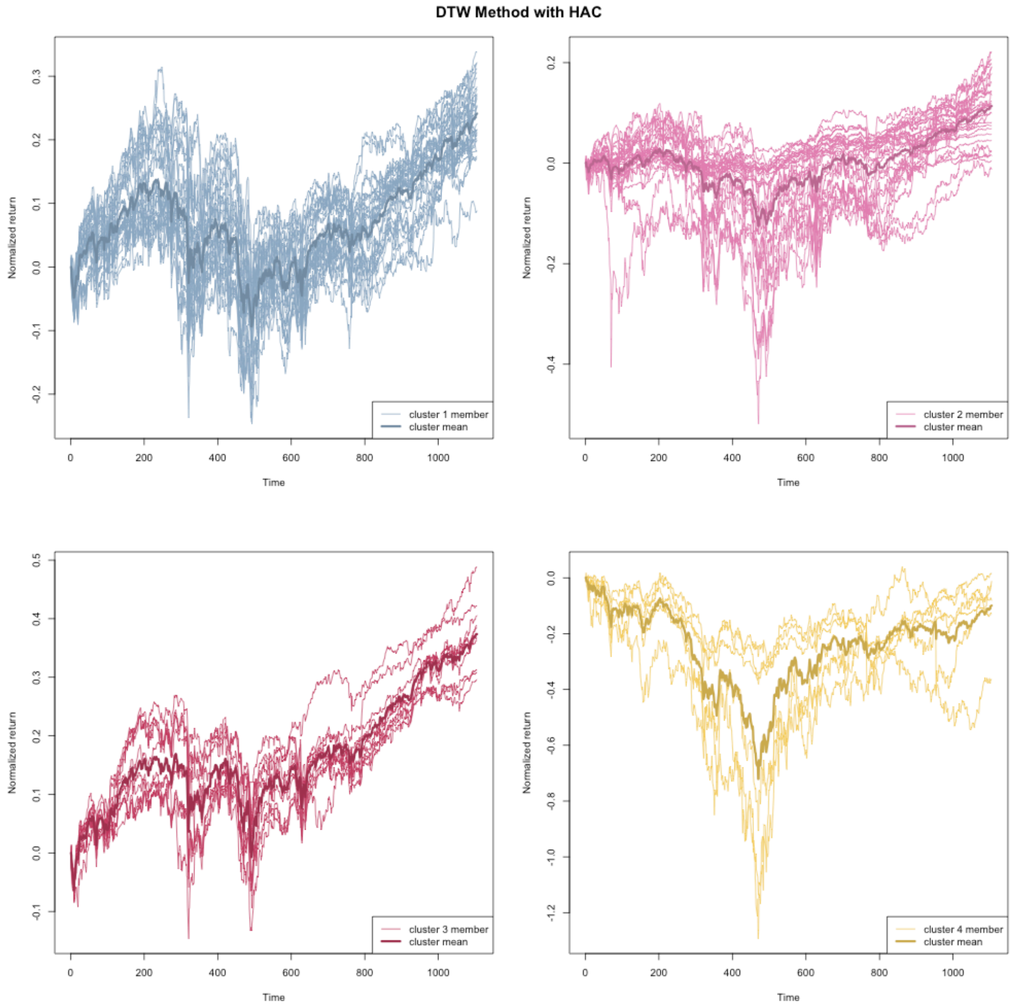
# Cepstral Method - Step

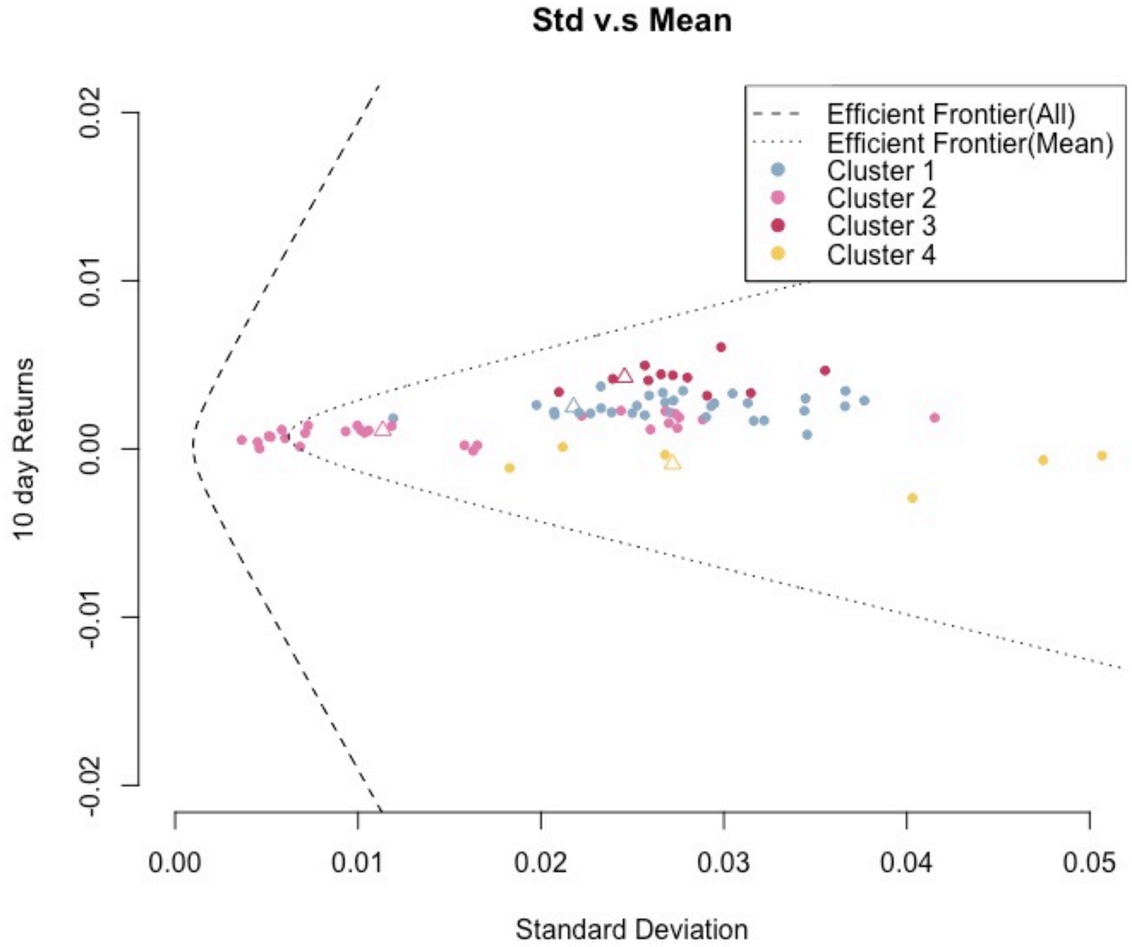
- Step 1: Given time series si , calculate the model ARIMA(15,1,0) and define $x_n = [\alpha_{i1}, \alpha_{i2}, ..., \alpha_{i15}]$
- Step 2: Convert $x_i$ to $c_i$ via the cepstral formula:

$$c_{in} = \begin{cases} -\alpha_{i1}, if\ n = 1 \\ -\alpha_{in} - \sum_{m=1}^{n-1}(1 - m/n)\alpha_{im}c_{i,n-m}, if\ 1 < n \leq p \\ -\sum_{m=1}^{p}(1 - m/n)\alpha_{im}c_{i,n-m}, if\ p < n \end{cases}$$

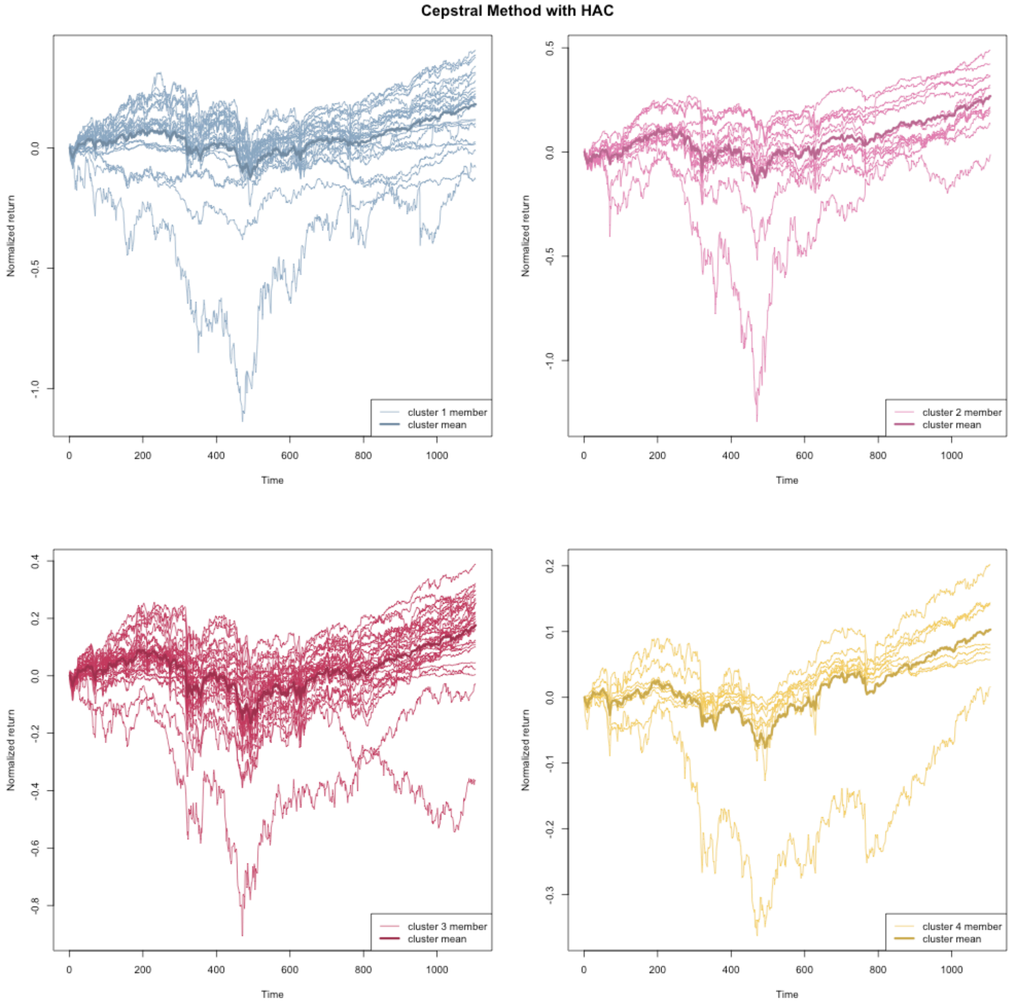
- Step 3: Apply Clustering method on set $\{c_1, c_2, ....., c_k\}$

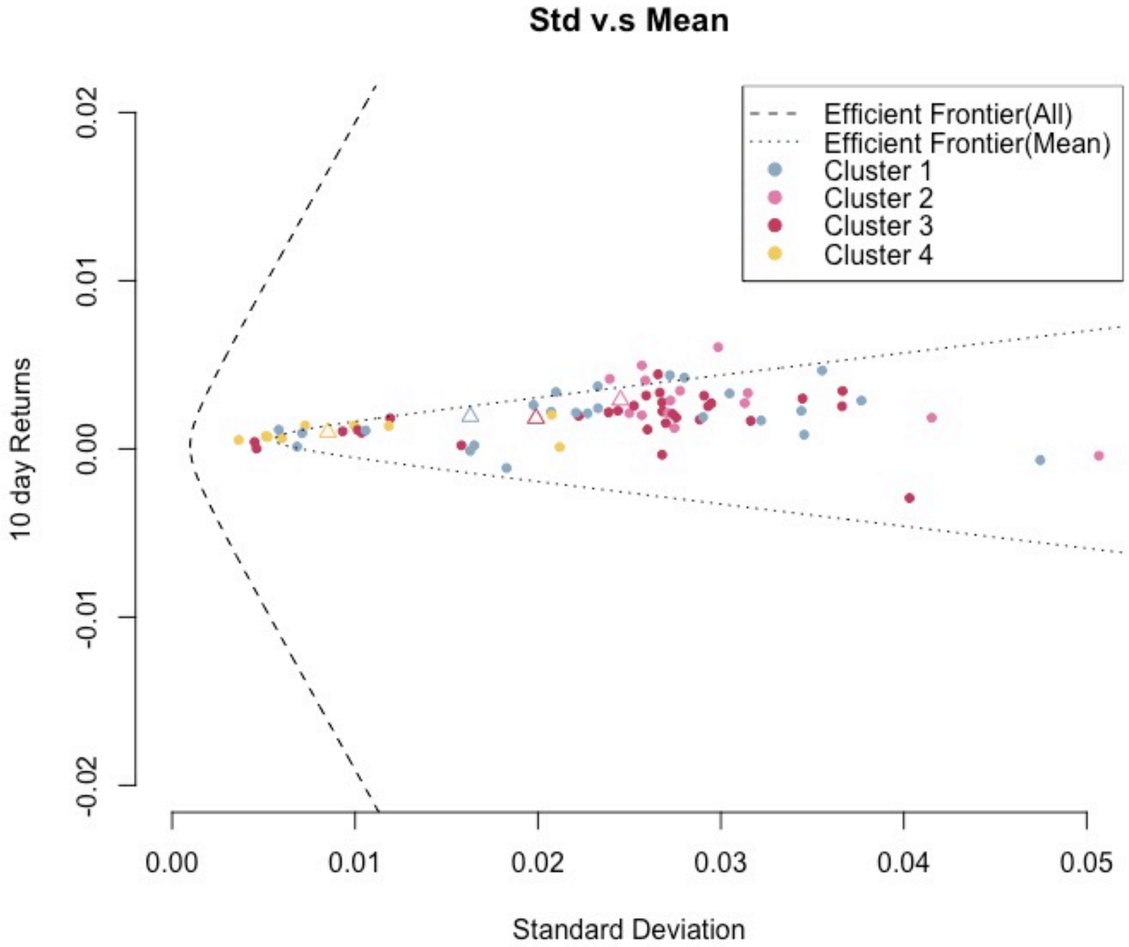
# 4.Result : DTW-HAC



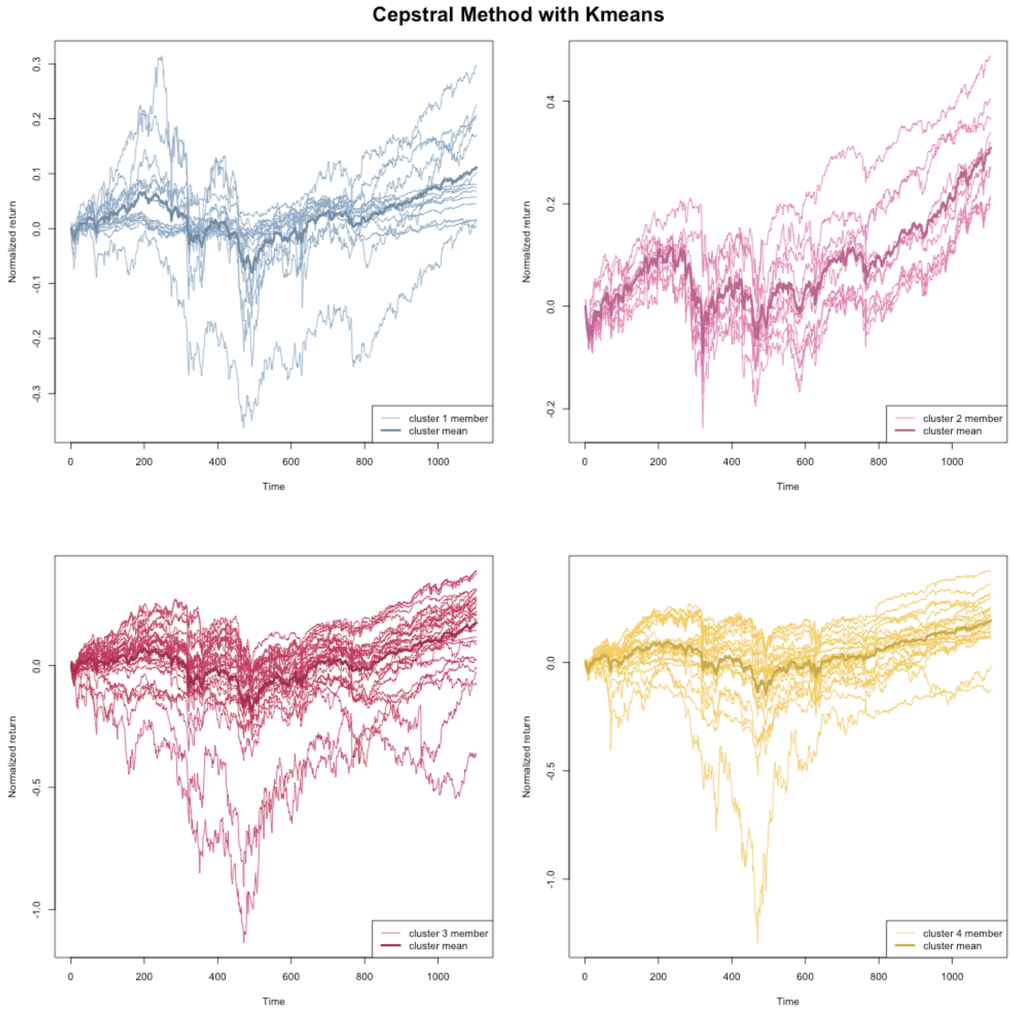



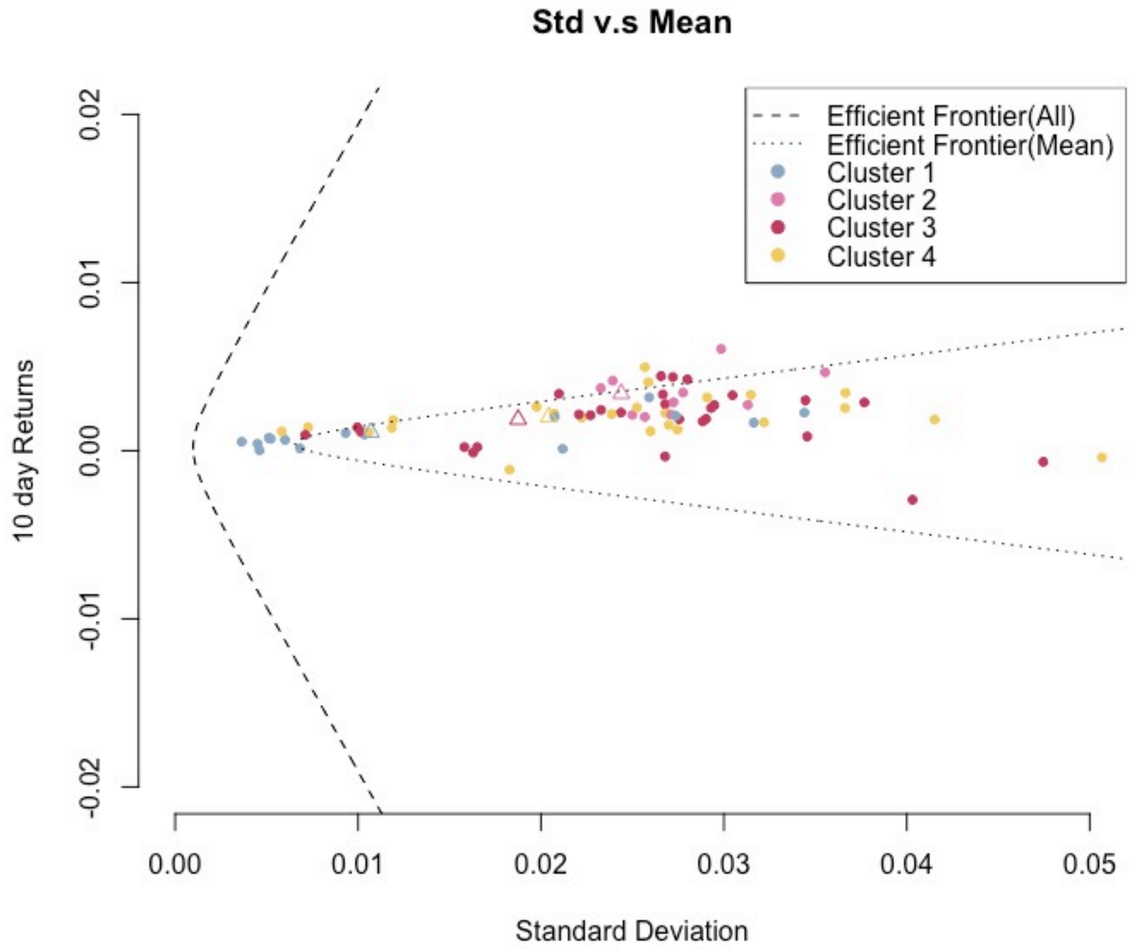
# Result : Cepstral-HAC



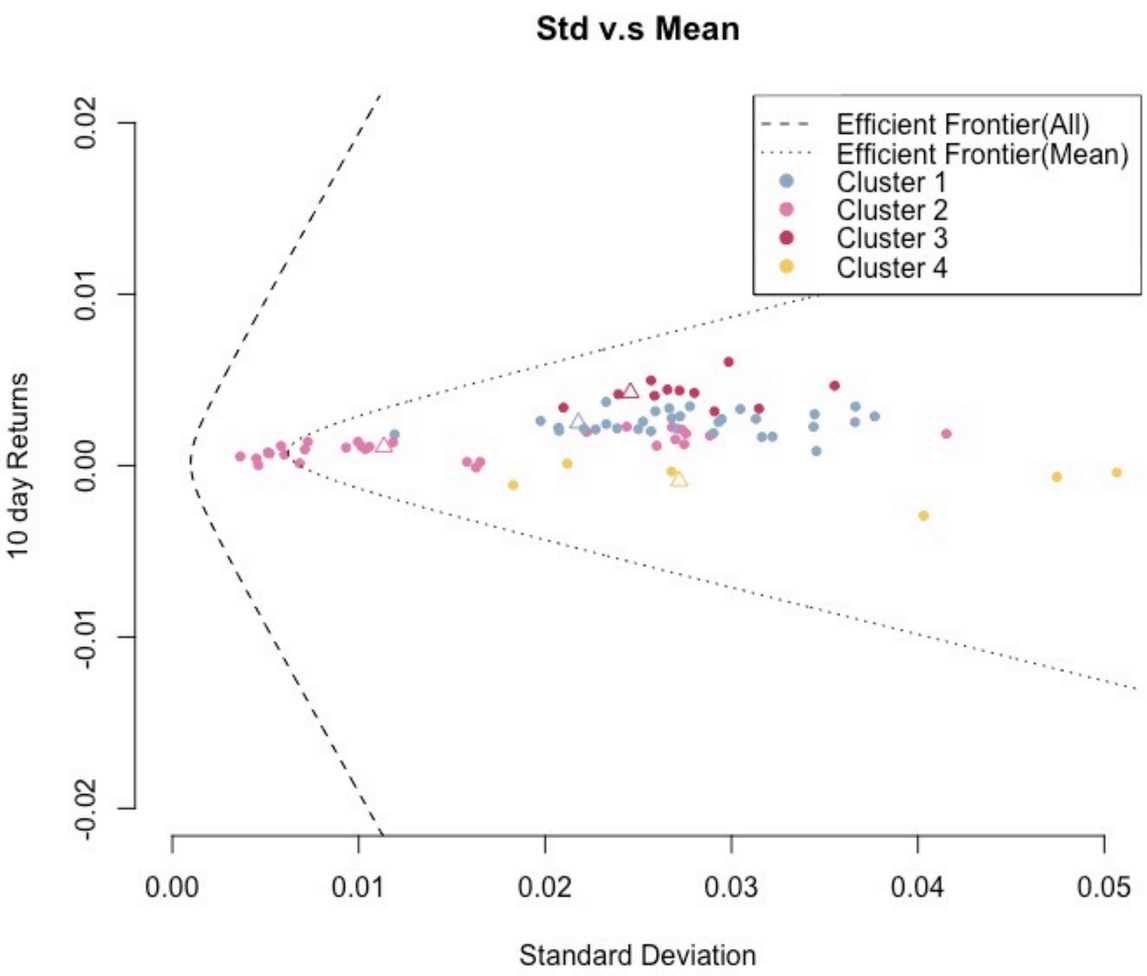



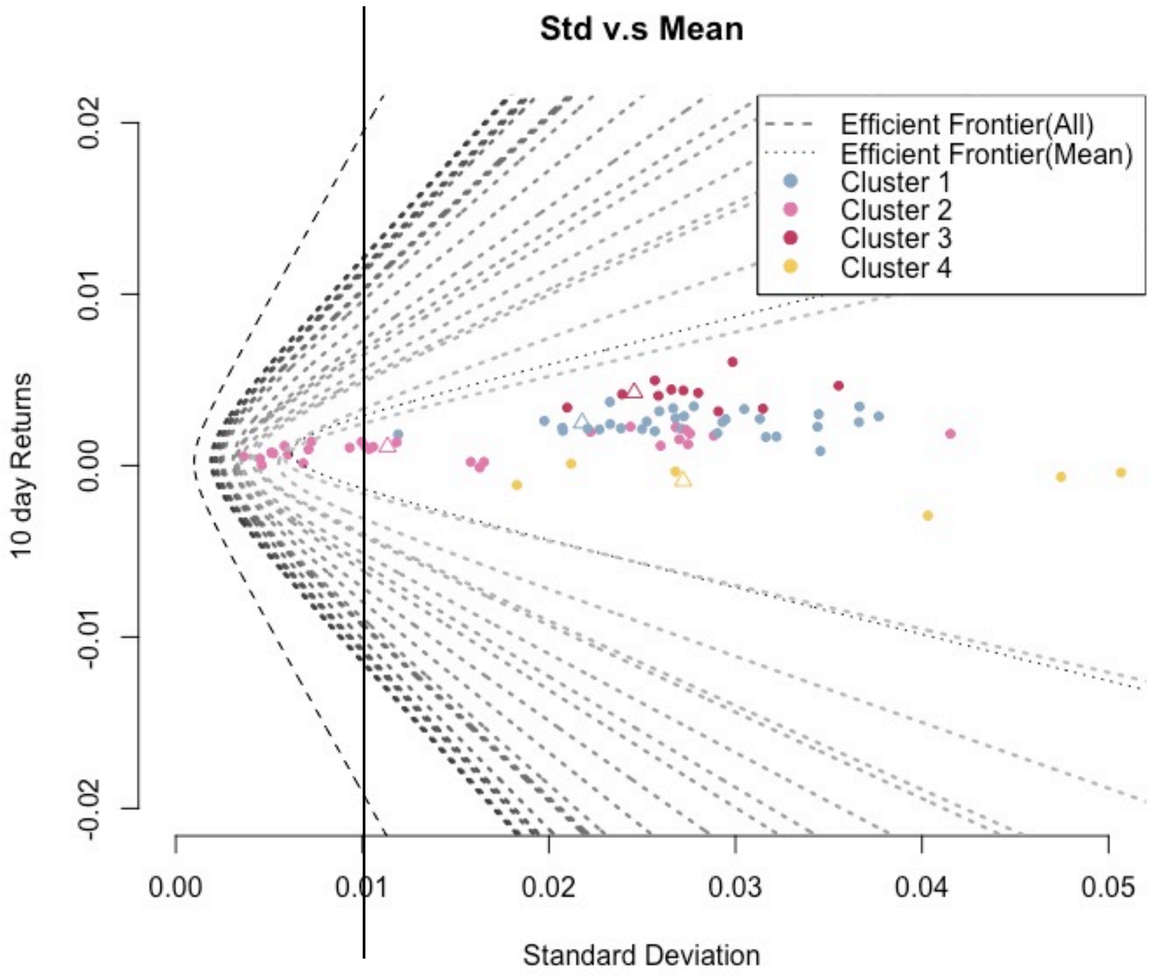
# Result : Cepstral-Kmeans



# Application N - Nearest Portfolio




We can consider the Transaction cost between return and standard deviation

# Conclusion

- DTW - HAC has better performance than the others.

- There are several reason
  - 4 is not the best group number for cepstral method.
  - ARIMA(15,1,0) is not a adequate model for the data set.
  - Eucledian distance is not suitable to describe the relation betweens the cepstral coefficient
  - Cepstral method catch the different characteristic of the series .

# Reference


- [1] Distance Measures for Effective Clustering of ARIMA Time-Series
  - Konstantinos Kalpakis

- [2] Computing and Visualizing Dynamic Time Warping Alignments in R: The dtw Package.
  - Toni Giorgino

- [3] K-Shape: Efficient and Accurate Clustering of Time Series
  - John Paparrizos

- [4]《影像學習筆記》：https://dotblogs.com.tw/dragon229/2013/02/04/89919

- [5] Clustering of time series data—a survey
  - T. Warren Liao