Multivariate Statistics (I)

5. Cluster Analysis (CA)

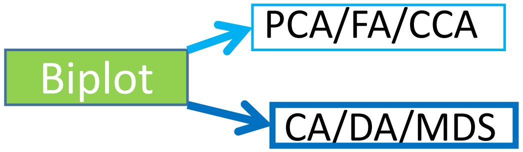
Contents

5.1 Comprehension of CA

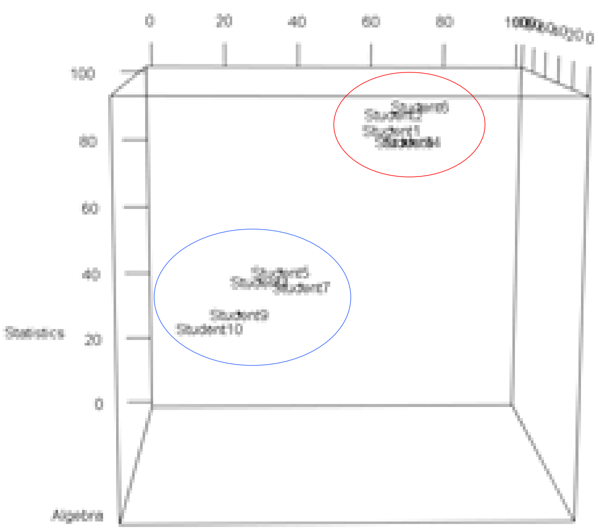
Geometrical Representations of 3-dimensinal space

R-Techniques : Analyses based on the matrix of covariance or correlations between variables.

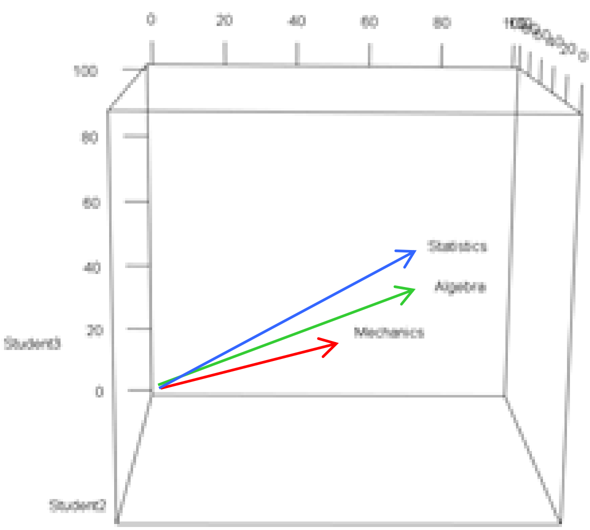
Q-Techniques : Analyses based on the matrix of distances between observations.



Students	Mechanics	Algebra	Statistics
Student1	65	85	85
Student2	65	80	90
Student3	30	40	50
Student4	70	83	82
Student5	35	43	52
Student6	72	82	92
Student7	40	43	48
Student8	68	83	82
Student9	25	32	43
Student10	17	51	35



a) n = 10 points in p–space



b) p = 3 points in n–space

5.1 Comprehension of CA

Methods for CA

Hierarchical Clustering Methods

- single linkage, complete linkage, average linkage, centroid linkage, median linkage, Ward's linkage

Nonhierarchical Clustering Methods

- K-means method, K-median method
- K-medoids method : PAM(partitioning around medoids)

Statistical Model

- EII model , VII model, VEI model, ...

5.2 Association measurements

Association measurement :

Quantitative scale for measuring the proximity or closeness between observations or variables.

similarity

What kind of measurements are used in clustering?

- For clustering observations, the measurements are used by some sort of **distances**. [Table 1.5.1] Euclidean, Standardized Euclidean, Mahalanobis, City-Block Distances

  → Dissimilarity

- For clustering variables, they are usually grouped on the basis of **correlation coefficients** or like **measurement of association of binary data.**

  → Similarity

5.2 Association measurements

Dissimilarity

- D1) Distances bt two observations : [Table 1.5.1]
  - *r th* and *s th* observations: $\boldsymbol{x}_r = (x_{r1}, \dots, x_{rp})'$, $\boldsymbol{x}_s = (x_{s1}, \dots, x_{sp})'$

Basic Conditions

1) $d_{r} > 0,\ r \neq s$.
2) $d_{rr} = 0$.
3) $d_{rs} = d_{sr}$

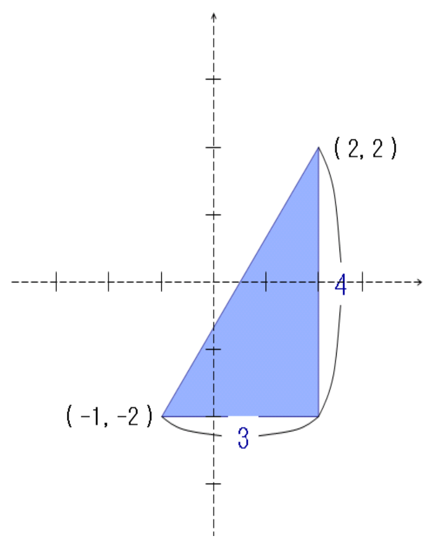
- Euclidean distance $d_{rs} = \left[\sum_{j=1}^{p} \left(x_{rj} - x_{sj} \right)^2 \right]^{1/2}$

$k = 2$

- Minkowski distance $d_{rs} = \left[\sum_{j=1}^{p} \left| x_{rj} - x_{sj} \right|^k \right]^{1/k}, k \geq 1$

$k = 1$

- City-block distance



- Euclidean distance

$= \sqrt{(-1-2)^2 + (-2-2)^2}$

$= \sqrt{3^2 + 4^2}$

$= 5$

- City block distance

$= 3 + 4$

$= 7$

5.2 Association measurements

- D2) Distances bt two groups

Karl Pearson distance	Mahalanobis distance
$d_{hk} = \left[\frac{1}{p} \sum_{j=1}^{p} \frac{(\bar{x}_{hj} - \bar{x}_{kj})^2}{(s_{hj}^2/n_h) + (s_{kj}^2/n_k)} \right]^{1/2}$	$d_{hk} = [(\bar{\boldsymbol{x}}_h - \bar{\boldsymbol{x}}_k)' S^{-1} (\bar{\boldsymbol{x}}_h - \bar{\boldsymbol{x}}_k)]^{1/2}$

- D3) Distance bt two variables : $\boldsymbol{x}_l = (x_{1l}, \ldots, x_{nl})'$, $\boldsymbol{x}_m = (x_{1m}, \ldots, x_{nm})'$

$$\boxed{d_{lm} = 1 - r_{lm} = 1 - \frac{s_{lm}}{s_l s_m}}$$

- D4) Weighted Euclidean distances bt two rows in $F = (f_{rc}),\ r = 1, \ldots, I;\ c = 1, \ldots, J$

$$\boldsymbol{r}_i = (f_{i1}, \ldots, f_{iJ})'/f_{i\cdot}\ ,\quad \boldsymbol{r}_j = (f_{j1}, \ldots, f_{jJ})'/f_{j\cdot}$$

$$\boxed{d_{ij} = \left[\sum_{c=1}^{J} \frac{1}{f_{\cdot c}} \left(\frac{f_{ic}}{f_{i\cdot}} - \frac{f_{jc}}{f_{j\cdot}} \right)^2 \right]^{1/2}}$$

5.2 Association measurements

Similarity

- S1) Similarity of two observations of binary variable
  - Binary data matrix: $X = (x_{rj}),\ r = 1,\ \ldots,\ n; j = 1,\ \ldots,\ p$

$$x_{ij} = \begin{cases} 1\text{: the characteristic is present.} \\ 0\text{: otherwise} \end{cases}$$

[Table 5.2.2] 2 x 2 Association table

		object k		Totals
		1	0	
object i	1	a	b	$a + b$
	0	c	d	$c + d$
Totals		$a + c$	$b + d$	p

5.2 Association measurements

[Table 5.2.3] Similarity coefficients for binary data

- Simple matching : $\frac{a+d}{p}$ Equal weights for 1−1 matches and 0−0 matches
- Double matching : $\frac{2(a+d)}{2(a+d)+b+c}$ Double weights for 1−1 matches and 0−0 matches
- Roser-Tanimoto : $\frac{a+d}{a+d+2(b+c)}$ Double weights for unmatched pairs

- Rusell-Rao : $\frac{a}{p}$ No 0−0 matches in numerator
- Jaccard : $\frac{a}{a+b+c}$ No 0−0 matches in numerator or denominator

Note : For the single linkage and complete linkage , any choice of the coefficients in red (or blue) box will produce the same grouping.

5.2 Association Measurements

- S2) Similarity for two variables in binary data

		Variable m		Total
		1	0	
Variable l	1	n_{11}	n_{12}	$n_{11}+n_{12}$
	0	n_{21}	n_{22}	$n_{21}+n_{22}$
Total		$n_{11}+n_{21}$	$n_{12}+n_{22}$	n

- → Product moment correlation coefficient

$$r = \frac{n_{11}n_{22} - n_{12}n_{21}}{[(n_{11}+n_{12})(n_{21}+n_{22})(n_{11}+n_{21})(n_{12}+n_{22})]^{1/2}}$$

↔ $\chi^2 = nr^2$: Chi-square test statistics

5.2 Association measurements

[Table 1.3.1] Economic Views Data

[Example 5.2.1]

[Table 5.2.1] Binary data for economic views of two institutes

기관	경제 전망									
	1	2	3	4	5	6	7	8	9	10
한국은행	0	1	0	0	0	1	1	1	1	1
동서증권	1	1	0	1	1	0	1	1	1	1

$$d_{rs}^2 = \sum_{j=1}^{10} (x_{rj} - x_{sj})^2 = (0-1)^2 + (1-1)^2 + \cdots + (1-1)^2 = 4$$

[Example 5.2.2] simple matching

[Table 5.2.2] 2 x 2 association table

		동서증권		Total
		1	0	
한국은행	1	5	1	6
	0	3	1	4
Total		8	2	10

$$d_{rs} = \sqrt{p(1-c_{rs})}$$

simple matching coefficient: $c_{rs} = (a+d)/p$ = (5+1)/10 = 0.6

Similarity vs. Dissimilarity

$$c_{rs} = \frac{1}{1+d_{rs}}$$

$$0 < c_{rs} \le 1$$

$$d_{rs} = \sqrt{c_{rr} - 2c_{rs} + c_{ss}} \xrightarrow{c_{rr} = c_{ss} = 1} d_{rs} = \sqrt{2(1-c_{rs})}$$

$$d_{rs} > 0, \ r \neq s$$

5.2 Association measurements

◆ Product moment correlation coefficient

[Table 5.2.6] 2 x 2 association table bt two economic views

		2=GNP		Total
		1	0	
1=성장률	1	6	4	10
	0	1	3	4
Total		7	7	14

$$r = \frac{n_{11}n_{22} - n_{12}n_{21}}{[(n_{11}+n_{12})(n_{21}+n_{22})(n_{11}+n_{21})(n_{12}+n_{22})]^{1/2}} \rightarrow r = \frac{6\times 3 - 4\times 1}{[10\times 4\times 7\times 7]^{1/2}} = \frac{\sqrt{10}}{10} \simeq 0.316$$

$$\chi^2 = n \times r^2 = 14 \times (0.316)^2 \simeq 1.4$$

$$\chi^2 = \left(\frac{6-5}{\sqrt{5}}\right)^2 + \left(\frac{4-5}{\sqrt{5}}\right)^2 + \left(\frac{1-2}{\sqrt{2}}\right)^2 + \left(\frac{3-2}{\sqrt{2}}\right)^2 = 1.4 < 3.84 = \chi_1^2(0.05)$$

Not reject H_0 : Growth rates and GNP are not related at 5%

5.3 Hierarchical clustering methods – single linkage

Nearest Neighbor Method

◆ [Example 5.3.1] Single linkage between 5 individuals and Dendrogram

- [STEP 1]-[STEP 2] Find the distance matrix between 5 individuals.

$$D = (d_{rs}) = \begin{array}{c} 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{array} \begin{pmatrix} 0 & & & & \\ 9 & 0 & & & \\ 3 & 7 & 0 & & \\ 6 & 5 & 9 & 0 & \\ 11 & 10 & \boxed{2} & 8 & 0 \end{pmatrix}$$

- [STEP 3] Find the shortest distance of two clusters, calculate their distance, and merge the two clusters.

$\min_{i,k}(d_{ik}) = d_{53} = 2$: new cluster (35)

- [STEP 4] New distance matrix

$$d_{(35)1} = \min(d_{31}, d_{51}) = \min(3, 11) = 3$$
$$d_{(35)2} = \min(d_{32}, d_{52}) = \min(7, 10) = 7$$
$$d_{(35)4} = \min(d_{34}, d_{54}) = \min(9, 8) = 8$$

$$\boxed{d_{(UV)W} = \min(d_{UW}, d_{VW})}$$

$$D_{(35)} = \begin{array}{c} (35) \\ 1 \\ 2 \\ 4 \end{array} \begin{pmatrix} 0 & & & \\ \boxed{3} & 0 & & \\ 7 & 9 & 0 & \\ 8 & 6 & 5 & 0 \end{pmatrix} \rightarrow \min_{i,k}(d_{ik}) = d_{(35)1} = 3$$

: new cluster (135)

[STEP 5] Repeat [STEP 3] and [STEP 4]

$$d_{(135)2} = \min(d_{(35)2}, d_{12}) = \min(7, 9) = 7$$
$$d_{(135)4} = \min(d_{(35)4}, d_{14}) = \min(8, 6) = 6$$

$$D_{(135)} = \begin{array}{c} (135) \\ 2 \\ 4 \end{array} \begin{pmatrix} 0 & & \\ 7 & 0 & \\ 6 & \boxed{5} & 0 \end{pmatrix}$$

$$d_{(135)(24)} = \min(d_{(135)2}, d_{(135)4}) = \min(7, 6) = 6$$

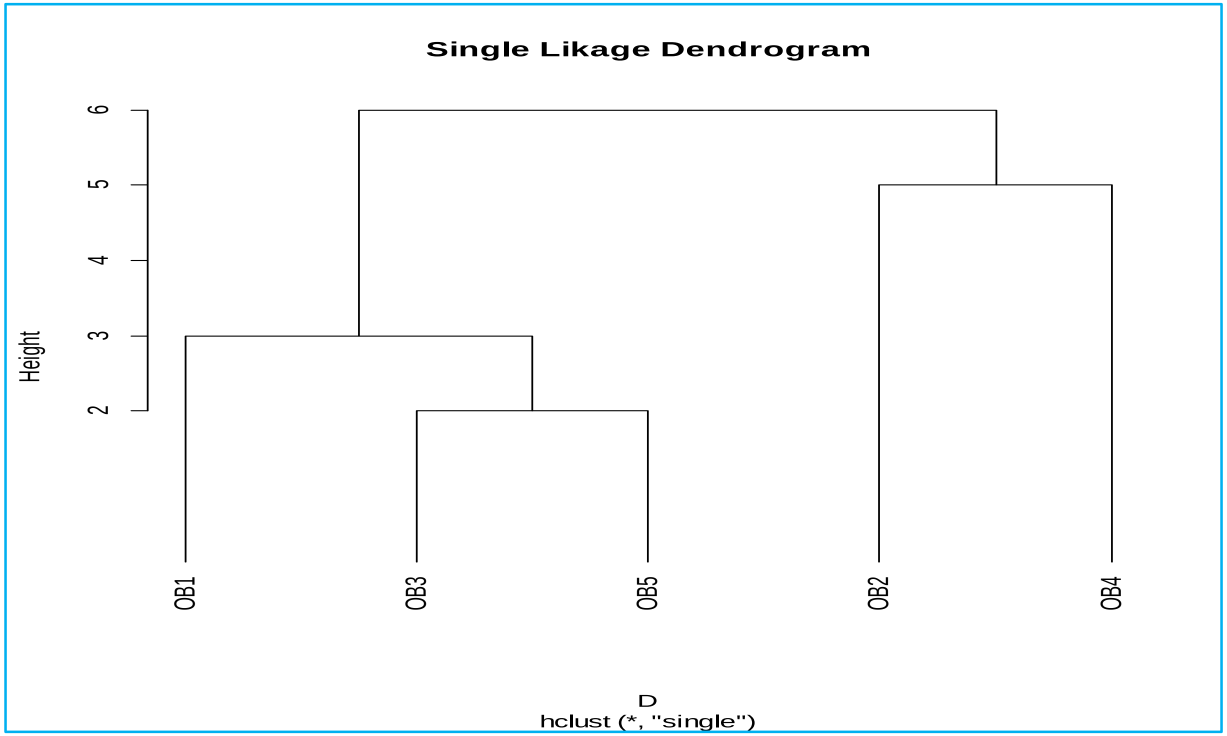
$$D_{(135,\ 24)} = \begin{array}{c} (135) \\ (24) \end{array} \begin{pmatrix} 0 & \\ 6 & 0 \end{pmatrix}$$

(135), (2,4) ⟶ (13524)

5.3 Hierarchical clustering methods – single linkage

- [STEP 6] Visually look at the merging of the clusters through the Dendrogram.

OB1	OB2	OB3	OB4	OB5
0	9	3	6	11
9	0	7	5	10
3	7	0	9	2
6	5	9	0	8
11	10	2	8	0



[R-code 5.3.1] single linkage and dendrogram(5obsdisit-CAsingle.R)

```
# Hierachical Cluster Analysis
D<-as.dist(read.table("5obsdist.txt", header=T))
single=hclust(D, method="single") # Single Linkage
plot(single, hang=-1, main="Single Likage Dendrogram")
```

5.3 Hierarchical clustering methods – complete linkage

Farthest Neighbor Method

[Example 5.3.2] Complete linkage between 5 individuals and Dendrogram

- [STEP 1]-[STEP 2] Find the distance matrix between 5 individuals.

$$D = (d_{rs}) = \begin{array}{c} 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{array} \begin{pmatrix} 0 & & & & \\ 9 & 0 & & & \\ 3 & 7 & 0 & & \\ 6 & 5 & 9 & 0 & \\ 11 & 10 & \boxed{2} & 8 & 0 \end{pmatrix}$$

- [STEP 3] Find the shortest distance of two clusters, calculate their distance, and merge the two clusters.

$\boxed{\min_{i,k}(d_{ik}) = d_{53} = 2}$: new cluster (35)

- [STEP 4] New distance matrix

$$d_{(35)1} = \max(d_{31}, d_{51}) = \max(3, 11) = 11$$
$$d_{(35)2} = \max(d_{32}, d_{52}) = \max(7, 10) = 10$$
$$d_{(35)4} = \max(d_{34}, d_{54}) = \max(9, 8) = 9$$

$$\boxed{d_{(UV)W} = \max(d_{UW},\ d_{VW})}$$

$$D_{(35)} = \begin{array}{c} (35) \\ 1 \\ 2 \\ 4 \end{array} \begin{pmatrix} 0 & & & \\ 11 & 0 & & \\ 10 & 9 & 0 & \\ 9 & 6 & \boxed{5} & 0 \end{pmatrix}.$$

$\min_{i,k}(d_{ik}) = d_{42} = 5$

: new cluster (24)

[STEP 5] Repeat [STEP 3] and [STEP 4]

$$d_{(24)(35)} = \max(d_{2(35)}, d_{4(35)}) = \max(10, 9) = 10$$
$$d_{(24)1} = \max(d_{21}, d_{41}) = \max(9, 6) = 9$$

$$D_{(35,24)} = \begin{array}{c} (35) \\ (24) \\ 1 \end{array} \begin{pmatrix} 0 & & \\ 10 & 0 & \\ 11 & \boxed{9} & 0 \end{pmatrix}$$

$\min_{i,k}(d_{ik}) = d_{(24)1} = 9$: new cluster (124)

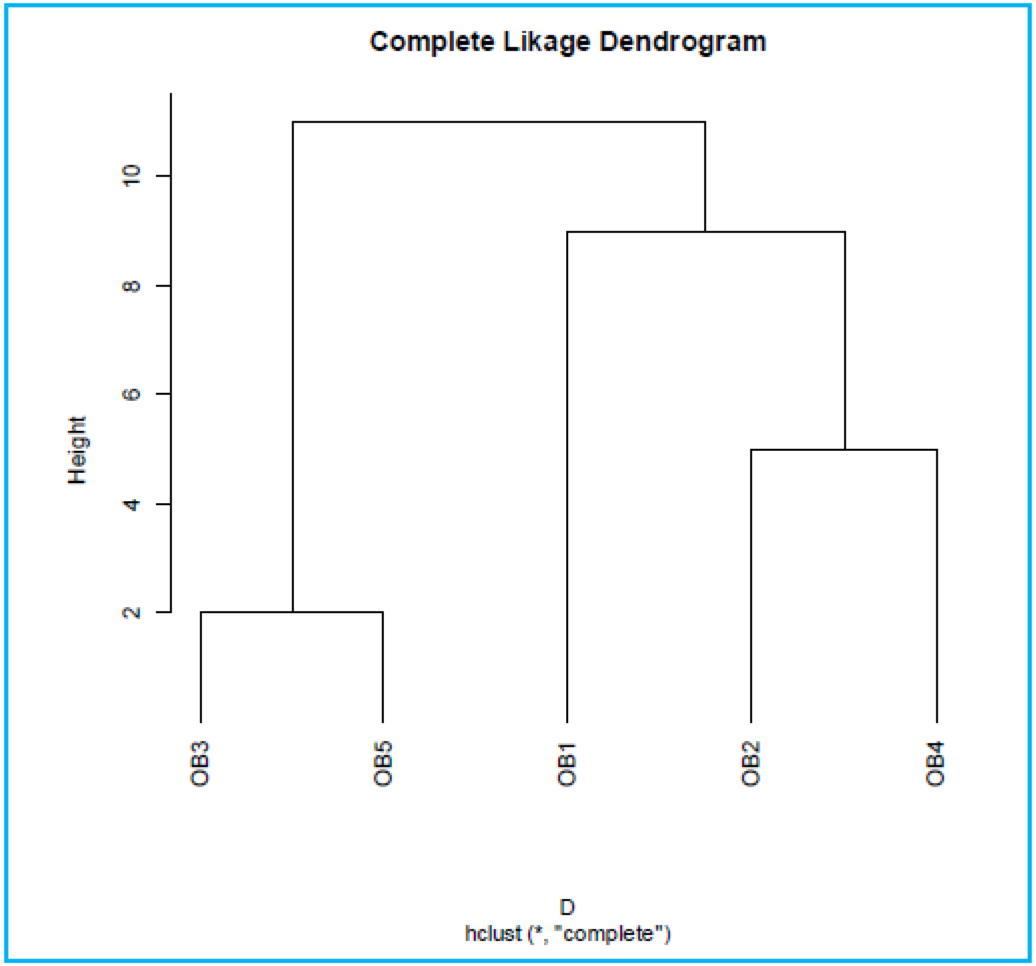
$$d_{(124)(35)} = \max(d_{1(35)}, d_{(24)(35)}) = \max(11, 10) = 11$$

➡ (124), (3.5) $\longrightarrow$ (12435)

5.3 Hierarchical clustering methods - complete linkage

- [STEP 6] Visually look at the merging of the clusters through the Dendrogram.

method="complete"



5.3 Hierarchical clustering methods – average linkage

$$D=(d_{rs})=\begin{array}{c}1\\2\\3\\4\\5\end{array}\begin{pmatrix}0&&&&\\9&0&&&\\3&7&0&&\\6&5&9&0&\\11&10&2&8&0\end{pmatrix}$$

$\min_{i,k}(d_{ik})=d_{53}=2$: new cluster (35)

$$d_{(35)1}=ave(d_{31},\,d_{51})=ave(3,\,11)=\frac{1}{2\times 1}(3+11)=7$$

$$d_{(35)2}=ave(d_{32},\,d_{52})=ave(7,\,10)=\frac{1}{2\times 1}(7+10)=8.5$$

$$d_{(35)4}=ave(d_{34},\,d_{54})=ave(9,\,8)=\frac{1}{2\times 1}(9+8)=8.5$$

$$D_{(35)}=\begin{array}{c}(35)\\1\\2\\4\end{array}\begin{pmatrix}0&&&\\7&0&&\\8.5&9&0&\\8.5&6&5&0\end{pmatrix}$$

$\min_{i,k}(d_{ik})=d_{(24)}=5$: new cluster (24)

$$d_{(24)(35)}=ave(d_{2(35)},\,d_{4(35)})=ave(d_{23}+d_{25}+d_{43}+d_{45})$$
$$=\frac{1}{2\times 2}(7+10+9+8)=8.5$$

$$d_{(24)1}=ave(d_{21},\,d_{41})=\frac{1}{2\times 1}(9+6)=7.5$$

$$D_{(35,\,24)}=\begin{array}{c}(35)\\(24)\\1\end{array}\begin{pmatrix}0&&\\8.5&0&\\7&7.5&0\end{pmatrix}$$

$\min_{i,k}(d_{ik})=d_{(35)1}=7$: new cluster (135)

$$d_{(135)(24)}=ave(d_{(135)2},\,d_{(135)4})=ave(d_{12}+d_{32}+d_{52}+d_{14}+d_{34}+d_{54})$$
$$=\frac{1}{3\times 2}(9+7+10+6+9+8)$$
$$=8.17$$

$$D_{(135,\,24)}=\begin{array}{c}(135)\\(24)\end{array}\begin{pmatrix}0&\\8.17&0\end{pmatrix}$$

(135), (2,4) $\longrightarrow$ (13524)

$$d_{(UV)W}=ave(d_{UW},\,d_{VW})=\frac{\sum_i\sum_k d_{ik}}{n_{(UV)}n_W}$$

method="average"

5.3 Hierarchical clustering methods –Ward linkage

$n_k \times p$ Data matrix of kth cluster C_k : $X_k = (x_{kij}),\ k=1, ..., g,\ i=1, ..., n_k,\ j=1, ..., p$

$$\boldsymbol{x}_{ki} = (x_{ki1}, ..., x_{kip})^t \longrightarrow \overline{\boldsymbol{x}}_k = (\overline{x}_{k1}, ..., \overline{x}_{kp})^t$$

Error sum of squares : $\boxed{ESS_k = \sum_{i=1}^{n_k} ||\boldsymbol{x}_{ki} - \overline{\boldsymbol{x}}_k||^2 = \sum_{i=1}^{n_k}\sum_{j=1}^{p} (x_{kij} - \overline{x}_{kj})^2}$

Min $ESS = \sum_{k=1}^{g} ESS_k$

method="ward"

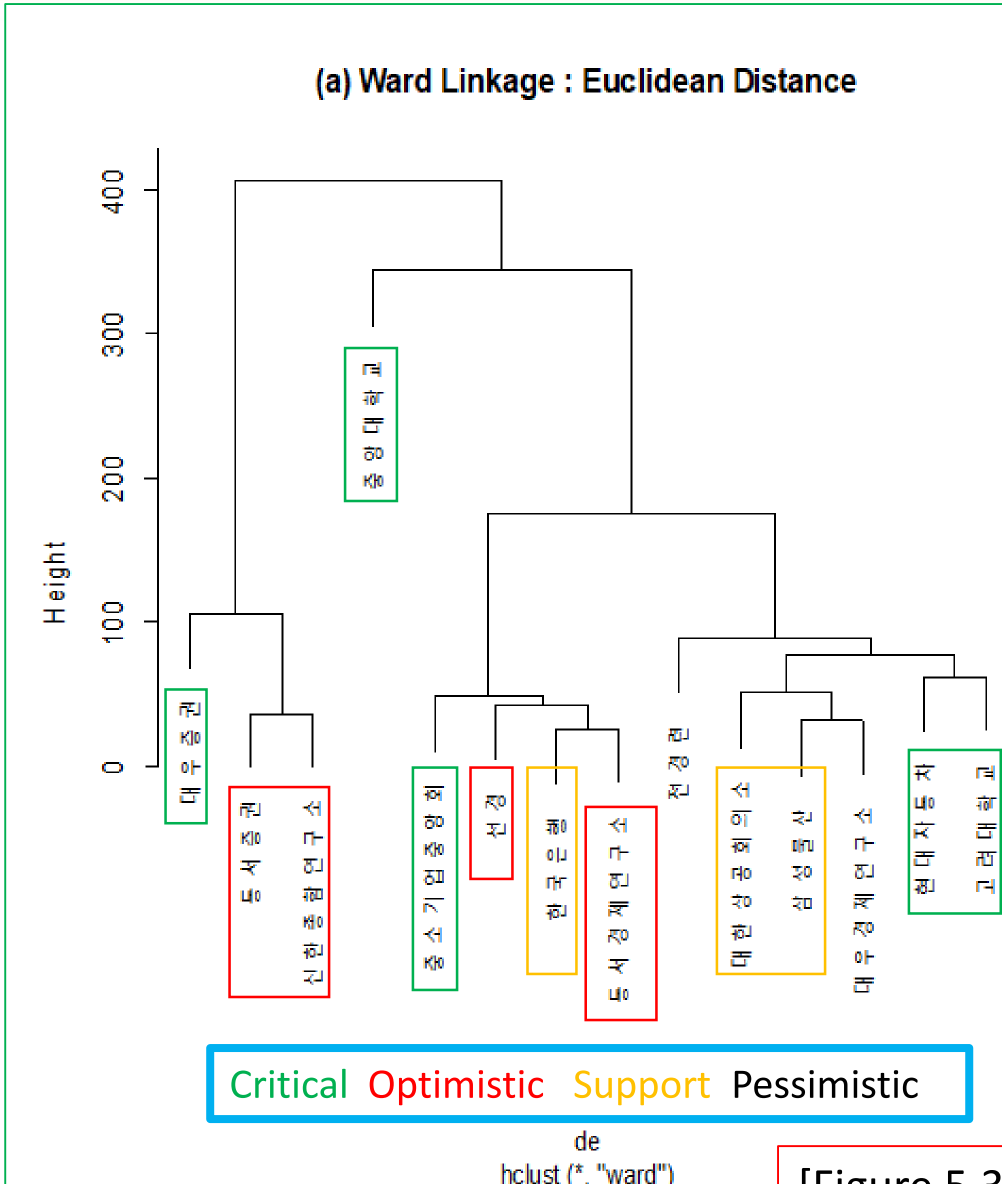
[Example 5.3.4] Ward linkage of institutes in economic views [Table 1.3.5] ⟵ [R-code 5.3.2]

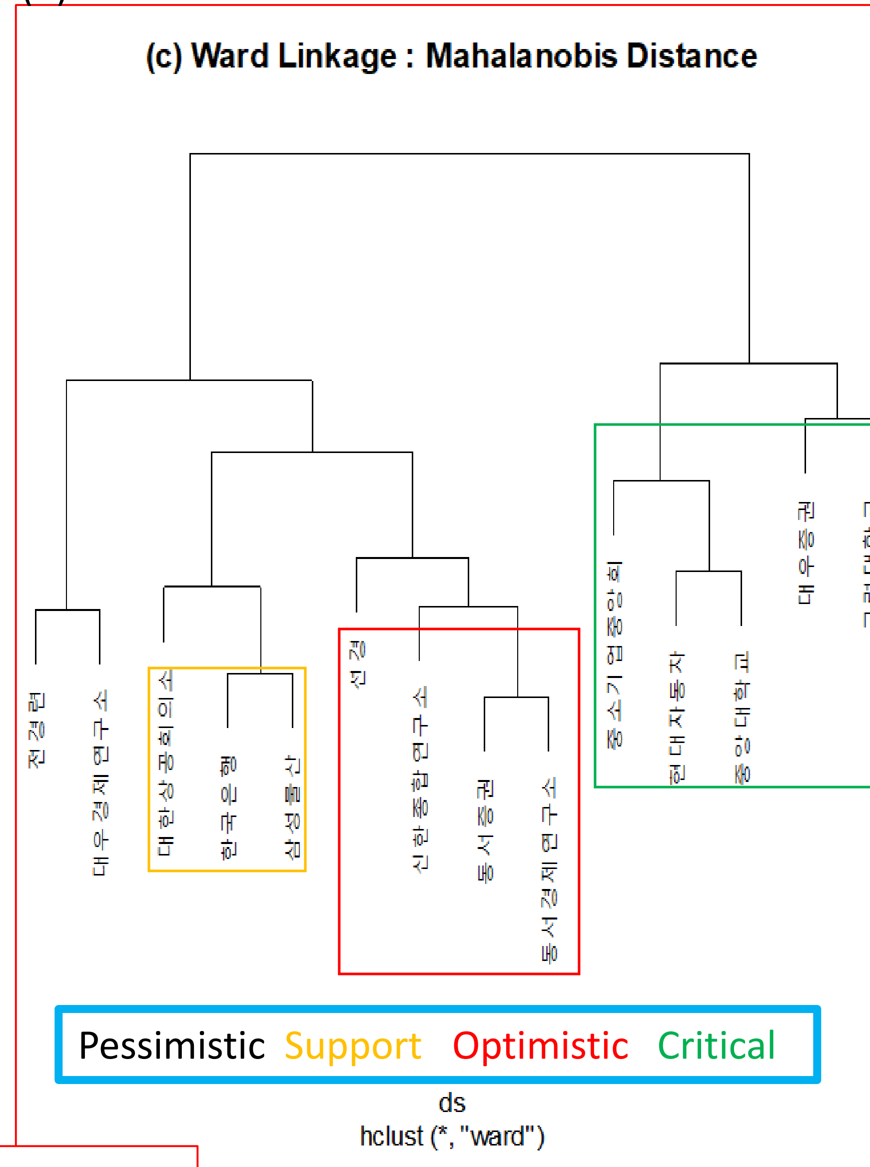
기관	성장률	GNP	수출	수입	국제 흑자	연말 외채	연말 환율	실업률	소비 물가	임금 상승
한국은행	8.2	1950	698	650	98	280	630	3.0	5.7	12.0
대우증권	9.5	2100	710	620	110	270	640	2.7	4.5	12.0
동서증권	9.0	2000	690	630	100	290	630	2.6	6.0	12.0
전경련	7.8	1850	668	660	88	280	620	3.0	6.4	13.8
대한상공회의소	8.5	1928	710	670	90	290	620	3.0	6.0	10.0
중소기업중앙회	9.0	1958	710	615	95	280	603	4.0	6.0	10.0
현대자동차	8.5	1900	700	610	100	250	620	3.2	5.5	14.0
삼성물산	8.0	1900	700	640	100	280	640	2.7	5.5	10.0
선경	8.5	1950	700	620	120	300	630	2.7	7.0	12.0
대우경제연구소	7.9	1900	697	645	95	280	610	2.9	6.8	15.0
신한종합연구소	9.0	2030	700	620	100	275	630	3.0	7.0	13.0
동서경제연구소	8.5	1950	690	630	90	290	630	2.9	6.0	12.0
고려대학교	9.9	1870	729	649	123	260	616	3.4	6.1	15.8
중앙대학교	8.5	1700	700	630	90	260	620	4.0	6.0	14.0

Growth /Export/Import/ Black-ink balance/ Foreign debt /Exchange rate/ Unemployed rate /Consumer price rate /Wage increase rate

5.3 Hierarchical clustering methods –Ward linkage

(b) Standardized Euclidean Distance from Z





[Figure 5.3.3] (a), (c)

5.3 Hierarchical clustering methods –Ward linkage

[Data 1.3.2]

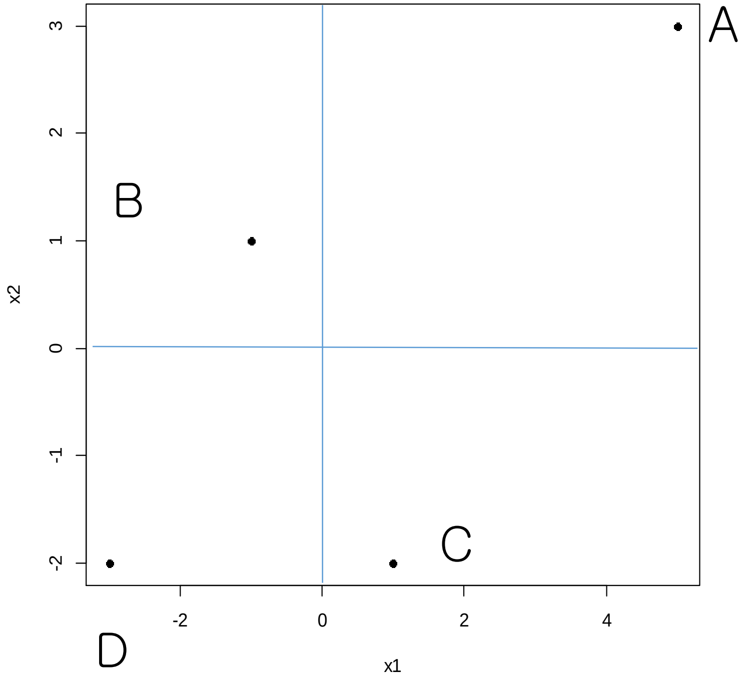
[Example 5.3.5] Hierarchical CA for KLPGA Data from [Figure 5.3.4] : [Table 5.3.3] ⟵ [R-code 5.3.3]

Linkage Method	Group C_1	Group C_2	Group C_3
Single	1, 2, 3, 4, 5, 6, 7, 10	10위권 - 40위권	
Complete	1, 2, 3, 4, 6	10위권 - 20위권 30, 46	30위권-40위권 11, 12, 14, 17, 18, 19, 20, 22, 23, 24, 27
Average	1, 2, 3, 4, 6, 7, 8, 9, 10, 12, 16,	10위권 - 20위권 40, 41, 48, 50	30위권-40위권 21, 23, 24, 55
Ward	1, 2, 3, 4, 6,	10위권 - 20위권 40, 41, 46, 48, 50	30위권-40위권 23
Characteristics	Top ranked players	Middle ranked players	Lower ranked players

5.4 Non-hierarchical clustering methods

◆ [Example 5.4.1] K-means method

Observation	variable X_1	X_2
A	5	3
B	-1	1
C	1	-2
D	-3	-2



[STEP 1] Divide n observations into k initial clusters. (AB), (CD)

[STEP 2] The centroid of each cluster is calculated as the average of the observations belonging to the cluster.

Cluster	$\bar{x}_1$	$\bar{x}_2$
(AB)	$\frac{5+(-1)}{2}=2$	$\frac{3+1}{2}=2$
(CD)	$\frac{1+(-3)}{2}=-1$	$\frac{-2+(-2)}{2}=-2$

Coordinates of Centroid :

$$C_{(AB)}=(2,2),\ \ C_{(CD)}=(-1,-2)$$

5.4 Non-hierarchical clustering methods

◆ [Example 5.4.1] k-means method

[STEP 3] From the Euclidean distance bt centroids and observations, assign the entity to the cluster with the closest centroid.

$$d^2(A, C_{(AB)}) = (5-2)^2 + (3-2)^2 = 10$$

$$d^2(A, C_{(CD)}) = (5+1)^2 + (3+2)^2 = 61$$

→ A → (A)

$$d^2(B, C_{(AB)}) = (-1-2)^2 + (1-2)^2 = 10$$

$$d^2(B, C_{(CD)}) = (-1+1)^2 + (1+2)^2 = 9$$

→ B → (CD) → (BCD)

[STEP 4] Repeat [STEP 2] – [STEP 3]

Cluster	$\bar{x}_1$	$\bar{x}_2$
(A)	5	3
(BCD)	$\frac{-1+1+(-3)}{3} = -1$	$\frac{1+(-2)+(-2)}{3} = -1$

Squared distances to cluster centroids

Cluster	A	B	C	D
(A)	0	40	41	89
(BCD)	52	4	5	5

$$d^2(A, C_{(A)}) = (5-5)^2 + (3-3)^2 = 0$$

$$d^2(A, C_{(BCD)}) = (5+1)^2 + (3+1)^2 = 52$$

→ (A), (BCD)

5.4 Non-hierarchical clustering methods

◆ [Example 5.4.2] K-median method

Observation	variable X_1	X_2
A	5	3
B	−1	1
C	1	−2
D	−3	−2

[STEP 1] Divide *n* observations into K initial clusters : (AB), (CD)

[STEP 2] The centroid of each cluster is calculated as the median of the observations in the cluster.

Cluster	x_1^M	x_2^M
(AB)	$\frac{5+(-1)}{2}=2$	$\frac{3+1}{2}=2$
(CD)	$\frac{1+(-3)}{2}=-1$	$\frac{-2+(-2)}{2}=-2$

Centre coordinate : $C_{(AB)}=(2,2),\ C_{(CD)}=(-1,-2)$

[STEP 3] After the Euclidean distance, assign the entity to the cluster with the closest centroid.

Ref: [Example 5.4.1] –[Step 3]

A → (A), B → (BCD)

[STEP 4] Repeat [STEP 2] – [STEP 3]

Cluster	x_1^M	x_2^M
(A)	5	3
(BCD)	-1	-2

$\boxed{-1,\ 1,\ -3}$ $\boxed{1,\ -2,\ -2}$

Coordinates of centroid : $C_{(A)}=(5,\ 3),\ C_{(BCD)}=(-1,\ -2)$

Cluster	A	B	C	D
(A)	0	40	41	89
(BCD)	61	9	20	20

5.4 Non-hierarchical clustering methods

◆ [Example 5.4.4] Public Utility Data in [Data 5.3.1]

Company	x1	x2	x3	x4	x5	x6	x7	x8
1. Arizona Public Service	1.06	9.2	151	54.4	1.6	9077	0.0	0.628
2. Boston Edison Co.	0.89	10.3	202	57.9	2.2	5088	25.3	1.555
3. Central Louisiana Electric Co.	1.43	15.4	113	53.0	3.4	9212	0.0	1.058
4. Commonwealth Edison Co.	1.02	11.2	168	56.0	0.3	6423	34.3	0.700
5. Consolidated Edison Co. (NY)	1.49	8.8	192	51.2	1.0	3300	15.6	2.044
6. Florida Power & Light Co.	1.32	13.5	111	60.0	-2.2	11127	22.5	1.241
7. Hawaiian Electric Co.	1.22	12.2	175	67.6	2.2	7642	0.0	1.652
8. Idaho Power Co.	1.10	9.2	245	57.0	3.3	13082	0.0	0.309
9. Kentucky Utilities Co.	1.34	13.0	168	60.4	7.2	8406	0.0	0.862
10. Madison Gas & Electric Co.	1.12	12.4	197	53.0	2.7	6455	39.2	0.623
11. Nevada Power Co.	0.75	7.5	173	51.5	6.5	17441	0.0	0.768
12. New England Electric Co.	1.13	10.9	178	62.0	3.7	6154	0.0	1.897
13. Northern States Power Co.	1.15	12.7	199	53.7	6.4	7179	50.2	0.527
14. Oklahoma Gas & Electric Co	1.09	12.0	96	49.8	1.4	9673	0.0	0.588
15. Pacific Gas & Electric Co	0.96	7.6	164	62.2	-0.1	6468	0.9	1.400
16 Puget Sound Power & Light Co.	1.16	9.9	252	56.0	9.2	15991	0.0	0.620
17. San Diego Gas & Electric Co.	0.76	6.4	136	61.9	9.0	5714	8.3	1.920
18. The Southern Co.	1.05	12.6	150	56.7	2.7	10140	0.0	1.108
19. Texas Utilities Co.	1.16	11.7	104	54.0	-2.1	13507	0.0	0.636
20. Wisconsin Electric Power Co.	1.20	11.8	148	59.9	3.5	7287	41.1	0.702
21. United Illuminating Co.	1.04	8.6	204	61.0	3.5	6650	0.0	2.116
22. Virginia Electric & Power Co.	1.07	9.3	174	54.3	5.9	10093	26.6	1.306

X1: Fixed-charge coverage ratio (income/debt)
X2: Rate of return on capital
X3: Cost per KW capacity in place
X4: Annual load factor
X5: Peak KWH demand growth from 197 to 1975
X6: Sales (KWH use per year)
X7: Percent nuclear
X8: Total fuel costs (cents per KWH)

[R-code 5.4.1]

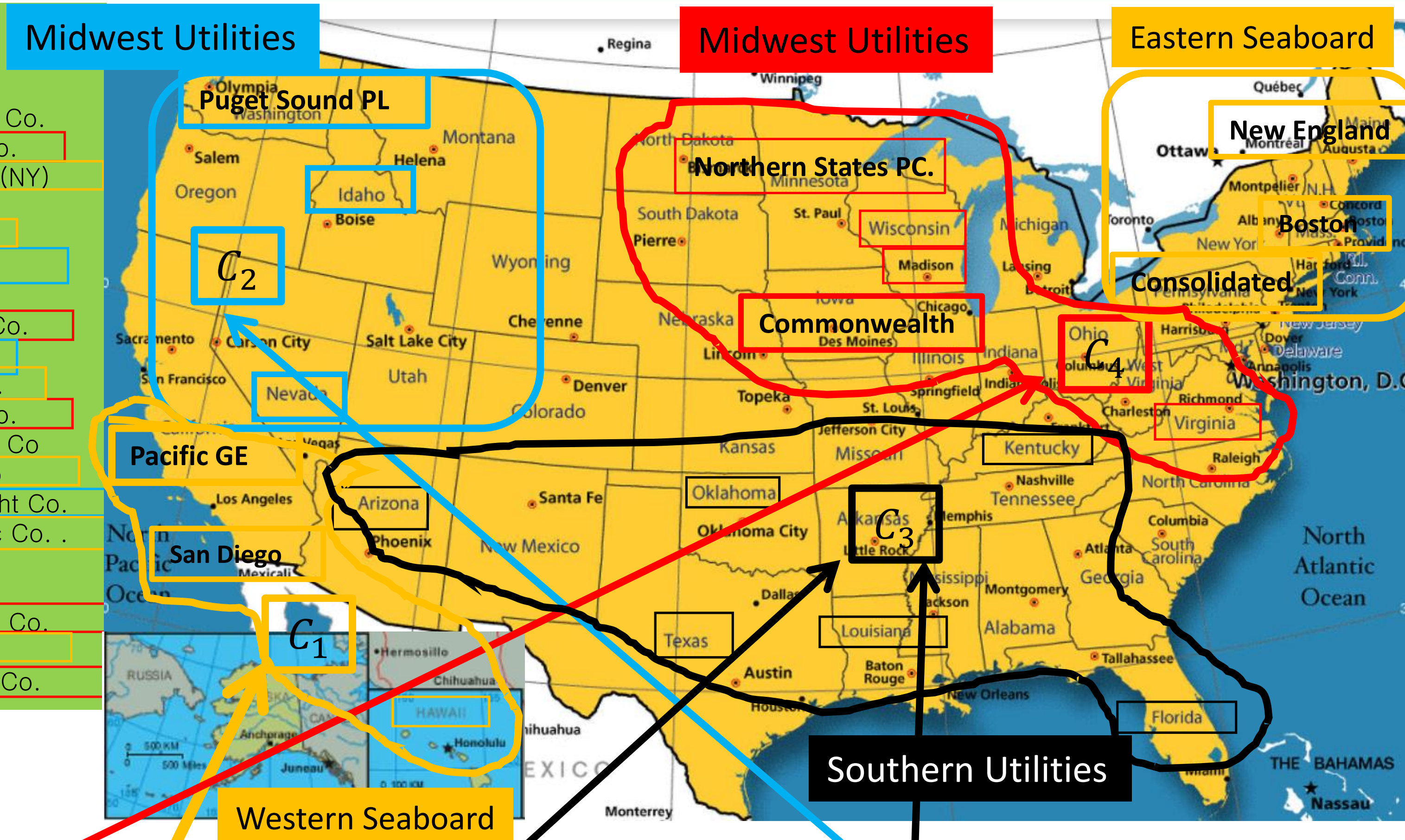
K-means	C_1	C_2	C_3	C_4
	2보스톤 5콘솔 7하와이안 12뉴잉글랜드 15퍼시픽 17샌디에고 21유나이티드	8아이다호 11네바다 16푸젯싸운드	1애리조나 3루이지애나 6플로리다 9캔터키 14오클라호마 18서던캄퍼니 19텍사스	4코몬웰스 10매디슨 13노던스테이트 20위스콘신 22버지니아
K-평균법 군집특성	서쪽 태평양 연안 극 서부 지방에 위치 X_8: 총연료비용이 높다.	X_3: KW당 비용이 높다.	주로 남부에 위치하며 X_2: 자금수익률과 X_6: 판매량이 높다.	X_7: 핵발전 비율이 높다.

5.4 Non-hierarchical clustering methods

[Example 5.4.4] Geographical Locations of Utility Companies

Company

1. Arizona Public Service
2. Boston Edison Co.
3. Central Louisiana Electric Co.
4. Commonwealth Edison Co.
5. Consolidated Edison Co. (NY)
6. Florida Power & Light Co.
7. Hawaiian Electric Co.
8. Idaho Power Co.
9. Kentucky Utilities Co.
10. Madison Gas & Electric Co.
11. Nevada Power Co.
12. New England Electric Co.
13. Northern States Power Co.
14. Oklahoma Gas & Electric Co
15. Pacific Gas & Electric Co
16. Puget Sound Power & Light Co.
17. San Diego Gas & Electric Co. .
18. The Southern Co.
19. Texas Utilities Co.
20. Wisconsin Electric Power Co.
21. United Illuminating Co.
22. Virginia Electric & Power Co.



X1: Fixed-charge coverage ratio (income/debt), X2: Rate of return on capital, X3: Cost per KW capacity in place, X4: Annual load factor, X5: Peak KWH demand growth from 197 to 1975, X6: Sales (KWH use per year), X7: Percent nuclear, X8: Total fuel costs (cents per KWH)

5.4 Non-hierarchical clustering methods

[Table 5.4.3] Mean of Variables for 4 Clusters : [R-code 5.4.1] aggregate()

군집	고정 요금	자금 수익률	KW당 비용	연 부하율	수요 성장	판매량	핵비율	총연료 비용	특 성
C_1	1.49	8.80	192.00	51.20	1.00	3300.00	15.60	2.04	총연료비용이 높다(태평양연안)
C_2	1.00	9.33	176.50	62.10	3.42	6286.00	5.75	1.76	KW당 비용이 높다 **높은 연부하율**
C_3	1.06	10.16	168.13	54.21	3.56	12375.50	3.33	0.75	자금수익률, 판매량 높다(남부)
C_4	1.23	12.86	157.71	56.57	3.04	8012.71	26.76	0.82	핵발전 비율이 높다

X1: Fixed-charge coverage ratio (income/debt)

X2: Rate of return on capital

X3: Cost per KW capacity in place

X4: Annual load factor

X5: Peak KWH demand growth from 197 to 1975

X6: Sales (KWH use per year)

X7: Percent nuclear

X8: Total fuel costs (cents per KWH)

5.4 Non-hierarchical clustering methods

◆ [Example 5.4.5] Economic View Data in [Data 1.3.5]

◆ [Table 5.4.4] Mean of 10 Variables for 4 Clusters

Growth /Export/Import/ Black-ink balance/ Foreign debt /Exchange rate/ Unemployed rate /Consumer price rate /Wage increase rate

군집	성장률	GNP	수출	수입	국제 흑자	연말 외채	연말 환율	실업률	소비 물가	임금 상승	특성
C_1	8.650	1976	699.75	635.0	101.0	284.38	631.25	2.83	5.960	11.63	Optimistic- Support
C_2	8.667	1853	703.33	618.3	95.0	263.33	614.33	3.73	5.833	12.67	Pessimistic
C_3	8.975	1870	729.00	649.0	123.0	260.00	616.00	3.40	6.100	15.80	Critical
C_4	7.850	1875	682.50	652.5	91.5	280.00	615.00	2.95	6.600	14.40	Critical

Ward linkage

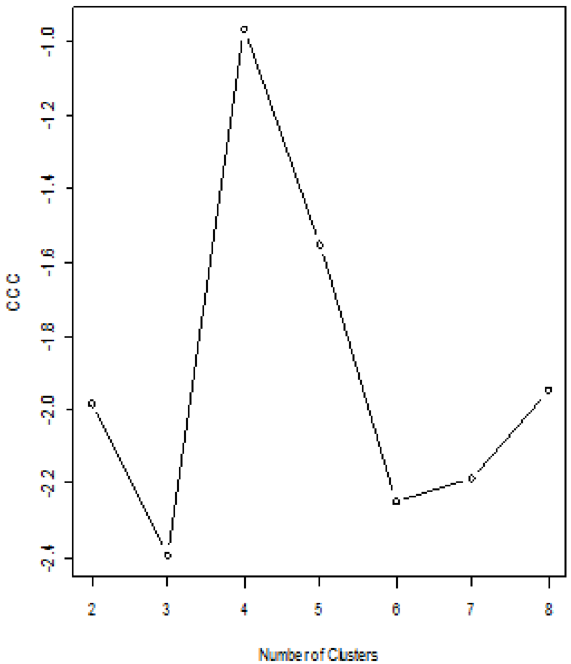
K-means

[R-code 5.4.2]

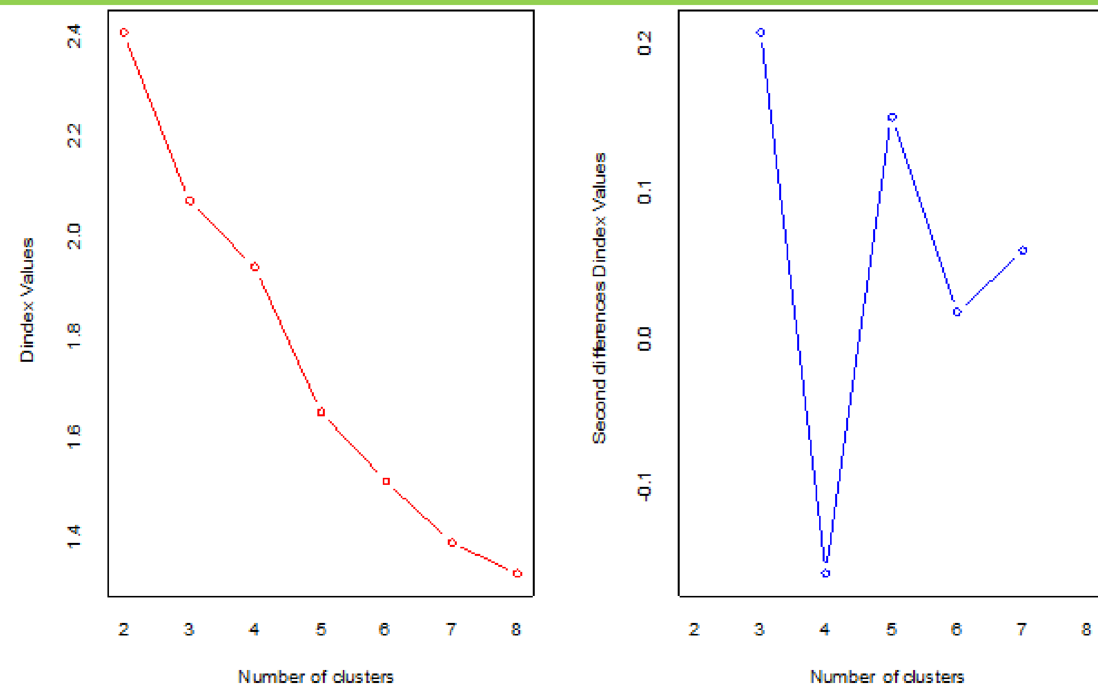
	대우증권 한국은행 대한상공회의소 삼성물산 동서증권 선경 신한종합연구소 동서경제연구소	전경련 대우경제연구소	고려대학교	중소기업중앙회 현대자동차 중앙대학교
K-평균법 군집특성	낙관론-정책옹호	비관론	정책비판	정책비판

5.5 Numbers of Clusters

◆ [Example 5.5.1] Utility Data : initial number k of clusters in K-means



(a) CCC



(b) Dindex

$$CCC = \ln\left[\frac{1-E(R^2)}{1-R^2}\right]\frac{\sqrt{np^*}}{(0.001+E(R^2))^{1.2}}$$

$$w(P^g) = \frac{1}{g}\sum_{k=1}^{g}\frac{1}{n_k}\sum_{x_{ki}\in C_k} d(x_{ki},\ c_k)$$

$$w(P^{g-1}) - w(P^g)$$

Max CCC k

Slope drops very sharply after k

Slope increases very sharply after k

5.5 Numbers of Clusters

◆ [Result 5.5.1] [R-code 5.5.1](utility-CAnclusterindex.R) : option index = “all ”

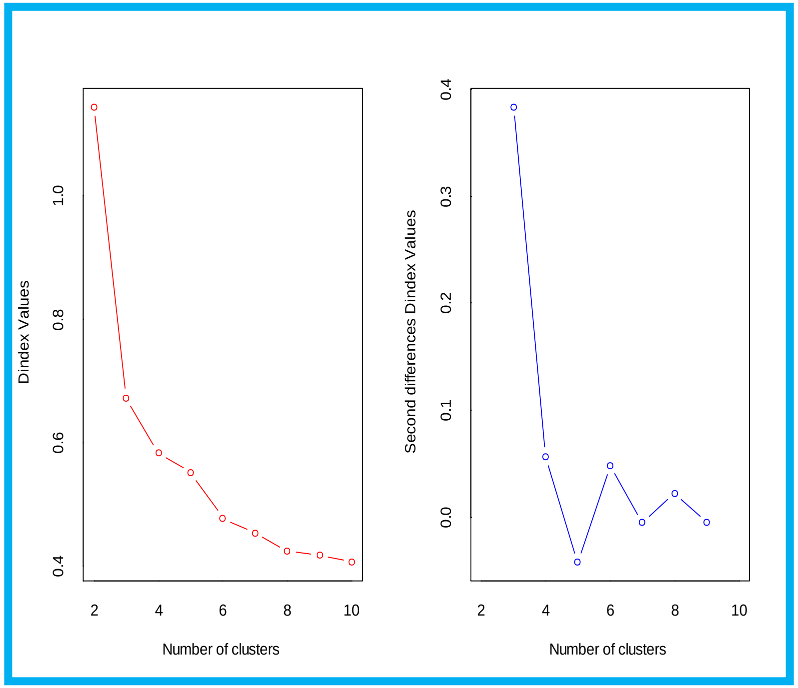
30 Indices

```
# Index for the Number of Clusters in K-Means CA: Public Utilities
library(NbClust)
Data5.3.1<-read.table("utility.txt", header=T)
X<-Data5.3.1[,-1]
Z<-scale(X)
company=Data5.3.1[,1]

#CCC Index
ccc<-NbClust(Z, distance="euclidean", min.nc = 2, max.nc = 8,
 method = "kmeans", index = "ccc")
ccc
plot(2:8, type="b", ccc$All.index, xlab="Number of Clusters",
   ylab="CCC")

#Dindex Index
dindex<-NbClust(Z, distance="euclidean", min.nc = 2, max.nc = 8,
 method = "kmeans", index = "dindex")
dindex

#All Indices
allindex<-NbClust(Z, distance="euclidean", min.nc = 2, max.nc = 8,
method = "kmeans", index = "all", )
allindex
```

```
*******************************************************************
* Among all indices:
* 4 proposed 2 as the best number of clusters
* 1 proposed 3 as the best number of clusters
* 10 proposed 4 as the best number of clusters
* 1 proposed 6 as the best number of clusters
* 4 proposed 7 as the best number of clusters
* 3 proposed 8 as the best number of clusters
                   ***** Conclusion *****
* According to the majority rule, the best number of clusters is 4
*******************************************************************
```

5.5 Numbers of Clusters

[Example 5.5.2] Iris flower data in [Data 1.3.4]



****

* Among all indices:
* 2 proposed 2 as the best number of clusters
* 15 proposed 3 as the best number of clusters
* 5 proposed 4 as the best number of clusters
* 1 proposed 6 as the best number of clusters
* 1 proposed 8 as the best number of clusters
* 3 proposed 10 as the best number of clusters

***** Conclusion *****

* According to the majority rule, the best number of clusters is 3

****

Table 5.5.1] clusters of 3-means method

iristype	setosa	versicolor	virginica
1	0	2	36
2	0	48	14
3	50	0	0

*APER*3 = (2 + 14 + 0)/150 x 100% = 10.67%

5.7 R for CA : Practice Time

R-Code:

CA(Hierarchical, Nonhierarchical, Statistical model) and Number of clusters	
hcluster(, method="single") method="complete"/"average"/"ward"	Hierarchical CA
kmeans(), aggregate() library(cluster), pam()	Non-hierarchical CA - k-means method - K-median method - K-medoids(partitioning around) method
library(NbClust), NbClust(, index="all"	CCC(cubic clustering criterion), Dindex

5.7 R for CA : Practice Time

R-code list of Chapter 3 Cluster Analysis

economicview-distances.R	[R-코드 5.2.1]	경제관련기관 경제전망의 연관성측도인 4가지 거리계산
5obsdisit-CAsingle.R	[R-코드 5.3.1]	5명 개체간의 단일연결법과 덴드로그램
economicview-CAward.R	[R-코드 5.3.2]	경제관련기관 경제전망의 3가지 거리에 대한 와드연결법의 덴드로그램
klpga-CAamlinkages.R	[R-코드 5.3.3]	KLPGA 선수 성적의 표준화 유클리드 거리에 대한 계층적 군집분석의 덴드로그램
utility-CAward.R	[R-코드 5.3.4]	공익회사 자료의 표준화 유클리드 거리에 대한 와드연결법의 덴드로그램
utility-CAKmeansKmedoids.R	[R-코드 5.4.1]	공익회사 자료의 표준화 유클리드 거리에 대한 K-평균법과 K-대표개체법
economicview-CAKmeansKmedoids.R	[R-코드 5.4.2]	경제관련기관의 표준화 유클리드 거리에 대한 K-평균법과 K-대표개체법
utility-CAnclusterindex.R	[R-코드 5.5.1]	공익회사 자료의 표준화 유클리드 거리에 대한 K-평균법에서 군집 수를 위한 지수 구하기
iris-CAindex.R	[R-코드 5.5.2]	[그림 5.5.2]을 위한 R-코드
iris-CAmodel.R	[R-코드 5.6.1]	붓꽃(iris flower) 자료의 통계모형에 의한 군집분석

5.7 R for CA : Practice Time

[R-code 5.3.3] klpga-CAamlinkages.R based on the Hierarchical Methods

```
# AMCA : AM Linkages
Data1.3.2<-read.table("klpga.txt", header=T)
X<-Data1.3.2
X<-as.matrix(Data1.3.2)
선수<-rownames(X)

n<-nrow(X)
xbar<-t(X)%*%matrix(1,n,1)/n  # 평균벡터
I<-diag(n)
J<-matrix(1,n,n)
H<-I-1/n*J                    # 중심화행렬
Y<-H%*%X                      # 중심화 자료행렬
S<-t(Y)%*%Y/(n-1)             # 공분산행렬
D<-diag(1/sqrt(diag(S)))      # 표준편차행렬의 역
Z<-Y%*%D                      # 표준화자료행렬
colnames(Z)<-colnames(X)
```

```
# 표준화 유클리드 거리
ds <- as.matrix(dist(Z, method="euclidean"))
ds <- as.dist(ds)
round(ds, 3)
#단일연결법
sinle=hclust(ds, method="single")
plot(sinle, labels=선수, hang=-1, main="(a) Sinle Linkage")
#완전연결법
complete=hclust(ds, method="complete")
plot(complete, labels=선수, hang=-1, main="(b) Complete Linkage")
#평균연결법
average=hclust(ds, method="average")
plot(average, labels=선수, hang=-1, main="(c) Average Linkage")
#와드연결법
ward=hclust(ds, method="ward")
plot(ward, labels=선수, hang=-1, main="(d) Ward Linkage")
```

5.7 R for CA : Practice Time

[R-code 5.4.1] utility-CAKmeansKmedoids.R based on the Non-hierarchical Methods

```
# K-Means & K-Medoids(Partitioning Around Medoids)CA for Public Utilities
Data5.3.1<-read.table("utility.txt", header=T)
X<-Data5.3.1[,-1]
Z<-scale(X)
company=Data5.3.1[,1]

# K-means Method
kmeans <- kmeans(Z, 4) # 4 cluster solution
cluster=data.frame(company,cluster=kmeans$cluster)
C1=cluster[(cluster[,2]==1),]
C2=cluster[(cluster[,2]==2),]
C3=cluster[(cluster[,2]==3),]
C4=cluster[(cluster[,2]==4),]
C1;C2;C3;C4

# Get cluster means
aggregate(X, by=list(kmeans$cluster),FUN=mean)

# K-medoids Method
library(cluster)
kmedoids <- pam(Z, 4, metric="euclidean") # 4 cluster solution
cluster=data.frame(company,cluster=kmedoids$cluster)
C1=cluster[(cluster[,2]==1),]
C2=cluster[(cluster[,2]==2),]
C3=cluster[(cluster[,2]==3),]
C4=cluster[(cluster[,2]==4),]
C1;C2;C3;C4
# Get cluster means
aggregate(X,by=list(kmedoids$cluster),FUN=mean)
```