

Visualizations of Big Data



: **UND** Understanding Data

CDA Categorical Data Analysis

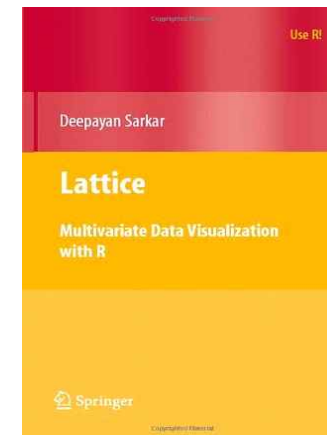
MDA Multivariate Data Analysis



최 용 석 (부산대학교 통계학과)

yschoi.pusan.ac.kr

2019.08.22.



부산대학교 통계연구소 워크숍

주제 1 : 자료의 이해

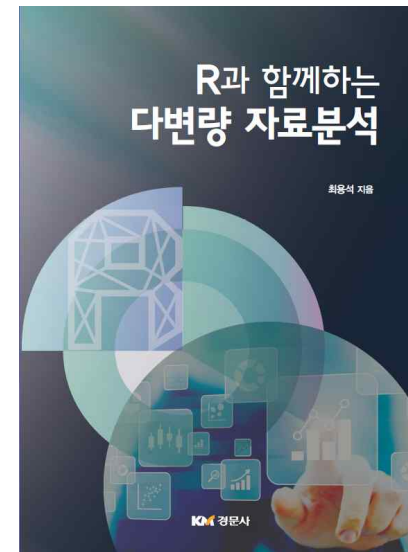
(Understanding Data)

Prolog

1.1 다변량자료 Multivariate Data

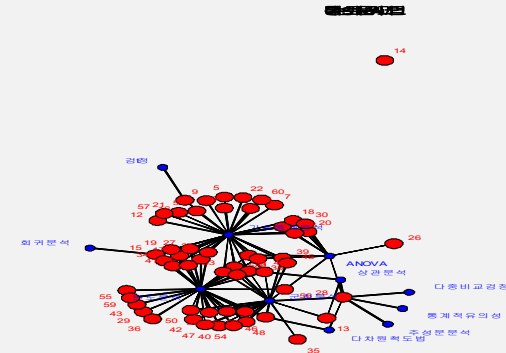
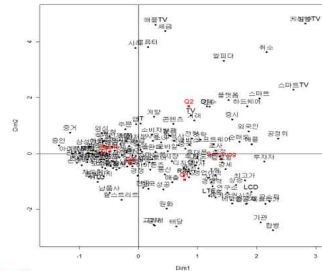
1.2 범주형자료 Categorical Data

1.3 텍스트자료 Text Data



Prolog : Unstructured Data(Text Data) ⇒ Structured Data(EDA, CDA, MDA)

통계적 기법 적용을 위한 비정형 빅 데이터의 정형화와 분석 및 시각화



MapReduce

워드 클라우드(Word Cloud)

a. row data

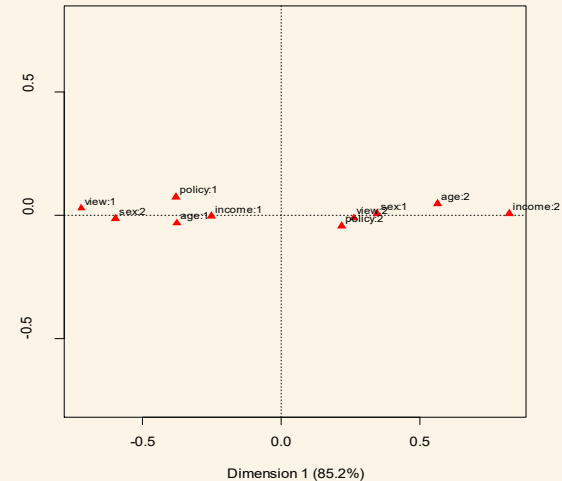
번호	년도	키워드
논문1	2015	남해, 동영, 화성출산업, 글, 라디칼, DPPH
논문2	2015	외포산업, 해양유류추출물, 알조식
논문3	2014	미세조류, 외포산업, 알조식, 알조식
논문4	2014	외포산업, 알조식, 알조식, 알조식
...	...	...
논문60	1989	생물유화제, 부산항, 울산항, 충무항, 소재산업

b. 빈도표

키워드	빈도
1. 지해	18
2. 저서동물	17
3. 남해	17
4. 군집	16
5. 외포	13
6. 산업	13
7. 외포	10
8. 분포	10
9. 갯벌	10
10. 해조류	8



MDA 결과도:



1.1 다변량자료 Multivariate Data

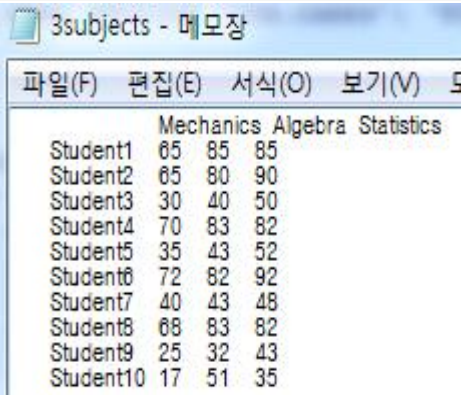
○ 자연과학, 사회과학, 의학, 경제학, 경영학, 교육학 등 여러 분야에서 관심이 되는 현상을 설명하기 위해 **하나의 측정변수**에 의해서 결과를 관찰할 수도 있지만 **대부분 많은 변수로 다양한 관점에서 자료를 수집**한다.

○ 일반적으로 **n 개의 개체(objects)**에 대해 **p 개의 변수(variables)**로 측정하며 이를 **다변량 자료(multivariate data)**라 한다.

○ **다변량 자료의 표현**: $n \times p$ Data Matrix $X = \begin{bmatrix} x_{11} & \cdots & x_{1j} & \cdots & x_{1p} \\ \vdots & & \vdots & & \vdots \\ \vdots & & \vdots & & \vdots \\ x_{i1} & \cdots & x_{ij} & \cdots & x_{ip} \\ \vdots & & \vdots & & \vdots \\ \vdots & & \vdots & & \vdots \\ x_{n1} & \cdots & x_{nj} & \cdots & x_{np} \end{bmatrix} = (x_{ij}), i = 1, \dots, n, j = 1, \dots, p$

[자료 1.1.0] 10명의 학생들의 3과목 시험성적 : 자료 요약과 차원 공간의 기하적 표현

3subjects.txt				3subjects-3d.R			
Students	Mechanics	Algebra	Statistics	<pre> setwd("C:/빅데이터/R_data") Data1.1.1<-read.table("3subjects.txt", header=T) X<-Data1.1.1 # Descriptive Statistics summary(X) # Covariance Matrix cov(X) # Correlation Matrix cor(X) # Multiple Scatter plot plot(X) # Boxplot of 3 Subjects boxplot(X, xlab="3 Subjets", ylab="Exam Marks") # Stem-and-leaf plot stem(X[,1]) stem(X[,2]) stem(X[,3]) </pre>			
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				



[1] 기술통계

Mechanics	Algebra	Statistics
Min. : 17.00	Min. : 32.00	Min. : 35.00
1st Qu. : 31.25	1st Qu. : 43.00	1st Qu. : 48.50
Median : 52.50	Median : 65.50	Median : 67.00
Mean : 48.70	Mean : 62.20	Mean : 65.90
3rd Qu. : 67.25	3rd Qu. : 82.75	3rd Qu. : 84.25
Max. : 72.00	Max. : 85.00	Max. : 92.00

[6] 줄기-잎 그림

Mechanics	Algebra	Statistics
0 7	2 2	2 5
2 505	4 0331	4 3802
4 0	6	6
6 55802	8 02335	8 22502

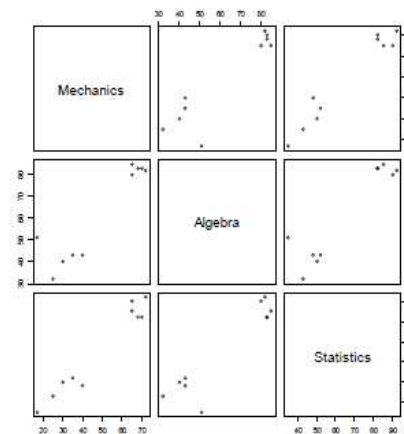
[2] 공분산행렬

	Mechanics	Algebra	Statistics
Mechanics	453.3444	431.5111	459.0778
Algebra	431.5111	484.6222	450.2444
Statistics	459.0778	450.2444	487.8778

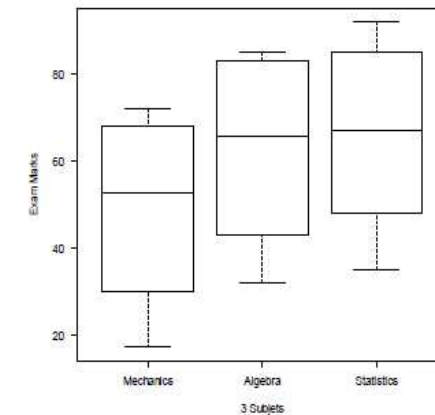
[3] 상관행렬

	Mechanics	Algebra	Statistics
Mechanics	1.0000000	0.9206110	0.9761501
Algebra	0.9206110	1.0000000	0.9259578
Statistics	0.9761501	0.9259578	1.0000000

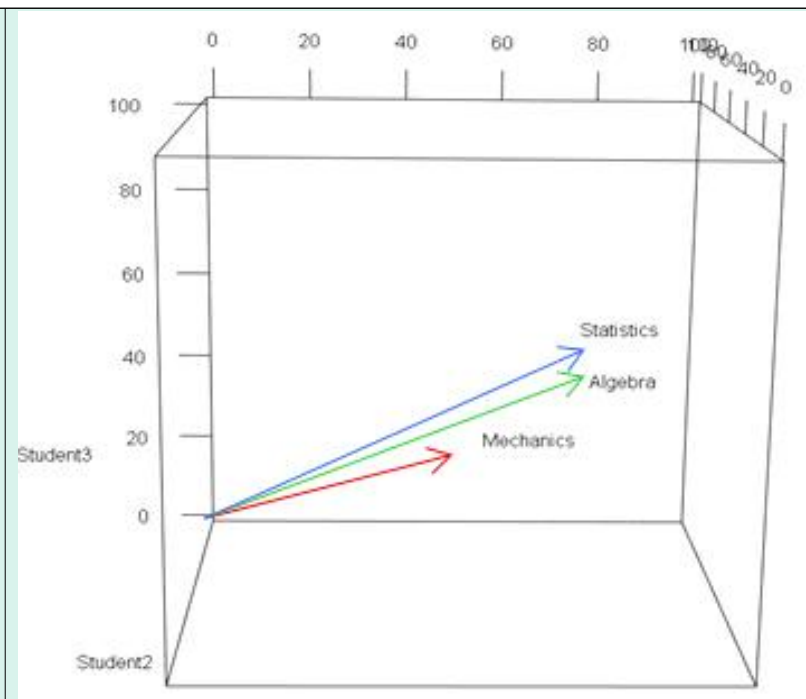
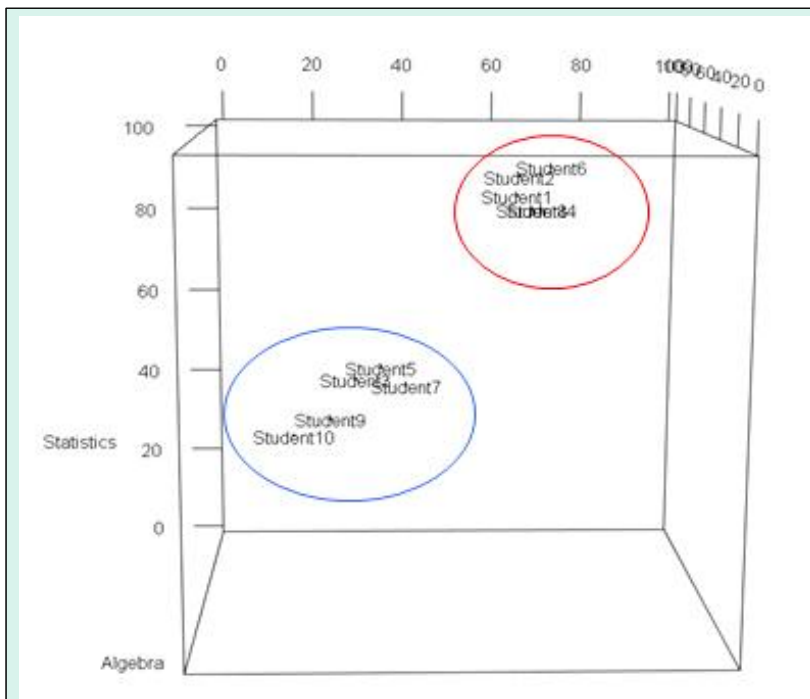
[4] 다중 산점도



[5] 상자 그림



다변량 자료의 요약 : Summary Statistics, Covariance Matrix, Correlation Matrix, Multiple Scatter Plot, Box Plot, Stem-and-Leaf Plot



변수 공간의 개체 : $Q^p - Space$

```
library(rgl)
# Observations in Variables Space
lim<-c(0, 100)
plot3d(X[,1], X[,2], X[,3],xlim=lim, ylim=lim, zlim=lim,
xlab="Mechanics", ylab="Algebra", zlab="Statistics")
text3d(X[,1], X[,2], X[,3],rownames(X))
```

개체 공간의 변수 : $R^n - Space$

```
# Variables in Observations Space
plot3d(X[1,], X[2,], X[3,], xlim=lim, ylim=lim, zlim=lim,
xlab="Student1", ylab="Student2", zlab="Student3")
text3d(X[1,], X[2,], X[3,], colnames(X))
```

$Q^p - Space$ 의 자료 점	$R^n - Space$ 의 자료 점
$X = \begin{bmatrix} \mathbf{x}_1^t \\ \vdots \\ \mathbf{x}_i^t \\ \vdots \\ \mathbf{x}_n^t \end{bmatrix}, \quad \mathbf{x}_i^t = (x_{i1}, \dots, x_{ip}), \quad i = 1, \dots, n$	$X = [\mathbf{x}_{(1)}, \dots, \mathbf{x}_{(j)}, \dots, \mathbf{x}_{(p)}]$ $\mathbf{x}_{(j)} = (x_{1j}, \dots, x_{nj})^t, \quad j = 1, \dots, p$
p 차원 변수공간의 개체들에 대한 좌표점 : $\mathbf{x}_i \in Q^p, \quad i = 1, \dots, n$	n 차원 개체공간의 변수들에 대한 좌표점 : $\mathbf{x}_{(j)} \in R^n, \quad j = 1, \dots, p$
$d_{rs} = \ \mathbf{x}_r - \mathbf{x}_s\ = \left[(\mathbf{x}_r - \mathbf{x}_s)^t (\mathbf{x}_r - \mathbf{x}_s) \right]^{1/2}$ $= \left[\sum_{j=1}^p (x_{rj} - x_{sj})^2 \right]^{1/2}$: Euclidean Distance $d_{rs} = \ \mathbf{x}_r - \mathbf{x}_s\ _{S^{-1}} = \left[(\mathbf{x}_r - \mathbf{x}_s)^t S^{-1} (\mathbf{x}_r - \mathbf{x}_s) \right]^{1/2}$: Mahalanobis Distance	$\mathbf{y}_{(j)} = (y_{1j}, \dots, y_{nj})^t = (x_{1j} - \bar{x}_j, \dots, x_{nj} - \bar{x}_j)^t,$ $\cos \theta_{kj} = \frac{\mathbf{y}_{(k)}^t \mathbf{y}_{(j)}}{\ \mathbf{y}_{(k)}\ \ \mathbf{y}_{(j)}\ }$ $= \frac{\sum_{i=1}^n (x_{ik} - \bar{x}_k)(x_{ij} - \bar{x}_j)}{\sqrt{\sum_{i=1}^n (x_{ik} - \bar{x}_k)^2} \sqrt{\sum_{i=1}^n (x_{ij} - \bar{x}_j)^2}},$ $= r_{kj}.$

[자료 1.1.1] 두 가지 방식의 시험성적(Mardia 외 2인, 1979, pp. 3-4)

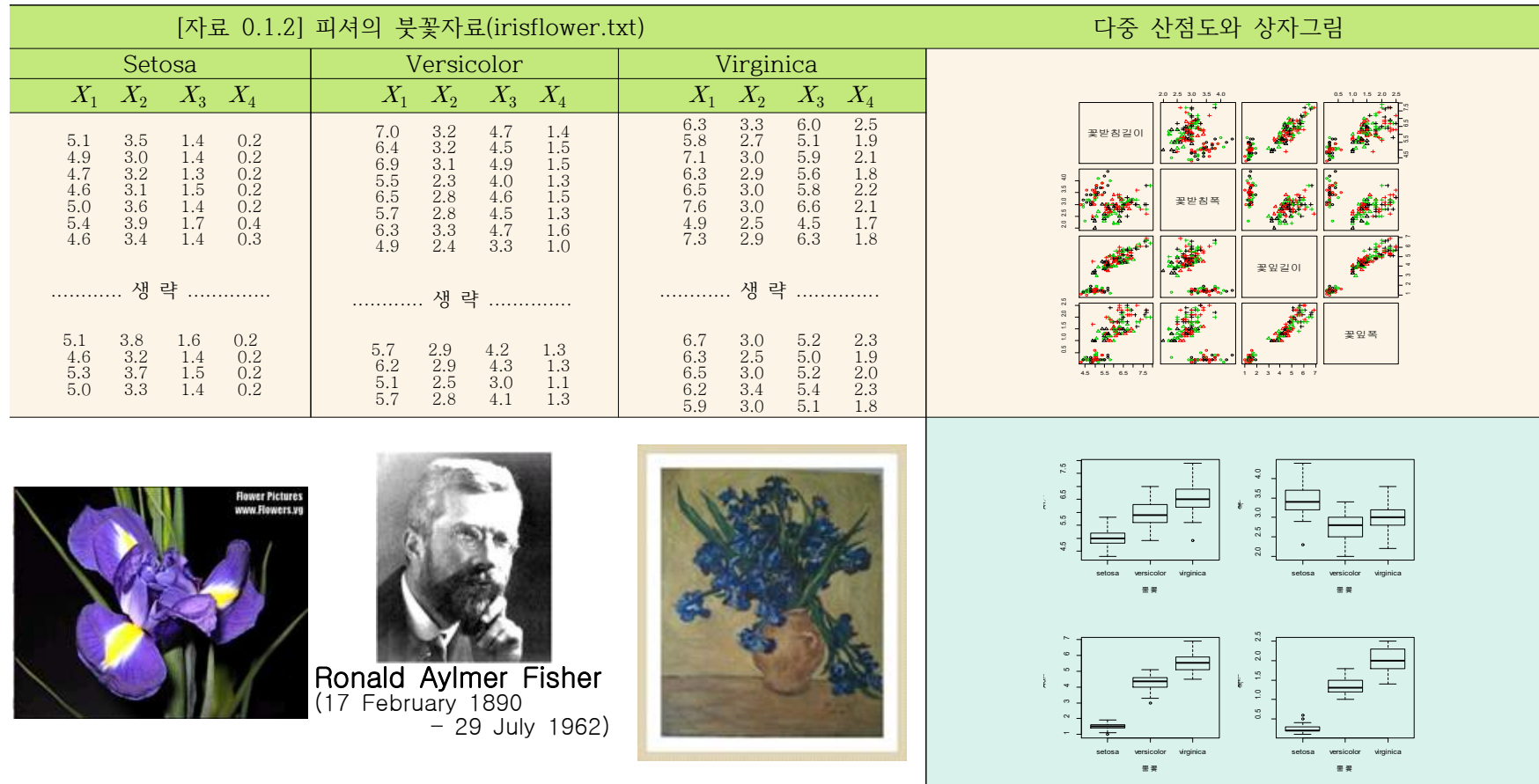
- R에서 오픈소스로 제공(library(MVT); data(examScor))
- 학생 88명을 대상으로 책을 참고하는 경우(open-book), 책을 참고하지 않는 경우(closed-book)에 따라 치른 다섯 과목 Mechanics(c), Vectors(c), Algebra(o), Analysis(o), Statistics(o)의 성적

[자료 0.1.1] 두 가지 방식의 시험성적(5subjects.txt)					
Student	Mechanics(c)	Vectors(c)	Algebra(o)	Analysis(o)	Statistics(o)
1	77	82	67	67	81
2	63	78	80	70	81
3	75	73	71	66	81
4	55	72	63	70	68
5	63	63	65	70	63
6	53	61	72	64	73
7	51	67	65	65	68
8	59	70	68	62	56
9	62	60	58	62	70
10	64	72	60	62	45
11	52	64	60	63	54
12	55	67	59	62	44
13	50	50	64	55	63
14	65	63	58	56	37
15	31	55	60	57	73
16	60	64	56	54	40
17	44	69	53	53	53
18	42	69	61	55	45
19	62	46	61	57	45
20	31	49	62	63	62
: 생략					
40	40	63	53	54	25
41	23	55	59	53	44
42	48	48	49	51	37
43	41	63	49	46	34
44	46	52	53	41	40

Student	Mechanics(c)	Vectors(c)	Algebra(o)	Analysis(o)	Statistics(o)	다중 산점도와 상자그림
45	46	61	46	38	41	
46	40	57	51	52	31	
47	49	49	45	48	39	
48	22	58	53	56	41	
49	35	60	47	54	33	
50	48	56	49	42	32	
51	31	57	50	54	34	
52	17	53	57	43	51	
53	49	57	47	39	26	
54	59	50	47	15	46	
55	37	56	49	28	45	
56	40	43	48	21	61	
57	35	35	41	51	50	
58	38	44	54	47	24	
59	43	43	38	34	49	
60	39	46	46	32	43	
: 생략						
80	30	24	43	33	25	
81	3	9	51	47	40	
82	7	51	43	17	22	
83	15	40	43	23	18	
84	15	38	39	28	17	
85	5	30	44	36	18	
86	12	30	32	35	21	
87	5	26	15	20	20	
88	0	40	21	9	14	

[자료 1.1.2] 피셔(Ronald Aylmer Fisher, 1890-1962)의 붓꽃자료(Johnson과 Wichern, 2007, pp. 658-659)

- R에서 오픈소스로 제공(data(iris)) : Sepal.Length, Sepal.Width, Petal.Length, Petal.Width, Species(변수명)
- 붓꽃 50포기씩을 임의로 추출: 꽃받침 길이(x_1), 꽃받침 폭(x_2), 꽃잎 길이(x_3), 꽃잎 폭(x_4).



[자료 1.1.3] 유럽 25개국의 단백질 섭취원(단위: 그램/일) 자료(Gabriel, 1981, pp. 150-154)

[자료 0.1.3] 유럽 25개국의 단백질 섭취원 자료(protein.txt)

국가	소고기	돼지·닭	계란	우유	생선	곡식	전분	콩·견과	과일·채소	Biplot : 행렬도
1 알바니아	10.1	1.4	0.5	8.9	0.2	42.3	0.6	5.5	1.7	
2 오스트리아	8.9	14.0	4.3	19.9	2.1	28.0	3.6	1.3	4.3	
3 벨기에-룩셈	13.5	9.3	4.1	17.5	4.5	26.6	5.7	2.1	4.0	
4 불가리아	7.8	6.0	1.6	8.3	1.2	56.7	1.1	3.7	4.2	
5 체코	9.7	11.4	2.8	12.5	2.0	34.3	5.0	1.1	4.0	
6 덴마크	10.6	10.8	3.7	25.0	9.9	21.9	4.8	0.7	2.4	
7 동독	8.4	11.6	3.7	11.1	5.4	24.6	6.5	0.8	3.6	
8 핀란드	9.5	4.9	2.7	33.7	5.8	26.3	5.1	1.0	1.4	
9 프랑스	18.0	9.9	3.3	19.5	5.7	28.1	4.8	2.4	6.5	
10 그리스	10.2	3.0	2.8	17.6	5.9	41.7	2.2	7.8	6.5	
11 헝가리	5.3	12.4	2.9	9.7	0.3	40.1	4.0	5.4	4.2	
12 아일랜드	13.9	10.0	4.7	25.8	2.2	24.0	6.2	1.6	2.9	
13 이탈리아	9.0	5.1	2.9	13.7	3.4	36.8	2.1	4.3	6.7	
14 네덜란드	9.5	13.6	3.6	23.4	2.5	22.4	4.2	1.8	3.7	
15 노르웨이	9.4	4.7	2.7	23.3	9.7	23.0	4.6	1.6	2.7	
16 폴란드	6.9	10.2	2.7	19.3	3.0	36.1	5.9	2.0	6.6	
17 포르투갈	6.2	3.7	1.1	4.9	14.2	27.0	5.9	4.7	7.9	
18 루마니아	6.2	6.3	1.5	11.1	1.0	49.6	3.1	5.3	2.8	
19 스페인	7.1	3.4	3.1	8.6	7.0	29.2	5.7	5.9	7.2	
20 스웨덴	9.9	7.8	3.5	24.7	7.5	19.5	3.7	1.4	2.0	
21 스위스	13.1	10.1	3.1	23.8	2.3	25.6	2.8	2.4	4.9	
22 영국	17.4	5.7	4.7	20.6	4.3	24.3	4.7	3.4	3.3	
23 소련	9.3	4.6	2.1	16.6	3.0	43.6	6.4	3.4	2.9	
24 서독	11.4	12.5	4.1	18.8	3.4	18.6	5.2	1.5	3.8	
25 유고슬라비아	4.4	5.0	1.2	9.5	0.6	55.9	3.0	5.7	3.2	

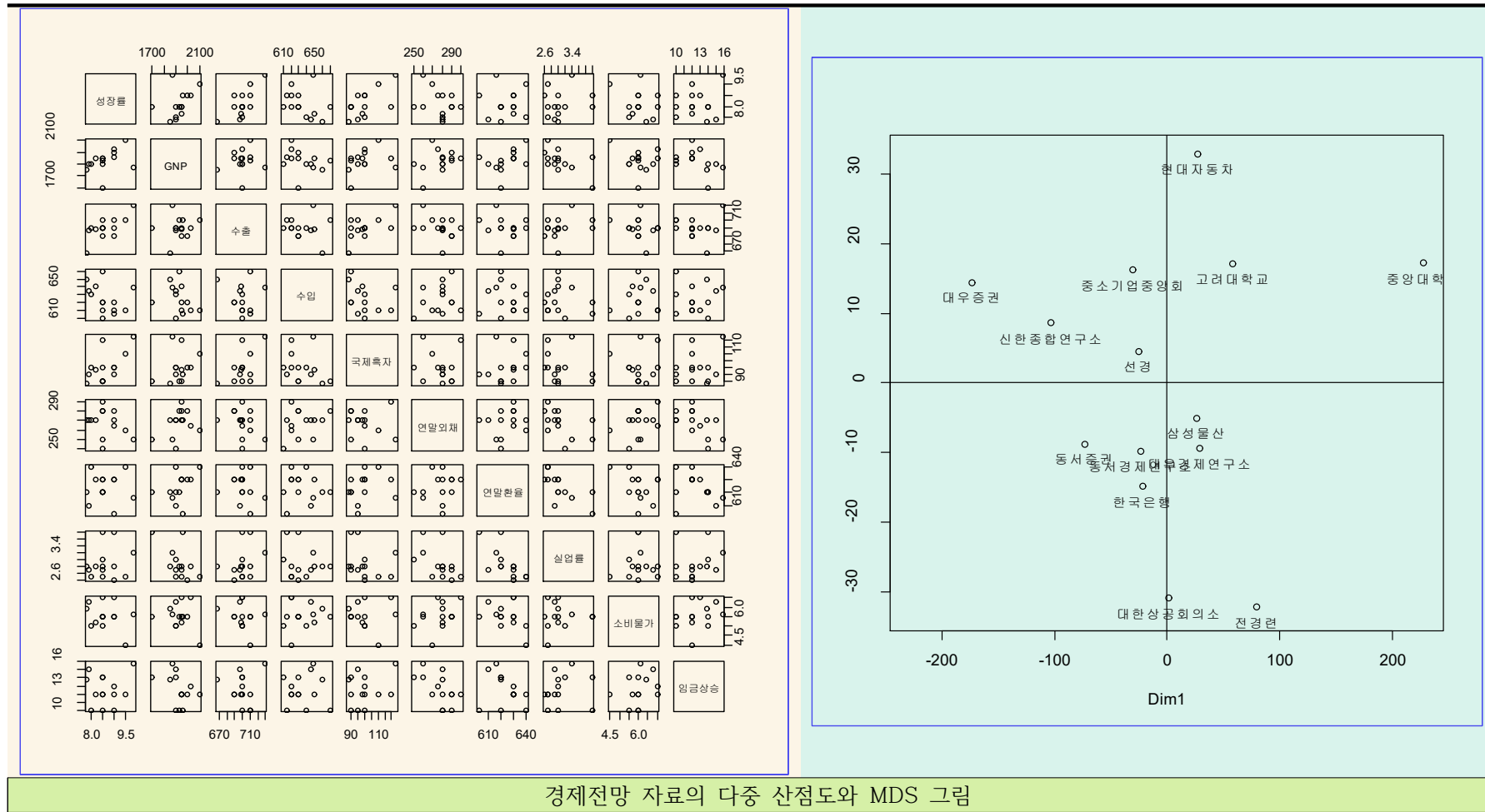
[자료 1.1.4] 경제관련 기관의 경제전망 자료(동아일보, 1989.1.4)

- 우리나라 14개 경제관련기관의 전문가들의 향후 1년간 10가지 경제전망(성장률, GNP, 수출, 수입, 국제흑자, 연말외채, 연말환율, 실업률, 소비자물가, 임금상승)에 관한 조사표
- (1) 성장률이 8.5% 이상 (2) GNP는 1950불 이상 (3) 수출은 700억불 이상 (4) 수입은 630억불 이하 (5) 국제흑자는 100억불 이하 (6) 연말외채는 280억불 이하 (7) 연말환율은 1불당 630원 이상 (8) 실업률은 3.0% 이하 (9) 소비자 물가의 상승률은 6.0% 이하 (10) 임금 상승률은 12.5% 이하

[자료 0.1.4] 우리나라 14개 경제관련기관의 향후 1년간 경제전망 자료(economicview.txt)

기관	성장률	GNP	수출	수입	국제흑자	연말외채	연말환율	실업률	소비물가	임금상승
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

이진수자료(economicviewbinary.txt)									
1	2	3	4	5	6	7	8	9	10
0	1	0	0	0	1	1	1	1	1
1	1	1	1	1	1	1	1	1	1
1	1	0	1	1	0	1	1	1	1
0	0	0	0	0	1	0	1	0	0
1	0	1	0	0	0	0	1	1	1
1	1	1	1	0	1	0	0	1	1
1	0	1	1	1	1	0	0	1	0
0	0	1	0	1	1	1	1	1	1
1	1	1	1	1	0	1	1	0	1
0	0	0	0	0	1	0	1	0	0
1	1	1	1	1	1	1	1	0	0
1	1	0	1	0	0	1	1	1	1
1	0	1	0	1	1	0	0	0	0
1	0	1	1	0	1	0	0	1	0



1.2 범주형자료 Categorical Data

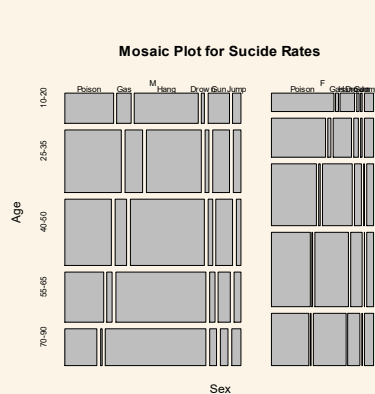
- 사회과학, 보건의학, 경제학, 경영학, 교육학 등 여러 분야에서 관심이 되는 현상에 대해 범주형 척도로 이루어진 측정변수에 의해서 자료를 수집한다.
- 이러한 자료는 일반적으로 n 개의 개체(object)에 대해 p 개의 범주변수(categorical variable)로 측정하며 각 범주변수는 여러 개의 범주(category)를 가지는 범주형 자료(categorical data)라 한다.
- 행동과학 : 정신질환유형(정신분열, 우울증, 신경질환)
정치학 : 정치철학(자유, 보수, 진보, 중도)
보건의학 : 진단결과(징후 있음, 징후 없음)

[자료 1.2.1] 자살유형에 따른 성별과 연령별 분할표 자료(최용석, 2017)

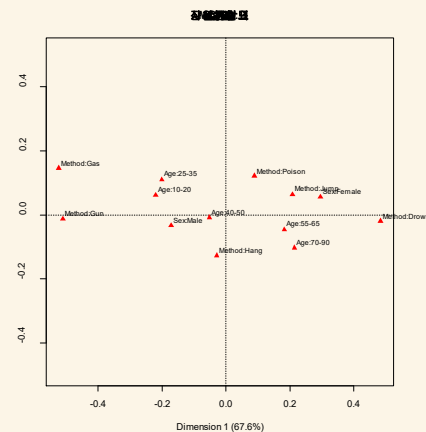
- 서독(West Germany), 성별과 연령별 여섯 가지 자살유형(Poison 독약, Gas 가스, Hang 목 매기, Drown 익사, Gun 총기, Jump 뛰어내림)에 따른 빈도수(frequency)를 나타내는 범주형자료.

[자료 0.2.1] $2 \times 5 \times 6$ 성별과 연령별 자살유형의 분할표(sucidefreq.txt)

Sex	Age	Poison	Gas	Hang	Drown	Gun	Jump
M	10-20	1160	335	1524	67	512	189
M	25-35	2823	883	2751	213	852	366
M	40-50	2465	625	3936	247	875	244
M	55-65	1531	201	3581	207	477	273
M	70-90	938	45	2948	212	229	268
F	10-20	921	40	212	30	25	131
F	25-35	1672	113	575	139	64	276
F	40-50	2224	91	1481	354	52	327
F	55-65	2283	45	2014	679	29	388
F	70-90	1548	29	1355	501	3	383



모자이크 그림



MCRA 그림

$2 \times 5 \times 6$ 삼원분할표(sucidefreq.txt)

Sex	Age	Method	Freq
Male	10-20	Poison	1160
Female	10-20	Poison	921
Male	25-35	Poison	2823
Female	25-35	Poison	1672
Male	40-50	Poison	2465
Female	40-50	Poison	2224
Male	55-65	Poison	1531
Female	55-65	Poison	2283
Male	70-90	Poison	938
Female	70-90	Poison	1548
Male	10-20	Gas	335
Female	10-20	Gas	40
----- 중간 생략 -----			
Female	55-65	Jump	388
Male	70-90	Jump	268
Female	70-90	Jump	383

세 가지 범주변수(Sex, Age, Method)의 범주에 의한 분류변수의 조합에 의한 자료

[자료 1.2.2] 분류변수 자료 : 외상 후 스트레스증후군(ptsd)

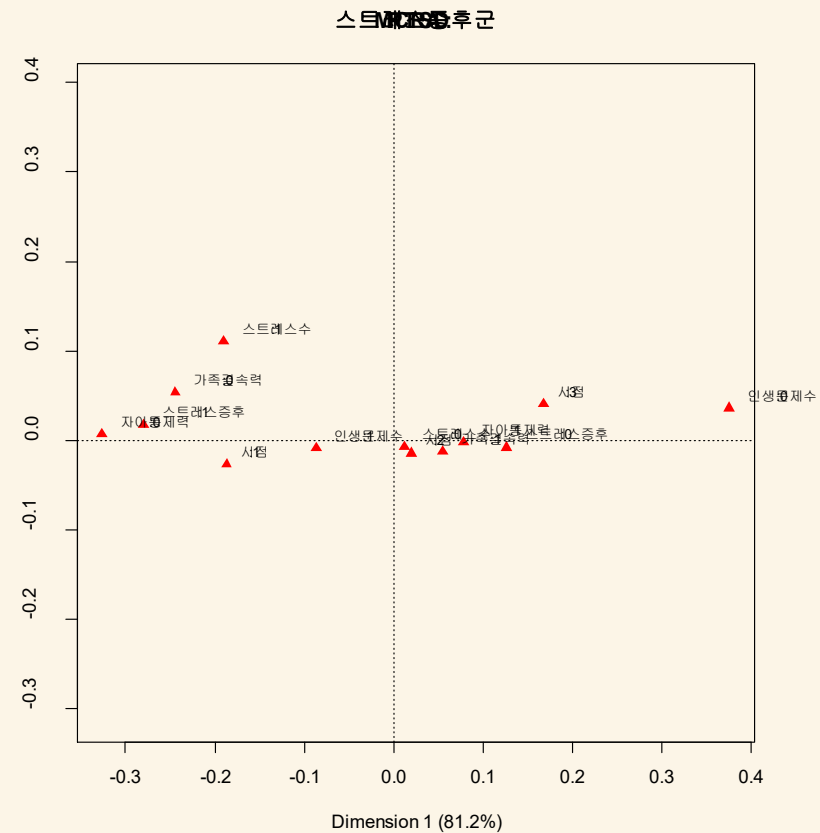
- － 화재에서 생존한 316명의 부상자 : 3개월, 6개월 그리고 12개월 후 면접을 통해 ptsd에 대한 다시점 자료(longitudinal data)

범주변수	분류변수 값	범주변수 명
개체고유번호	정수	subject
시점	1, 2, 3	time
스트레스증후	사례(1), 대조(0)	ptsd
자아통제력	실수	control
인생문제횟수	실수	problems
스트레스횟수	정수	sevent
가족결속력	정수	cohes

[자료 0.2.2] 외상 후 스트레스증후군에 대한 다시점자료(ptsd.txt)

subject	time	ptsd	control	problems	sevent	cohes
15	1	0	3.22222	5.625	1	8
15	2	0	3.16667	5.375	0	8
15	3	0	3.27778	3.750	1	8
18	1	1	2.55556	9.250	0	8
18	2	0	3.44444	4.375	0	8
18	3	0	3.33333	2.375	0	8
----- 중 간 생 략 -----						
571	1	0	3.55556	3.000	0	7
571	2	0	2.94444	1.875	0	7
571	3	0	3.50000	2.750	0	7

----- 중 간 생 략 -----



[자료 1.2.3] 표시행렬 자료와 버트행렬 자료

어느 가게 고객 $n = 12$ 명을 대상으로 한 통계조사에서 범주변수인 성별, 나이, 키에 따른 $p = 3$ 인 삼원분할표(최용석과 정광모, 2003, 8장).

세 범주변수(성별, 나이, 키)에 따른 삼원분할표			
성별	나이	키	
		크다	작다
남	중장년	1	3
	청소년	2	1
여	중장년	1	2
	청소년	1	1

남	여	중장년	청소년	키큼	키작음
1	0	1	0	1	0
1	0	1	0	0	1
1	0	1	0	0	1
1	0	1	0	0	1
1	0	0	1	1	0
1	0	0	1	1	0
1	0	0	1	0	1
0	1	1	0	1	0
0	1	1	0	0	1
0	1	1	0	0	1
0	1	0	1	1	0
0	1	0	1	0	1

$$Z = [Z_1 | Z_2 | Z_3] =$$

표시행렬(indicator matrix) 자료

$$B = Z^t Z = \begin{pmatrix} Z_1^t Z_1 & Z_1^t Z_2 & Z_1^t Z_3 \\ Z_2^t Z_1 & Z_2^t Z_2 & Z_2^t Z_3 \\ Z_3^t Z_1 & Z_3^t Z_2 & Z_3^t Z_3 \end{pmatrix}$$

	남	여	중장년	청소년	키큼	키작음
남	7	0	4	3	3	4
여	0	5	3	2	2	3
중장년	4	3	7	0	2	5
청소년	3	2	0	5	3	2
키큼	3	2	2	3	5	0
키작음	4	3	5	2	0	7

$B =$

버트행렬(Burt matrix) 자료

1.3 텍스트자료 Text Data

- 여러 분야의 기관 보고서나 인터넷 상의 기사 문서에 대해 문자로 된 수집된 **비정형화자료(unstructured data)**이다.
- 이러한 자료는 **문서-용어(document-term)** 정보를 측정하며 자료는 빈도(frequency)로 이루어진다.
 - 용어의 시각적 연관성 분석 : SCA, MCA
- 통계분석을 위해서는 **정형화자료(structured data)**로 전환이 필요
 - 근접성 데이터(proximity data)
 - 군집화를 위한 시각화 기법 : MDS
 - 군집분석

- 부산발전 연구원 Big Data 워크숍
- Visualizations of Big Data Using the Biplots with R.



- 19 -

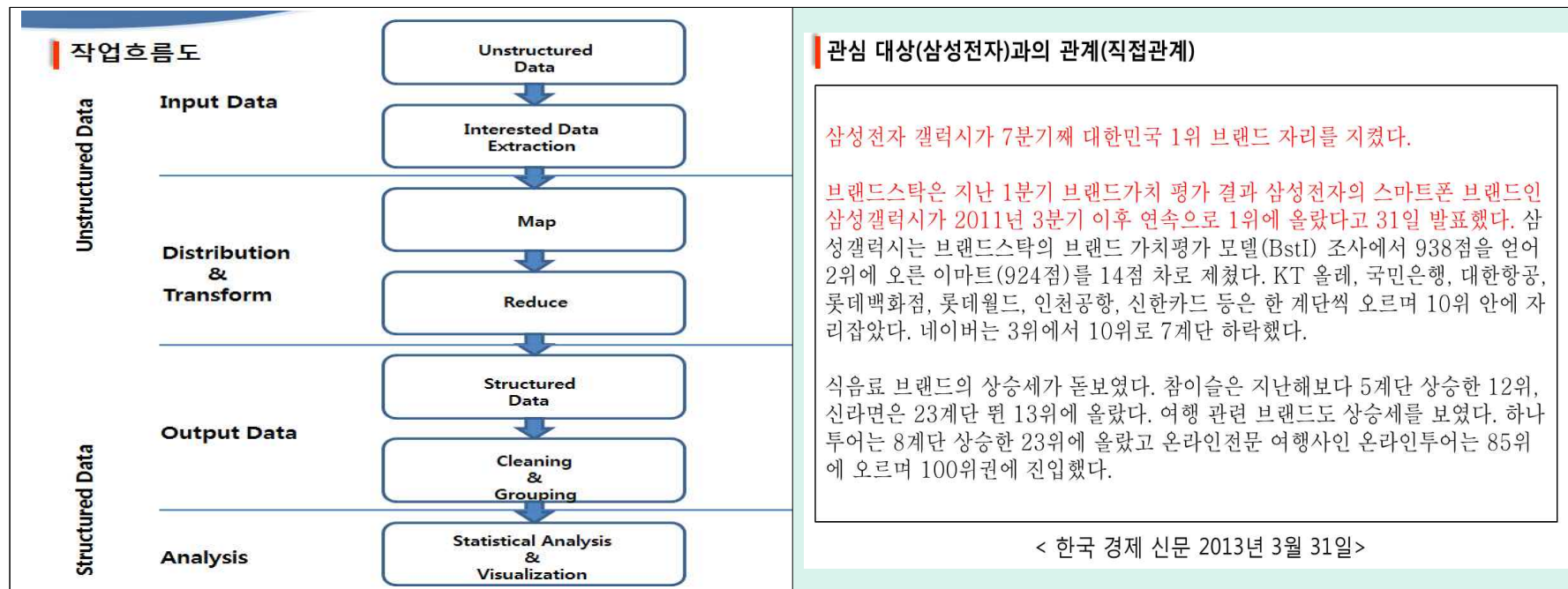
[자료 1.3.2] 비정형 텍스트 자료

- 삼성과 애플의 기사의 맵리듀스와 MCA를 활용한 동향 분석

-Choi, J. and Choi, Y.S. (2014). Standardizing Unstructured Big Data and Visual Interpretation Using MapReduce and Correspondence Analysis, *The Korean Journal of Applied Statistics*, 27(2), 169-183.

1) 단어 세분화 및 군집화 (Structured Data) : 두 기업의 분기별 동향과 이슈(2012.01-2012.12, 한국경제신문)

Hadoop Distributed Filesystem

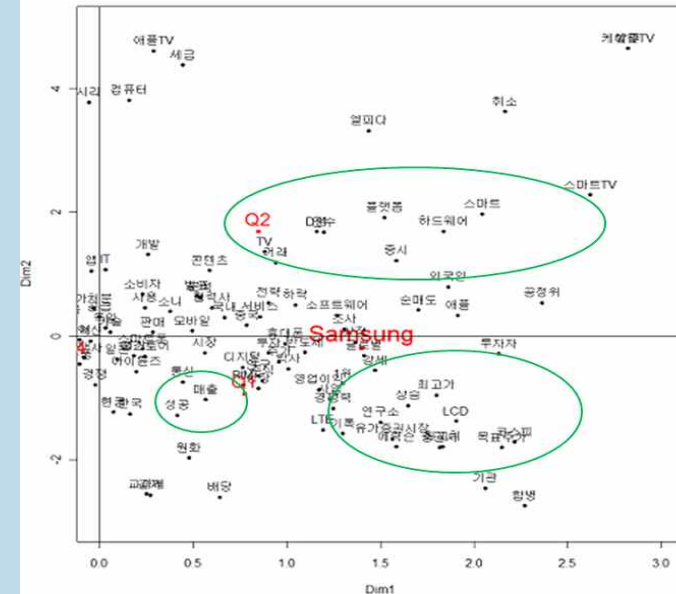
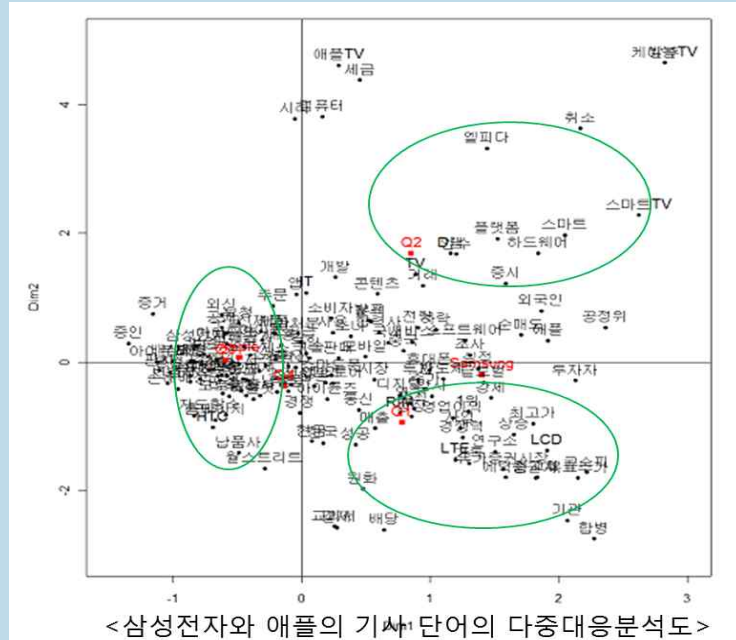


삼성과 애플 관련 기사 빈도 및 관련 단어 수

분기	1	2	3	4	총
삼성	61	58	47	40	206
단어수	2,769	2,407	2,395	2,060	9,631
애플	45	43	100	75	263
단어수	2,903	2,745	5,106	4,088	15,042
총 단어수	5,672	4,152	7,501	6,148	24,673

업체	빈도수 상위 5%	세분화 후 빈도수 상위 5% (분기별)
삼성	153	70
애플	186	105
삼성&애플	256	152

2) 대응분석(Correspondence Analysis)



기업	분기	사분면	단어	의미
Samsung (삼성전자)	Q1(1분기)	제4사분면	매출, 원화, 성공, 상승, 최고가, 영업이익, 1위, 목표주가, 교과서, 경쟁력, 코스피, 한국, 현금 등	성장
	Q2(2분기)	제1사분면	D램, 하드웨어, 스마트, 스마트TV, 애플TV, 콘텐츠, 개발, 모바일, 소프트웨어 등	제품, 사업확장
Apple (애플)	Q3(3분기)	제2사분면	외신, 증거, 증인, 삼성전자, 요청, 제소, 소송, 판매금지, 배심원 등	특허 소송
	Q4(4분기)	제4사분면	지도, 홈페이지, 손해배상, 변호, 제기, 경쟁사, 모토로라, 판결, 경쟁 등	특허 소송

- 현황분석 : 2012년

애플 : 특허 관련 소송 주력

삼성전자 : 사업 확장과 성장 주력

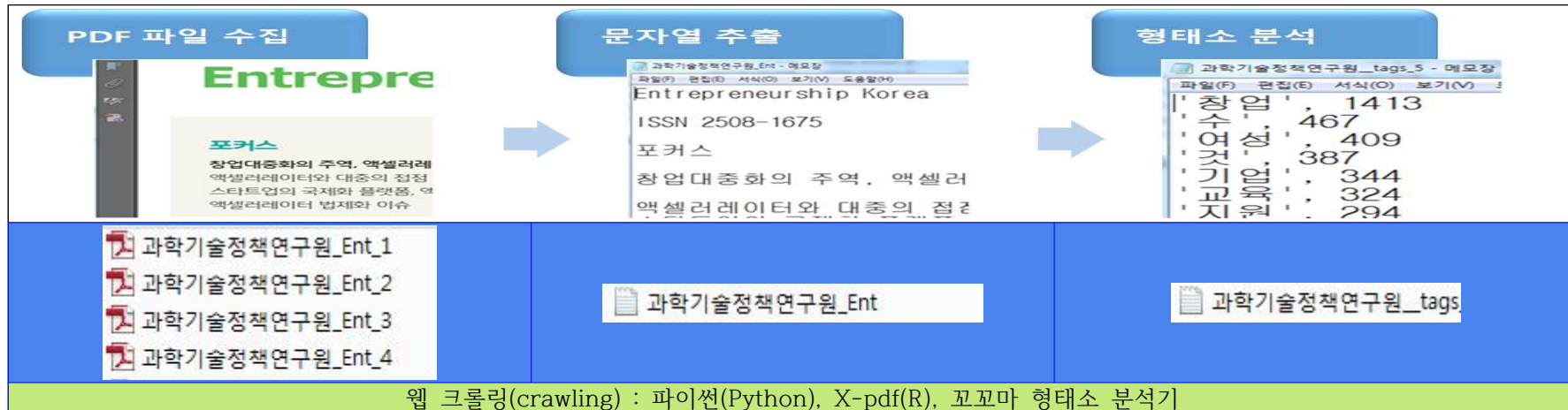
- 예측분석 : 2013년 상반기

삼성전자 : 지속적 성장과 사업 확장 -> 상승세 전망

애플 : 계속되는 특허 관련 소송 -> 슬로건: 혁신적 기업 이미지에 타격 -> 성장 주춤 전망

[자료 1.3.3] 비정형 텍스트 자료: 19개 정부 출연지원 연구기관(국가과학기술: 25, 경제, 인문 사회 26)

- 정민지(2017). A Study on **Clustering Methods** for Proximity Data in Text Mining,
- 정호영(2018). **Text classification** with various term weighting in Text Mining.
- 이수진(2019). **Document Classification** Using **Deep Neural Network** in Text Mining.



웹 크롤링(crawling) : 파이썬(Python), X-pdf(R), 꼬꼬마 형태소 분석기

> #2016년도 경제인문사회연구회 출연 10개 연구기관들의 정기간행물 발간 개수					가가 가게 가격 가격대 가공 가능 가능성 가동 가맹점 가연														
> govern m																			
기관명		기관명 정기간행물발간개수																	
[1,]	"과학기술정책연구원"	"A"	"12"		A01	1	1	9	3	2	36	7	4	1	1				
[2,]	"대외경제정책연구원"	"B"	"23"		A02	0	0	1	0	4	27	16	2	0	0				
[3,]	"산업연구원"	"C"	"12"		A03	0	0	1	0	1	14	14	4	0	0				
[4,]	"육아정책연구소"	"D"	"4"		A04	1	1	1	0	0	28	5	0	0	0				
[5,]	"정보통신정책연구원"	"E"	"11"		A05	0	0	5	0	3	41	37	0	0	0				
[6,]	"통일연구원"	"F"	"4"		A06	0	0	2	0	2	13	3	0	0	0				
[7,]	"한국개발연구원"	"G"	"12"		A07	0	0	2	1	1	50	30	0	0	0				
[8,]	"한국교육개발원"	"H"	"12"		A08	0	0	4	0	1	23	13	0	0	0				
[9,]	"한국교육과정평가원"	"I"	"4"																
[10,]	"한국교통연구원"	"J"	"12"																
10개 기관 문서의 발간 개수					343 × 117,820					343 × 26,915					문서-용어 빈도행렬				

TF-IDF(Term Frequency-Inverse Document Frequency)

- 용어 가중치 기법 중 가장 대표적으로 알려진 알고리즘
- 특정 문서 내에서 용어 빈도가 높을수록, 전체 문서들에는 그 용어를 포함한 문서가 적을수록 TF-IDF 값이 높아짐

$$w_{x,y} = tf_{x,y} \times \log \left(\frac{N}{df_x} \right)$$

- N : 총 문서의 수
- $tf_{x,y}$: y 번째 문서에서 x 번째 용어에 대한 빈도
- df_x : x 번째 용어가 포함된 문서의 수,

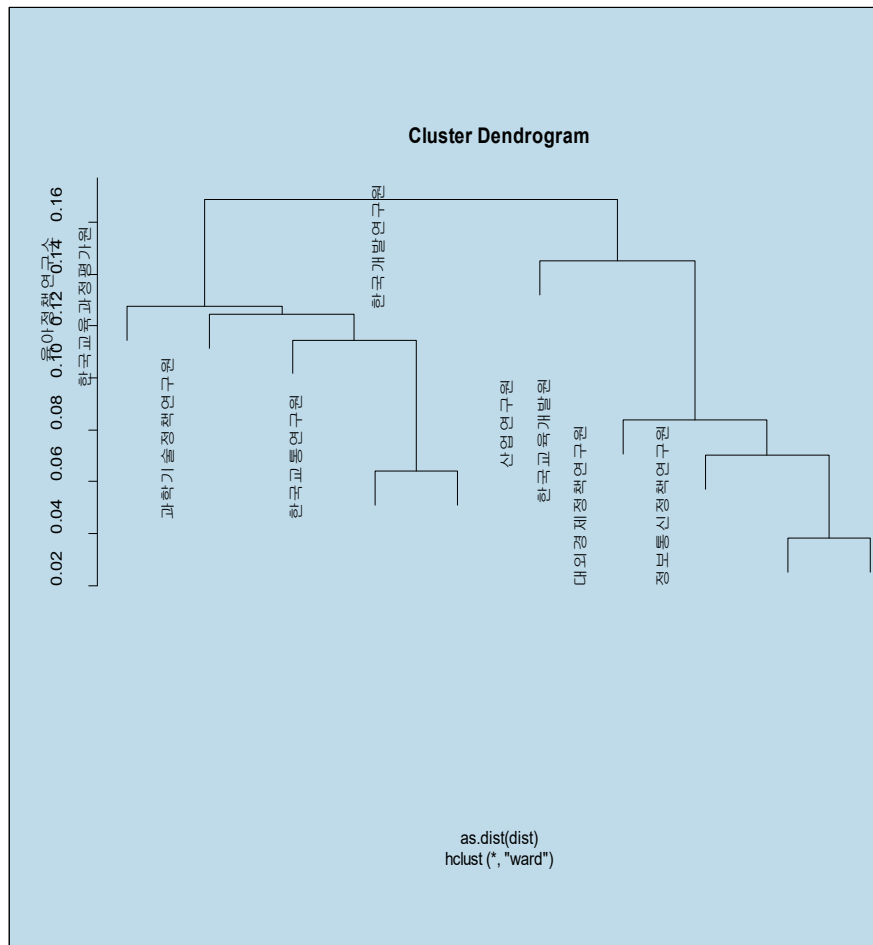
	가	가	가	가	가	가	가	가		가	가	가	가	가	가	가	가
	가	가	가	가	가	가	가	가		가	가	가	가	가	가	가	가
A01	2.599	3.493	8.667	13.775	2.942	7.251	1.586	8.107	A01	8.667	7.251	3.347	2.855	2.599	0.666	10.197	1.414
A02	0.000	0.000	0.963	0.000	5.883	5.438	3.624	4.054	A02	0.963	5.438	65.823	0.714	0.000	0.000	6.207	1.414
A03	0.000	0.000	0.963	0.000	1.471	2.820	3.171	8.107	A03	0.963	2.820	0.000	12.845	0.000	0.000	3.103	0.000
A04	2.599	3.493	0.963	0.000	0.000	5.640	1.133	0.000	A04	0.963	5.640	2.231	2.141	5.199	1.999	6.207	1.414
A05	0.000	0.000	4.815	0.000	4.413	8.258	8.382	0.000	A05	4.815	8.258	8.925	4.995	0.000	1.333	3.547	1.414
A06	0.000	0.000	1.926	0.000	2.942	2.618	0.680	0.000	A06	1.926	2.618	1.116	1.427	0.000	3.332	2.217	2.827
A07	0.000	0.000	1.926	4.592	1.471	10.071	6.796	0.000	A07	1.926	10.071	1.116	5.709	0.000	0.666	8.867	0.000
A08	0.000	0.000	3.852	0.000	1.471	4.633	2.945	0.000	A08	3.852	4.633	2.231	3.568	2.599	0.000	7.093	0.000
A09	2.599	0.000	5.778	0.000	1.471	4.028	4.984	2.027	A09	5.778	4.028	10.041	0.714	0.000	3.332	3.547	0.000
A10	0.000	0.000	0.963	0.000	2.942	3.021	1.359	10.134	A10	0.963	3.021	1.116	1.427	0.000	0.000	5.763	0.000
343 × 26,915 문서-용어 가중행렬(TF-IDF)									343 × 1,158 문서-핵심어 가중행렬								

> m3Data_CD																		
	A01	A02	A03	A04	A05	A06	A07	A08	A09	A10	A11	A12	B01	B02	B03	B04	B05	B06
A01	0.000000	3.206905	3.446756	3.135831	3.112932	3.425774	3.164416	3.239197	3.120057	3.624077	3.458488	3.243560	5.951523	6.060868	4.809896	6.102394	5.427850	6.851673
A02	3.206905	0.000000	3.447395	3.136805	3.249183	3.372051	3.121716	3.138945	3.078453	3.590831	3.373510	3.224884	5.888912	5.973620	4.765283	6.103320	5.337483	6.658822
A03	3.446756	3.447395	0.000000	3.235597	3.445325	3.479209	3.208754	3.371477	3.282020	3.629488	3.534736	3.306393	6.007731	6.026458	4.875460	6.174279	5.452535	6.920198
A04	3.135831	3.136805	3.235597	0.000000	3.150504	3.212510	2.943465	3.014069	2.964491	3.429519	3.248745	3.072847	5.837448	5.951662	4.716090	6.067601	5.327308	6.807162
A05	3.112932	3.249183	3.445325	3.150504	0.000000	3.330516	3.093575	3.207166	3.151760	3.604940	3.474772	3.318196	5.925739	6.015682	4.785269	6.002871	5.379061	6.837892
A06	3.425774	3.372051	3.479209	3.212510	3.330516	0.000000	3.148407	3.304958	3.220470	3.636024	3.513455	3.370602	5.984741	6.015938	4.869386	6.094700	5.445845	6.894796
A07	3.164416	3.121716	3.208754	2.943465	3.093575	3.148407	0.000000	3.041731	2.938896	3.408736	3.221334	3.112927	5.833370	5.936283	4.673042	6.040274	5.309467	6.765517
A08	3.239197	3.138945	3.371477	3.014069	3.207166	3.304958	3.041731	0.000000	3.029635	3.520874	3.372225	3.181459	5.840920	5.963889	4.647631	6.089577	5.389903	6.832247
A09	3.120057	3.078453	3.282020	2.964491	3.151760	3.220470	2.938896	3.029635	0.000000	3.390363	3.224326	3.002332	5.826901	5.899427	4.664486	6.042246	5.285818	6.761967
A10	3.624077	3.590831	3.629488	3.429519	3.604940	3.636024	3.408736	3.520874	3.390363	0.000000	3.671097	3.573772	6.115112	6.240692	4.957642	6.304374	5.650009	7.017427
A11	3.458488	3.373510	3.534736	3.248745	3.474772	3.513455	3.221334	3.372225	3.224326	3.671097	0.000000	3.407937	6.020923	6.114658	4.902031	6.188610	5.500348	6.898208
A12	3.243560	3.224884	3.306393	3.072847	3.318196	3.370602	3.112927	3.181459	3.002332	3.573772	3.407937	0.000000	5.959055	6.041648	4.763086	6.119602	5.441414	6.852891
B01	5.951523	5.888912	6.007731	5.837448	5.925739	5.984741	5.833370	5.840920	5.826901	6.115112	6.020923	5.959055	0.000000	7.623281	6.747794	7.731151	7.163997	8.347855
B02	6.060868	5.973620	6.026458	5.951662	6.015682	6.015938	5.936283	5.963889	5.899427	6.240692	6.114658	6.041648	7.623281	0.000000	6.779784	7.627627	6.545571	8.325572
B03	4.809896	4.765283	4.875460	4.716090	4.785269	4.869386	4.673042	4.647631	4.664486	4.957642	4.902031	4.763086	6.747794	6.779784	0.000000	6.711388	6.295994	7.588896
B04	6.102394	6.103320	6.174279	6.067601	6.002871	6.094700	6.040274	6.089577	6.042246	6.304374	6.188610	6.119602	7.731151	7.627627	6.711388	0.000000	7.068456	8.491877
B05	5.427850	5.337483	5.452535	5.327308	5.379061	5.445845	5.309467	5.389903	5.285818	5.650009	5.500348	5.441414	7.163997	6.545571	6.295994	7.068456	0.000000	7.975529
B06	6.851673	6.658822	6.920198	6.807162	6.837892	6.894796	6.765517	6.832247	6.761967	7.017427	6.898208	6.852891	8.347855	8.325572	7.588896	8.491877	7.975529	0.000000

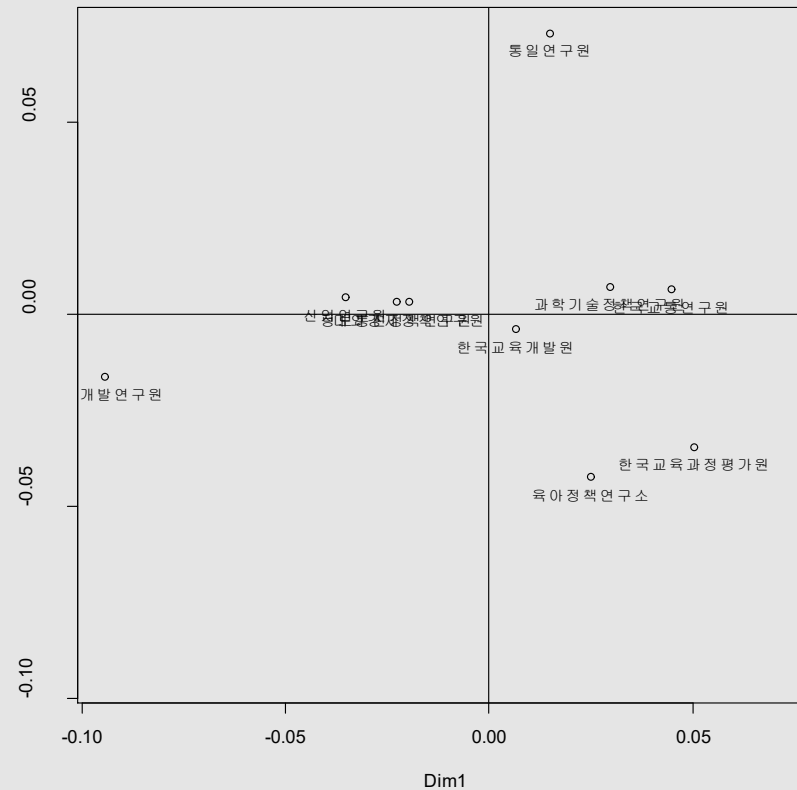
간행물 간의 유사성 : 유클리드거리, 가중유클리드거리

> round(dist,3)											
	과학기술정책연구원	대외경제정책연구원	산업연구원	육아정책연구소	정보통신정책연구원	통일연구원	한국개발연구원	한국교육개발원	한국교육과정평가원	한국교통연구원	
과학기술정책연구원		0.000	0.077	0.100	0.102	0.084	0.099	0.140	0.066	0.099	0.064
대외경제정책연구원	0.077		0.000	0.063	0.101	0.038	0.097	0.107	0.061	0.107	0.097
산업연구원	0.100	0.063		0.000	0.122	0.068	0.121	0.121	0.090	0.132	0.118
육아정책연구소	0.102	0.101	0.122		0.000	0.105	0.130	0.152	0.086	0.125	0.111
정보통신정책연구원	0.084	0.038	0.068	0.105		0.000	0.103	0.111	0.064	0.114	0.103
통일연구원	0.099	0.097	0.121	0.130	0.103		0.000	0.150	0.095	0.132	0.110
한국개발연구원	0.140	0.107	0.121	0.152	0.111	0.150		0.000	0.120	0.161	0.152
한국교육개발원	0.066	0.061	0.090	0.086	0.064	0.095	0.120		0.000	0.100	0.081
한국교육과정평가원	0.099	0.107	0.132	0.125	0.114	0.132	0.161	0.100		0.000	0.105
한국교통연구원	0.064	0.097	0.118	0.111	0.103	0.110	0.152	0.081	0.105		0.000

기관별 유사성 거리 : 기관 내와 기관 간의 거리 평균



계층적군집분석의 덴드로그램 : 와드연결법



다차원 척도법(Multidimensional Scaling)

SVM(Support Vector Machine) 활용한 판별분석

- 문서-핵심어 가중행렬을 바탕으로 을 이용하여 10개 기관 문서의 텍스트 분류.
- train:test = 80:20의 비율로 기관별 문서에 대해 증화추출.

예측	실제									
	A	B	C	D	E	F	G	H	I	J
A	3	0	0	0	0	0	0	0	0	0
B	0	4	0	0	0	0	0	0	0	0
C	0	0	3	0	0	0	0	0	0	0
D	0	0	0	3	0	0	0	0	0	0
E	0	0	0	0	3	0	0	0	0	0
F	0	0	0	0	0	2	0	0	0	0
G	0	0	0	0	0	0	3	0	0	0
H	0	1	0	0	0	0	0	26	0	1
I	0	0	0	0	0	0	0	0	8	0
J	0	0	0	0	0	0	0	0	0	5

Accuracy : 0.9677 = 60/62

서포트 벡터 머신에 의한 10개 기관 문서의 분류표

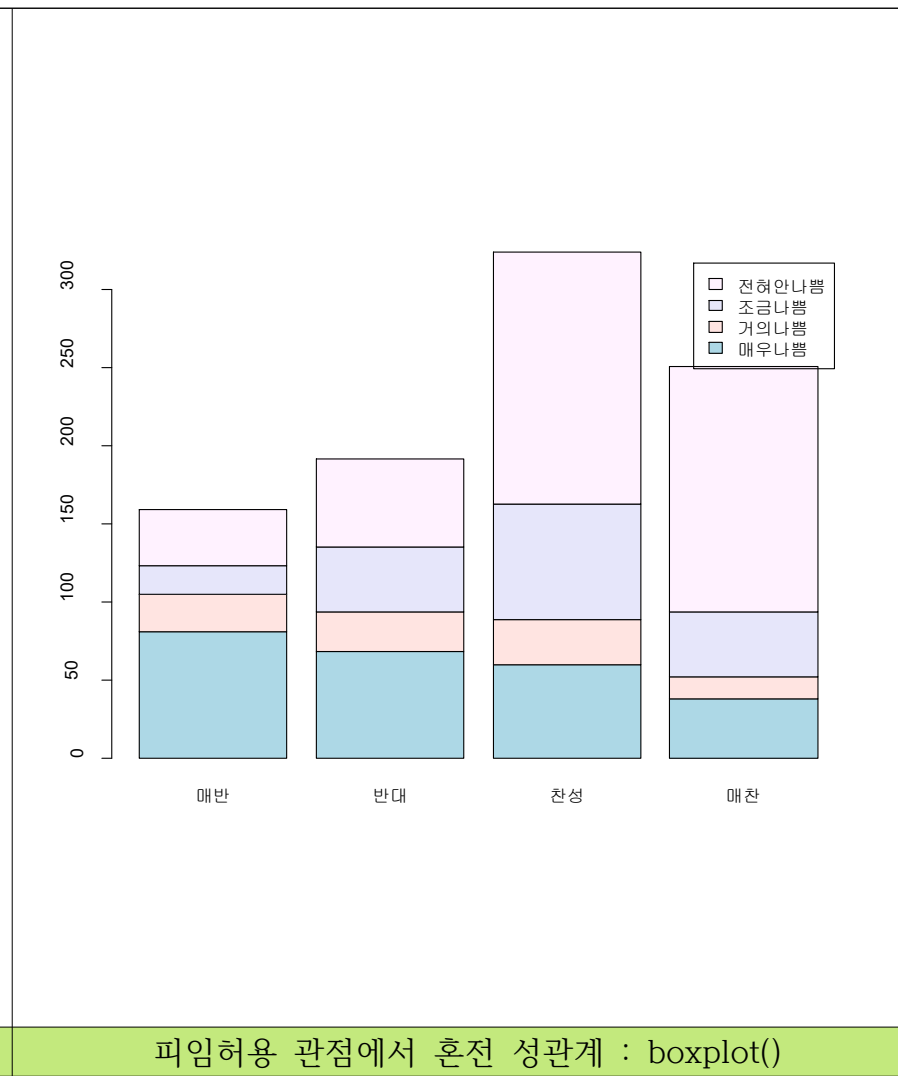
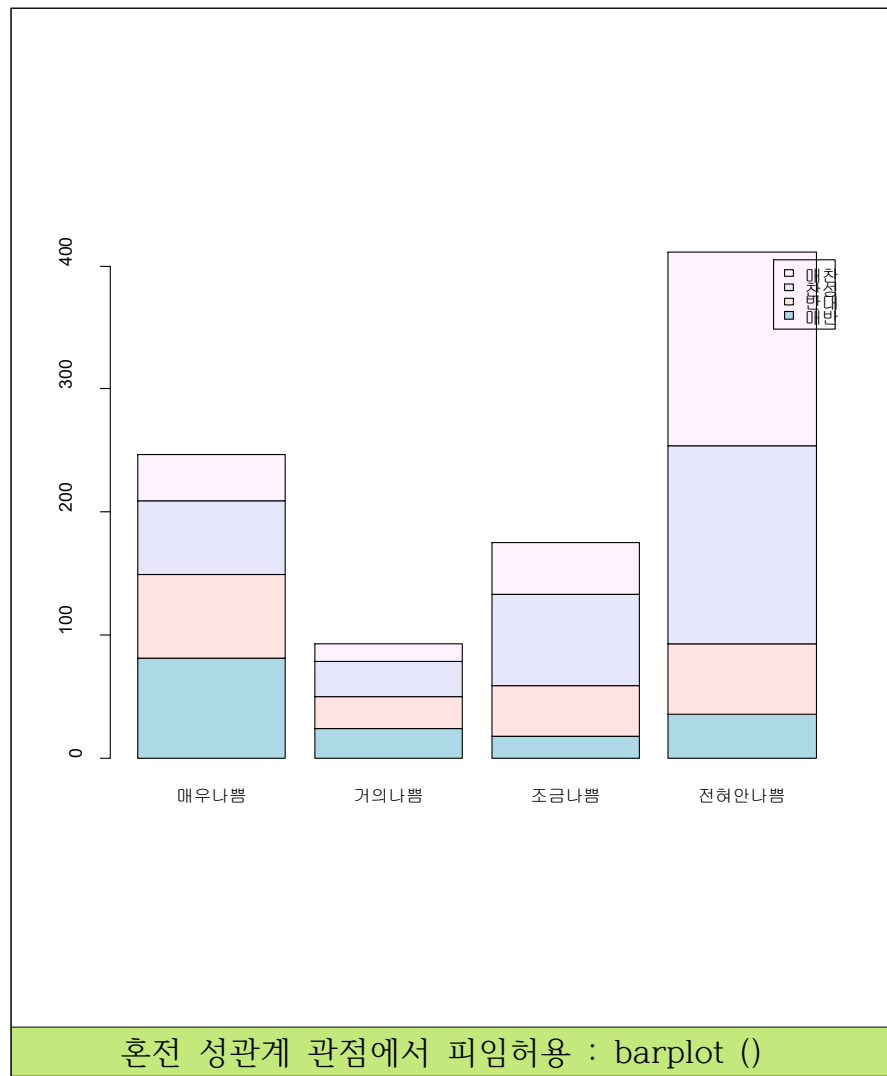
주제 2 : 범주형자료의 시각화 (Visualizations for CDA)

- 2.1 막대 그림
- 2.2 모자이크 그림
- 2.3 단순 대응분석 그림
- 2.4 다중 대응분석 그림
- 2.5 사회네트워크 분석
- 2.6 워드 클라우드

2.1 막대 그림 Bar Plot

이원분할표 Tow-Way Contingency Table

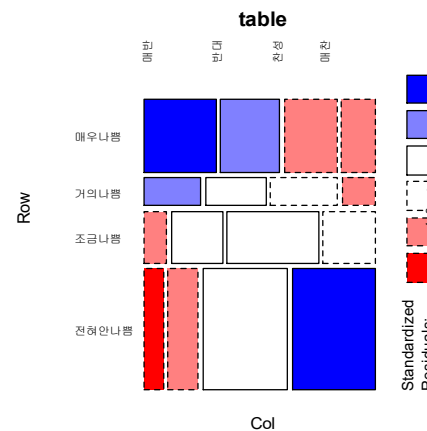
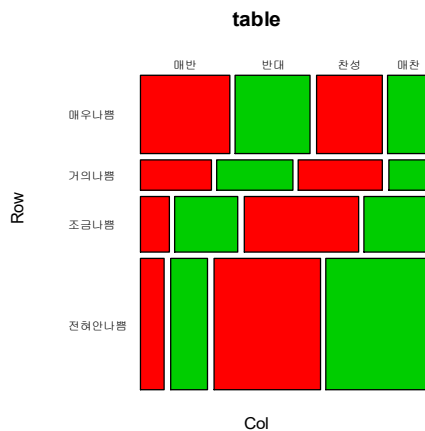
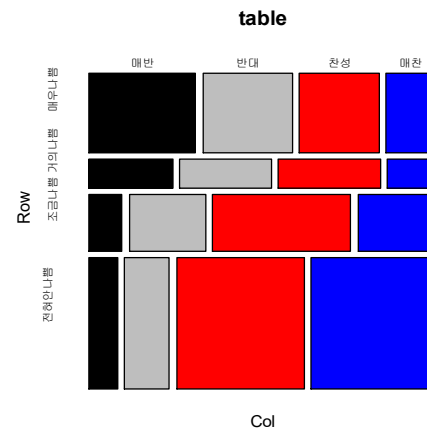
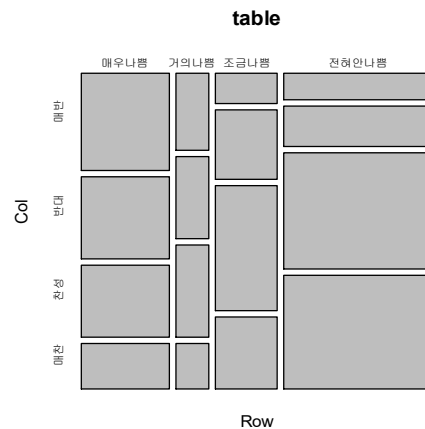
presexfreq.txt	presexfreq-barplot.R																													
<table><tr><th rowspan="2">혼전 성관계</th><th colspan="4">10대 임신조절</th></tr><tr><th>매우 반대</th><th>반 대</th><th>찬 성</th><th>매우 찬성</th></tr><tr><td>매우 나쁘다</td><td>81</td><td>68</td><td>60</td><td>38</td></tr><tr><td>거의 나쁘다</td><td>24</td><td>26</td><td>29</td><td>14</td></tr><tr><td>조금 나쁘다</td><td>18</td><td>41</td><td>74</td><td>42</td></tr><tr><td>전혀 나쁘지 않다</td><td>36</td><td>57</td><td>161</td><td>157</td></tr></table>	혼전 성관계	10대 임신조절				매우 반대	반 대	찬 성	매우 찬성	매우 나쁘다	81	68	60	38	거의 나쁘다	24	26	29	14	조금 나쁘다	18	41	74	42	전혀 나쁘지 않다	36	57	161	157	<pre># Simple CRA : Two-Way Table O<-matrix(c(81, 68, 60, 38, 24, 26, 29, 14, 18, 41, 74, 42, 36, 57, 161, 157), byrow=T, nrow=4) rownames(O) <- c("매우나쁨", "거의나쁨", "조금나쁨", "전혀안 나쁨") colnames(O) <- c("매반", "반대", "찬성", "매찬") col=c("lightblue", "mistyrose", "lavender", "#FFF2FFFF") # Bar Plot barplot(t(O), legend=colnames(O), col=col) barplot(O, legend=rownames(O), col=col)</pre>
혼전 성관계		10대 임신조절																												
	매우 반대	반 대	찬 성	매우 찬성																										
매우 나쁘다	81	68	60	38																										
거의 나쁘다	24	26	29	14																										
조금 나쁘다	18	41	74	42																										
전혀 나쁘지 않다	36	57	161	157																										



2.2 모자이크그림 Mosaic Plot

2.2.1 이원분할표 Tow-Way Contingency Table

presexfreq.txt		presexfreq-mosaic.R	
[자료 2.1.1] 결혼 전 성관계에 대한 반응과 피임허용의 이원분할표		<pre> setwd("C:/빅데이터/R_data") freq<-read.table("presexfreq.txt") F<-as.matrix(freq) row<-c("매우나쁨", "거의나쁨", "조금나쁨", "전혀안나쁨") col<-c("매반", "반대", "찬성", "매찬") data<-expand.grid(Row=row, Col=col) data<-cbind(data, F) table<-xtabs(F~Row+Col, data=data) # Mosaicplot par(mfrow=c(2,2)) mosaicplot(table) mosaicplot(table, color=F, dir=c("h", "v")) mosaicplot(table, color= 2:3, dir=c("h", "v"), las=1) mosaicplot(table, shade=T, dir=c("h", "v"), las=2) </pre>	
혼전 성관계	10대 임신조절		
	매우 반대 반 대 찬 성 매우 찬성		
매우 나쁘다	81 68 60 38		
거의 나쁘다	24 26 29 14		
조금 나쁘다	18 41 74 42		
전혀 나쁘지 않다	36 57 161 157		



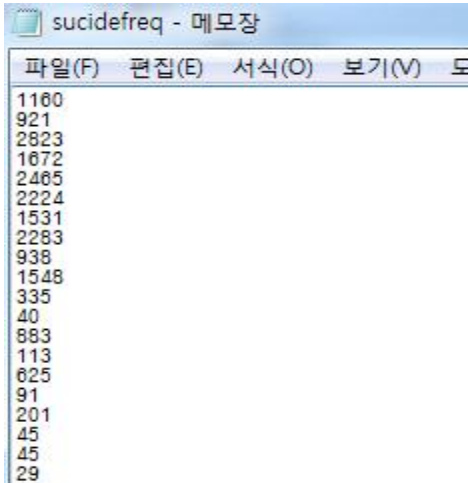
```
> # Chi-square test of independence
> summary(table)
Call: xtabs(formula = F ~ Row + Col, data = data)
Number of cases in table: 926
Number of factors: 2
Test for independence of all factors:
    Chisq = 128.68, df = 9, p-value = 2.212e-23
```

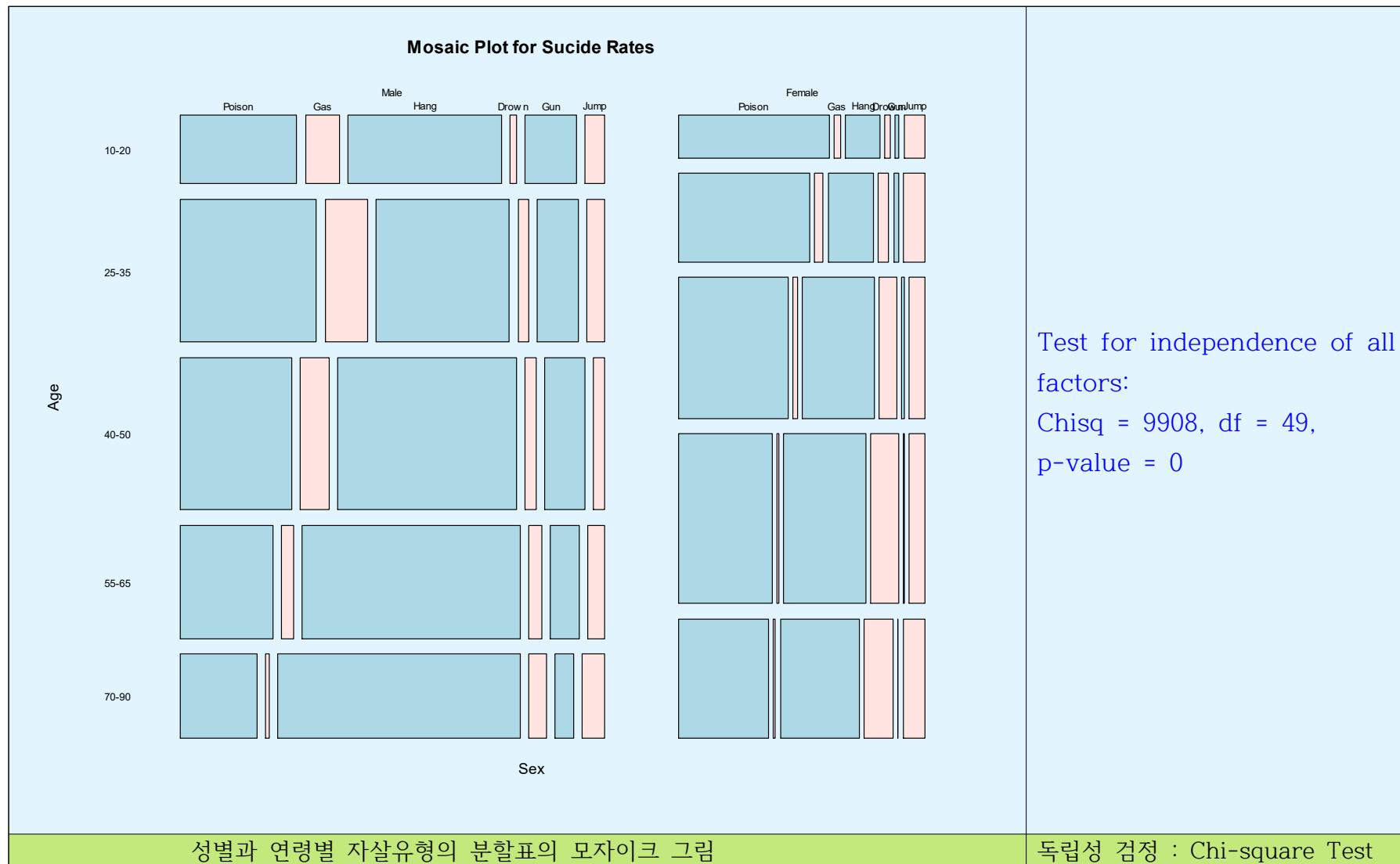
모자이크 그림 : mosaicplot()

독립성 검정 : Chi-square Test

2.2.2 삼원분할표 Three-Way Contingency Table

sucidefreq.txt								sucidefreq-mosaicplot.R
[자료 0.2.1] 2×5×6 성별과 연령별 자살유형의 분할표								<pre> setwd("C:/빅데이터/R_data") Data0.2.1<-read.table("sucidefreq.txt") freq<-as.matrix(Data0.2.1) sex<-c("Male", "Female") age<- c("10-20", "25-35", "40-50", "55-65", "70-90") method<-c("Poison", "Gas", "Hang", "Drown", "Gun", "Jump") data<-expand.grid(Sex=sex, Age=age, Method=method) data<-cbind(data, freq) # Makeing a 3-way table table<-xtabs(freq~Sex+Age+Method, data=data) # Print table ftable(table) # Chi-square test of indepedence summary(table) win.graph() # Mosaic Plot mosaicplot(table, dir=c("h", "v"), color=c("lightblue", "mistyrose"), las=1, main="Mosaic Plot for Sucide Rates") </pre>
Sex	Age	Poison	Gas	Hang	Drown	Gun	Jump	
M	10-20	1160	335	1524	67	512	189	
M	25-35	2823	883	2751	213	852	366	
M	40-50	2465	625	3936	247	875	244	
M	55-65	1531	201	3581	207	477	273	
M	70-90	938	45	2948	212	229	268	
F	10-20	921	40	212	30	25	131	
F	25-35	1672	113	575	139	64	276	
F	40-50	2224	91	1481	354	52	327	
F	55-65	2283	45	2014	679	29	388	
F	70-90	1548	29	1355	501	3	383	





2.3 단순 대응분석 그림 Simple Correspondence Analysis Map

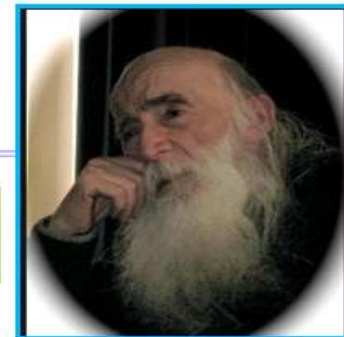
- **Definition**

- Technique for displaying the associations among the rows and columns of contingency table in a type of scatterplot or map
- In general, chi-square statistic providing a testing for the independence bt rows and columns
 - In addition, CRA accounts more clearly for the relationship bt rows and columns



- **Applications**

- Educational Psychology, Medicine, Survey, Social Science, etc.
- Categorical Data Analysis, **Text Mining, Big Data**



Jean-Paul Benzécri is a [French statistician](#) and was born in 1932. He is most famous for the development of the [Correspondence analysis](#), a statistical technique for analyzing [contingency tables](#). It is widely used in sociological studies of categorical data. (Wikipedia)

대응분석의 종류 : 단순 대응분석과 다중 대응분석

Types of CRA : Simple CRA and Multiple CRA

- Simple CRA is the CRA of the two-way contingency table

3 x 3 Table for the preference of the sizes of car in 3 countries

	Large	Medium	Small
American	37	61	30
European	4	20	21
Japanese	2	61	102

- Multiple CRA is the CRA of the multi-way contingency table

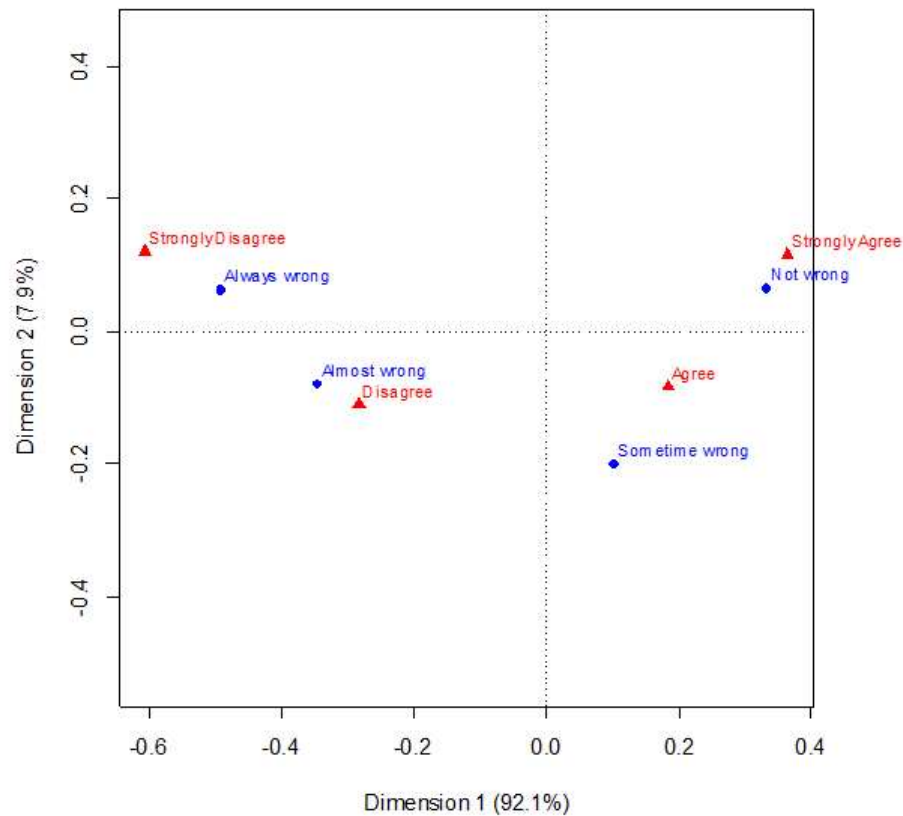
[Data 8.4.1] 2 x 2 x 2 Table for customers

Sex	Age	Height	
		tall	small
Male	old	1	3
	young	2	1
Female	old	1	2
	young	1	1

2.3.1 지위와 흡연 습관의 연관성

data(smoke)						smoke-SCRA.R																																																
<table><tr><th colspan="6">Contingency Table</th></tr><tr><th></th><th>none</th><th>light</th><th>medium</th><th>heavy</th><th>Sum</th></tr><tr><th>SM</th><td>4</td><td>2</td><td>3</td><td>2</td><td>11</td></tr><tr><th>JM</th><td>4</td><td>3</td><td>7</td><td>4</td><td>18</td></tr><tr><th>SE</th><td>25</td><td>10</td><td>12</td><td>4</td><td>51</td></tr><tr><th>JE</th><td>18</td><td>24</td><td>33</td><td>13</td><td>88</td></tr><tr><th>SC</th><td>10</td><td>6</td><td>7</td><td>2</td><td>25</td></tr><tr><th>Sum</th><td>61</td><td>45</td><td>62</td><td>25</td><td>193</td></tr></table>						Contingency Table							none	light	medium	heavy	Sum	SM	4	2	3	2	11	JM	4	3	7	4	18	SE	25	10	12	4	51	JE	18	24	33	13	88	SC	10	6	7	2	25	Sum	61	45	62	25	193	<pre># Simple CRA : smoker data library(ca) data(smoke) O=smoke sca=ca(O) sca win.graph() par(pty="s") plot(sca, main="SCRA : 지위와 흡연습관 분할표")</pre>
Contingency Table																																																						
	none	light	medium	heavy	Sum																																																	
SM	4	2	3	2	11																																																	
JM	4	3	7	4	18																																																	
SE	25	10	12	4	51																																																	
JE	18	24	33	13	88																																																	
SC	10	6	7	2	25																																																	
Sum	61	45	62	25	193																																																	
Staff group SM: Senior managers JM: Junior managers SE: Senior employees JE: Junior employees SC: Secretaries																																																						
Smoking category None : does not smoke Light : smokes 1-10 cigarettes a day Medium : smokes 11-20 cigarettes a day Heavy : smokes more than 20 cigarettes a day																																																						

SCRA package ca : Two-Way Table



- Relative positions of the points : similarities and differences among rows and columns categories.

Most negative:

Always wrong - Strongly Disagree

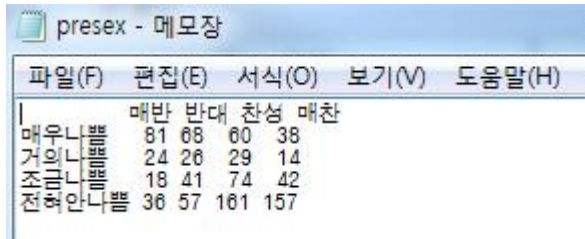
Most positive:

Not wrong - Strongly Agree

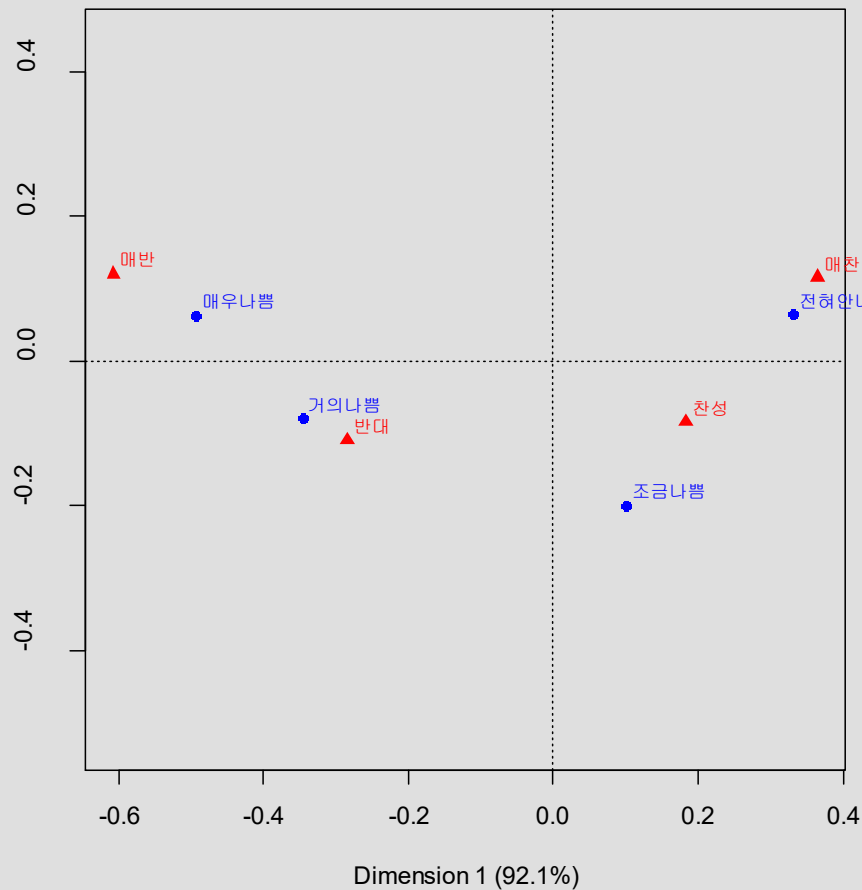
단순 대응분석 그림 : library(ca), ca()

해석

2.3.2 혼전 성관계와 피임허용 연관성

presex.txt	presex-SCRA.R																																		
[자료 2.1.1] 결혼 전 성관계에 대한 반응과 피임허용의 이원분할표 <table><thead><tr><th rowspan="2">혼전 성관계</th><th colspan="4">10대 임신조절</th></tr><tr><th>매우 반대</th><th>반</th><th>대</th><th>찬성</th><th>매우 찬성</th></tr></thead><tbody><tr><td>매우 나쁘다</td><td>81</td><td>68</td><td></td><td>60</td><td>38</td></tr><tr><td>거의 나쁘다</td><td>24</td><td>26</td><td></td><td>29</td><td>14</td></tr><tr><td>조금 나쁘다</td><td>18</td><td>41</td><td></td><td>74</td><td>42</td></tr><tr><td>전혀 나쁘지 않다</td><td>36</td><td>57</td><td></td><td>161</td><td>157</td></tr></tbody></table> 	혼전 성관계	10대 임신조절				매우 반대	반	대	찬성	매우 찬성	매우 나쁘다	81	68		60	38	거의 나쁘다	24	26		29	14	조금 나쁘다	18	41		74	42	전혀 나쁘지 않다	36	57		161	157	<pre># Simple CRA ca() : Matrix for Two-Way Table O<-matrix(c(81, 68, 60, 38, 24, 26, 29, 14, 18, 41, 74, 42, 36, 57, 161, 157), byrow=T, nrow=4) rownames(O)<-c("매우나쁨", "거의나쁨", "조금나쁨", "전혀안나쁨") colnames(O)<-c("매반", "반대", "찬성", "매찬") library(ca) sca=ca(O) par(pty="s") plot(sca, main="SCRA package ca : 이원분할표") # Simple CRA ca() : Text for Two-Way Table setwd("C:/2017빅데이터/R_data") O<-read.table("presex.txt", header=T) library(ca) sca=ca(O) sca win.graph() par(pty="s") plot(sca, main="SCRA package ca : 이원분할표")</pre>
혼전 성관계		10대 임신조절																																	
	매우 반대	반	대	찬성	매우 찬성																														
매우 나쁘다	81	68		60	38																														
거의 나쁘다	24	26		29	14																														
조금 나쁘다	18	41		74	42																														
전혀 나쁘지 않다	36	57		161	157																														

SCRA의 행범주 ca :



- 수평축(Dim1)에 대하여 왼편의 행범주 (매우 나쁨, 거의나쁨)은 열범주 (매반, 반대)방향으로 놓여져 있다. 이는 **혼전 성관계가 나쁘다고** 여기는 사람들은 **10대에 피임을 반대**하는 입장에 있음을 나타낸다.

- 오른편의 행범주 (조금나쁨, 전혀안나쁨)은 열범주 (찬성, 매찬) 방향으로 놓여져 있어 이를 서로 대응한다고 한다. 이는 **혼전 성관계에 대해 부정적이지 않은** 사람들은 **10대에 피임을 찬성**하고 있다.

- **혼전 성경험에 대해서도 관대한** 사람들은 **10대에게 피임을 허용**해야 한다고 좀 더 강하게 느끼는 입장을 보이고 있다.

단순 대응분석 그림 : library(ca), ca()

해석

2.3.3 단순 대응분석 알고리즘

[1단계] $n \times p$ 이원분할표 자료행렬: $O = (o_{ij})$, $o_{ij} \geq 0$, $i = 1, \dots, n$; $j = 1, \dots, p$

[2단계] 대응행렬: $F = (f_{ij})$, $f_{ij} = \frac{o_{ij}}{o_{..}}$, $i = 1, \dots, n$; $j = 1, \dots, p$

- $o_{..} = \sum_{i=1}^n \sum_{j=1}^p o_{ij}$: O 의 모든 원소들의 총합
- 행과 열 프로파일의 중심: $c = F^t \mathbf{1}_n$, $r = F \mathbf{1}_p$

$\Rightarrow$ 중심화 대응행렬: $\bar{F} = F - rc^t$

[3단계] 잔차행렬의 비정칙값분해: $Y = D_r^{-1/2} \bar{F} D_c^{-1/2} = U \Lambda V^t$

- $D_r = \text{diag}(f_{1.}, \dots, f_{n.})$, 행주변비율의 합 $f_{i.} = \sum_{j=1}^p f_{ij}$, $i = 1, \dots, n$
- $D_c = \text{diag}(f_{.1}, \dots, f_{.p})$, 열주변비율의 합 $f_{.j} = \sum_{i=1}^n f_{ij}$, $j = 1, \dots, p$
- $\Lambda = \text{diag}(\lambda_1, \dots, \lambda_p)$: Y 의 비정칙값 $\lambda_1, \dots, \lambda_p$ 에 의한 대각행렬
- U 와 V 는 각각 $U^t U = I_p$ 와 $V^t V = I_p$ 를 만족한다.

[4단계] 단순 CRA 그림의 행과 열 좌표점:

$$A = D_r^{-1/2} U \Lambda, \quad B = D_c^{-1/2} V \Lambda$$

$\Rightarrow s$ 차원의 단순 CRA 그림의 행과 열 좌표점:

$$A_{(s)} = D_{r(s)}^{-1/2} U_{(s)} \Lambda_{(s)}, \quad B_{(s)} = D_{c(s)}^{-1/2} V_{(s)} \Lambda_{(s)}$$

[5단계] $s (\geq 2)$ 차원의 단순 CRA 그림의 설명력:

$$fit = \frac{\sum_{k=1}^s \lambda_k^2}{\sum_{k=1}^{p-1} \lambda_k^2} \times 100$$

2.4 다중 대응분석 그림 Multiple Correspondence Analysis Map

2.4.1 삼원분할표 Three-Way Contingency Table

sucidefreq.txt								sucidefreq-MCRA.R
[자료 0.2.1] 2×5×6 성별과 연령별 자살유형의 분할표								
Sex	Age	Poison	Gas	Hang	Drown	Gun	Jump	
M	10-20	1160	335	1524	67	512	189	setwd("C:/빅데이터/R_data")
M	25-35	2823	883	2751	213	852	366	Data0.2.1<-read.table("sucidefreq.txt")
M	40-50	2465	625	3936	247	875	244	Freq<-as.matrix(Data0.2.1)
M	55-65	1531	201	3581	207	477	273	
M	70-90	938	45	2948	212	229	268	
F	10-20	921	40	212	30	25	131	sex<-c("Male", "Female")
F	25-35	1672	113	575	139	64	276	age<- c("10-20", "25-35", "40-50", "55-65", "70-90")
F	40-50	2224	91	1481	354	52	327	method<-c("Poison", "Gas", "Hang", "Drown", "Gun", "Jump")
F	55-65	2283	45	2014	679	29	388	data<-expand.grid(Sex=sex, Age=age, Method=method)
F	70-90	1548	29	1355	501	3	383	data<-cbind(data, Freq)

sucidefreq - 메모장

파일(F) 편집(E) 서식(O) 보기(V) 도

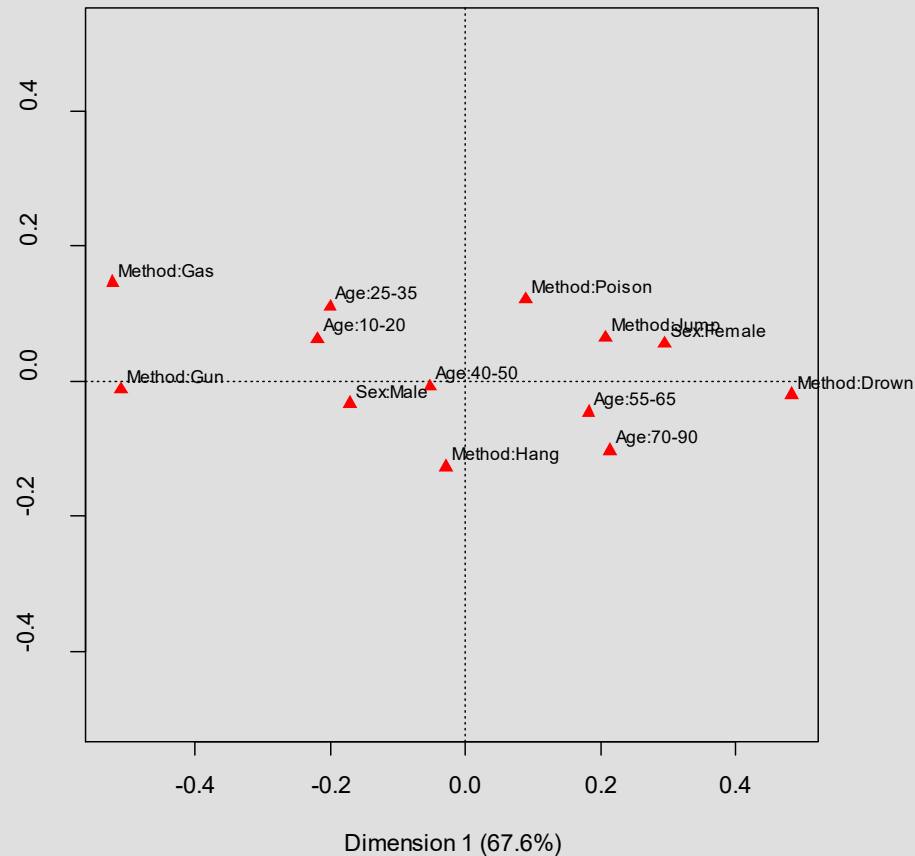
1160
921
2823
1672
2465
2224
1531
2283
938
1548
335
40
883
113
625
91
201
45
45
29

```

# Makeing a 3-way table
table<-xtabs(Freq~Sex+Age+Method, data=data)
plot(mjca(table), main="MCRA : 성별, 나이, 자살유형의
삼원분할표")

```

삼원 분할표의



- 수평축(제1 축) : - 연령별(Age): 왼편에 10-20대와 25-35대가 같은 위치, 서로 대응하여 동질하며 40-50대는 다소 원점에 가깝지만 이들과 같이 대응. 오른편에 고 연령대(55-65, 70-90)가 대응

- 성별(Sex) : 제1축의 왼편 남자(Male), 오른편 여자(Female)가 있어 서로 동질하지 않음

- 자살유형(Method) : 제1축 왼편 Gun(총기)와 Gas(가스)가 있고 Hang(목매기)도 이들과 대응하나 뚜렷하지는 않다. 오른편 Poison(독약), Jump(뛰어내림), Drown(익사)가 서로 대응.

- 전체적 경향: 나이가 젊은 남성들은 자살 유형으로 Gun(총기)와 Gas(가스)가 있고 Hang(목매기)를 선택하고 나이가 많은 여성들은 Poison(독약), Jump(뛰어내림), Drown(익사)의 자살유형을 선택.

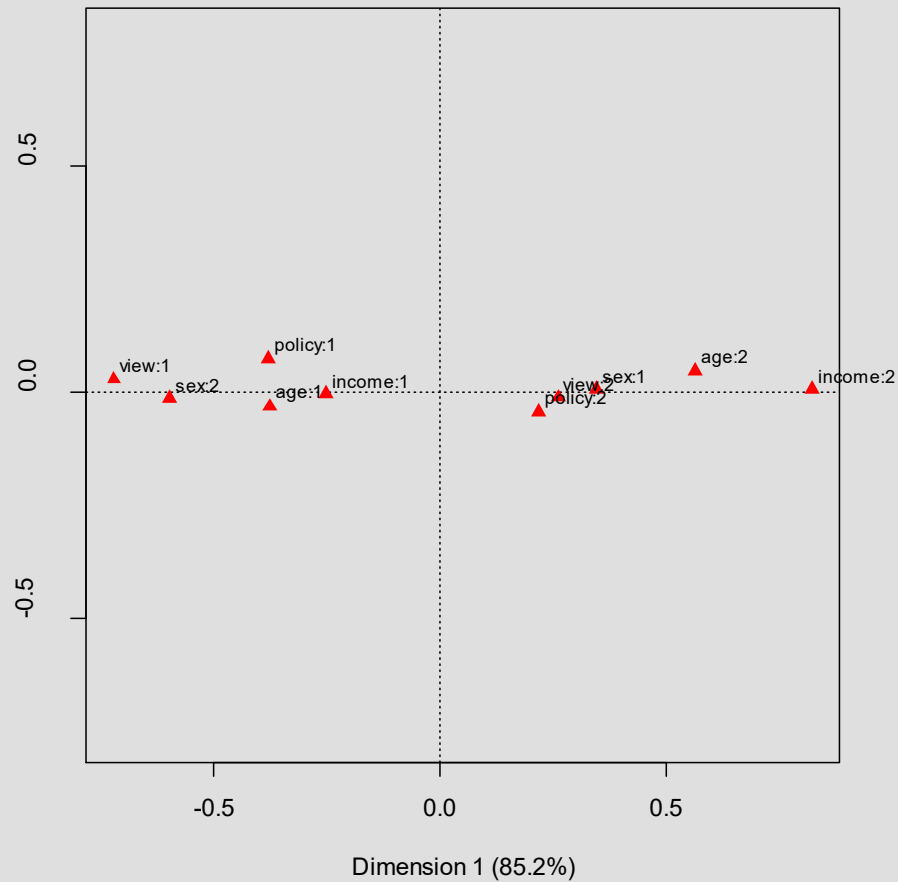
삼원 분할표의 다중 대응분석 그림 : library(ca), mjca()

해 석

2.4.2 오원 분할표 Five-Way Contingency Table

view.txt	view-MCRA.R																																																																																																																																																																																																																																					
[자료 1.4.1] 경제전망과 경제정책선택 <table><thead><tr><th>id</th><th>sex</th><th>age</th><th>income</th><th>view</th><th>policy</th></tr></thead><tbody><tr><td>1</td><td>1</td><td>2</td><td>2</td><td>2</td><td>2</td></tr><tr><td>2</td><td>1</td><td>1</td><td>1</td><td>1</td><td>2</td></tr><tr><td>3</td><td>2</td><td>1</td><td>1</td><td>1</td><td>1</td></tr><tr><td>4</td><td>1</td><td>2</td><td>2</td><td>2</td><td>2</td></tr><tr><td>5</td><td>2</td><td>1</td><td>1</td><td>2</td><td>2</td></tr><tr><td>6</td><td>1</td><td>2</td><td>1</td><td>2</td><td>1</td></tr><tr><td>7</td><td>1</td><td>1</td><td>1</td><td>2</td><td>2</td></tr><tr><td>8</td><td>1</td><td>2</td><td>2</td><td>2</td><td>2</td></tr><tr><td>9</td><td>1</td><td>1</td><td>1</td><td>1</td><td>2</td></tr><tr><td>10</td><td>2</td><td>1</td><td>1</td><td>1</td><td>1</td></tr></tbody></table> ----- 중 간 생 략 ----- <table><tbody><tr><td>115</td><td>2</td><td>1</td><td>1</td><td>1</td><td>1</td></tr><tr><td>116</td><td>1</td><td>2</td><td>2</td><td>2</td><td>2</td></tr><tr><td>117</td><td>2</td><td>1</td><td>1</td><td>2</td><td>2</td></tr><tr><td>118</td><td>1</td><td>2</td><td>1</td><td>2</td><td>1</td></tr><tr><td>119</td><td>1</td><td>1</td><td>1</td><td>2</td><td>2</td></tr><tr><td>120</td><td>2</td><td>1</td><td>1</td><td>1</td><td>1</td></tr></tbody></table>	id	sex	age	income	view	policy	1	1	2	2	2	2	2	1	1	1	1	2	3	2	1	1	1	1	4	1	2	2	2	2	5	2	1	1	2	2	6	1	2	1	2	1	7	1	1	1	2	2	8	1	2	2	2	2	9	1	1	1	1	2	10	2	1	1	1	1	115	2	1	1	1	1	116	1	2	2	2	2	117	2	1	1	2	2	118	1	2	1	2	1	119	1	1	1	2	2	120	2	1	1	1	1	view - 메모장 <table><thead><tr><th>파일(F)</th><th>편집(E)</th><th>서식(O)</th></tr></thead><tbody><tr><td colspan="3">id sex age income view policy</td></tr><tr><td>1</td><td>1</td><td>2</td><td>2</td><td>2</td><td>2</td></tr><tr><td>2</td><td>1</td><td>1</td><td>1</td><td>1</td><td>2</td></tr><tr><td>3</td><td>2</td><td>1</td><td>1</td><td>1</td><td>1</td></tr><tr><td>4</td><td>1</td><td>2</td><td>2</td><td>2</td><td>2</td></tr><tr><td>5</td><td>2</td><td>1</td><td>1</td><td>2</td><td>2</td></tr><tr><td>6</td><td>1</td><td>2</td><td>1</td><td>2</td><td>1</td></tr><tr><td>7</td><td>1</td><td>1</td><td>1</td><td>2</td><td>2</td></tr><tr><td>8</td><td>1</td><td>2</td><td>2</td><td>2</td><td>2</td></tr><tr><td>9</td><td>1</td><td>1</td><td>1</td><td>1</td><td>2</td></tr><tr><td>10</td><td>2</td><td>1</td><td>1</td><td>1</td><td>1</td></tr><tr><td>11</td><td>1</td><td>2</td><td>2</td><td>2</td><td>2</td></tr><tr><td>12</td><td>2</td><td>1</td><td>1</td><td>2</td><td>2</td></tr><tr><td>13</td><td>1</td><td>2</td><td>1</td><td>2</td><td>1</td></tr><tr><td>14</td><td>1</td><td>1</td><td>1</td><td>2</td><td>2</td></tr><tr><td>15</td><td>2</td><td>1</td><td>1</td><td>1</td><td>1</td></tr><tr><td>16</td><td>1</td><td>2</td><td>2</td><td>2</td><td>2</td></tr><tr><td>17</td><td>2</td><td>1</td><td>1</td><td>2</td><td>2</td></tr><tr><td>18</td><td>1</td><td>2</td><td>1</td><td>2</td><td>1</td></tr><tr><td>19</td><td>1</td><td>1</td><td>1</td><td>2</td><td>2</td></tr><tr><td>20</td><td>2</td><td>1</td><td>1</td><td>1</td><td>1</td></tr></tbody></table>	파일(F)	편집(E)	서식(O)	id sex age income view policy			1	1	2	2	2	2	2	1	1	1	1	2	3	2	1	1	1	1	4	1	2	2	2	2	5	2	1	1	2	2	6	1	2	1	2	1	7	1	1	1	2	2	8	1	2	2	2	2	9	1	1	1	1	2	10	2	1	1	1	1	11	1	2	2	2	2	12	2	1	1	2	2	13	1	2	1	2	1	14	1	1	1	2	2	15	2	1	1	1	1	16	1	2	2	2	2	17	2	1	1	2	2	18	1	2	1	2	1	19	1	1	1	2	2	20	2	1	1	1	1	setwd("C:/빅데이터/R_data") view<-read.table("view.txt", header=T) view<-view[, -1] ##Multiple CA Plot library(ca) mjca(view) par(pty="s") plot(mjca(view), main="MCRA : 성별, 나이 수입에 따른 경제전망과 정책선택도")
id	sex	age	income	view	policy																																																																																																																																																																																																																																	
1	1	2	2	2	2																																																																																																																																																																																																																																	
2	1	1	1	1	2																																																																																																																																																																																																																																	
3	2	1	1	1	1																																																																																																																																																																																																																																	
4	1	2	2	2	2																																																																																																																																																																																																																																	
5	2	1	1	2	2																																																																																																																																																																																																																																	
6	1	2	1	2	1																																																																																																																																																																																																																																	
7	1	1	1	2	2																																																																																																																																																																																																																																	
8	1	2	2	2	2																																																																																																																																																																																																																																	
9	1	1	1	1	2																																																																																																																																																																																																																																	
10	2	1	1	1	1																																																																																																																																																																																																																																	
115	2	1	1	1	1																																																																																																																																																																																																																																	
116	1	2	2	2	2																																																																																																																																																																																																																																	
117	2	1	1	2	2																																																																																																																																																																																																																																	
118	1	2	1	2	1																																																																																																																																																																																																																																	
119	1	1	1	2	2																																																																																																																																																																																																																																	
120	2	1	1	1	1																																																																																																																																																																																																																																	
파일(F)	편집(E)	서식(O)																																																																																																																																																																																																																																				
id sex age income view policy																																																																																																																																																																																																																																						
1	1	2	2	2	2																																																																																																																																																																																																																																	
2	1	1	1	1	2																																																																																																																																																																																																																																	
3	2	1	1	1	1																																																																																																																																																																																																																																	
4	1	2	2	2	2																																																																																																																																																																																																																																	
5	2	1	1	2	2																																																																																																																																																																																																																																	
6	1	2	1	2	1																																																																																																																																																																																																																																	
7	1	1	1	2	2																																																																																																																																																																																																																																	
8	1	2	2	2	2																																																																																																																																																																																																																																	
9	1	1	1	1	2																																																																																																																																																																																																																																	
10	2	1	1	1	1																																																																																																																																																																																																																																	
11	1	2	2	2	2																																																																																																																																																																																																																																	
12	2	1	1	2	2																																																																																																																																																																																																																																	
13	1	2	1	2	1																																																																																																																																																																																																																																	
14	1	1	1	2	2																																																																																																																																																																																																																																	
15	2	1	1	1	1																																																																																																																																																																																																																																	
16	1	2	2	2	2																																																																																																																																																																																																																																	
17	2	1	1	2	2																																																																																																																																																																																																																																	
18	1	2	1	2	1																																																																																																																																																																																																																																	
19	1	1	1	2	2																																																																																																																																																																																																																																	
20	2	1	1	1	1																																																																																																																																																																																																																																	
(1) sex(성별) : 1(male), 2(female) (2) age(나이) : 1(30세 이하), 2(30세 이상) (3) income(연수입) : 1(1000만원 이하), 2(1000만원 이상) (4) view(경제전망) : 1(낙관), 2(비관) (5) policy(정부의 경제정책) : 1(찬성), 2(반대)																																																																																																																																																																																																																																						

경제전망에 대한 다중 대응분석 그래프



- 제1축의 왼편에 연령(age:1)과 수입(income:1)이 낮은 여성(sex:2)들이 경제전망을 낙관(view:1)하고 정부의 경제정책을 찬성(policy:1)하고 있다.

- 제1축의 오른편에는 연령(age:2)과 수입(income:2)이 높은 남성(sex:1)들이 경제전망을 비판(view:2)하고 경제 정책을 반대(policy:2)하고 있음을 보여주고 있다.

오원분할표의 다중 대응분석 그림 : library(ca), mja()

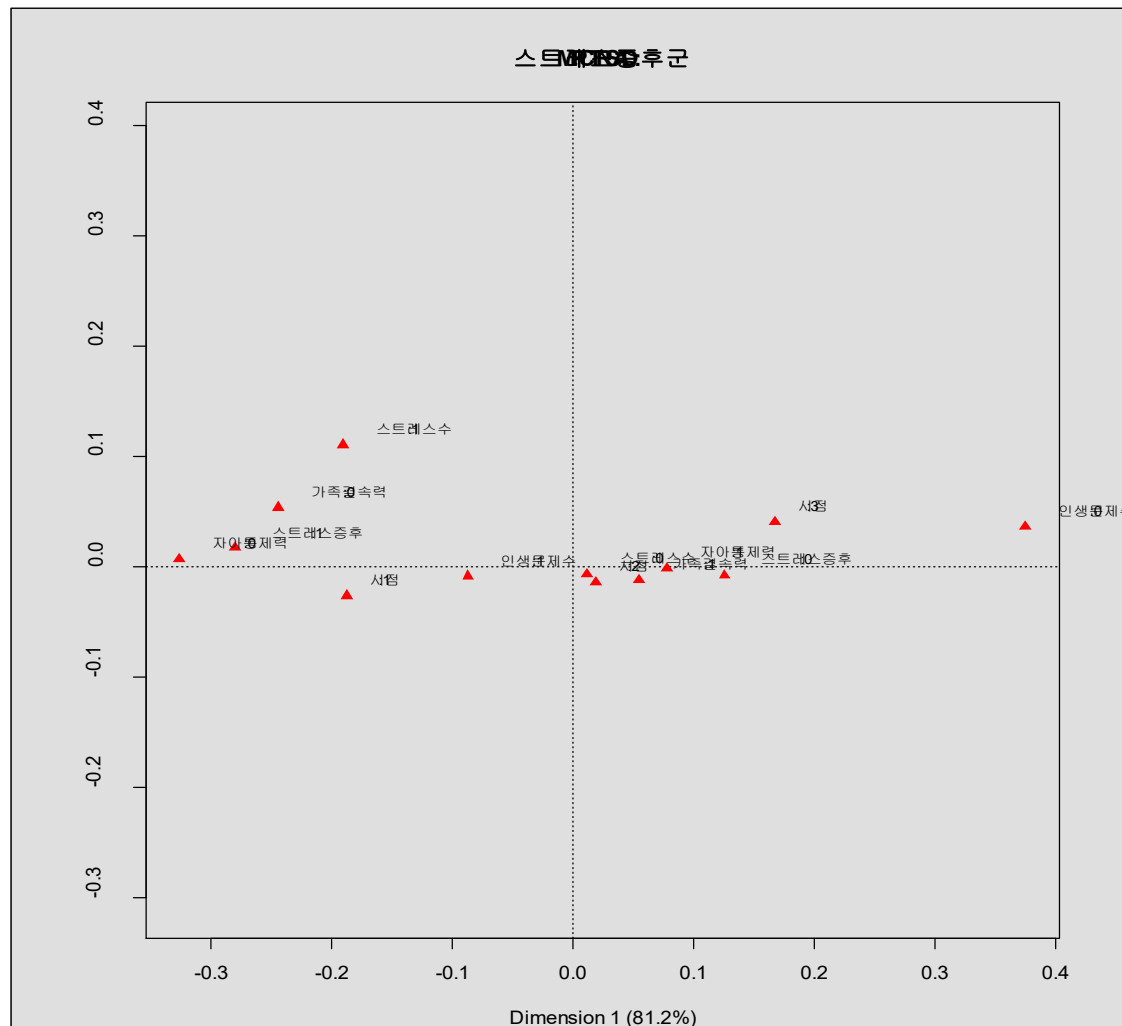
해 석

2.4.3 다원 분할표 Multi-Way Contingency Table

ptsd.txt							pstd-MCRA.R
[자료 1.4.2] 외상 후 스트레스증후군에 대한 다시점자료							<pre>setwd("C:/빅데이터/R_data") ptsd<-read.table("ptsd.txt", header=T) v1 <- as.numeric(ptsd\$자아통제력>=3) v2 <- as.numeric(ptsd\$인생문제수>=3) v3<- as.numeric(ptsd\$스트레스수>=3) v4 <- as.numeric(ptsd\$가족결속력>=6) nptsd <- data.frame(ptsd[,1:3], 자아통제력=v1,인생문제수=v2, 스트레스수=v3, 가족결속 력=v4) nptsd<-nptsd[, -1] library("ca") mjca(nptsd) par(pty="s") plot(mjca(nptsd), main="MCRA : 스트레스증후군 PTSD")</pre>
subject	time	ptsd	control	problems	sevent	cohes	
15	1	0	3.22222	5.625	1	8	
15	2	0	3.16667	5.375	0	8	
15	3	0	3.27778	3.750	1	8	
18	1	1	2.55556	9.250	0	8	
18	2	0	3.44444	4.375	0	8	
18	3	0	3.33333	2.375	0	8	
----- 중 간 생 락 -----							
571	1	0	3.55556	3.000	0	7	
571	2	0	2.94444	1.875	0	7	
571	3	0	3.50000	2.750	0	7	

ptsd - 메모장

파일(F)	편집(E)	서식(O)	보기(V)	도움말(H)		
개체	시점	스트레스증후	자아통제력	인생문제수	스트레스수	가족결속력
15	1	0	3.22222	5.625	1	8
15	2	0	3.16667	5.375	0	8
15	3	0	3.27778	3.750	1	8
18	1	1	2.55556	9.250	0	8
18	2	0	3.44444	4.375	0	8
18	3	0	3.33333	2.375	0	8
19	1	1	2.72222	7.750	1	7
19	2	1	2.77778	7.750	1	7
19	3	0	2.77778	7.500	1	7



다원분할표의 다중 대응분석 그림 : library(ca), mjca()

- 제1축 왼편 : 시점:1(3개월 후), 자아통제력:0 과 가족결속력:0, 인생문제수:1과 스트레스 수:1이 서로 대응한다.
- 환자가 외상 3개월 후 자아통제력과 가족결속력이 낮으며 인생문제와 스트레스가 높음을 보여주고 있다.
- 제2축의 오른편 : 외상 후 환자가 시점 2-3(3-12개월 후)에서 자아통제력(1)과 가족결속력(1)이 높고, 인생문제수(0)과 스트레스 (0)가 낮아졌음을 보여주고 있다.

해 석

[자료 1.4.3] 조건대 저널의 키워드와 통계기법 자료

[illegible]

origin.csv

	A	B	C	D	E	F
1	1	2	3	4	5	6
2	1	0	1	1	1	1
3	0	0	1	1	0	1
4	0	0	0	1	0	1
5	0	0	0	0	0	0
6	0	0	0	0	0	0
7	1	0	1	0	0	0
8	0	0	0	0	0	0
9	0	0	0	0	0	0
10	0	0	0	0	0	0

	A	B	C	D	E	F	G	H	I	J	K	L
1	기초통계분석	빈도분석	군집분석	다차원척도	주성분분석	ANOVA	통계적유의	다중비교	상관분석	회귀분석	t검정	서해

origin-MCRA.R

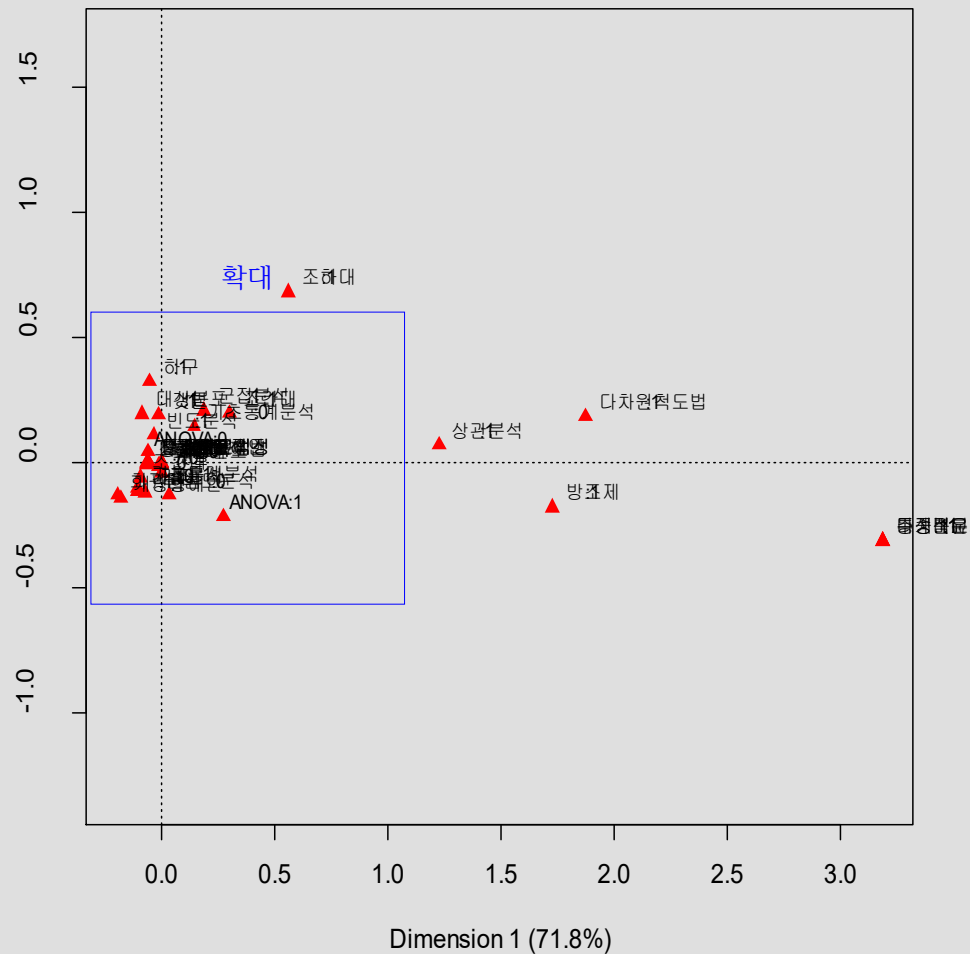
```
# 데이터 불러오기(파일명 : origin.data.csv, colname.csv)
setwd("C:/빅데이터/R_data")
origin.data <- read.csv("origin.data.csv",header=TRUE)
dim(origin.data) #138,60
```

```
data <- t(origin.data)
rownames(data) <- c(1:60)
colnames(data)<-
as.matrix(read.csv("colname.csv",header=FALSE))
data <- data.frame(data)
dim(data) #60,138
```

```
library(ca)
```

```
# 통계기법과 Zonation의 연관성
#먼저 합이 0이 되는 열이 있는지 확인, 없으면 그 데이터
그대로 사용
apply(data[,c(1:11,45:51)],2,sum)
x9 <- data[,c(1:11,45:51)] #60,18
mjca(x9)
plot(mjca(x9))
```

```
# 영역 확대
plot(mjca(x9), xlim=c(-0.3,0.3), ylim=c(-0.3,0.3))
```



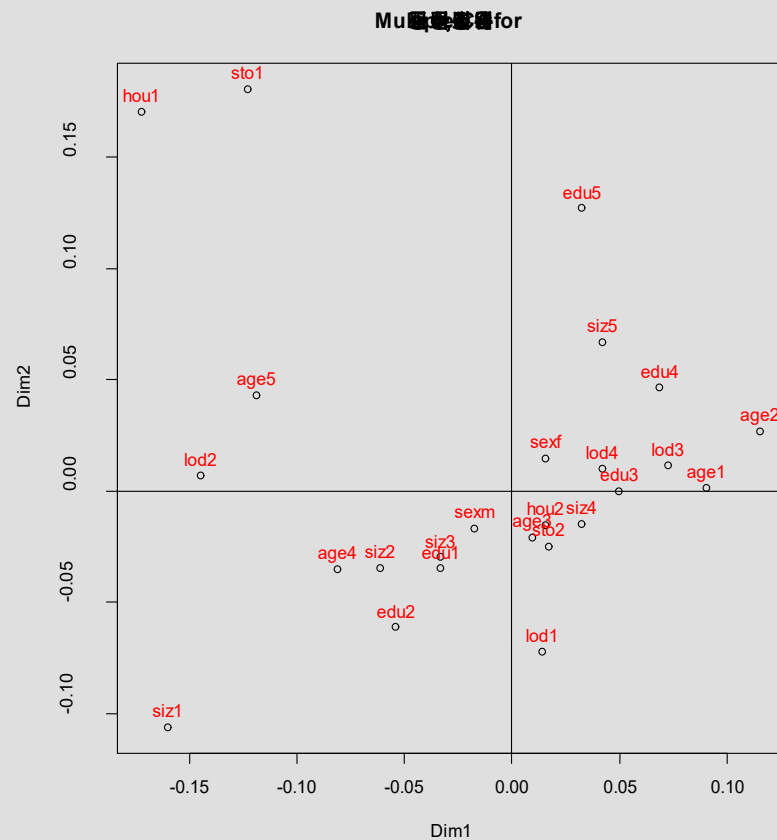
통계기법과 Zonation의 연관성을 위한 다중 대응분석 그림

```
> apply(data[,c(1:11,45:51)],2,sum)
```

기초통계분석	빈도분석	군집분석	다차원척도법	주성분분석
35	32	21	2	1
ANOVA	통계적유의성	다중비교검정	상관분석	회귀분석
10	1	1	3	1
t검정	대상분포	조건대	조하대	방조제
1	2	14	1	2
갯벌	개방형해안	하구		
10	1	1		

빈도 : apply(, sum)

2.4.5 버트행렬 자료 : 교육수준, 나이, 거주지역의 규모 등 인구통계학적 변인과 사회경제적 변인간의 관계



: Goodness-of-fit

```
> rbind(round(lam, 3), round(fit, 1))
      [,1] [,2] [,3] [,4] [,5]
[1,]  0.004  0.002  0.001  0.001  0.001
[2,] 31.200 18.100  9.300  7.400  5.400
```

사분면	대응 관계	해 석
1	age1-2, edu3-5, siz5, lod3-4, sexf	35세미만/50만명대도시 교육수준상/임대주택/여
2	age5, hou1, lod2, sto1	65세이상/주식 부동산소유
3	age4, edu1-2, siz1-3, sexm	5-60대/교육수준하, 주민수 10만명이하/남
4	age3, hou2, siz4, lod1, sto2	3-40대/10만이상도시 주식동산무소유

다중 대응분석 그림 : library(ca), mja()

해 석

2.4.6 다중 대응분석 알고리즘

[1단계] $n \times q$ 표시행렬: $Z = [Z_1 | Z_2 | \dots | Z_p]$,

• Z_j : $n \times q_j$, $j = 1, \dots, p$

• $q = q_1 + q_2 + \dots + q_p$

[2단계] $q \times q$ 버트행렬: $B = Z^t Z = (b_{ij})$, $i, j = 1, \dots, q$

[3단계] 대응행렬: $P = (p_{ij})$, $p_{ij} = \frac{b_{ij}}{b_{..}}$, $i, j = 1, \dots, q$

• $b_{..} = \sum_{i=1}^q \sum_{j=1}^q b_{ij}$: B 의 모든 원소들의 총합

• 행과 열 프로파일의 중심: $c = P^t \mathbf{1}_q$, $r = P \mathbf{1}_q$

$\Rightarrow$ 중심화 대응행렬: $\tilde{P} = P - rc^t$

[4단계] 잔차행렬의 비정칙값분해: $Y = D_r^{-1/2} \tilde{P} D_c^{-1/2} = U \Lambda V^t$

• $D_r = \text{diag}(p_{.1}, \dots, p_{.q})$, 행주변비율의 합 $p_{.i} = \sum_{j=1}^q p_{ij}$, $i = 1, \dots, q$

• $D_c = \text{diag}(p_{1.}, \dots, p_{q.})$, 열주변비율의 합 $p_{j.} = \sum_{i=1}^q p_{ij}$, $j = 1, \dots, q$

• $\Lambda = \text{diag}(\lambda_1, \dots, \lambda_q)$: Y 의 비정칙값 $\lambda_1, \dots, \lambda_q$ 에 의한 대각행렬

• U 와 V 는 각각 $U^t U = I_q$ 와 $V^t V = I_q$ 를 만족한다.

[5단계] 다중 CRA 그림의 좌표점: $C_B = D_c^{-1/2} V \Lambda$

$\Rightarrow s$ 차원의 다중 CRA 그림의 좌표점: $C_{B(s)} = D_{c(s)}^{-1/2} V_{(s)} \Lambda_{(s)}$

[6단계] $s(\geq 2)$ 차원의 다중 CRA 그림의 설명력:

$$fit = \frac{\sum_{k=1}^s \lambda_k^2}{\sum_{k=1}^q \lambda_k^2} \times 100$$

Indicator Matrix/ Burt Matrix

2 x 2 Contingency Table

	c1	c2	Total
r1	2	2	4
r2	1	3	4
Total	3	5	8

8 X 4 Indicator Table

	Rows		Columns		Total
	r1	r2	c1	c2	
1	0	1	0	0	2
1	0	1	0	0	2
1	0	0	1	0	2
1	0	0	0	1	2
0	1	1	1	0	2
0	1	0	0	1	2
0	1	0	0	1	2
0	1	0	0	1	2
Total	4	4	3	5	16

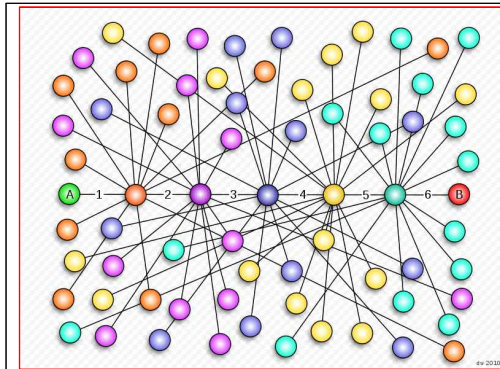
$Z = [Z_1 | Z_2]$: Indicator Matrix

$B = Z Z^t = \begin{pmatrix} Z_1^t Z_1 & Z_1^t Z_2 \\ Z_2^t Z_1 & Z_2^t Z_2 \end{pmatrix} = \begin{bmatrix} 4 & 0 & 2 & 2 \\ 0 & 4 & 1 & 3 \\ 2 & 1 & 3 & 0 \\ 2 & 3 & 0 & 5 \end{bmatrix}$: Burt Matrix

$Z_1^t Z_2 = N$

2.5 사회네트워크분석 Social Network Analysis

- Kim, J.G., Choi, Y.S. and Cho i, S.K.(2012). Social Network Analysis and It's Applications for Co-authors and Key-words Networks in the JKSS, *Communications of the Korean Statistical Society* , 19(4), 547-558.



Six degrees of separation is the theory that everyone and everything is six or fewer steps away, by way of introduction, from any other person in the world(Stanley Milgram, 1967).

- a chain of "a friend of a friend"

- 세상이 참 좁다 : small world

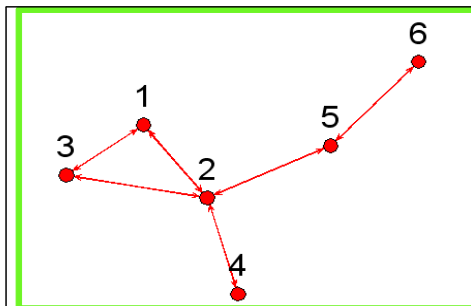
- n 10 100 1000 ----> 급히 증가

- 거리의 중간값 2 3 4 ----> 천천히 증가

Year	Distance	
2008	4.28	
2011	3.74	
2016	3.57	
Distances as reported in Feb 2016 [39]		

◇ 다수의 연결된 또는 연결되지 않은 개인으로 구성된 사회적 구조

◇ 연결 여부는 친족/친구 관계, 공통 관심, 신뢰도 등 다양하게 정의



네트워크(Network)

- **노드(node)** : 개인, 사회, 집단, 조직 등으로 정의
- **선(link)** : 정보거래, 친구관계, 직무거래 등 다양한 관계로 정의

중심성(centrality)

■ 근접 중심성(closeness centrality)

하나의 노드에서 전체 노드들의 거리를 산출하여 각 노드마다
전체 네트워크에서 중심에 얼마나 위치하고 있는지 거리를 측정하는 방법

■ 매개 중심성(betweenness centrality)

네트워크 내에서 한 노드가 담당하는 매개자 혹은 중재자 역할의 정도를
측정하는 방법

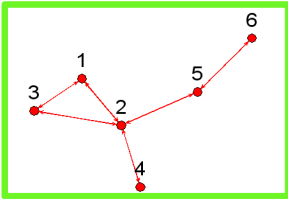
네트워크의 근접 중심성

$d(i, j)$: 노드 $i(= 1, 2, \dots, n)$ 에서 노드 $j(\neq i)$ 에 이르는 거리

◆ 노드 i 의 절대적 근접 중심성 :
$$C_C(i) = \frac{1}{\sum_{j \neq i} d(i, j)}$$

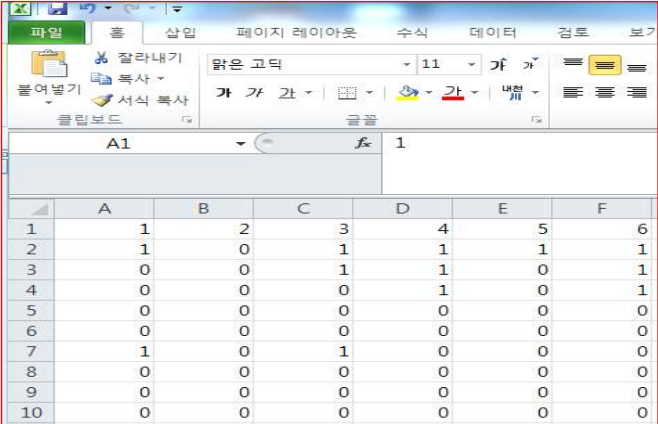
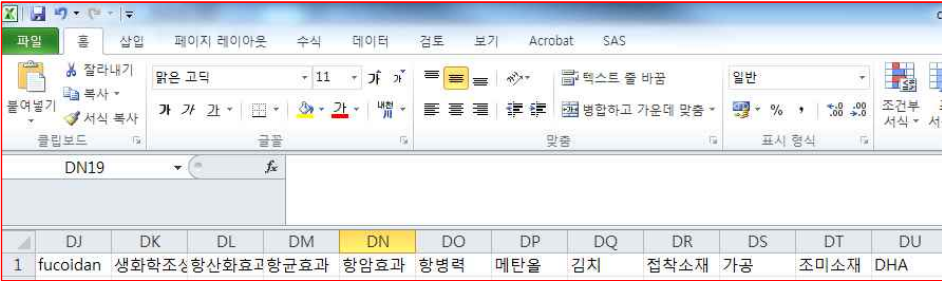
- 노드 i 의 상대적 근접 중심성 : $C_{C^*}(i) = \frac{n-1}{\sum_{j \neq i} d(i,j)}$: closeness(G.w)

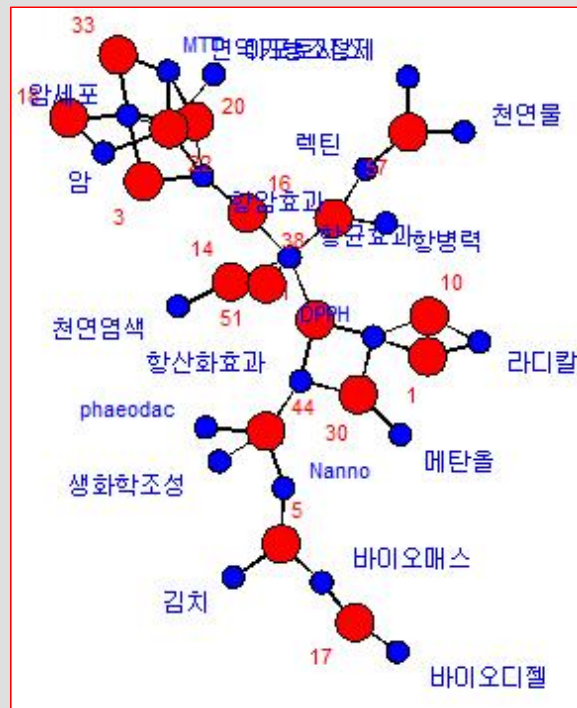
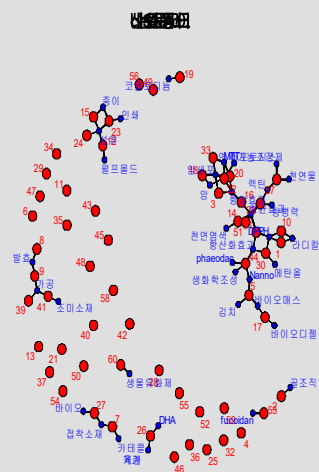
- 노드 i 의 근접 중심성 : $C_{\tilde{C}}(i) = \frac{\sum_{j \neq i} (1/d(i,j))}{n-1}$: closeness(G.w, cmode="suminvundir")



노드	1	2	3	4	5	6	$\sum_{j \neq i}^6 d(i,j)$	절대 근접중심성	상대적 근접중심성	근접 중심성
1	0	1	1	2	2	3	9	1/9	5/9	$1 + 1 + 1/2 + 1/2 + 1/3 = 3.33, \quad 3.33/5=0.67$
2	1	0	1	1	1	2	6	1/6	5/6	$1 + 1 + 1 + 1 + 1/2 = 4.5, \quad 4.5/5 = 0.9$

2.5.1 [자료 1.4.3] 조건대 저널의 키워드와 산업용어 자료

origin.csv	origin-industry-SNA.R
 	<pre> library(sna) setwd("C:/빅데이터/R_data") A<- as.matrix(read.csv("origin.data.csv",header=TRUE)) A=t(A) p = nrow(A) q = ncol(A) rownames(A) = c(1:60) colnames(A) <- as.matrix(read.csv("colname.csv",header=FALSE)) ##### 산업용어 A9=A[,102:138] A9 p9 = nrow(A9) q9 = ncol(A9) row=colnames(A9) B9 = rbind(cbind(matrix(0,p9,p9),A9),cbind(t(A9),matrix(0,q9,q9))) vertex.col = c(rep("red",p9), rep("blue",q9)) vertex.cex = c(rep(1.2,p9), rep(0.7,q9)) gplot(B9, mode="kamadakawai", displaylabels=T , boxed.label=F, vertex.col=vertex.col, vertex.cex=vertex.cex, label.col=vertex.col, label.cex=0.7, usearrows=F, main="SNA:논문과 산업용어 네트워크") ### 근접 중심성 ### close=round(closeness(B9,cmode="suminvundir"))[p9+1:q9],4) ### 중개 중심성 ### between=round(betweenness(B9)[p9+1:q9],2) rbind(row, close, between) </pre>



키워드의 빈도, 근접중심성, 중개중심성

키워드	빈도 (순위)	근접중심성 (순위)	중개중심성 (순위)
항균효과	5 (1)	0.1523 (1)	918.0 (1)
항산화효과	3 (6)	0.1306 (2)	548.0 (2)
항암효과	4 (3)	0.1304 (3)	510.3 (3)
DPPH	4 (3)	0.1282 (4)	249.0 (5)
암세포	5 (1)	0.1176 (5)	102.0 (8)
:	:	:	:

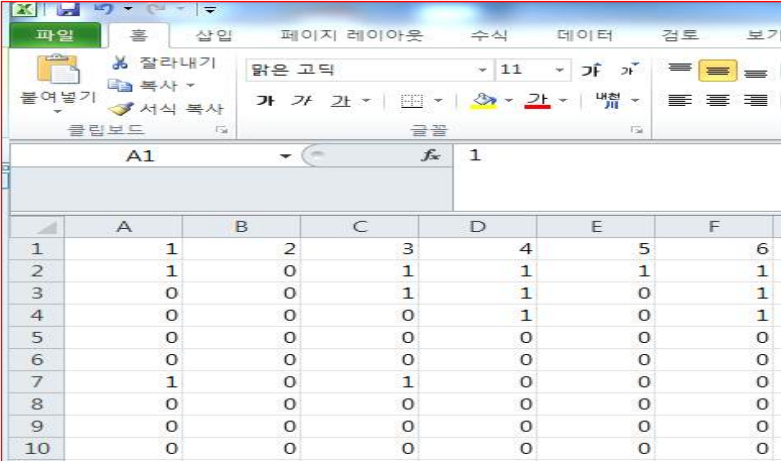
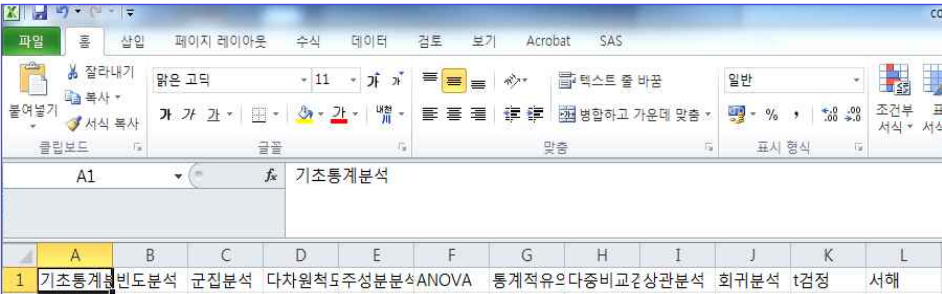
- 근접중심성과 중개중심성이 높게 나타난 키워드들이 오른쪽위쪽 하단에서 가장 큰 네트워크를 이루고 있음
- 주로 의료산업을 다루고 있는 논문 10편들이 위 키워드들과 네트워크를 형성하고 있음
- 그 외에 소재산업, 기능성식품산업과 관련된 논문과 키워드들이 작게 네트워크를 형성하고 있음

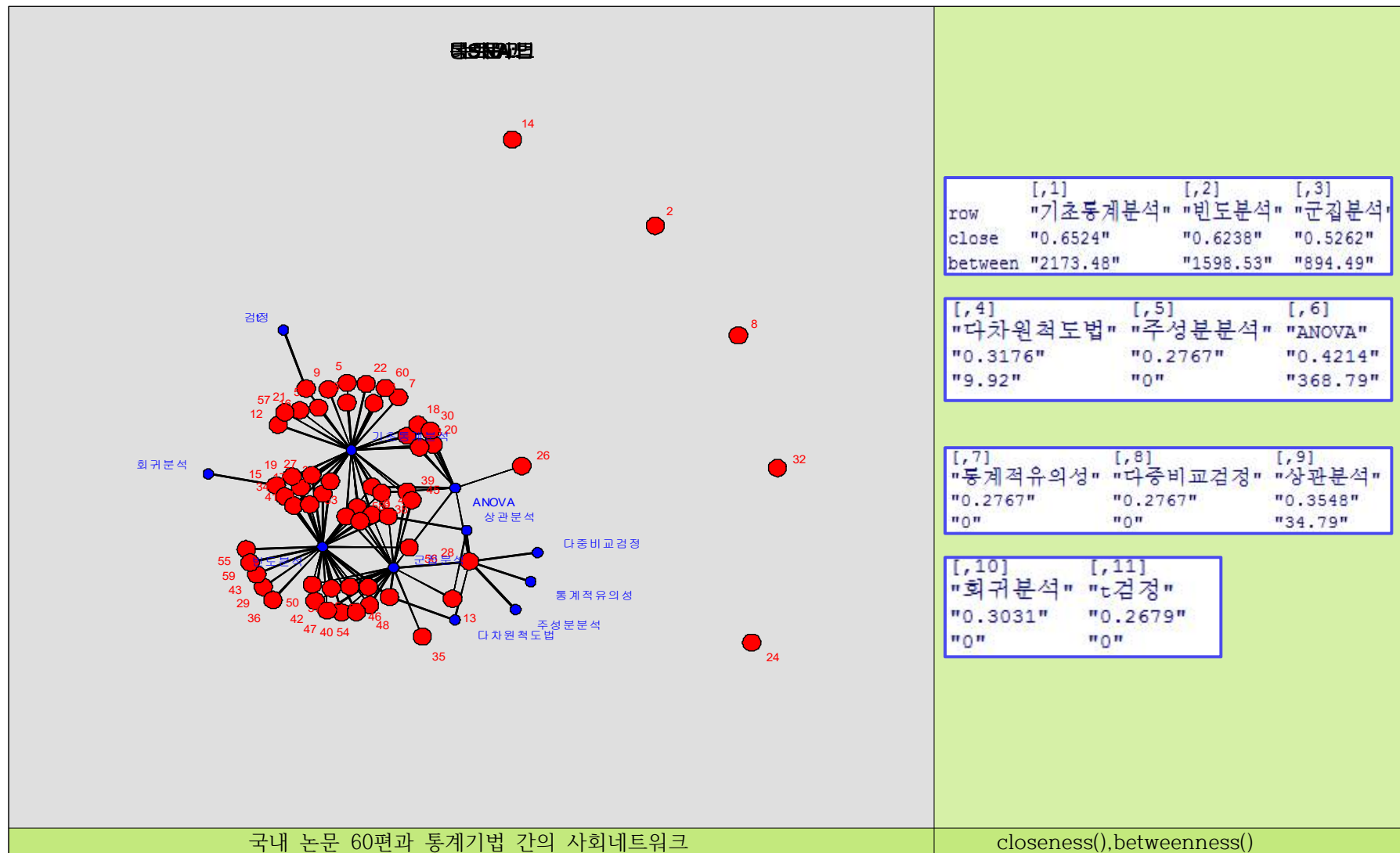
국내 논문 60편과 산업용어 37개 간의
사회네트워크 :

library(sna), gplot()

closeness(),betweenness()

2.5.2 [자료 1.4.3] 조건대 저널의 키워드와 통계기법 자료

origin.csv	origin-stat-SNA.R
 	<pre> library(sna) setwd("C:/빅데이터/R_data") A<- as.matrix(read.csv("origin.data.csv",header=TRUE)) A=t(A) p = nrow(A) q = ncol(A) rownames(A) = c(1:60) colnames(A) <- as.matrix(read.csv("colname.csv",header=FALSE)) ##### 통계기법 A1=A[,1:11] p1 = nrow(A1) q1 = ncol(A1) row=colnames(A1) B1= rbind(cbind(matrix(0,p1,p1),A1),cbind(t(A1),matrix(0,q1,q1))) vertex.col = c(rep("red",p1), rep("blue",q1)) vertex.cex = c(rep(1.2,p1), rep(0.7,q1)) gplot(B1, mode="kamadakawai", displaylabels=T , boxed.label=F, vertex.col=vertex.col, vertex.cex=vertex.cex, label.col=vertex.col, label.cex=0.7, usearrows=F, main="SNA:논문과 통계기법 네트워크") ### 근접 중심성 ### close=round(closeness(B1,cmode="suminvundir")[p1+1:q1],4) ### 중개 중심성 ### between=round(betweenness(B1)[p1+1:q1],2) rbind(row, close, between) </pre>



2.6 워드 클라우드 Word Cloud

2017bigdata.txt	2017bigdata-cloud.R
2017bigdata - 메모장 파일(F) 편집(E) 서식(O) 보기(V) 도움말(H) 부산대통계연구소, 빅데이터, 빅데이터, 빅데이터, Big Data 사회네트워크, 통계학과, 로지스틱스, 빅데이터, 시간화, 시간화 데이터, 수집, 저장, 관리, 분석, Big Data, 빅데이터, 빅데이터, 빅데이터, 빅데이터, 수집, 발굴, 빅데이터, Big Data, 정형화자료, 비정형화자료, 텍스트자료, 데이터 베이스, 빅데이터, Big Data, 정형화자료, 비정형화자료, 텍스트자료, 관측, 수집, 데이터, 정보, 추출, 지식, 추출, 비정형화자료, 텍스트자료, 관계, 빅데이터, 음악, 의사결정, 표본, 통계조사, Big Data, 시간화, 민주통, 자료, 다변량, 다변량통계학, Big Data, 정형화자료, 비정형화자료, 텍스트자료, 다변량, EDA, 회귀분석, Machine learning, 알고리즘, 통계학, 빅데이터, 빅데이터, 빅데이터, 스토리텔링, 데이터 마이닝, 빅데이터, 시간화, 비정형화자료, 텍스트자료, 정형화자료, 비정형화자료, 텍스트자료, 비정형화자료, 텍스트자료, 멀리투스, 정형화, 비정형화, 멀리투스, 사회네트워크, 멀리투스, 빅데이터, 빅데이터, 노드, 선, 중심성, 근접, 중심성, 매개 중심성, 네트워크, 시간화, 시간화, 민주통, 민주통, 빅데이터, 시간화, 데이터베이스, 빅데이터, 멀리투스, 다변량통계, 사회네트워크, 통계학, 탐색적, 자료분석기법, 시간화, 시간화, 다변량, 다변량, 통계학, 중심성, 회귀분석, 다변량통계학, 다변량통계학, 다변량통계학, 회귀분석, 회귀분석, 회귀분석, 회귀분석, 회귀분석, 회귀분석, 부산대통계연구소, 사회네트워크, 회귀분석, 부산대통계연구소, 사회네트워크, 회귀분석, Big Data 부산대통계연구소, 사회네트워크, 회귀분석, Big Data 비정형화자료, 텍스트자료, 나이, 인구통계, 이력통계, 다원통계, 시간화, 부산대통계연구소, 부산대통계연구소, 부산대통계연구소, 부산대통계연구소	<pre> install.packages("KoNLP")#KoNLP 패키지 설치 #install.packages("wordcloud") #wordcloud 패키지 설치 library(KoNLP) #KoNLP 패키지 로딩 library(wordcloud) #wordcloud 패키지 로딩 library(RColorBrewer) #컬러 출력 라이브러리 로딩 useSejongDic() #한글 사전 사용 setwd("C:/빅데이터/R_data") data1 <- readLines("2017bigdata.txt") #텍스트파일을 읽어서 data1 변수에 저장 data1 #data1에 저장된 내용 확인 data2<-sapply(data1,extractNoun,USE.NAMES = F) # data1에서 명사만 data2에 저장 data2 #data2에 저장된 내용 확인 data3 <- unlist(data2)#두글자 이상 필터링을 위해 unlist작업하여 저장 data3 <- Filter(function(x){nchar(x)>= 2}, data3)# 두 글자 이상 단어만 저장 data3 #data2에 저장된 내용 확인, 변경된 걸 확인하실 수 있을 겁니다. data3 <- gsub("\\d+", "", data3) #숫자 제외, 정규식 data3 <- gsub("\\\\(", "", data3) # 괄호 제외 data3 <- gsub("\\\\)", "", data3) #)괄호 제외 data3 <- gsub("[A-Za-z]", "", data3) # 영문 제외, 정규식 write(unlist(data3), "bigdata2.txt") #data3를 txt파일로 저장 data4 <- read.table("bigdata2.txt") #read.table로 읽으면서 공백 제거 data4 wordcount <- table(data4)# 단어별 count wordcount # 단어별 count 확인 palette <- brewer.pal(9, "Set1") # 글자 색깔 지정 wordcloud(names(wordcount), freq=wordcount, scale=c(5,1),rot.per = 0.1, min.freq=1,random.order=F, color=T, colors=palette) </pre>



data4	관리	관측	근접	네트워크
1	1	1	1	7
노드	다변량	다변량분산분석		다변량통계
1	4	1	1	1
다변량통계학	다원	대응		대칭
3	1	2		1
데이터	토지소식	마이닝		매개
3	1	1		1
맵리뷰스	모형	발굴		범주
4	1	1		2
범주형자료	베이스	부산대통계연구소		분산
1	1	5		4
분석	분할	비정형		빅데이터
7	3	1		15
빅데이터, 빅데이터, 빅데이터		사회		수집
1	1	6		3
스플데이터, 데이터	시각화	알고리즘		용약
1	10	1		1
용어	의사결정	이원		자료
2	1	1		12
자료분석기	저장	정보		정형화
1	1	1		13
조간	조사	구성분		중심
2	1	1		3
지식	창출	추출		탐색
1	1	1		1
데이터	텍스터자료	통계		통계학
1	8	4		1
통계학과	표본	행렬		회귀
1	1	5		1
회귀분석				
3				

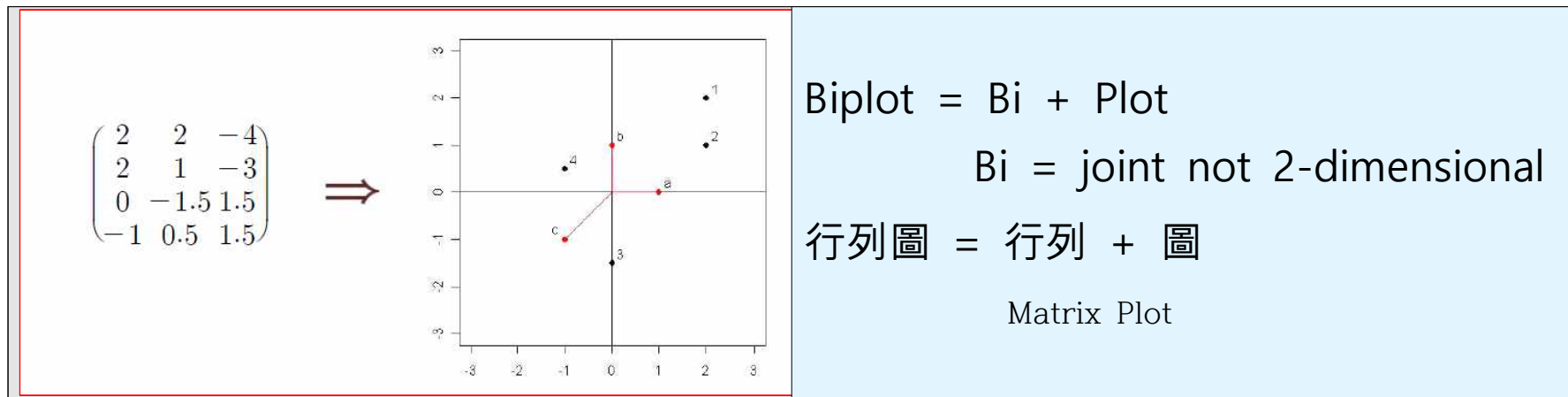
주제 3 : 다변량자료의 시각화 (Visualizations for MDA)

3.1 행렬도

3.2 다차원척도법

3.1 행렬도 Biplot

Geometric Visualization of Biplot



◇ 행렬도 개발

Gabriel(1971, 1982) : 주성분행렬도, JK' 행렬도, 대칭행렬도

◇ 국내 행렬도 연구

Choi(1991) : 행렬도의 저항성 버전 연구, 행렬도를 처음 소개

허명회(1993) : Biplot을 行列圖라 처음 명명함

3.1.1 행렬도(주성분, 대칭, JK)을 위한 알고리즘

[1단계] 자료행렬 Y : 크기가 $n \times p$ 이고 계수(rank)가 r 이다.

[2단계] Y 의 비정칙값분해: $Y = U \Lambda V^t = \sum_{k=1}^r \mathbf{u}_k \lambda_k \mathbf{v}_k^t = \sum_{k=1}^r (\mathbf{u}_k \lambda_k^m) (\lambda_k^{1-m} \mathbf{v}_k^t)$ (9.2.1)

- $U = (\mathbf{u}_1, \dots, \mathbf{u}_r)$: 직교 열벡터 $\mathbf{u}_k = (u_{1k}, \dots, u_{nk})^t$, $k = 1, \dots, r$
- $V = (\mathbf{v}_1, \dots, \mathbf{v}_r)$: 직교 열벡터 $\mathbf{v}_k = (v_{1k}, \dots, v_{pk})^t$, $k = 1, \dots, r$
- $\Lambda = \text{diag}(\lambda_1, \dots, \lambda_r)$: $\lambda_1 \geq \dots \geq \lambda_r > 0$ 을 만족하는 비정칙값의 대각행렬
- $0 \leq m \leq 1$: 상수

[3단계] 행렬도의 좌표점:

- $n \times r$ 행 좌표점: $A = U \Lambda^m = (\mathbf{u}_1 \lambda_1^m, \dots, \mathbf{u}_r \lambda_r^m)$
- $p \times r$ 열 좌표점: $B = V \Lambda^{1-m} = (\mathbf{v}_1 \lambda_1^{1-m}, \dots, \mathbf{v}_r \lambda_r^{1-m})$

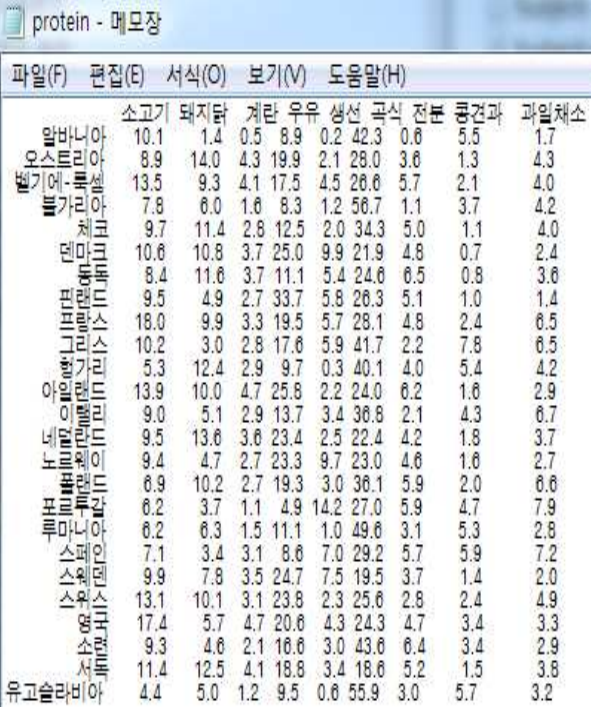
$\Rightarrow s$ 차원 행렬도의 행과 열 좌표점: $A_{(s)} = U_{(s)} \Lambda_{(s)}^m$, $B_{(s)} = V_{(s)} \Lambda_{(s)}^{1-m}$

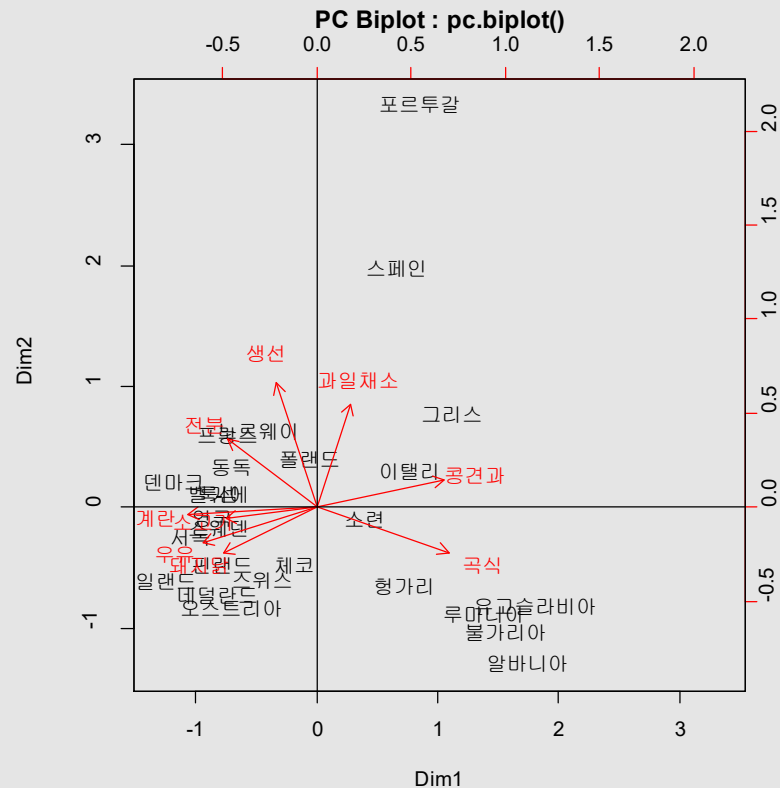
[4단계] 행렬도의 선택:

- $m = 0$: 주성분 행렬도, $m = 1/2$: 대칭 행렬도, $m = 1$: JK^t 행렬도

[5단계] $s (\geq 2)$ 차원 행렬도의 설명력: $fit = \frac{\sum_{k=1}^s \lambda_k^2}{\sum_{k=1}^r \lambda_k^2} \times 100$

3.1.2 [자료 0.1.3] 유럽 25개국의 단백질 섭취원(단위: 그램/일) 자료

protein.txt	protein-PCbiplot.R
	<pre> setwd("C:/빅데이터/R_data") Y<-read.table("protein.txt", header=T) pcasvd<-prcomp(Y, scale=T) summary(pcasvd) biplot(pcasvd, pc.biplot=T, xlab="Dim1", ylab="Dim2", main=" PC Biplot : pc.biplot() ", cex=1.0, pch=16) abline(v=0,h=0) </pre>



주성분행렬도 : prcomp(), biplot(, pc.biplot=T)

- 설명력 : 제1축 : 44.52% + 제2축 : 18.17%

- 제3사분면 : 단백질 섭취원인 우유는 계란, 소고기, 돼지, 닭과 매우 상관이 높다. 이들로부터 단백질을 주로 섭취하는 나라 : 핀란드, 아일랜드, 영국, 스웨덴, 서독, 네덜란드, 덴마크, 오스트리아, 스위스이다.

- 반면에 이들 나라는 제4사분면의 곡식과 제1사분면의 과일채소, 콩과류로부터 상대적으로 낮은 단백질 섭취를 하고 있다. 실제, 우유와 곡식은 동물성과 식물성단백질의 섭취원을 대표하며, 이들 두 좌표가 이루는 각이 거의 90° 이상으로 음의 상관을 보여주고 있다.

- 생선이 주된 단백질 섭취원인 나라는 스페인과 포르투갈이다. 특히, 생선은 과일, 채소와 매우 상관이 높아 이들은 스페인과 포르투갈의 주요 단백질 섭취원

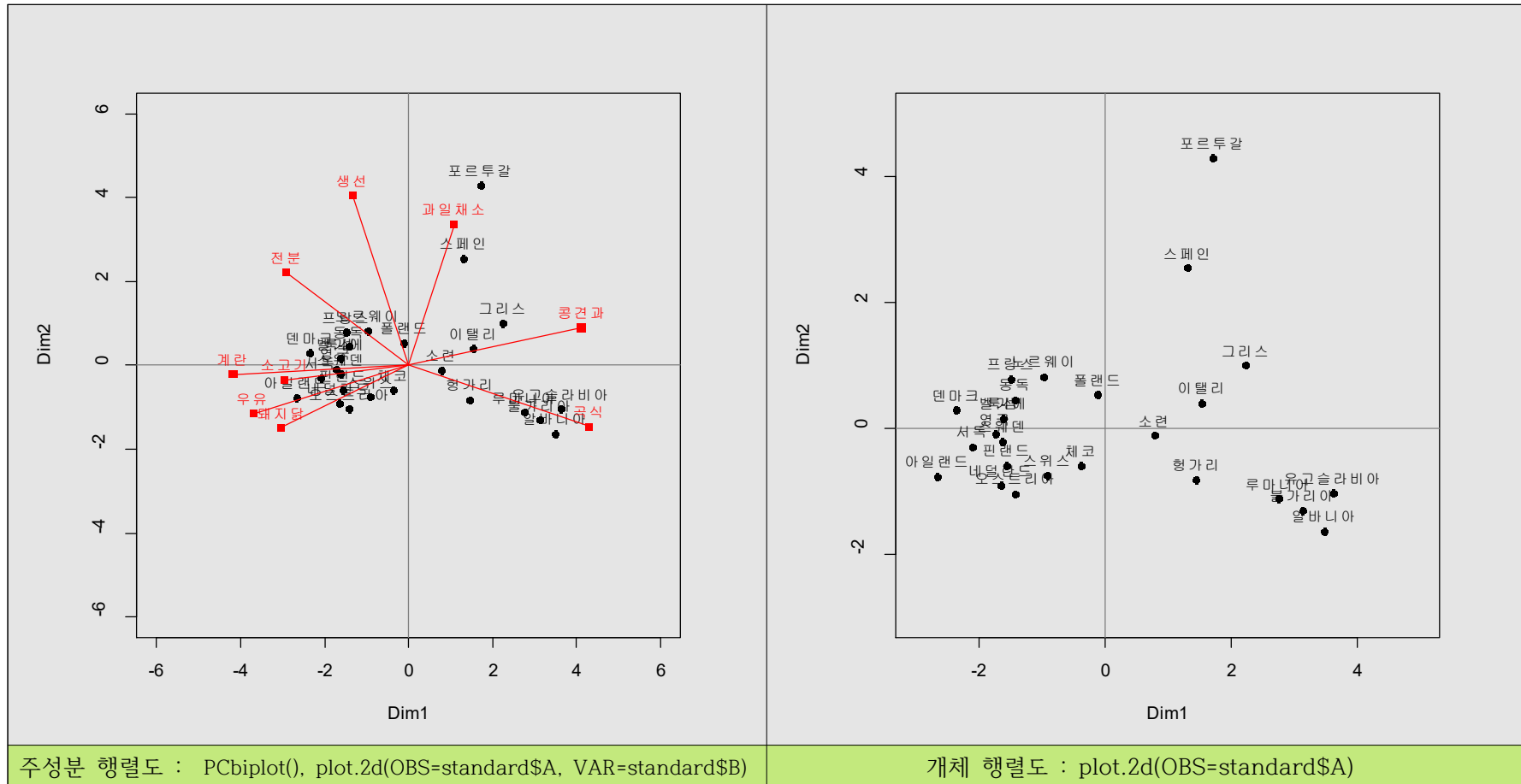
- 제4사분면 : 곡식은 콩과류와 상관이 높으며 동유럽 국가인 소련, 헝가리, 루마니아, 유고슬라비아, 불가리아, 알바니아가 이들로부터 주로 단백질을 섭취하고 있다.

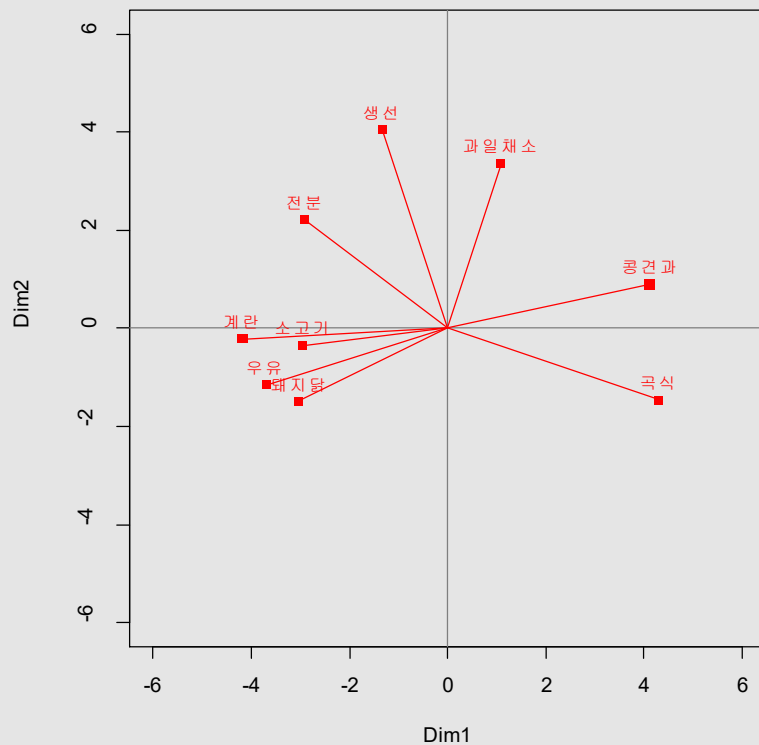
- 제4사분면 : 동유럽 국가이고 단백질 섭취원은 곡식,
제3사분면 : 북서 유럽국가로 동물성단백질(우유, 소고기, 돼지, 닭, 계란) 섭취원
제1과 제2사분면 위쪽 : 지중해 연안국가- 과일,채소와 콩,견과류, 생선 단백질 섭취.

해 석

3.1.2 [자료 0.1.3] 유럽 25개국의 단백질 섭취원(단위: 그램/일) 자료 : Algebraic Algorithm

protein-PCbiplotAlgorithm.R	
<pre> setwd("C:/빅데이터/R_data") X<-read.table("protein.txt", header=T) colnames(X) rownames(X) ##### Functions ##### PCbiplot <- function(X,std=0) { if(!is.matrix(X)) { X <- as.matrix(X) } if(is.na(match(std,c(1,0)))){stop("std : Select One of 1:Standardized / 0:Centred","\n")} n <- nrow(X) p <- ncol(X) if(std==0) { Y <- scale(X,scale=F);std <- "Centred"} if(std==1) { Y <- scale(X);std <- "Standardized"} svd.Y <- svd(Y) U <- svd.Y\$u V <- svd.Y\$v D <- diag(svd.Y\$d) r <- length(svd.Y\$d) A <- U%*%D B <- V%*%D dimnames(A) <- list(rownames(X),paste("Dim",c(1:r),sep="")) dimnames(B) <- list(colnames(X),paste("Dim",c(1:r),sep="")) method <- paste(std,"PCplot") gof <- gof.2d(svd.Y\$d) return(list(method=method,Eigen=gof\$per,GOF=gof\$gof,A=A[,1:2],B=B[,1:2])) } gof.2d <- function(s.val){ e.val <- s.val^2 per <- e.val/sum(e.val)*100 e.mat <- round(cbind(e.val,per),4) colnames(e.mat) <- c("Eigenvalue", "Percent") gof <- paste("Goodness of Fit : ",round(sum(per[1:2]),3),"%",sep="") return(list(per=e.mat,gof=gof)) } plot.2d <- function(OBS=NULL,VAR=NULL,Line=T,v.lab=T,o.lab=T,v.pos=3,o.pos=3,lim =NULL,Title=NULL){ if(!is.logical(Line)){ stop("Line : Select One of T:Line Draw / F:Only Points","\n") } </pre>	<pre> if(!is.logical(v.lab)){ stop("v.lab : Select One of T:Var label write / F:Not","\n") } if(!is.logical(o.lab)){ stop("o.lab : Select One of T:Obs label write / F:Not","\n") } if(is.null(lim)){ if(is.null(OBS)){ lim <- range(pretty(VAR)) } if(is.null(VAR)){ lim <- range(pretty(OBS)) } if(!is.null(OBS)&!is.null(VAR)){ lim <- range(pretty(range(OBS,VAR))) } } par(pty="s") if(!is.null(OBS)){ n <- nrow(OBS) plot(OBS[,1],OBS[,2],xlab="Dim1",ylab="Dim2",xlim=lim,ylim=lim,pch=16,main=Title) abline(v=0,h=0,col="gray50") if(o.lab) { if(is.null(rownames(OBS))) { text(OBS[,1],OBS[,2],c(1:n),cex=0.8,pos=o.pos) } if(!is.null(rownames(OBS))){ text(OBS[,1],OBS[,2],rownames(OBS),cex=0.8,p=o.pos) } } if(!is.null(VAR)){ points(VAR[,1],VAR[,2],pch=15,col=2) if(v.lab){ text(VAR[,1],VAR[,2],rownames(VAR),cex=0.8,col=2,pos=v.pos) } if(Line){ arrows(0,0,VAR[,1],VAR[,2],col=2,code=0) } } } if(is.null(OBS)){ plot(VAR[,1],VAR[,2],xlab="Dim1",ylab="Dim2",xlim=lim,ylim=lim,pch=15,col=2,main=Title) abline(v=0,h=0,col="gray50") if(v.lab){ text(VAR[,1],VAR[,2],rownames(VAR),cex=0.8,col=2,pos=v.pos) } if(Line){ arrows(0,0,VAR[,1],VAR[,2],col=2,code=0) } } } ##### PCA, PCplot ##### standard <- PCbiplot(X, std=1) plot.2d(OBS=standard\$A, VAR=standard\$B) plot.2d(OBS=standard\$A) plot.2d(VAR=standard\$B) PCbiplot(X, std=1) </pre>





변수 행렬도 : plot.2d(VAR=standard\$B)

	Eigenvalue	Percent
[1,]	96.1545	44.5160
[2,]	39.2400	18.1667
[3,]	27.0701	12.5324
[4,]	22.9119	10.6074
[5,]	11.1321	5.1538
[6,]	7.8031	3.6126
[7,]	6.5186	3.0178
[8,]	2.7910	1.2921
[9,]	2.3787	1.1012

\$GOF	
[1]	"Goodness of Fit : 62.683%"

\$A		Dim1	Dim2
알바니아	3.4853673	-1.63047985	
오스트리아	-1.4226694	-1.04123130	
벨기에-룩셈	-1.6220323	0.15949557	
불가리아	3.1340813	-1.30106563	
체코	-0.3704646	-0.60266842	
덴마크	-2.3652688	0.28544582	
독일	-1.4222108	0.45030085	
핀란드	-1.5638563	-0.59600255	
프랑스	-1.4879824	0.78536517	
그리스	2.2397000	1.00105887	
헝가리	1.4574398	-0.81595115	
아일랜드	-2.6634775	-0.76370648	
이탈리아	1.5345653	0.39898708	
네덜란드	-1.6414454	-0.91199089	
노르웨이	-0.9747029	0.82202867	
폴란드	-0.1218695	0.53174194	
포르투갈	1.7058540	4.28893399	
루마니아	2.7568124	-1.11878536	
스페인	1.3118074	2.55352416	
스웨덴	-1.6337300	-0.20738445	
스위스	-0.9123182	-0.75105865	
영국	-1.7353682	-0.09397944	
소련	0.7825965	-0.11077014	
서독	-2.0938353	-0.29377901	
유고슬라비아	3.6230077	-1.03802883	

\$B		Dim1	Dim2
소고기	-2.967339	-0.3523707	
패지탕	-3.045265	-1.4836926	
계란	-4.183941	-0.2213496	
우유	-3.703933	-1.1562978	
생선	-1.330161	4.0517969	
곡식	4.292442	-1.4625932	
전분	-2.914763	2.2101643	
콩견과	4.121732	0.8977236	
과일채소	1.082760	3.3587925	

행렬도를 위한 결과 : PCbiplot(X, std=1)

3.1.3 일본과 구미 선진국의 고령화율 추이 (단위 : %) : 대도시 고령사회의 현상과 대응(금성근 · 한승욱 · 한성일(창의연구 2015-12-657))

agingt.txt											agingt-PCbiplot.R	
연도	일본	캐나다	미국	이탈리아	네덜란드	스웨덴	스위스	영국	독일			
1970	7.1	7.9	9.8	11.2	10.2	13.7	11.4	13.0	13.7			
1980	9.1	9.4	11.2	13.5	11.5	16.3	13.8	14.9	15.6			
1990	12.1	11.3	12.3	15.2	12.8	17.8	14.6	15.7	15.0			
2000	17.4	12.6	12.4	18.4	13.6	17.2	15.4	15.9	16.4			
2005	20.2	13.1	12.4	19.7	14.2	17.2	16.0	16.1	18.9			
2010	23.1	14.1	13.0	20.4	15.4	18.3	17.3	16.6	20.5			
2020	29.3	18.1	16.1	23.0	19.8	21.0	20.2	18.5	23.0			
2030	31.8	22.7	19.8	26.8	23.8	22.6	24.1	20.9	28.2			
2040	36.5	24.5	21.0	31.8	26.3	24.1	26.3	22.6	31.8			
2050	39.6	25.5	21.6	33.3	25.6	24.1	26.0	22.9	32.5			

agingt - 메모장

파일(F) 편집(E) 서식(O) 보기(V) 도움말(H)

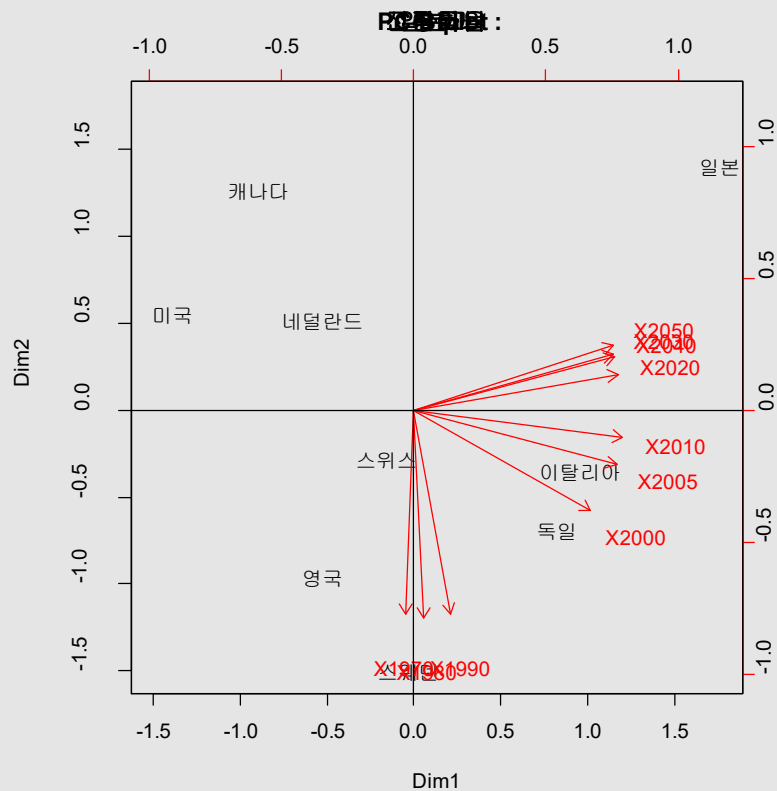
	1970	1980	1990	2000	2005	2010	2020	2030	2040	2050
일본	7.1	9.1	12.1	17.4	20.2	23.1	29.3	31.8	36.5	39.6
캐나다	7.9	9.4	11.3	12.6	13.1	14.1	18.1	22.7	24.5	25.5
미국	9.8	11.2	12.3	12.4	12.4	13.0	16.1	19.8	21.0	21.6
이탈리아	11.2	13.5	15.2	18.4	19.7	20.4	23.0	26.8	31.8	33.3
네덜란드	10.2	11.5	12.8	13.6	14.2	15.4	19.8	23.8	26.3	25.6
스웨덴	13.7	16.3	17.8	17.2	17.2	18.3	21.0	22.6	24.1	24.1
스위스	11.4	13.8	14.6	15.4	16.0	17.3	20.2	24.1	26.3	26.0
영국	13.0	14.9	15.7	15.9	16.1	16.6	18.5	20.9	22.6	22.9
독일	13.7	15.6	15.0	16.4	18.9	20.5	23.0	28.2	31.8	32.5

```

setwd("C:/빅데이터/R_data")
Y<-read.table("agingt.txt", header=T)

pcasvd<-prcomp(Y, scale=T)
summary(pcasvd)
biplot(pcasvd, pc.biplot=T, xlab="Dim1", ylab="Dim2",
       main=" PC Biplot :일본과 구미 선진국의 고령화율 추이 ", cex=1.0, pch=16)
abline(v=0,h=0)

```



```
> summary(pcasvd)
```

Importance of components:

	PC1	PC2	PC3	PC4	PC5	PC6	PC7
Standard deviation	2.5011	1.8488	0.48344	0.25675	0.1343	0.07379	0.05536
Proportion of Variance	0.6255	0.3418	0.02337	0.00659	0.0018	0.00054	0.00031
Cumulative Proportion	0.6255	0.9673	0.99072	0.99731	0.9991	0.99966	0.99996

	PC8	PC9
Standard deviation	0.01917	3.824e-16
Proportion of Variance	0.00004	0.000e+00
Cumulative Proportion	1.00000	1.000e+00

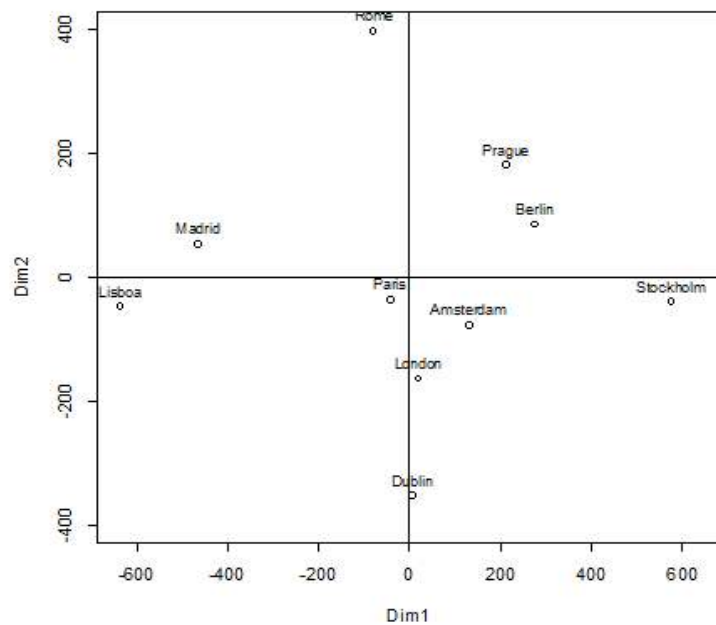
주성분행렬도 : prcomp(), biplot(, pc.biplot=T)

결 과

3.2 다차원척도법 Multidimensional Scaling

Metric MDS of distance between 10 European cities

London	0									
Stockholm	569	0								
Lisboa	667	1212	0							
Madrid	530	1043	201	0						
Paris	141	617	596	431	0					
Amsterdam	140	446	768	608	177	0				
Berlin	357	325	923	740	340	218	0			
Prague	396	423	882	690	337	272	114	0		
Rome	569	787	714	516	436	519	472	364	0	
Dublin	190	648	714	622	320	302	514	573	755	0



3.2.1 Torgerson Algorithm for Metric MDS

[Step 1] Calculate an $n \times n$ matrix $A = (a_{ij})$ from $D = (d_{ij})$, $i, j = 1, \dots, n$

$$\text{where, } a_{ij} = -\frac{1}{2}d_{ij}^2, \quad i, j = 1, \dots, n.$$

[Step 2] Calculate $B = (b_{ij}) = H A H$ from $A = (a_{ij})$

$$\text{where, } b_{ij} = a_{ij} - \bar{a}_{i.} - \bar{a}_{.j} + \bar{a}_{..}$$

$$a_{ij} = -\frac{1}{2}d_{ij}^2, \quad \bar{a}_{i.} = \frac{1}{n} \sum_{j=1}^n a_{ij}, \quad \bar{a}_{.j} = \frac{1}{n} \sum_{i=1}^n a_{ij}, \quad \bar{a}_{..} = \frac{1}{n^2} \sum_{i=1}^n \sum_{j=1}^n a_{ij}.$$

$$H = I_n - J_n/n : \text{centering matrix}$$

GOF

$$fit_1 = \frac{\sum_{i=1}^s l_i^2}{\sum_{i=1}^n l_i^2} \times 100$$

[Step 3] Calculate spectral decomposition: $B = V D_l V^t$

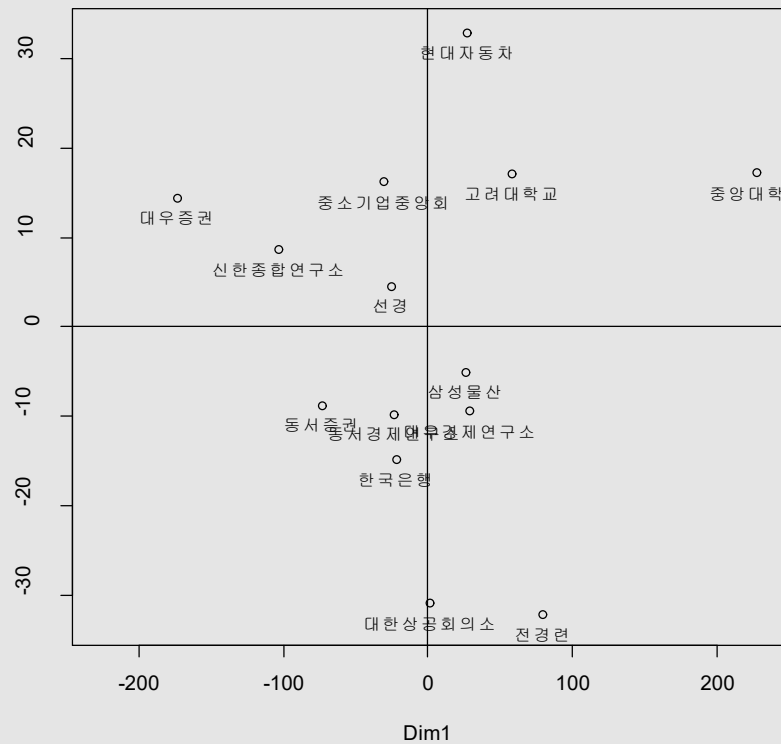
where, D_l : Diagonal matrix, V : Orthogonal matrix

[Step 4] Obtain an $n \times s$ coordinate matrix:

$$C_{(s)} = V_{(s)} D_{l(s)}^{1/2} = (\mathbf{v}_1 l_1^{1/2}, \dots, \mathbf{v}_s l_s^{1/2})$$

3.2.2 [자료 0.1.4] 경제관련 기관의 경제전망 자료

economicview.txt	economicview-MDS.R																																																																																																																																																																					
economicview - 메모장 파일(F) 편집(E) 서식(O) 보기(V) 도움말(H) <table><thead><tr><th>기관</th><th>성장률</th><th>GNP</th><th>수출</th><th>수입</th><th>국제흑자</th><th>연말외채</th><th>연말환율</th><th>실업률</th><th>소비물가</th><th>임금상승</th></tr></thead><tbody><tr><td>한국은행</td><td>8.2</td><td>1950</td><td>698</td><td>650</td><td>98</td><td>280</td><td>630</td><td>3.0</td><td>5.7</td><td>12.0</td></tr><tr><td>대우증권</td><td>9.5</td><td>2100</td><td>710</td><td>620</td><td>110</td><td>270</td><td>640</td><td>2.7</td><td>4.5</td><td>12.0</td></tr><tr><td>동서증권</td><td>9.0</td><td>2000</td><td>690</td><td>630</td><td>100</td><td>290</td><td>630</td><td>2.8</td><td>6.0</td><td>12.0</td></tr><tr><td>전경련</td><td>7.8</td><td>1850</td><td>668</td><td>660</td><td>88</td><td>280</td><td>620</td><td>3.0</td><td>6.4</td><td>13.8</td></tr><tr><td>대한상공회의소</td><td>8.5</td><td>1928</td><td>710</td><td>670</td><td>90</td><td>290</td><td>620</td><td>3.0</td><td>6.0</td><td>10.0</td></tr><tr><td>중소기업중앙회</td><td>9.0</td><td>1958</td><td>710</td><td>615</td><td>95</td><td>280</td><td>603</td><td>4.0</td><td>6.0</td><td>10.0</td></tr><tr><td>현대자동차</td><td>8.5</td><td>1900</td><td>700</td><td>610</td><td>100</td><td>250</td><td>620</td><td>3.2</td><td>5.5</td><td>14.0</td></tr><tr><td>삼성물산</td><td>8.0</td><td>1900</td><td>700</td><td>640</td><td>100</td><td>280</td><td>640</td><td>2.7</td><td>5.5</td><td>10.0</td></tr><tr><td>선경</td><td>8.5</td><td>1950</td><td>700</td><td>620</td><td>120</td><td>300</td><td>630</td><td>2.7</td><td>7.0</td><td>12.0</td></tr><tr><td>대우경제연구소</td><td>7.9</td><td>1900</td><td>697</td><td>645</td><td>95</td><td>280</td><td>610</td><td>2.9</td><td>6.8</td><td>15.0</td></tr><tr><td>신한종합연구소</td><td>9.0</td><td>2030</td><td>700</td><td>620</td><td>100</td><td>275</td><td>630</td><td>3.0</td><td>7.0</td><td>13.0</td></tr><tr><td>영서경제연구소</td><td>8.5</td><td>1950</td><td>690</td><td>630</td><td>90</td><td>290</td><td>630</td><td>2.9</td><td>6.0</td><td>12.0</td></tr><tr><td>고려대학교</td><td>9.9</td><td>1870</td><td>729</td><td>649</td><td>123</td><td>280</td><td>616</td><td>3.4</td><td>6.1</td><td>15.8</td></tr><tr><td>중앙대학교</td><td>8.5</td><td>1700</td><td>700</td><td>630</td><td>90</td><td>280</td><td>620</td><td>4.0</td><td>6.0</td><td>14.0</td></tr></tbody></table>	기관	성장률	GNP	수출	수입	국제흑자	연말외채	연말환율	실업률	소비물가	임금상승	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	2.8	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	영서경제연구소	8.5	1950	690	630	90	290	630	2.9	6.0	12.0	고려대학교	9.9	1870	729	649	123	280	616	3.4	6.1	15.8	중앙대학교	8.5	1700	700	630	90	280	620	4.0	6.0	14.0	<pre>setwd("C:/빅데이터/R_data") Data0.1.4<-read.table("economicview.txt", header=T) X<-scale(Data0.1.4[, -1], scale=T) X<-as.matrix(Data0.1.4[, -1]) 기관=Data0.1.4[,1] #표준화 유클리드거리 D <- as.matrix(dist(X, method="euclidean")) # Metric MDS con<-cmdscale(D, k=2, eig=T) con x<-con\$points[,1] y<-con\$points[,2] #lim<-c(-max(abs(con\$points)), max(abs(con\$points))) lim1<-c(-max(abs(x)), max(abs(x))) lim2<-c(-max(abs(y)), max(abs(y))) plot(x,y, xlab="Dim1", ylab="Dim2", xlim=lim1, ylim=lim2) text(x,y, 기관, cex=0.8, pos=1) abline(v=0, h=0)</pre>
기관	성장률	GNP	수출	수입	국제흑자	연말외채	연말환율	실업률	소비물가	임금상승																																																																																																																																																												
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	2.8	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																																																																																																																																																												
영서경제연구소	8.5	1950	690	630	90	290	630	2.9	6.0	12.0																																																																																																																																																												
고려대학교	9.9	1870	729	649	123	280	616	3.4	6.1	15.8																																																																																																																																																												
중앙대학교	8.5	1700	700	630	90	280	620	4.0	6.0	14.0																																																																																																																																																												



MDS Map : cmdscale(D, k=2, eig=T)

\$GOF

[1] 0.9500013 0.9500013

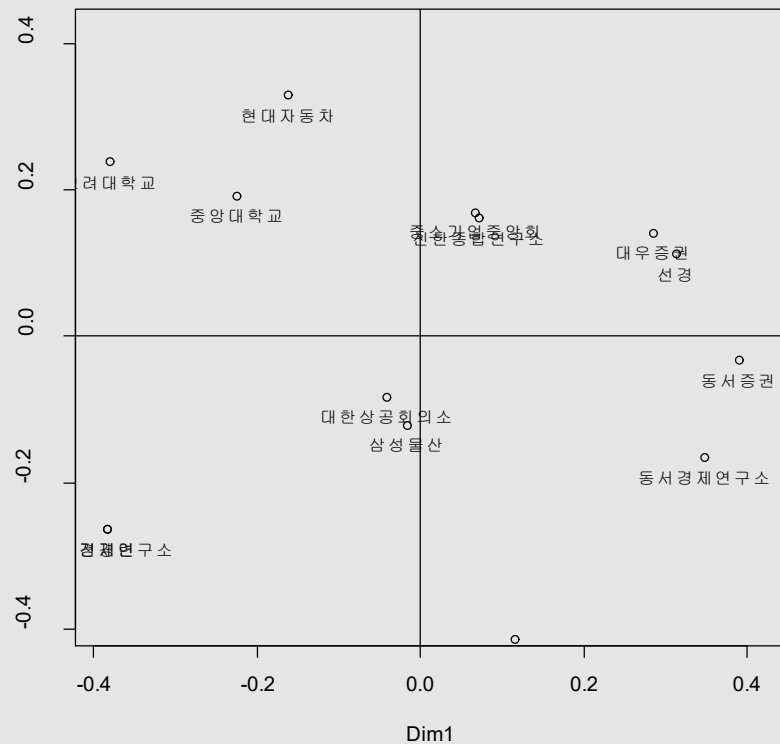
경제전망에 대한 MDS 그림의 군집 및 특성

사분면	기 관	특 성
제1	현대자동차, 중앙대학교, 고려대학교	정책비판
제2	중소기업중앙회, 대우증권, 신한종합연구소, 선정,	낙관론
제3	동서증권, 동서경제연구소, 한국은행	정책옹호
제4	삼성물산, 대우경제연구소, 전경련, 대한상공회의소	비관론

상자그림 : boxplot()

3.2.3 경제관련 기관의 경제전망 자료 : 이진수자료(Binary Data)

economicviewbinary.txt										economicviewbinary-MDS.R
이진수자료										<pre># Dissimilarity Matrix from Binary Data setwd("C:/빅데이터/R_data") binary<-read.table("economicviewbinary.txt", header=T) X<-binary[,-1] 기관<-binary[,1] n<-nrow(X) p<-ncol(X) # Dissimilarity Matrix from Binary Data m <-as.matrix(dist(X, method="euclidean", diag=T)) D<-round(m^2, 3)/p # Metric MDS win.graph() con<-cmdscale(D, k=2, eig=T) list(con\$points, con\$GOF) x<-con\$points[,1] y<-con\$points[,2] lim<-c(-max(abs(x)), max(abs(y))) plot(x,y, xlab="Dim1", ylab="Dim2", xlim=lim, ylim=lim) text(x,y,기관, cex=0.8, pos=1) abline(v=0, h=0)</pre>
1	2	3	4	5	6	7	8	9	10	
0	1	0	0	0	1	1	1	1	1	
1	1	1	1	1	1	1	1	1	1	
1	1	0	1	1	0	1	1	1	1	
0	0	0	0	0	1	0	1	0	0	
1	0	1	0	0	0	0	1	1	1	
1	1	1	1	0	1	0	0	1	1	
1	0	1	1	1	1	0	0	1	0	
0	0	1	0	1	1	1	1	1	1	
1	1	1	1	1	0	1	1	0	1	
0	0	0	0	0	1	0	1	0	0	
1	1	1	1	1	1	1	1	0	0	
1	1	0	1	0	0	1	1	1	1	
1	0	1	0	1	1	0	0	0	0	
1	0	1	1	0	1	0	0	1	0	
(1) 성장률이 8.5% 이상 (2) GNP는 1950불 이상 (3) 수출은 700억불 이상 (4) 수입은 630억불 이하 (5) 국제흑자는 100억불 이하 (6) 연말외채는 280억불 이하 (7) 연말환율은 1불당 630원 이상 (8) 실업률은 3.0% 이하 (9) 소비자 물가의 상승률은 6.0% 이하 (10) 임금 상승률은 12.5% 이하										



```
> list(con$points, con$GOF)
[[1]]
      [,1]      [,2]
 1  0.11501325 -0.41472290
 2  0.28486680  0.14126082
 3  0.38988825 -0.03242467
 4 -0.38263182 -0.26387796
 5 -0.04057270 -0.08267867
 6  0.06736468  0.16944139
 7 -0.16226512  0.32930325
 8 -0.01688406 -0.12131904
 9  0.31340651  0.11213031
10 -0.38263182 -0.26387796
11  0.07180745  0.16150207
12  0.34811650 -0.16601263
13 -0.38006950  0.23893100
14 -0.22540843  0.19234500

[[2]]
[1] 0.5666653 0.6940827
```

경제전망에 대한 MDS 그림의 군집 및 특성

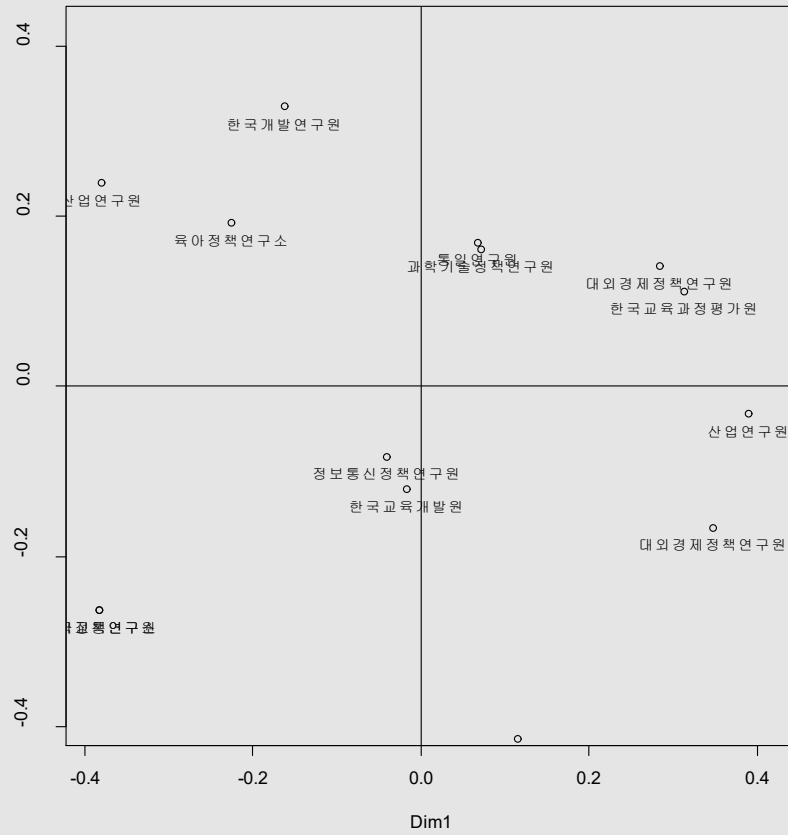
사분면	기 관	특 성
제2	현대자동차, 중앙대학교, 고려대학교	정책비판
제1	중소기업중앙회, 대우증권, 신한종합연구소, 선경,	낙관론
제4	동서증권, 동서경제연구소, 한국은행	정책옹호
제3	삼성물산, 대우경제연구소, 전경련, 대한상공회의소	비관론

결과 및 해석

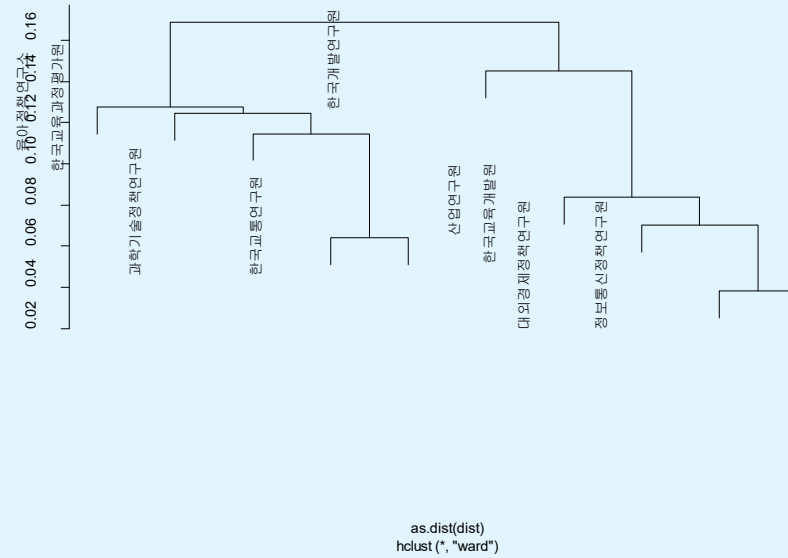
3.2.4 [자료 0.3.3] 비정형 텍스트 자료

- 정민지(2017). A Study on Clustering Methods for Proximity Data in Text Mining.

textdist.txt										
> round(dist,3)	과학기술정책연구원	대외경제정책연구원	산업연구원	육아정책연구소	정보통신정책연구원	통일연구원	한국개발연구원	한국교육개발원	한국교육과정평가원	한국교통연구원
과학기술정책연구원	0.000	0.077	0.100	0.102	0.084	0.099	0.140	0.066	0.099	0.064
대외경제정책연구원	0.077	0.000	0.063	0.101	0.038	0.097	0.107	0.061	0.107	0.097
산업연구원	0.100	0.063	0.000	0.122	0.068	0.121	0.121	0.090	0.132	0.118
육아정책연구소	0.102	0.101	0.122	0.000	0.105	0.130	0.152	0.086	0.125	0.111
정보통신정책연구원	0.084	0.038	0.068	0.105	0.000	0.103	0.111	0.064	0.114	0.103
통일연구원	0.099	0.097	0.121	0.130	0.103	0.000	0.150	0.095	0.132	0.110
한국개발연구원	0.140	0.107	0.121	0.152	0.111	0.150	0.000	0.120	0.161	0.152
한국교육개발원	0.066	0.061	0.090	0.086	0.064	0.095	0.120	0.000	0.100	0.081
한국교육과정평가원	0.099	0.107	0.132	0.125	0.114	0.132	0.161	0.100	0.000	0.105
한국교통연구원	0.064	0.097	0.118	0.111	0.103	0.110	0.152	0.081	0.105	0.000
기관별 유사성 거리 : 기관 내와 기관 간의 거리 평균										
textdist-MDS-CA.R										
<pre> setwd("C:/빅데이터/R_data") dist <- read.table("testdist.txt", header=T) 기관<-rownames(dist) # 1. distance matrix (Console Font: 편집-GUI설정-Font-DotumChe) # dist round(dist,3) # 2. MDS con<-cmdscale(dist, k=2, eig=T) list(con\$points, con\$GOF) x<-con\$points[,1] y<-con\$points[,2] lim<-c(-max(abs(x)), max(abs(y))) plot(x,y, xlab="Dim1", ylab="Dim2", xlim=lim, ylim=lim) text(x,y,기관, cex=0.8, pos=1) abline(v=0, h=0) # 3. cluster analysis # ward <- hclust(as.dist(dist),method="ward") plot(ward) </pre>										



MDS Map : `cmdscale(D, k=2, eig=T),
dist(X, method="euclidean", diag=T)`



군집분석 WARD법 : `hclust(,method="ward")`

Epilog : 허명회(2015), 최용석(2016)

Big Data : Big data is the data not touched and went yet

이제까지 다루지도 가본적도 없는 자료



Data Scientist : Frontier and Pioneer Spirit

- 능동적 태도 proactive attitude
- 문제 해결 능력 problem solving skill
- 창의력 creative mind
- 소통력 communication skill

⇨ 탐구/발표/협업/소통력/정보교류
(읽기, 말하기, 듣기, 쓰기 능력)

- 통계학 : 다변량통계학, EDA, 시각화기법,
Machine learning(알고리즘 중심 통계학)
<- 패키지 중심의 통계학 폐단
- 컴퓨터 : dB(Data Base), Web 기술 (Java, Python),
R은 필수
- 수학 : 미적분학, 행렬대수(대용량 자료의 표현)
- 적용분야 지식 : 역사, 경제, 사회과학, 공학
- 실용주의 철학 : 응용과 실천

References

범주형자료의 시각화 (Visualizations for CDA)

- 정광모, 최용석(2009). 범주형자료분석-SAS의 응용 및 해석-제2판, 자유아카데미.
- 최용석(2001). SAS 대응분석의 이해와 응용. 서울: 자유아카데미.
- 최용석(2018). R과 함께하는 다변량 자료분석, 경문사.
- 허명희(2010). R을 활용한 사회네트워크분석 입문, 자유아카데미
- 유충현, 홍성학(2015). R을 활용한 데이터시각화, 인사이트.
- Keen, K.J. (2010). Graphics for Statistics and Data Analysis with R, Chapman&Hall/CRC, New York.
- Wickham, H.(2016). ggplot2 - Elegant Graphics fir Data Analysis, Second Ed., Springer, AG.
- Sarkar, D.(2008). Lattice - Multivariate Data Visualization with R, Springer, New York.
- Choi, J. and Choi, Y.S. (2014). Standardizing Unstructured Big Data and Visual Interpretation Using MapReduce and Correspondence Analysis, *The Korean Journal of Applied Statistics*, 27(2), 169-183.
- Lee, C. and Choi, Y.S. (2014). Visualizations for Matched Pairs Models Using Modified Correspondence Analysis, *Communications for Statistical Applications and Methods*, 21(4), 275-284 .
- Kim, J.G., Choi, Y.S. and Cho i, S.K.(2012). Social Network Analysis and It's Applications for Co-authors and Key-words Networks in the JKSS, *Communications of the Korean Statistical Society* , 19(4), 547-558.
- 2005, Dynamic Simple Correspondence Analysis (with G.H. Hyun and M.R. Seo), *The Korean Communications in Statistics*, 12(1), 199-205 .
- 2005, A Numerical Comparison of Map Variability in SCA Using the Procrustes Analysis (with G.H. Hyun and J.G. Ki m), *Journal of the Korean Data Analysis Society*, 7(5), 1531-1538.
- 2002, Reinterpretation of Multiple Correspondence Analysis using the K-means Cluster Analysis(with G.H. Hyun and K. H. Kim), *The Korean Communications in Statistics*, 9(2), 505-514.
- 1999, Robust Simple Correspondence Analysis (with M.H. Huh), *Journal of the Korean Statistical Society*, 28(3), 337-346.

다변량자료의 시각화 (Visualizations for MDA)

- 유충현, 홍성학(2015). R을 활용한 데이터시각화, 인사이트.
- 최용석(2018). R과 함께하는 다변량 자료분석, 경문사.
- 최용석(2014). 다차원척도법의 산책, 자유아카데미.
- 최용석, 신상민 (2013). R을 활용한 행렬도 분석, 자유아카데미.
- Keen, K.J. (2010). Graphics for Statistics and Data Analysis with R, Chapman&Hall/CRC, New York.
- Wickham, H.(2016). ggplot2 - Elegant Graphics fir Data Analysis, Second Ed., Springer, AG.
- Sarkar, D.(2008). Lattice - Multivariate Data Visualization with R, Springer, New York.
- 정민지(2017). A Study on Clustering Methods for Proximity Data in Text Mining, 석사학위논문, 부산대학교.
- Shin, S.M., Park , J.S., Choi, Y.S. (2015). The mixed h-plot for continuous and categorical data, 18(1), Journal of the Korean Data Analysis Society
- Shin, S.M., Kim, E.S. and Choi, Y.S. (2015). Multidimensional scaling using the pseudo-points based on partition method, The Korean Journal of Applied Statistics, 28(6), 1171-1180.DOI: <http://dx.doi.org/10.5351/KJAS.2015.28.6.1171> (2015.12)
- Lee, S.G., Choi, Y.S. and Lee, B.H. (2014). Visualizations of Asymmetric Multidimensional Scaling, The Korean Journal of Applied Statistics, 27(4), 619-627.
- Lee, S.J. and Choi, Y.S. (2014). Independent Component Biplot, The Korean Journal of Applied Statistics, 27(1), 31-41 .
- Choi, T.H. and Choi, Y.S. (2013). Relationship between Physical Fitness and Basic Skill Factors for KTA Tennis Players Using the Semi-Partial Cannonical Correlation Biplot That Removed the Linear Effect of the Set of Covariate Variables from Physical Fitness Factor and Procrustes Analysis, Journal of the Korean Data Analysis Society , 15(6), 3067-3076.
- Choi, T.H. and Choi, Y.S. (2012). Relationship between Physical Fitness and Basic Skill Factors for KTA Players Using the Partial Cannonical Correlation Biplot Removing the Linear Effect of the Set of Covariate Variables and Procrustes Analysis, Communications of the Korean Statistical Society , 19(1), 97-106 .

- Lee, B.H., Choi, Y.S. and Shin, S.M.(2012). Semi-Partial Canonical Correlation Biplot, The Korean Journal of Applied Statistics , 25(3), 521-529.
- Yeom, A.R . and Choi, Y.S. (2011). Partial Canonical Correlation Biplot, The Korean Journal of Applied Statistics , 24(3),559-566 .
- Choi, T.H. and Choi, Y.S. (2010). A Study on the Relationship between Physique, Physical Fitness and Basic Skill Factors of Tennis Players of Korea Tennis Association Using the Generalized Canonical Correlation Biplot and Procrustes Analysis, The Korean Communications in Statistics , 17(6), 1-9.
- Hong, H.U., Choi, Y.S., Shin, S.M. and Kang, C.W . (2010). A Study on Shape Variability in Canonical Correlation Biplot with Missing Values, The Korean Journal of Applied Statistics , 23(5), 955-966
- Choi, T.H., Choi, Y.S. and Shin, S.M.(2009). A Study on the Relationship between Player Characteristic Factors and Competitive Factors of Tennis Grand Slams Competition Using Canonical Correlation Biplot and Procrustes Analysis, The Korean Journal of Applied Statistics, 22(4), 855-864.
- Shin, S.M., Choi, Y.S. and Lee, N.Y. (2008). Comparison of Shape Variability in Principal Component Biplot with Missing Values, The Korean Journal of Applied Statistics, 21(6), 1109-1116.
- Kim, H.Y. and Choi, Y.S (2008). Applications of Cluster Analysis in Biplots, Communications of The Korean Statistical Society, 15(1), 65-76.
- Choi, T.H. and Choi, Y.S. (2008). A Study on the Relationship between Skill and Competition Score Factors of KLPGA Players Using Canonical Correlation Biplot and Cluster Analysis, The Korean Journal of Applied Statistics, 21(3), 429-439.
- 2005, Biplot Analysis of Multinomial Logistic Regression Models (with C.W. Kang and K.D. Kim), Journal of the Korean Data Analysis Society, 7(3), 839-849.
- 2005, Biplot of the Parameter Matrix in MANOVA (with G.H. Hyun and S.M. Jung), Journal of the Korean Data Analysis Society, 7(3), 851-858.
- 2005, Biplots' Variability based on the Procrustes Analysis (with G.H. Hyun and W.J. Woo), Journal of the Korean Data Analysis Society, 7(6), 1925-1933.
- 2005, MANCOVA Biplot (with G.H. Hyun and S.M. Jung), The Korean Communications in Statistics, 12(3), 705-712 .
- 2000, The Development of Biplots System (with G.H. Hyun), The Korean Journal of Applied Statistics, 3(2), 297-306.
- 2000, The Exploratory Approach of Baseline-Category Logit Models using the Biplot (with K.D. Kim), Journal of Data Sciences and

Classifications, 4, 35-47.

- 1998, Robust Weighted Nonmetric Multidimensional Scaling (with I.T. Jolliffe and J.R. Lee), Journal of Data Sciences and Classifications, 2, 33-47.
- 1995, Resistant h-plot for a Sample Variance-Covariance Matrix. Journal of The Korean Statistical Society, 24(2), 407-417.