Overview of Clustering

2022/5/6

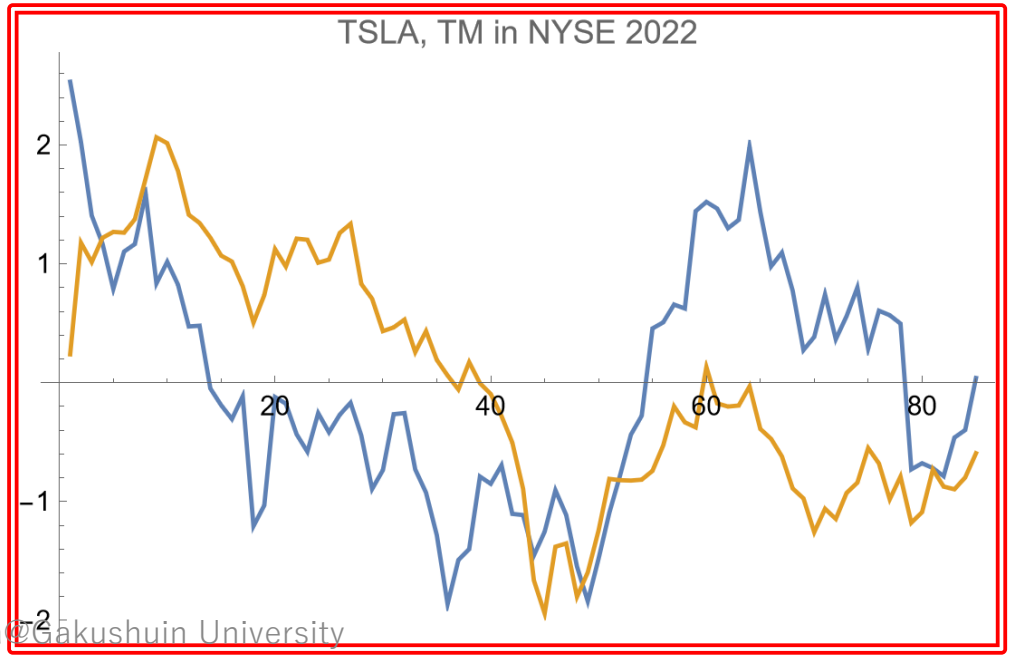
Gakushuin University

Prof Yukari Shirota

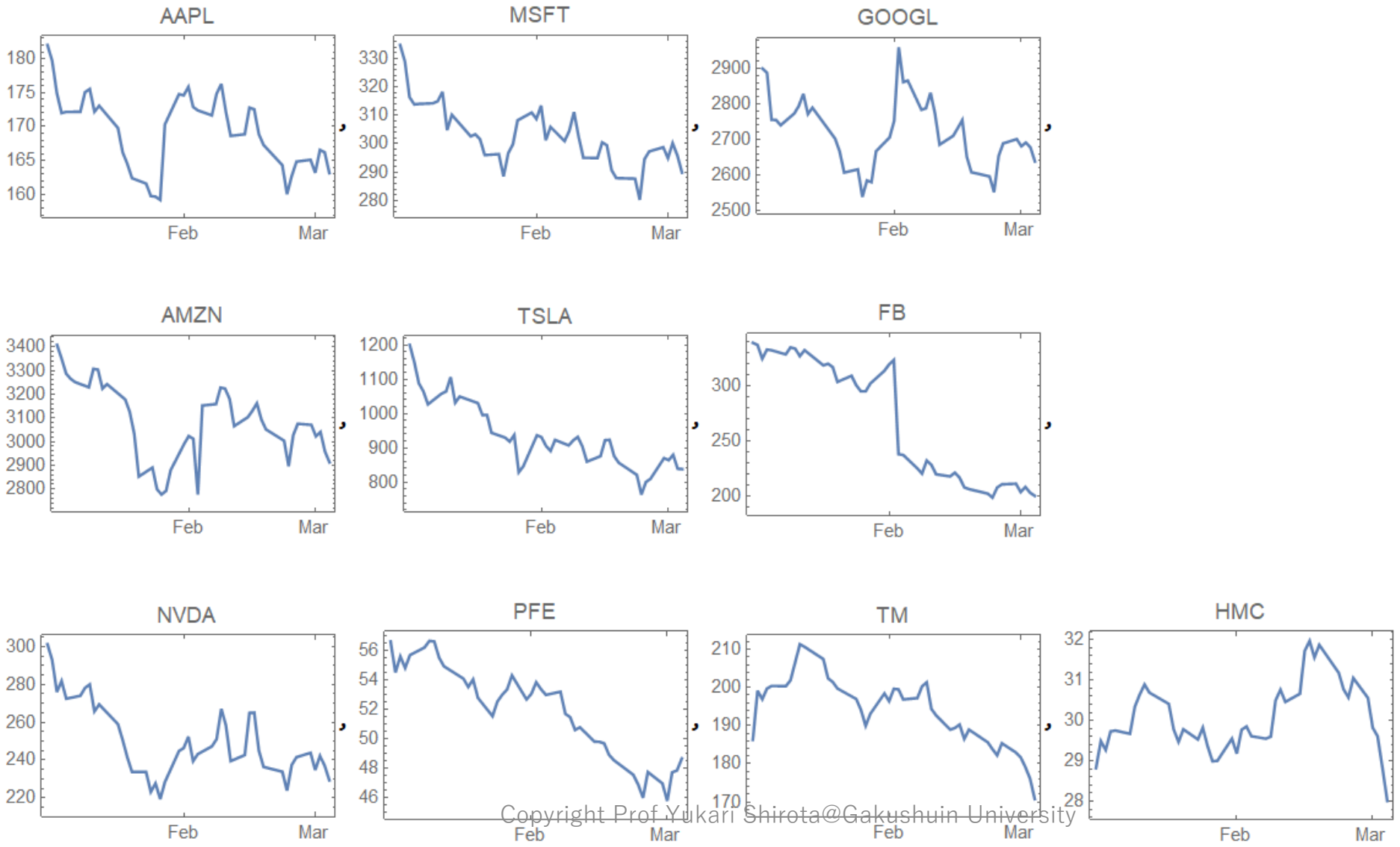
Stock Price Movement

- Currency is different
- Original scale is different
- Sometimes stock division occur

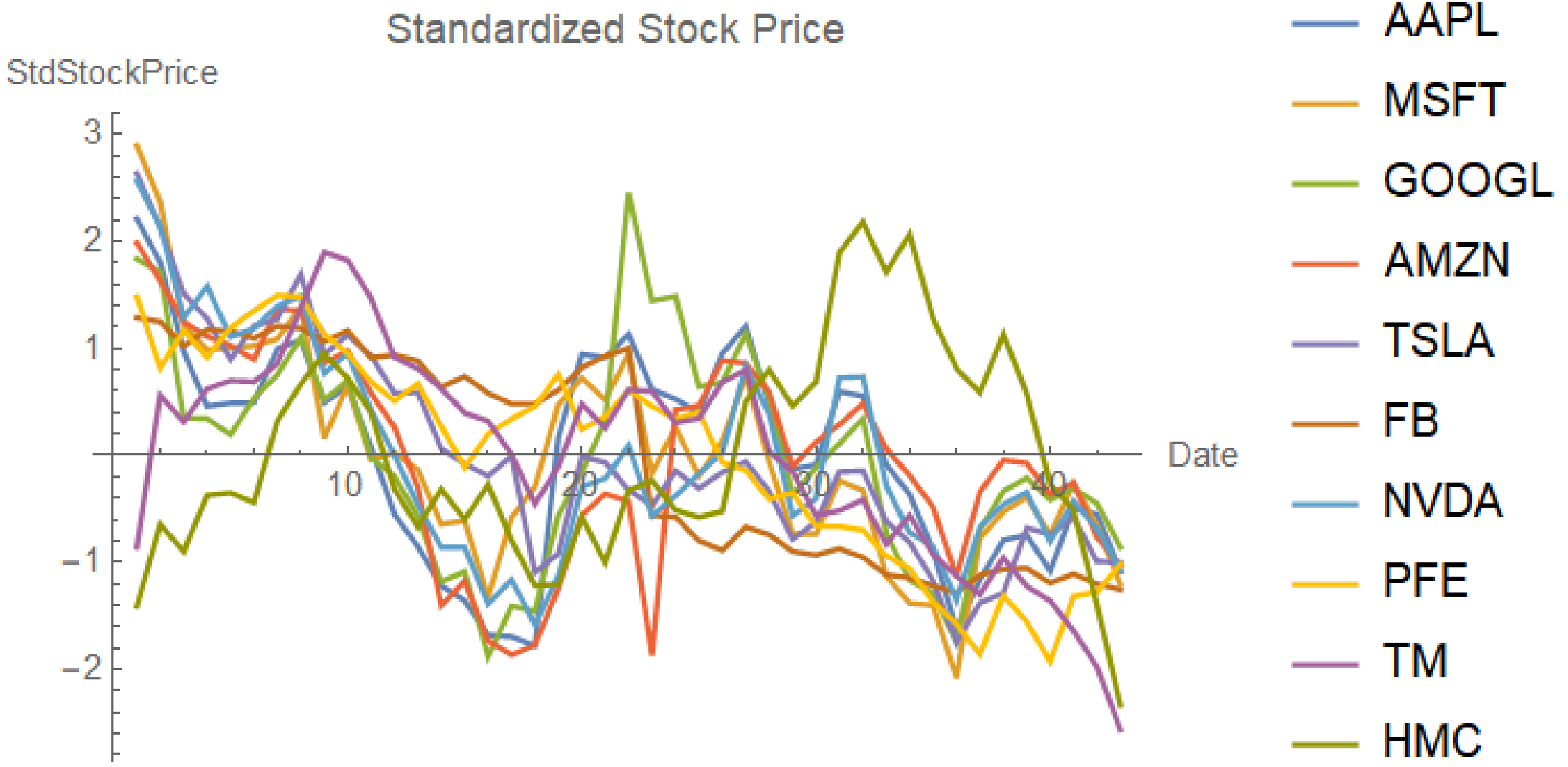
-> STANDARDIZATION is needed for COMPARISON



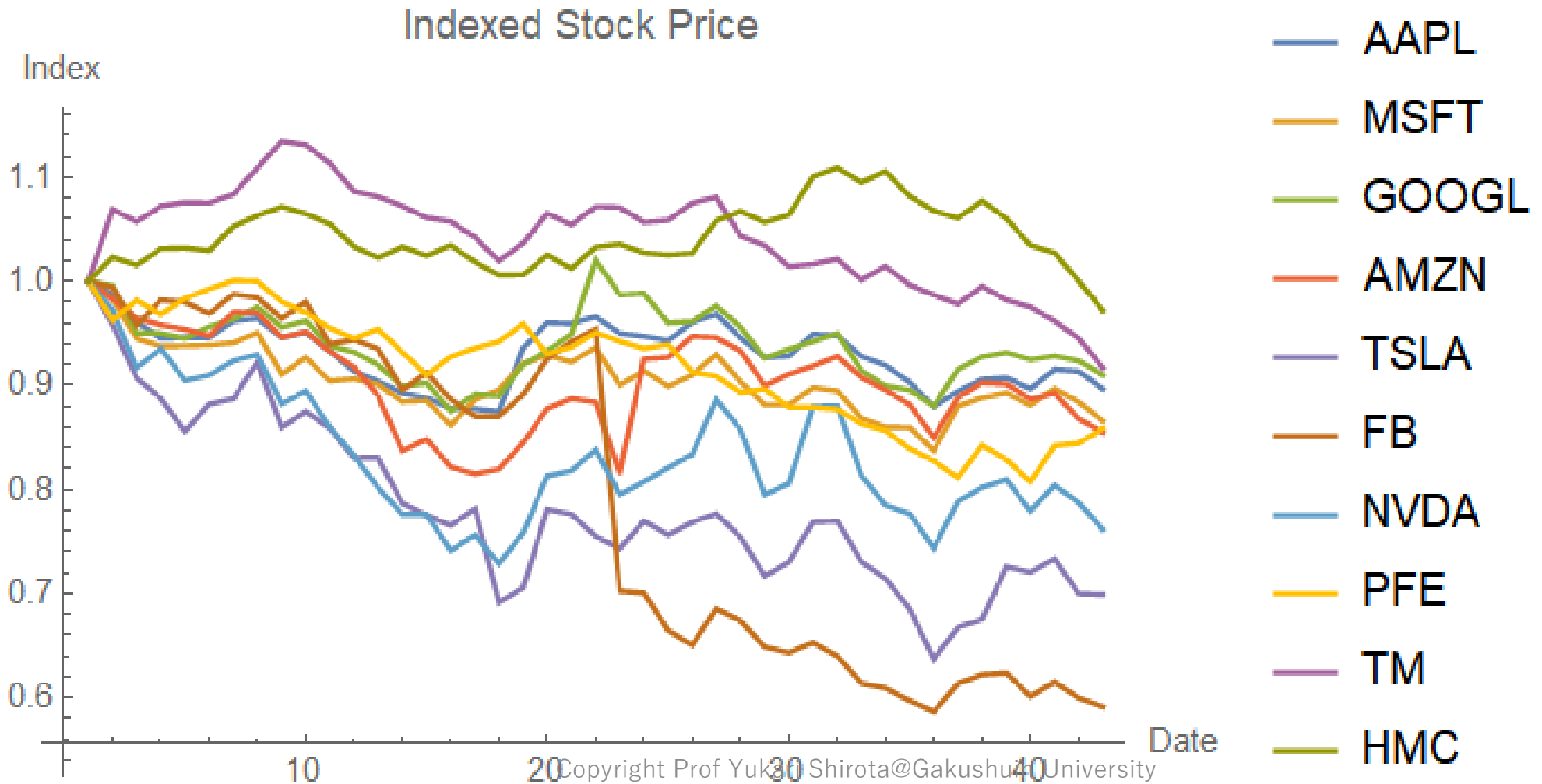
NYSE　Stock Price 2022/01/03/ to 03/04



Standardized Stock Price Movement



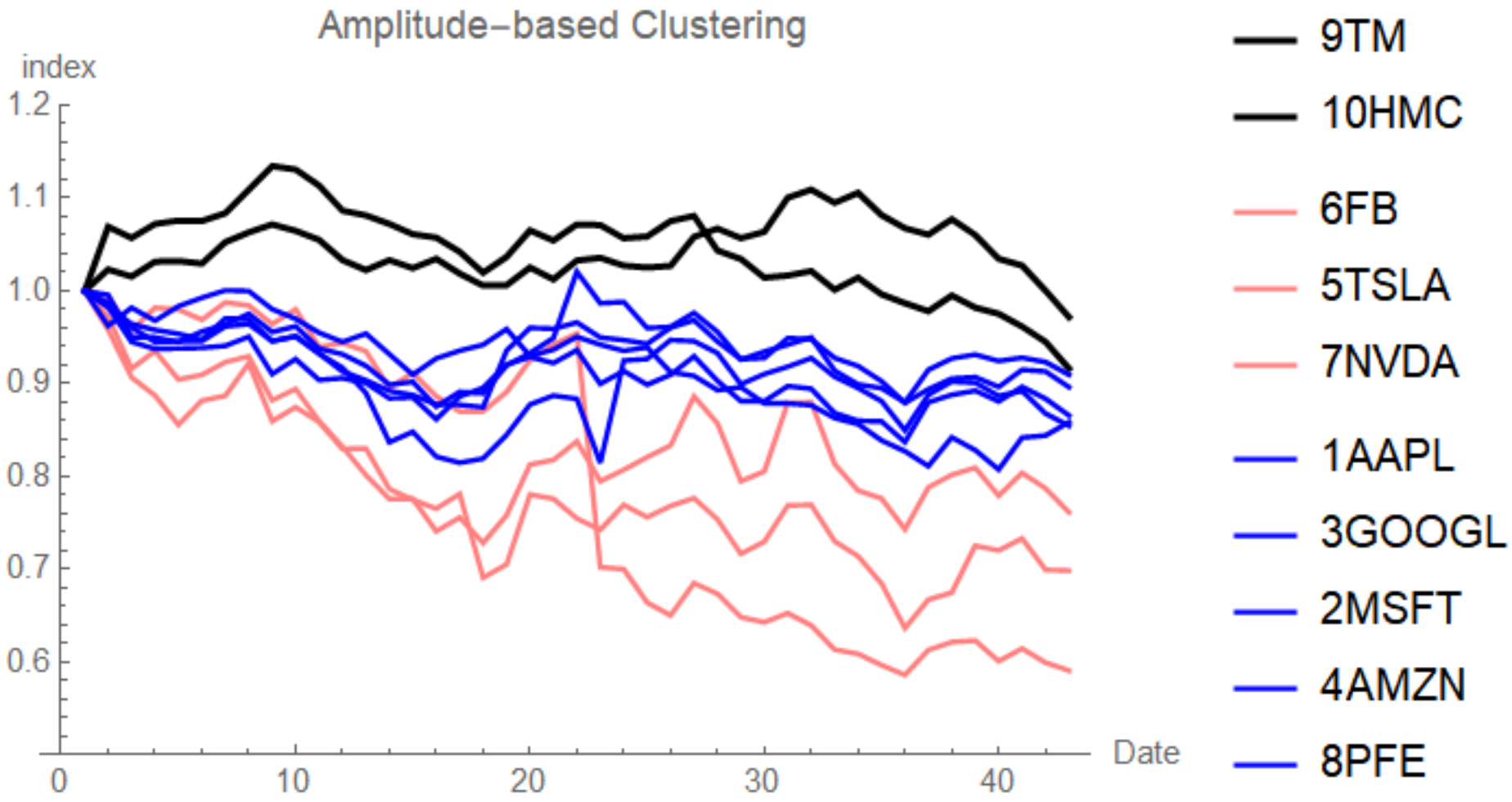
Indexed Stock Price Movement with the first one being 1



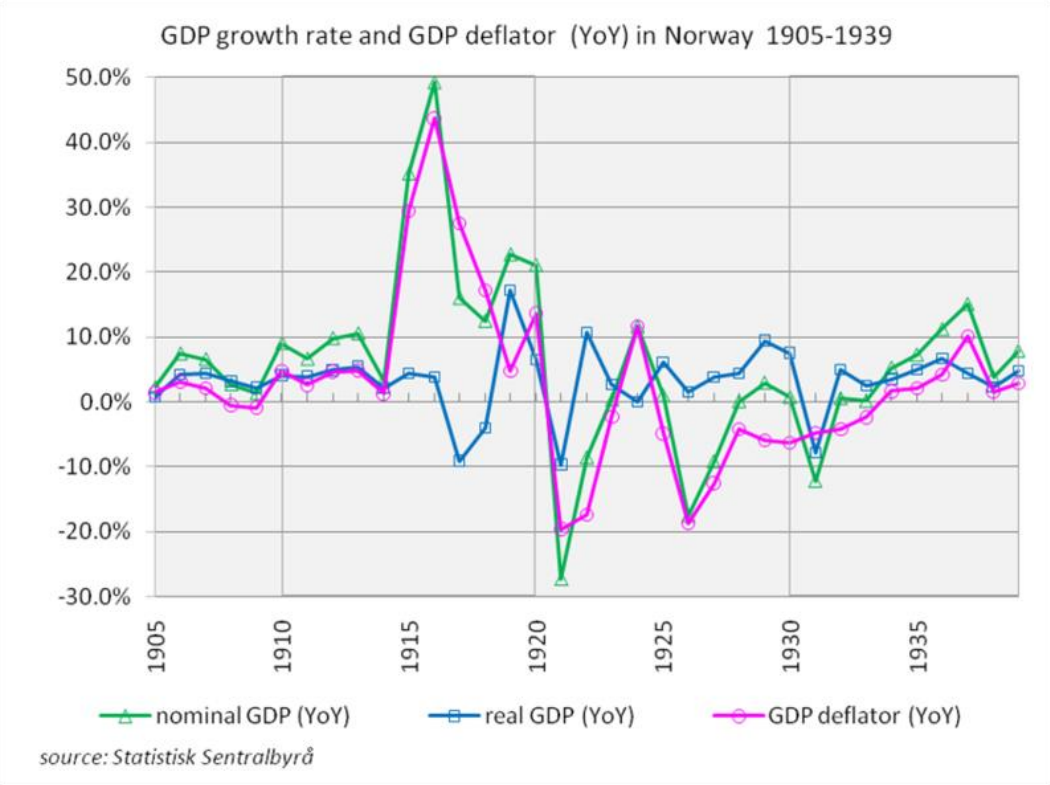
Another Requirement

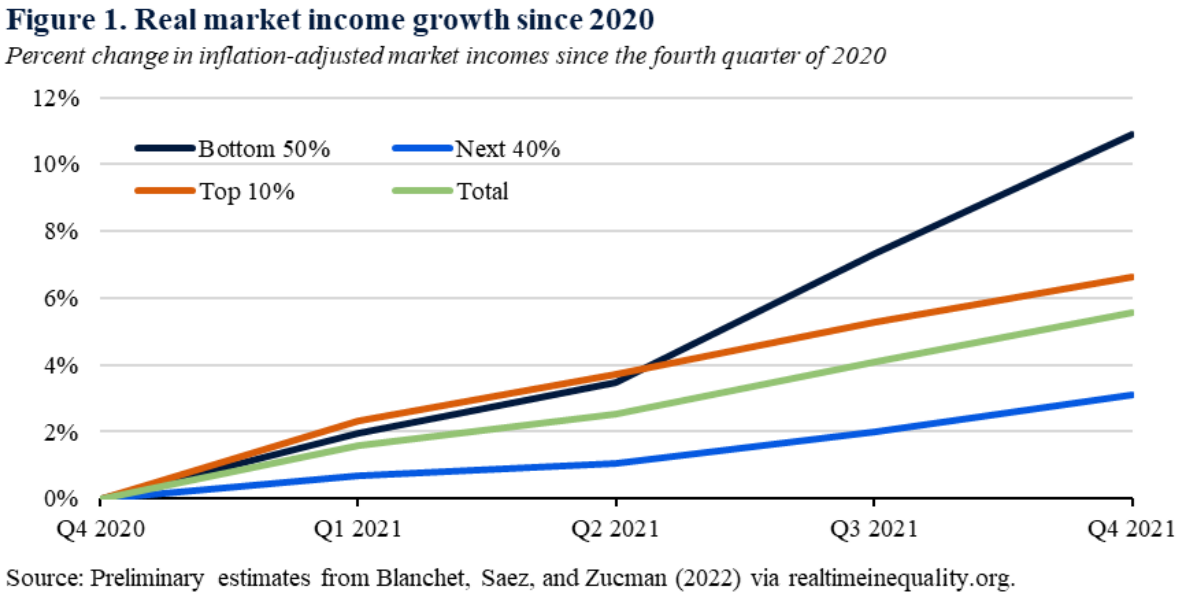
- We want to compare the growth rate/amplitude

Amplitude-based Clustering k=3



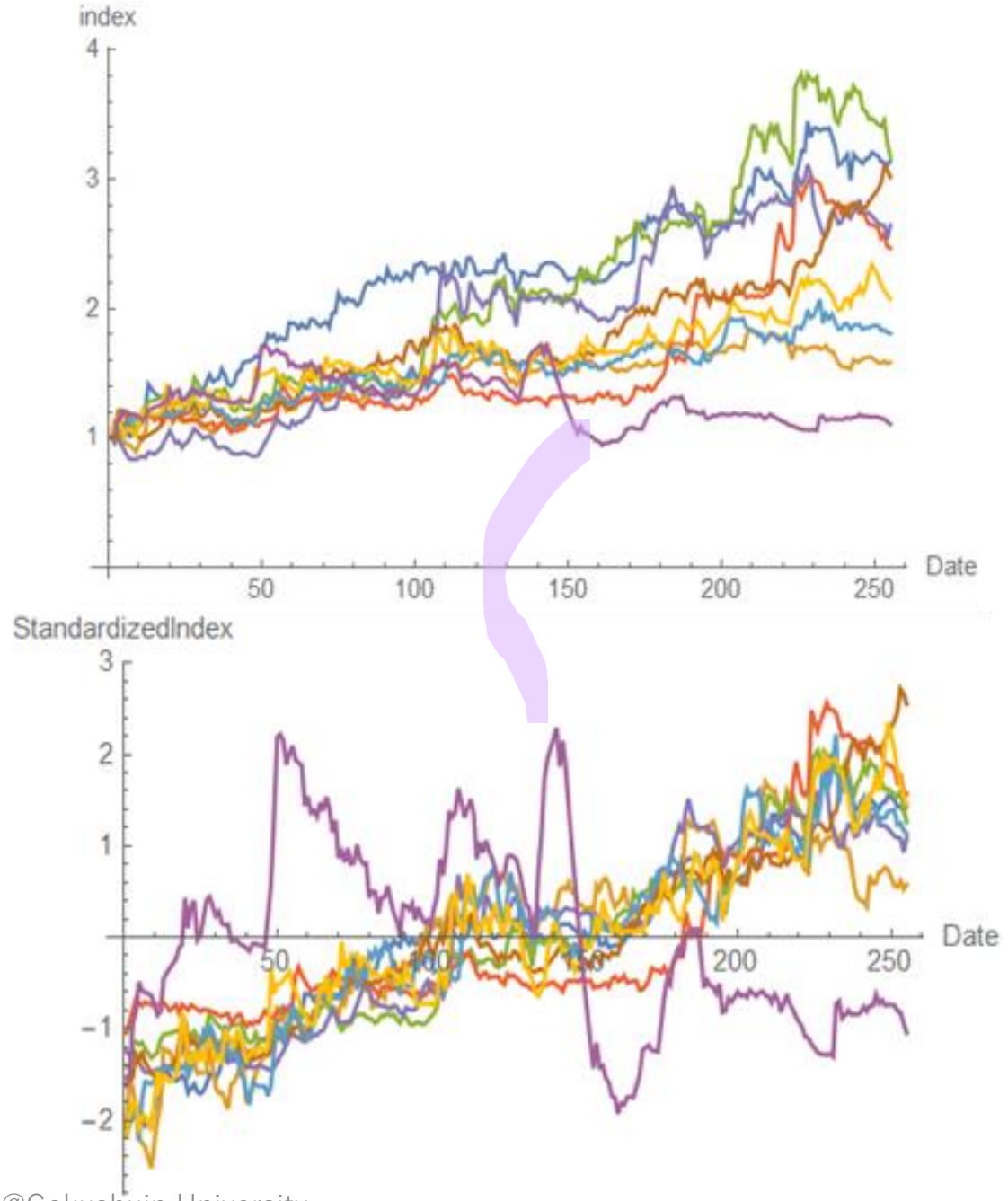
Application sample of Amplitude-based : GDP growth ratio, Economic growth





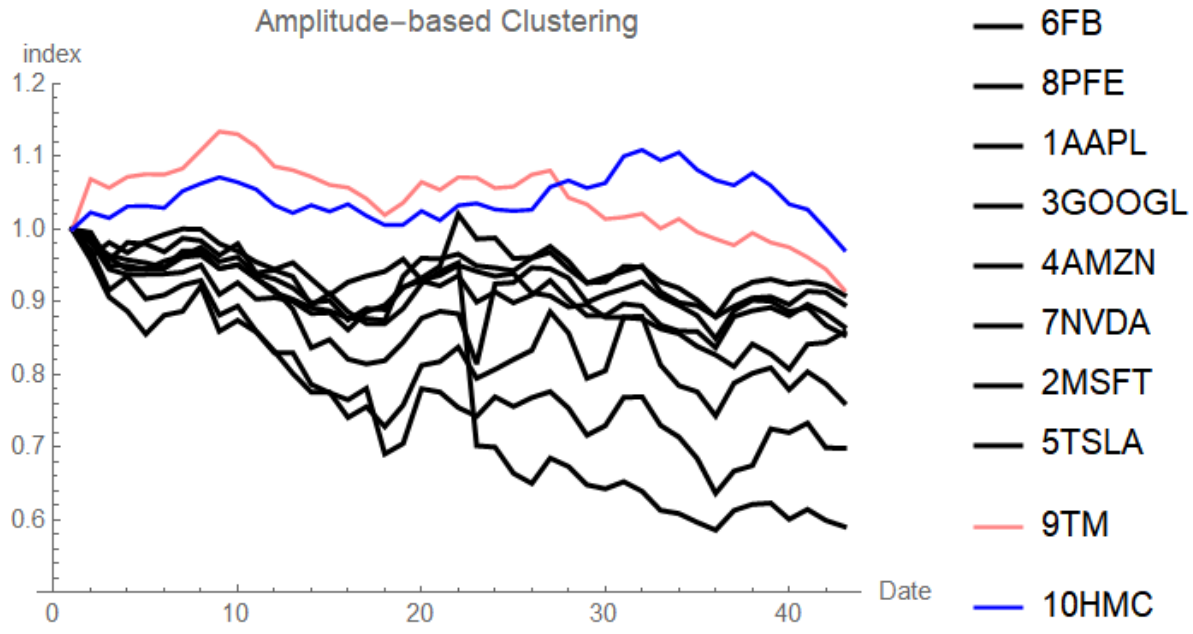
Lack of Ampl. Info.

- Standardization makes a small fluctuation
  that with a large variance
- No amplitude info.

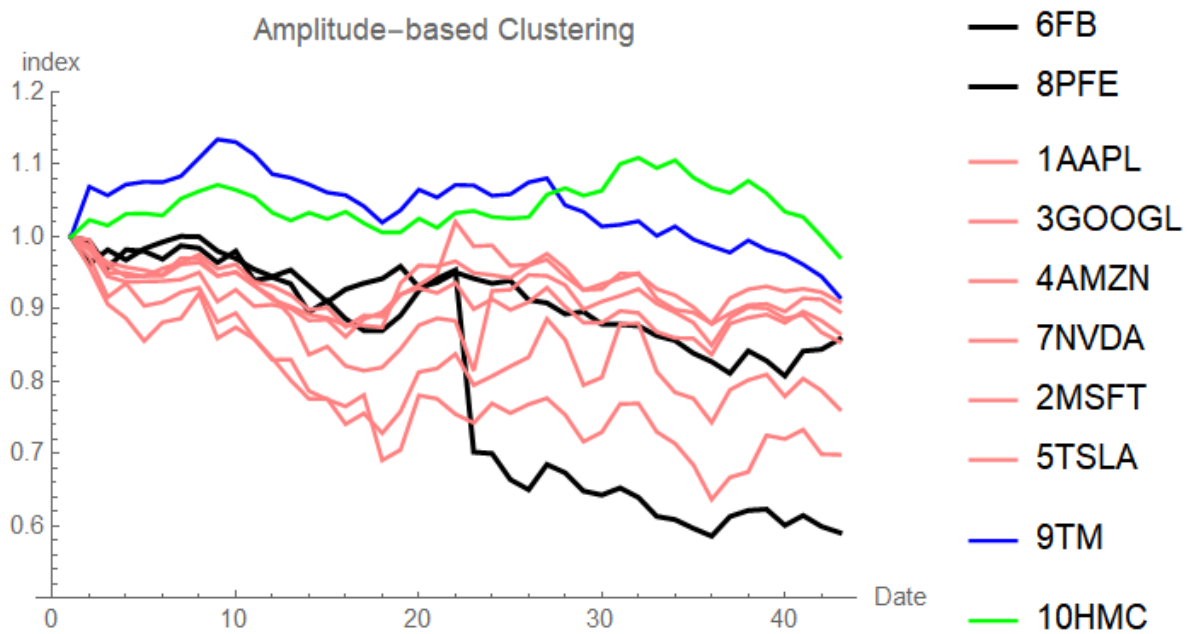


HRP's Clustering
It does not meet our requirements.

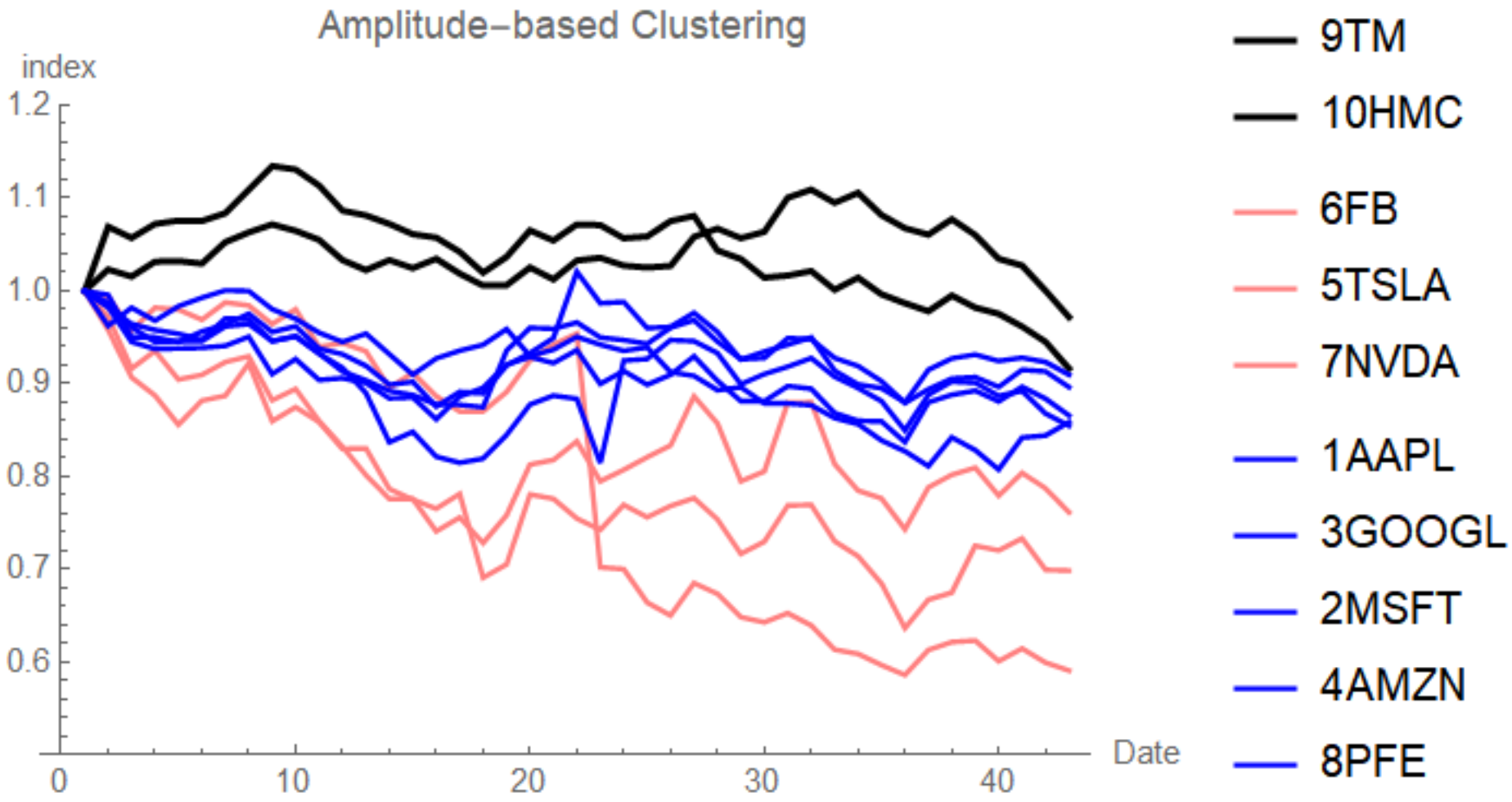
K=3



K=4



Amplitude-based Clustering k=3



Contents of Standardization Problem

1. Standardization Requirement
2. Comparison by Standardized Stock Price Sequences
3. Standardization Problem : Lack of Amplitude Info.
4. Amplitude-based Clustering
   - Algorithm and Distance
   - Advantages
5. Conclusion

Our Requirement

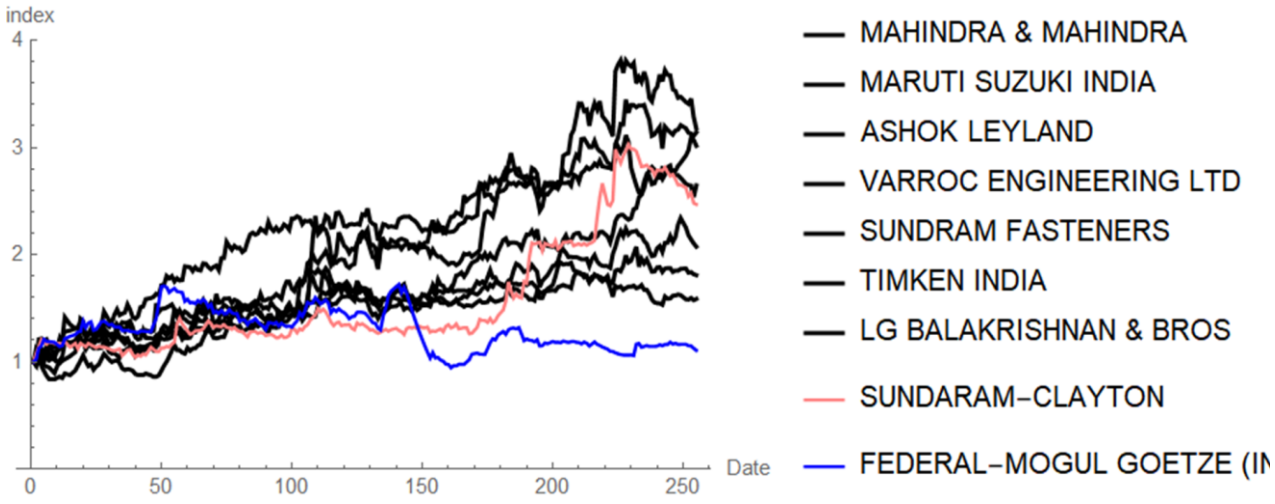
- Focus on growth rates
- Keeping amplitude info., conduct clustering
- How about k-Shape ?
  NO because input data standardization is needed
- k-Shape and DTW
  suppose almost same scaling
- We cannot use the existing clustering methods
- Developed a new one “Amplitude-based clustering”

Amplitude-Based Time Series Data Clustering Method

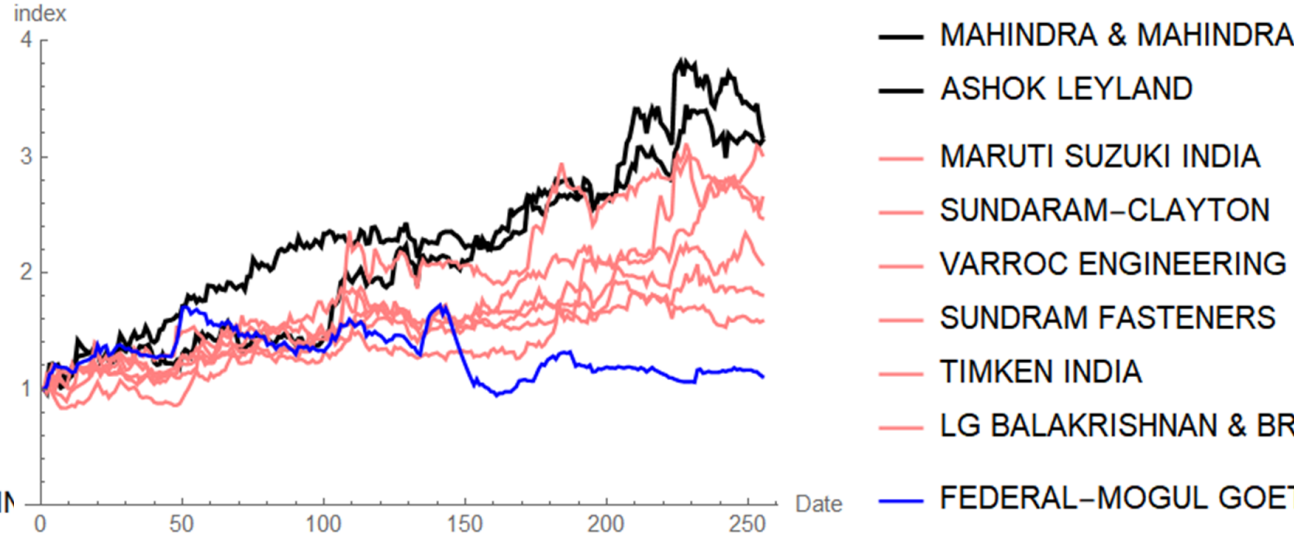
- **Y. Shirota, and B. Chakraborty,** “Amplitude-Based Time Series Data Clustering Method," ***Gakushuin Economics Papers,*** **vol. 59, no. 2, 2022.**

Which is a result we wanted ?

HRP clustering



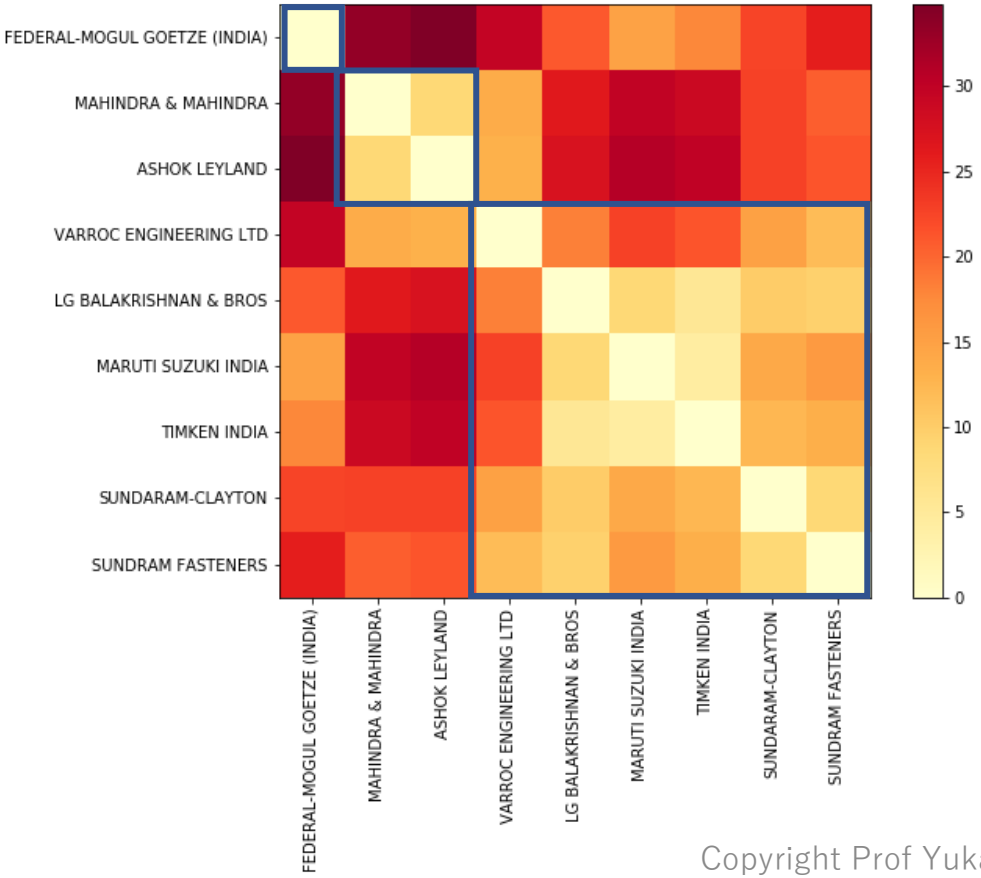
Amplitude-based clustering



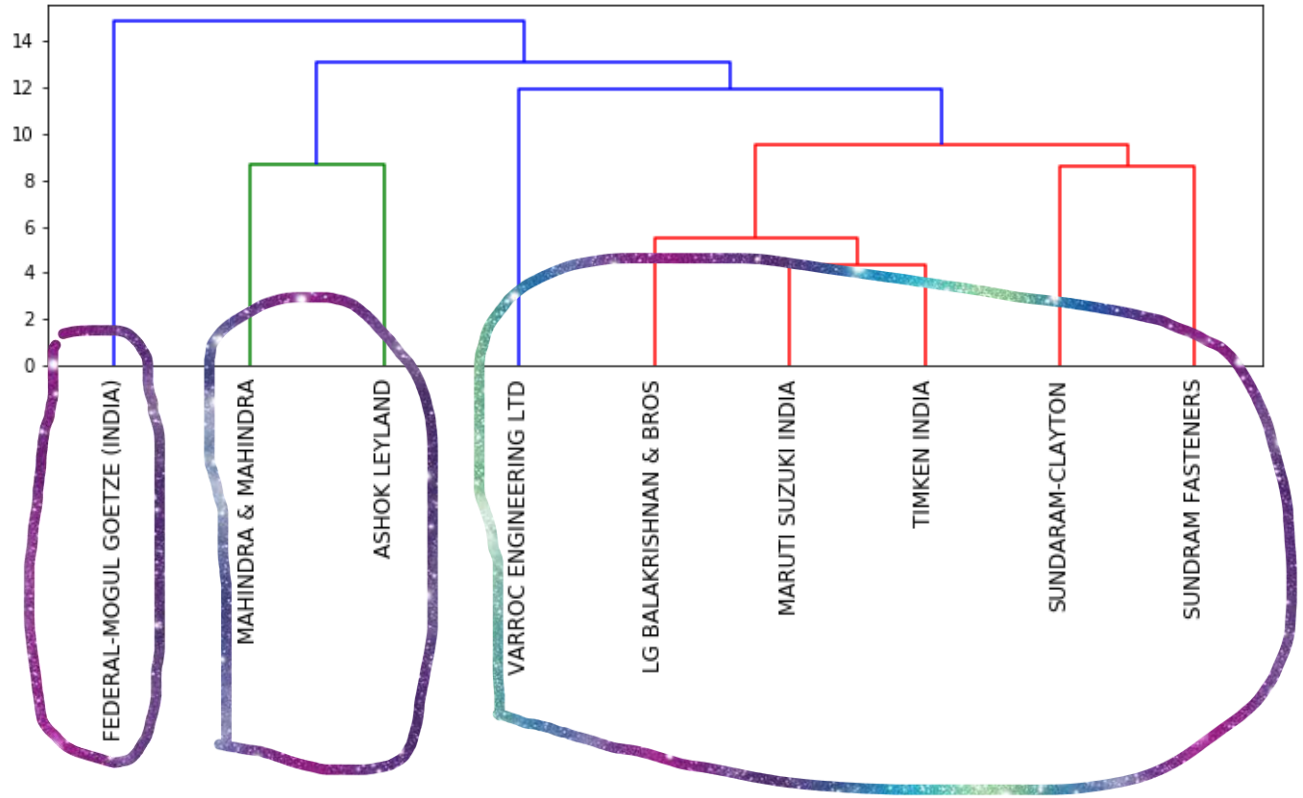
Because it uses standard deviations

Amplitude-based clustering result

Distance matrix



Dendrogram



Contents of Standardization Problem

1. Standardization Requirement
2. Comparison by Standardized Stock Price Sequences
3. Standardization Problem : Lack of Amplitude Info.
4. Amplitude-based Clustering
   - Algorithm and Distance
   - Advantages
5. Conclusion

Amplitude-based clustering Algorithm (1) Hierarchical clustering

Euclidean distance between i, j

- $ED_{i,j} = \sqrt{\sum_{k=1}^{T}\left(G_{i,k} - G_{j,k}\right)^2}$

$G_{j,k}$ is the index data of j-th company on k-th day.

T is the number of days.

Euclidean distance of distance

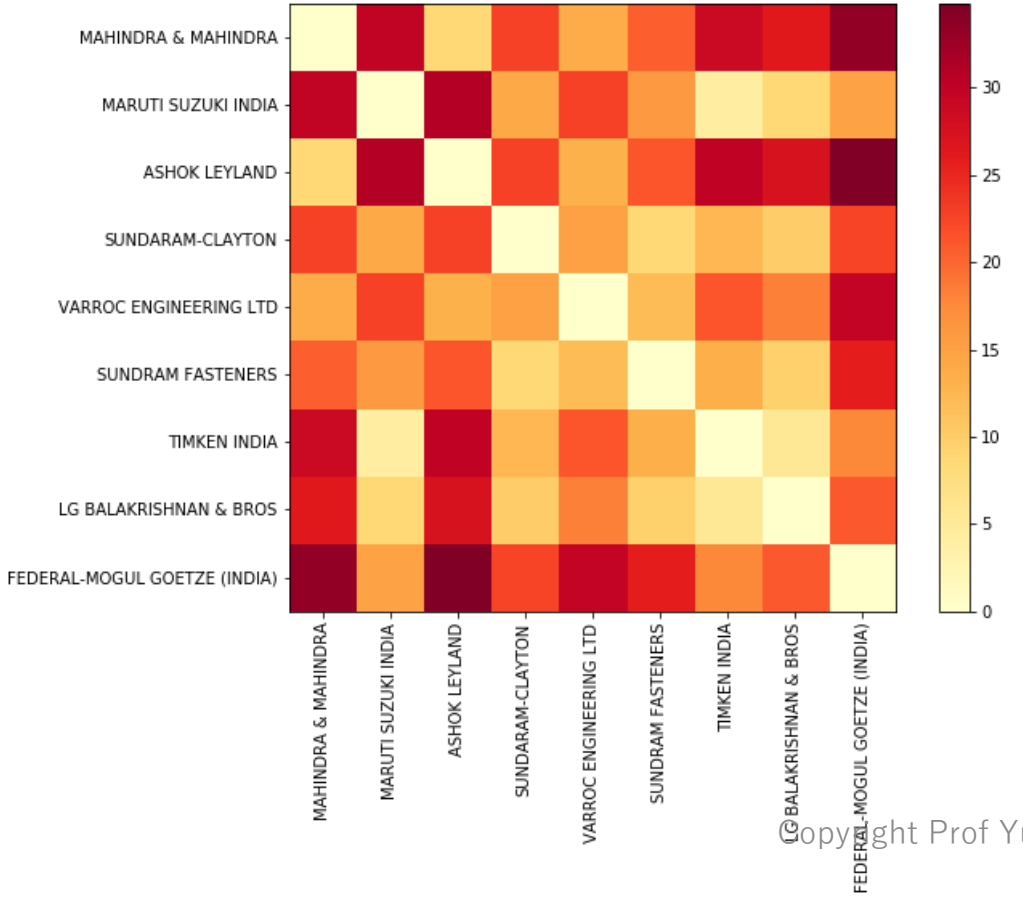
- $\widetilde{ED}_{i,j} = \sqrt{\sum_{n=1}^{N}\left(ED_{n,i} - ED_{n,j}\right)^2}$

N is the number of companies.

Q: How about just using Euclidean distances ?
A: Cannot obtain the good result.
Distance of distance includes the whole company data

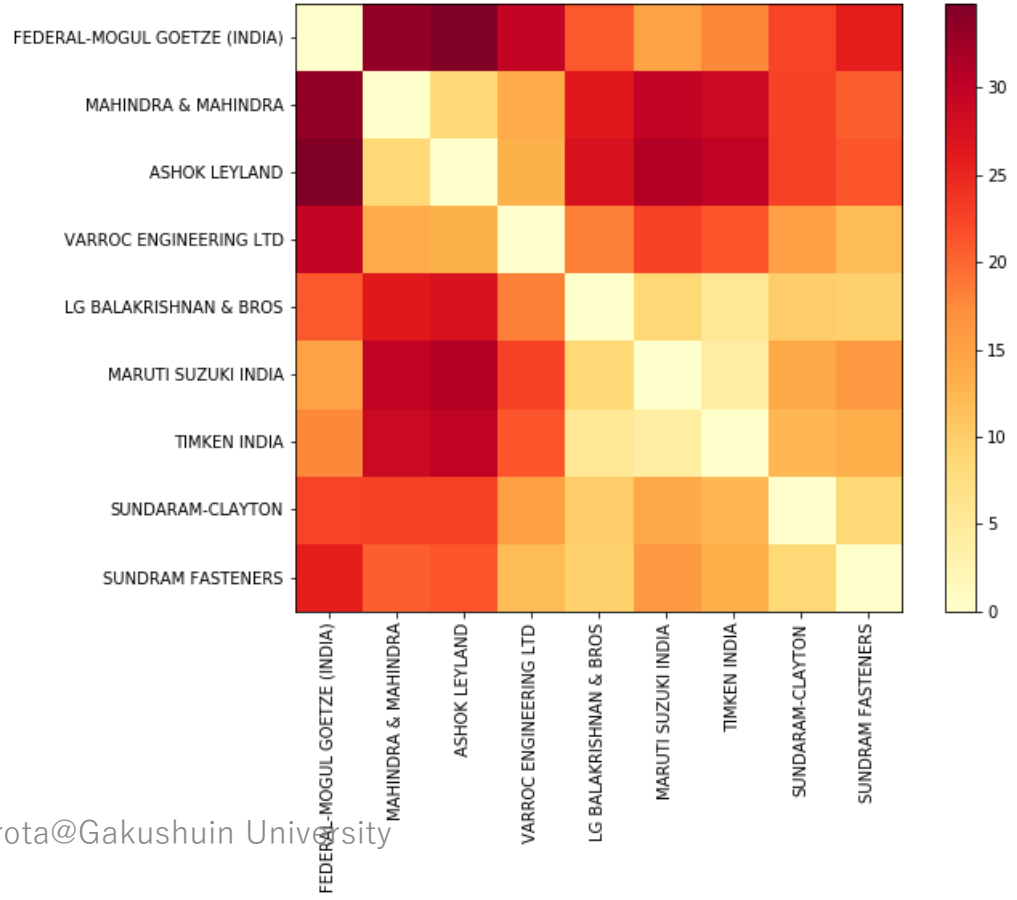
Amplitude-based clustering Algorithm (2) Serialization (Diagonalization)

BEFORE



AFTER

Change the company order for diagonalization

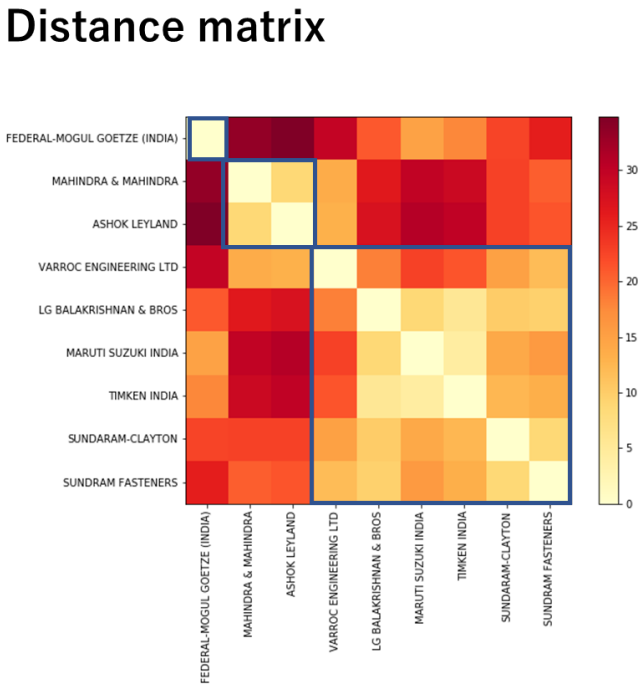


Contents of Standardization Problem

1. Standardization Requirement
2. Comparison by Standardized Stock Price Sequences
3. Standardization Problem : Lack of Amplitude Info.
4. Amplitude-based Clustering
   - Algorithm and Distance
   - Advantages
5. Conclusion

Advantages of Amplitude-based Clustering

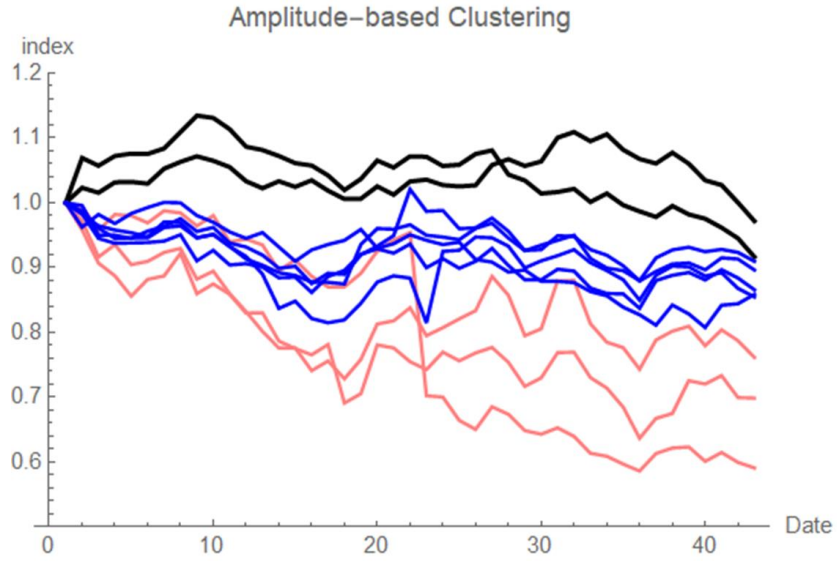
- High performance company makes one cluster
- Other than the extraordinary companies, Interested in the middle large cluster
- In other words, remove the top group and the tail group
- Serialization (quasi-diagonalization) Effect
- We can divide the middle large cluster into several groups.



Contents of Standardization Problem

1. Standardization Requirement
2. Comparison by Standardized Stock Price Sequences
3. Standardization Problem : Lack of Amplitude Info.
4. Amplitude-based Clustering
   - Algorithm and Distance
   - Advantages
5. Conclusion

Summary
Amplitude-based clustering



- Keeping amplitude info., compare the movement patterns
- Application fields: economics/stock growth comparison
- Then cannot use k-Shape because of data standardization
- Very simple algorithm, but helpful in real data analyses.
- SEE ALSO

Y. Shirota, T. Sakura, and B. Chakraborty, "世界自動車製造業2014年度株価成長の時系列分析," ***Gakushuin Economics Papers,*** **vol. 59, no. 2, 2022.**