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 만듬

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

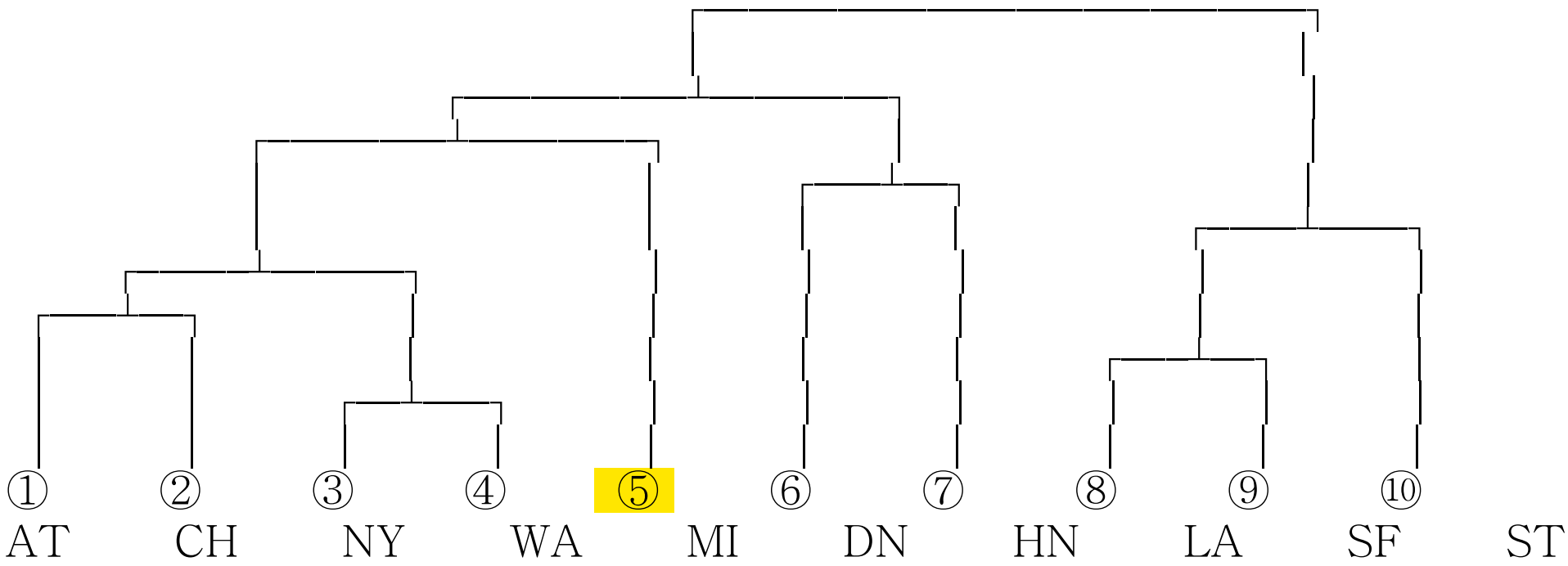
비행거리가 가까운 도시끼리 묶는다면 어떤 순서

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 $\boldsymbol{x}_i' = (X_{i1}, \cdots, X_{ip})$와 $\boldsymbol{x}_j' = (X_{j1}, \cdots, X_{jp})$

사이의 거리의 예:

①유클리드(Euclid)거리 :

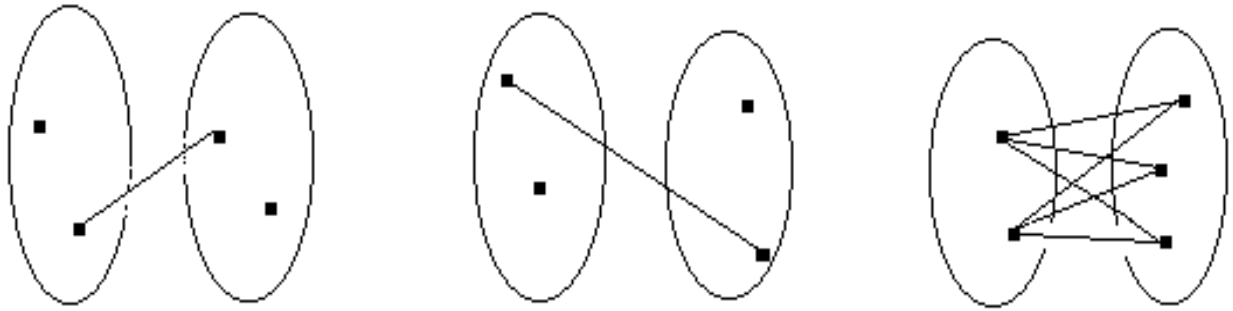
$$d(\boldsymbol{x}_i, \boldsymbol{x}_j) = \sqrt{(X_{i1} - X_{j1})^2 + \cdots + (X_{ip} - X_{jp})^2}$$

②마할라노비스(Mahalanobis)거리 :

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

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,

N개의 관측값 $x_1,\cdots, x_N$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, B가 가장 가까우므로 A와 B를 병합시키면 새로운 거리행렬 D_1은 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, E가 가장 가까우므로 D와 E를 병합시킨 거리행렬은 다음과 같다.

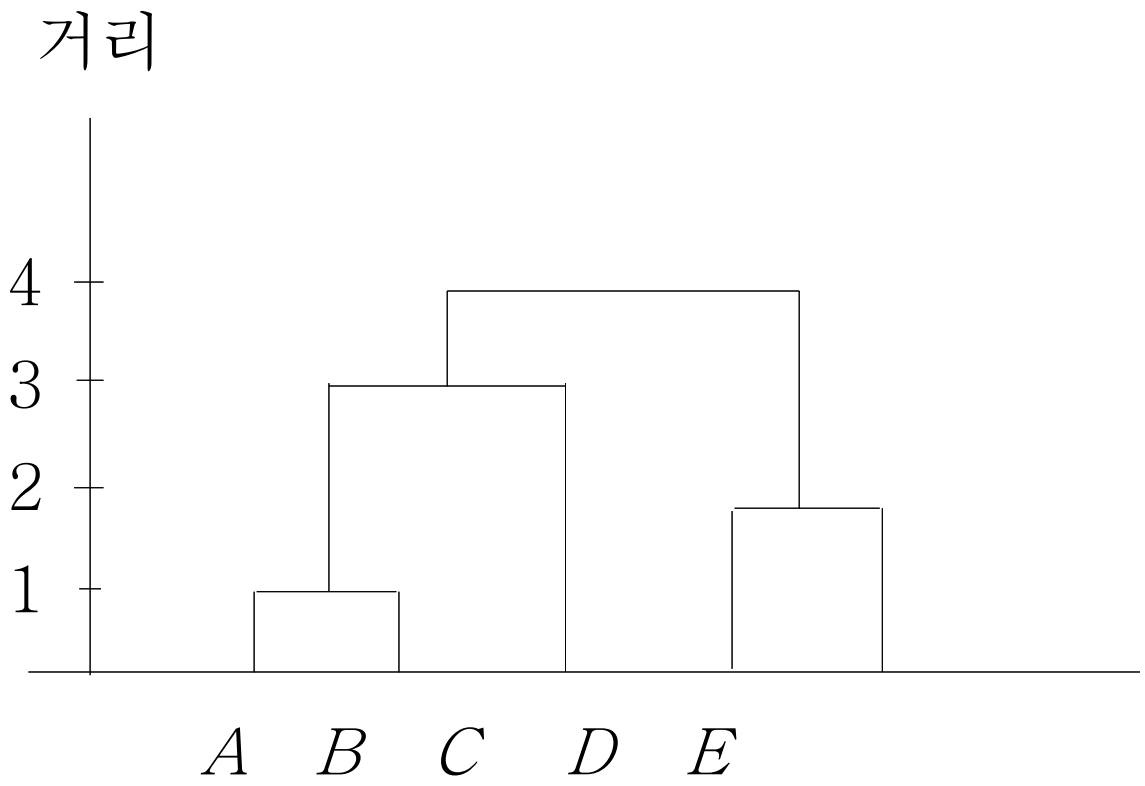
$$D_2 = \begin{pmatrix} 0 & & \\ 3 & 0 & \\ 6 & 4 & 0 \end{pmatrix} \begin{matrix} AB \\ C \\ DE \end{matrix}$$

3단계 AB와 C가 가장 가까우므로 AB와 C를 병합시킨 거리행렬은 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 A와 B가 가장 가까우므로 A와 B를 병합시키고, 새로운 거리행렬을 다음과 같이 구한다.

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

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

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　D와 E를 병합시킨 결과는 다음과 같다.

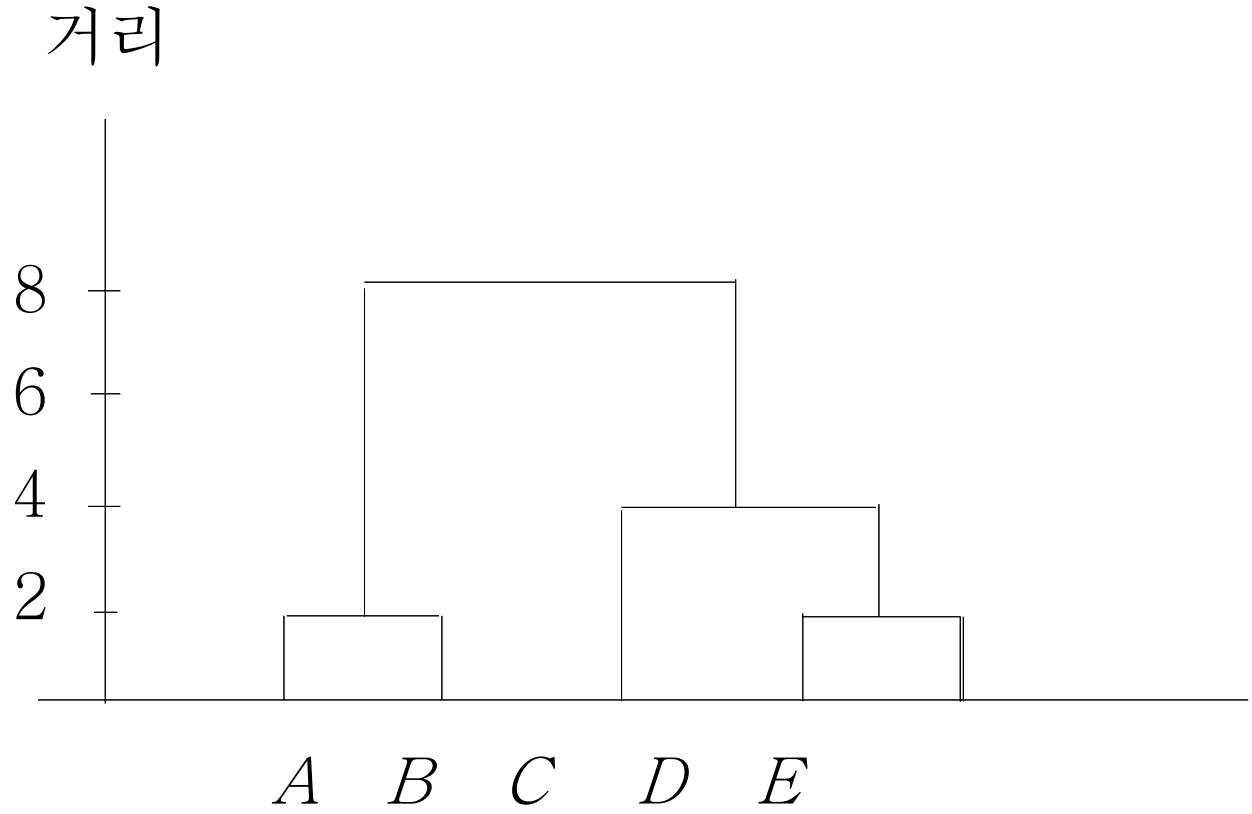
$$D_2=\begin{pmatrix}0 & & \\ 5 & 0 & \\ 8 & 4 & 0\end{pmatrix}\begin{matrix}AB\\ C\\ DE\end{matrix}$$

3단계　C와 DE를 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,

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

(5) 메디안연결법(Median Linkage Method)

중심연결법에서는 크기가 n_1, n_2인 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**

- i번째 군집내의 잔차제곱합 :

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

(j : 개체번호, k : 변수번호)

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

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

두 군집 C_1 과 C_2가 병합될 때에 ESS의 증분 :

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

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                                                    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;
```

```
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.

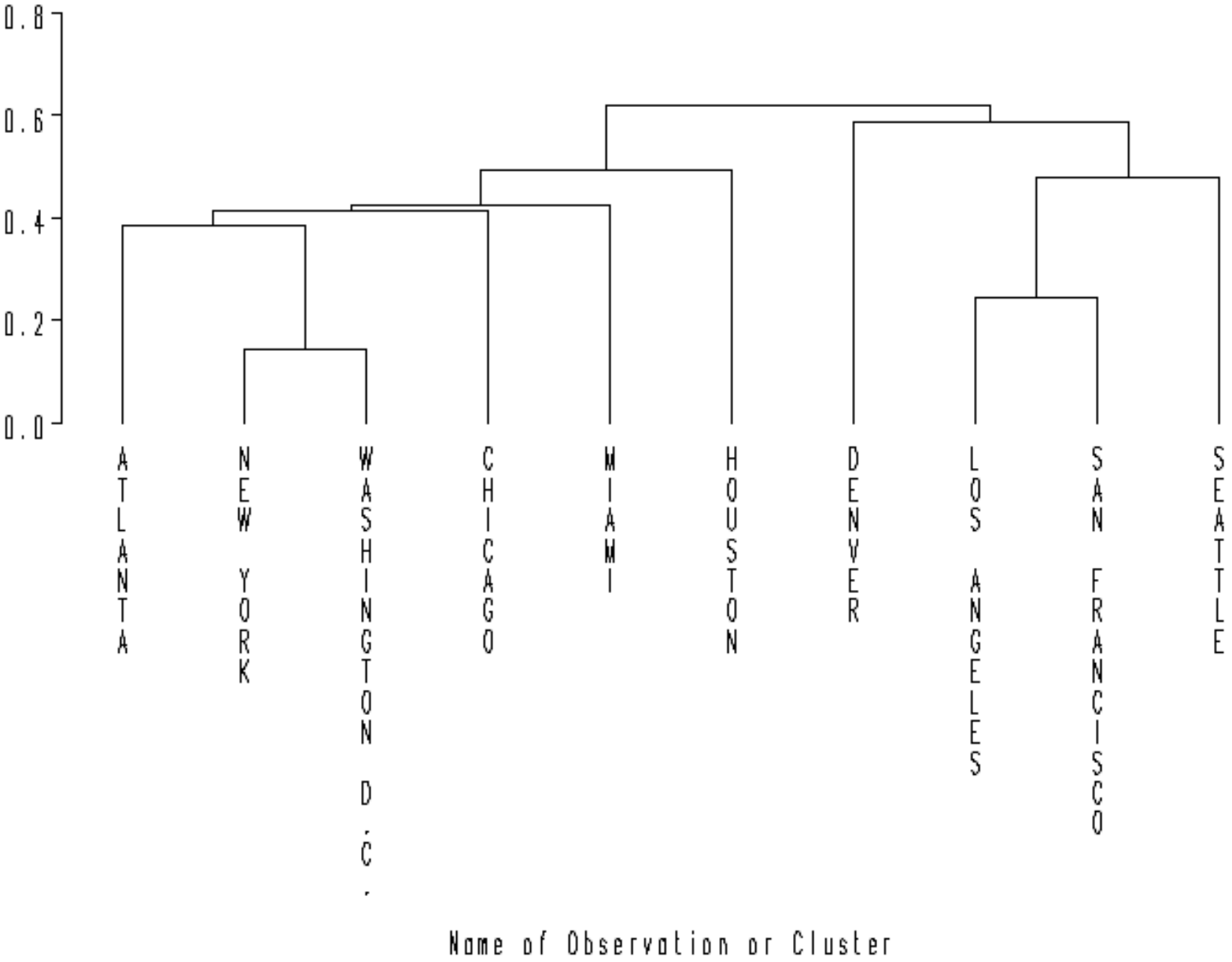
Single Linkage Cluster Analysis

Mean Distance Between Observations = 1417.133

Cluster History

NCL	Clusters Joined		FREQ	Norm Min Dist	Tie
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

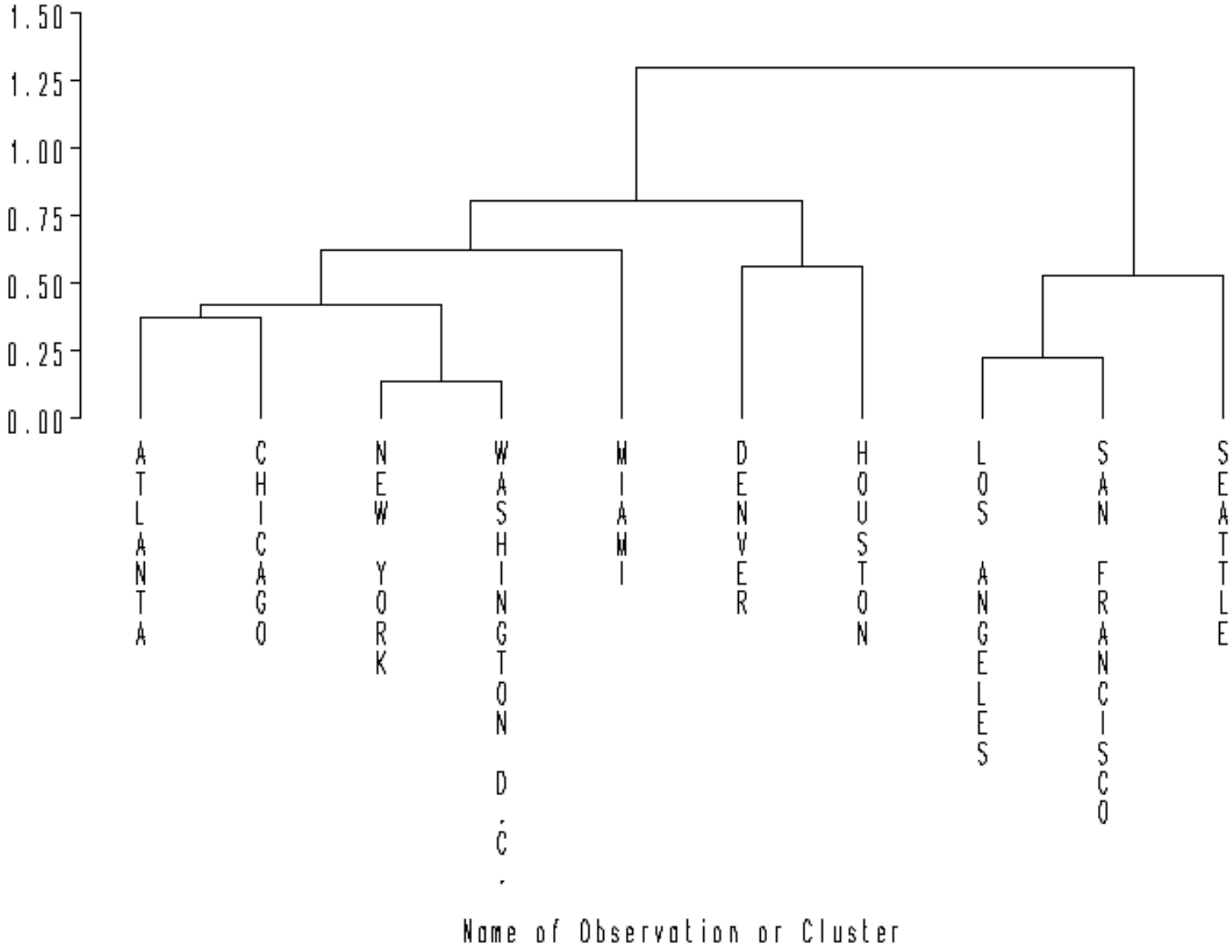
Average Linkage Cluster Analysis

Root-Mean-Square Distance Between Observations = 1580.242

Cluster History

NCL	Clusters Joined		FREQ	Norm RMS Dist	T i 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


1.50
1.25
1.00
0.75
0.50
0.25
0.00
ATLANTA
CHICAGO
NEW YORK
WASHINGTON D.C.
MIAMI
DENVER
HOUSTON
LOS ANGELES
SAN FRANCISCO
SEATTLE
Name of Observation or Cluster

[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_4와 SPECIES를 추가하여 OUT이라는 자료가 만듬.

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

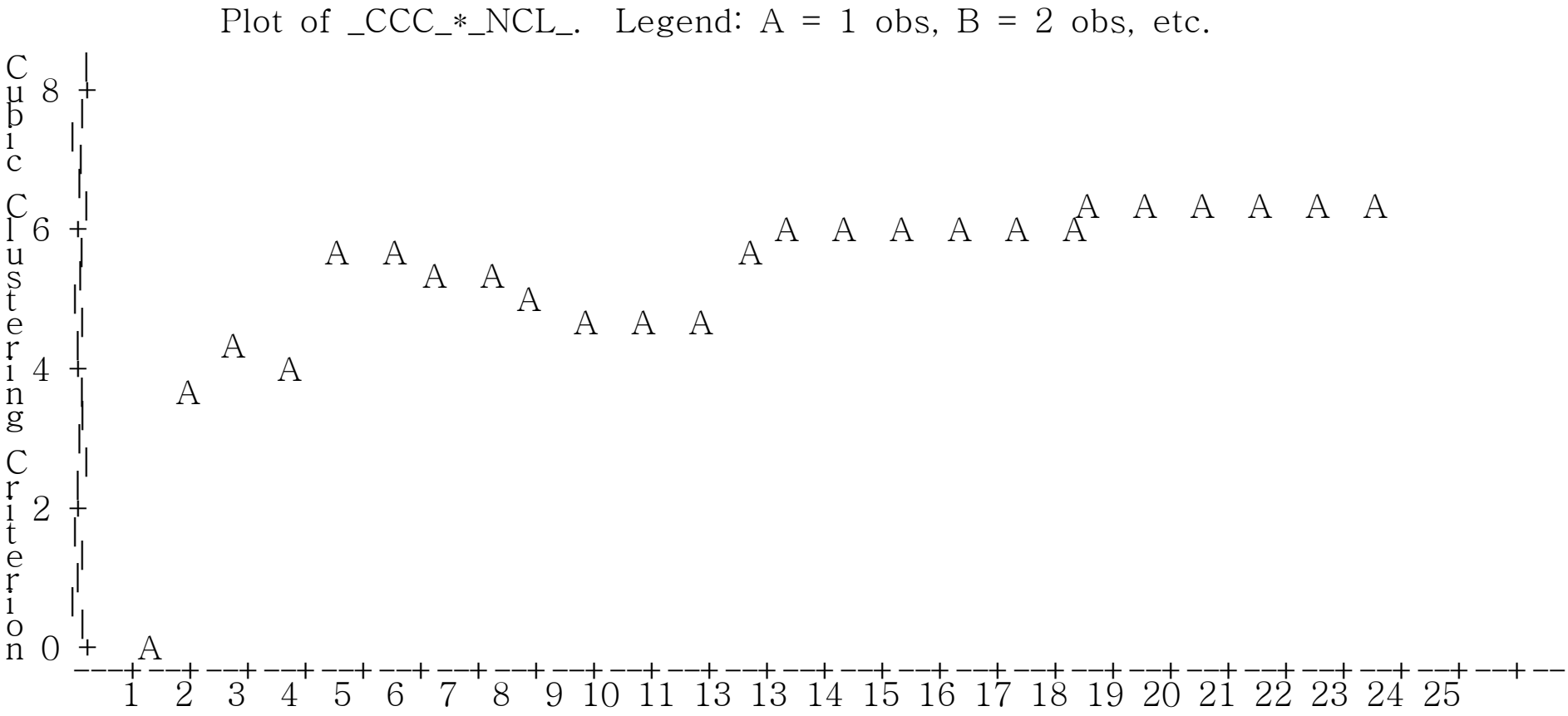
	Eigenvalue	Difference	Proportion	Cumulative
1	422.824171	398.557096	0.9246	0.9246
2	24.267075	16.446125	0.0531	0.9777
3	7.820950	5.437441	0.0171	0.9948
4	2.383509		0.0052	1.0000

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	Ti 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

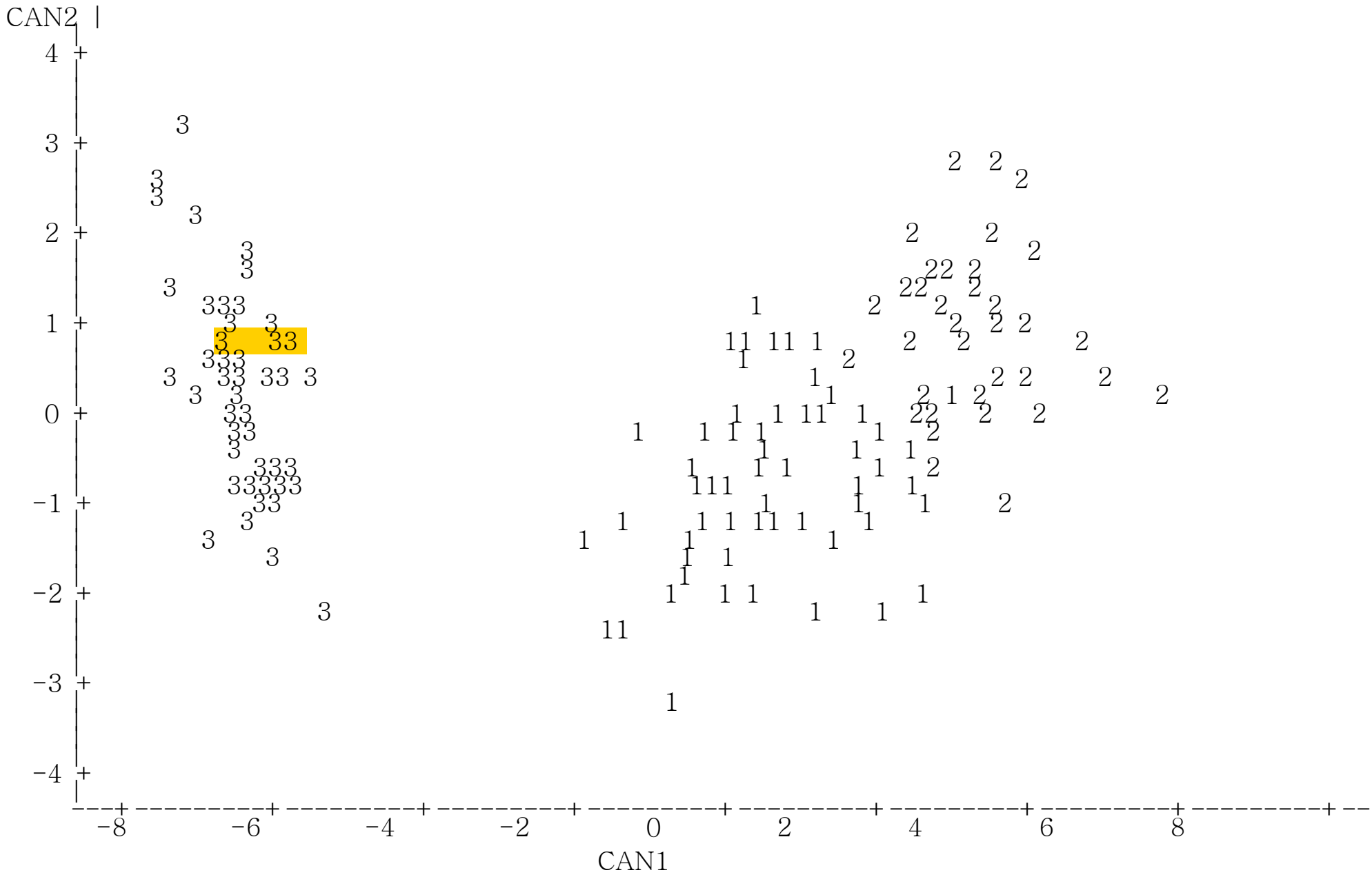


<그림 13.8> 군집개수에 대한 CCC값의 산점도

CLUSTER * SPECIES 교차표

CLUSTER	SPECIES			
빈도 / 백분율 / 행 백분율 / 칼럼 백분율	SETOSA	VERSICOLOR	VIRGINICA	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.
CAN2
4
3
2
1
0
-1
-2
-3
-4
-8
-6
-4
-2
0
2
4
6
8
CAN1

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 W의 크기를 최소화하는 방법

① $\mathrm{tr}(W)$의 최소화

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

② $|W|$의 최소화

W의 행렬식도 행렬의 크기를 나타내는 측도

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				
1	AE	(3.5, 2)	BCD	(-1, 0.33)

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

ID	V1	V2	V3	V4	V5	V6	V7	V8	V9	V10
1	162	142	42	73	51	102	33	81	67	91
2	162	137	37	74	54	103	32	80	62	87
3	156	134	37	72	51	99	32	87	70	93
4	163	143	38	75	56	108	31	88	69	96
5	165	140	41	74	54	103	29	75	60	81
6	160	137	38	72	52	103	30	80	64	94
7	160	138	40	70	52	102	31	81	64	88
8	161	139	38	77	56	104	34	96	78	100
9	161	140	40	72	51	102	34	98	81	100
10	157	138	42	72	51	98	32	91	76	98
11	162	140	38	73	50	104	30	76	61	87
12	158	136	37	68	48	98	30	82	65	91
13	157	131	35	68	49	97	32	88	66	89
14	163	140	38	71	51	104	31	80	66	93
15	168	150	39	74	53	109	32	82	66	93
16	163	143	39	74	54	105	31	83	66	94
17	162	141	38	71	53	104	31	83	69	93
18	160	136	38	71	53	102	30	83	67	87
19	159	135	37	70	54	100	31	85	65	85
20	158	136	36	73	52	92	30	83	68	89
21	164	142	40	72	52	103	31	76	64	87
22	160	137	38	71	51	101	31	83	65	89
23	155	135	39	69	48	99	29	88	62	86
24	158	136	38	73	48	99	32	85	66	90
25	163	140	38	72	51	104	33	89	72	97
26	161	136	37	69	50	101	33	90	75	96
27	161	138	41	72	50	100	32	83	67	88
28	160	136	36	69	50	103	33	82	66	92
29	159	135	38	69	50	100	31	81	68	93
30	157	134	38	69	49	98	30	80	64	88
31	156	134	36	70	51	100	29	80	65	86
32	162	140	38	73	52	104	30	80	65	91
33	157	133	37	72	52	97	31	83	66	84
34	163	143	39	70	52	101	31	78	69	90
35	160	135	38	73	47	98	30	82	68	88
36	155	131	37	68	52	97	28	75	57	83
37	158	134	36	70	53	98	31	83	69	91
38	166	143	42	73	52	103	32	82	65	87
39	162	140	41	71	50	101	31	85	70	89
40	153	134	38	69	48	97	31	87	72	90
41	163	140	42	72	51	101	31	80	65	87
42	165	141	42	73	52	102	31	87	71	95
43	156	136	37	68	50	99	31	79	64	85
44	163	141	39	71	50	105	28	71	56	80
45	165	142	40	69	52	105	30	82	64	90
46	163	139	39	70	52	102	31	81	64	95
47	160	139	40	68	50	99	31	87	69	96
48	164	143	39	71	50	104	33	84	74	96
49	171	144	43	73	53	107	30	88	71	97
50	153	133	35	70	48	98	32	81	68	91
51	164	141	41	73	52	101	32	92	76	96
52	160	138	41	69	52	101	31	83	70	93
53	165	144	41	77	53	103	31	82	72	93
54	157	135	37	67	48	99	32	86	75	92
55	157	136	37	71	50	99	31	79	68	94
56	156	133	37	71	52	97	30	86	68	94
57	159	135	38	71	52	98	32	85	70	94
58	162	139	39	73	53	101	32	85	69	95
59	163	140	39	72	53	102	31	80	65	95
60	154	132	38	71	52	95	32	87	72	90
61	160	138	38	73	53	100	32	85	76	95
62	155	132	37	69	50	96	28	70	62	84
63	165	143	41	74	53	101	34	98	81	104
64	153	132	36	68	49	95	30	78	65	86
65	172	148	45	74	51	105	30	81	68	97
66	162	138	39	71	53	99	32	77	69	87
67	162	142	42	72	52	99	34	89	75	97
68	159	135	39	67	49	98	30	81	64	89
69	159	132	37	69	50	97	31	80	65	91
70	168	146	39	74	53	108	31	89	73	98
71	162	137	36	72	52	103	29	74	58	80
72	152	133	37	70	48	97	32	85	71	89
73	163	139	39	70	51	101	30	83	69	91
74	165	140	39	71	52	101	30	80	65	86
75	166	148	34	72	54	104	31	89	71	96
76	167	147	42	75	55	108	32	81	69	91
77	160	137	39	67	48	100	29	72	63	86
78	157	133	37	65	47	97	29	80	64	87
79	150	127	33	66	48	93	29	76	61	84
80	172	144	41	72	53	107	33	85	71	96
81	155	132	40	69	48	93	31	87	69	91
82	159	137	41	67	50	98	31	77	63	89
83	160	135	39	72	51	97	33	86	69	93
84	163	141	37	69	51	103	31	82	68	91
85	167	146	40	75	55	107	31	82	70	91
86	158	137	40	72	50	99	33	82	74	93
87	162	141	42	72	49	102	33	85	72	93
88	169	148	41	74	54	108	31	81	67	90
89	163	139	40	72	50	100	31	80	67	91
90	161	139	39	72	53	100	31	91	76	93
91	157	135	39	68	47	98	30	76	61	88
92	163	141	37	72	51	105	30	82	64	89
93	164	140	37	74	52	103	30	77	62	90
94	156	132	37	65	48	96	29	75	62	83
95	159	139	39	69	51	101	30	77	66	91
96	163	140	38	71	52	104	31	86	67	92
97	168	145	40	74	54	103	30	77	67	92
98	158	137	39	75	54	100	33	89	72	93
99	158	134	38	65	46	96	33	89	75	95
100	159	137	40	72	52	101	28	82	65	90
101	159	139	40	73	52	101	30	76	61	89
102	152	130	37	68	48	94	31	80	65	89
103	160	139	38	72	53	101	31	77	63	86
104	154	133	39	69	49	98	29	83	64	88
105	162	141	39	70	51	102	32	87	73	95
106	160	138	39	73	52	100	31	80	65	88
107	163	142	40	70	54	102	31	85	68	90
108	162	139	36	71	50	103	30	77	65	86
109	162	140	41	71	51	103	31	76	63	89
110	156	134	38	69	51	98	30	77	69	94
111	160	136	37	71	51	100	31	83	69	90
112	162	140	38	74	54	103	32	82	69	91
113	157	134	38	69	49	96	32	83	69	89
114	161	139	39	70	50	101	31	82	65	91
115	156	134	39	69	49	97	29	80	63	88
116	155	134	37	68	49	100	30	81	65	90
117	160	135	38	69	49	101	31	82	69	90
118	163	139	38	71	52	103	33	89	72	89
119	157	135	35	68	50	101	29	75	63	87
120	157	133	36	66	47	99	31	75	63	92
121	160	136	38	72	51	99	33	92	75	97
122	158	134	36	73	53	99	32	84	69	96
123	159	135	37	69	52	102	31	83	66	91
124	159	136	37	72	54	104	29	80	64	90
125	174	149	44	75	56	106	35	81	75	93
126	170	149	43	78	57	106	37	93	84	103
127	168	145	40	78	57	105	39	102	89	103
128	174	151	43	79	58	110	37	86	76	98
129	173	151	43	78	59	112	37	92	78	96
130	170	147	42	81	59	108	38	93	74	94
131	170	146	41	76	56	107	36	91	83	94
132	164	143	38	75	54	104	33	86	70	97
133	158	135	37	70	51	102	30	85	69	96
134	168	146	39	78	58	109	34	92	78	98
135	168	145	38	78	56	111	32	82	66	94
136	164	140	38	72	53	105	30	79	67	93
137	172	147	38	77	58	112	32	83	72	101
138	152	130	35	69	51	98	31	80	66	92
139	165	141	36	72	55	107	30	85	67	89
140	152	132	36	68	50	96	32	84	71	94
141	160	137	38	68	52	101	29	74	62	89
142	158	135	35	71	52	102	33	79	65	94
143	159	139	38	72	53	103	30	80	68	89
144	161	138	36	70	52	104	29	77	65	88
145	152	132	32	70	50	101	30	81	68	92
146	161	139	35	71	53	105	30	80	66	90
147	163	138	38	72	52	100	32	82	68	94
148	156	133	35	68	49	100	31	75	63	89
149	156	132	36	67	48	97	30	82	66	88
150	160	140	42	74	54	100	35	101	85	102
151	154	131	38	70	50	95	31	84	71	94
152	158	134	36	68	50	101	31	84	64	92
153	165	141	40	73	53	105	34	90	75	98
154	160	138	39	69	50	100	30	81	63	92
155	160	139	38	71	54	104	30	75	60	88
156	157	134	38	69	51	98	32	86	72	94
157	165	140	38	69	51	105	32	86	70	97
158	171	151	44	79	57	109	34	96	77	100
159	163	140	36	72	52	105	31	85	70	94
160	153	130	36	69	50	94	29	75	60	87
161	168	146	40	77	57	108	32	81	65	93
163	160	137	38	73	53	103	29	76	60	90
164	163	139	39	72	52	101	32	79	64	93
165	158	138	37	74	55	101	31	78	65	91
166	157	133	40	69	49	97	32	84	61	89
167	164	142	41	72	53	104	32	81	66	93
168	158	132	38	67	49	95	35	94	80	95
169	159	135	39	71	53	99	33	86	69	95
170	155	130	36	69	51	95	31	84	68	90
171	161	137	39	72	52	101	33	86	68	92
172	158	135	39	72	51	97	31	90	76	94
173	160	136	36	73	54	104	31	80	66	90
174	162	138	39	71	52	100	30	85	69	91
175	153	129	37	68	50	95	28	75	59	83
176	155	135	40	70	51	100	32	89	79	93
177	162	142	37	73	53	106	31	85	69	94
178	168	146	41	78	56	108	35	99	79	108
179	162	139	37	75	55	104	34	93	78	97
180	160	139	34	70	55	107	29	75	59	88
181	152	131	39	68	52	95	31	86	73	90
182	158	136	35	71	54	103	33	86	73	96
183	156	135	38	70	51	98	31	87	70	96
184	155	132	35	68	50	99	33	84	69	97
185	166	144	39	77	58	109	33	81	67	96
186	160	135	35	69	52	102	30	81	66	94
187	160	138	37	71	52	103	34	84	72	98
188	165	142	37	75	54	106	33	82	70	95
189	168	145	39	73	54	109	30	77	60	87
190	158	135	38	70	52	102	28	80	64	89
191	160	138	38	72	53	103	30	82	67	92
192	157	137	36	69	50	103	31	83	69	91
193	158	138	39	70	51	99	31	82	64	95
194	158	137	39	73	52	101	32	81	61	91
195	162	141	37	73	54	107	31	89	74	94
196	157	135	41	72	53	97	31	85	65	92
197	162	139	41	72	52	101	31	78	61	90
198	164	141	41	73	53	104	34	98	80	100
199	163	140	38	72	52	104	32	81	68	94
200	160	139	39	70	53	103	30	81	63	91
201	166	145	39	76	57	109	32	87	77	100
202	170	148	40	78	59	114	32	89	71	102
203	154	136	41	71	51	98	32	85	71	101
204	155	135	39	71	53	98	30	78	62	86
205	157	134	39	69	51	97	31	78	63	92
206	158	136	37	69	50	101	31	79	59	89
207	157	136	38	70	52	99	33	88	74	94
208	158	136	38	73	54	101	31	86	72	96
209	158	138	39	74	52	103	33	90	73	96
210	161	140	39	73	52	103	33	86	70	95
211	160	138	38	73	54	103	30	79	63	90
212	164	142	43	73	52	103	35	84	70	98
213	163	140	42	70	51	100	32	88	72	98
214	160	136	37	71	52	99	32	82	68	96
215	155	133	37	72	52	98	32	92	72	98
216	157	137	36	69	50	102	31	85	71	96
217	164	144	41	73	55	105	34	89	72	98
218	160	134	36	70	51	100	31	87	69	91
219	158	134	36	70	51	99	30	81	67	89
220	160	137	36	70	51	103	31	89	73	93
221	158	136	38	69	51	99	33	86	74	100
222	162	140	39	73	54	103	33	89	75	99
223	165	140	38	75	56	102	33	99	80	104
224	156	133	35	69	52	101	31	80	65	90
225	162	140	36	73	53	104	32	87	70	96
226	155	132	35	69	52	101	32	82	68	89
227	151	127	36	69	51	95	31	97	80	96
228	156	136	37	70	51	100	32	84	67	90
229	154	132	37	69	50	96	31	88	69	97
230	163	143	39	77	56	105	36	102	83	106
231	168	139	41	74	55	107	30	81	63	91
232	160	136	35	71	52	103	29	80	61	86
233	157	134	36	70	52	102	31	85	73	93
234	157	136	40	68	50	100	31	85	73	97
235	155	131	36	69	50	97	31	85	66	88
236	158	138	39	68	52	103	29	77	63	92
237	163	140	38	72	53	105	32	86	70	97
238	155	132	37	66	48	96	30	76	61	88
239	163	141	39	72	52	101	33	84	69	96
240	162	138	40	70	53	100	29	75	61	92

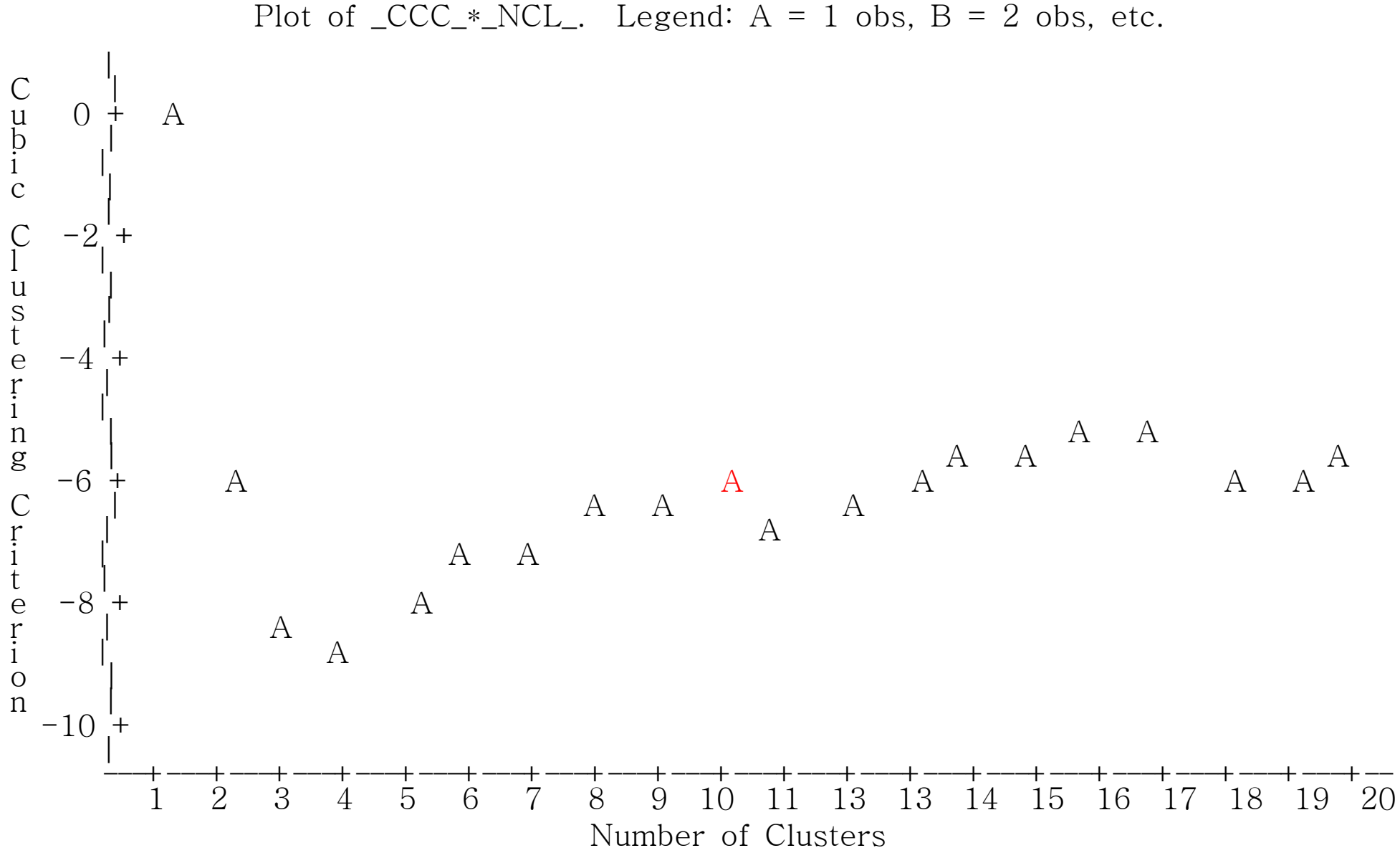
[예제 13.3](계속) 인자분석에서 인자1과 인자2에 속하는 변수들 가운데 치수를 결정하는 중요 변수

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 MISSOVER;
  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;
```



<그림 13.11> 군집개수에 대한 CCC의 산점도

10개의 군집수에서 국소최고점이 있으므로 10개의 군집을 고려

이와 같은 문제에서는 군집 개수를 치수체계의 특성 및 제조회사의 요청에 의해 재조정하는 것이 바람직함.

FASTCLUS 절차를 이용하여 10개의 군집으로 분리시키고, 각 군집내에서 변수들의 평균을 구하여 출력

```
/* CLUSTER4.SAS : ANALYSIS OF BODY SIZE DATA  */
DATA PHYSICAL;
  INFILE 'C:₩SASPGM₩PHYSICAL.DAT' FIRSTOBS=2 MISSOVER;
  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 FASTCLUS DATA=PHYSICAL MAXCLUSTERS=10 OUT=OUTCLUS;   ①
  VAR V1 V2 V5 V6 V8 V9 V10;
RUN;
```

```
PROC SORT DATA=OUTCLUS;                    ②
  BY CLUSTER;
RUN;
PROC MEANS DATA=OUTCLUS NOPRINT;
  VAR V1 V2 V5 V6 V8 V9 V10;
  BY CLUSTER;
  OUTPUT OUT=CLMEAN MEAN=MEANV1 MEANV2 MEANV5 MEANV6 MEANV8-MEANV10;  ③
RUN;
PROC SORT DATA=CLMEAN OUT=CLMEAN;    ④
  BY MEANV1;
RUN;
PROC PRINT DATA=CLMEAN;                    ⑤
  VAR MEANV1 MEANV2 MEANV5 MEANV6 MEANV8-MEANV10 _FREQ_;
RUN; QUIT;
```

① MAXCLUS=10 에서 군집수를 10개로 지정

② OUTCLUS에 기억된 군집 내용을 CLUSTER 번호 순서대로 재배열

③ 각 군집내에서 변수들의 평균을 구하고 이들에 MEANV1… 등의 이름을 부여

④ 각 군집별 평균이 기억된 자료명인 CLMEAN을 첫 번째 변수인 신장(MEANV1)에 의해 재배열

⑤ CLMEAN의 내용을 각 군집의 도수와 함께 출력

FASTCLUS 절차에 의한 출력결과 :

군집 초기값(initial seeds)은 디폴트로 출력 (생략)

이 외에 각 군집별로 기술통계량의 값들과

군집 사이의 정보가 출력 (그림에서는 생략)

10개의 군집을 평균신장(MEANV1)의 크기순으로 출력한 내용이

최종적인 군집 내용

이들로부터 다음과 같은 특징을 지적할 수 있다.

FASTCLUS Procedure: Replace=FULL Radius=0 Maxclusters=10 Maxiter=1

OBS	MEANV1	MEANV2	MEANV5	MEANV6	MEANV8	MEANV9	MEANV10	_FREQ_
1	154.500	129.500	50.0000	95.000	95.5000	80.0000	95.500	2
2	154.571	131.857	49.2143	96.571	75.6429	61.7143	86.357	14
3	156.825	134.750	50.9500	98.500	86.3750	71.5500	94.425	40
4	157.255	134.490	50.3137	99.020	81.8824	65.6471	89.686	51
5	161.727	139.364	52.1818	103.455	74.7273	60.2727	85.545	11
6	162.094	139.358	52.3962	102.698	79.8868	65.1132	90.509	53
7	162.632	140.579	52.3947	103.105	86.9211	71.8421	95.184	38
8	163.889	141.889	54.6667	103.444	99.2222	81.7778	103.000	9
9	169.714	146.786	55.0000	108.214	83.0714	69.7143	94.857	14
10	169.750	147.875	57.7500	109.250	91.6250	77.7500	98.375	8

<그림 13.12> 군집분석에서 최종적으로 얻은 군집

신장이 154인 그룹이 두 군집으로 나누어진 것은 둘레를 나타내는 변수들이 많이 다르기 때문이다. 이와 같은 현상은 신장이 156~157인 그룹, 161~163인 그룹, 169인 그룹들에서도 일어난다. 즉, 신장이 같더라도 가슴둘레, 허리둘레, 엉덩이둘레가 심각하게 다른 군집이 존재함을 뜻하며, 이에 대한 충분한 고려를 해서 치수체계를 만들어야만 좀더 많은 사용자들이 만족할 것이다.