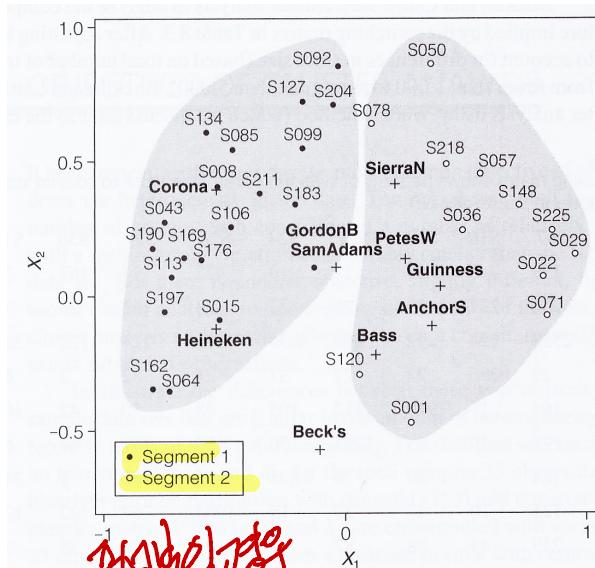
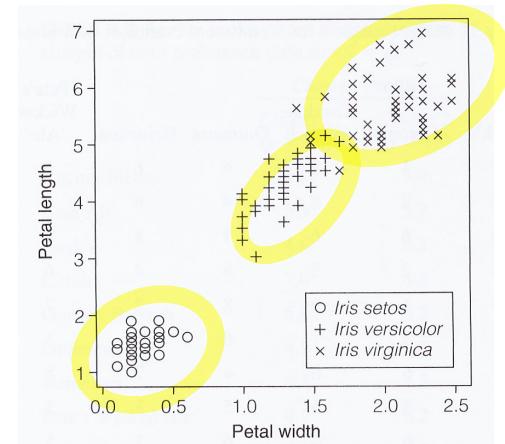


군집분석

Cluster Analysis

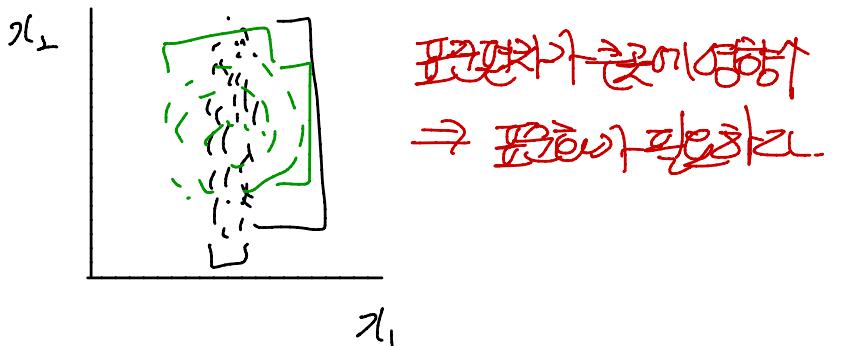
군집분석이란?

- 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 관찰치는 서로 상이한 속성을 갖도록 군집 형성
- 군집분석의 활용
 - 생물의 종 구분
 - 시장세분화 *Segmentation*
 - 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



시장세분화
그룹화의 목표는?

유사성 측도



- 유clidean 거리(Euclidean Distance)

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

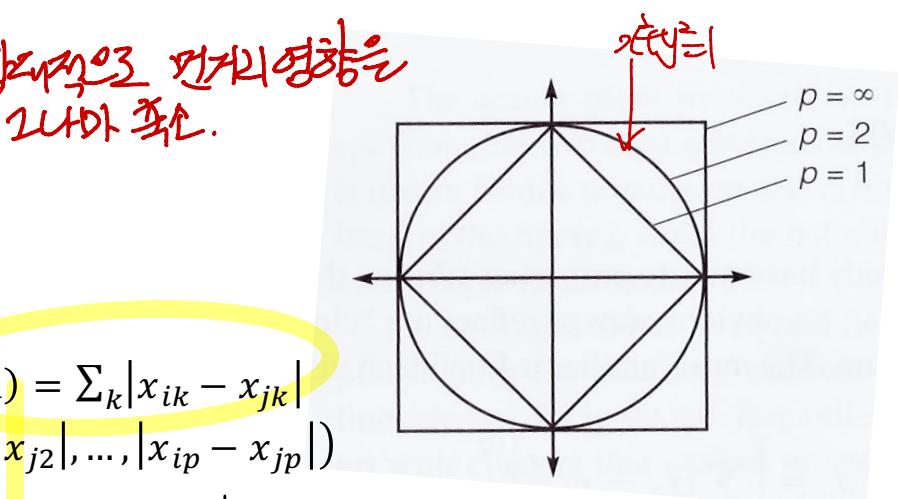
일반화 ($p=2$)

- 변수들의 서로 다른 단위를 가진 경우가 많으므로 주로 표준화된 자료에 사용
- 민코우스키 거리(Minkowski p-Metric)

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

상수적으로 편의영향을
제거해 주는 것.

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

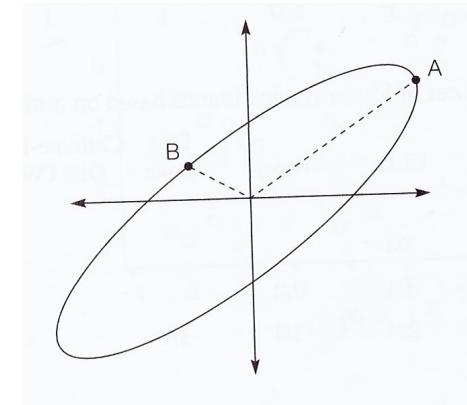


$(x_1 - y_1) + (x_2 - y_2) + \dots + (x_p - y_p)$

- 마할라노비스 거리 (Mahalanobis Distance)

$$D_{ij} = (\mathbf{x}_i - \mathbf{x}_j)' \Sigma^{-1} (\mathbf{x}_i - \mathbf{x}_j)$$

- 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 거리
 - 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 $d_{C_i C_j} = d_{ij}$

ex) 50주가 촉 50개의 군집을 가지고 시작

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

step 2. 새 군집과 기존 군집들 간의 거리 계산

→ 군집과 군집간의 거리를 계산하기

step 3. 하나의 군집이 남을 때 까지 step 1 & 2 반복

· 구간마다 따라가 방향이 달라진다. (Next)

- 군집의 개수 결정**

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

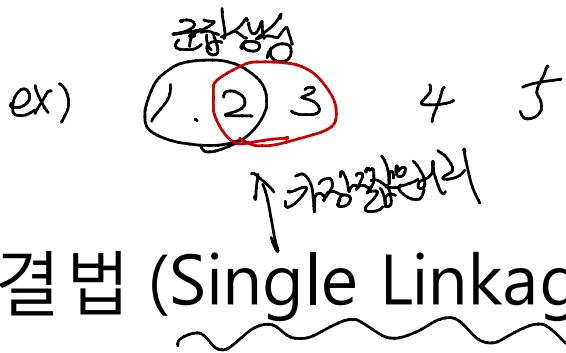
계층적 군집분석: 최단연결법 (Single Linkage)

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

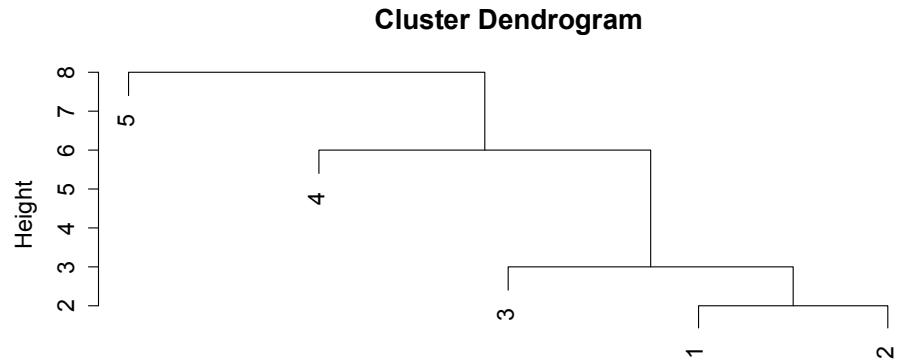
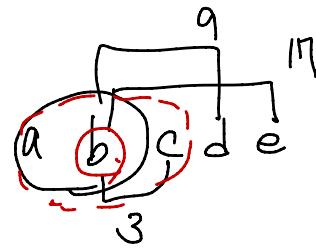
```

> 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)

```



a	1
b	3
c	6
d	12
e	20.



dist(x, method = "euclidean")
hclust (*, "single")

- 계산이 효율적
- 근시안적: 길다란 형태의 군집 형성 가능

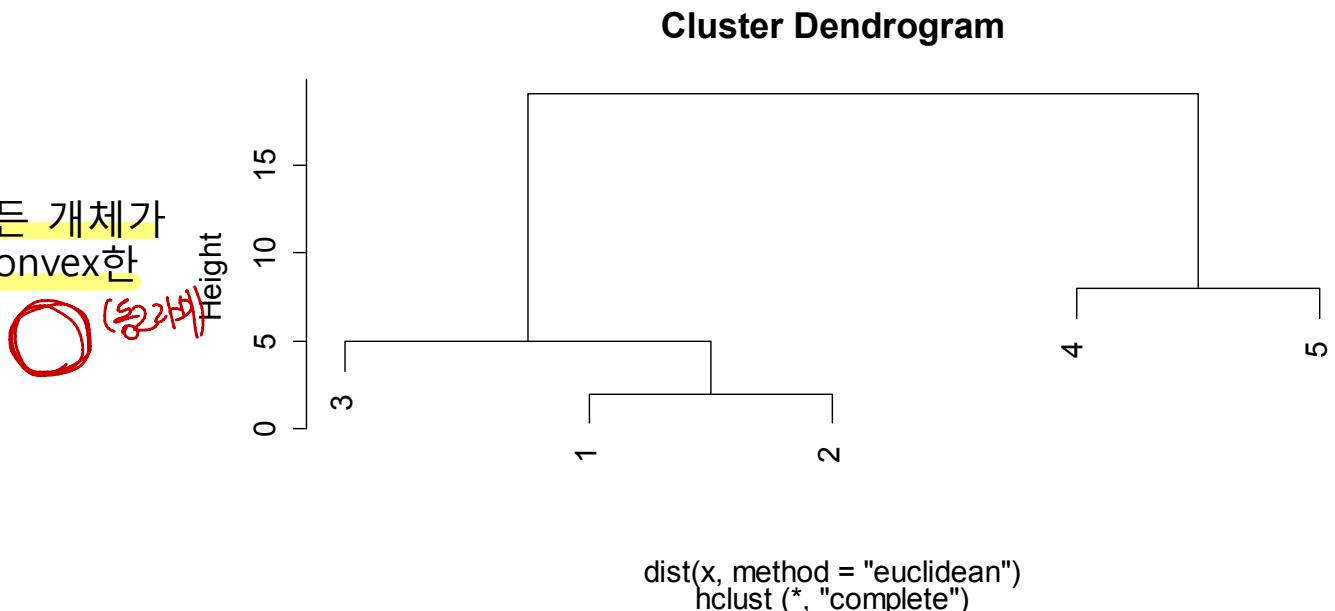
계층적 군집분석: 최장연결법 (Complete Linkage)

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

가장먼저(?) 있는 두개의 군집이 합쳐지면
군집형성이 안될수 있다.

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

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



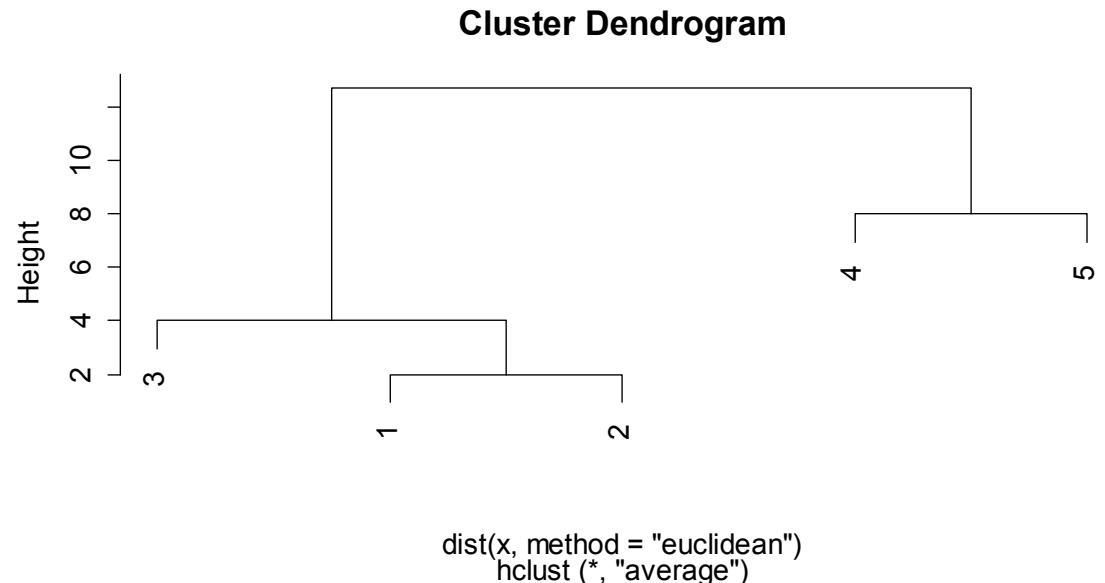
계층적 군집분석: 평균연결법 (Average Linkage)

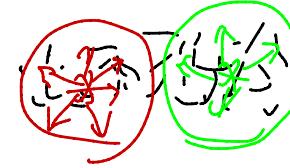
각 군집간의
평균연결

$$\cdot d_{C_1 C_2} = \frac{1}{n_1 n_2} \sum_i \sum_j d_{ij} \max\{d(x, y) | x \in C_i, y \in C_j\}$$

```
> 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)

$$ESS_C = \sum_j (x_{Cj} - \bar{x}_C)'(x_{Cj} - \bar{x}_C)$$

정상화된 내분산 거리의 합

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

계층적 군집분석: 군집 개수의 결정

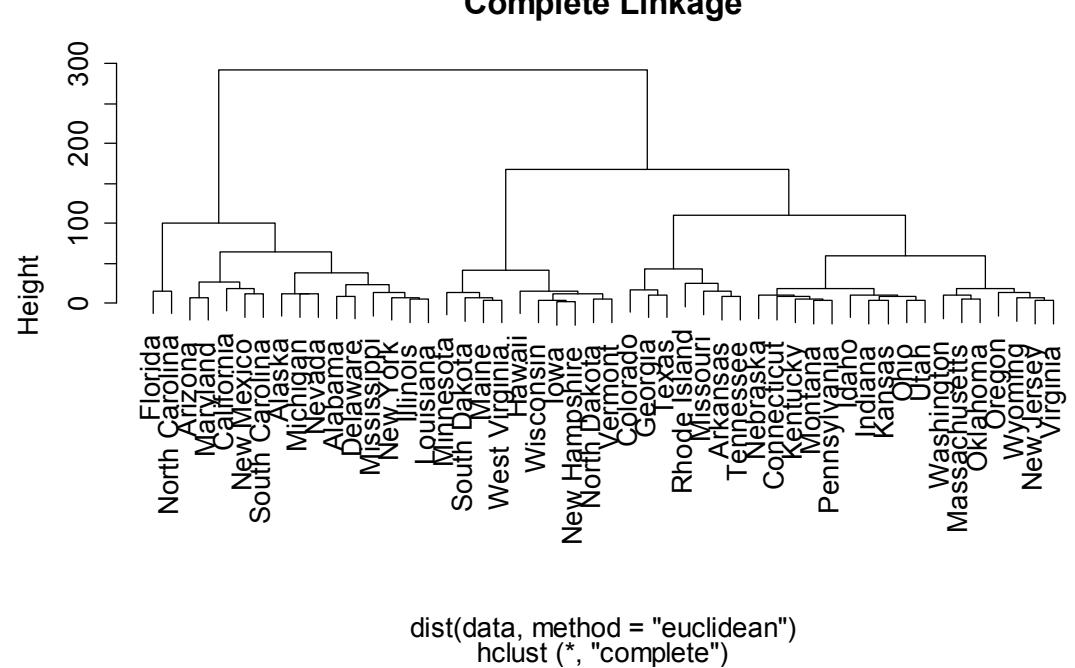
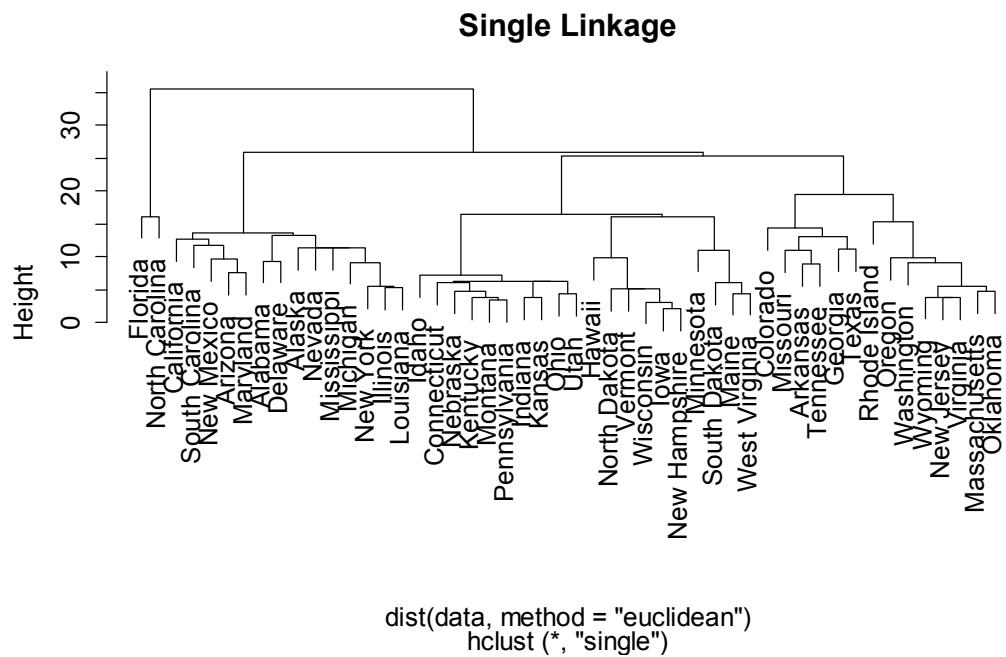
- 덴드로그램을 어디서 자를 것인가? 데이터에 최적인 특정한 그룹의 개수를 어떻게 정할 것인가?
- 덴드로그램 높이 변화의 크기를 조사하여 큰 변화가 있는 곳에서 자름
- 적절한 군집의 개수가 알려져 있을 경우 해당 군집 개수를 얻는 위치에서 자름
- `cutree` 함수 사용하여 군집분석의 해 얻음

Example: USArrests

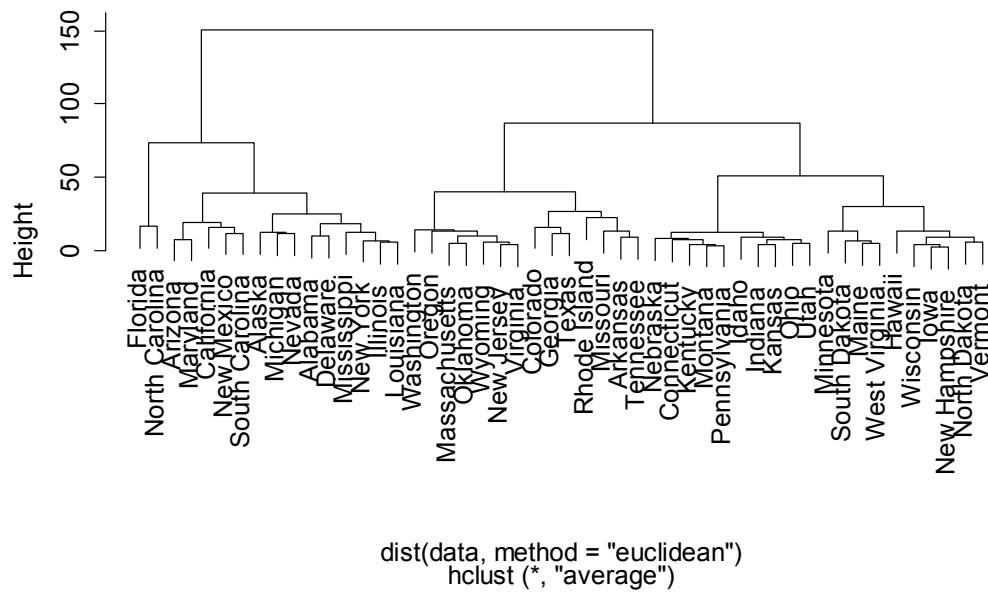
- This data set contains statistics, in arrests per 100,000 residents for assault, murder, and rape in each of the 50 US states in 1973. Also given is the percent of the population living in urban areas.
- A data frame with 50 observations on 4 variables.

```
[,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)
```

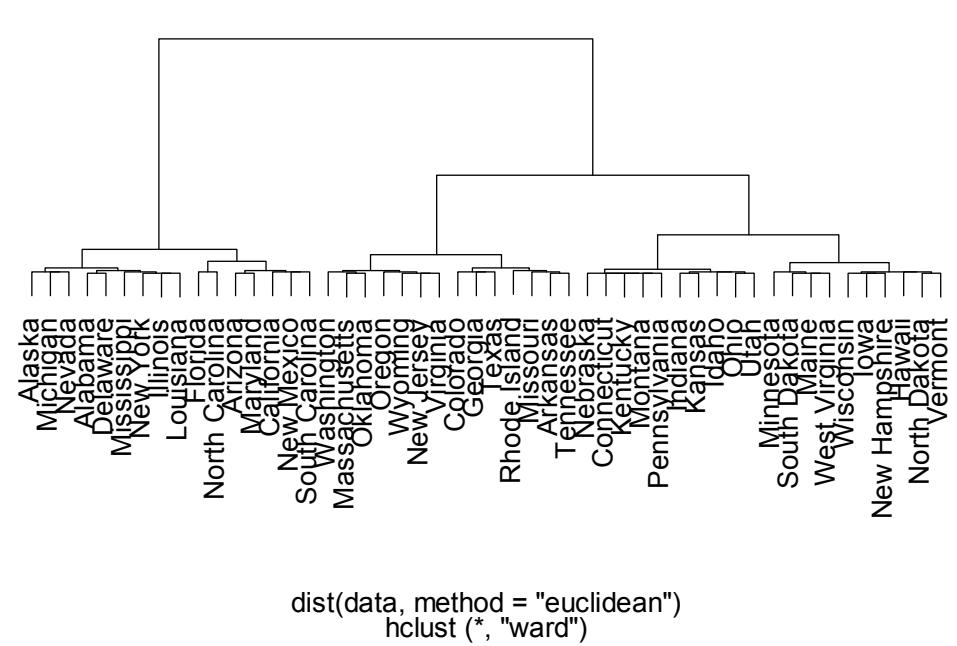
- 전 미 50개 주의 1973년 인구 10만명 당 체포범죄건수
- 변수: Murder, Assault, Rape



Average Linkage



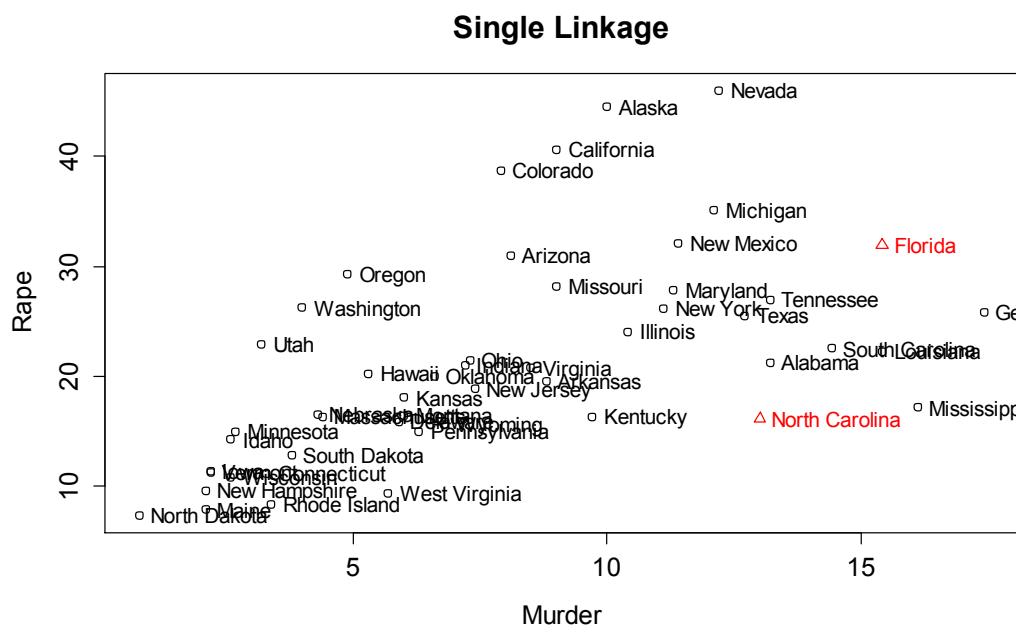
Ward's Method



- 그룹별로 Boxplot 을 그려보는것이 좋다.

• 그룹별로
[그림]

Confidence Interval
그룹별로 그려보자.



비계층적군집분석: K-means Clustering (고려하지만 사용자↑)

- 군집의 개수를 미리 설정 (k개)

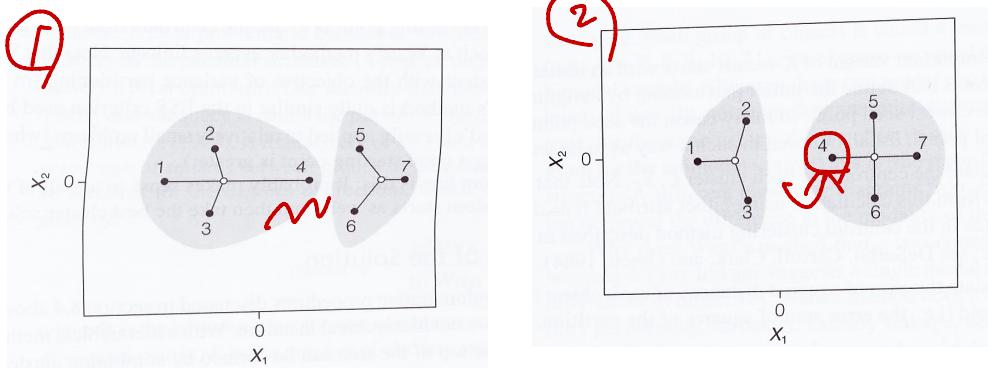
Step 1. 관측치를 Initial Partition k개로 나누고 각 군집의 중심인 seed 계산

(
둘일 때마다 결과가
다를 수 있다.)

Step 2. 각 관측치로부터 각 seed 사이의 거리 계산 ($n \times k$ 개)하여 가장 가까운 seed에 할당 Why?

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

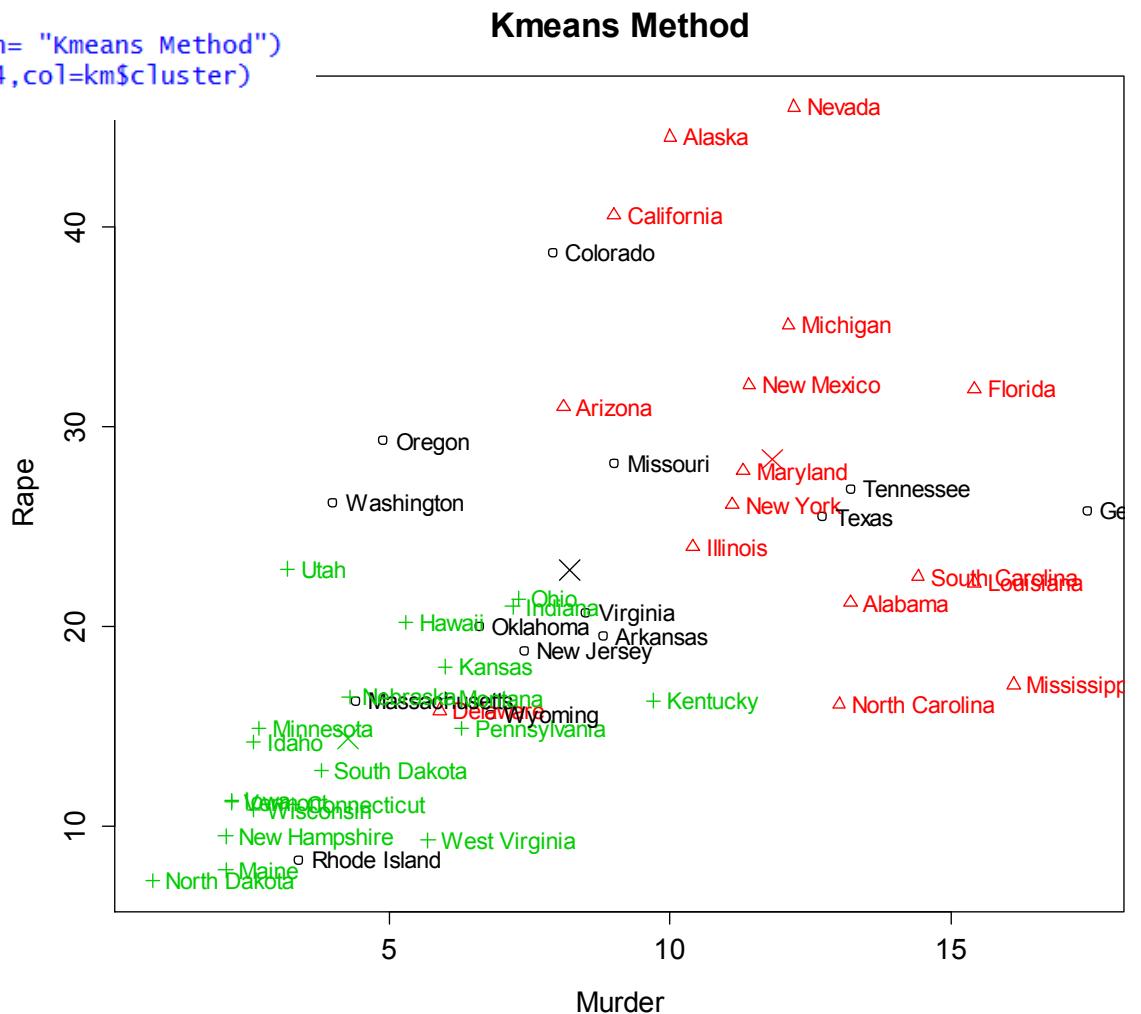
Initialization
Random 초기화
기본적으로 초기화하는
방법

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)

```



Barplot은 Boxplot에 비해 사용하기는 애는

Standard Error Bar

→ 모집단에서의 예측값과 같은데.

Standard error Bar가 표수로 표기

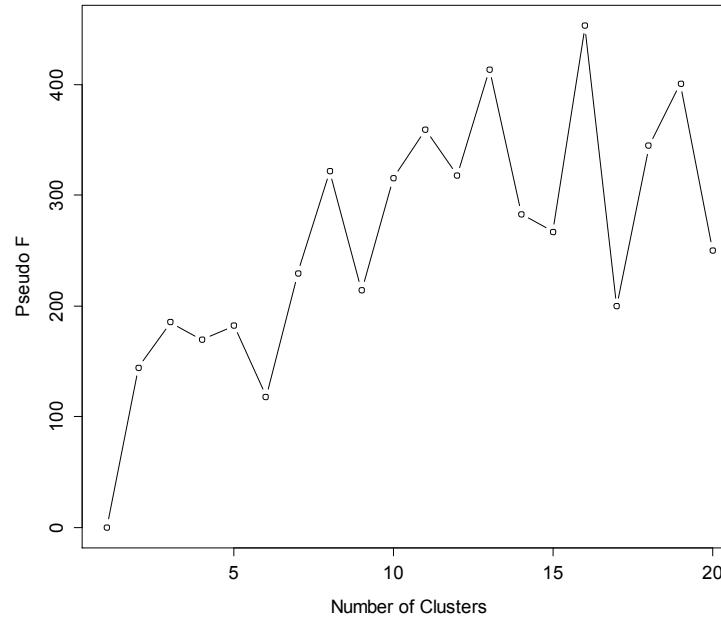
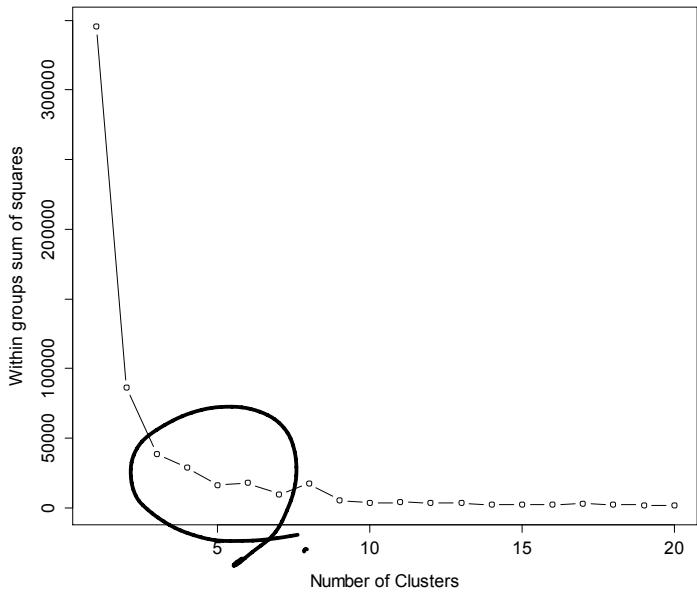


비계층적군집화: 군집의 개수 결정

어떻게 정하는지.

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

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



cluster↑ ⇒ cluster 중심까지 거리↓.

통계에서 주로 사용

모형 기반 군집분석(Model-based Clustering)

- 확률분포에 대한 정보가 있을 경우 이를 활용하여 군집분석
- k 번째 군집에 속한 관측치 x 가 확률밀도함수 $f(x; \theta)$ 를 가진다고 가정
- 주로 $f(x; \theta)$ 를 평균이 μ_k , 공분산행렬이 Σ_k 인 다변량 정규분포로 가정

$$f(x) = \sum_{k=1}^K \pi_k f(x; \mu_k, \Sigma_k), \quad \sum_{k=1}^K \pi_k = 1$$

~~시각화~~ ~~우치(어느정도) k~~

- 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	λDAD^\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

각 그룹의 확률밀도. 종합밀도를 이룰 것이다.

Σ_k
(공분산행렬) ~~다음~~

ex)
 $\lambda = 2$.

$f(x) = \pi_1 f(x_1, \mu_1, \Sigma_1)$
 $+ \pi_2 f(x_2, \mu_2, \Sigma_2)$

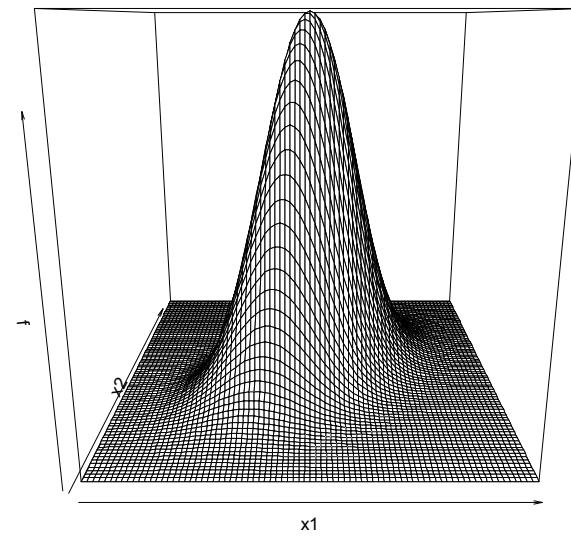
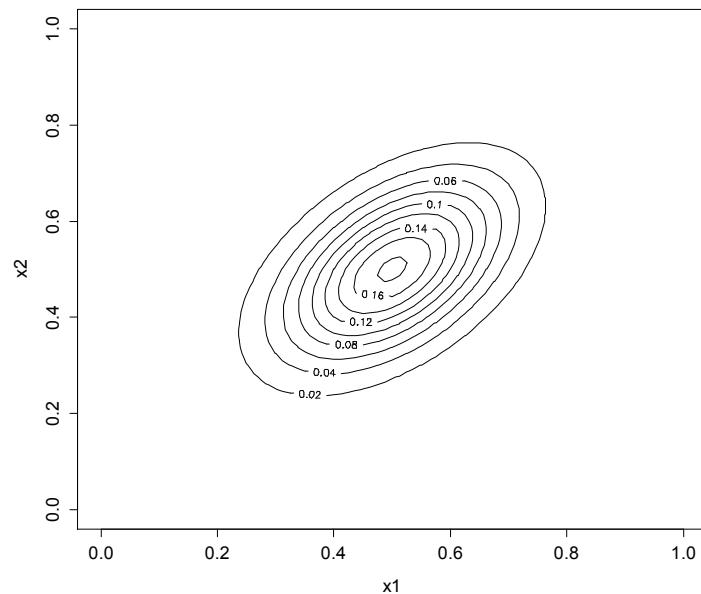
$\Sigma_1 = \Sigma_2 = \lambda I$. (Identity matrix)

공분산 행렬
= 측정자. $= \begin{bmatrix} \lambda & x_1 \\ x_1 & ? \end{bmatrix}$

- 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
- $\mu = E(X), \Sigma = cov(X)$

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



- 모형 비교 기준은 학파수로에 독립변수의 개수에 대한
존재성을 관여하는 방법.
소식가중치를 계산할 때 따라 AIC / BIC. 두 가지 사용.

여러 사정 종류는 하나로 묶어

$$f(x) = \sum_{k=1}^K \pi_k f(x; \mu_k, \Sigma_k),$$

각 모가지가 증가하면

$$\sum_k \pi_k = 1$$

여기서 Σ 가 나온점은 가 어느정도

AIC : Akaike information criterion
 \Rightarrow a measure of the relative quality
 of statistical models for a given
 Data.

- EM algorithm을 사용해 각 그룹의 π_k, μ_k, Σ_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
 - BIC 값이 최대가 되는 모형 선택 가장 적합한 모형 선택 기준을 설정하는 것이다.

$$BIC = 2 \cdot \log L - (\# \text{ of parameters})$$

후진적 model 선택에서 활용하는 (forward, Backward, Aitken)

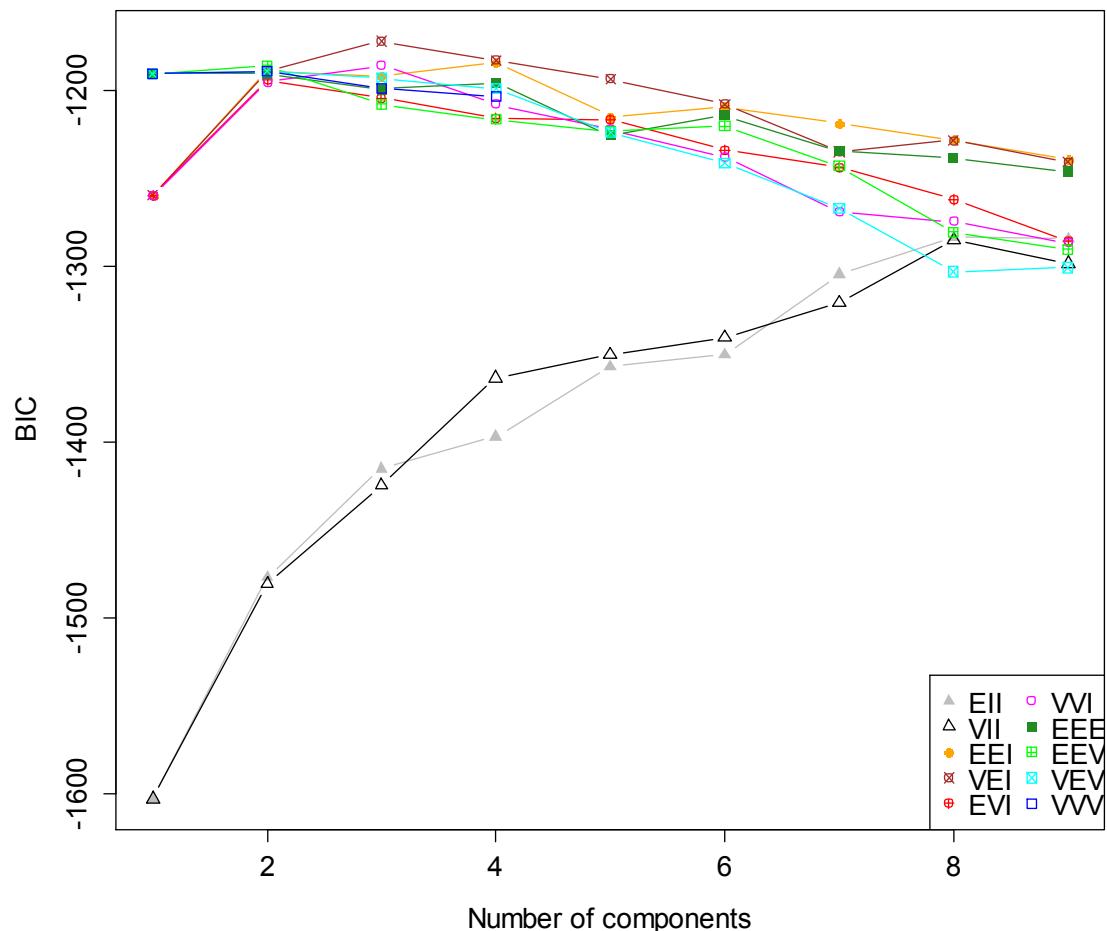
- $AIC = -2 \cdot \log L + 2k$.
독립변수의 수.
- $BIC = -2 \cdot \log L + k \log n$.

예: USArrests

```
> library(mclust)
> mc=Mclust(data)
> summary(mc)
Gaussian finite mixture model fitted by EM algorithm
-----
Mclust VEV (diagonal, equal shape) model with 3 components:
log.likelihood n df      BIC      ICL
-554.5497 50 16 -1171.692 -1174.609

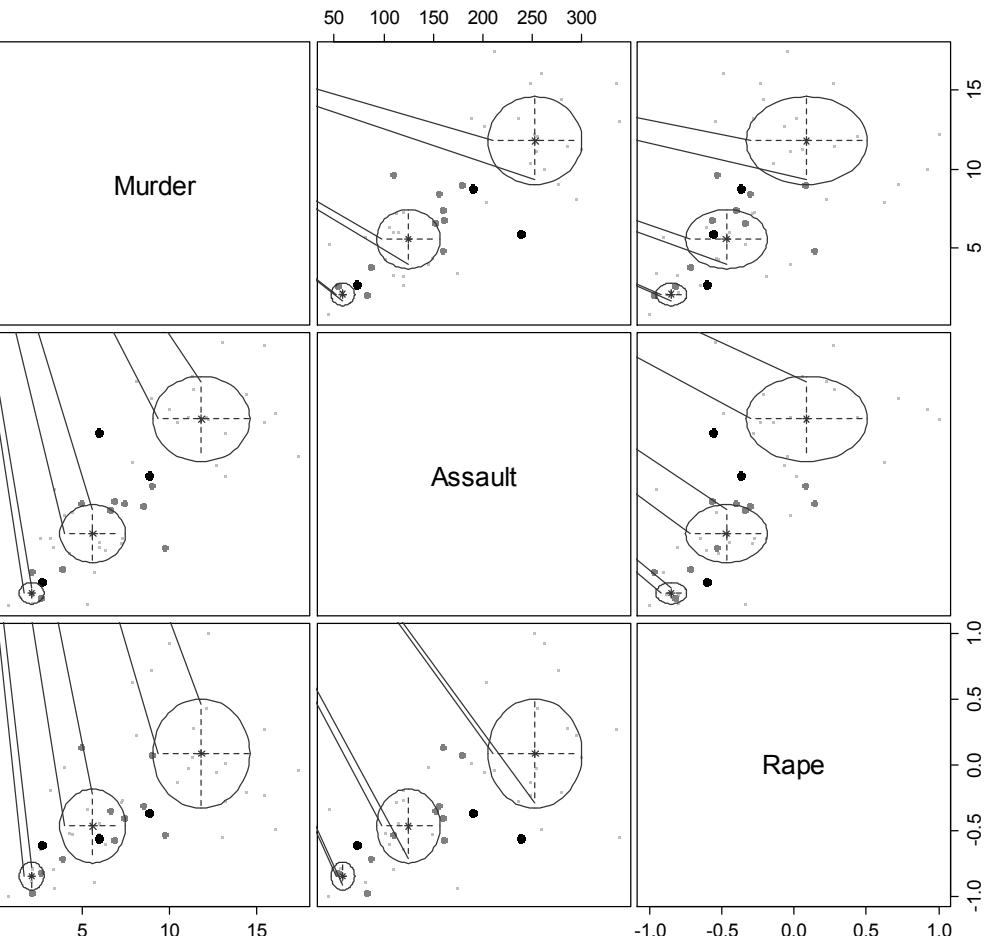
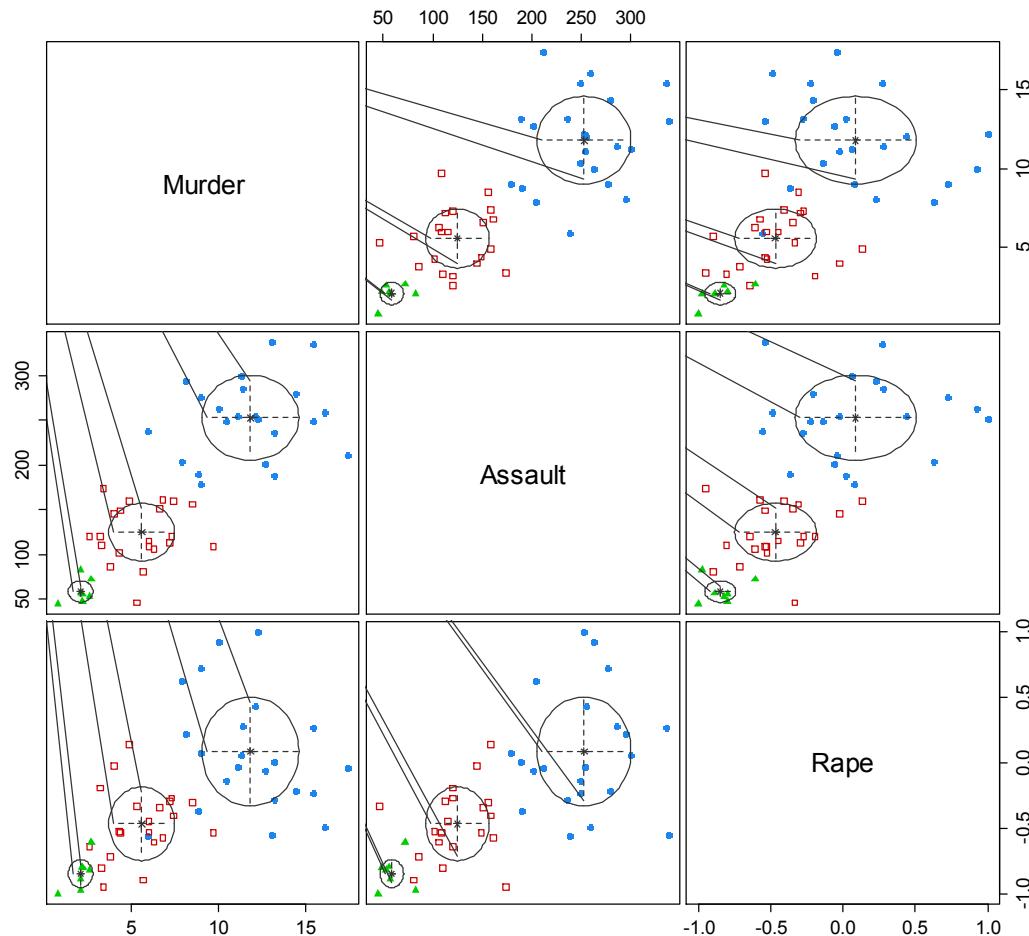
Clustering table:
 1 2 3
22 21 7

> plot(mc)
```

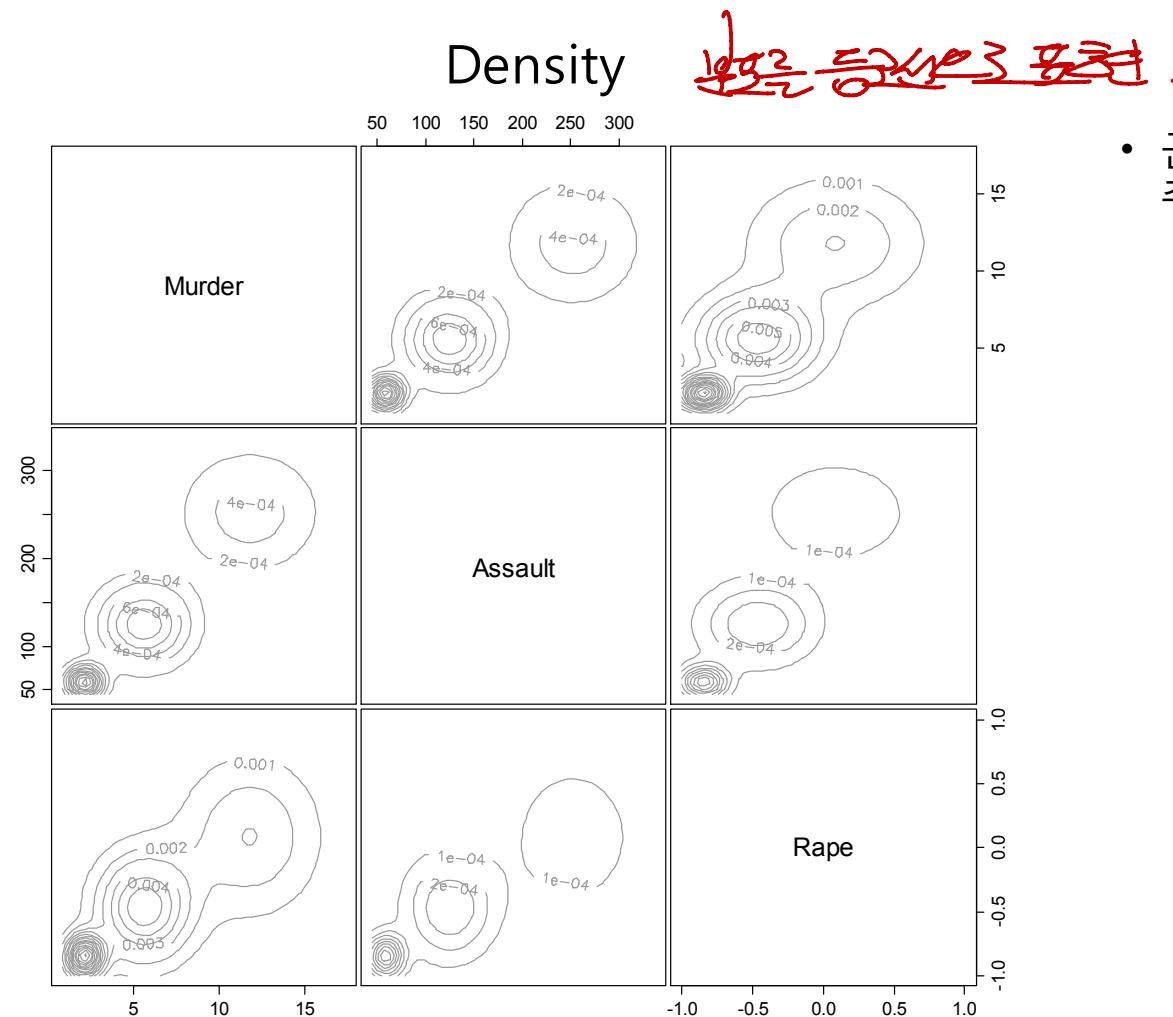


• **Uncertainty** | 경계(이웃)는 그림에 표시되는 것과는 다른 것

Clustering result



Uncertainty
: symbol이 클수록 어느 군집에 속하는지에 대한 불확실성이 큼



- 군집분석과 동시에 각 군집에 대한 분포 추정이 가능함

1군집을 등록하는 과정 .