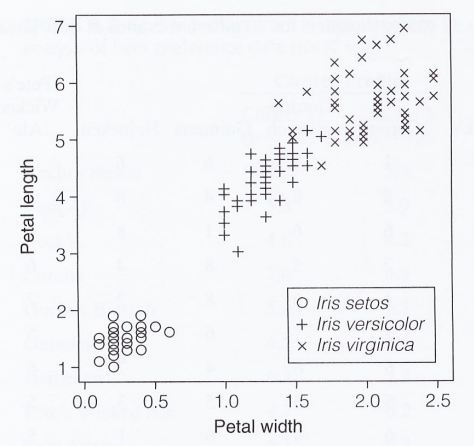
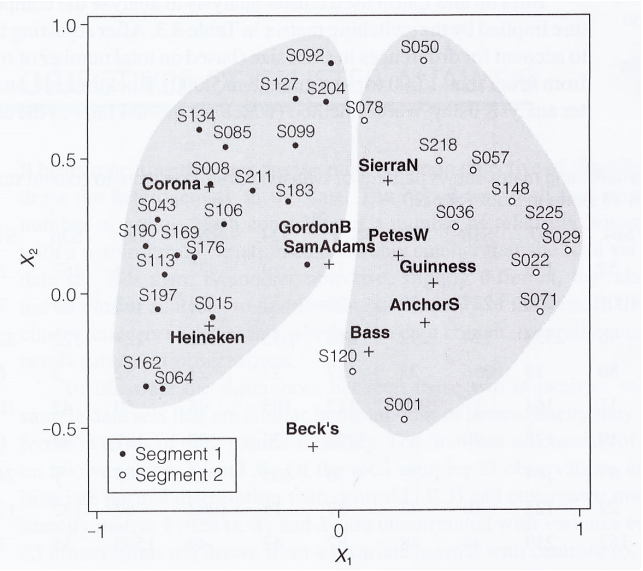
군집분석

Cluster Analysis

군집분석이란?

- 관측치들을 여러 개의 서로 배타적인 그룹으로 분류하여 서로 유사한 것을 같은 그룹으로 모이도록 하는 것
- 범주(그룹)에 대한 사전 정보가 없음
  - 그룹의 개수나 성질에 대한 사전정보가 있는 경우 판별분석(Classification) 사용
- 동일한 군집 내의 관찰치는 서로 비슷한 속성을 갖도록, 다른 군집의 관칠치는 서로 상이한 속성을 갖도록 군집 형성
- 군집분석의 활용
  - 생물의 종 구분
  - 시장세분화
    - 구매태도, 구매성향, 매체사용습관 등과 같은 특성에 따라 공통적인 특성을 공유하는 사람, 시장, 조직의 군집 발견
  - 잠재적인 신제품 기회 발견
    - 시장에서 경쟁관계에 있는 상품이나 기업의 속성에 따라 군집화





유사성 측도

- 유클리드 거리(Euclidean Distance)

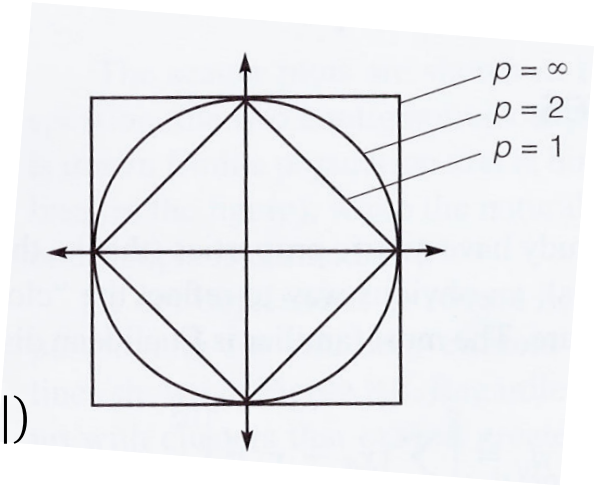
$$d_{ij} = \sqrt{\sum_k \left(x_{ik} - x_{jk}\right)^2}$$

  - 변수 들의 서로 다른 단위를 가진 경우가 많으므로 주로 표준화된 자료에 사용

- 민코우스키 거리(Minkowski p-Metric)

$$d_{ij}(p) = \left[\sum_k \left|x_{ik} - x_{jk}\right|^p\right]^{1/p}$$

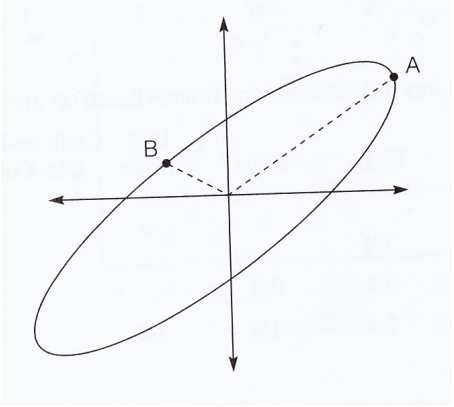
  - p=2: 유클리드 거리
  - p=1: City-block metric (manhattan distance) $d_{ij}(1) = \sum_k \left|x_{ik} - x_{jk}\right|$
  - p= ∞ : sup metric $d_{ij}(\infty) = \max\left(\left|x_{i1} - x_{j1}\right|, \left|x_{i2} - x_{j2}\right|, \ldots, \left|x_{ip} - x_{jp}\right|\right)$



- 마할라노비스 거리 (Mahalanobis Distance)

$$D_{ij} = (\boldsymbol{x_i} - \boldsymbol{x_j})' \Sigma^{-1} (\boldsymbol{x_i} - \boldsymbol{x_j})$$

  - 공분산 행렬을 고려한 거리측도
  - 유클리드 거리 측도에 의하면 B가 A에 비해 원점에 가깝지만 마할라노비스 거리 측도에 의하면 거리가 같음
  - Rotation과 Scale에 의해 자료를 변환하여 공분산행렬을 I로 만드는 효과




```
d=dist(x,method="euclidean")
```

- "dist" function
  - method="euclidean": 유클리드 거
  - method="manhattan": 맨하탄 거리
  - method="minkowski": 민코우스키 거리
- "mahalanobis" function
  - 마할라노비스 거리

계층적 군집분석(Hierarchical Clustering, Agglomerative Clustering)

- 관측치 간의 유사성을 계산해 가까운 개체들을 군집화
- 한번 한 군집에 소속되면 이동 불가능
- 덴드로그램(Dendrogram)을 사용해 군집형성 과정 파악 가능
- 알고리즘

step 0. 각 관측치가 하나의 군집을 형성하는 상태에서 시작($C_1, \dots, C_n$: n개의 군집)

두 군집 사이의 거리 계산 $d_{C_i C_j} = d_{ij}$

step 1. 가장 거리가 가까운 두 군집을 찾아 합쳐 새로운 군집 생성

step 2. 새 군집과 기존 군집들 간의 거리 계산

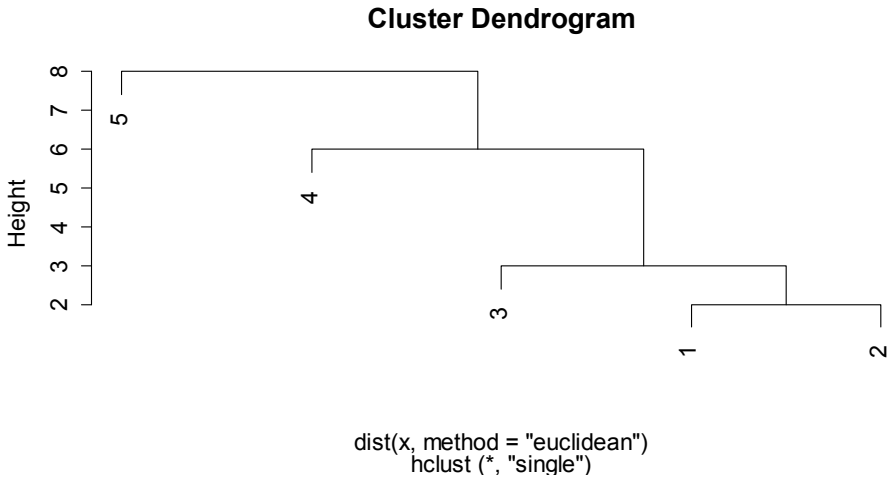
step 3. 하나의 군집이 남을 때 까지 step 1 & 2 반복

- **<u>군집의 개수 결정</u>**
  - 군집 간의 거리 차이에 큰 변화를 보이는 경우를 고려해 군집 개수 결정
  - Ward's method: ESS의 증가가 급격한 위치에서 군집의 개수 결정

계층적 군집분석: 최단연결법 (Single Linkage)

- $d_{C_i C_j} = \min\{d(x,y) | x \in C_i, y \in C_j\}$

```
> x=c(1,3,6,12,20)
> dist(x,method="euclidean")
   1  2  3  4
2  2
3  5  3
4 11  9  6
5 19 17 14  8
> hc1=hclust(dist(x,method="euclidean"),method="single") # single linkage
> plot(hc1)
```



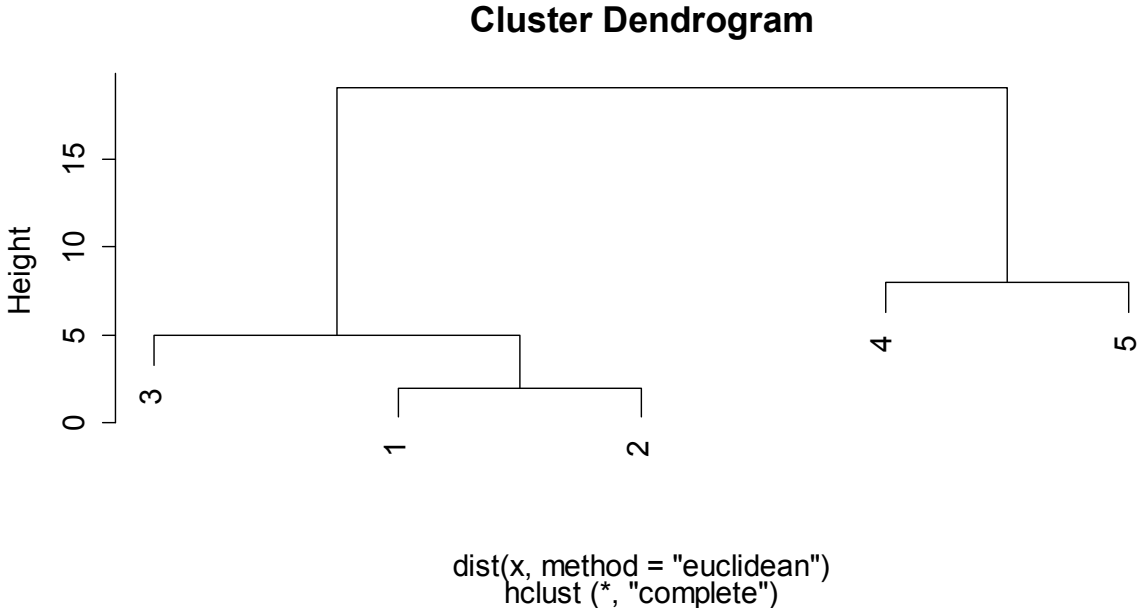
- 계산이 효율적
- 근시안적: 길다란 형태의 군집 형성 가능

계층적 군집분석: 최장연결법 (Complete Linkage)

- $d_{C_i C_j} = \max\{d(x,y) | x \in C_i, y \in C_j\}$

```
> hc2=hclust(dist(x,method="euclidean"),method="complete") # complete linkage
> plot(hc2)
```

- 군집과 군집을 합할 때 군집의 모든 개체가 서로 가까Single linkage에 비해 convex한 군집 형성
- 이상치에 민감

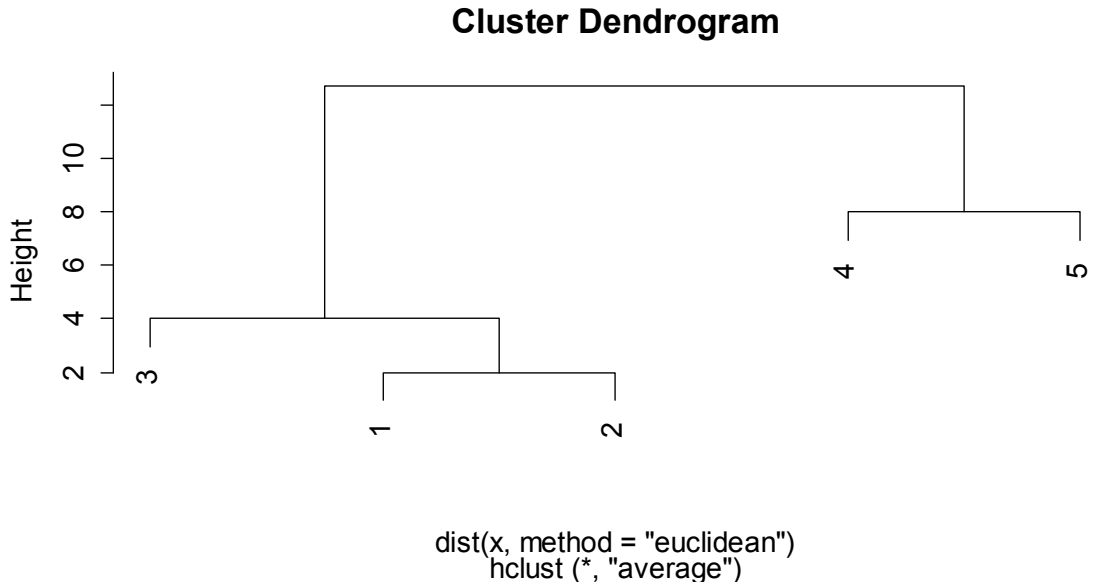


계층적 군집분석: 평균연결법 (Average Linkage)

- $d_{C_1C_2} = \frac{1}{n_1 n_2}\sum_i \sum_j d_{ij} \max\{d(x,y)|x \in C_i, y \in C_j\}$

```
> hc3=hclust(dist(x,method="euclidean"),method="average") # average linkage
> plot(hc3)
```

- single linkage와 complete linkage의 중간



계층적 군집분석: Ward's Method

- 군집 간 정보의 손실을 최소화 하는 군집화
- Minimum variance method
- 클러스터 C의 군집내거리(within-cluster distance)

$$ESS_C = \sum_j (x_{Cj} - \overline{x_C})'(x_{Cj} - \overline{x_C})$$

- 군집내거리를 최소화 하는 군집화 찾음
- 비슷한 크기의 군집을 생성하는 경향
- 비계층적 군집화 방법과 비슷한 결과를 생성하는 경향

계층적 군집분석: 군집 개수의 결정

- 덴드로그램을 어디서 자를 것인가? 데이터에 최적인 특정한 그룹의 개수를 어떻게 정할 것인가?

- 덴드로그램 높이 변화의 크기를 조사하여 큰 변화가 있는 곳에서 자름
- 적절한 군집의 개수가 알려져 있을 경우 해당 군집 개수를 얻는 위치에서 자름

- cutree 함수 사용하여 군집분석의 해 얻음

Example: USArrests

- This data set contains statistics, in arrests per 100,000 residents for assault, murder, and rape in each of the 50 US states in 1973. Also given is the percent of the population living in urban areas.
- A data frame with 50 observations on 4 variables.

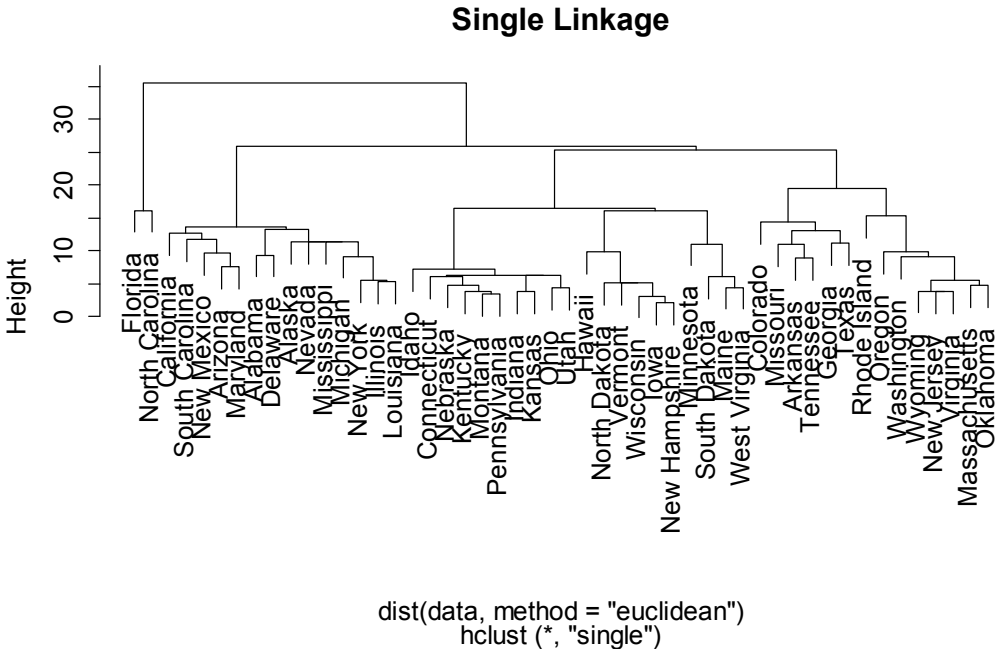
[,1] Murder numeric Murder arrests (per 100,000)

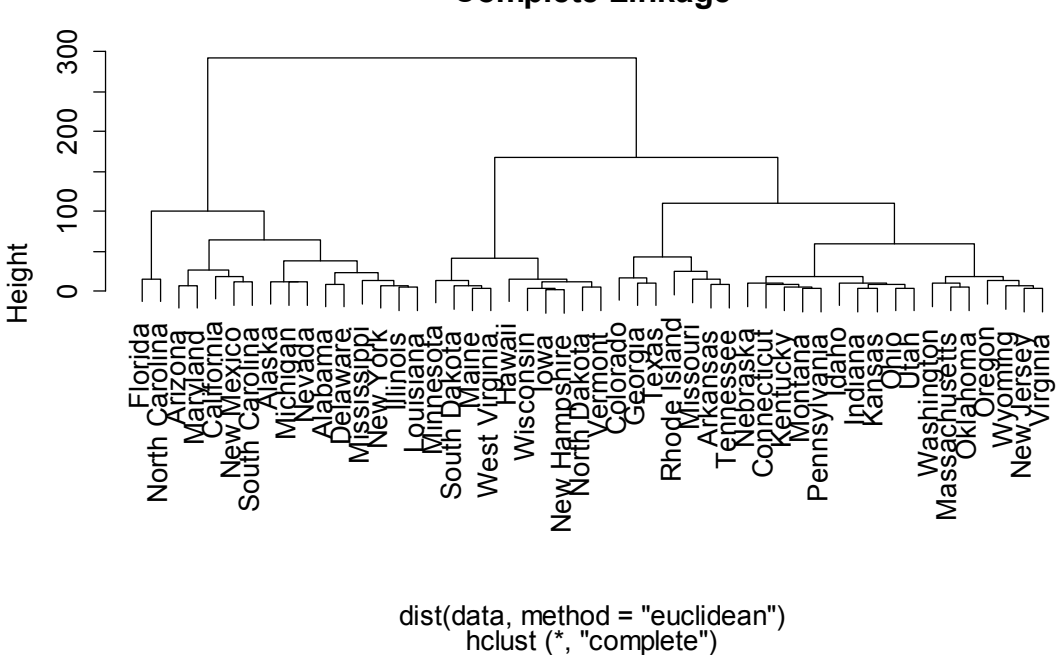
[,2] Assault numeric Assault arrests (per 100,000)

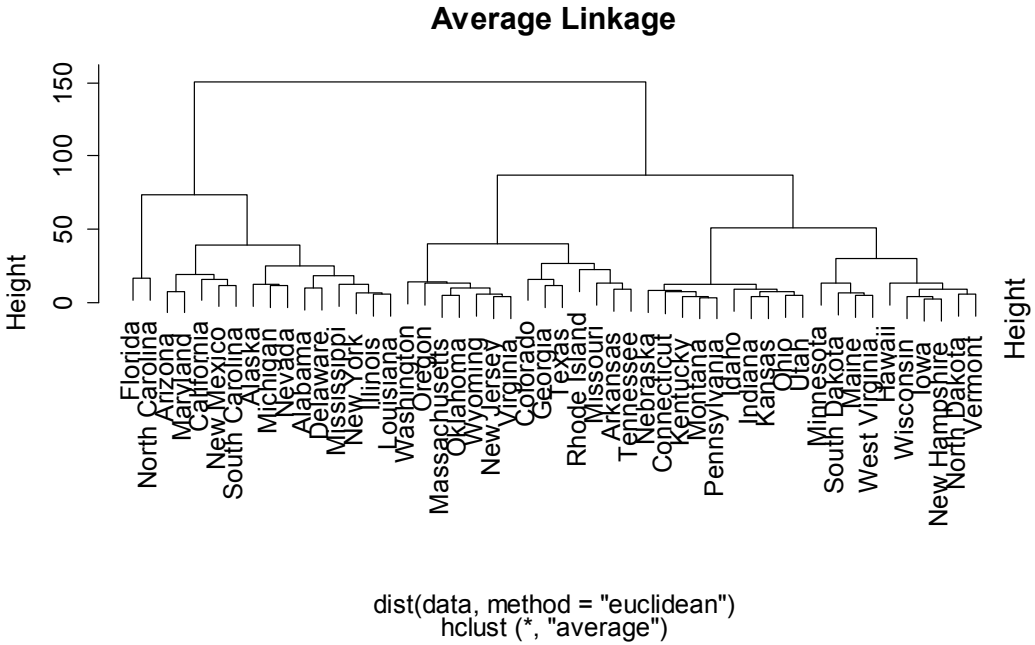
[,3] UrbanPop numeric Percent urban population

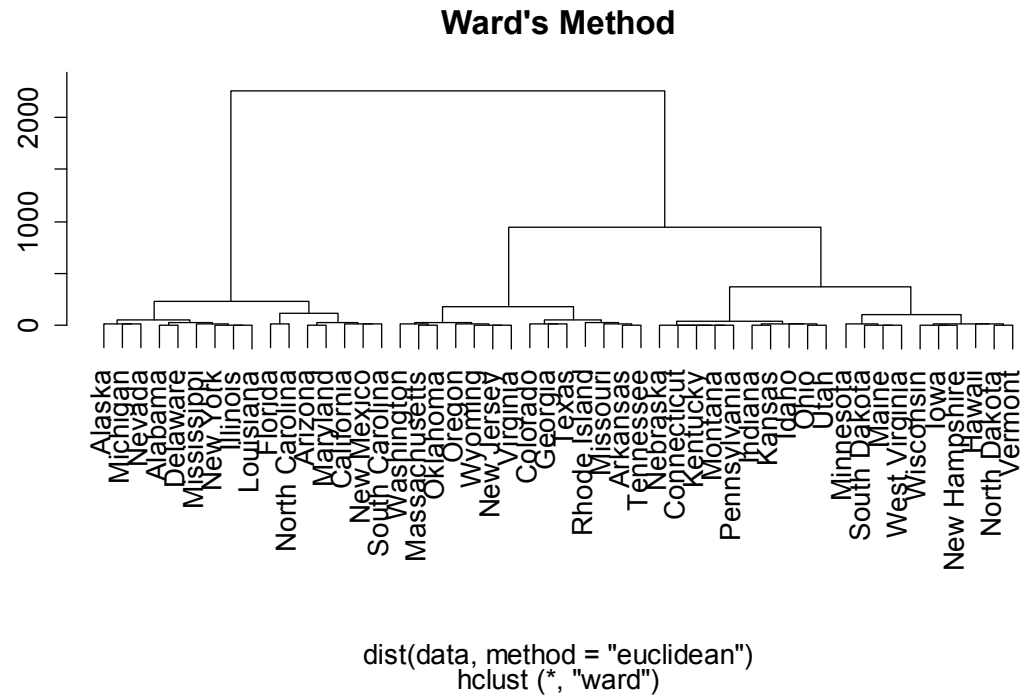
[,4] Rape numeric Rape arrests (per 100,000)

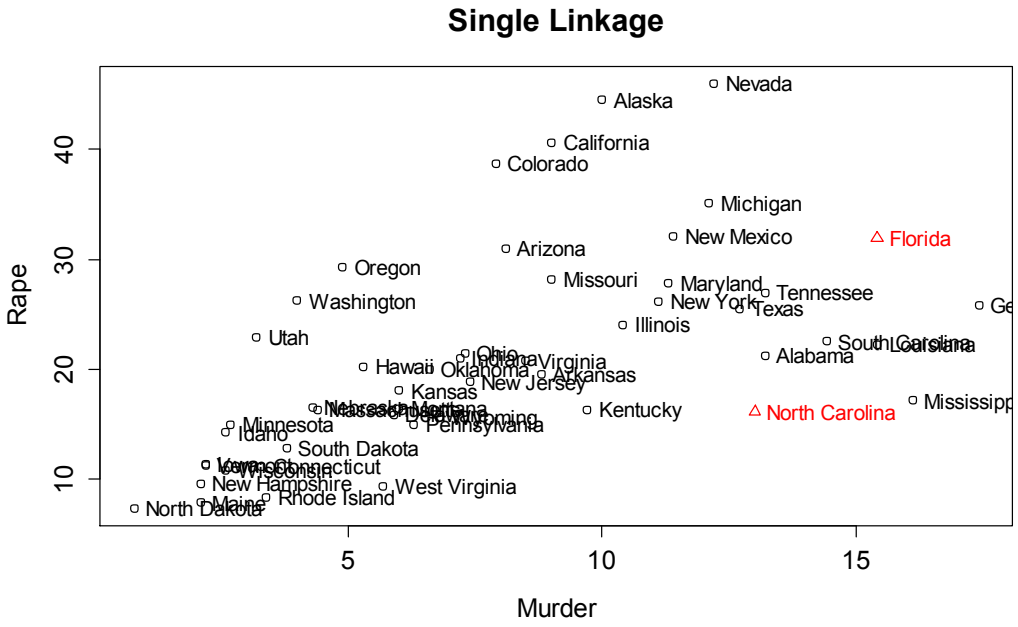
- 전 미 50개 주의 1973년 인구 10만명 당 체포범죄건수
- 변수: Murder, Assault, Rape

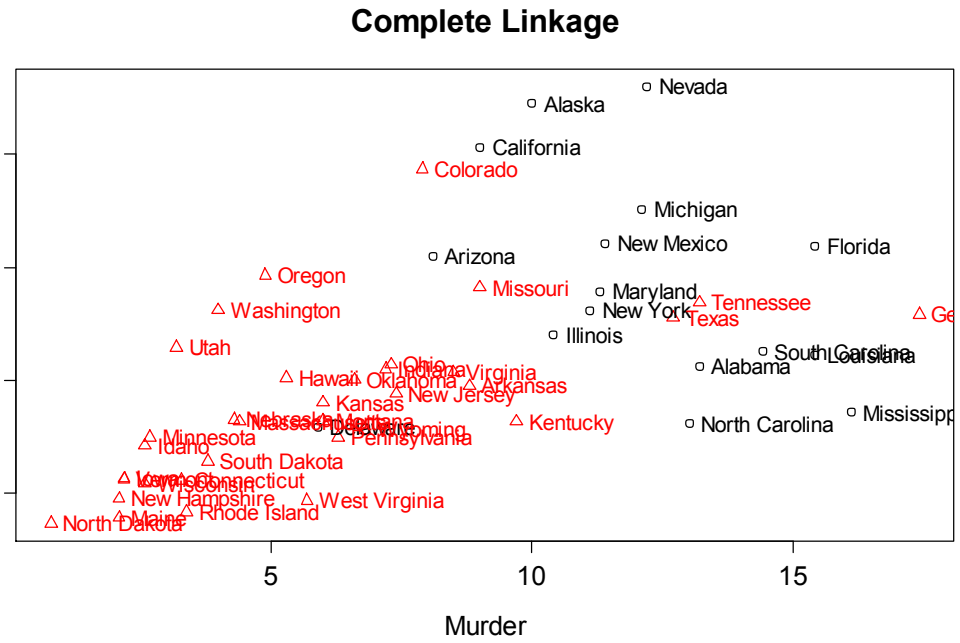





Average Linkage
Height
0
50
100
150
Florida
North Carolina
Arizona
Maryland
California
New Mexico
South Carolina
Alaska
Michigan
Nevada
Alabama
Delaware
Mississippi
New York
Illinois
Louisiana
Washington
Oregon
Massachusetts
Oklahoma
Wyoming
New Jersey
Virginia
Colorado
Georgia
Texas
Rhode Island
Missouri
Arkansas
Tennessee
Nebraska
Connecticut
Kentucky
Montana
Pennsylvania
Idaho
Indiana
Kansas
Ohio
Utah
Minnesota
South Dakota
Maine
West Virginia
Hawaii
Wisconsin
Iowa
New Hampshire
North Dakota
Vermont
dist(data, method = "euclidean")
hclust (*, "average")


Ward's Method
Height
0
1000
2000
Alaska
Michigan
Nevada
Alabama
Delaware
Mississippi
New York
Illinois
Louisiana
Florida
North Carolina
Arizona
Maryland
California
New Mexico
South Carolina
Washington
Massachusetts
Oklahoma
Oregon
Wyoming
New Jersey
Virginia
Colorado
Georgia
Texas
Rhode Island
Missouri
Arkansas
Tennessee
Nebraska
Connecticut
Kentucky
Montana
Pennsylvania
Indiana
Kansas
Idaho
Ohio
Utah
Minnesota
South Dakota
Maine
West Virginia
Wisconsin
Iowa
New Hampshire
Hawaii
North Dakota
Vermont
dist(data, method = "euclidean")
hclust (*, "ward")


Single Linkage
Rape
Murder
10
20
30
40
5
10
15
Nevada
Alaska
California
Colorado
Michigan
New Mexico
Florida
Arizona
Oregon
Missouri
Maryland
Tennessee
Washington
New York
Texas
Illinois
Utah
Alabama
Hawaii
Oklahoma
Arkansas
New Jersey
Kansas
Kentucky
North Carolina
Minnesota
Idaho
South Dakota
Connecticut
New Hampshire
West Virginia
Rhode Island
North Dakota


Complete Linkage
Murder
5
10
15
Nevada
Alaska
California
Colorado
Michigan
New Mexico
Florida
Arizona
Oregon
Missouri
Maryland
Tennessee
Washington
New York
Texas
Illinois
Utah
Alabama
Hawaii
Oklahoma
Arkansas
New Jersey
Kansas
Kentucky
North Carolina
Minnesota
Idaho
South Dakota
Connecticut
New Hampshire
West Virginia
Rhode Island
North Dakota

비계층적군집분석:K-means Clustering

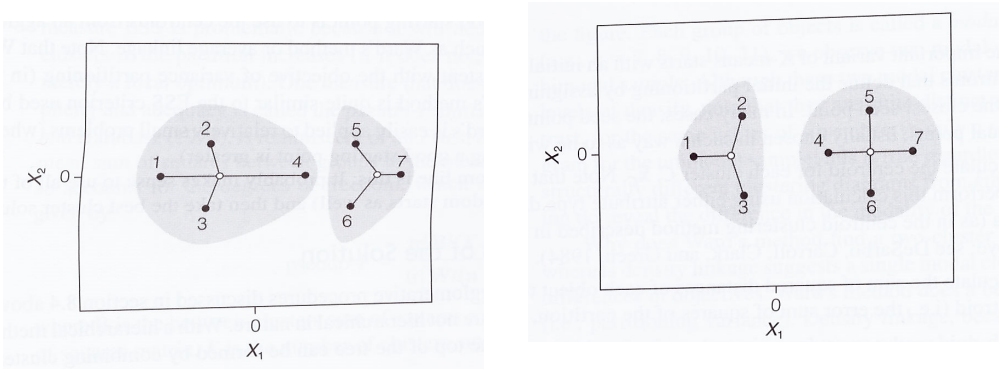
- 군집의 개수를 미리 설정 (k개)

Step 1. 관측치를 Initial Partition k개로 나누고 각 군집의 중심인 seed 계산

Step 2. 각 관측치로부터 각 seed사이의 거리 계산 (n*k개)하여 가장 가까운 seed에 할당

Step 3. 군집의 seed를 군집에 속한 관측치의 평균(중심)으로 업데이트

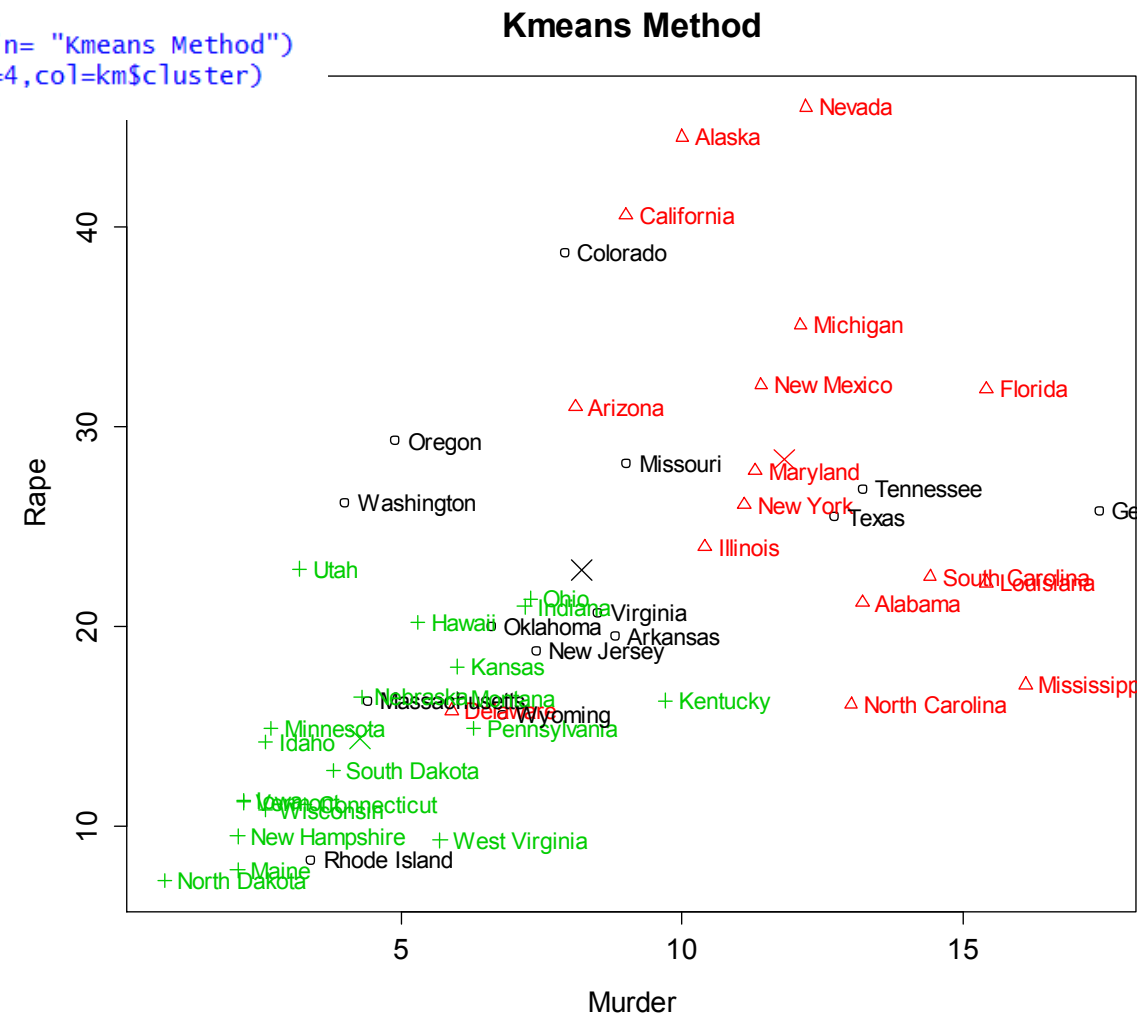
더이상 개체의 군집 이동이 없을 때 까지 반복



- 특징
  - 척도불변성이 아님 (표준화 하면 다른 결과)
  - 데이터에 대해 spherical 구조를 강요

예: USArrests

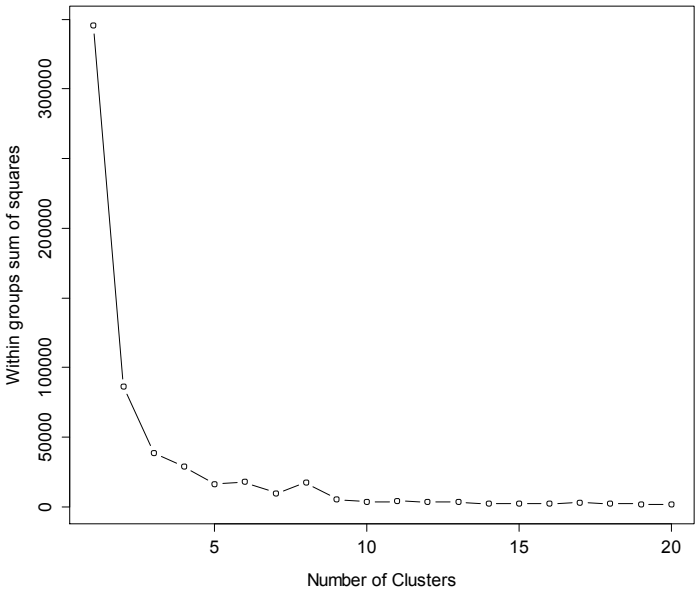
```
> km=kmeans(data,centers=3)
> plot(Murder,Rape,pch=km$cluster,col=km$cluster,main= "Kmeans Method")
> text(Murder,Rape,labels=state, cex=0.8, adj=0,pos=4,col=km$cluster)
> points(km$centers[,c(1,3)],col=1:3,pch=4,cex=2)
```

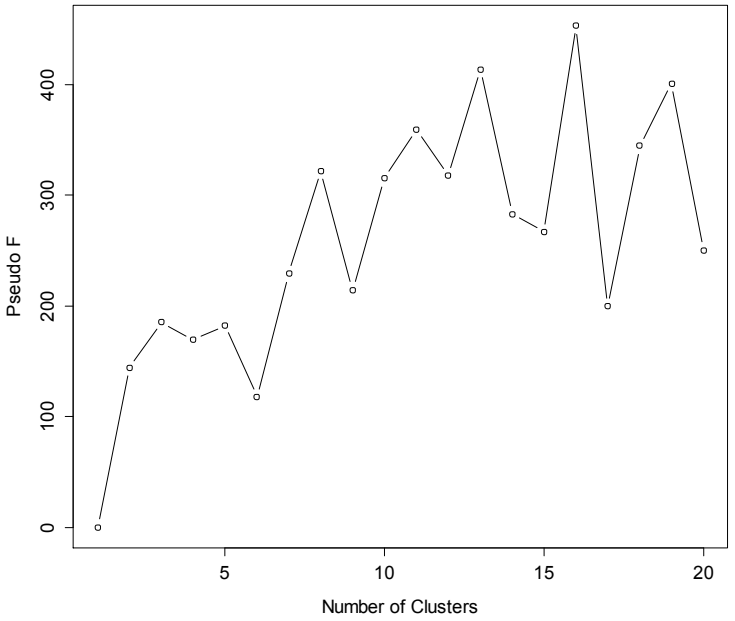


비계층적군집화: 군집의 개수 결정

- ESS (Within-group sum of squares)는 군집의 개수가 늘어날때 감소
  - ESS의 감소가 완만해 지는 '팔꿈치 ' 에서 결정
  - Pseudo-F=$\frac{tr[B/(k-1)]}{tr[W/(n-K)]}$

  B: between-clusters sum of squares matrix, W: within-clusters sum of squares matrix
  pseudo-F가 클수록 효과적인 partitioning





모형 기반 군집분석(Model-based Clustering)

- 확률분포에 대한 정보가 있을 경우 이를 활용하여 군집분석
- k번째 군집에 속한 관측치 x가 확률밀도함수 $f(\boldsymbol{x};\boldsymbol{\theta})$를 가진다고 가정
- 주로 $f(\boldsymbol{x};\boldsymbol{\theta})$를 평균이 $\boldsymbol{\mu_k}$, 공분산행렬이 Σ_k 인 다변량 정규분포로 가정

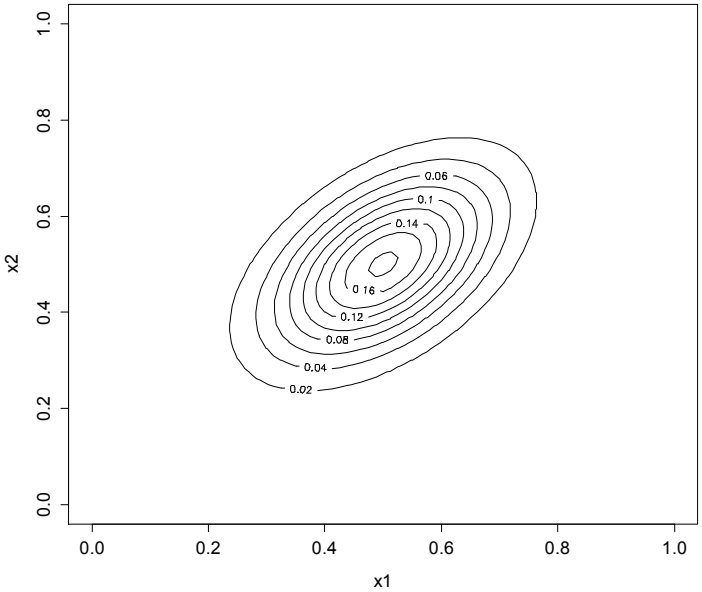
$$f(x) = \sum_{k=1}^{G} \pi_k f(\boldsymbol{x};\boldsymbol{\mu_k},\Sigma_k), \qquad \sum_k \pi_k = 1$$

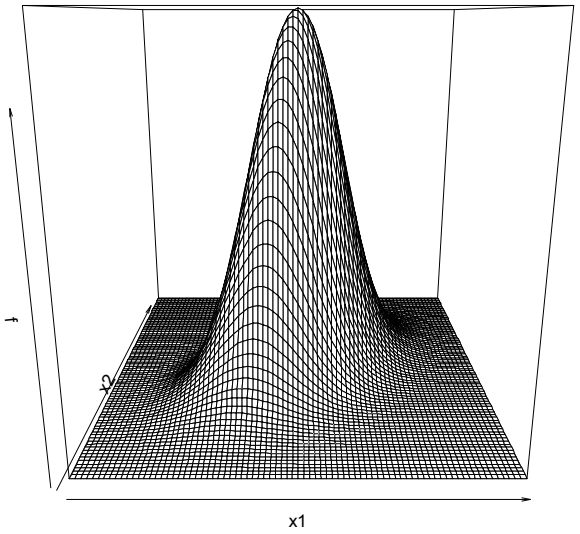
- Σ_k 의 형태에 대한 가정 (G:그룹의 개수, d:변수의 개수)

Identifier	Model	# Covariance parameters	Distribution
EII	λI	1	Spherical
VII	$\lambda_k I$	G	Spherical
EEI	λA	d	Diagonal
VEI	$\lambda_k A$	$G + (d-1)$	Diagonal
EVI	λA_k	$1 + G(d-1)$	Diagonal
VVI	$\lambda_k A_k$	Gd	Diagonal
EEE	$\lambda DAD^\top$	$d(d+1)/2$	Ellipsoidal
EEV	$\lambda D_k A D_k^\top$	$1 + (d-1) + G[d(d-1)/2]$	Ellipsoidal
VEV	$\lambda_k D_k A D_k^\top$	$G + (d-1) + G[d(d-1)/2]$	Ellipsoidal
VVV	$\lambda_k D_k A_k D_k^\top$	$G[d(d+1)/2]$	Ellipsoidal

- Recall: 다변량 정규분포
  - 일변량 정규분포 $f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \ -\infty < x < \infty$
  - X: p변량 확률벡터
  - $\boldsymbol{\mu} = E(X), \Sigma = cov(X)$

$f(x) = \frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}} e^{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})'\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu})}, \quad x = (x_1, \ldots, x_p)', \ -\infty < x < \infty, i = 1,2, \ldots, p$





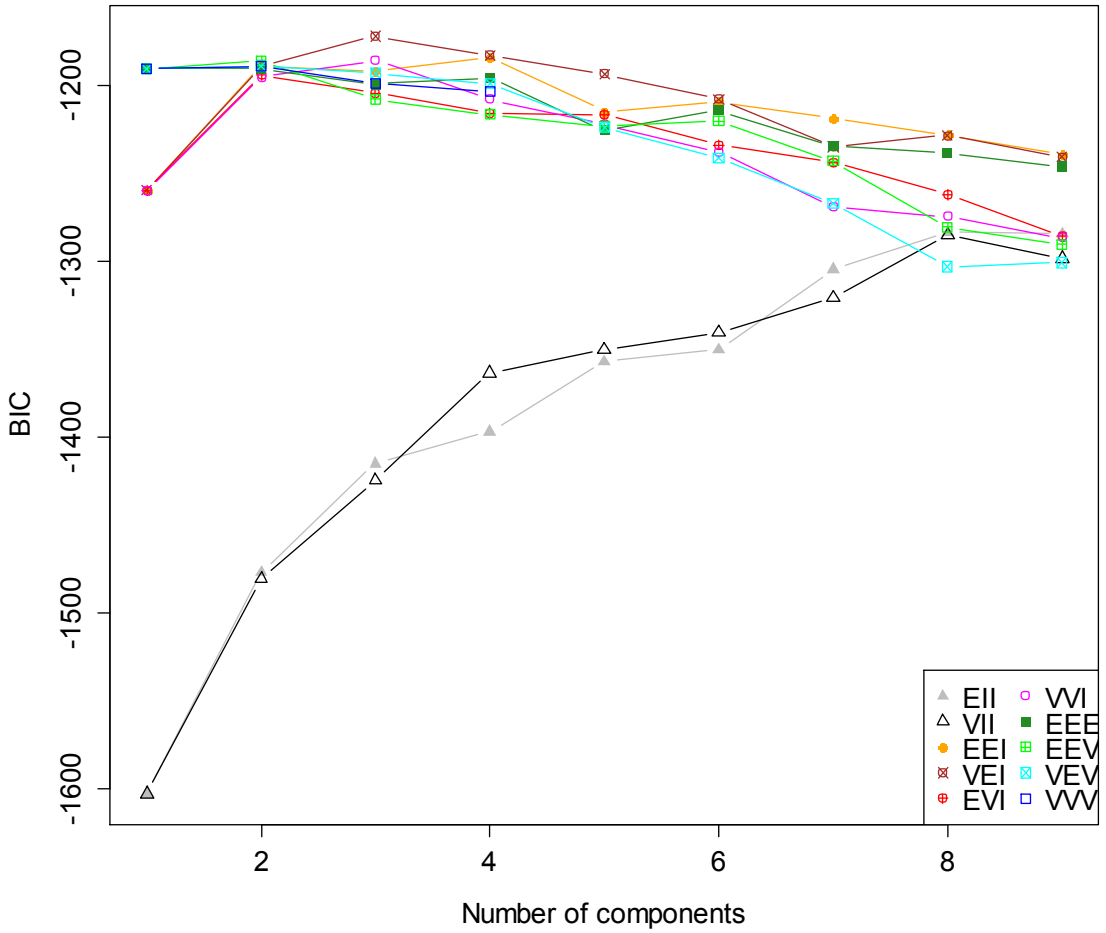
$$f(x) = \sum_{k=1}^{G} \pi_k f(\boldsymbol{x}; \boldsymbol{\mu_k}, \Sigma_k), \qquad \sum_k \pi_k = 1$$

- EM algorithm을 사용해 각 그룹의 $\pi_k, \boldsymbol{\mu}_k, \Sigma_k$ 추정
- 각 관측치에 대해 각 군집에 속할 사후 확률을 계산하여 가장 확률이 높은 군집으로 할당
- 모형선택방법 BIC (Bayesian information criterion)을 사용해 군집의 개수 및 적절한 공분산행렬의 모양 선택 가능
  - BIC값이 최대가 되는 모형 선택

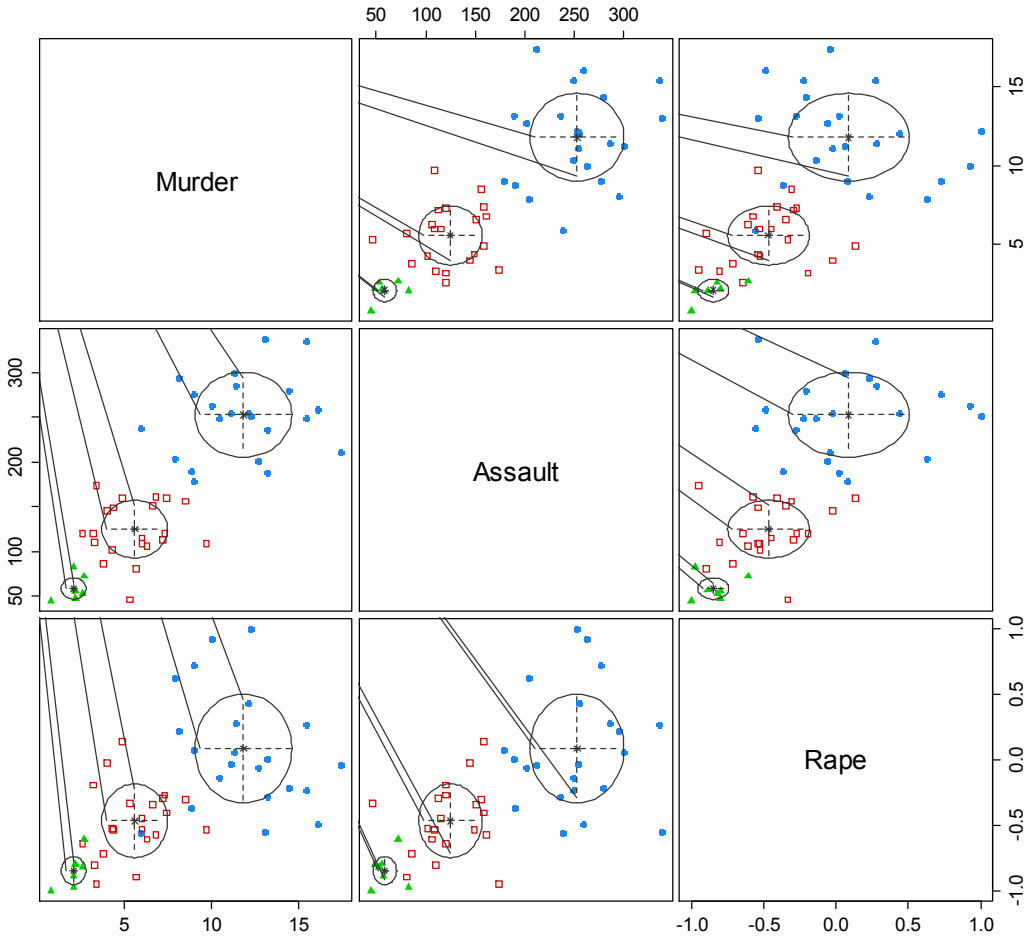
$$BIC = 2 \cdot loglikelihood - (\#\ of\ parameters)$$

예: USArrests

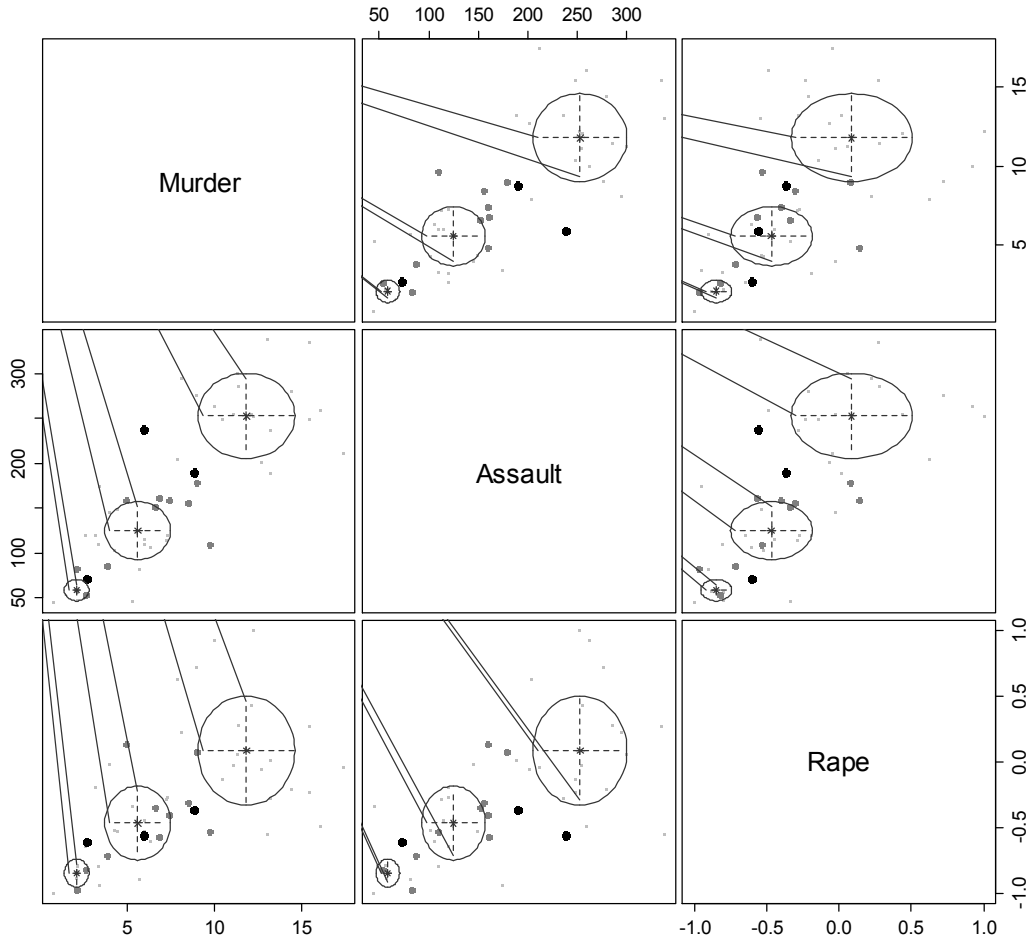
```
> library(mclust)
> mc=Mclust(data)
> summary(mc)
----------------------------------------------------
Gaussian finite mixture model fitted by EM algorithm
----------------------------------------------------

Mclust VEI (diagonal, equal shape) model with 3 components:

 log.likelihood  n df       BIC       ICL
      -554.5497 50 16 -1171.692 -1174.609

Clustering table:
 1  2  3
22 21  7


> plot(mc)
```

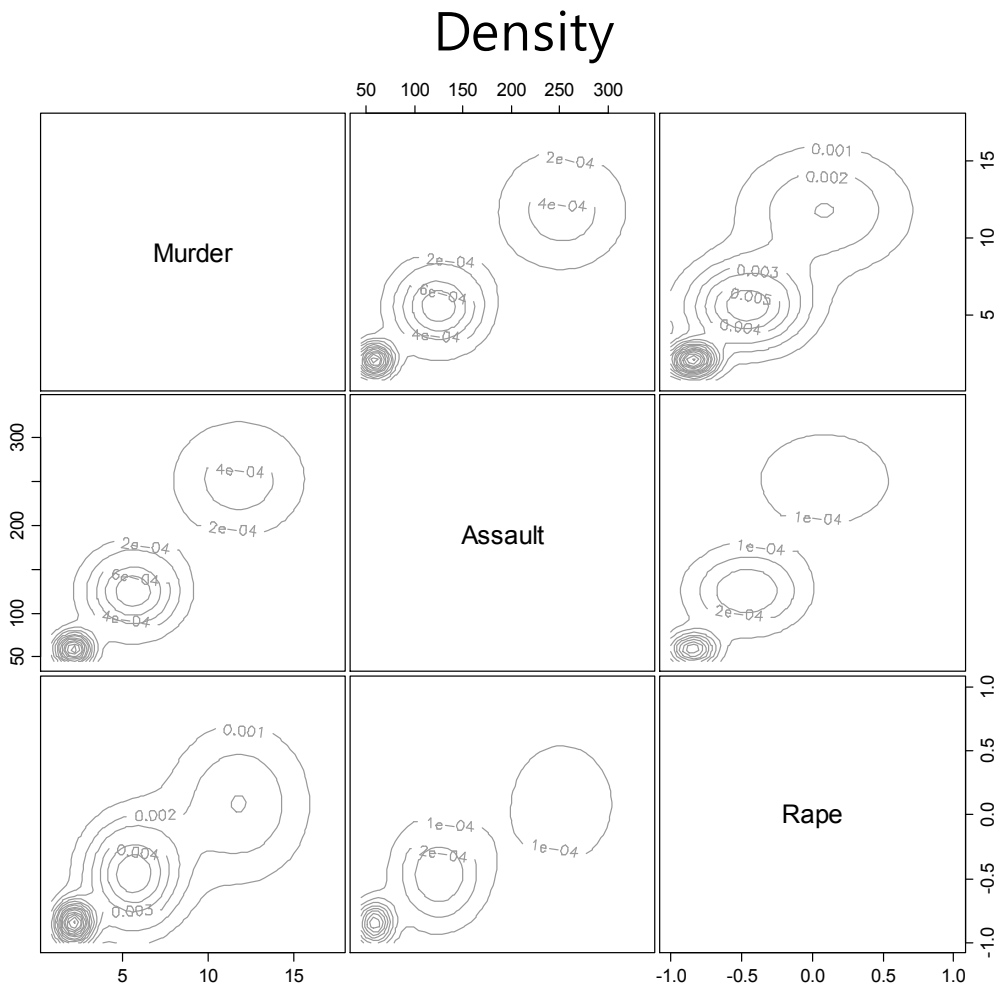


Clustering result



Uncertainty
: symbol이 클수록 어느 군집에 속하는지데 대한 불확실성이 큼





- 군집분석과 동시에 각 군집에 대한 분포 추정이 가능함