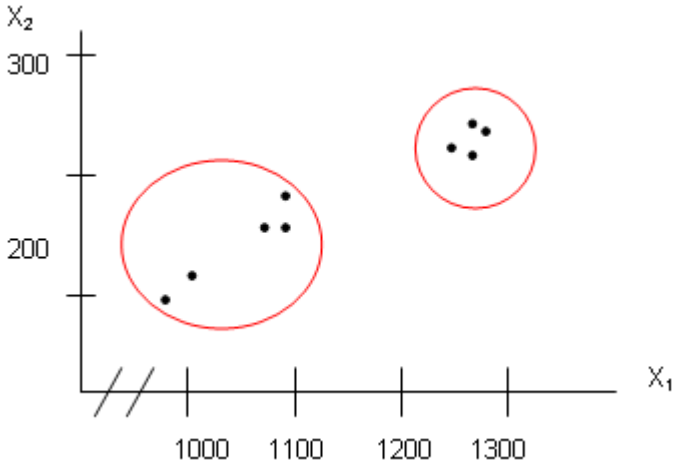
CHAPTER 7

군집 분석

마케팅 담당자가 자사 고객들을 분류하기 위하여 나이, 학력, 소득, 결혼 상태, 자녀 수, 직업 등에 대한 정보를 수집하였다. 이를 이용하여 고객들을 집단(cluster)으로 분류하고자 할 때 사용되는 다변량 분석 방법이 군집 분석 (Clustering Analysis)이다. 판별 분석과 군집 분석의 다른 점은 판별 분석은 조사된 데이터에 개체의 집단 변수가 이미 포함되어 있으나 군집 분석은 개체들에 대해 측정된 변수에 의해 집단을 분류하게 되므로 집단의 개수와 집단의 종류(이름)는 분류 후 정해지게 된다. 즉 군집 분석은 개체들이 분석 전에는 어떤 그룹에 속하는지 알려져 있지 않다. 군집 분석은 grouping 혹은 classification 이라 불리기도 한다.

개체를 분류한다? 아래 산점도와 같이 측정 변수가 2 개라면 거리가 가까운 개체들끼리 묶으면 될 것이다. 12 개의 개체가 아마 2 개 군집(집단)으로 분류될 것이다. 두 개체간의 Euclidean 거리를 계산(측정)하고 거리가 가까운(유사성이 높음) 개체끼리 묶으면 된다.



이렇게 개체를 분류하는데 도움을 주는 그림은 1)측정 변수가 2 개인 경우 산점도(scatter plot) 2)측정 변수가 3 개인 경우는 Bubble Plot 3)측정 변수가 4 개 이상이 경우는 주성분 변수를 이용하여 2-3 개의 주성분에 대한 산점도가 있다. 그러나 이런 그래프들을 이용하여 시각적 분류는 가능하지만 실제 유사성에 의해 개체를 분류하려면 유사성이 값으로 계산되어야 한다.

7.1. 유사성과 비 유사성

7.1.1. Euclidean 거리

두 개체 사이의 우사 정도를 거리로 표현할 수 있다. 거리가 멀면 유사성(similarity)이 떨어진다. 다음 식은 r 번째 개체와 s 번째 개체의 Euclidean 거리이다.

$$d_{rs} = [(\underline{x}_r - \underline{x}_s)'(\underline{x}_r - \underline{x}_s)]^{1/2}$$

7.1.2. 표준화 Euclidean 거리

개체들에 대해 측정된 변수들이 단위가 다르거나 분산이 다를 경우 변수를 표준화한 후 거리를 구하는 것이 더 적절하다. 다음 식은 r 번째 개체와 s 번째 개체의 표준화 Euclidean 거리이다.

$$d_{rs} = [(\underline{z}_r - \underline{z}_s)'(\underline{z}_r - \underline{z}_s)]^{1/2}$$

7.1.3. Mahalanobis 거리

다음 식이 r 번째 개체와 s 번째 개체의 Mahalanobis 거리이고 Σ는 within 군집 분산-공분산 행렬 추정치이다. 거리를 이와 같이 정의하는데 문제점은 개체를 다 분류하기 전에는 Σ의 추정치를 구할 수 없다는 것이다.

$$d_{rs} = [(\underline{x}_r - \underline{x}_s)'\Sigma^{-1}(\underline{x}_r - \underline{x}_s)]^{1/2}$$

7.2. 군집 분석 방법

7.2.1. 비계층적 방법

군집의 중심이 되는 seed 점들 집합을 선택하여 그 seed 점과 유사성이 높은(거리가 가까운) 개체들을 묶는(그룹화) 방법이다. 이 방법은 다음 3 가지 문제점을 갖고 있다. 1)사전에 군집(그룹) 수에 대한 예상이 필요하다. 2)개체 분류는 처음 선정한 seed 점들에 의해 영향을 많이 받고 분석자 마다 분류가 다를 가능성이 있다. 3)군집의 수와 seed 값의 위치의 결합 조건이 너무 많아 계산이 분류를 위한 계산이 용이하지 않다. 이 방법에 대한 SAS procedure 은 FASTCLUS 이다.

7.2.2. 계층적 방법

유사성이 가까운 순서대로 개체들을 묶어(군집화) 가는 방법으로 single-linkage clustering 방법이 이 방법 중 가장 효율적이다. Neighbor Method 은 single-linkage clustering 방법 중 하나로 다음 순서에 의해 개체를 분류한다.

(1)처음에는 개체의 수(n)만큼의 군집이 있다. 예를 들어 개체 6 개가 있고 다음은 각 개체 간 Euclidean 거리(유사성)를 계산한 표이다. 처음에는 군집은 6 개이다.

	1	2	3	4
1		0.1	0.7	0.2
2			0.4	0.6
3				0.3
4				

두 개체간의 거리이므로 대각 원소는 동일하다.

(2)유사성이 가장 가까운(거리가 가장 가까운) 개체를 군집으로 묶는다. 예제에서는 (3.5)가 묶인다.

	(1, 2)	3	4
(1, 2)		?	?
3			0.3
4			

?: 어떤 값으로 개체 집단(1, 2)와 개체)의 유사성(거리)을 측정할 것인가?

(3)개체가 군집으로 묶이면 개체와 새로 만들어진 군집과의 유사성을 계산한다. 군집과 군집(혹은 개체)의 유사성(거리)을 측정하는 방법은 다음 5 가지가 있다.

①Nearest neighbor: 두 군집의 각 개체 중 가장 가까이 있는 개체의 거리(유사성)

②Furthest neighbor: 두 군집의 각 개체 중 가장 멀리 있는 개체의 거리

③Centroid neighbor: 군집의 평균 간의 거리

④Average neighbor: 한 군집의 개체와 다른 군집 개체들의 각 거리 평균

⑤Ward's minimum variance: 군집의 평균간 거리를 각 군집의 개체 개수의 역의 합으로 나눈 제곱근을 구한 거리이다.

Nearest, Furthest, Centroid neighbor, Average neighbor, Ward's minimum variance 중 어떤 방법을 사용하는 것이 좋은가? Nearest 방법은 개체간의 거리가 가까워 개체를 묶는 경향이 있어 군집의 수가 줄어들고 Furthest 는 군집간 거리를 최소화 하는 경향이 있어 개체 수가 적은 군집을 얻게 한다. 그러므로 각 방법의 장단점이 있으므로 2-3 개 방법을 사용하여 개

체의 군집화가 보다 잘되는 방법을 선택하는 것이 좋다. 가장 많이 사용하는 방법은 Average neighbor 방법이다.

(4)다음은 Nearest neighbor 방법에 의해 개체를 군집화 하는 과정이다.

1 과 3 의 거리는 0.7, 2 와 3 의 거리는 0.4 이므로 (1, 2)와 3 의 거리는 0.4 가 된다. 1 과 4 의 거리는 0.2 이고 2 와 4 의 거리는 0.6 이므로 작은 거리 0.2 가 (1, 2)와 4 의 거리이다.

	(1, 2)	3	4
(1, 2)		0.4	0.2
3			0.3
4			

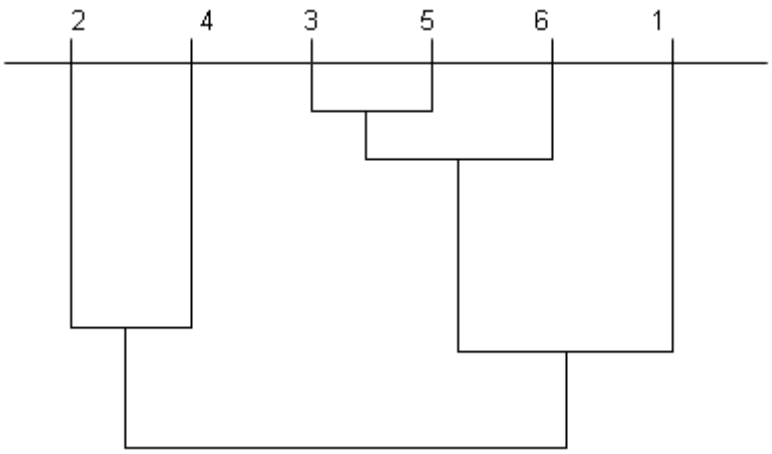
(1, 2)의 거리는 0.4 이고 3 과 4 의 거리는 0.3 이므로 (1, 2, 4)와 3 의 거리는 0.3 이 된다.

	(1, 2, 4)	3
(1, 2, 4)		0.3
3		

7.3. 군집 개수

군집의 개수를 몇 개로 하면 좋은가? 그래프적 방법으로는 Tree diagram 이 있고 검정 통계량을 이용하는 방법으로는 Hotel ling's T^2 검정이나 Cubic Clustering Criterion 방법을 이용하면 된다.

7.3.1. 계층적 나무 다이어그램



위의 그림은 나무 그림(Tree Diagram)이라 하여 선의 길이는 개체간, 개체와 군집간, 군집간 유사성(거리)이다. 이 다이어그램(diagram)에 의하면 2 개의 군집 (2, 4), (3, 5, 6, 1)으로 분류하는 것이 타당해 보인다.

7.3.2. Pseudo Hotel ling's T^2 검정

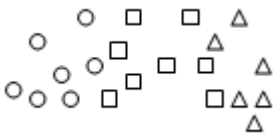
Hotel ling's T^2 검정 통계량은 두 집단 다변량 평균의 차이를 보는 통계량이다. 이를 군집분석에 이용하는데…… 이와 유사 개념의 검정 통계량을 이용하여 개체의 군집간 평균의 차이가 유의하지 않으면 두 군집을 합치고 유의하면 군집 그대로 유지하는 방법이다.

7.3.3. CCC

Searle(1983)이 제안한 방법으로 군집의 개수와 CCC(Cubic Clustering Criterion)의 산점도를 그려 CCC 의 값이 3 이상이고 최대 값인 경우 그 때의 군집의 개수가 적당하다. 이용 방법은 예제에서 살펴보기로 한다.

7.4. 판별 분석과 군집 분석의 비교

	판별 분석(Discriminant Analysis)	군집 분석(Clustering Analysis)

분석 초기	개체들은 이미 분류되어 있다. 	개체들을 측정 변수에 의해 분류한다. 
	판별 변수 $(X_1, X_2, ..., X_P)$ 분류 변수	
목적	새로운 개체를 분류 ▸개체를 잘 판별할 수 있는 판별 변수 선택이 관건이다.	위의 개체들을 분류 ▸개체들의 특성을 나타내는 변수들을 선택하는 것이 관건
분석 순서	(1)판별 분석 방법 선택 →오분류가 적은 방법 사용 •Fisher method (판별 변수 선택 방법 이용, 유의 수준을 다소 높게 설정) •K Nearest Discriminant Analysis •Logistic Regression 판별 분석(변수 선택 방법 사용, 유의수준 다소 높게 설정) (2)개체 분류 경향을 파악하기 위하여 판별 변수들에 산점도(by 그룹)를 그린다. 판별 변수가 2 개 이상이면 주성분 분석을 이용하여 산점도를 그리면 된다. (3)최종적으로 구해진 판별식에 의해 새로운 개체를 (□△○) 중 하나로 분류한다.	(1)개체 분류 방법을 선택한다. •Nearest neighbor •Furthest neighbor •Centroid neighbor •Average neighbor •Ward's minimum variance (2)군집의 개수를 정한다. •CCC •Pseudo Hotel ling's T^2 •Tree Diagram (3)개체 분류가 잘 되었는지 알아보기 위하여 산점도를 그린다. 변수가 3 개 이상인 경우는 주성분 분석을 이용하여 산점도 그린다. 군집 결과는 개체 분류 방법과 군집 개수에 의해 결정된다. (4)각 군집에 적절한 이름을 붙인다. 

7.5. 예제

56 개 피자 제품에 대해 MOIS(수분 함유), PROT(단백질 함유량), FAT(지방 함유량), ASH(ash 함유량), SODIUM(나트륨 함유량), CARB(탄수화물 함유량), CAL(칼로리)를 조사하였다. 이를 이용하여 56 개 피자 제품을 분류하여 보자. PIZZA.txt/[Applied Multivariate Methods for Data Analysts, Dallas E. Johnson, 1998, p331]

7.5.1. 프로그램

```
DATA PIZZA;
    INFILE 'D:\TEMP\PIZZA.TXT';
    INPUT ID MOIS PROT FAT ASH SODIUM CARB CAL;
RUN;

PROC CLUSTER DATA=PIZZA STANDARD METHOD=AVERAGE
                        CCC PSEUDO OUTTREE=OUT1;
    VAR MOIS PROT FAT ASH SODIUM CARB CAL;
    ID ID;
RUN;
```

(1)STANDARD 옵션은 Standardized Euclidean Distance 거리에 의해 개체간 유사성을 측정한다. 각 분류 변수의 측정 단위가 다르므로 STANDARD 옵션을 사용하였다.

(2)군집과 개체를 묶어갈 때 Average neighbor(군집내의 개체 각각의 거리 평균) 방법을 사용하였다. (METHOD=AVERAGE)

①AVERAGE | AVE average linkage vs. CENTROID | CEN centroid method

②COMPLETE | COM complete linkage (**furthest neighbor**, maximum method, diameter method, rank order typal analysis).

③DENSITY | DEN density linkage, which is a class of clustering methods using nonparametric probability density estimation. You must also specify one of the K=, R=, or HYBRID options

④EML maximum-likelihood hierarchical clustering

⑤MEDIAN | MED Gower's median method

⑥SINGLE | SIN single linkage (**nearest neighbor**, minimum method, connectedness method, elementary linkage analysis, or dendritic method).

⑦WARD | WAR Ward's minimum-variance method

(3)CCC 는 Cubic Clustering Criterion 값을, PSEUDO 는 Pseudo Hotel ling's T^2 검정 통계량을 출력하라는 것이다. 이 통계량은 군집 개수를 결정하는데 사용된다.

(4)군집 분석 결과가 TREE 라는 SAS data 에 저장된다.

7.5.2. 출력 결과

분류 변수 7 개의 표본 상관 행렬로부터 주성분 분석 결과이다. 7 개 변수의 총 변동을 설명하려면 주성분 개수가 2 개이면 충분하다. 개체 분류 후 주성분 변수의 산점도를 이용하여 분류 결과를 시각적으로 표현 할 수 있다. 주성분 분석을 실시한 이유는 주성분 변수를 이용하여 산점도를 그리기 위해서이다. 이 산점도는 개체들을 군집화(그룹화) 한 후 각 개체 집단의 이름을 부여하거나 군집 간 차이를 보는데 사용된다.

	Eigenvalue	Difference	Proportion	Cumulative
1	3.90915703	1.38581840	0.5585	0.5585
2	2.52333862	2.08087020	0.3605	0.9189
3	0.44246842	0.34408358	0.0632	0.9821
4	0.09838484	0.07180589	0.0141	0.9962
5	0.02657895	0.02651569	0.0038	1.0000
6	0.00006327	0.00005440	0.0000	1.0000
7	0.00000887		0.0000	1.0000

The data have been standardized to mean 0 and variance 1

7.5.3. 개체 군집 과정

Cluster History

NCL	Clusters Joined		FREQ	Norm RMS Dist	Time
55	34021	34026	2	0.0203	
54	24107	34022	2	0.0304	
53	14072	24030	2	0.0366	
52	CL54	CL55	4	0.0386	
51	24049	24033	2	0.0402	
50	14118	14143	2	0.0418	
49	CL52	14067	5	0.0426	
48	34037	34034	2	0.0427	
47	14047	14074	2	0.0432	
46	14099	14122	2	0.0507	
45	14100	24100	2	0.0524	
44	24056	24069	2	0.0567	
43	24071	34039	2	0.0618	
42	CL48	CL51	4	0.0619	
41	14025	14164	2	0.068	
40	14166	24119	2	0.073	

(1)중간에 필요 없는 부분은 제외하였다.

(2)개체 분류 순서가 나타난다. 제일 먼저 피자 34021 와 피자 34026 가 묶인다. 두 피자(개체)의 유사성(Norm Distance): 즉 거리)은 0.0203 이다.

(3)두 번째는 피자 24107 와 피자 34022 가 묶인다. 유사성은 0.0304

(4)세 번째는 피자 14072 와 피자 24030 가 묶인다. 유사성은 0.0366

(5)네 번째는 CL54(군집 54: 피자 24107, 파자 34022)와 군집 55(피자 34021, 피자 34026)이 서로 묶인다.

(6)다섯 번째는 피자 24049, 피자 24033 이 여섯 번째는 피자 14118, 피자 14143 이 묶인다.

(7)일곱 번째는 군집 52(군집 54, 군집 55)에 피자 14067 이 묶인다. ....

(8)FREQ 는 군집에 들어가는 개체의 개수를 나타낸다. 개체끼리 묶이는 경우는 2 개이고 집단과 집단이 묶이는 경우는 집단 내의 개체 수이다.

7.5.4. 군집 개수 결정

다음은 군집의 개수를 결정하기 위한 통계량 부분을 출력한 것이다. 위의 결과 바로 아래 출력된다. NCL 은 Number of Clustering 의 약어로 군집의 개수이다. 만약 유사성이 Tie(동점)이 되지 않는 한 개체는 하나씩 분류된다. Tie 가 발생하면 제일 마지막 열(Tie)에 표시된다.

Cluster History

NCL	------Clusters Joined-------		FI	CCC	PSF	PST2	Norm RMS Dist
14	CL22	CL44		.	235	9.2	0.2099
13	CL24	14126		.	235	6.2	0.2284
12	CL28	CL20		.	231	9.1	0.2318
11	CL13	CL30		17.8	227	5.8	0.2426
10	CL15	CL25		17.8	226	5.5	0.2579
9	CL11	CL18		15.8	187	17.2	0.2743
8	CL16	CL29		16.2	196	9.1	0.2942
7	CL12	14140		14.1	215	4.1	0.323
6	CL37	CL14		12.3	178	56.7	0.379
5	CL10	CL7		11.5	164	20.1	0.4551
4	CL8	CL5		9.36	131	21.1	0.6532
3	CL9	CL6		3.64	72.9	126	0.7578
2	CL4	CL3		-.81	36.5	61.0	1.0267
1	CL17	CL2		0.00	.	36.5	1.6634

(1)CCC(Cubic Clustering Criterion, Searle; 1983) 값은 자료 수의 20%까지만 출력된다. CCC 값이 3 이상이고 갑자기 줄어드는 부분이면 적당하다. CCC 에 의하면 군집의 개수는 4 개가 적당하다고 결론 내릴 수 있다.

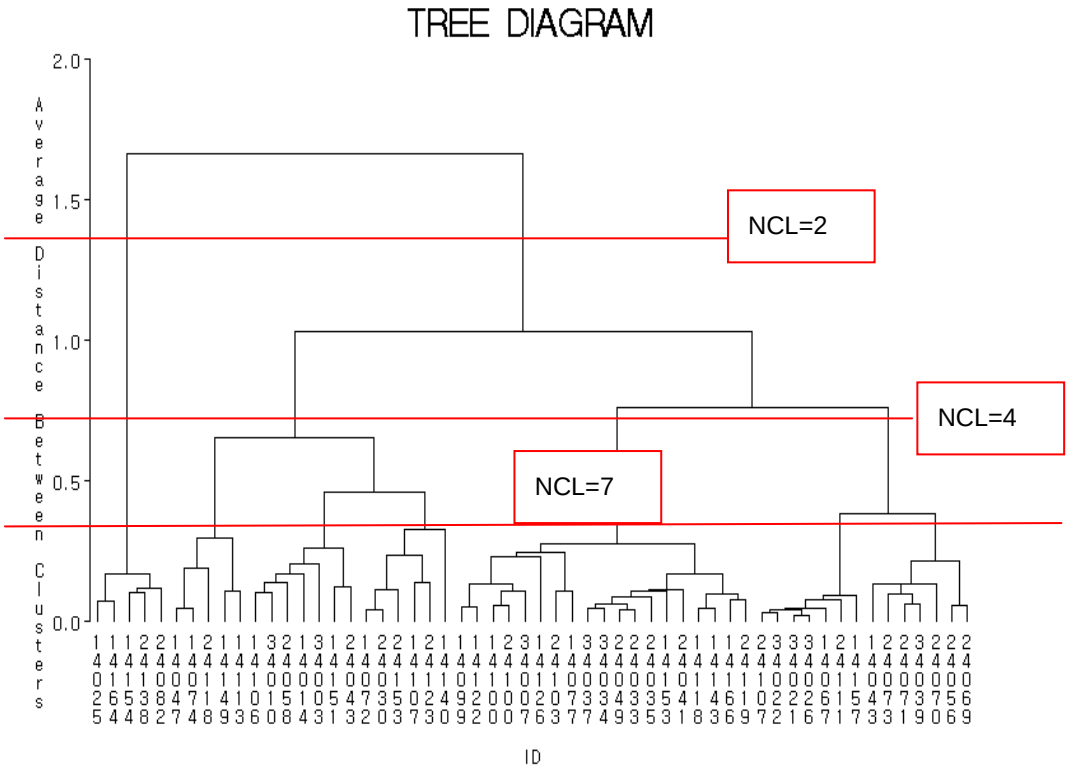
(2)Hotelling's T^2 검정 통계량은 두 다변량 정규 분포 집단의 평균을 비교하는데 사용된다. 이 개념을 이용하여 PST2(Pseudo Hotelling's T^2)는 두 군집을 하나로 합칠 수 있는가를 알아보는 기준이 된다. PST2 값이 크다는 것은 군집간 거리가 멀다는 것을 의미하므로 군집을 나누는 것이 좋다. 값이 적다는 것은 군집간 거리가 가깝다는 것으로 군집으로 합치는 것이 좋다. PST2 의 경우 NCL=6, NCL=3 인 경우 주변 값들보다 크므로 이 값 바로 전의 군집 개수(NCL)가 적절하다. 즉 NCL=6 를 고르면 최적 군집의 개수는 7 이고 NCL=3 를 고르면 최적 군집의 개수는 7 이다.

(3)PSF 는 Pseudo F 통계량으로 PST2 Pseudo Hotelling's T^2 검정 통계량과 유사하다. 그러나 일반적으로 CCC 와 PST2 에 의해 군집 수를 결정한다.

(4)CCC 에 의해서는 4 개, PST2 에 의해 4 개나 7 개가 적당하다.

7.5.5. Hierarchical Tree 다이어그램

```
PROC TREE DATA=OUT1 OUT=OUT2;
    COPY MOIS PROT FAT ASH SODIUM CARB CAL;
    ID ID;
RUN;
```



7.5.6. 개체 분류 결과 출력하기

군집의 개수가 정해지면(예제는 4 개나 7 개) 개체 군집의 이름을 부여하기 위하여 개체의 분류 변수와 군집을 출력하거나 그래프(산점도)를 이용해야 한다. 다음은 개체를 7 개로 분류한 경우 결과를 출력하거나 그래프를 그려 군집의 이름을 부여하는 프로그램이다.

```
TITLE 'No. of CLUSTER=7';
PROC TREE DATA=TREE OUT=TREEOUT NCLUSTERS=7;
 COPY MOIS PROT FAT ASH SODIUM CARB CAL;
 ID ID;
 RUN;

PROC SORT DATA=TREEOUT; BY CLUSTER;
PROC PRINT DATA=TREEOUT;
 VARIABLES ID CLUSTER;
RUN;
```

군집 개수=7

Obs	ID	CLUSTER
1	34021	1
2	34026	1
3	24107	1
4	34022	1
5	14067	1
6	24111	1
7	14157	1
8	14072	2
9	24030	2
10	24153	2
11	14107	2
12	24123	2
13	14140	2

군집 1 로 분류된 피자 7 개를 보고 군집의 이름을 붙이면 된다. 예를 들어 이 피자들이 2001 피자이면 군집 1 은 2001 피자 군집이라 이름 붙인다. 그러나 변수가 많은 경우 묶여진 개체들을 보고 이름을 붙이는 것은 쉽지 않다. 그러므로 변수를 축약한 주성분 변수를 이용하여 군집에 대한 이름을 붙으면 편리하다.

7.5.7. 군집 이름 부여하기

개체 군집 결과를 위의 출력 형태보다는 그래프로 보면 이름 붙이기 더 쉬울 것이다. 개체 군집이 나타날 수 있도록 산점도를 그려 보자. 그런데 불행히도 각 피자에 대해 측정 변수가 7 개(MOIS, PROT, FAT, ASH, SODIUM, CARB, CAL)이므로 하나의 산점도로는 표현할 수 없다. 그리고 군집 분석에서는 분류 변수(군집 분석에 이용되는 변수)를 선택한다는 것은 전혀 의미가 없다.

유용한 그래프 산점도를 어떻게 그리지? 변수가 2 개라면 몰라도... 아 어떻게 하지. 고민하지 말자. 우리는 변수의 개수를 축약하는 방법인 주성분 분석을 알고 있다. 주성분 분석에

의해 분류 변수를 축약하고 주성분 분석으로 산점도를 그리자. 주성분 개수가 2 개 이하이면 더 말할 나위 없이 좋지만 3 개라도 산점도 2 개만 그리면 되니 별 문제는 없다.

```
TITLE 'Scatter plot of PRIN1*PRIN2 NCL=7';
PROC PRINCOMP DATA=PIZZA OUT=SCORES;
 VAR MOIS PROT FAT ASH SODIUM CARB CAL;
 RUN;
PROC SORT DATA=TREEOUT; BY ID;
PROC SORT DATA=SCORES; BY ID;
DATA COMB; MERGE SCORES TREEOUT; BY ID;
PROC PLOT DATA=COMB;
 PLOT PRIN1*PRIN2=CLUSTER/VPOS=25 HPOS=50;
RUN;
```

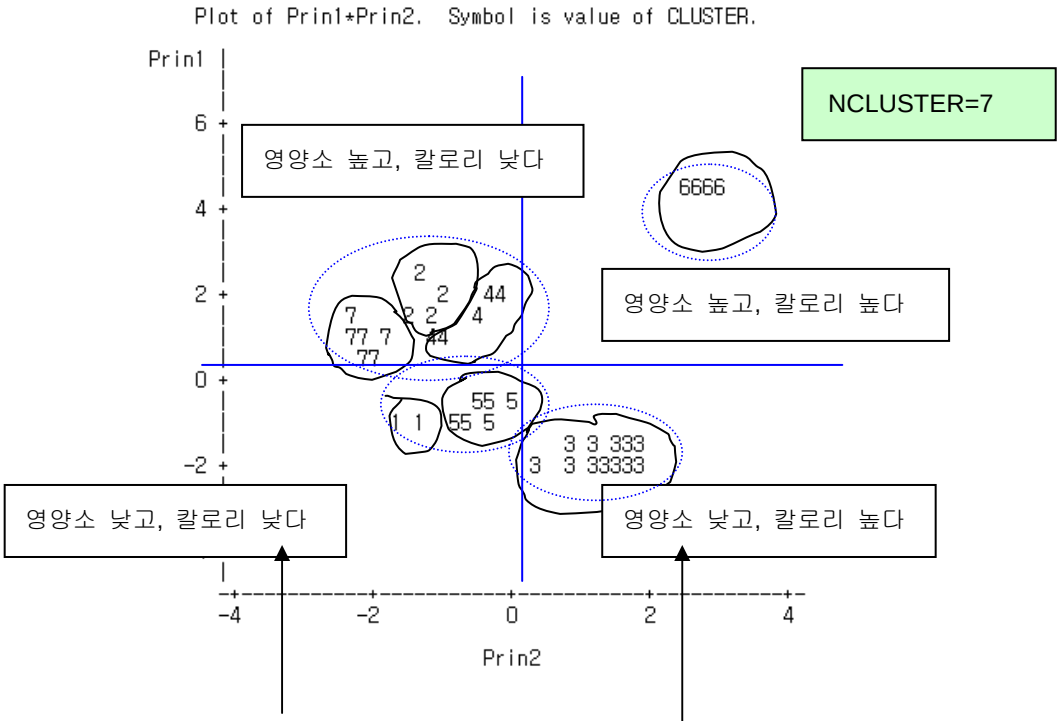
HPOS=50 옵션은 산점도 그림 크기를 줄이는 옵션으로 그래프를 수평(H)은 50%로 VPOS=50 옵션은 수직(V)은 25% 크기로 축소하여 그린다.

주성분 분석(상관 계수 이용: 측정 단위가 다르므로) 결과 중 주성분 계수를 출력한 것이다. 페이지 152 에서 우리는 주성분 변수 2 개로 전체 6 개 변수의 변동 중 91.8%가 설명됨을 알았다. Prin1 은 영양소(PROT, FAT, ASH, SOLDIUM, CARB(탄수화물 반대 작용)) 함유량 주성분이고 Print2 는 수분과 칼로리(서로 반대 개념) 주성분이다. 그래서 주성분 1 을 영양소 변수, 주성분 2 를 칼로리 변수로 이름 붙였다.

Eigenvectors

	Prin1	Prin2	Prin3	Prin
MOIS	0.072468	-.591440	0.457949	-.2013
PROT	0.376602	-.287747	-.724057	-.0640
FAT	0.439164	0.281253	0.210489	-.5125
ASH	0.486412	-.108591	-.093574	0.5523
SODIUM	0.440202	0.227130	0.432383	0.4494
CARB	-.431373	0.319985	-.074430	0.3455
CAL	0.208799	0.567914	-.143065	-.2571

주성분 변수에 이름을 붙이는 이유는 개체(군집)에 대한 산점도를 그린 후 주성분 변수에 의해 군집에 적절한 이름을 부여하기 위함이다. 다시 강조하지만 군집 분석에는 어디에도 집단을 나타내는 변수가 측정(조사)되지 않는다. 군집 분석 후 각 개체 군집에 적절한 이름이 부여되는 것이다.

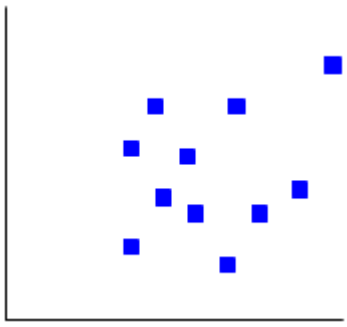


산점도의 점들은 군집 번호로 표시되었으므로 이를 이용하여 각 군집에 적절한 이름을 부여하면 된다. 만약 군집을 4 개로 분류한다면 원과 같이 개체가 분류될 것이다. 위의 산점도에 의하면 7 개 집단보다는 4 개 집단으로 하는 것이 더 적절할 것이다. 파란 색 직선에 의해 각 집단의 이름 부여하기도 4 개 집단이 더 적절할 것이다.

7.6. Faster Cluster 군집 분석

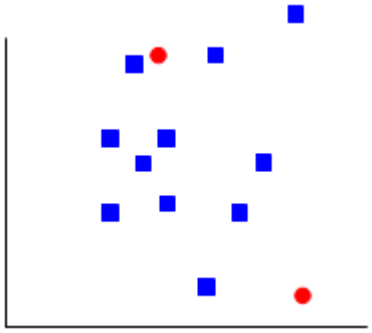
Faster clustering 방법은 앞 절에서 살펴보았던 계층적 군집(hierarchical clustering) 방법과는 [유사성(거리)이 가까운 개체들을 차례로 군집으로 묶어가는 방법] 달리 비계층적 군집(non-hierarchical clustering) 방법이다. 우선 seed 를 정하고 이 seed 에 가까운 개체들을 군집으로 묶는다. 그러므로 군집의 개수를 분석 전에 정해야 하며(number of clusters) 군집의 크기(size: 이는 radius 로 설정)를 정해 주어야 한다. 비계층적 군집 방법의 순서는 다음과 같으며 SAS 에서는 FASTCLUS 에서 이 방법으로 군집 분석할 수 있다.

Faster Cluster 방법을 이해하기 위하여 예를 들어 설명하기로 한다.

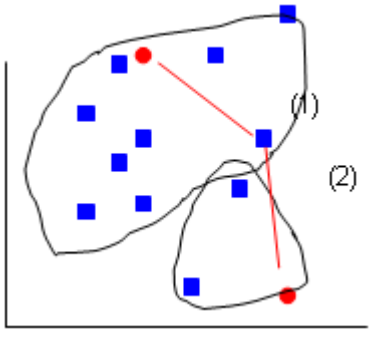


(STEP1)군집 seeds 가 선택된다. 초기 seed 의 개수는 분석자가 정해 주는 개수대로 된다.

MAXCLUSETRS=2 옵션을 사용했을 경우 초기 SEED 개수는 2 개이다.

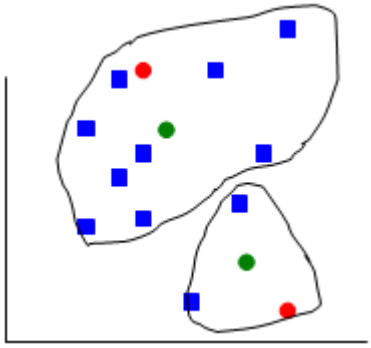


(STEP2)개체들을 가장 가까운 군집 seeds 에 묶는다. 만약 이 경우 DRIFT 옵션을 쓰면 seed 가 임시 군집의 평균으로 옮겨가며 개체를 묶는다.

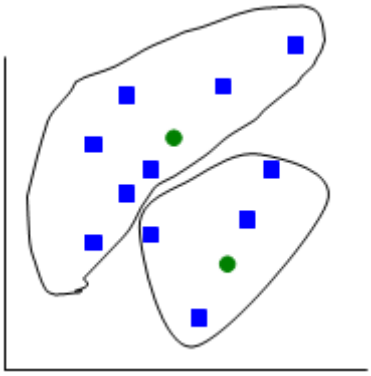


(1)이 (2)보다 거리가 가깝다.

(STEP3)일단 개체 군집이 끝나면 군집 개체의 평균을 SEED 로 하여 개체를 다시 분류한다.



(STEP4)군집이 끝나면 STEP 2)-STEP3)을 반복한다. 다음 STEP 에서는 SEED 가 빨강에서 초록색으로 옮겨간다. 그리고 위의 STEP 을 반복한다. MAXITER 옵션이 없으면 개체가 군집 분류가 이전과 같을 때까지 반복한다. MAXITER=3 이라 하면 위의 STEP 을 3 번 반복한다.



Non-hierarchical clustering 방법의 또 하나의 결점은 자료 입력 순서에 따라 군집이 달라지므로 RANDOM 이라는 옵션을 사용해 자료 입력 순서를 바꾸어 가며 군집하게 한다. 이처럼 FASTCLUS 방법도 하나의 방법이 있는 것은 아니다. 군집 분류 결과의 산점도(변수가 3 개 이상인 경우는 주성분 분석을 이용하여 PRIN1, PRIN2 를 이용한다.) 살펴보고 군집 내 개체간 거리 분산이 작은 것들을 택하면 된다.

7.6.1. 예제 자료

Fisher 의 Iris(꽃) 자료[IRIS.TXT]이다. 이것에 대해 4 가지 특성을 조사하였다. 물론 이 꽃은 어떤 종인지 (VARIETY: S, C, V ► 3 종류) 이미 알고 있지만 종을 모른다고 가정하고 군집 분석을 실시해 보자.

꽃잎 길이(petal length), 꽃잎 넓이(petal width), 수술 길이(stamen length), 수술 넓이(stamen width)

```
DATA IRIS;
    INFILE "D:\TEMP\IRIS.TXT";
    INPUT TYPE $ SL SW PL PW;
RUN;

PROC PRINT DATA=IRIS;
RUN;
```

Obs	TYPE	SL	SW	PL	PW
1	S	5.1	3.5	1.4	0.2
2	S	4.9	3.0	1.4	0.2
3	S	4.7	3.2	1.3	0.2

7.6.2. 변수 표준화

PROC CLUSTER 와는 달리 FASTCLUS 에는 변수(자료) 표준화 옵션이 없다. 그러므로 변수 측정 단위가 다르거나 분산 차이가 많으면 분석자가 직접 표준화 procedure 를 사용하여 표준화 해야 한다.

```
PROC STANDARD DATA=IRIS OUT=IRISO MEAN=0 STD=1;
    VAR SL SW PL PW;
RUN;
PROC PRINT DATA=IRISO;
RUN;
```

Obs	TYPE	SL	SW	PL	PW
1	S	-0.89767	1.01560	-1.33575	-1.31105
2	S	-1.13920	-0.13154	-1.33575	-1.31105
3	S	-1.38073	0.32732	-1.39240	-1.31105
4	S	-1.50149	0.09789	-1.27910	-1.31105

(SL, SW), (PL, PW)의 측정 단위가 차이가 있으므로 표준화 변수를 사용하여 군집을 분석해 보자.

7.6.3. 주성분 분석

Fast cluster 방법을 사용하려면 군집의 개수에 대한 정보가 필요하다. IRIS 자료처럼 집단의 수(3 개: S, C, V)를 알고 있다면 이를 사용하면 되지만 그렇지 않은 경우는 변수들간의 산점도를 그려 개체 분류 개수를 예상하면 된다. 그러나 변수가 많은 경우는 산점도를 동시에 고려한 집단 분류는 쉬운 문제가 아니므로 주성분 분석을 이용하여 2~3 개 정도의 주성분 변수간 산점도를 그리면 된다.

```
PROC PRINCOMP DATA=IRISO OUT=IRIS_PRIN;
    VAR SL SW PL PW;
RUN;
```

Eigenvalues of the Correlation Matrix

	Eigenvalue	Difference	Proportion	Cumulative
1	2.91849782	2.00446735	0.7296	0.7296
2	0.91403047	0.76727360	0.2285	0.9581
3	0.14675688	0.12604204	0.0367	0.9948
4	0.02071484		0.0052	1.0000

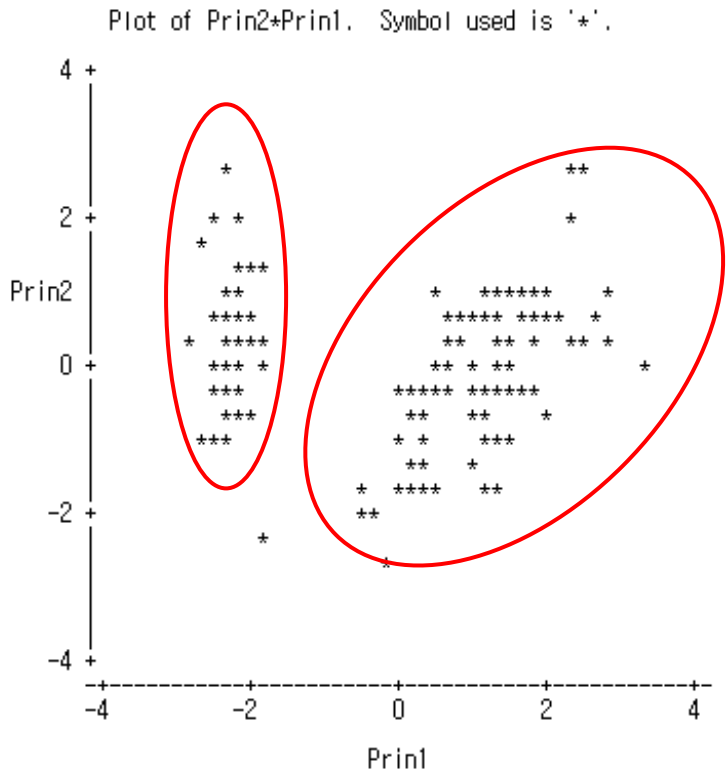
주성분이 2 개면 충분

Eigenvectors

	Prin1	Prin2	Prin3	Prin4
SL	0.521066	0.377418	-.719566	-.261286
SW	-.269347	0.923296	0.244382	0.123510
PL	0.580413	0.024492	0.142126	0.801449
PW	0.564857	0.066942	0.634273	-.523597

주성분 1 은 (SL, PL, PW: 꽃잎의 크기), 주성분 2 는 SW(수술의 넓이)에 의해 결정된다. VAXIS 는 Y 축, HAXIS 는 X 축에 대한 설정이다. VPOS 는 가로 25%, HPOS 세로 50% 크기의 PLOT 을 그리라는 명령이다.

```
PROC PLOT;
    PLOT PRIN2*PRIN1='*'/VAXIS=-4 TO 4 BY 2 HAXIS=-4 TO 4 BY 2
    VPOS=25 HPOS=50;
RUN;
```



위의 주성분 산점도를 살펴보면 PRIN1 에 의해 군집이 잘 나누어 진다. 즉 꽃잎의 크기(넓이, 길이)가 개체를 분류하는데 결정적 역할을 하고 있음을 알 수있다.

7.6.4. Fast clustering 군집 분석

비계층적 군집 방법을 사용하되 최대 군집의 개수는 3 개로 하여 개체를 군집하여 보자. 주성분 분석 결과 2 개 집단일 가능성이 높지만 우리는 이미 집단이 3 개인지 알고 있고 예상되는 집단보다는 1 나 정도 더 잡아 주는 것이 좋다. 왜냐하면 최대 군집 개수이므로.....

주성분 분석 결과가 저장된 SAS data 자료(IRIS_PRIN)를 이용하여 군집 분석하면 이미 PRIN1, PRIN2 등 주성분 변수가 있으니 군집 분석 결과에 대한 산점도 그리는데 용이하다.

```
PROC FASTCLUS DATA=IRIS_PRIN OUT=CLUS MAXCLUSTERS=3;
    VAR SL SW PL PW;
RUN;
```

①

① Cluster Summary

Cluster	Frequency	RMS Std Deviation	Maximum Distance from Seed to Observation	Radius Exceeded	Nearest Cluster	Distance Between Cluster Centroids
1	75	0.5644	2.6275		2	2.3679
2	36	0.6514	1.9424		3	2.3387
3	39	0.4032	1.8118		2	2.3387

Statistics for Variables

Variable	Total STD	Within STD	R-Square	RSQ/(1-RSQ)
SL	1.00000	0.57548	0.673271	2.060640
SW	1.00000	0.69102	0.528904	1.122711
PL	1.00000	0.43202	0.815862	4.430711
PW	1.00000	0.46730	0.784559	3.641633
OVER-ALL	1.00000	0.55084	0.700649	2.340560

①RMS Std Deviation

군집 내의 개체들간의 거리 평균 제곱근으로 작을수록 개체 군집화가 잘 된 것이다.

②Maximum Distance from Seed to Observation

군집 내 개체 중 seed 와 거리가 가장 먼 개체와 seed 간 거리

③R-Square

군집에 의해 변수를 예측할 때 결정 계수

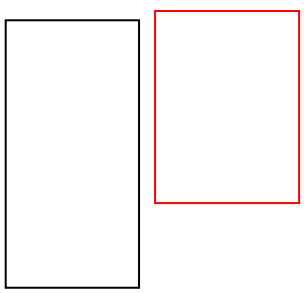
④RSQ/(1-RSQ)

군집에 의해 변수를 얼마나 잘 예측할 수 있나?

⑤Distance Between Cluster Centroids

가장 가까이 있는 군집과 현재 군집의 중심(평균)간의 거리

PL 이 개체들을 분류(군집)하는데 가장 큰 역할을 하고 있음을 알 수 있다. 그 다음이 PW 이다. 이는 주성분 변수 1 이 개체를 분류하는 역할을 하고 있음을 주성분 변수 산점도에서 알았다.



Cluster Means

Cluster	SL	SW	PL	PW
1	0.815554834	-0.217191992	0.834986391	0.831761570
2	-0.726592535	-0.800704280	-0.339697087	-0.396347653
3	-0.897673879	1.156788551	-1.292176517	-1.233682108

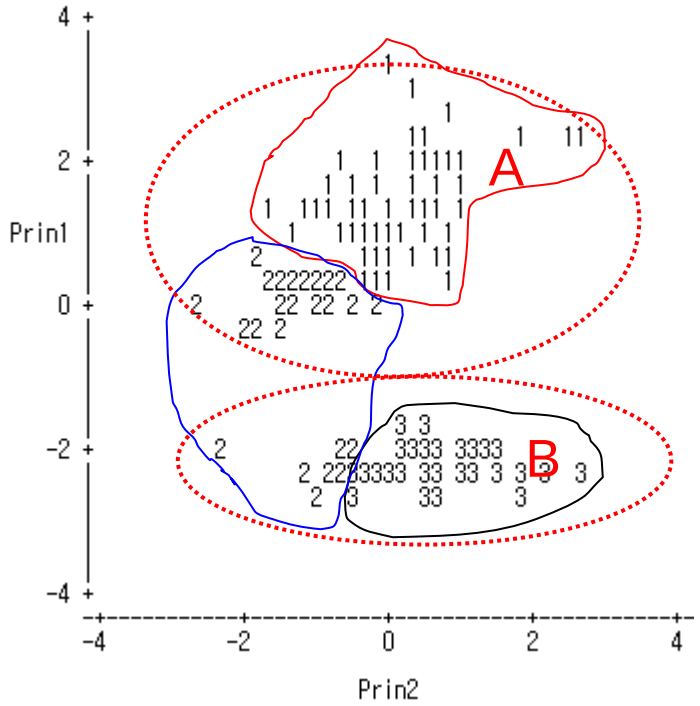
Cluster Standard Deviations

Cluster	SL	SW	PL	PW
1	0.6508432352	0.6983117908	0.3717155663	0.4741678502
2	0.5754614371	0.6836255246	0.6931777480	0.6468942077
3	0.3888582277	0.6834685284	0.1018841969	0.1465788652

각 군집에서 분류 변수들의 평균과 표준 편차 값이 출력되므로 이것에 의해 정보를 이용하여 어떤 변수들에 의해 나누어 지고 있는지 알 수 있으므로 군집의 이름을 붙이는데 유용하다.

```
PROC PLOT DATA=CLUS;
    PLOT PRIN1*PRIN2=CLUSTER/VPOS=25 HPOS=50;
RUN;
```

Plot of Prin1*Prin2. Symbol is value of CLUSTER.



만약 최대 군집 개수를 2 개로 하였다면 빨간 점선으로 2 개의 개체 분류가 될 것이다.(A 는 꽃잎의 크기가 큰 군집, B 는 꽃잎의 크기가 작은 군집) 이 경우 주성분 변수 1, 2 의 산점도에 의해 개체를 분류하는 것과 동일하다. 그러나 주성분 변수에 의해 개체를 분류하면 개체의 ID 를 모르므로 군집 분석 방법을 사용하여 개체를 분류하고 개체의 이름을 부일 때 주성분 산점도를 사용하는 것이 좋다.

7.6.5. 다른 옵션 사용 1

최대 군집 개수 3 개, SEED 는 매번 옮겨 가게(DRIFT), 최대 반복 수는 3 번

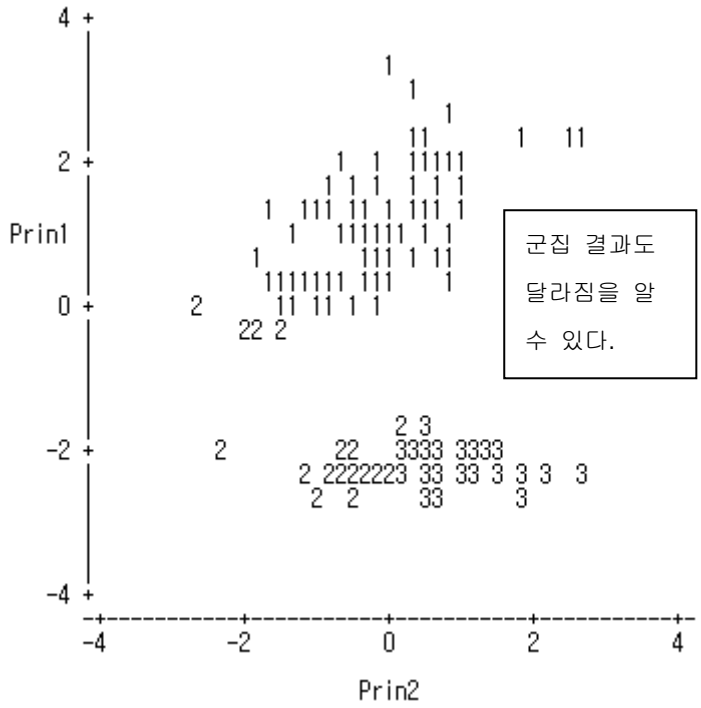
```
PROC FASTCLUS DATA=IRIS_PRIN OUT=CLUS0 RANDOM=1 DRIFT MAXCLUSTERS=3;
    VAR SL SW PL PW;
RUN;
```

Cluster Summary

Cluster	Frequency	RMS Std Deviation	Maximum Distance from Seed to Observation	Radius Exceeded	Nearest Cluster	Distance Between Cluster Centroids
1	96	0.6267	2.9910		2	3.1801
2	26	0.5325	2.8213		3	1.6929
3	28	0.3570	1.6787		2	1.6929

페이지 168 의 결과에 비해 RMS 편차가 크므로 더 좋은 방법은 아니다.

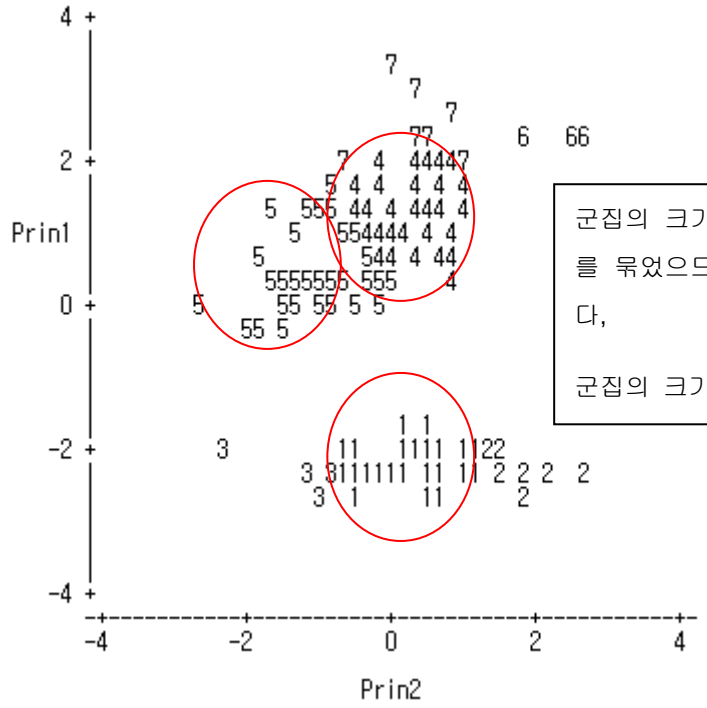
```
PROC PLOT DATA=CLUS0;
    PLOT PRIN1*PRIN2=CLUSTER/VPOS=25 HPOS=50;
RUN;
```



7.6.6. 다른 옵션 사용 2

군집의 반지름을 정해주는 방법으로 군집의 개수가 반지름에 의해 결정된다. 다음은 군집의 반지름을 2 으로 설정하여 한 결과이다.

```
PROC FASTCLUS DATA=IRIS_PRIN OUT=CLUS1 RADIUS=2;
    VAR SL SW PL PW;
RUN;
PROC PLOT DATA=CLUS1;
    PLOT PRIN1*PRIN2=CLUSTER/VPOS=25 HPOS=50;
RUN;
```



군집의 크기가 반지름 2 이하 되게 개체를 묶었으므로 군집의 수가 7 개가 되었다,

군집의 크기가 매우 작음을 알 수 있다.

Cluster	Frequency	RMS Std Deviation	Maximum Distance from Seed to Observation
1	49	0.4571	2.3375
2	70	0.5392	2.6469
3	31	0.4724	2.2554

반지름을 3(RADIUS=3) 설정하였더니 군집의 개수는 3 개이고 RMS 편차도 가장 적음을 알 수 있다. 어떤 옵션을 사용해야 하나? Intuition & Trial Error

7.7. 다차원 척도법

다차원 척도법(MDS: Multi Dimensional Scaling)이란 n 개의 개체를 저 차원 가시적 공간(일반적으로 2 차원)에 나타낼 수 있도록 하는 방법이다. 각 개체간 유사성(similarity) 혹은 거리는 저 차원으로 옮겨지더라도 원래 유사성 크기를 갖고 있어야 한다. 예를 들어 보자. Under-arm Deodorant 생산 회사에서 판매 전략을 세우기 위하여 각 제품들이 서로 얼마나 유사한지(가까운지) 알아보려고 한다. 이를 위하여 소비자를 임의로 선택하여 제품의 각 분야에 대한 평가를 하게 하였다. 즉 향기, 냄새 제거 정도, 사용 편리 정도, 옷에 묻어나는 정도 등을 10 점 만점으로 평가하였다. 이 점수들을 이용하여 제품의 유사성 정도를 2 차원 공간에 표현하는 방법이다.

개체간 유사성을 측정하는 방법은 metric 방법과 non-metric 방법이 있다.

(1)Euclidean distance ► 측정형 변수 거리 (Metric 방법)

(2)각 개체의 유사성(거리)을 사람들이 평가하도록 한다. (Metric/non-Metric 방법)

(3)평가자 들이 개체를 마음대로 분류하게 하고 빈도로부터 유사성을 측정한다. (non-Metric 방법)

응용 범위를 살펴보면 다음과 같다.

(1)회사들의 이미지 측정을 통한 고객 분류

(2)소비자들이 인지하고 있는 유사한 상품 속성이나 상품 분류에 사용

(3)인구 학적 특성, 경제적 특성을 기초하여 도시간 동질성 파악

7.7.1. 개체간의 거리 측정

다차원 척도법이란 n 개의 개체를 저 차원 가시적 공간(일반적으로 2 차원)에 나타낼 수 있도록 하는 방법이므로 각 개체간 거리(유사성)를 측정해야 한다. 군집 분석과 유사해 보이지만 다차원 척도법은 개체의 유사성을 이차원에 표시하는 것이고 군집 분석은 개체간의 거리(유사성)가 가까운 것끼리 묶어 가는 방법이다.

(1)metric 방법

metric 방법은 두 개체간의 거리(유사성)를 Euclidean distance 로 나타낸다.

측정변수 / 개체	X_1	X_2	⇢	X_p
1	x_{11}	x_{12}	⇢	x_{1p}
2	x_{21}	x_{22}	⇢	x_{2p}
⇣	⇣	⇣	⇢	⇣
n	x_{n1}	x_{n2}	⇢	x_{np}

두 개체 (D_1 , D_2)간 거리는 $\sqrt{(x_{i1}-x_{j1})^2+(x_{i2}-x_{j2})^2+\cdots+(x_{ip}-x_{jp})^2}$ (Euclidean distance)

측정이 불가능한 경우는 각 개체간의 유사성(거리)을 리커드 척도(10 점, 100 점)를 이용하여 사람들이 평가하도록 하는 방법이 있다. 즉 각 x_1, x_2,…, x_p 가 사용자들의 주관적인 평가 점수(리커드 척도)가 된다. Deodorant 제품을 분류할 경우 향기, 냄새 제거 정도, 사용 편리 정도, 옷에 묻어나는 정도 등을 10 점 만점으로 평가하였다면 이 점수가 개체들간의 거리를 측정하는데 사용된다.

(2)non-metric 방법

개체간의 거리를 사람들이 임의로 분류한 결과로부터 만들어진 빈도로 측정하는 방법이다. 이 방법을 이해하려면 예를 들어보는 것이 더 편리할 것이다. 50 명이 20 개의 개체를 임의로 분류하여 (1, 2)를 하나의 군집으로 분류한 사람이 30, (1, 20)을 하나의 군집으로 분류한 사람이 25 명, (2, 20)을 하나의 군집으로 분류한 사람이 45 명이라면

측정변수 / 개체	1	2	⇢	20
1	0			
2	30	0		
⇣	⇣	⇣	⇢	
20	25	45	⇢	0

이 경우 숫자가 클수록 거리는 가깝다. 즉 개체간 유사성이 높다. MDS 분석을 위하여 자료를 입력할 때는 (1-빈도/평가자수) 이것이 유사성이 된다.

7.7.2. 기본 알고리즘

각 개체간 유사성을 측정한다. 개체의 개수가 n 개인 경우 k=n(n-1)/n 개 유사성 그룹이 존재한다.

(1)유사성이 작은 것부터 크기 순으로 배열한다. $S_{i1j1} < S_{i2j2} < \ldots < S_{ikjk}$

(2)개체를 m(일반적으로 2)차원으로 공간으로 줄일 경우 개체간의 거리를 구한다. 이는 물론 측정 변수 전체를 가지고 유사성을 측정한 2)와 다를 것이다. 2 차원 공간으로 줄일 수 있는지를 알아 보는 것이 STRESS 값이다.

$$STRESS = \sqrt{\frac{\sum\sum_{i<j}(S_{ij}^{2)} - S_{ij}^{3)})^2}{\sum_{i<j}(S_{ij}^{2)})^2}}$$

2 차원 공간으로 줄이는 것이 목적이므로 Stress 검정은 하지 않는다.

Stress	Goodness of fits
20%	Poor
10%	Fair
5%	Good
2.5%	Excellent

7.7.3. 예제 1 (Metric)

변수가 하나되고 측정형인 경우 다차원 척도법을 적용해보자.

```
data city;
input  (atlanta chicago denver houston losangel
       miami newyork sanfran seattle washdc) (5.)
       @56 city $15.;
cards;
   0                                                Atlanta
 587    0                                          Chicago
1212  920    0                                     Denver
 701  940  879    0                                Houston
1936 1745  831 1374    0                           Los Angeles
 604 1188 1726  968 2339    0                      Miami
 748  713 1631 1420 2451 1092    0                 New York
2139 1858  949 1645  347 2594 2571    0            San Francisco
2182 1737 1021 1891  959 2734 2408  678    0       Seattle
 543  597 1494 1220 2300  923  205 2442 2329    0  Washington D.C.
run;
```

(5.)은 5 칸씩 읽어 들이고 @56 은 56 번째 줄부터 city 변수 값을 읽으라는 의미이다.

위 프로그램은 미국 10 개 도시간 거리를 조사한 후 MDS 방법을 이용하여 도시를 2 차원 공간에 표현하기 위하여 입력한 자료 형태이다. 것이다. 만약 non-metric 의 경우에는 거리 자리에 (1-빈도/평가자수)를 넣어주면 된다.

(1)MDS 프로그램

①LEVEL=absolute 옵션은 측정 변수가 측정형인 경우 사용된다. 만약 리커드 척도면 LEVEL=ordinal 을 사용하면 된다. default 는 ordinal 이다.

②SHAPE=square 를 쓰면 각 전체 값이 다 나타난 경우이다. 위와 같이 아래 부분에만 자료를 입력한 경우는 SHAPE=triangle 이다. triangle 이 default 이다.

③PLOTIT 은 결과를 그린 것이다. 도시간 유사성(거리가) 2 차원 공간에 나타난다.

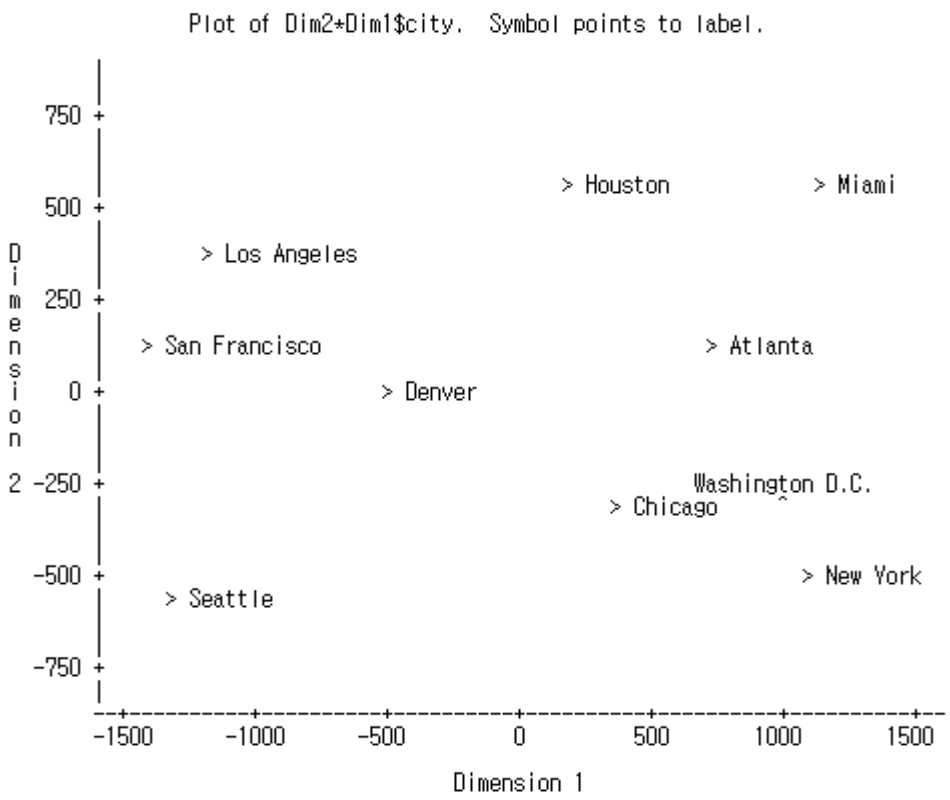
④MDS 의 결과는 개체 간의 유사성을 나타낸 그래프(산점도)에 대한 해석이면 충분하다.

```
proc mds data=city fit=2 level=absolute out=out;
id city;
run;

proc plot data=out;
plot dim2 * dim1 $ city;
run;

%plotit(data=out,datatype=mds,labelvar=city,color=black,vtoh=1.5);
quit;
```

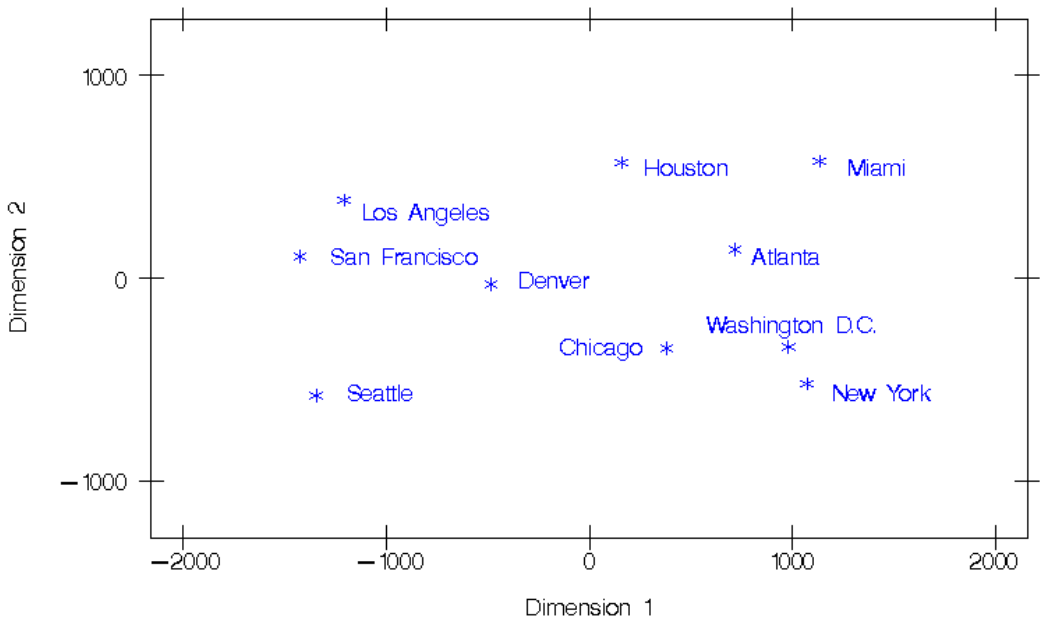
2 차원 공간으로 줄이는 경우에는 STRESS 검정을 하지 않으므로 산점도 이외 특별히 주의해서 관찰해야 할 출력 결과는 없다. 다음은 SAS data OUT 에 저장되어 변수들이다.

Obs	_DIMENS_	_MATRIX_	_TYPE_	city	_NAME_	Dim1	Dim2
1	2	.	CRITERION			0.00	.
2	2	.	CONFIG	Atlanta	atlanta	718.65	143.759
3	2	.	CONFIG	Chicago	chicago	382.01	-341.229
4	2	.	CONFIG	Denver	denver	-482.46	-25.691
5	2	.	CONFIG	Houston	houston	161.11	571.996
6	2	.	CONFIG	Los Angeles	losangel	-1202.80	387.114
7	2	.	CONFIG	Miami	miami	1132.39	580.375
8	2	.	CONFIG	New York	newyork	1071.81	-517.473
9	2	.	CONFIG	San Francisco	sanfran	-1420.15	111.397
10	2	.	CONFIG	Seattle	seattle	-1339.92	-575.694
11	2	.	CONFIG	Washington D.C.	washdc	979.36	-334.554

(2)MDS 산점도 해석



실제 도시간 거리와 산점도의 거리는 일치하지 않는다. 이는 모든 개체들의 거리(유사성)를 고려하여 개체의 좌표(Dim1, Dim2)가 결정되었기 때문이다.



7.7.4. 예제 2

개체간의 유사성이 측정치가 아니라 사람들에게 설문하여 측정된 경우 다음과 같이 분석하면 된다. 다음 자료는 6 살 아이 15 명을 대상으로 몸 부위 15 개에 대해 각각 관련이 깊은 순서대로 나열하게 하여 다음 자료를 얻었다. 행 15 개씩 한 아이의 평가 결과이다. 첫 열은 cheek(턱)과 가장 관련이 깊은 순서대로 순위를 매긴 것이다. mouth=>face=>head .. 순이다.

```
data body;
   title 'First 15 subjects are 6 year-old children';
   input  cheek face mouth head ear body arm elbow hand
          palm finger leg knee foot toe;
   datalines;
 0  2  1  3  4 10  5  9  6  7  8 11 12 13 14
 2  0 12  1 13  3  8 10 11  9  7  4  5  6 14
 3  2  0  1  4  9  5 11  6  7  8 10 13 12 14
 2  1  3  0  4  9  5  6 11  7  8 10 12 13 14
10  1 11  2  0  6  3  4  5 12 13  7  8 14  9
14 12  9  6 13  0  8  7  5 10 11  1  4  2  3
12 14 11 10 13  5  0  4  1  3  2  6  9  7  8
 5  7 14  8  6  9  1  0  2  3  4 10 11 12 13
13 11 12 10 14  9  3  4  0  1  2  6  5  7  8
 8  6  7  9  4  5  3 10  1  0  2 12 11 13 14
14  5 13  6  9 12  3  4  1  2  0  7  8 10 11
14 12 13 11  9  7  4  6  5  3 10  0  8  1  2
12 11 14 10 13  4  5  8  6  7  9  1  0  2  3
12 14 10 13 11  9  4  5  8  6  7  2  3  0  1
13  8  9 11 14  3  6  5  7 10 12  2  4  1  0
 0  4  2  3 11  9 14 12  1  7 13  8  6  5 10
```

```
proc mds data=body condition=row level=ordinal
    dimension=3 out=out pfinal;
run;
```

(1)CONDITION 옵션은 data 의 입력 상태 설정으로 ROW 는 행으로 인식, MATRIX(default)는 행렬 상태로 인식하라는 것이다. 일반적으로 행렬로 입력된다. (대칭 행렬)

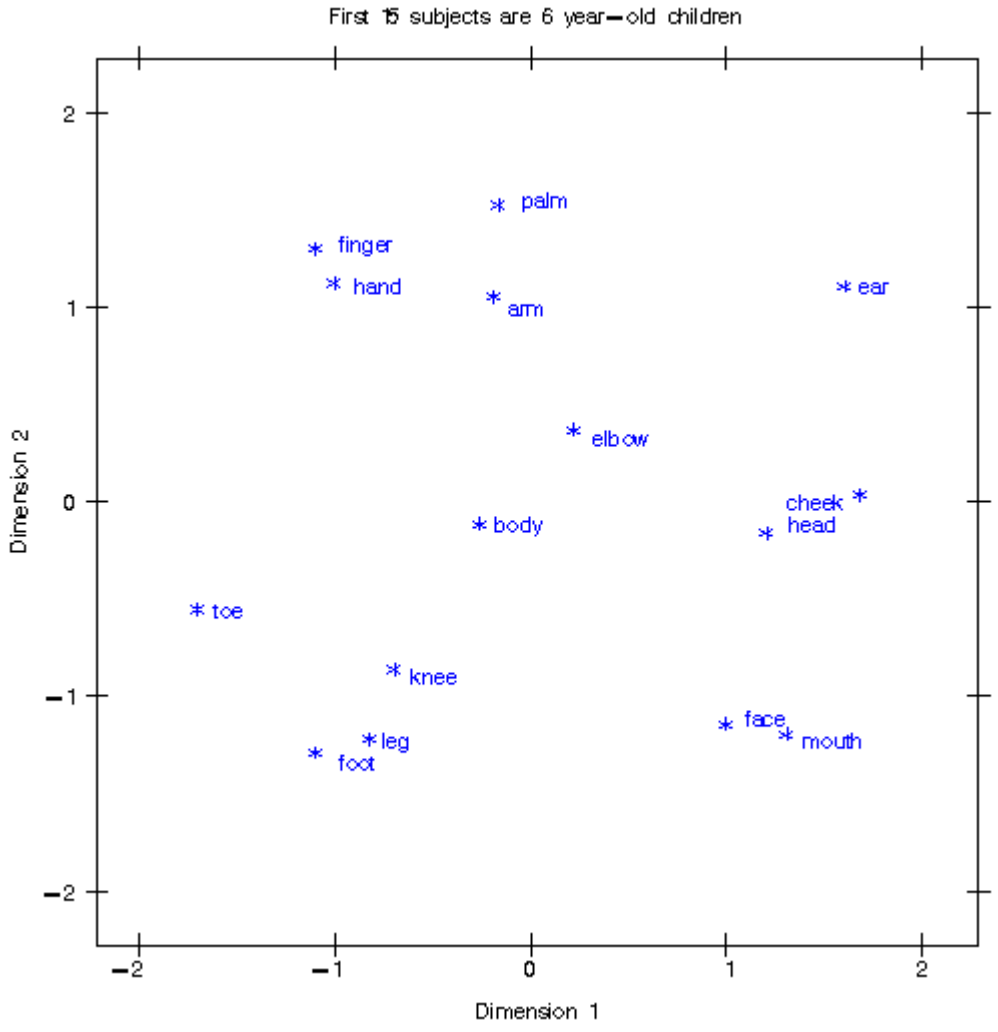
(2)LEVEL 옵션은 data 형태를 지정하는데 ORDINAL(default)은 순서형, ABSOLUTE 는 측정형이다.

(3)DIMENSION 은 차수를 지정한다.

(4)PFINAL 은 최종 예측치가 출력되는데 이는 OUT 옵션에 의해 저장된다.

```
proc print data=out;
run;
```

TYPE	_LABEL_	_NAME_	Dim1	Dim2	Dim3
CRITERION			0.19889	.	.
CONFIG	cheek	cheek	1.68256	0.03609	-1.07337
CONFIG	face	face	0.99937	-1.14243	0.79431
CONFIG	mouth	mouth	1.30754	-1.19176	-0.53881
CONFIG	head	head	1.20611	-0.15902	0.96118
CONFIG	ear	ear	1.60413	1.10846	0.46651
CONFIG	body	body	-0.25922	-0.11156	1.87045
CONFIG	arm	arm	-0.18827	1.05366	-1.46103
CONFIG	elbow	elbow	0.22302	0.37175	-0.56833
CONFIG	hand	hand	-1.00126	1.12422	0.69991
CONFIG	palm	palm	-0.16411	1.52583	0.45577
CONFIG	finger	finger	-1.09911	1.30240	-0.41552
CONFIG	leg	leg	-0.82345	-1.21679	-0.41693
CONFIG	knee	knee	-0.69464	-0.86098	-1.56571
CONFIG	foot	foot	-1.09436	-1.28767	0.62574
CONFIG	toe	toe	-1.69830	-0.55221	0.16583



```
%plotit(data=out, datatype=mds,
        plotvars=dim2 dim1,labelvar=_name_, vtoh=1.5,color=black);
%plotit(data=out, datatype=mds,
        plotvars=dim3 dim1, labelvar=_name_, vtoh=1.5,color=black);
%plotit(data=out, datatype=mds,
        plotvars=dim3 dim2, labelvar=_name_, vtoh=1.5, color=black);
run;
```

7.7.5. 예제 3

미국에서 생산되는 34 개 차들의 HP(마력), TIM1(시속 60 마일 걸리는 시간:초) TIM2(1/4 마일 걸리는 시간:초), TS(최대속력) BRAK1(60 마일 속력에서 제동 거리) BRAK2 (80 마일 속력에서 제동 거리) SP(손잡이 흔들림 없는 최대 속력) SS(회전 코스 최대 속력) MPG(가스 마일리지)를 조사하였다. MDS 분석하시오. [CARS.txt]

Obs	MAKE	ET	HP	TIM1	TIM2	TS	BRAK1	BRAK2	SP	SS	MPG
1	Acura NSX	V-6	270	5.8	14.0	168	120	200	0.87	62.3	18.0
2	Alfa Rom	V-6	230	7.6	15.8	150	150	268	0.77	59.3	17.5
3	Audi S4	I-5t	227	6.5	15.0	130	147	246	0.83	60.0	16.0
4	Bentley T	V-8t	315	7.1	15.5	128	147	253	0.73	56.8	11.5

```
DATA CARS;
    INFILE "D:\TEMP\CARS.TXT";
    INPUT MAKE $ 1-11 ET $ HP TIM1 TIM2 TS BRAK1 BRAK2 SP SS MPG;
RUN;

DATA CARS4;
    SET CARS;
KEEP MAKE;

PROC STANDARD DATA=CARS MEAN=0 STD=1 OUT=CARS2;
    VAR HP--MPG;
RUN;
```

CARS4 에는 자동차 회사 변수만 있고 CARS2 는 측정 단위가 다르므로 변수를 표준화 한 결과가 들어 있다.

```
DATA ONE; SET CARS2;
 I = _N_;
 KEEP HP--MPG I;
RUN;

DATA ORIG; SET ONE;
 DO J=1 TO 34;
 OUTPUT; END;
RUN;
```

```
DATA DUP; SET ORIG;
    II=J;  JJ=I;  I=II;  J=JJ;  DROP JJ II;
    Y1=HP; Y2=TIM1; Y3=TIM2; Y4=TS; Y5=BRAK1;
    Y6=BRAK2; Y7=SP; Y8=SS; Y9=MPG;
NN = _N_;
DROP HP--MPG;
RUN;
```

```
PROC SORT DATA=DUP;
    BY I J;
RUN;
```

ONE 에는 I 변수가 있고 이것이 자료의 관측치 번호이다. ORIG 는 각 관측치 번호에 J 라는 변수를 만들어 (i, j)=(1,2), (1,3), ……(34,34) 총 34*34 행이 존재한다.

DUP 데이터는 i 와 j 값을 바꾸고 행을 나타내는 NN 변수 생성한다. 측정 변수와 동일하게 새로운 변수들을 만든다.

```
DATA COMB; MERGE ORIG DUP; BY I J;
    DROP I J;
    H = SQRT((HP-Y1)**2 + (TIM1-Y2)**2 +(TIM2-Y3)**2
        +(TS-Y4)**2 +(BRAK1-Y5)**2 +(BRAK2-Y6)**2
        +(SP-Y7)**2 +(SS-Y8)**2 +(MPG-Y9)**2);
    KEEP H;
RUN;
```

각 개체들간 거리를 계산한다. 그리고 거리만 남겨두고 모든 변수는 제외한다.

```
PROC IML; RESET NOLOG;
    USE  COMB;
    READ ALL INTO DIST;
    USE CARS4;
    READ ALL VAR _CHAR_ INTO CARS;
    PRINT CARS;
    N=SQRT(NROW(DIST));
    DD=SHAPE(DIST,N,N);
    *PRINT 'DISTANCE MATRIX = ',DD;
    CREATE DIST_DAT FROM DD[COLNAME=CARS];
    APPEND FROM DD;
QUIT;
```

```
DATA FINAL;
MERGE CARS4 DIST_DAT;
RUN;
```

```
PROC PRINT DATA=FINAL;
TITLE 'DISTANCES BETWEEN CAR MAKES';
RUN;
```

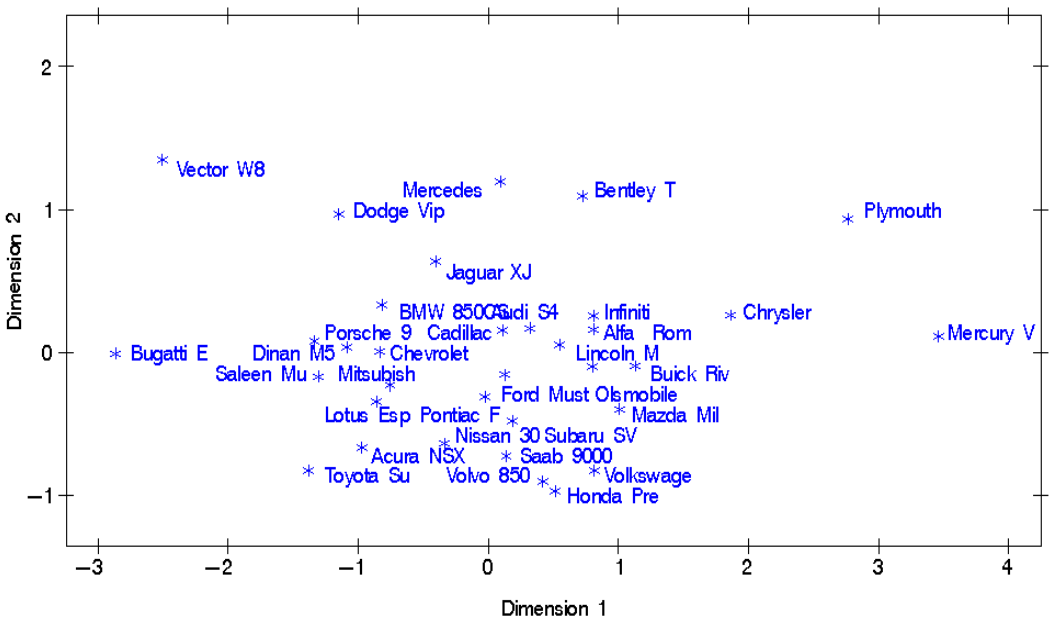
```
DROP HP--MPG;
RUN;
```

새로운 행렬 이름=SHAPE(기존 행렬 이름, nrow, ncol) ➔ 기존 행렬로부터 (nrow, ncol) 차수를 갖는 새로운 행렬을 만든다. (열 벡터 DIST 를 $n \times n$ 행렬 DD 를 만든다.)

Obs	MAKE	Acura_ NSX	Alfa__ Rom	Audi_S4	Bentley_ T	BMW_ 850CS	Bugatti_ E
1	Acura NSX	0.00000	4.23485	3.40093	4.86142	2.38311	4.2833
2	Alfa Rom	4.23485	0.00000	1.77467	2.59301	3.72810	7.1009
3	Audi S4	3.40093	1.77467	0.00000	2.42033	2.65289	6.4896
4	Bentley T	4.86142	2.59301	2.42033	0.00000	3.53591	7.3232
5	BMW 850CS	2.38311	3.72810	2.65289	3.53591	0.00000	4.2496
6	Bugatti E	4.28335	7.10095	6.48957	7.32320	4.24963	0.0000

```
PROC MDS DATA=FINAL OUT=OUT OUTRES=RES SHAPE=SQUARE;
    ID MAKE;
    TITLE 'MDS Analyses and Plots';
RUN;
```

```
goptions reset all;
%PLOTIT(DATA=OUT, DATATYPE=MDS, LABELVAR=MAKE,COLOR=BLACK);
RUN;
```


Vector W8
Mercedes
Bentley T
Dodge Vip
Plymouth
Jaguar XJ
BMW 850CSi
Audi S4
Infiniti
Chrysler
Porsche 9
Cadillac
Alfa Rom
Mercury V
Bugatti E
Dinan M5
Chevrolet
Lincoln M
Saleen Mu
Mitsubish
Buick Riv
Ford Must
Oldsmobile
Lotus Esp
Pontiac F
Mazda Mil
Nissan 30
Subaru SV
Acura NSX
Saab 9000
Toyota Su
Volvo 850
Volkswage
Honda Pre
Dimension 2
Dimension 1
2
1
0
−1
−3
−2
−1
0
1
2
3
4

[EXERCISE]

(1)ORANGE.txt 자료는 5 개국으로부터 생산된 오렌지들의 화학 성분을 조사한 것이다. Boron(B), Barium(BA), Calcium(CA), Potassium(PO), Magnesium(MA), Manganese(MN), Phosphorous(PH), Rubidium(RU), Zinc(ZN)

①군집 분석(개체 분류 방법 Centroid(학번 홀수), Average(학번 짝수) 방법 사용)을 이용하여 개체를 분류하시오.

②개체 군집 결과와 원산지가 일치하는지 알아보시오.

③개체가 분류된 군집에 적절한 이름을 붙이시오. (주성분 분석을 이용할 수 밖에 없지 않나?)

(2)ORANGE.txt 자료 FASTCLUSTER 방법 중 하나로 군집 분석하고 (1)에서 Hierarchical 방법과 비교 해석하시오.

(3)밀 자료(WHEAT.txt)에서 개체 집단 변수(지역, 밀 종류, 그룹)가 없다고 가정하고 문제 (1)과 (2)를 하시오.

(4)국내 승용차에 대해 다차원 분석을 실시해 보자.

①Metric 방법을 이용하시오. ► 측정 변수는 수집 가능한 변수만을 이용하시오.

②Non-metric 방법을 이용하시오. ► 사람들에게 설문 응답