Hierarchical structure in financial markets

R.N. Mantegna

경상국립대학교 수학과 석사과정 오서영

The paradigm of mathematical finance

The time series of stock returns are unpredictable. (Samuelson, 2016)
-> Random processes

Key point

1. If the random processes of stock returns time series of different stocks are uncorrelated.
2. If economic factors are present in financial markets and are **driving several stocks at the same time.**

Modeling of financial markets

Modeling of financial markets
need to quantify a distance between different stocks traded in a financial markets.
- Hierarchical arrangement of the stocks of a given portfolio is the synchronous **correlation coefficient** of the daily difference of logarithm of closure price of stocks.
-> The correlation coefficient is computed between all the possible pairs of stocks present in the portfolio in a given time period.
→ The goal of the present study is to obtain the taxonomy of a portfolio of stocks traded in a financial market by using the information of time series of stock prices only.

Dataset
Dow Jones Industrial Average (DJIA) index and the portfolio of stocks used to compute the Standard and Poor's 500 (S&P 500) index in the time period from July 1989 to October 1995. Both indices describe the performance of the New York Stock Exchange.

Modeling of financial markets

Quantify the degree of similarity between the synchronous time evolution of a pair of stock price by the correlation coefficient:

$$\rho_{ij} = \frac{\langle Y_i Y_j \rangle - \langle Y_i \rangle \langle Y_j \rangle}{\sqrt{(\langle Y_i^2 \rangle - \langle Y_i \rangle^2)(\langle Y_j^2 \rangle - \langle Y_j \rangle^2)}} \rightarrow$$

n × n matrix of correlation coefficients for daily logarithm price differences

where i and j are the numerical labels of stocks,

$Y_i = \ln P_i(t) - \ln P_i(t-1)$ and $P_i(t)$ is the closure price of the stock i at the day t.

The statistical average <…> is a temporal average performed on all the trading days of the investigated time period

$$-1 \leq \rho_{ij} \leq 1$$

Completely anti-correlated pair of stocks   Completely correlated pair of stocks

$$\rho_{ij} = 0$$

Two stocks are uncorrelated

Modeling of financial markets

The correlation coefficient:

$$\rho_{ij} = \frac{\langle Y_i Y_j \rangle - \langle Y_i \rangle \langle Y_j \rangle}{\sqrt{(\langle Y_i^2 \rangle - \langle Y_i \rangle^2)(\langle Y_j^2 \rangle - \langle Y_j \rangle^2)}}$$

a symmetric matrix with $\rho_{ii} = 1$ in the main diagonal.

→ In each portfolio, $\frac{n(n-1)}{2}$ correlation coefficients characterize the matrix completely.

→ Detecting the hierarchical organization present inside a portfolio of stocks

Modeling of financial markets

The correlation coefficient of a pair of stocks cannot be used as a **distance** between the two stocks.
-> Because it does not fulfill the **three axioms that define a metric**.
However a metric can be defined using as distance a function of the correlation coefficient.

An appropriate function:

$$d(i,j) = \sqrt{2\,(1-\rho_{ij})}. \qquad (2)$$

(i) $d(i,j) = 0 \; if \; and \; only \; if \; i = j$
(ii) $d(i,j) = d(j,i)$
(iii) $d(i,j) \leq d(i,k) + d(k,j)$

(i) $d(i,j) = 0 \; if \; and \; only \; if$ the correlation is total ($\rho = 1$, namely only if the two stocks perform the same stochastic process)

(ii) The correlation coefficient matrix and the distance matrix D is symmetric

(iii) (2) is equivalent to the Euclidean distance between $\widetilde{Y}_i$ and $\widetilde{Y}_j$ which are obtained from the time series Y_i and Y_j by considering each record of the time series.
(The vector obtained must have a unitary norm, namely it is obtained by subtracting to each record the average value and by normalizing it to its standard deviation.)

Minimal spanning tree (MST)

Calculate the correlation coefficient matrix and distance coefficient matrix between nodes on the graph
→ select the most important correlation
→ build a graph that completely connects these nodes with the shortest distance
→ identify the clusters of nodes

Step 1 Select an arbitrary point and connect it to the least dissimilar neighbour. These two points constitute a subgraph of the MST.

Step 2 Connect the current subgraph to the least dissimilar neighbour of any of the members of the subgraph.

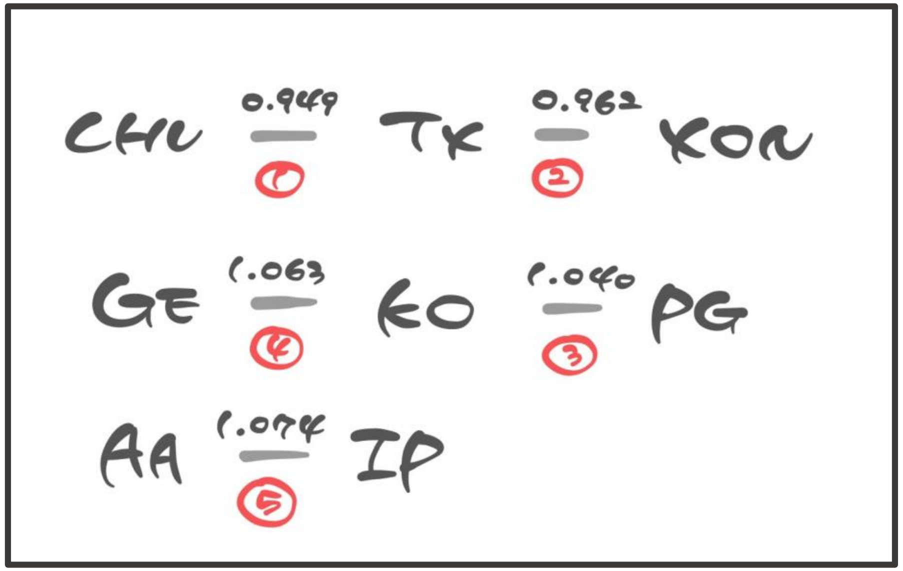
Step 3 Loop on Step 2, until all points are in the one subgraph: this, then, is the MST

The **MST** associated with an **Euclidean distance matrix**

Order the nondiagonal elements of the distance matrix D in increasing order.

Shortest 10 distances:
CHV–TX d = 0.949 | TX–XON d = 0.962 | CHV–XON d = 0.982 | KO–PG d = 1.040 | GE– KO d = 1.063 | AA–IP d = 1.074 | GE–MMM d = 1.078 | KO–MCD d = 1.084 | GE–T d = 1.090 | DD–GE d = 1.095.

-> The MST is progressively built up by linking all the elements of the set together in a graph characterized by a minimal distance between stocks.

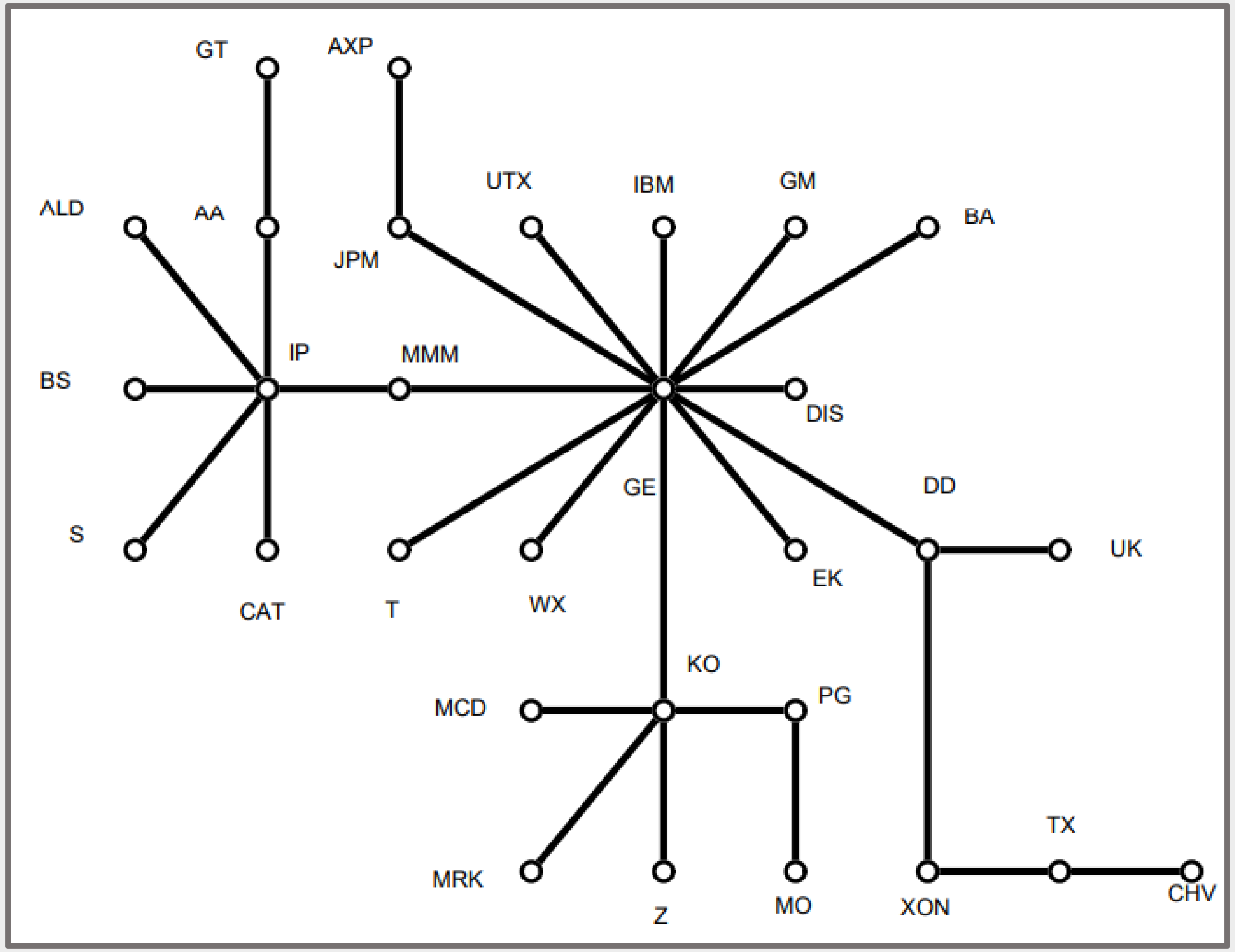


Three distinct groups

(MMM,MCD, T and DD) need to be linked to the GE–KO–PG.
MMM, T and DD → GE (which has at this stage 4 links)
MCD → KO (which has at this stage 3 links)

Different groups (IP-MMM d=1.110) is detected
→ the two groups containing MMM and IP are then linked together through the MMM-IP connection

The **MST** associated with an **Euclidean distance matrix**



Minimal spanning tree connecting the 30 stocks (DJIA)

By following the above illustrated procedure for all then $\frac{n(n-1)}{2}$ distances one eventually obtain the final MST.

The **MST** associated with an **Euclidean distance matrix**

Identifying the market's network with a **Euclidean distance** defined as a function of the correlation coefficient is too **complex** to extract useful information about the correlation coefficient between nodes.
→ The concept of **ultrametric space** is introduced. Among them, **subdominant hyperspace** is widely accepted as suitable for reasons of simplicity. (Onnela et al., 2003)

The complex interrelationship between assets measured by the Euclidean distance.

↓ greatly reduce (simplify complexity)

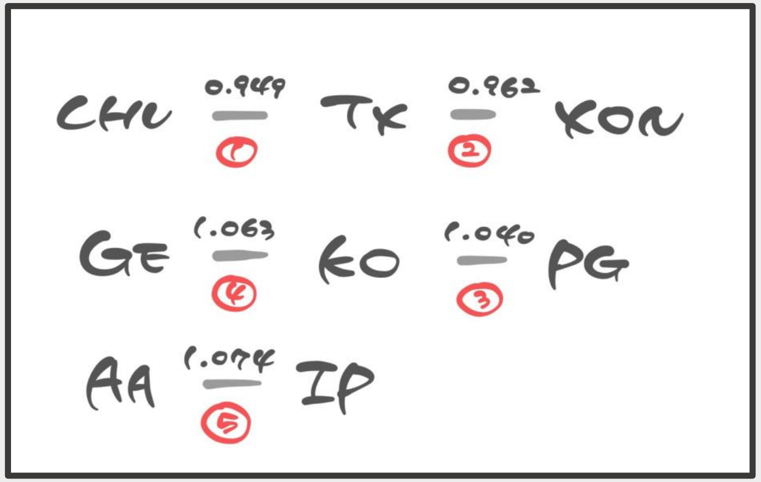
Correlation Structure Information between Assets in Ultrametric Space

The **MST** associated with a **subdominant ultrametric distance matrix**

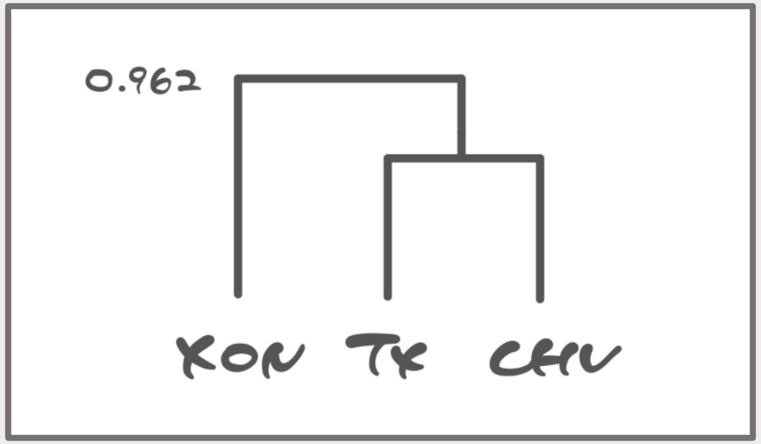
Subdominant ultrametric distance matrix $D^<$ obtained by defining the **subdominant ultrametric distance** $d^<(i,j)$ between i and j as the **maximum** value of any Euclidean distance $d(k,l)$ detected by moving in single steps from i to j through the shortest path connecting i and j in the MST.

Ultrametric distance between XON and CHV is $d^<$ = 0.962
(Euclidean distance between XON and CHV is d = 0.982)

The MST based on hyperspace can compress the information of $\frac{n(n-1)}{2}$ of the correlation coefficient matrix C^t and distance coefficient matrix D^t to (n-1) information.
(Maintain most of the important characteristics of the correlation coefficient distribution). (Onnela et al., 2003)

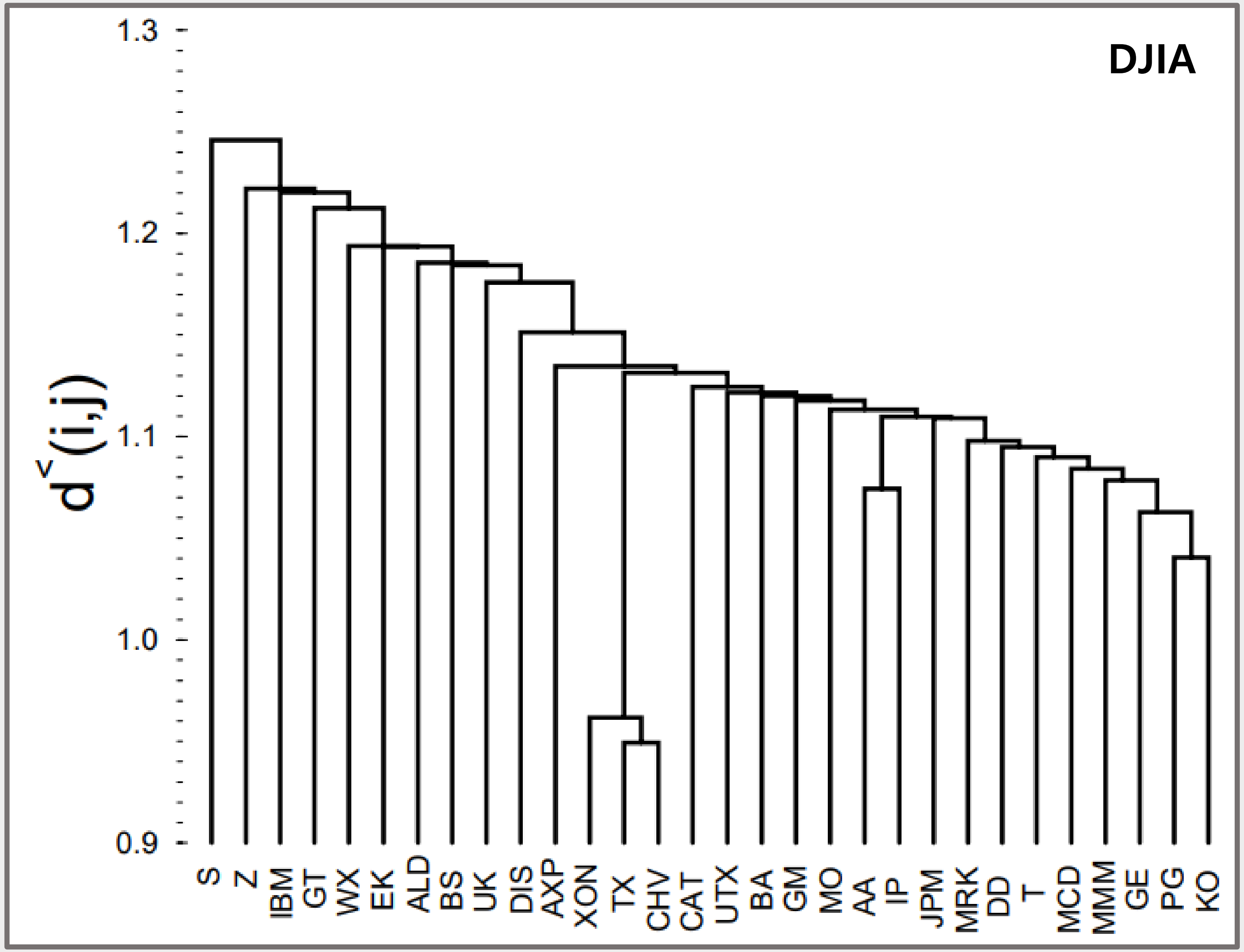


Euclidean



Ultrametric

The **MST** associated with a **subdominant ultrametric distance matrix**



Hierarchical tree of the subdominant ultrametric space associated with the MST

Existence of **three groups of stocks**
→ direct economic explanation

1. **CHV, TX and XON**
are working in the same industry (energy)

2. **AA and IP** provide
raw materials.

3. **PG and KO** deal with
consumer nondurables.

The **MST** associated with a **subdominant ultrametric distance matrix**

The detection of a **hierarchical structure** in a broad portfolio of stocks traded in a financial market is consistent with the assumption that the time series of returns of a stock is **affected by a number of economic factors**.

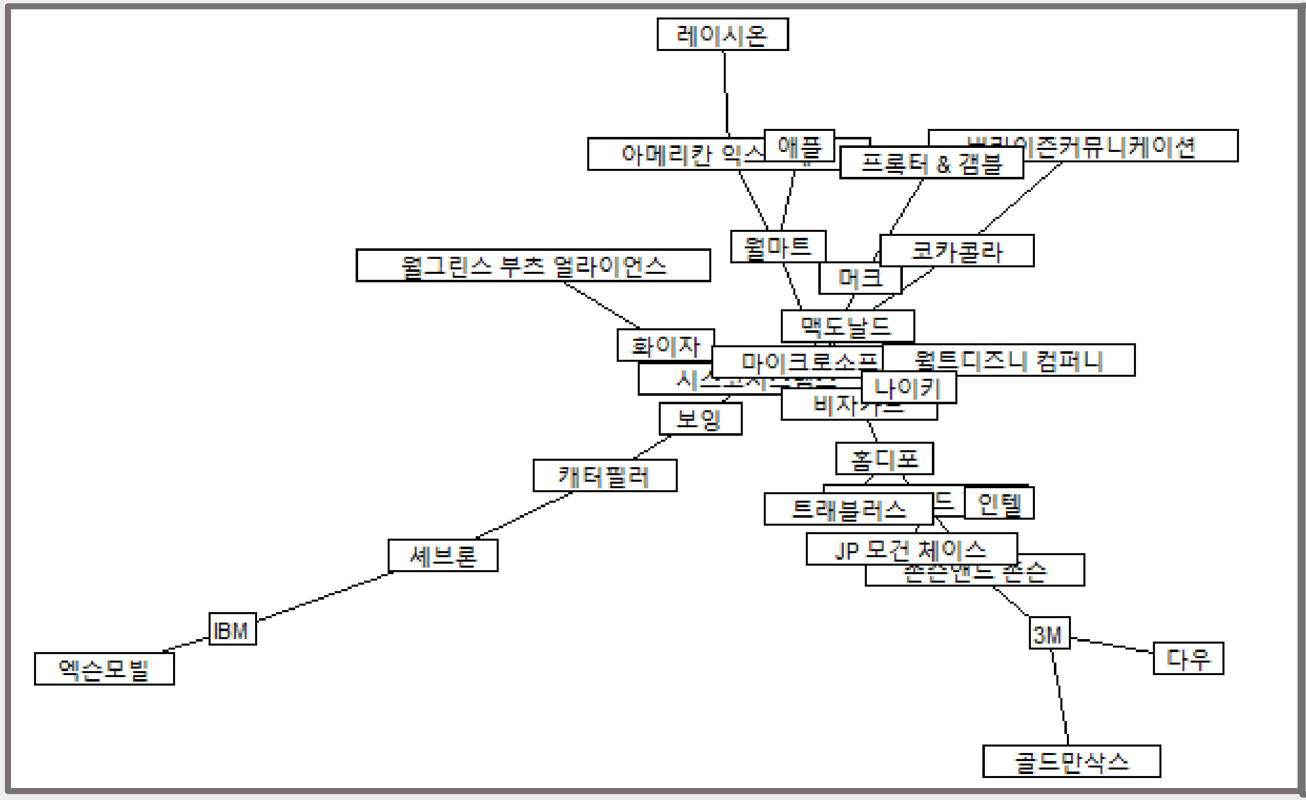
→ The number and the relative influence of these factors is specific to each stock.

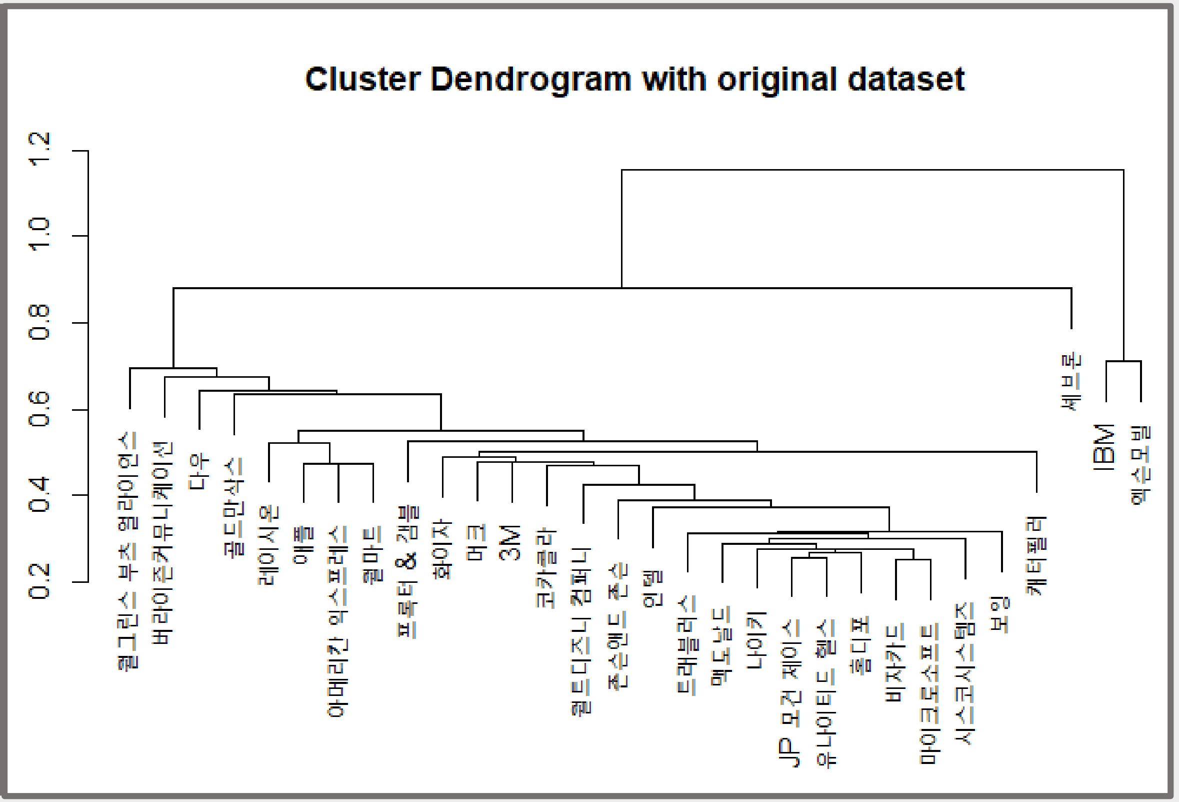
→ In general, stocks or groups of stocks departing early from the tree (at high values of the distance $d^{<}(i,j)$) are mainly controlled by economic factors which are specific to the considered group

→ When departure occurs for (moderately) low values of $d^{<}$, the stocks are affected either by economic factors which are common to all stocks and by other economic factors which are specific to the considered set of stocks. The relative relevance of these factors is quantified by the length of the segment (or segments) observed for each group from one branching to the successive one.

Experiments

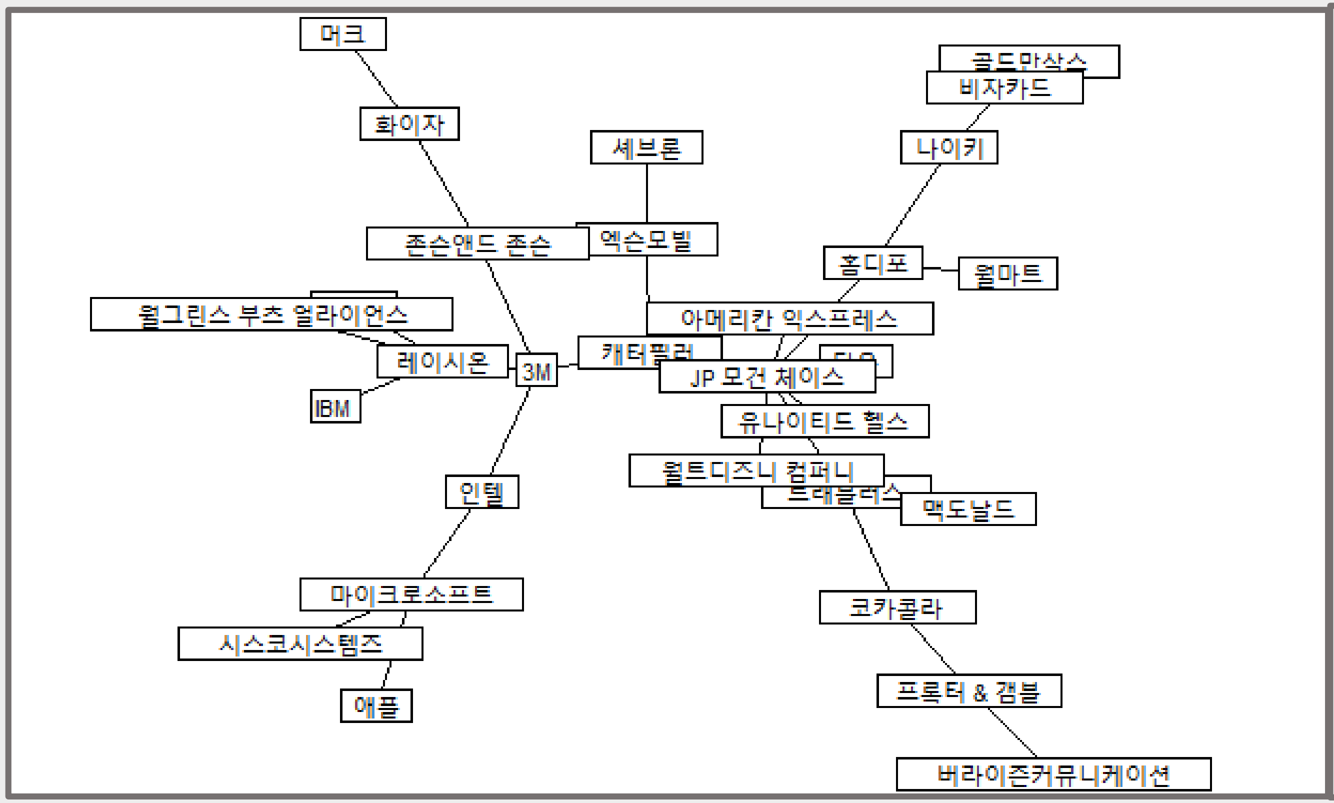
1. Original dataset

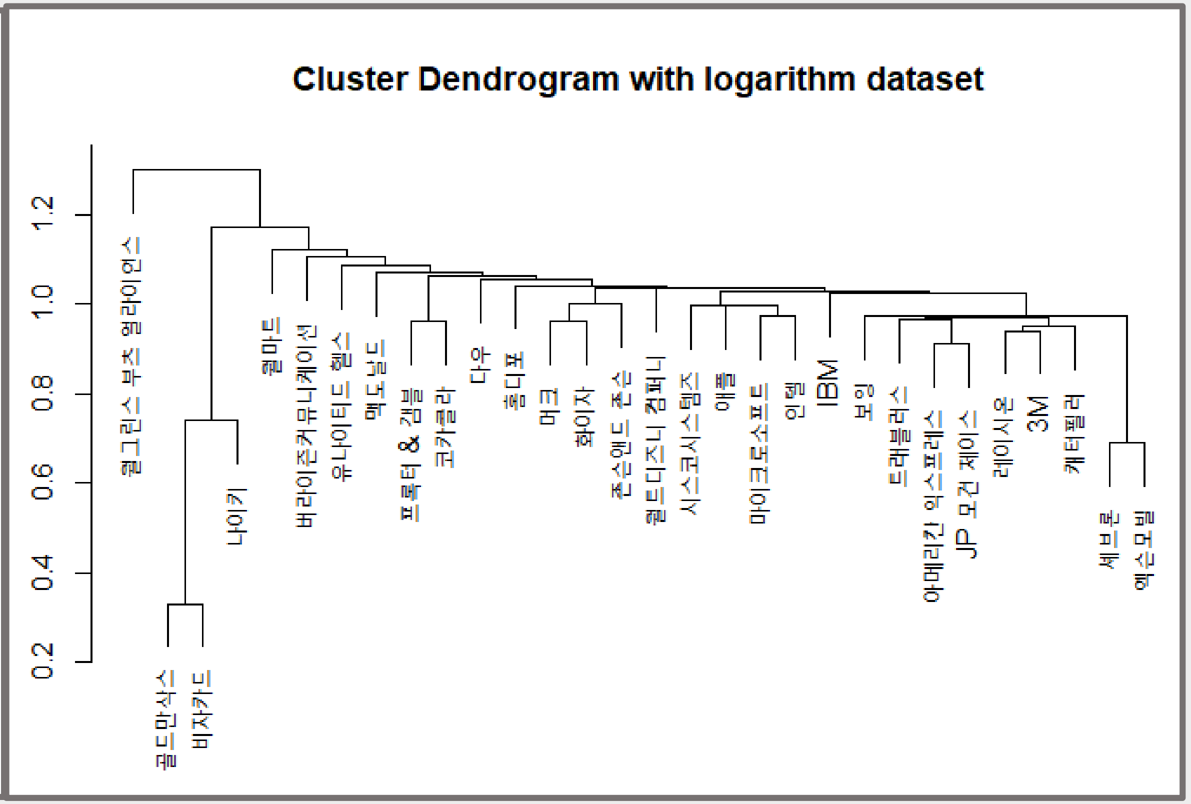




Experiments

2. Logarithm dataset $(Y = ln(P(t)) - ln(P(t-1)))$





Experiments

3. Logarithm dataset ($Y = \frac{P(t)-P(t-1)}{P(t-1)}$)