

13. 군 집 분 석

13.1 개념 및 목적

군집분석(cluster analysis) :

모집단 또는 범주에 대한 사전정보가 **없는** 경우에 관측값들 사이의 거리(또는 유사성)를 이용하여 전체를 몇 개의 그룹 또는 군집 (cluster)으로 나누는 분석법

예 생물분류학에서는 생물을 그의 특성에 따라 분류
의상학에서는 인체의 체형에 따라 치수체계를 만듦

군집분석에서 어려운 점 :

‘군집간의 거리’에 대한 정의는?

군집으로 나누는 방법은?

군집으로 나누는 방법 :

- ① 계층적(hierarchical) 방법
- ② 비계층적(nonhierarchical) 방법

계층적 방법은 가까운 개체끼리 차례로 묶거나 멀리 떨어진 개체를 차례로 분리해 가는 군집방법으로서, 한 번 병합된 개체는 다시 분리되지 않는 것이 특징.

비계층적 방법은 산포를 나타내는 여러 가지 측도를 이용하여 이들 판정기준을 최적화시키는 군집방법으로서, 한 번 분리된 개체도 반복적으로 시행하는 과정에서 재분류될 수 있는 것이 특징.

SAS에서 제공되는 군집분석의 절차 :

- CLUSTER : 계층적 군집분석
(여러 종류의 군집법이 있음)
- FASTCLUS : 최적분리 비계층적 군집분석
- TREE : 덴드로그램(dendrogram)으로 알려진
나무구조를 출력

[예제 13.1]

도시간 비행거리를 나타내는 자료 분석

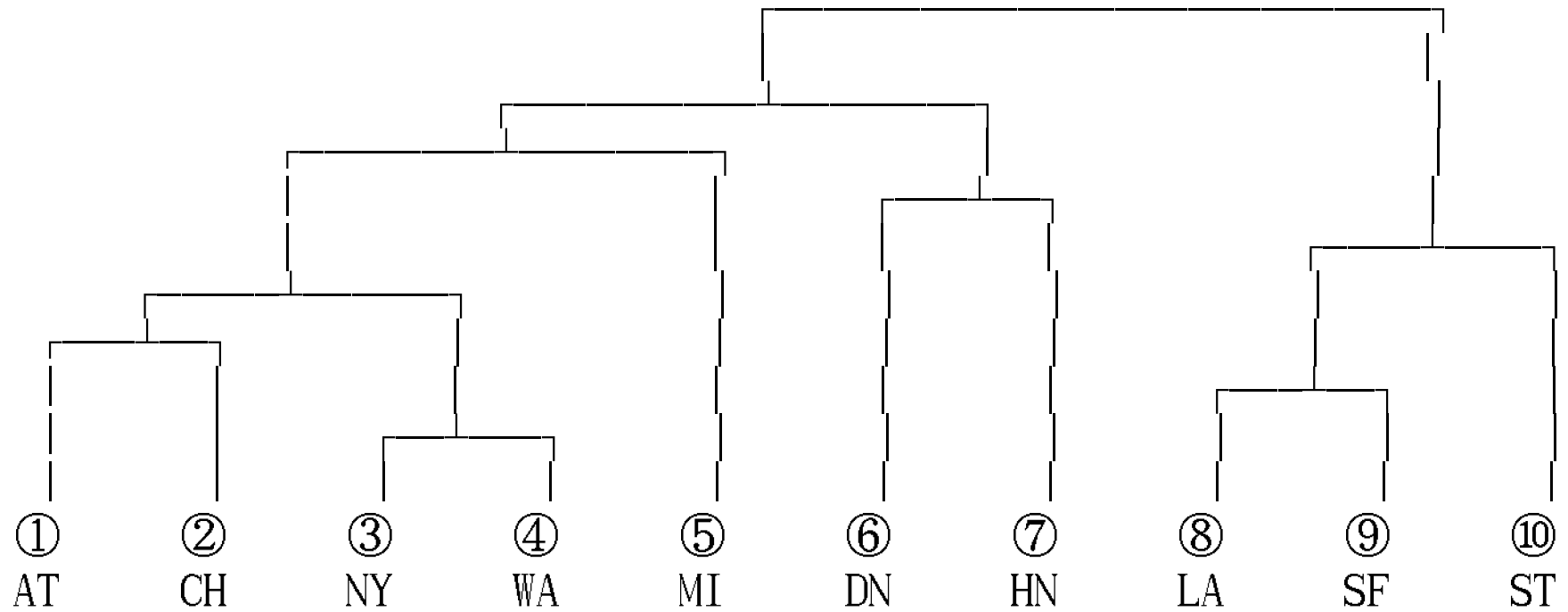
비행거리가 가까운 도시끼리 묶는다면 어떤 순서

CLUSTER 절차에서 평균연결법으로 병합되는 과정

(덴드로그램)

<표 13.1> 도시간 비행거리

①	②	③	④	⑤	⑥	⑦	⑧	⑨	⑩	
0										① ATLANTA
587	0									② CHICAGO
1212	920	0								③ DENVER
701	940	879	0							④ HOUSTON
1936	1745	831	1374	0						⑤ LOS ANGELES
604	1188	1726	968	2339	0					⑥ MIAMI
748	713	1631	1420	2451	1092	0				⑦ NEW YORK
2139	1858	949	1645	347	2594	2571	0			⑧ SAN FRANCISCO
2182	1737	1021	1891	959	2734	2408	678	0		⑨ SEATTLE
543	597	1494	1220	2300	923	205	2442	2329	0	⑩ WASHINGTON D. C.



<그림 13.1> 평균연결법의 덴드로그램

[예제 13.2] 예제 12.3에서 다룬 피셔의 붓꽃자료에 대하여 품종을
모른다는 가정 하에 군집분석을 시행

[예제 13.3] 예제 11.2에서 다룬 인체의 여러 부위에 대한 측정자료
앞에서는 인자분석을 이용하여 체형을 결정하는 요인을 찾고, 각
요인에서 주요 변수들의 기여도를 분석

이제 이들 관측값들에 군집분석을 적용시켜 치수체계를 구성

13.2 계층적 군집분석

13.2.1 이론적 배경

두 점 $\mathbf{x}_i' = (x_{i1}, \dots, x_{ip})$ 와 $\mathbf{x}_j' = (x_{j1}, \dots, x_{jp})$

사이의 거리의 예:

① 유클리드(Euclid)거리 :

$$d(\mathbf{x}_i, \mathbf{x}_j) = \sqrt{(x_{i1} - x_{j1})^2 + \dots + (x_{ip} - x_{jp})^2}$$

② 마할라노비스(Mahalanobis)거리 :

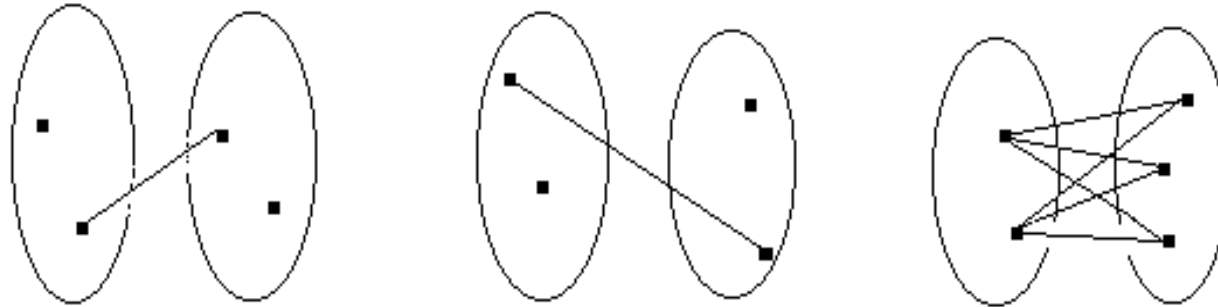
$$d(\mathbf{x}_i, \mathbf{x}_j) = \sqrt{(\mathbf{x}_i - \mathbf{x}_j)' \Sigma^{-1} (\mathbf{x}_i - \mathbf{x}_j)}$$

두 군집 사이의 거리를 나타내는 방법의 예 :

① 최단거리

② 최장거리

③ 평균거리



<그림 13.2> 군집 사이의 거리의 예

거리행렬 :

x_i, x_j 사이의 거리를 $d_{ij}=d(x_i, x_j)$ 라고 하면,

N 개의 관측값 x_1, \dots, x_N 에 대한 거리행렬

$$\begin{pmatrix} d_{11} & \cdots & d_{1N} \\ \vdots & & \vdots \\ d_{N1} & \cdots & d_{NN} \end{pmatrix}$$

(1) 단일연결법(Single Linkage Method)

두 군집 사이의 최단거리를 두 군집의 거리로 정의
즉, 두 군집 C_1, C_2 사이의 거리 :

$$d(C_1, C_2) = \min \{ d(x, y) : x \in C_1, y \in C_2 \}$$

[예제 13.4] 다음 거리행렬에 대하여 최단거리로 병합시키고 그 결과를 덴드로그램으로 나타내보자.

$$D = \begin{pmatrix} 0 & & & & \\ 1 & 0 & & & \\ 3 & 5 & 0 & & \\ 6 & 8 & 4 & 0 & \\ 8 & 7 & 4 & 2 & 0 \end{pmatrix} \begin{matrix} A \\ B \\ C \\ D \\ E \end{matrix}$$

1단계 A, B 가 가장 가까우므로 A 와 B 를 병합시키면 새로운 거리행렬 D_1 은 다음과 같다.

$$d(AB, C) = \min\{d(A, C), d(B, C)\} = \min\{3, 5\} = 3$$

$$d(AB, D) = \min\{d(A, D), d(B, D)\} = \min\{6, 8\} = 6$$

$$d(AB, E) = \min\{d(A, E), d(B, E)\} = \min\{7, 8\} = 7$$

따라서

$$D_1 = \begin{pmatrix} 0 & & & \\ 3 & 0 & & \\ 6 & 4 & 0 & \\ 7 & 4 & 2 & 0 \end{pmatrix} \begin{matrix} AB \\ C \\ D \\ E \end{matrix}$$

2단계 D, E 가 가장 가까우므로 D 와 E 를 병합시킨 거리행렬은 다음과 같다.

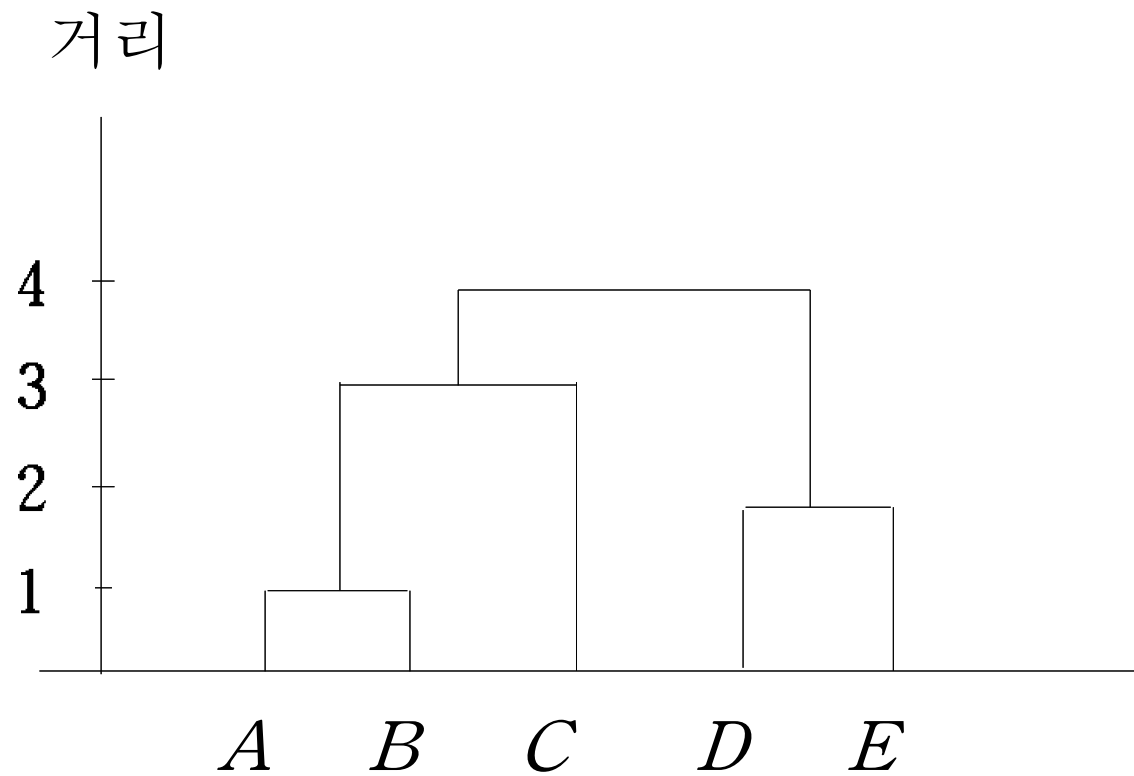
$$D_2 = \begin{pmatrix} 0 & & \\ 3 & 0 & \\ 6 & 4 & 0 \end{pmatrix} \begin{matrix} AB \\ C \\ DE \end{matrix}$$

3단계 AB 와 C 가 가장 가까우므로 AB 와 C 를 병합시킨 거리
행렬은 다음과 같다.

$$D_3 = \begin{pmatrix} 0 & \\ 4 & 0 \end{pmatrix} \begin{matrix} ABC \\ DE \end{matrix}$$

4단계 전체를 하나로 묶는다.

이들 병합과정을 나타내는 덴드로그램 :



(2) **완전연결법** (Complete Linkage Method)

두 군집 사이의 최장거리를 두 군집의 거리로 정의
즉, 두 군집 C_1, C_2 사이의 거리 :

$$d(C_1, C_2) = \max \{ d(x, y) : x \in C_1, y \in C_2 \}$$

완전연결법 : 가장 멀리 떨어진 개체 사이의 거리를
최소화

참고 단일연결법은 최단연결법으로서 고립된 군집을 찾는데 중점을 둔 군집법이며, 이에 대하여 완전연결법은 최장연결법으로서 내부적 응집성에 중점을 둔 군집법임.

[예제 13.5] 예제 13.3에서 사용한 거리행렬에 대하여 최장연결법의 병합과정은 다음과 같다.

1단계 A 와 B 가 가장 가까우므로 A 와 B 를 병합시키고, 새로운 거리행렬을 다음과 같이 구한다.

$$d(AB, C) = \max\{d(A, C), d(B, C)\} = \max\{3, 5\} = 5$$

$$d(AB, D) = 8, \quad d(AB, E) = 8$$

따라서

$$D_1 = \begin{pmatrix} 0 & & & \\ 5 & 0 & & \\ 8 & 4 & 0 & \\ 8 & 4 & 2 & 0 \end{pmatrix} \begin{matrix} AB \\ C \\ D \\ E \end{matrix}$$

2단계 D 와 E 를 병합시킨 결과는 다음과 같다.

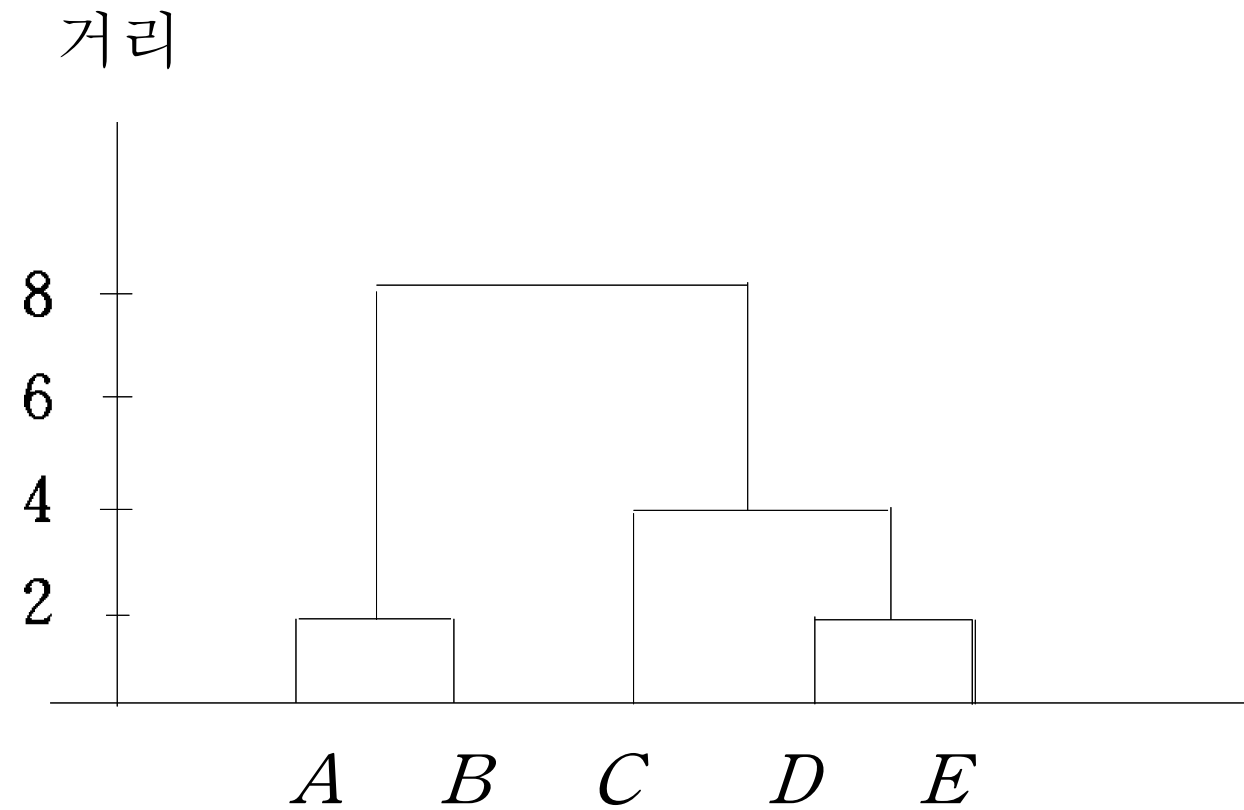
$$D_2 = \begin{pmatrix} 0 & & \\ 5 & 0 & \\ 8 & 4 & 0 \end{pmatrix} \begin{matrix} AB \\ C \\ DE \end{matrix}$$

3단계 C 와 DE 를 병합시킨 결과는 다음과 같다.

$$D_3 = \begin{pmatrix} 0 & & \\ 8 & 0 & \end{pmatrix} \begin{matrix} AB \\ CDE \end{matrix}$$

4단계 전체를 하나로 묶는다.

이들 병합과정의 덴드로그램은 다음과 같다.



(3) 평균연결법(Average Linkage Method)

두 군집에 있는 모든 개체들 사이의 거리의 평균을
두 군집 사이의 거리로 정의. 즉,

$$d(C_1, C_2) = \frac{1}{n_1 n_2} \sum_{x_i \in C_1} \sum_{x_j \in C_2} d(x_i, x_j)$$

(4) 중심연결법(Centroid Linkage Method)

군집평균 사이의 거리를 두 군집 사이의 거리로 정의
즉,

$$d(C_1, C_2) = \| \bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2 \|^2$$

(5) **메디안연결법** (Median Linkage Method)

중심연결법에서는 크기가 n_1, n_2 인 두 군집을 병합했을 때
새로운 군집의 중심은

$$\frac{(n_1 \bar{x}_1 + n_2 \bar{x}_2)}{(n_1 + n_2)}$$

단점 : 작은 군집은 지나치게 경시됨

메디안연결법에서 새로운 군집의 중심 :

$$\frac{(\bar{x}_1 + \bar{x}_2)}{2}$$

(6) Ward의 방법

각 병합의 단계에서 잔차제곱합(ESS)의 증분을 최소화

i 번째 군집내의 잔차제곱합 :

$$ESS_i = \sum_{j=1}^{n_i} \sum_{k=1}^p (x_{ijk} - \bar{x}_{ik})^2$$

(j : 개체번호, k : 변수번호)

$$\text{다만, } \bar{x}_{ik} = \frac{1}{n_i} \sum_{j=1}^{n_i} x_{ijk}$$

전체 g 개의 군집에 대한 잔차제곱합 : $ESS = \sum_{i=1}^g ESS_i$

두 군집 C_1 과 C_2 가 병합될 때에 ESS 의 증분 :

$$E(C_1, C_2) = \| \bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2 \| / \left(\frac{1}{N_1} + \frac{1}{N_2} \right)$$

따라서 Ward의 방법에서 두 군집사이의 거리 :

$$d(C_1, C_2) = E(C_1, C_2)$$

13.2.2 CLUSTER와 TREE 절차

CLUSTER와 TREE 절차의 일반형 :

① CLUSTER 절차

```
PROC CLUSTER METHOD=name <options>;  
  VAR variables;  
  ID variables;  
  BY variables;
```

② TREE 절차

```
PROC TREE <options>;  
  NAME variable;  
  PARENT variable;  
  ID variable;  
  HEIGHT variable;  
  BY variables;
```

각 절차에서 사용되는 명령문과 옵션들 :

① PROC CLUSTER <options>;

DATA=SAS dsn(data set name) : 자료명

SIMPLE(또는 S) : 변수별로 기술통계량을 출력

STANDARD(또는 STD) : 모든 변수를 표준화

METHOD=name : 군집연결법을 지정

SINGLE : 단일연결법 COMPLETE : 완전연결법

AVERAGE : 평균연결법 CENTROID : 중심연결법

MEDIAN : 메디안연결법 WARD : Ward연결법

PRINT=n : n개의 마지막 군집과정을 출력

CCC : Cubic Clustering Criterion으로서 군집의 개수를
결정하는 판정기준

특징

- ㉠ 군집당 표본수가 10 이하이면 CCC의 변동이 심함
- ㉡ $CCC > 2$ 이면서 국소최고점(local peak)이 있으면
그 점에 대응되는 군집수가 적당

② PROC TREE <options>;

DATA=SAS dsn;

LEVEL=n; 상호배반적 군집의 수를 지정

③ TREE 절차에서의 명령문

NAME variable : 각 관측값을 나타내는 노드 지정

PARENT variable : 나무구조의 노드를 지정

HEIGHT variable : 각 노드의 높이를 나타내는 변수
지정. 디폴트는 ‘군집의 개수’ 인 _NCL_

13.2.3 예제

[예제 13.1](계속) CLUSTER 절차를 이용하여 비행거리에 의한 도시들의 군집과정을 살펴보기 위해, 단일연결법과 평균연결법에 의한 군집분석


```

/* CLUSTER1.SAS : CLUSTER ANALYSIS OF FLYING MILEAGES BETWEEN CITIES */
OPTIONS PS=60;
DATA MILEAGES(TYPE=DISTANCE);    ①
    INPUT  (ATLANTA CHICAGO DENVER HOUSTON LOSANGEL
            MIAMI NEWYORK SANFRAN SEATTLE WASHDC) (5.)
            @55 CITY $15.;
    CARDS:
        0
    587      0
1212  920      0
    701  940  879      0
1936 1745  831 1374      0
    604 1188 1726  968 2339      0
    748  713 1631 1420 2451 1092      0
2139 1858  949 1645  347 2594 2571      0
2182 1737 1021 1891  959 2734 2408  678      0
    543  597 1494 1220 2300  923  205 2442 2329      0
;
RUN;

```

	ATLANTA
	CHICAGO
	DENVER
	HOUSTON
	LOS ANGELES
	MIAMI
	NEW YORK
	SAN FRANCISCO
	SEATTLE
	WASHINGTON D. C.

```
PROC CLUSTER DATA=MILEAGES METHOD=SINGLE;    ② ③
  ID CITY;
RUN;
PROC TREE;    ④
RUN;
PROC CLUSTER DATA=MILEAGES METHOD=AVERAGE;    ⑤
  ID CITY;
RUN;
PROC TREE;    ⑥
RUN; QUIT;
```

- ① TYPE=DISTANCE는 자료가 ‘거리’를 나타내는 자료임을 뜻하며, 따라서 자료는 거리행렬로 입력
TYPE=DISTANCE가 선언되지 않으면 자료를 좌표로 이해
- ② CLUSTER 절차에서는 METHOD 옵션을 반드시 사용

- ③ 단일연결법을 지정
- ④ 단일연결법에 의한 덴드로그램을 그린다.
- ⑤ 평균연결법을 지정
- ⑥ 평균연결법에 의한 덴드로그램을 그린다.

단일연결법과 평균연결법의 결과는 약간의 차이가 있으며, 군집과정이 거의 비슷함을 알 수 있다.

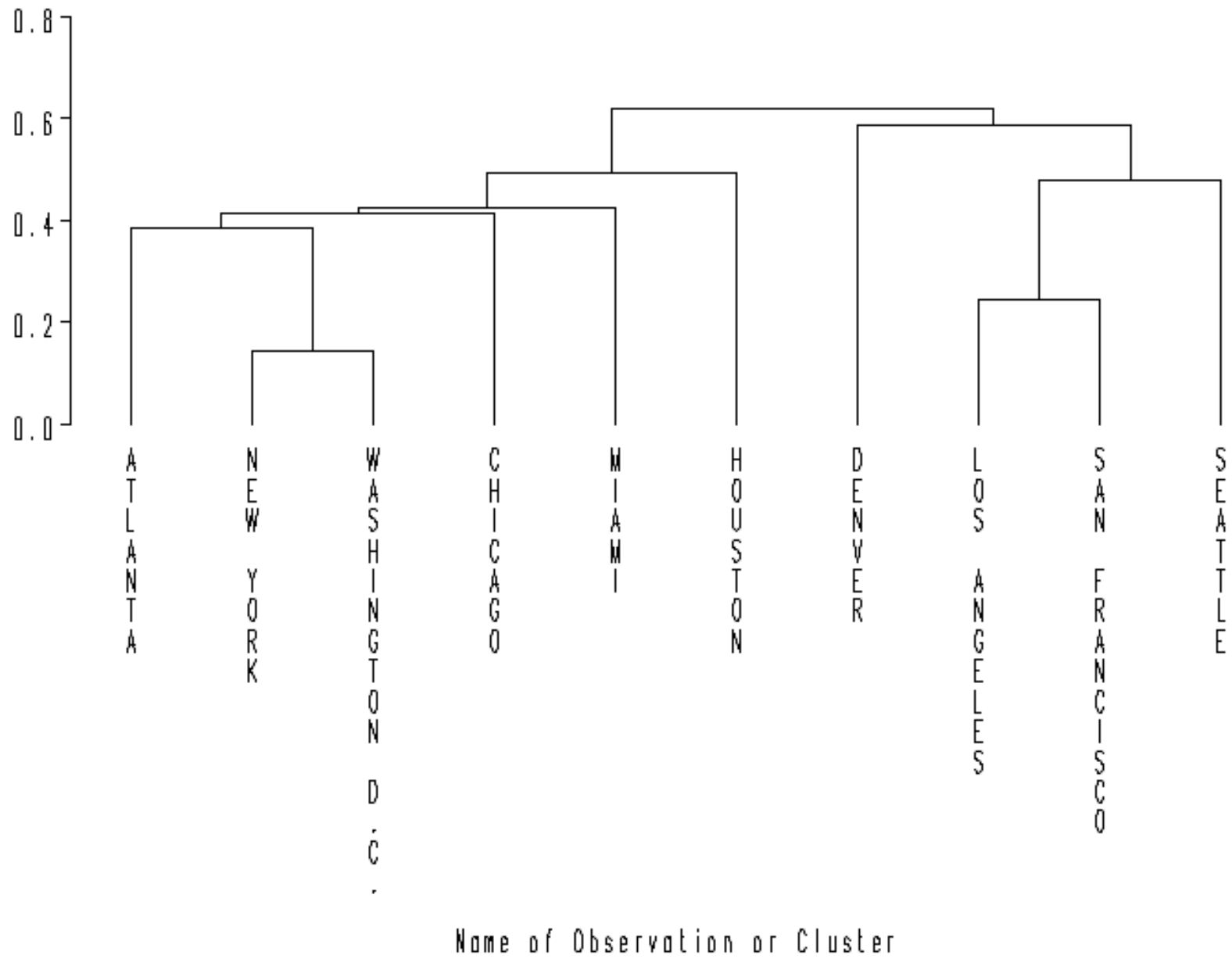
Single Linkage Cluster Analysis

Mean Distance Between Observations = 1417.133

Cluster History

NCL	-----Clusters Joined-----		FREQ	Norm	T
				Min	i
				Dist	e
9	NEW YORK	WASHINGTON D.C.	2	0.1447	
8	LOS ANGELES	SAN FRANCISCO	2	0.2449	
7	ATLANTA	CL9	3	0.3832	
6	CL7	CHICAGO	4	0.4142	
5	CL6	MIAMI	5	0.4262	
4	CL8	SEATTLE	3	0.4784	
3	CL5	HOUSTON	6	0.4947	
2	DENVER	CL4	4	0.5864	
1	CL3	CL2	10	0.6203	

<그림 13.3> 단일연결법 군집분석 결과



<그림 13.4> 단일연결법에 의한 덴드로그램

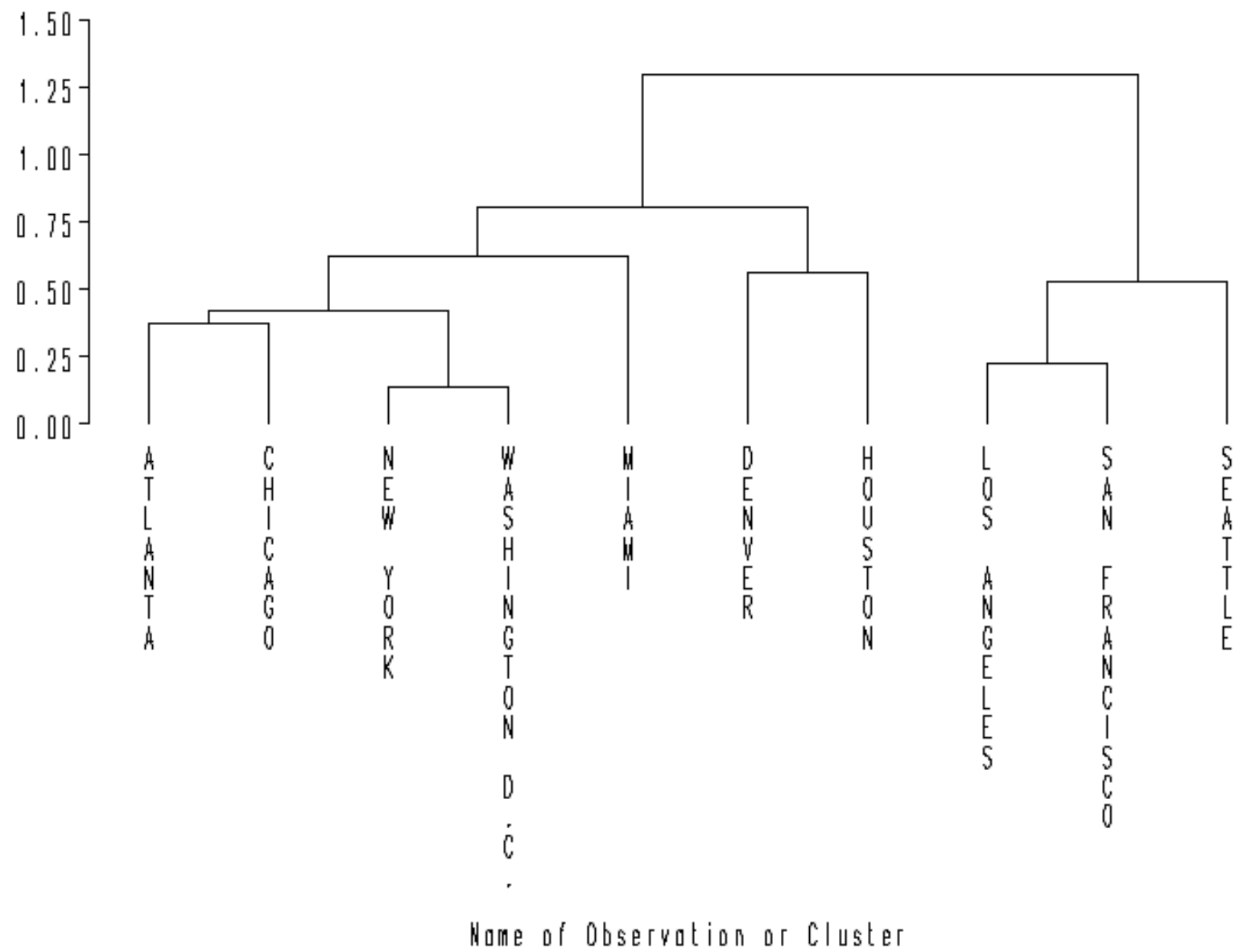
Average Linkage Cluster Analysis

Root-Mean-Square Distance Between Observations = 1580.242

Cluster History

				Norm	T
				RMS	i
NCL	-----Clusters Joined-----		FREQ	Dist	e
9	NEW YORK	WASHINGTON D.C.	2	0.1297	
8	LOS ANGELES	SAN FRANCISCO	2	0.2196	
7	ATLANTA	CHICAGO	2	0.3715	
6	CL7	CL9	4	0.4149	
5	CL8	SEATTLE	3	0.5255	
4	DENVER	HOUSTON	2	0.5562	
3	CL6	MIAMI	5	0.6185	
2	CL3	CL4	7	0.8005	
1	CL2	CL5	10	1.2967	

<그림 13.5> 평균연결법 군집분석의 결과



[예제 13.2](계속) 붓꽃자료의 군집분석

군집분석은 WARD 방법 사용

군집수를 결정하기 위해 CCC의 값을 구하고,

군집분석의 결과를 분할표로 나타내고,

군집의 내용을 정준 산점도로 나타낸다.

```
/* CLUSTER2.SAS : ANALYSIS OF FISHER'S IRIS DATA */
```

```
DATA IRIS;
```

```
  INFILE 'A:\IRIS.DAT';
```

```
  INPUT X1-X4 SPEC_NO @@;
```

```
  IF SPEC_NO=1 THEN SPECIES='SETOSA    ';
```

```
  IF SPEC_NO=2 THEN SPECIES='VERSICOLOR';
```

```
  IF SPEC_NO=3 THEN SPECIES='VIRGINICA ';
```

```
  LABEL X1='SEPAL LENGTH'
```

```
        X2='SEPAL WIDTH '
```

```
        X3='PETAL LENGTH'
```

```
        X4='PETAL WIDTH ';
```

```
RUN;
```

```
PROC CLUSTER DATA=IRIS METHOD=WARD PRINT=15 CCC;    ① ② ③
```

```
  VAR X1-X4;
```

```
  COPY SPECIES;    ④
```

```
RUN;
```

```
PROC PLOT;    ⑤
```

```
  PLOT _CCC_*_NCL_ / VPOS=26
```

```

HAXIS=1 TO 25 BY 1;
RUN;
PROC TREE NOPRINT NCL=3 OUT=OUT; ⑥
COPY X1-X4 SPECIES;
RUN;
PROC FREQ; ⑦
TABLES CLUSTER*SPECIES;
RUN;
PROC CANDISC NOPRINT OUT=CAN; ⑧
CLASS CLUSTER;
VAR X1-X4;
RUN;
PROC PLOT;
PLOT CAN2*CAN1=CLUSTER; ⑨
RUN; QUIT;

```

<그림 13.8 설명>

- ① Ward의 군집방법을 지정
- ② 최종 15개의 군집 과정을 출력
- ③ 판정기준 CCC (cubic clustering criterion)의 값
- ④ 군집의 개수(_NCL_)에 대하여 CCC의 값(_CCC_)의 산점도

⑤ NCL=3에 군집의 개수는 3으로 지정

OUT에는 관측값번호, 군집번호, 각 개체가 속한 군집이름이
기억되고, 이와 같은 변수들에 $X_1 - X_4$ 와 SPECIES를 추가
하여 OUT이라는 자료가 만듦.

⑥ ⑤번에서 얻은 OUT에서 각 개체의 군집번호와 실제 품종번호의 분할표를 작성

결과 :

setosa는 50개 모두 군집 3으로 옳게 분류
versicolor는 1개가 virginica로 잘못 분류
virginica는 15개가 setosa로 잘못 분류
따라서 모두 16(10.66%)가 잘못 분류

⑦ 정준판별분석을 실시

⑧ 처음 두 개의 정준에 따라 군집번호의 산점도

Ward's Minimum Variance Cluster Analysis

Eigenvalues of the Covariance Matrix

	Eigenvalue	Difference	Proportion	Cumulative
1	422.824171	398.557096	0.9246	0.9246
2	24.267075	16.446125	0.0531	0.9777
3	7.820950	5.437441	0.0171	0.9948
4	2.383509		0.0052	1.0000

Root-Mean-Square Total-Sample Standard Deviation = 10.69224

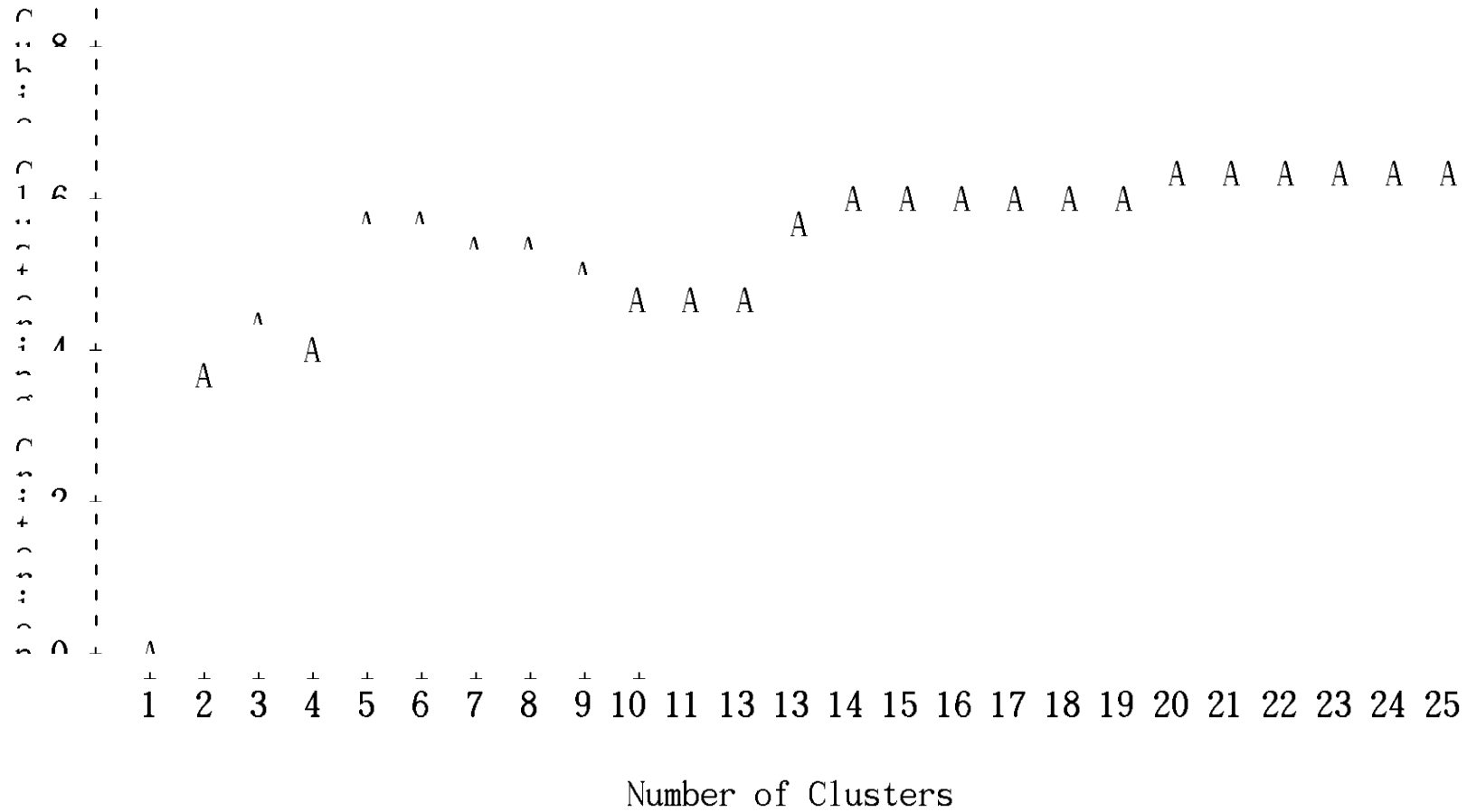
Root-Mean-Square Distance Between Observations = 30.24221

Cluster History

NCL	--Clusters Joined--		FREQ	SPRSQ	RSQ	ERSQ	CCC	T i e
15	CL24	CL28	15	0.0016	.971	.958	5.93	
14	CL21	CL53	7	0.0019	.969	.955	5.85	
13	CL18	CL48	15	0.0023	.967	.953	5.69	
12	CL16	CL23	24	0.0023	.965	.950	4.63	
11	CL14	CL43	12	0.0025	.962	.946	4.67	
10	CL26	CL20	22	0.0027	.959	.942	4.81	
9	CL27	CL17	31	0.0031	.956	.936	5.02	
8	CL35	CL15	23	0.0031	.953	.930	5.44	
7	CL10	CL47	26	0.0058	.947	.921	5.43	
6	CL8	CL13	38	0.0060	.941	.911	5.81	
5	CL9	CL19	50	0.0105	.931	.895	5.82	
4	CL12	CL11	36	0.0172	.914	.872	3.99	
3	CL6	CL7	64	0.0301	.884	.827	4.33	
2	CL4	CL3	100	0.1110	.773	.697	3.83	
1	CL5	CL2	150	0.7726	.000	.000	0.00	

<그림 13.7> 군집과정과 CCC값

Plot of _CCC*_NCL_. Legend: A = 1 obs, B = 2 obs, etc.

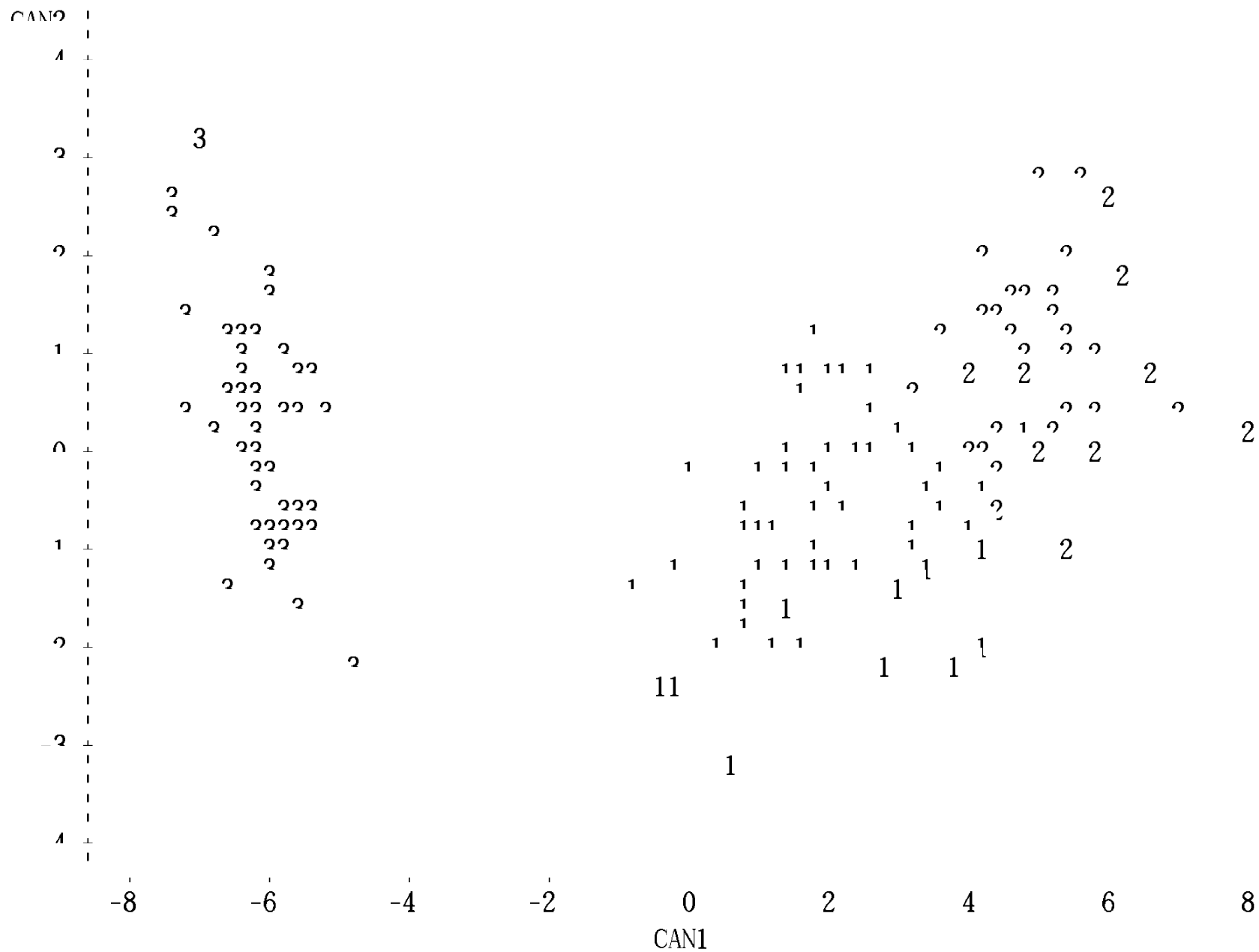


<그림 13.8> 군집개수에 대한 CCC값의 산점도

CLUSTER * SPECIES 교차표				
CLUSTER	SPECIES			
비도 백분율 행 백분율 컬럼 백분율	SETOSA	VERSICOLOR OR	VIRGINICA A	총합
1	0	49	15	64
	0.00	32.67	10.00	42.67
	0.00	76.56	23.44	
	0.00	98.00	30.00	
2	0	1	35	36
	0.00	0.67	23.33	24.00
	0.00	2.78	97.22	
	0.00	2.00	70.00	
3	50	0	0	50
	33.33	0.00	0.00	33.33
	100.00	0.00	0.00	
	100.00	0.00	0.00	
총합	50	50	50	150
	33.33	33.33	33.33	100.00

<그림 13.9> 군집번호와 실제 품종번호 분할표

Plot of CAN2*CAN1. Symbol is value of CLUSTER.



13.3 비계층적 군집분석

13.3.1 이론적 배경

비계층적 군집분석은 주어진 판정기준을 최적화하는 최적분리기법을 사용하며, 따라서 최적분리 군집분석이라고도 한다.

최적분리 군집분석에서는 흔히 군집의 개수를 미리 정하고 다음과 같이 군집을 만든다.

- ① 초기군집을 만들고, 각 개체를 군집에 할당
- ② 각 군집의 일부 또는 전부를 판정기준에 의해 재할당

• 판정기준 : 군집내 제곱합

군집내 산포행렬(within dispersion matrix) :

$$W = \sum_i \sum_j (x_{ij} - \bar{x}_i)(x_{ij} - \bar{x}_i)'$$

행렬인 W 의 크기를 최소화하는 방법

① $\text{tr}(W)$ 의 최소화

W 의 대각선 원소는 각 군집내의 제곱합이므로, $\text{tr}(W)$ 는 이들 제곱합의 총계에 해당

② $|W|$ 의 최소화

W 의 행렬식도 행렬의 크기를 나타내는 척도

FASTCLUS : 최적분리 군집분석을 시행하는 절차

CLUSTER 절차와는 달리 덴드로그램을 만들지 못하며, 군집의 수를 바꾸면 처음부터 모든 계산을 다시 한다.

다음의 예제 : 관측값이 군집으로 분리되는

k -평균 (k -means) 방법

[예제 13.6] 관측값은 2변량 벡터 (x_1, x_2) 로서 다음과 같이 주어질 때, 2개의 군집으로 분리하는 과정을 살펴보자.

개체	관측값 (x_1, x_2)
A	(5, 3)
B	(-1, 1)
C	(1, -2)
D	(-3, 2)
E	(2, 1)

두 군집의 초기값으로 거리가 가장 먼 개체인 A와 D를 선택.
다음에는 각 개체를 중심과의 거리가 가까운 군집으로 할당
다음의 <표 13.2>는 이들 과정의 요약

단계	군집1		군집2	
	군집	중심	군집	중심
반복1	A	$(5, 3)$	D	$(-3, 2)$
1	.	.	BD	$(-2, 1.5)$
2	.	.	BCD	$(-1, 0.33)$
3	.	.	$BCDE$	$(-0.25, 0.5)$
4				
반복2	AE	$(3.5, 2)$	BCD	$(-1, 0.33)$
1				

첫 번째 반복에서 개체 $E(2,1)$ 은 군집 2로 할당

그러나 두 번째 반복에서는 전체의 제곱합을 줄이기 위해서 $E(2,1)$ 은 군집 1로 재할당됨.

13.3.2 FASTCLUS 절차

FASTCLUS 절차의 일반형 :

```
PROC FASTCLUS MAXCLUSTERS=n ; RADIUS=t <options>;  
  VAR variables;  
  ID variables;  
  FREQ variables;  
  BY variables;
```

FASTCLUS 절차의 옵션들

SEED=SAS dsn : 군집의 초기값 자료명을 지정

OUT=SAS dsn : 분석결과를 보관할 자료명

MEAN=SAS dsn : 각종 통계량의 값을 보관할 자료명

CLUSTER=name : 군집의 이름을 지정

디폴트로는 CLUSTER를 사용

MAXCLUSTERS(또는 MAXC)=m : 최대 군집수를 지정

디폴트는 100개

13.3.3 예제

[예제 11.2] 방진복(防塵服) 자료

V1 : 신장 V2 : 총길이 V3 : 등길이 V4 : 화장

V5 : 소매길이 V6 : 바지길이 V7 : 밑위앞뒤

V8 : 가슴둘레 V9 : 허리둘레 V10 : 엉덩이둘레

– 62 –

[예제 13.3](계속) 인자분석에서 인자1과 인자2에 속하는 변수들 가운데 치수를 결정하는 중요 변수

- ① 신장 ② 총길이 ⑤ 소매길이 ⑥ 바지길이
- ⑧ 가슴둘레 ⑨ 허리둘레 ⑩ 엉덩이둘레

적당한 군집수를 알아보기 위하여 CCC를 이용한 분석

```

/* CLUSTER3.SAS : ANALYSIS OF BODY SIZE DATA */
OPTIONS PS=40;
DATA PHYSICAL;
    INFILE 'C:\SASPGM\PHYSICAL.DAT' FIRSTOBS=2 MISSEVER;
    INPUT ID V1-V10;
    LABEL V1='신장' V2='총길이' V3='등길이' V4='화장' V5='소매길이'
          V6='바지길이' V7='밑위앞뒤' V8='가슴둘레' V9='허리둘레'
          V10='엉덩이둘레';
RUN;
PROC CLUSTER DATA=PHYSICAL METHOD=WARD PRINT=0 CCC;
    VAR V1 V2 V5 V6 V8 V9 V10;
RUN;
PROC PLOT;
    PLOT _CCC*_NCL_/HAXIS=1 TO 20 BY 1;
RUN; QUIT;

```


Plot of _CCC*_NCL_. Legend: A = 1  = 2 

