Cluster Analysis

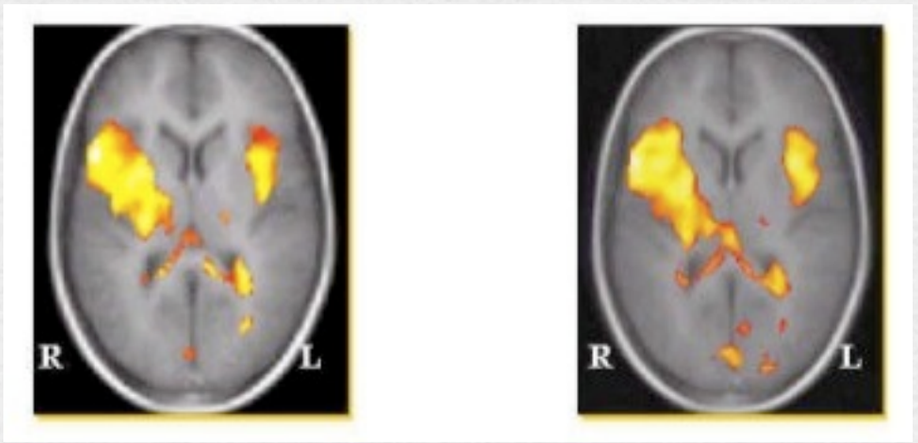
(군집분석)

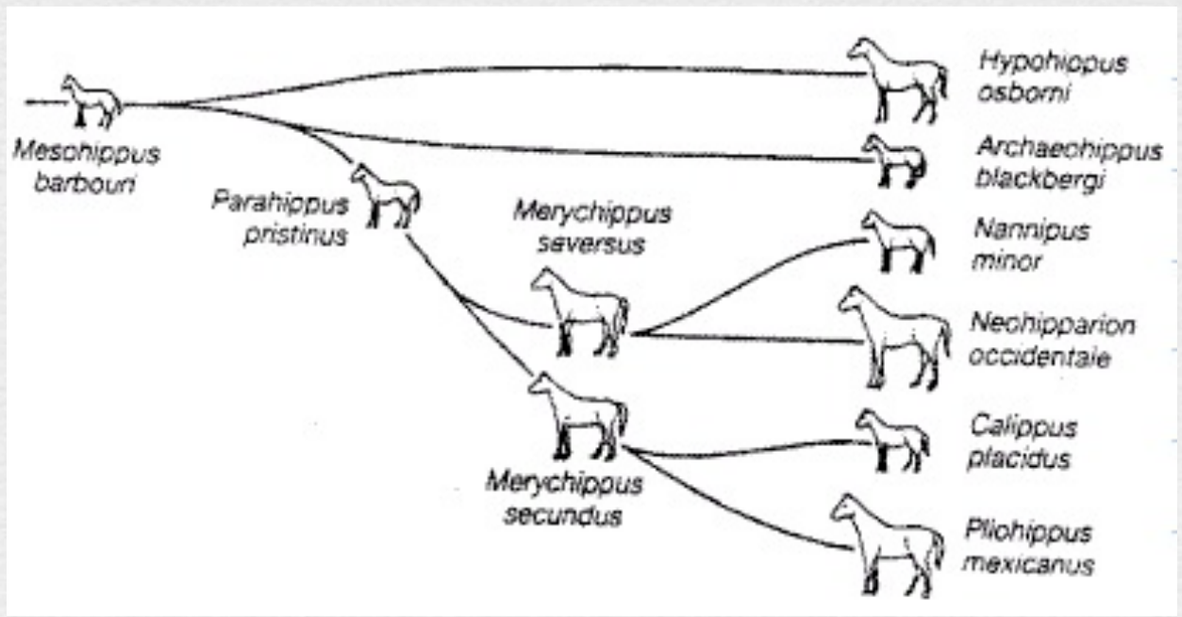
prof. Heungsun Park

hspark@hufs.ac.kr

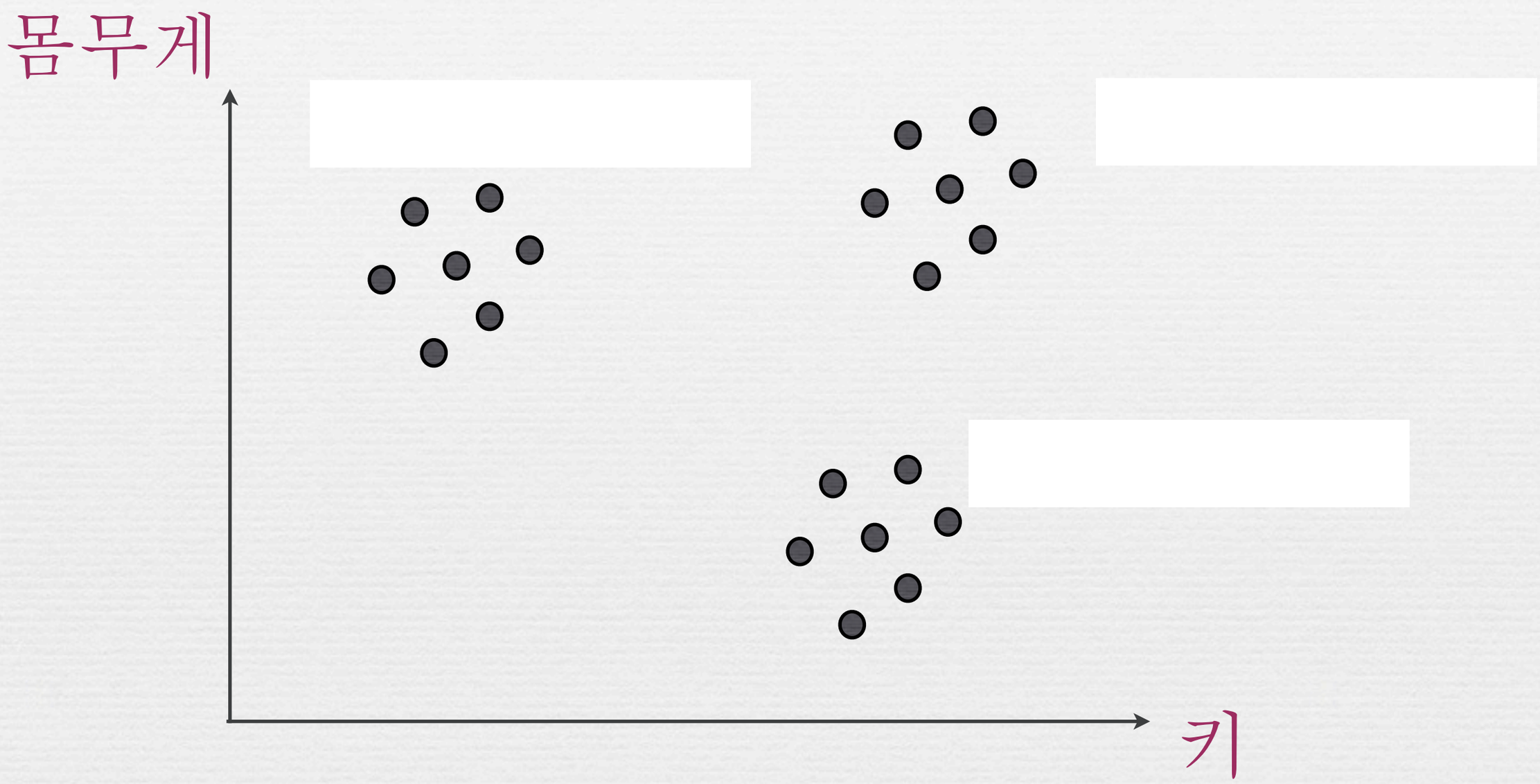
Cluster Analysis

- "다변량의 자료들을 서로 비슷한 자료끼리는 그룹을 짓고, 상이한 자료들은 서로 다른 그룹에 속하도록 분리하는 통계분석방법"
- (e.g.) 병원에 입원한 환자의 정보(x_1=체격, x_2=증상, ... x_q=운동습관)를 이용하여 환자들을 그룹화(건강체질, 허약체질 etc.)
-
-





example : 신체조건(키, 몸무게)



"주어진 자료를 가지고 가능한 그룹들을 생성함"

example: Market Segmentation

A is a sub-set of a market made up of people or organizations with one or more characteristics that cause them to demand similar product and/or services based on qualities of those products such as price or function.

-Wikipedia-

Examples:

- Gender
- Location
- Religion
- Income
- Household size

http://dongabiz.com/Business/Marketing/article_content.php?atno=1202033901&chap_no=1

Clustering

- Agglomerative / Bottom up-Clustering
- Divisive Clustering

Clustering

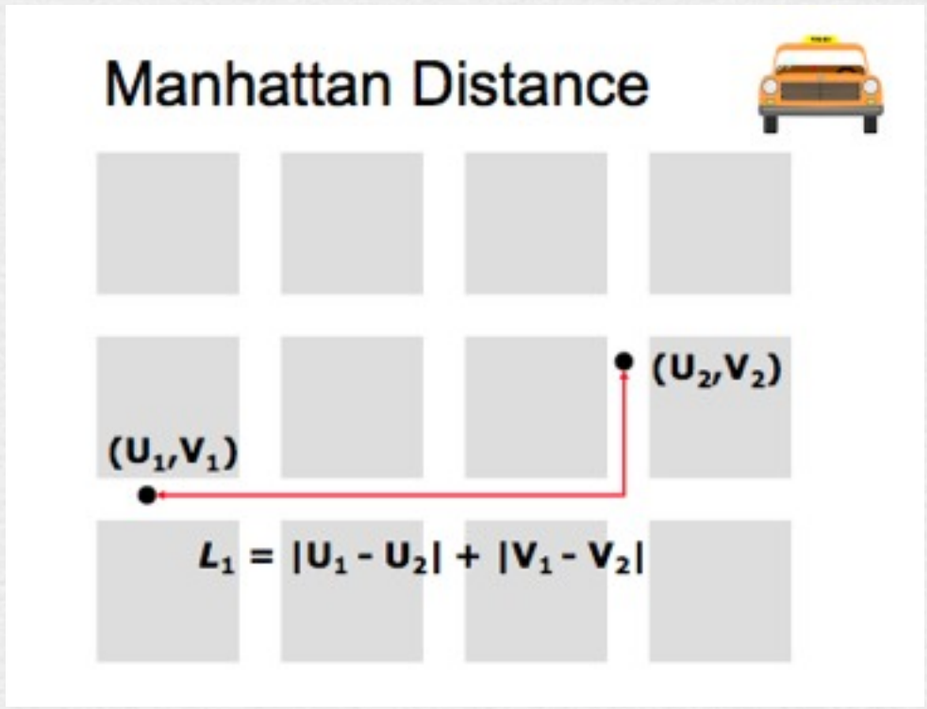
Distance

- Minkowski Distance

$$\begin{aligned} \boldsymbol{x}_i &= (x_{i1}, x_{i2}, \cdots, x_{ip}) \\ \boldsymbol{x}_j &= (x_{j1}, x_{j2}, \cdots, x_{jp}) \end{aligned}$$

- Euclidean Distance (m=2)

- Manhattan Distance (m=1)



- Mahalanobis Distance

$$d_{ij} =$$

Data Matrix

$$X = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix}$$

Dissimilarity Matrix

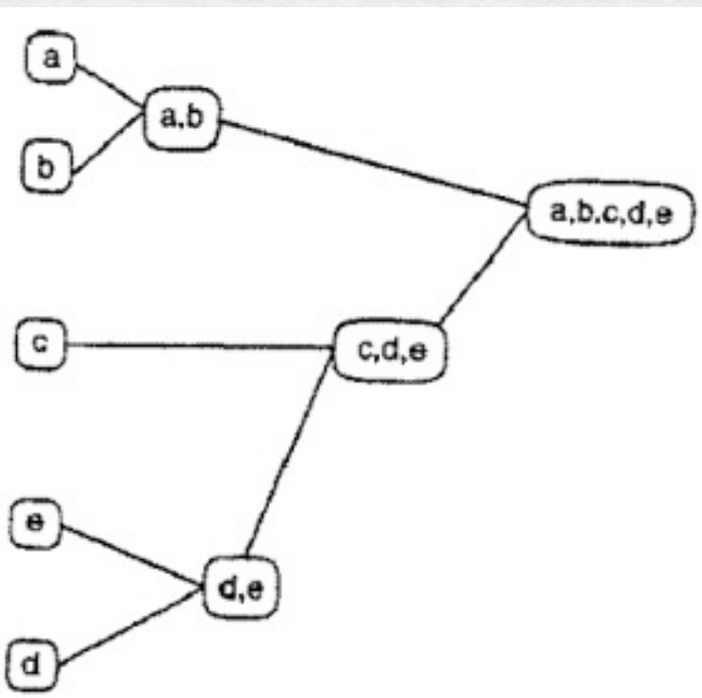
$$D = \begin{pmatrix} 0 & & & \\ d_{21} & 0 & & \\ \vdots & \vdots & \ddots & \vdots \\ d_{n1} & d_{n2} & \cdots & 0 \end{pmatrix}$$

응집적 군집
(agglomerative -bottom up-clustering)

1. n 개의 자료를 각각의 군집으로 간주한다. (c_1 , c_2 , . . . , c_n)
2. (c_i , c_j) 짝 가운데 [redacted] 에 있는 두 군집을 한개의 군집으로 합친다 >> n-1 개 군집형성
3. 다시 (2)를 반복하여 n-2 개 군집형성
4. 마지막 1개의 군집이 형성될 때 까지 계속 한다.

(Q) 3개의 군집을 만들면?

[redacted]

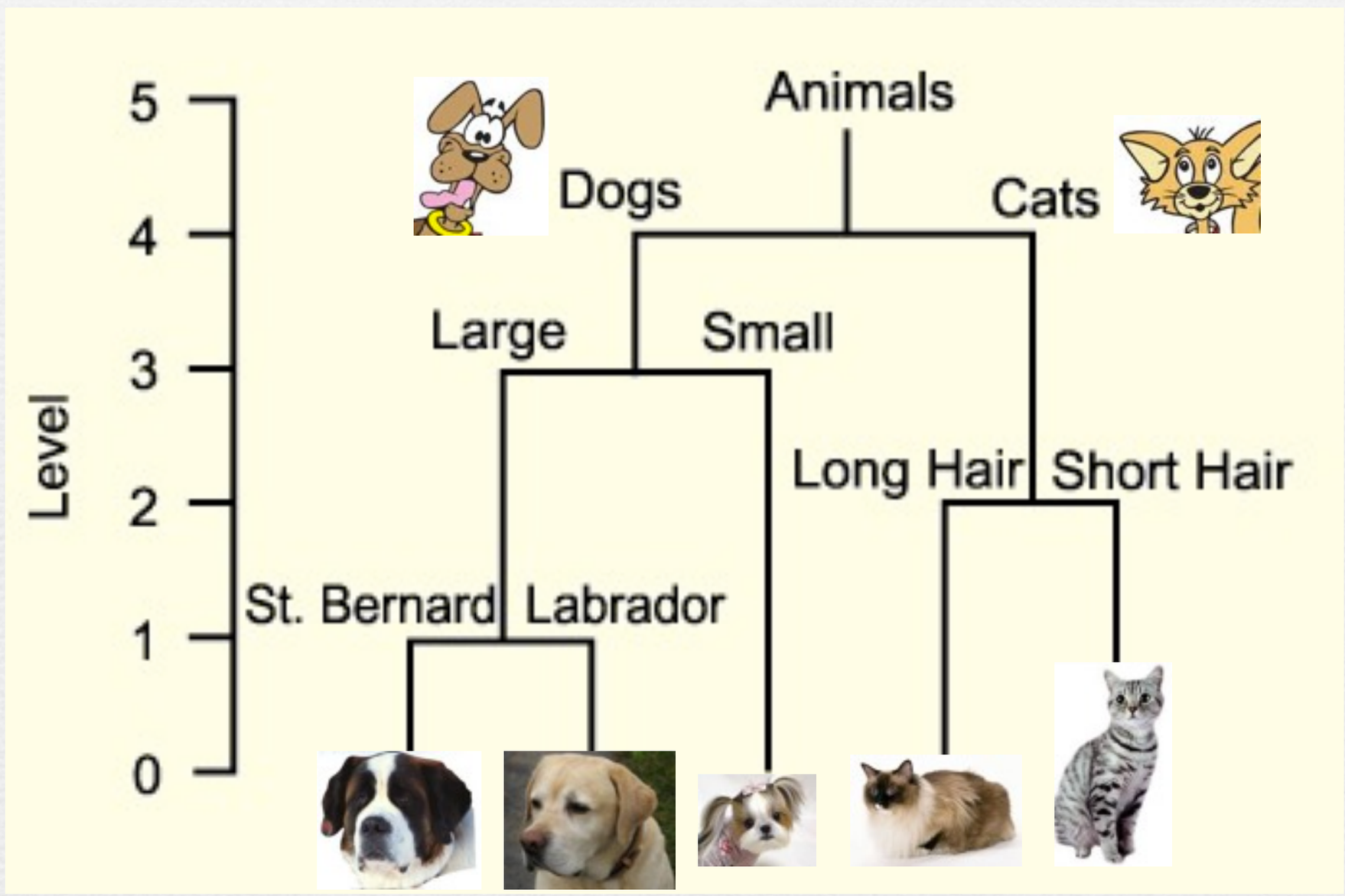


분열적 군집
(Divisive Clustering)

- 한 개의 군집에서 출발하여 분열해 나가는 방식

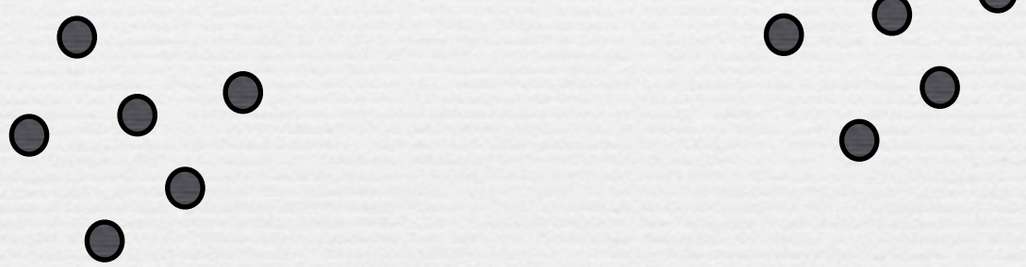
-

- 특징에 따라 구분해 나감.



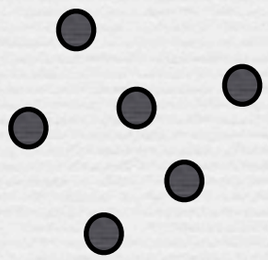
두 그룹간 거리를 측정하는 방법 (bottom-up)

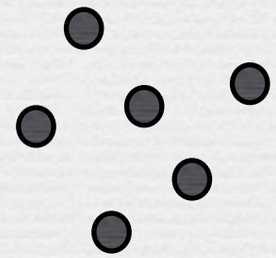
Single Linkage $d_{AB} =$



Complete Linkage $d_{AB} =$

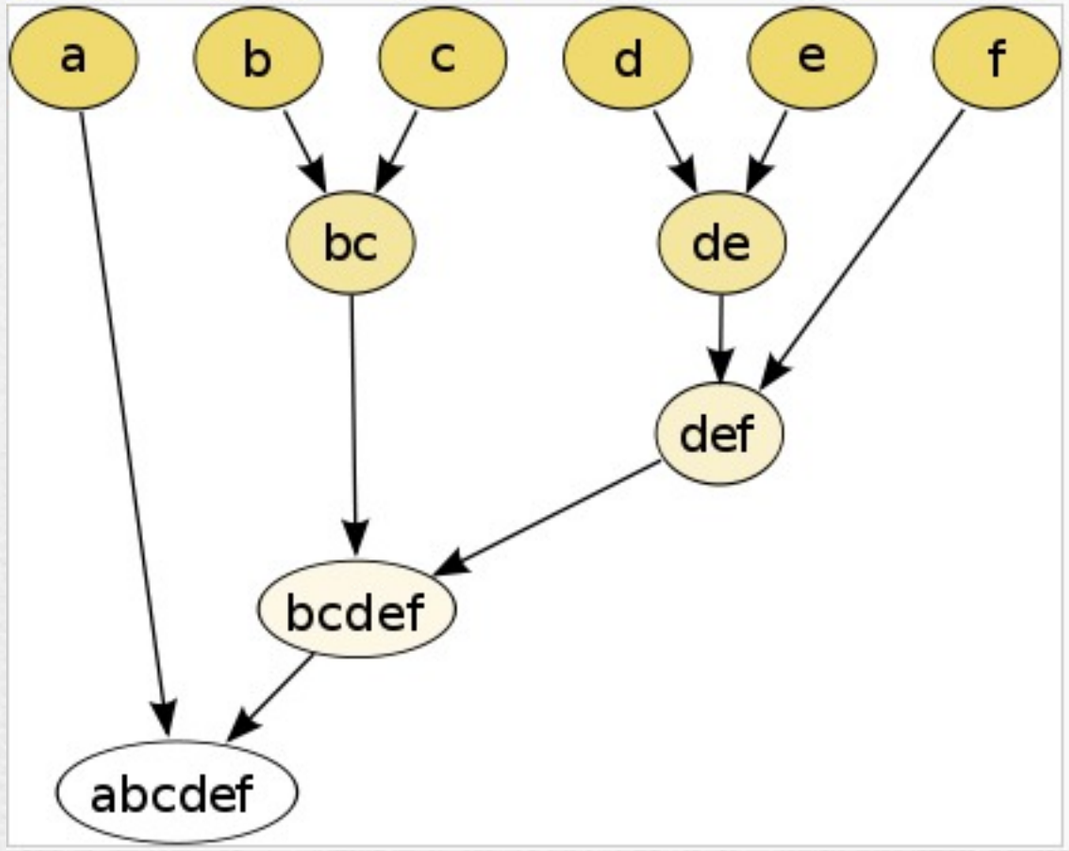
Average Linkage $d_{AB} =$





Dendrogram(tree plot)

- Tree Sturucture
- Height of a node = distance of two clusters
- Clustering 뿐 아니라 closeness (similarity) 제공
-
-

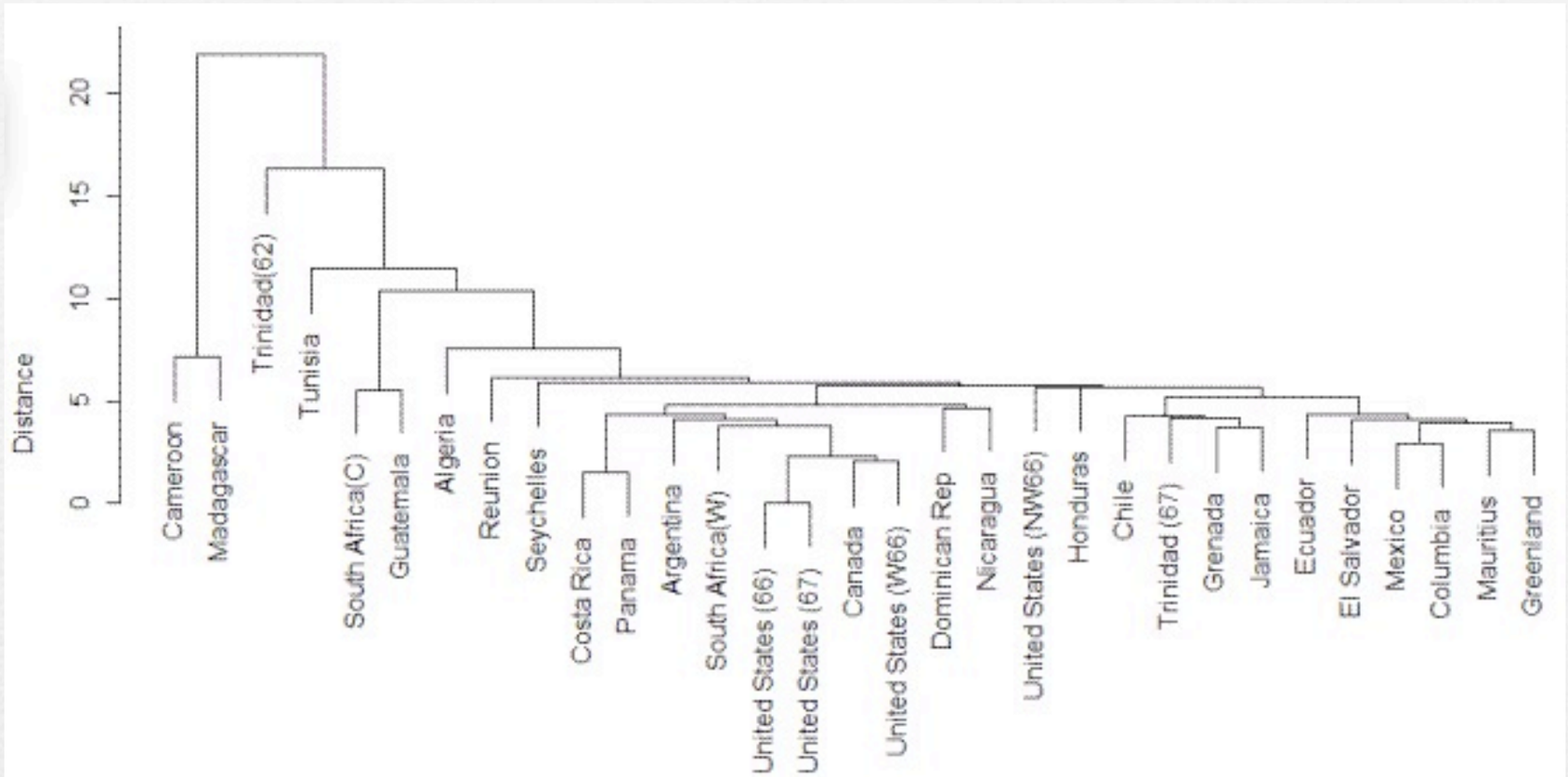


Cluster Analysis in R

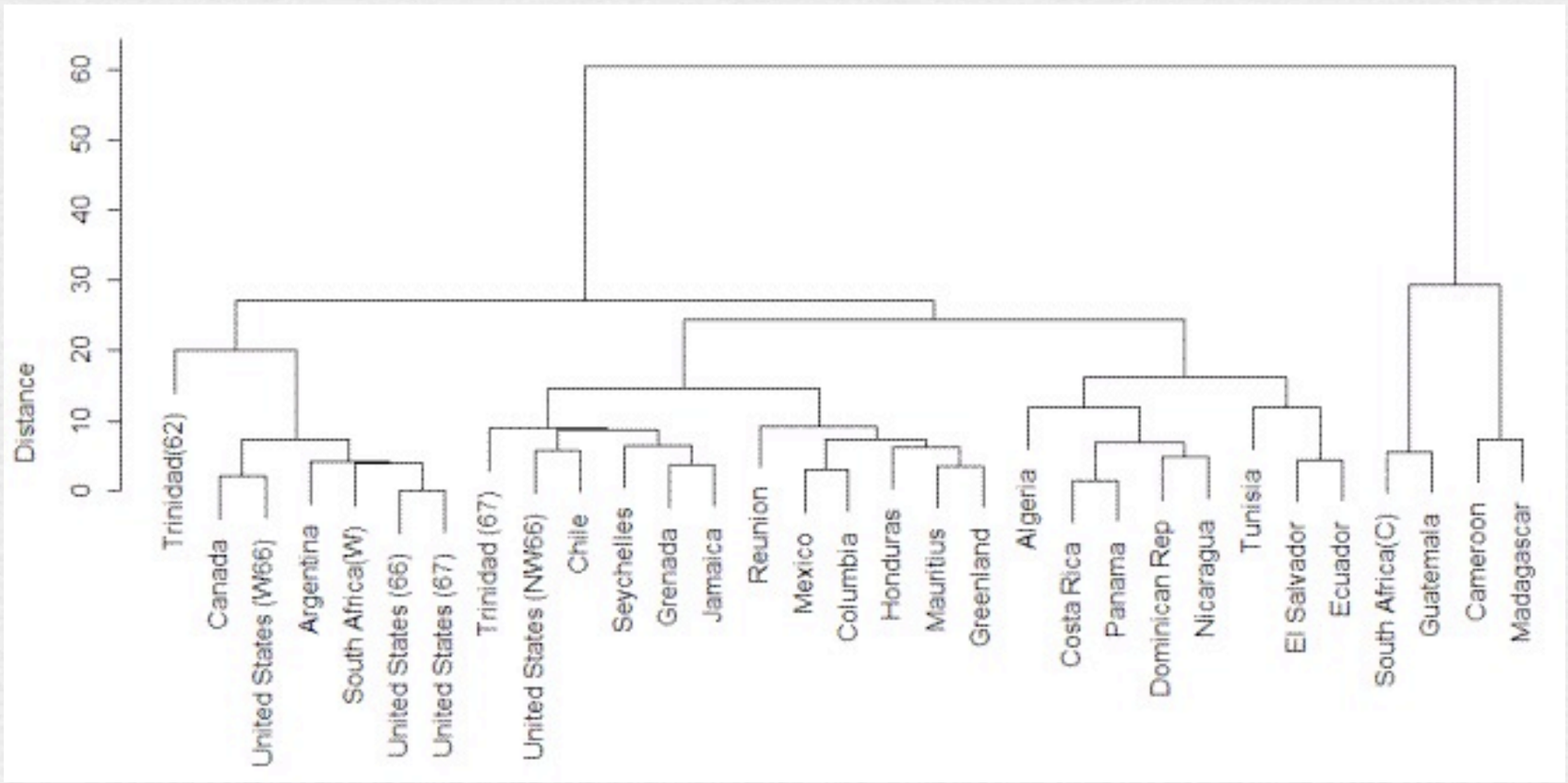
```
country <- row.names(life)

plclust( hclust(dist(life), method="single"), labels=country)
```

```
> source("c:\\mvar\\Data\\chap4lifeexp.dat")
> life
                  m0 m25 m50 m75 w0 w25 w50 w75
Algeria           63  51  30  13 67  54  34  15
Cameroon          34  29  13   5 38  32  17   6
Madagascar        38  30  17   7 38  34  20   7
Mauritius         59  42  20   6 64  46  25   8
Reunion           56  38  18   7 62  46  25  10
Seychelles        62  44  24   7 69  50  28  14
South Africa(C)   50  39  20   7 55  43  23   8
South Africa(W)   65  44  22   7 72  50  27   9
Tunisia           56  46  24  11 63  54  33  19
Canada            69  47  24   8 75  53  29  10
Costa Rica        65  48  26   9 68  50  27  10
```

single



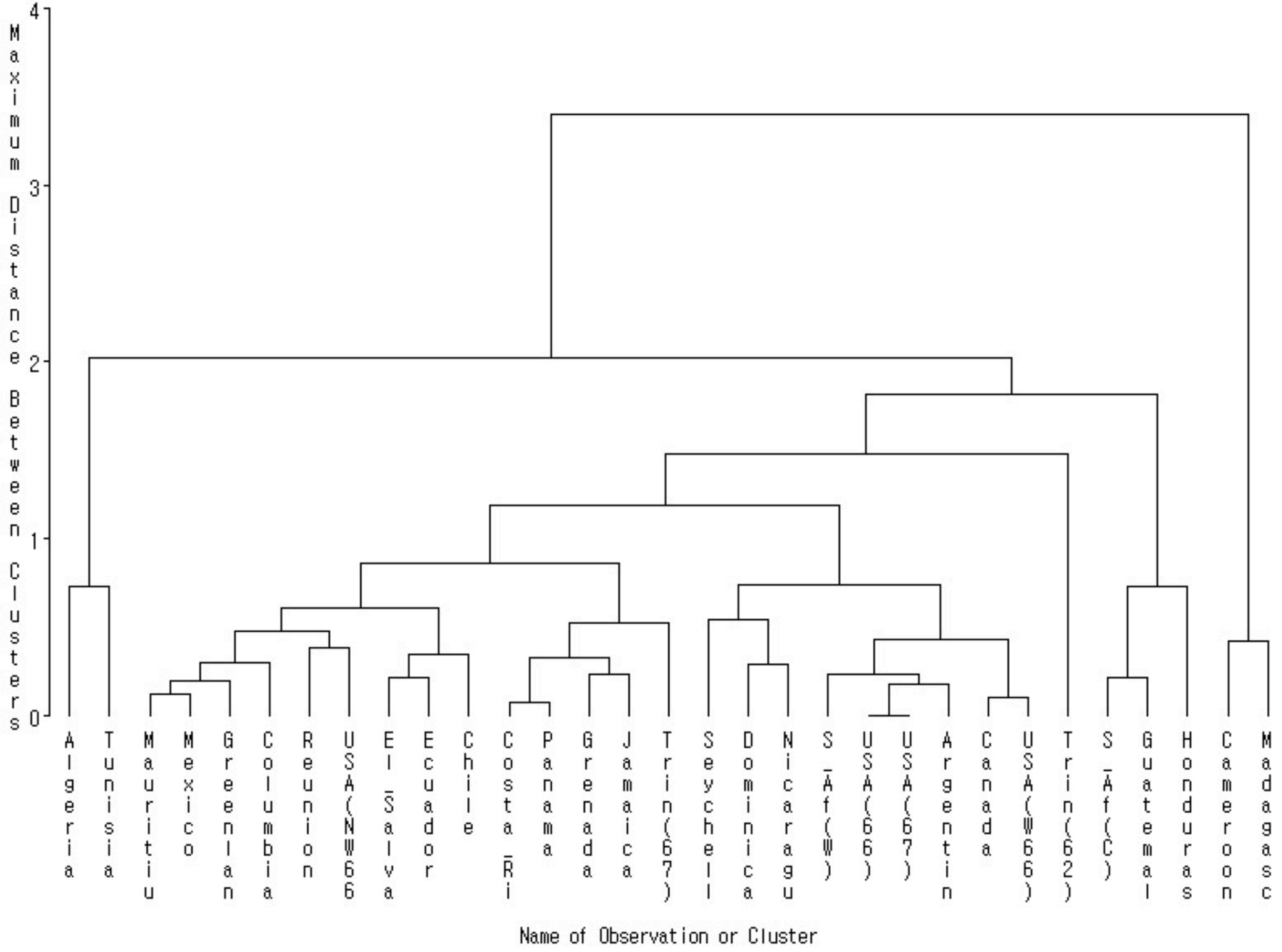
complete



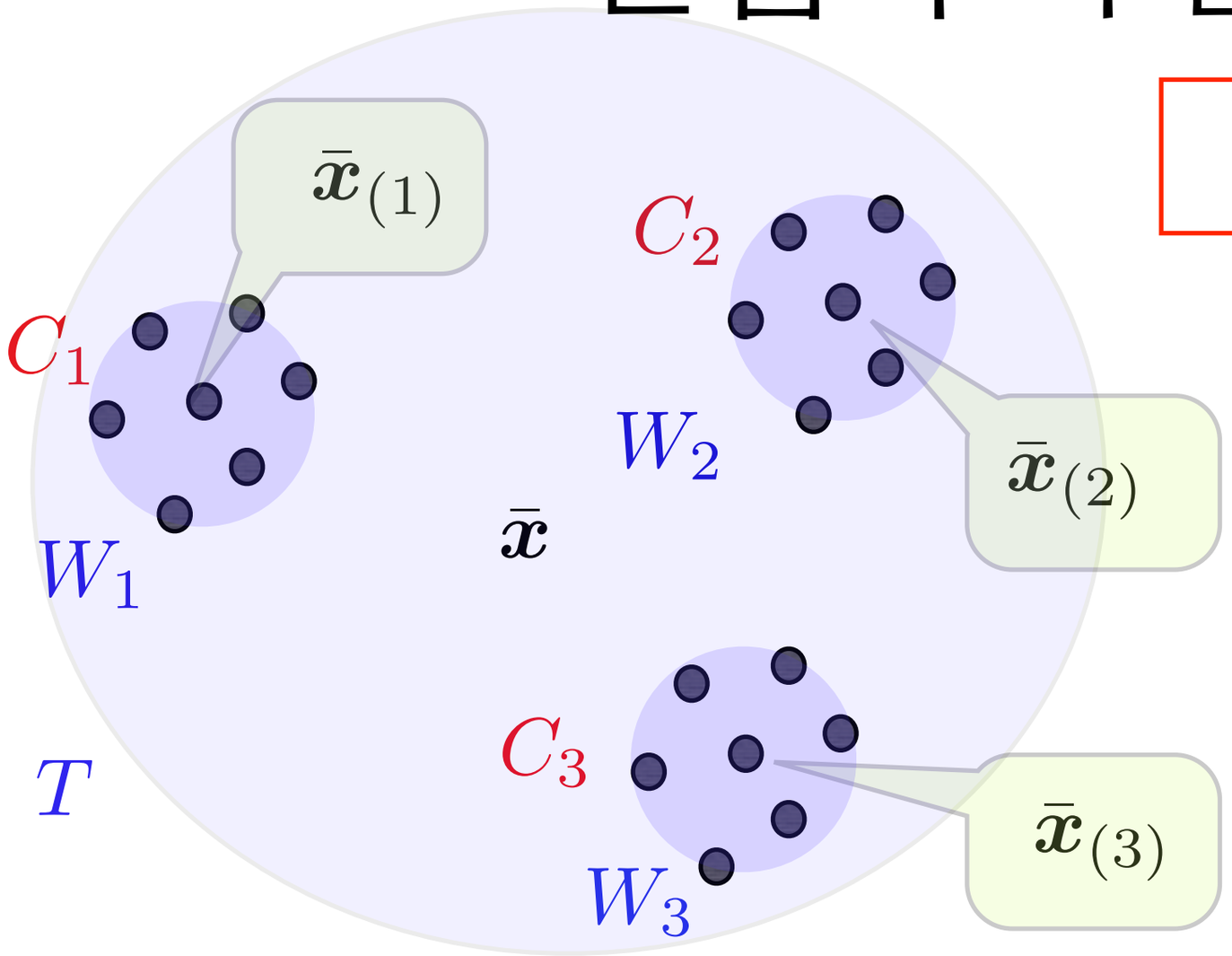
proc cluster 를 이용한 군집분석

```
proc cluster data=a
                    method=                    outtree=outa;
    id country;
    var m0 m25 … w75;
run;
proc tree data=outa;
run;
```

cluster 의 수 ?


Maximum Distance Between Clusters
4
3
2
1
0
Algeria
Tunisia
Mauritiu
Mexico
Greenlan
Columbia
Reunion
USA(NW66
El_Salva
Ecuador
Chile
Costa_Ri
Panama
Grenada
Jamaica
Trin(67)
Seychell
Dominica
Nicaragu
S_Af(W)
USA(66)
USA(67)
Argentin
Canada
USA(W66)
Trin(62)
S_Af(C)
Guatemal
Honduras
Cameroon
Madagasc
Name of Observation or Cluster

군집 수의 결정



$$\boldsymbol{x}_i = (x_{i1}, x_{i2}, \cdots, x_{ip})$$

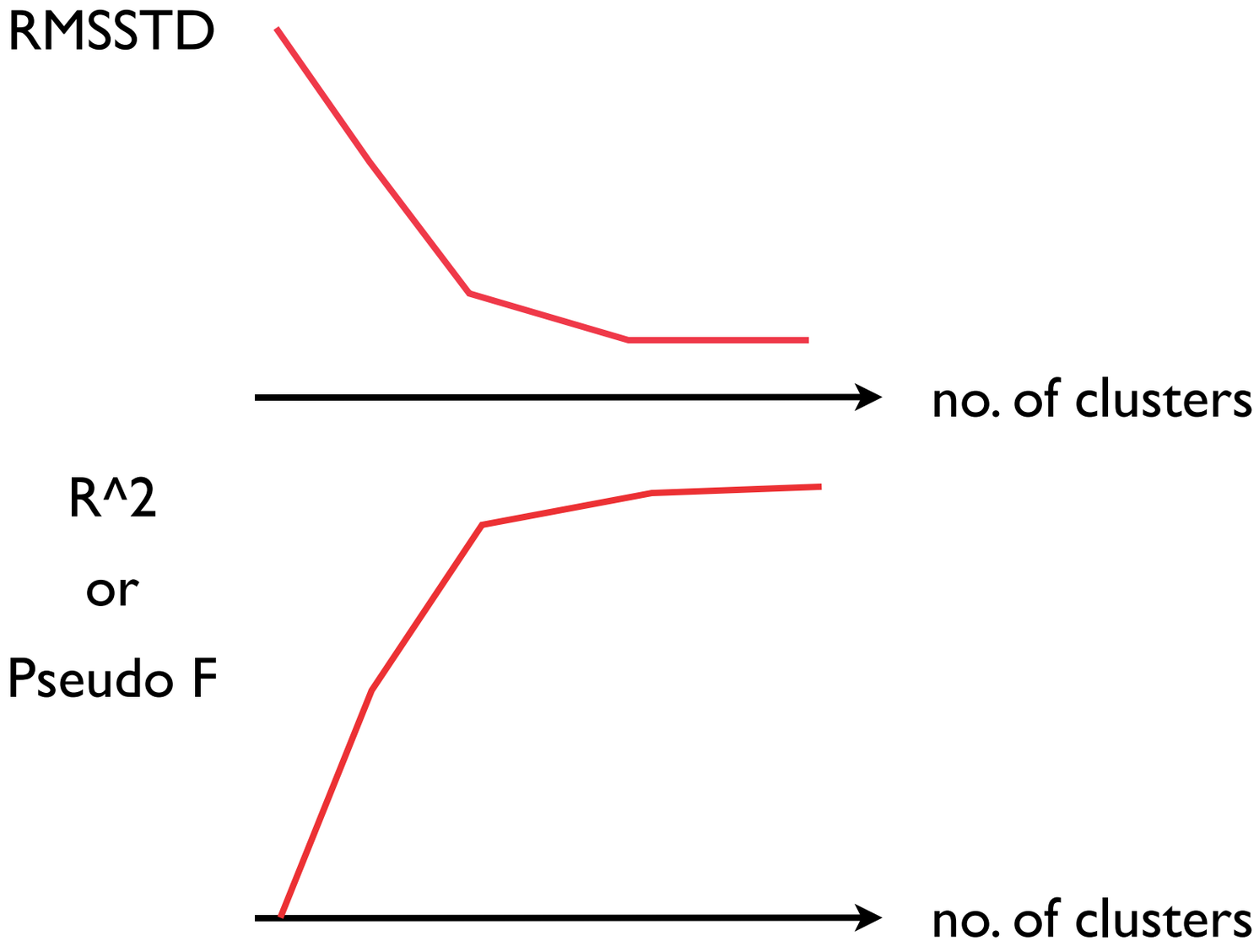
$T =$

$W_k =$

$P_G =$

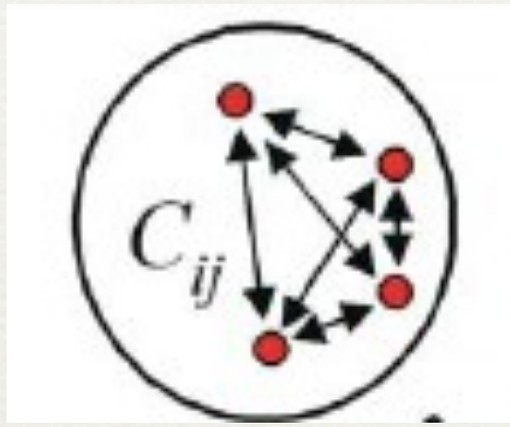
$RMSSTD =$

$$\text{Pseudo } F = \frac{\frac{T - P_G}{G - 1}}{\frac{P_G}{n - G}}$$



비계층적 군집분석
(Non-hierarchical Cluster Analysis)

- Data: n개 자료 k개 변수 (x_ij)
- K-Means 방법 사용 (k=군집의 수)
- x : 연속형, 표준화



(1) Correlation distance $C_{ij} = d_{ij} = 1 - s(\boldsymbol{x}_i, \boldsymbol{x}_j)$

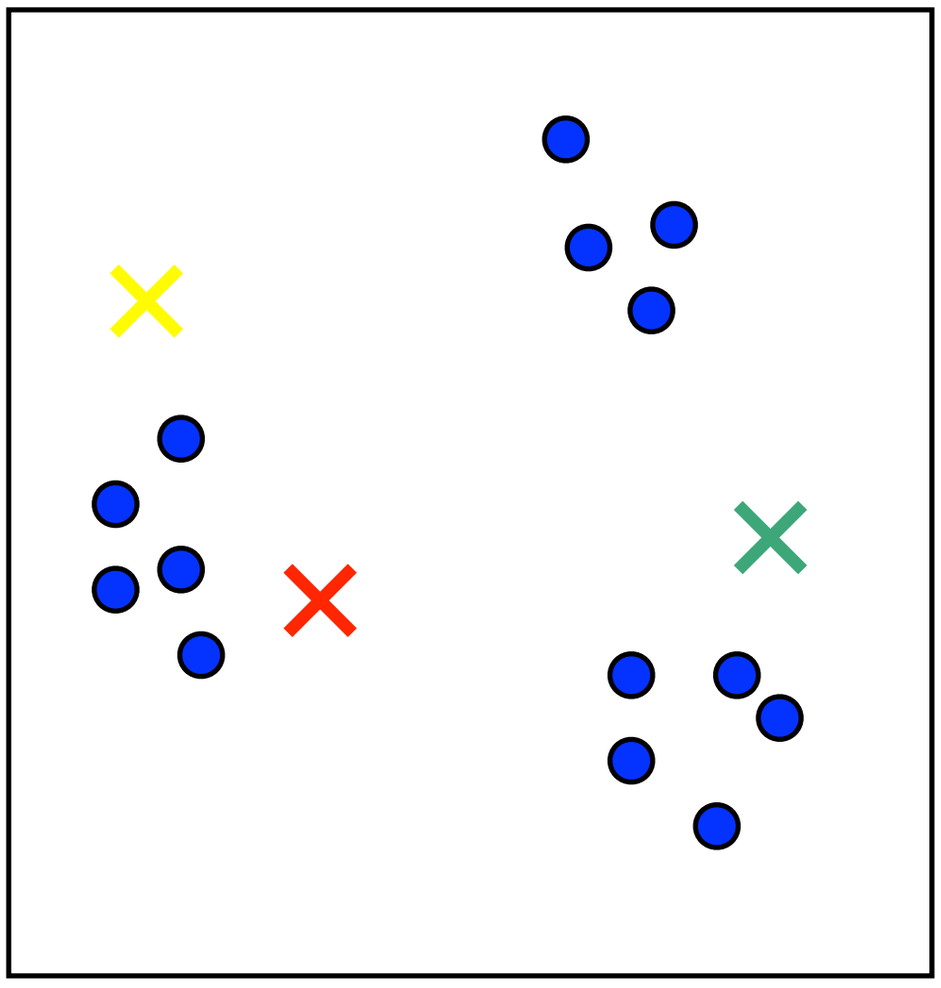
$$s(x_1, x_2) = \frac{\sum_{k=1}^{K}(x_{1k} - \bar{x}_1)(x_{2k} - \bar{x}_2)}{\sqrt{\sum_{k=1}^{K}(x_{1k} - \bar{x}_1)^2 \sum_{k=1}^{K}(x_{2k} - \bar{x}_2)^2}}$$

Correlation coef.

(2) Eucledian distance

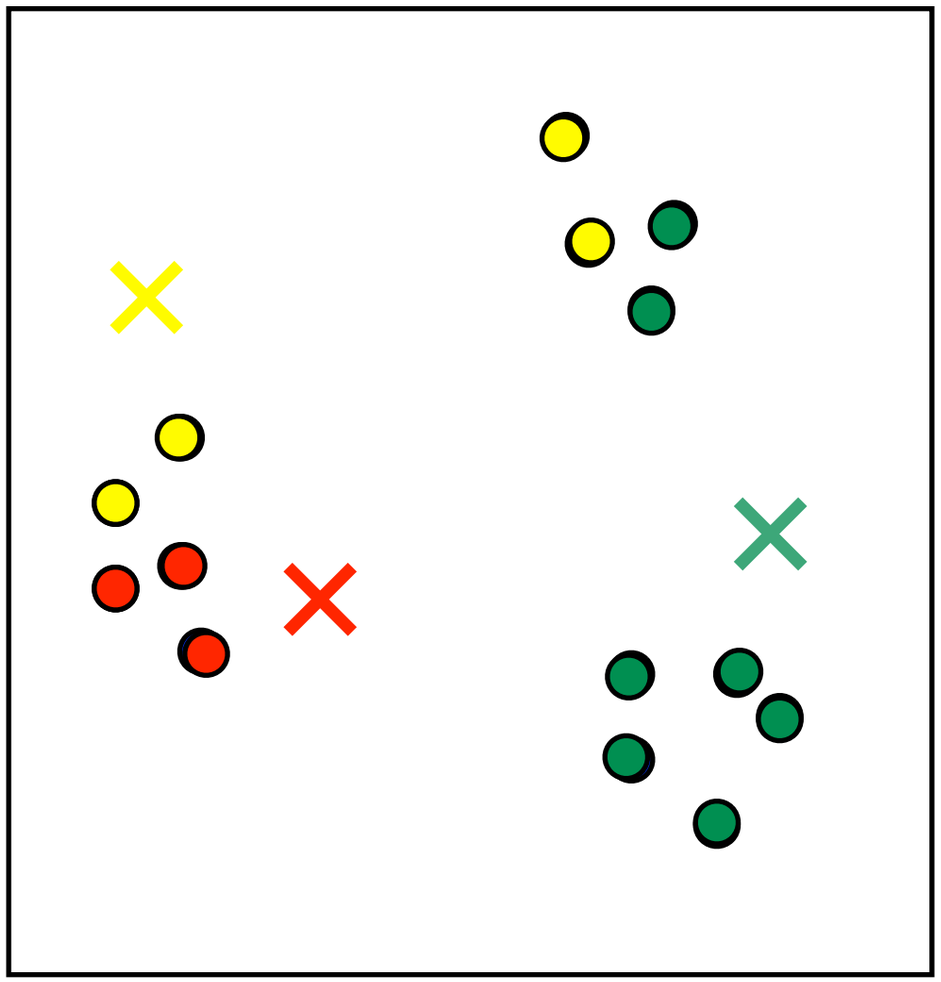
$$C_{ij} = d_{ij} = \sqrt{\sum_{l=1}^{k}(x_{il} - x_{jl})^2}$$

K-means (k=3점의 cluster 임의로 선택)



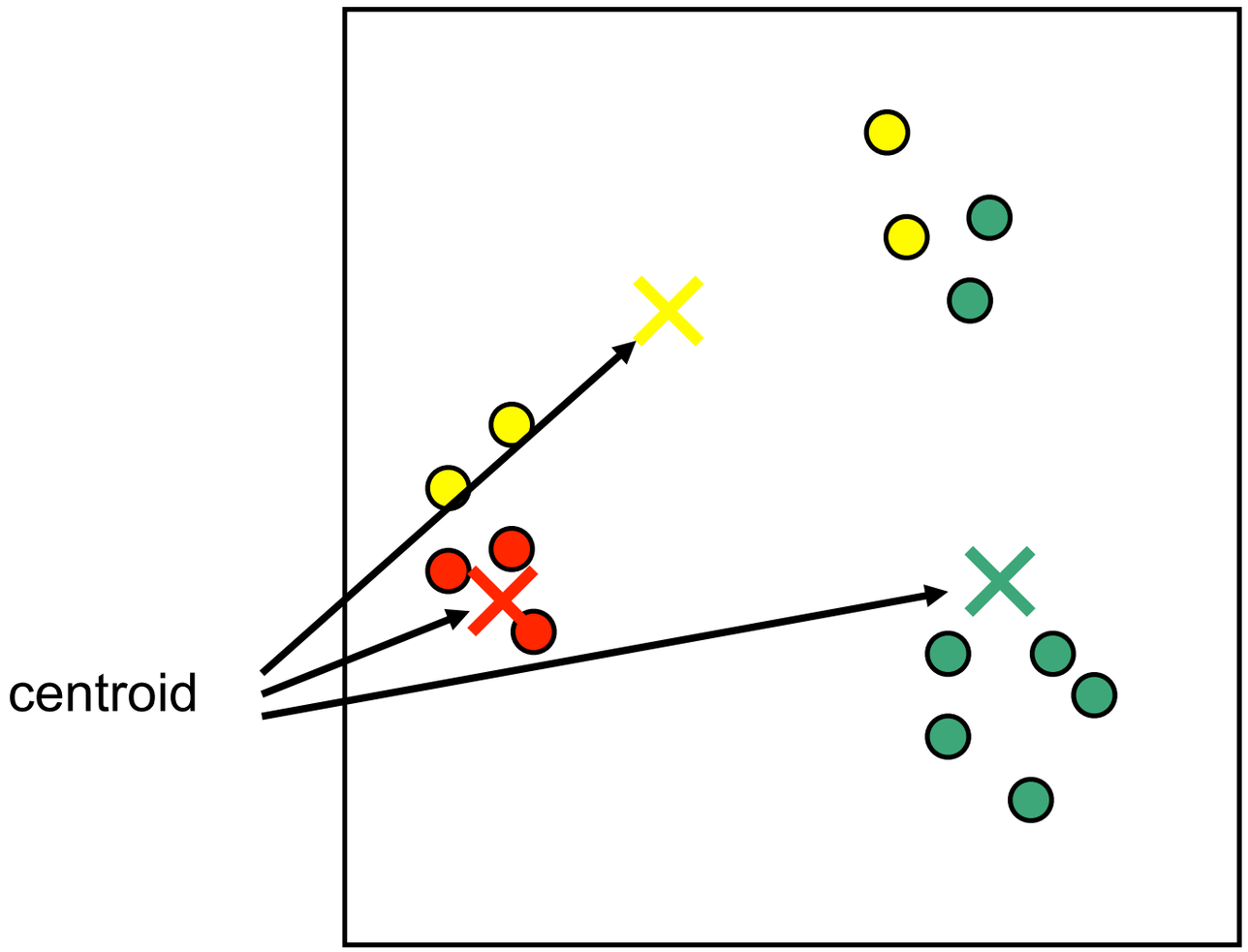
Iteration = 0

K-means (가장 가까운 **cluster** 찾음)



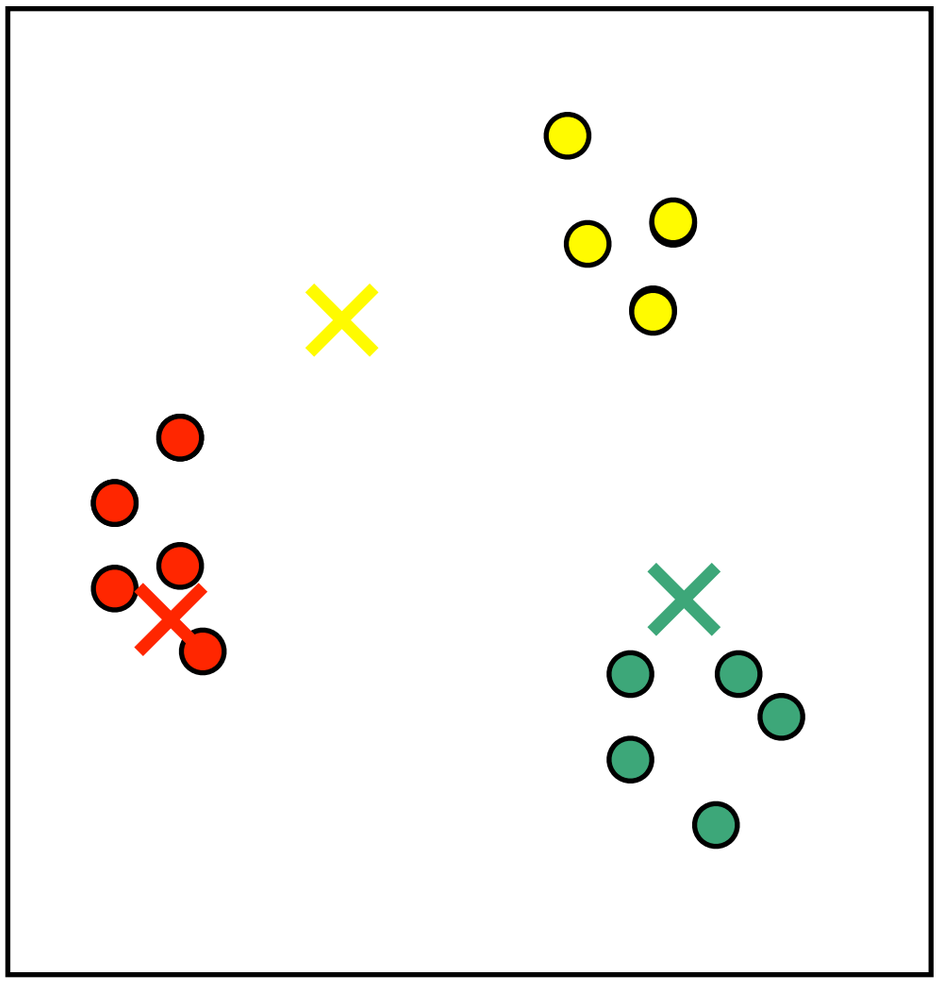
Iteration = 1

K-means (새로운 **centroid** 계산함)



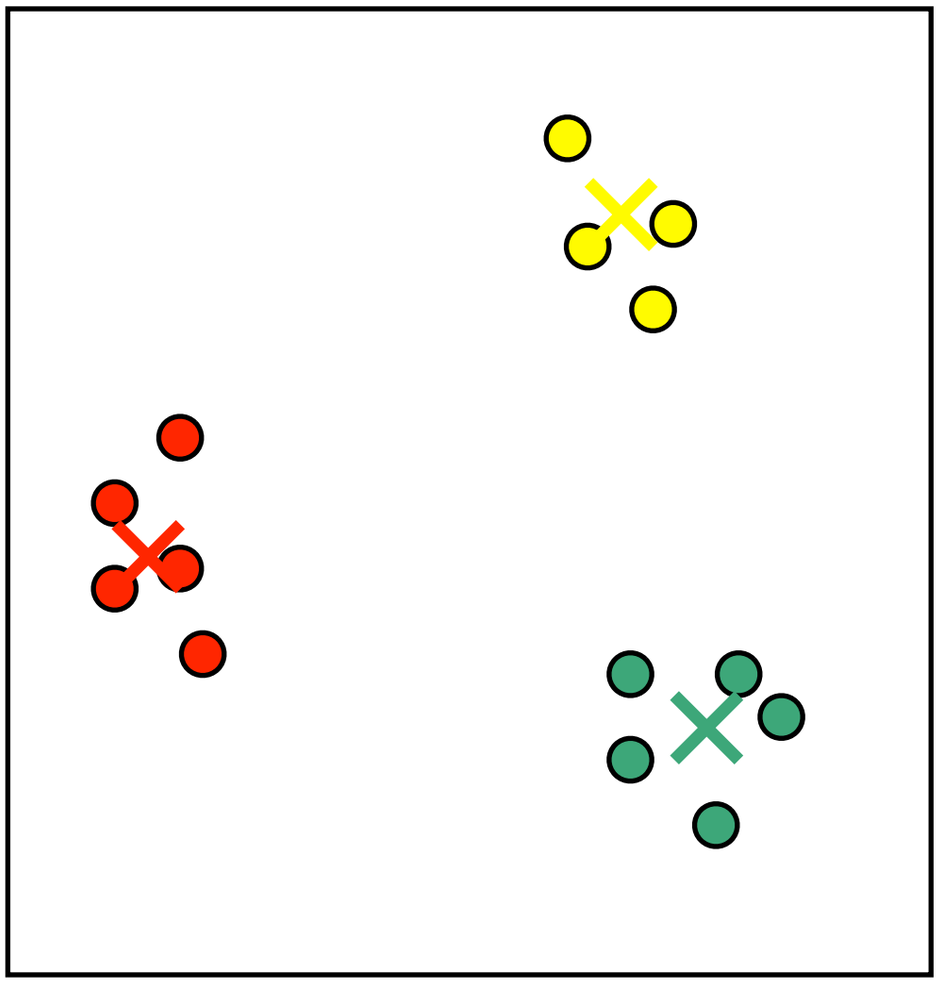
Iteration = 2

K-means (새로운 **centroid** 계산함)



Iteration = 2

K-means (가까운 **cluster** 다시 찾음)



Iteration = 3

Proc Fastclus (SAS)

- 자료간의 거리를 계산하여 군집화
- 관측자료는 적어도 한개의 군집에 속함
- **standardized** (반드시).
- TREE 를 생성하지 않음
- a excellent tool for segmentation of a population
- good for detecting “outliers”
- Euclidean distance (default)

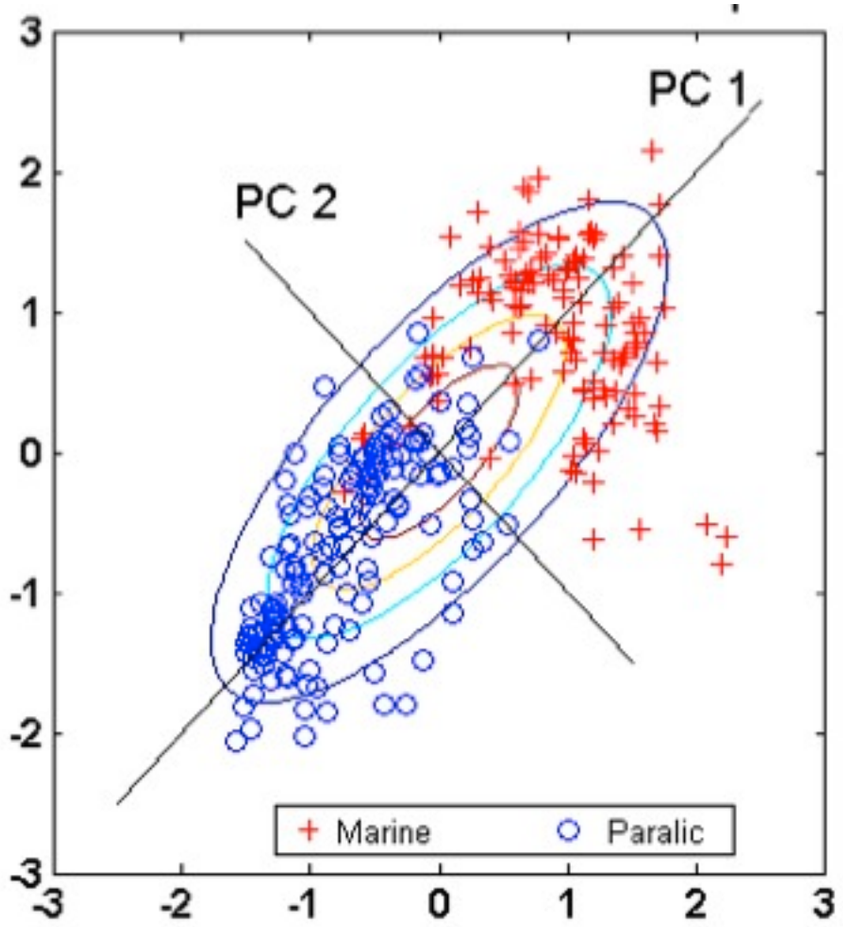
```
proc standard data=a out=b mean=0 std=1  replace;
        var height weight chest;
 run;

proc fastclus data=b out=outb;
        maxclusters=<..>   maxiter=<..>;
        var height weight chest;
run;

proc freq data=outb;
        tables gender*cluster;
run;
```

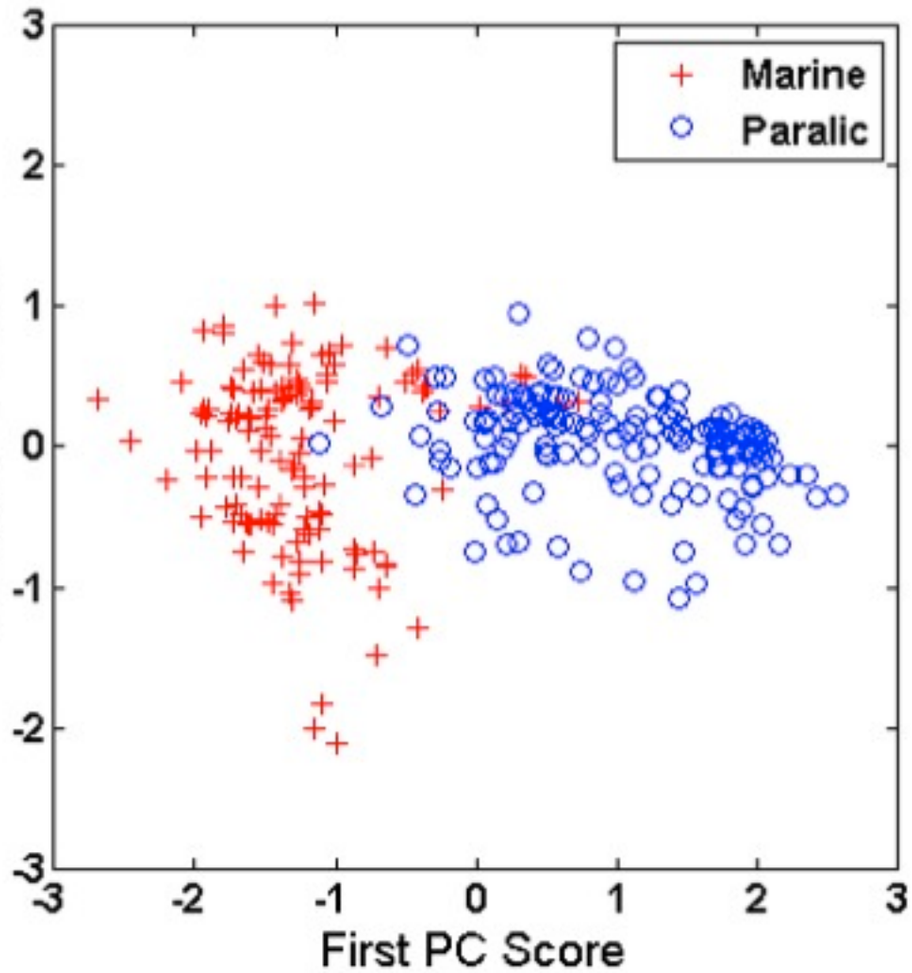
정준변수 (cannonical variable)

- 정준상관분석 :

- 군집분석 :

PCA & Groups

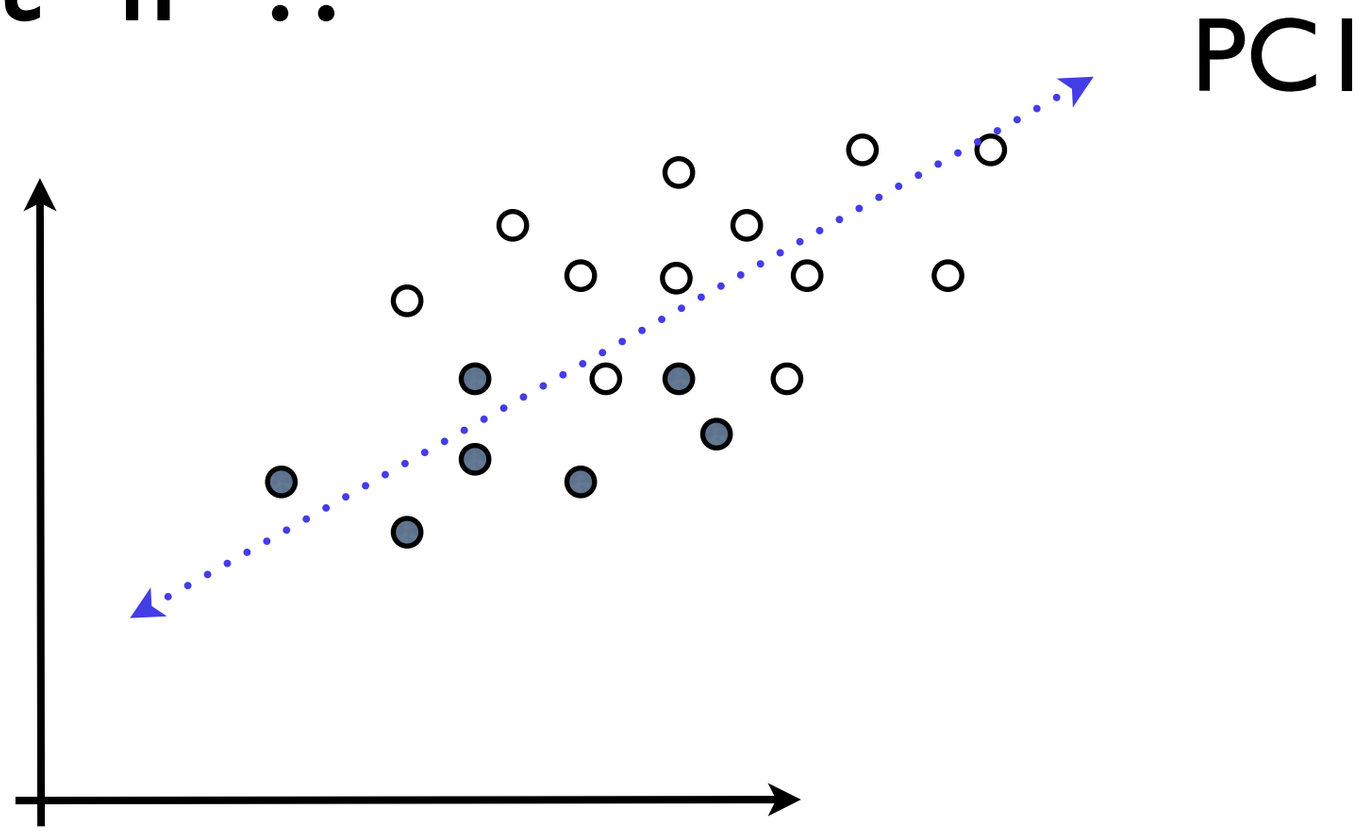


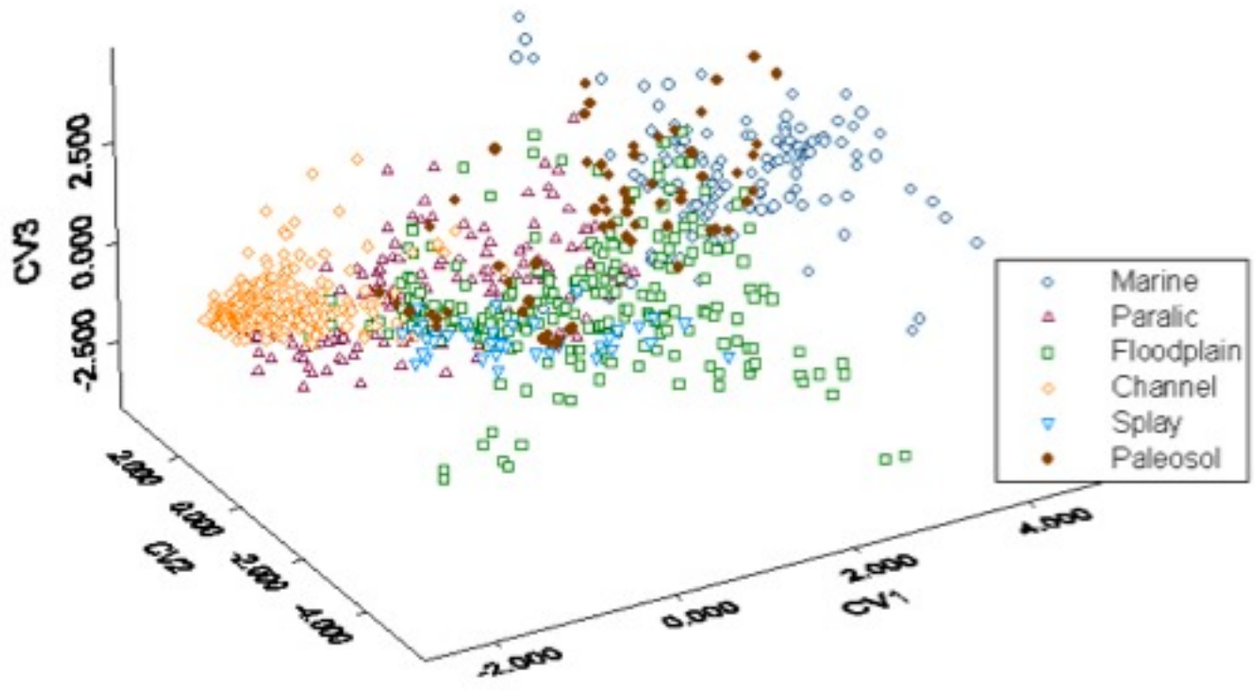


>> 변수 q개 대신 _____사용하여 그룹간 차이 규명
( )

>> _____ !!!!

What if ??


PC1


CV3
2.500
0.000
-2.500
2.000
0.000
-2.000
CV2
-4.000
-2.000
0.000
2.000
4.000
CV1
Marine
Paralic
Floodplain
Channel
Splay
Paleosol

How to get ??

The within-groups covariance matrix is just the pooled covariance matrix, **S**, that we discussed in the context of linear discriminant analysis, describing the average variation of each group about its respective group mean. The between-groups covariance matrix describes the variation of the group means about the global mean. The sample version of it is given by

$$\mathbf{B} = \frac{K}{n(K-1)} \sum_{k=1}^{K} n_k (\bar{\mathbf{x}}_k - \bar{\mathbf{x}})(\bar{\mathbf{x}}_k - \bar{\mathbf{x}})'$$

where K is the number of groups, n_k is the number of data in group k, n is the total number of data, $\bar{\mathbf{x}}_k$ is the mean vector for group k, and $\bar{\mathbf{x}}$ is the global mean vector.

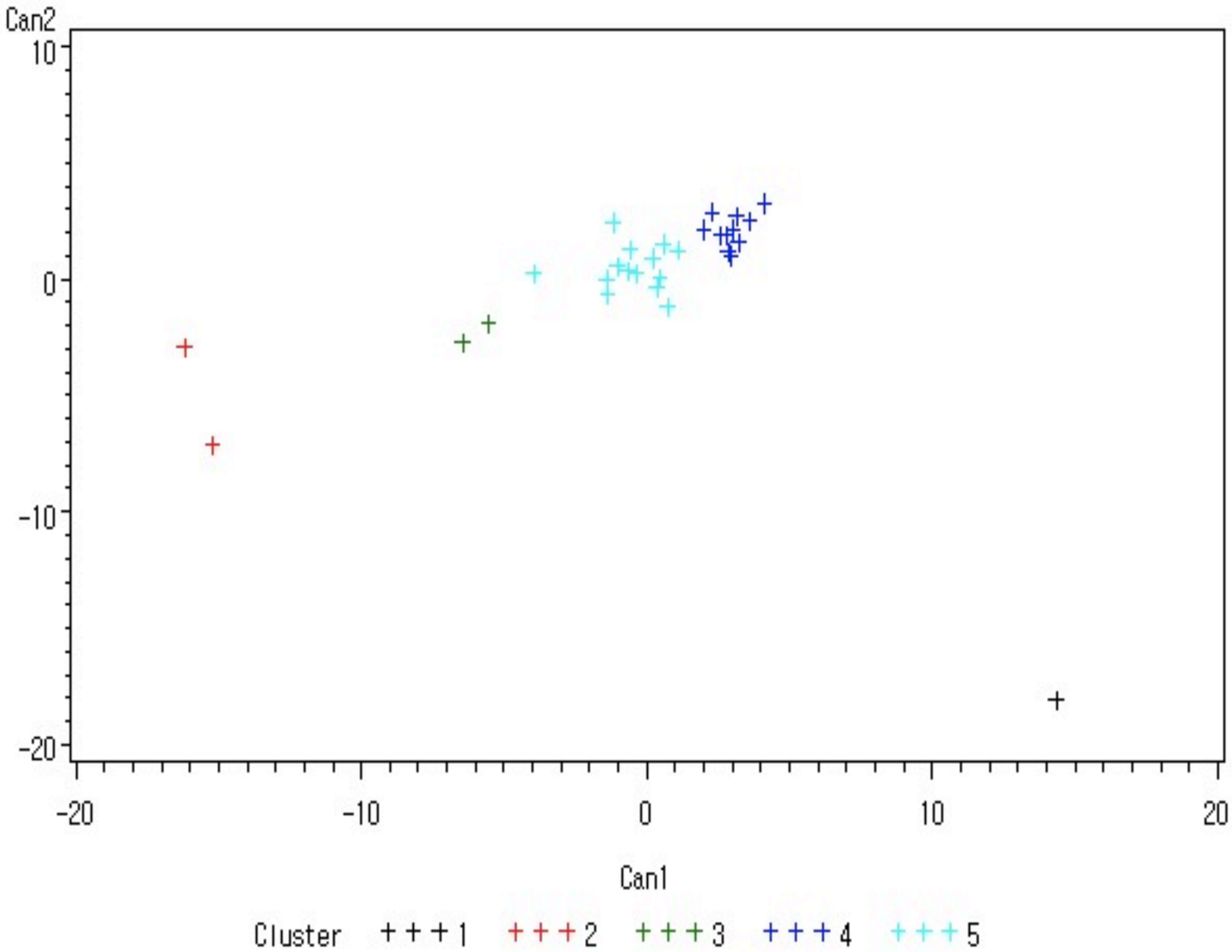
If $\mathbf{u}_1$ is the first eigenvector (associated with the largest eigenvalue) of $\mathbf{S}^{-1}\mathbf{B}$, then the first canonical variate, $\mathbf{v}_1 = \mathbf{u}_1' \mathbf{x}$, is the linear combination of variables that shows the maximum ratio of between-groups to within-groups variation. From there, results can be interpreted pretty much the same way as PCA, with "ratio of between- to within-groups" variation substituted for "total variation".

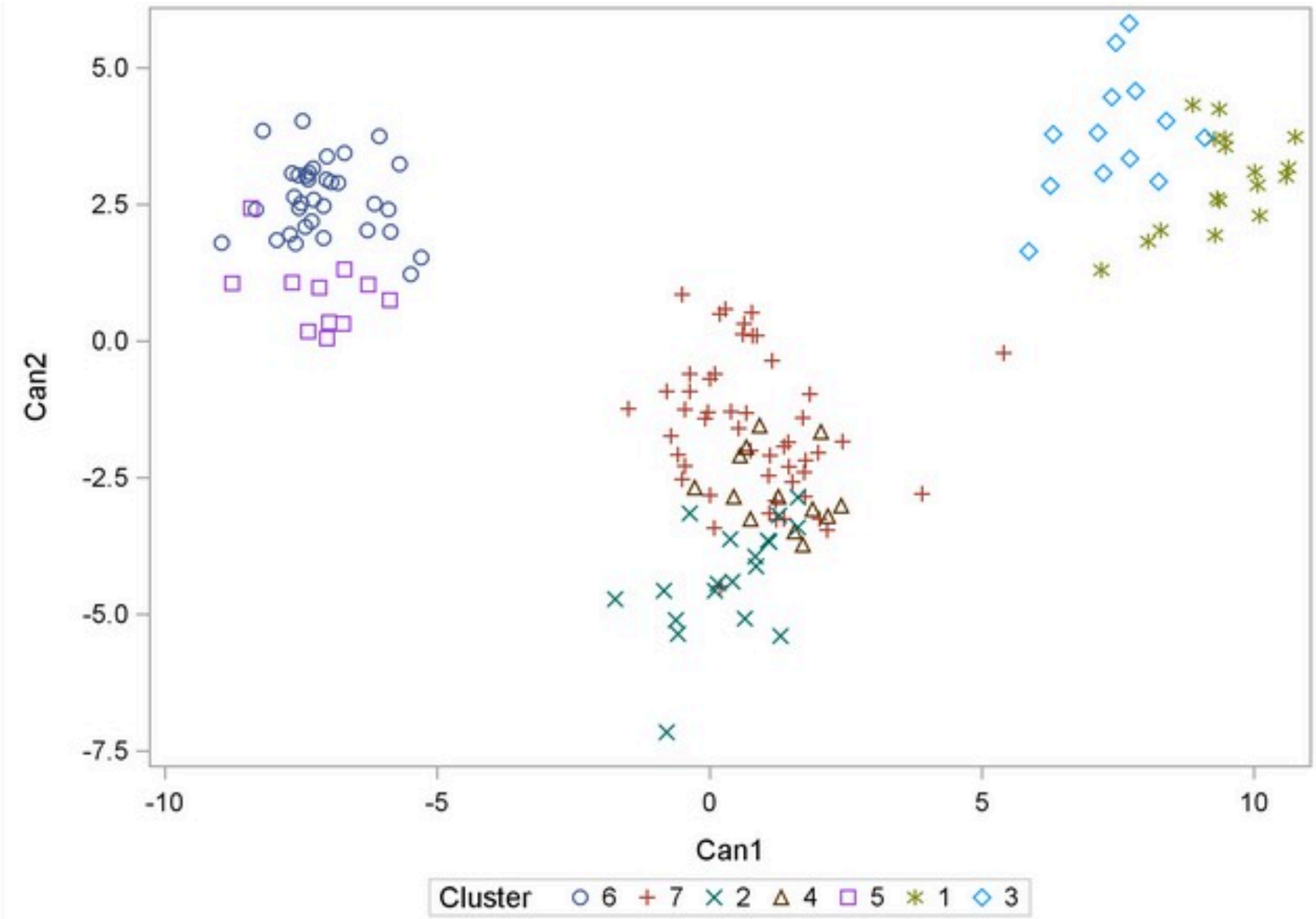
proc fastclus 이용한 군집분석

```
/* ==K-means clustering ============*/
proc fastclus data=a out=clust  maxclusters=5;
        var m0 m25 … w75;
```

```
proc candisc data=clust out=can;
 class cluster;
        var m0 m25 … w75;
```

```
proc gplot data=can;
        plot can2*can1=cluster;
 run;
```


Can2
10
0
-10
-20
-20
-10
0
10
20
Can1
Cluster +++1 +++2 +++3 +++4 +++5


5.0
2.5
0.0
-2.5
-5.0
-7.5
Can2
-10
-5
0
5
10
Can1
Cluster
6
7
2
4
5
1
3

군집분석의 활용

- VIP 고객들의 소비성향분석을 위한 고객세분화 (_____) 필요성 (마케팅전략)
- 결측치(_____)에 대한 보완책
- 특정영역의 여러 변수에 대한 군집화를 통해 변수의 개수 축소효과(_____)
- correlated 변수들은 _____