11장 군집분석

덕성여자대학교 정보통계학과

김 재희

11.8 모형기반 군집 방법

▶ Scott and Symons (1971)가 제안

Banfield and Raftery (1993), Fraley and Raftery (2002) 등이 발전시킨 방법으로 모형 기반 군집방법(model-based clustering)을 설명하고자 한다.

▶ 모집단이 G개의 군집으로 구성.

k번째 군집에 속한 $p-$차원 관측벡터 $\boldsymbol{x}$의 밀도함수는 $f_k(\boldsymbol{x},\boldsymbol{\theta})$라고 가정.

$\boldsymbol{\gamma}=(\gamma_1,\cdots,\gamma_n)'$: 여기서 $\boldsymbol{x_i}$가 k번째 군집에 속하였으면 $\gamma_i=k$이다.

▶ 가능도함수는

$$L(\boldsymbol{\theta},\boldsymbol{\gamma})=\prod_{i=1}^{n} f_{\gamma_i}(x_i;\theta_{\gamma_i})$$

가능도를 최대화하는 $\boldsymbol{\theta}=(\theta_1,\cdots,\theta_G)'$ 와 $\boldsymbol{\gamma}=(\gamma_1,\gamma_2,...,\gamma_n)'$ 를 선택하게 된다.

데이터가 속한 군집은 내재하는 확률분포로부터 형성되었다고 가정하고

혼합 모형(mixture model)

$$L_{mix}(\boldsymbol{\theta},\boldsymbol{\gamma}) = \prod_{i=1}^{n}\sum_{k=1}^{G}\tau_k f_k(\boldsymbol{x_i}|\boldsymbol{\theta_k})$$

을 고려한다. 여기서 $\boldsymbol{\theta}=(\theta_1,\cdots,\theta_G)'$는 모수벡터

τ_k는 관측벡터가 k 번째 군집에 속할 확률이며 $\tau_k \geq 0,\ \sum_{k=1}^{G}\tau_k = 1$.

▶ 관측벡터가 다변량 정규분포를 따른다고 가정하는 Gaussian 혼합모형을 고려.

$f_k(\boldsymbol{x},\boldsymbol{\theta})$: 평균벡터 $\boldsymbol{\mu_k}$와 공분산행렬 $\boldsymbol{\Sigma_k}$를 갖는 k 번째 군집의 다변량 정규밀도함수

가능도함수 :

$$L(\boldsymbol{\theta},\boldsymbol{\gamma}) = const \cdot \prod_{k=1}^{G}\prod_{i\in E_k}|\boldsymbol{\Sigma_k}|^{1/2}\exp\left\{-\frac{1}{2}(\boldsymbol{x_i}-\boldsymbol{\mu_k})'\boldsymbol{\Sigma_k}^{-1}(\boldsymbol{x_i}-\boldsymbol{\mu_k})\right\}$$

여기서 $E_k = \{i;\gamma_i = k\}$이다. $\boldsymbol{\mu_k}$의 최대가능도추정량은 $\overline{\boldsymbol{x_k}} = n_k^{-1}\Sigma_{i\in E_k}\boldsymbol{x_i}$

n_k는 E_k에 속한 원소의 개수.

▶ μ_k에 대한 최대우도추정량인 $\overline{x_k}$ 로 대체하여 로그가능도함수를 나타내면

$$l(\theta,\gamma) = \text{const} - \frac{1}{2}\sum_{k=1}^{G}\left\{tr(W_k\Sigma_k^{-1}) + n_k \log|\Sigma_k|\right\}$$

여기서 const는 상수이고 W_k는 k번째 군집의 표본교차곱행렬

$$W_k = \sum_{i\in E_k}(x_i - \overline{x_k})(x_i - \overline{x_k})'$$

▶ 각 군집을 형성하는 기하학적인 특징(shape, volume, orientation)은 공분산행렬 Σ_k에 의해 결정된다. (Banfield and Raftery(1993))

고유값 분해에 의하여 공분산행렬이 대표성을 갖는 일반적인 구조 :

$$\Sigma_k = \lambda_k D_k A_k D_k'$$

여기서 D_k는 고유벡터의 직교행렬,

A_k는 각 원소가 Σ_k의 고유값을 비례적으로 취하는 대각행렬, λ_k는 스칼라.

D_k는 군집의 orientation을 결정. A_k는 군집의 shape을 결정. λ_k는 군집의 volume을 결정.

▶ 예상되는 군집 개수 G가 정해지면 가능한 군집 개수 $1 \leq k \leq G$ 에 대해 $(\tau_k, \mu_k, \Sigma_k)$ 가 EM 알고리듬에 의해 추정된다.

EM 알고리듬은 E(expectation) 단계와 M(maximization) 단계로 이루어지며,

E 단계에서는 주어진 조건하에서 관측벡터가 각 군집에 속할 확률을 구하고

M 단계에서는 주어진 상황에서 모수가 추정된다.

각 개체가 최대 확률로 해당 그룹에 할당될 때 EM 알고리듬 결과로 수렴하게 된다.

모형 선택시 BIC(Bayesian Information Criterion)를 계산하여
BIC 값이 최대가 되는 군집 개수를 최종 모형으로 선택할 수 있다.

$$BIC = 2\text{loglikelihood}(x, \theta_k^*) - (\text{no of parameters})\log n.$$

이와 같은 계산과정은 R 시스템에서는 mclust(Fraley and Raftery, 1998) 패키지로 제공되고 있으며 이를 활용하여 결과를 얻을 수 있다.

▶표 11.1 Yeung et al. (2001)가 제안 다섯 개의 모형

모형	설명
equal volume spherical model	$\Sigma_k = \lambda I$ 가장 제한된 모형으로 모수의 개수가 가장 적음
unequal volume spherical model	$\Sigma_k = \lambda_k I$ volume을 결정하는 λ가 군집마다 다르므로 서로 다른 volume의 구형군집들이 형성됨
unconstraint model	$\Sigma_k = \lambda_k D_k A_k D_k'$ 가장 일반적인 모형이지만 모수가 최대개수로 추정되어야 한다는 단점이 있음
elliptical model	$\Sigma_k = \lambda D A D'$ 각 군집은 타원형이지만 모두 동일한 shape, volume, orientation을 가짐
diagonal model	$\Sigma_k = \lambda_k B_k$ 여기서 B_k는 $\lvert B_k \rvert = 1$ 을 만족하는 대각행렬 기하학적으로 한 축에 일직선으로 세워진 타원형의 군집들과 일치하게 됨

≪예제 11.4≫ 1975년 미국 대도시의 강력범죄에 관한 자료에 대해 통계적 분석을 하고자한다.

[표 11.2] 미국 주요도시 강력범죄 발생률 자료(인구 100,000명당)

번호	도시명	murder	rape
1	Atlanta	16.5	24.8
2	Boston	4.2	13.3
3	Chicago	11.6	24.7
4	Dallas	18.9	34.2
5	Denver	6.9	41.5
6	Detroit	13.0	35.7
7	Hartford	2.5	8.8
8	Honolulu	3.6	12.7
9	Houston	16.8	26.6
10	Kansas City	10.8	43.2
11	Los Angeles	9.7	51.8
12	New Orleans	10.3	39.7
13	New York	9.4	19.4
14	Portland	5.9	23.0
15	Tucson	5.1	22.9
16	Washington	12.5	27.6

[프로그램 11.1] 범죄자료에 대한 계층적 군집분석

```
crime=read.csv("C:/data/c[redacted]header=T)
crime
attach(crime)
x=crime[, 3:4]
dx=round(dist(x), digits=2)   # 표 11.2 distance
matrix
dx
D2= dist(x, method ="manhattan")
D2
```

```
#############################
#          Hierarchical cluster ananlysis          #
#############################
hc1=hclust(dist(x)^2,meth          )   # 최단연결법
  plot(hc1,labels=city,hang      main="dandrogram:single")
#dendrogram

hc2=hclust(dist(x)^2,meth              )  # 최장연결법
  plot(hc2,labels=city,hang      main="complete linkage")

hc3=hclust(dist(x)^2,meth          )      # Ward 방법
  plot(hc3,labels=city,hang      main="Ward Method")

hc4=hclust(dist(x)^2,meth              )   # 평균연결법
  plot(hc4,labels=FALSE,har      main="Average linkage")
```

```
cl.num=2    # number of clusters
colnames(x)=c("murder","r[redacted]

hc1.result=cutree(hc2,k=c[redacted]
   plot(x, pch=hc1.result)
   text(x,labels=city, adj=0, cex=0.5, main="single")

hc2.result=cutree(hc2,k=c[redacted]
   plot(x, pch=hc2.result)
   text(x,labels=city, adj=0, cex=0.5, main="complete")

hc3.result=cutree(hc3,k=c[redacted]
   plot(x, pch=hc3.result)
   text(x,labels=city, adj=0, cex=0.5, main="Ward")
```

▶표 11.3 거리 행렬

(i) 유클리드 거리행렬

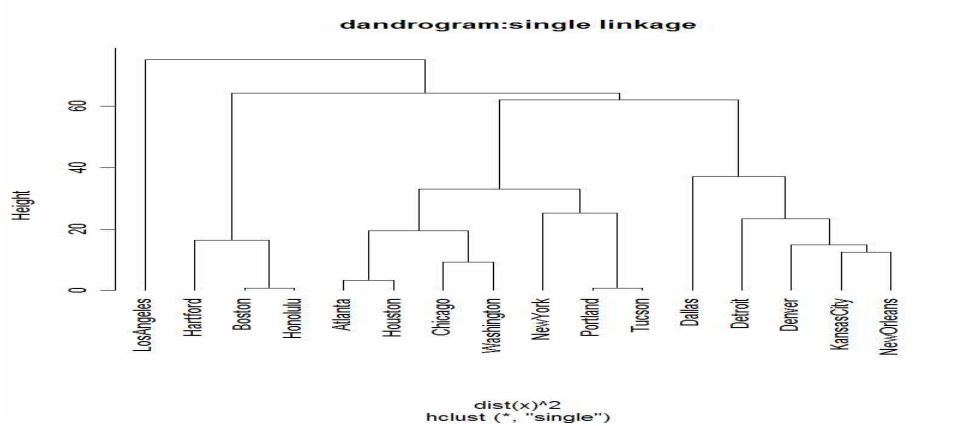
```
> dx=round(dist(x), digits=2)   # distance matrix
> dx
      1     2     3     4     5     6     7     8     9    10    11    12    13    14    15
2  16.84
3   4.90 13.59
4   9.70 25.55 11.98
5  19.26 28.33 17.45 14.05
6  11.45 24.07 11.09  6.09  8.42
7  21.26  4.81 18.32 30.23 32.99 28.88
8  17.69  0.85 14.42 26.39 28.99 24.85  4.05
9   1.82 18.32  5.54  7.88 17.89  9.86 22.83 19.17
10 19.26 30.62 18.52 12.11  4.25  7.82 35.39 31.34 17.65
11 27.84 38.89 27.17 19.86 10.67 16.43 43.60 39.57 26.18  8.67
12 16.14 27.10 15.06 10.21  3.85  4.83 31.87 27.82 14.62  3.54 12.11
13  8.92  8.02  5.74 17.59 22.24 16.69 12.65  8.86 10.32 23.84 32.40 20.32
14 10.75  9.85  5.95 17.16 18.53 14.55 14.60 10.55 11.48 20.79 29.05 17.27  5.02
15 11.56  9.64  6.74 17.84 18.69 15.04 14.34 10.31 12.27 21.09 29.26 17.59  5.54  0.81
16  4.88 16.53  3.04  9.19 14.99  8.12 21.29 17.36  4.41 15.69 24.36 12.30  8.77  8.04  8.77
```

(ii) 맨하탄 거리행렬

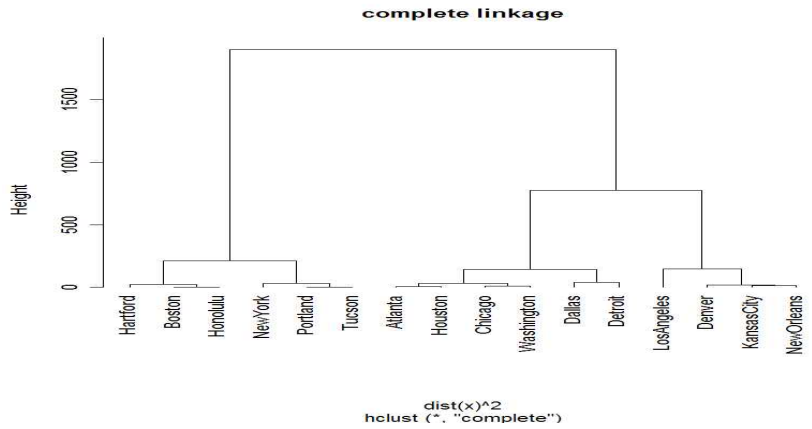
```
> D2 <- dist(x, method ="manhattan")
> D2
      1    2    3    4    5    6    7    8    9   10   11   12   13   14   15
2  23.8
3   5.0 18.8
4  11.8 35.6 16.8
5  26.3 30.9 21.5 19.3
6  14.4 31.2 12.4  7.4 11.9
7  30.0  6.2 25.0 41.8 37.1 37.4
8  25.0  1.2 20.0 36.8 32.1 32.4  5.0
9   2.1 25.9  7.1  9.7 24.8 12.9 32.1 27.1
10 24.1 36.5 19.3 17.1  5.6  9.7 42.7 37.7 22.6
11 33.8 44.0 29.0 26.8 13.1 19.4 50.2 45.2 32.3  9.7
12 21.1 32.5 16.3 14.1  5.2  6.7 38.7 33.7 19.6  4.0 12.7
13 12.5 11.3  7.5 24.3 24.6 19.9 17.5 12.5 14.6 25.2 32.7 21.2
14 12.4 11.4  7.4 24.2 19.5 19.8 17.6 12.6 14.5 25.1 32.6 21.1  7.1
15 13.3 10.5  8.3 25.1 20.4 20.7 16.7 11.7 15.4 26.0 33.5 22.0  7.8  0.9
16  6.8 22.6  3.8 13.0 19.5  8.6 28.8 23.8  5.3 17.3 27.0 14.3 11.3 11.2 12.1
```

[결과 11.1] 계층적 군집분석 결과

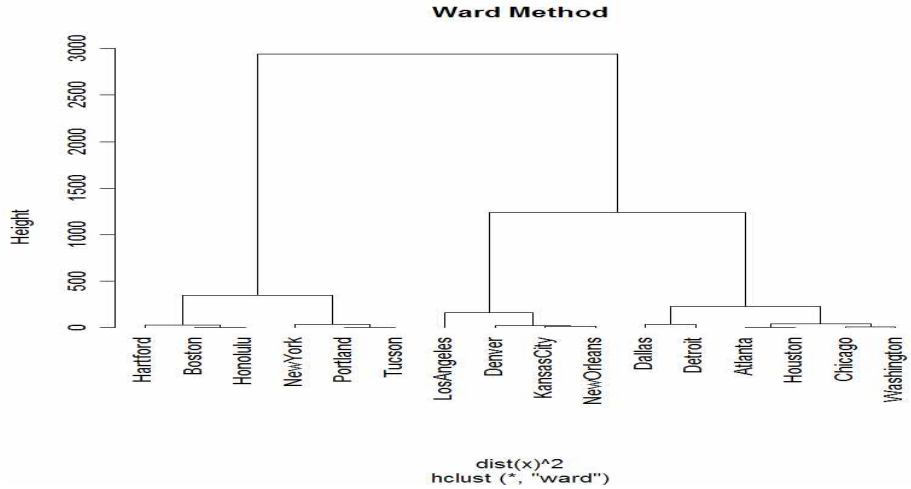
(1) 최단연결법을 이용한 경우 덴드로그램

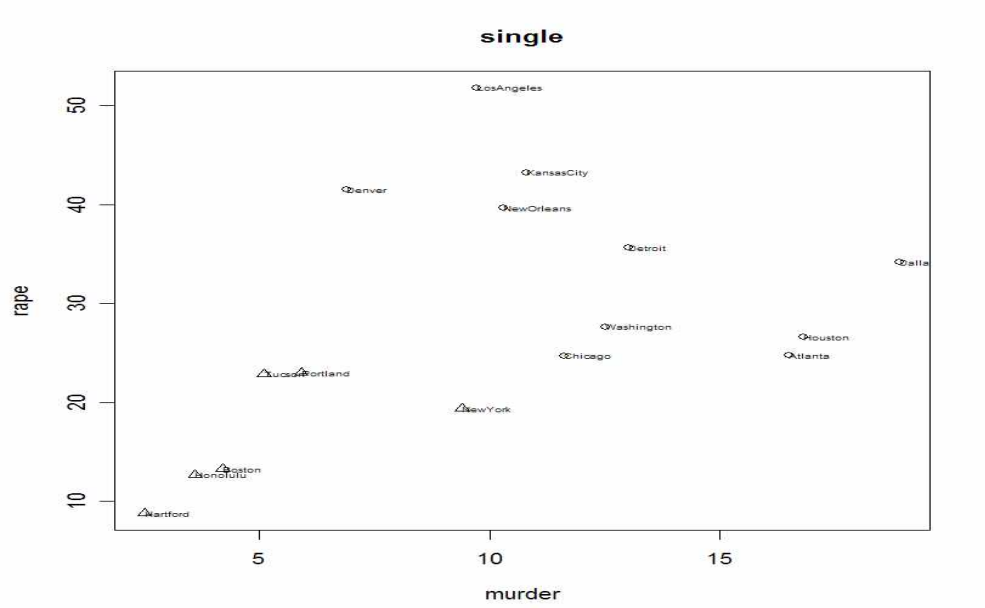


(2) 최장연결법을 이용한 경우 덴드로그램

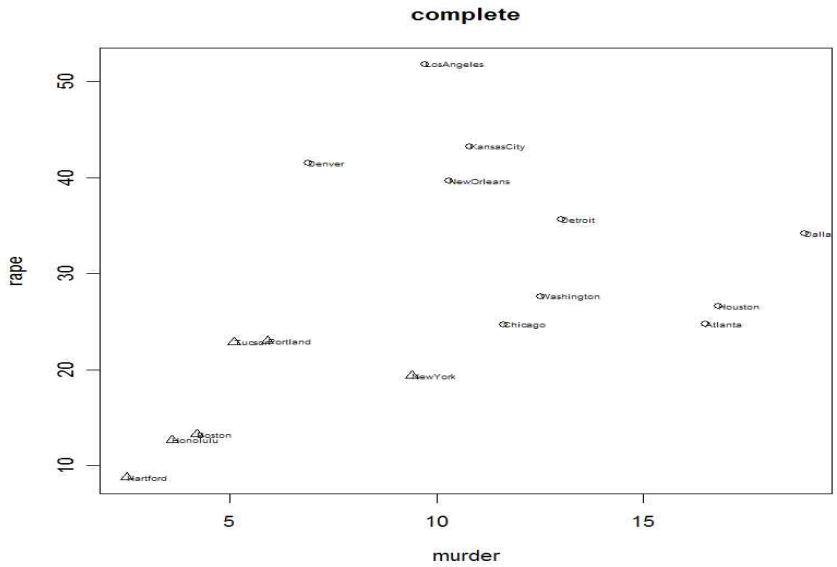


(3) Ward 방법을 이용한 경우 덴드로그램

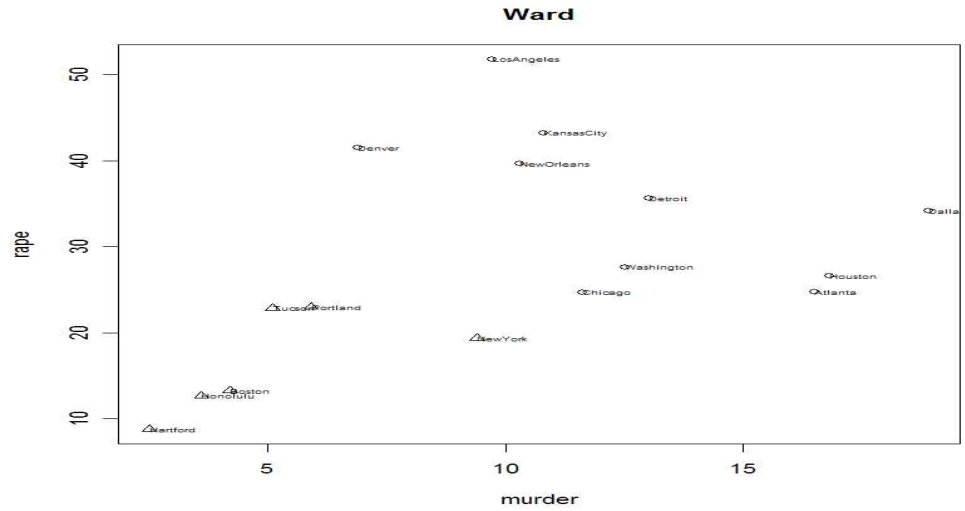




[그림 11.1] 최단연결법 이용 결과 군집



[그림 11.2] 최장연결법 이용 결과 군집



[그림 11.3] Ward 방법 이용 결과 군집

[프로그램 11.3] 범죄자료에 대한 K-means 군집분석

```
##############################
#          K-means  clustering                    #
##############################
crime_k=kmeans(x,centers=3)   # 3개 군집
attributes(crime_k)
crime_k$cluster

### grouping ###
clus=cbind(city,x,crime_k$clust
clus1=clus[(clus[,4]==1),]
clus1
clus2=clus[(clus[,4]==2),]
clus2
clus3=clus[(clus[,4]==3),]
clus3
kc=table(crime_k$cluster)   ## number of each cluster
 kc
plot(x, pch=crime_k$cluster,col=crime_k          main="K-means
clustering")
 text(x,labels=city, adj=0, cex=0.5)
ccent(x,crime_k$cluster)    # clusterwise info
```

[결과 11.3] K-means 군집분석 수행결과

```
> clus=cbind(city,x,crime_k$clust
> clus1=clus[(clus[,4]==1),]
> clus1
         city murder rape crime_k$cluster
4      Dallas   18.9 34.2               1
5      Denver    6.9 41.5               1
6     Detroit   13.0 35.7               1
10 KansasCity   10.8 43.2               1
11 LosAngeles    9.7 51.8               1
12 NewOrleans   10.3 39.7               1
```

```
> clus2=clus[(clus[,4]==2),]
> clus2
          city murder rape crime_k$cluster
1      Atlanta   16.5 24.8               2
3      Chicago   11.6 24.7               2
9      Houston   16.8 26.6               2
13     NewYork    9.4 19.4               2
14    Portland    5.9 23.0               2
15      Tucson    5.1 22.9               2
16 Washington   12.5 27.6               2
```

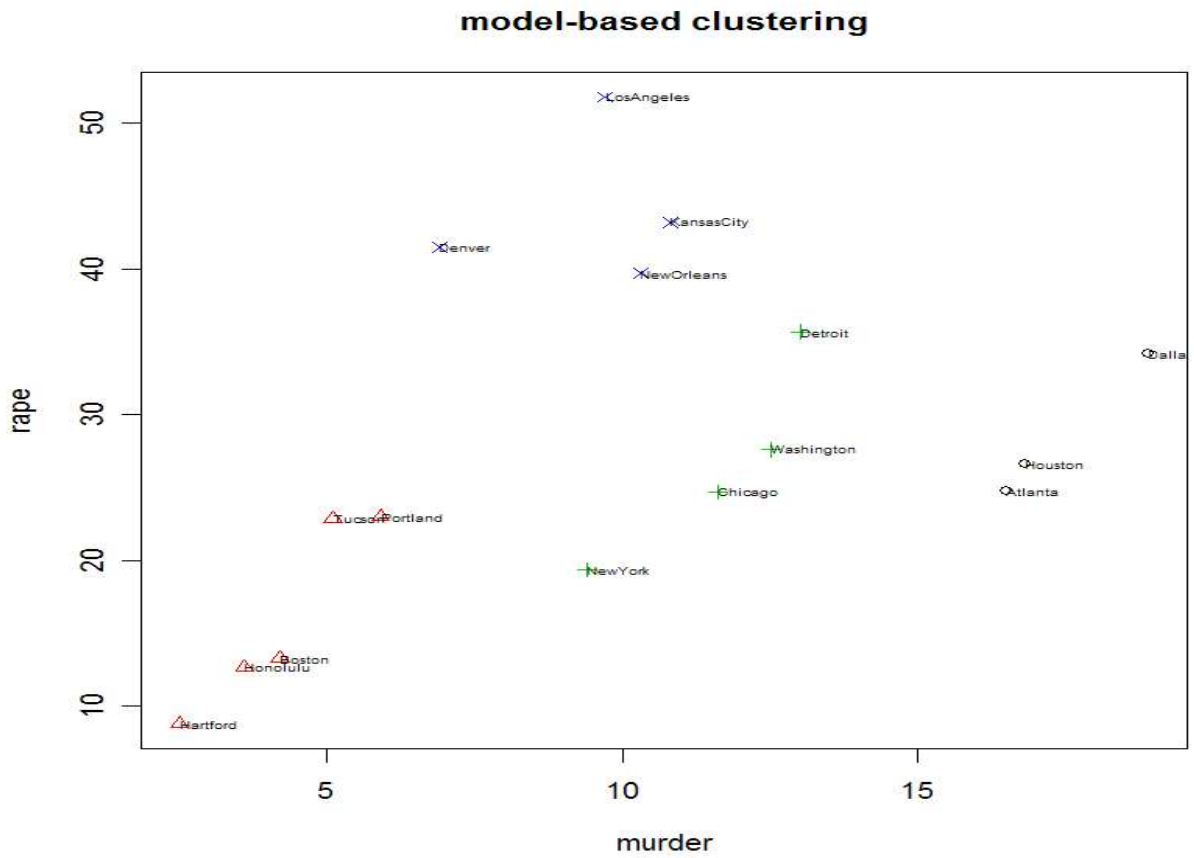
```
> clus3=clus[(clus[,4]==3),]
> clus3
      city murder rape crime_k$cluster
2   Boston    4.2 13.3               3
7 Hartford    2.5  8.8               3
8 Honolulu    3.6 12.7               3
> table(crime_k$cluster)    # number of each cluster
1 2 3
6 6 4
>plot(x,pch=crime_k$cluster,col[redacted]clustering")
>  text(x,labels=city, adj=0, cex=0.5)
> ccent(x,crime_k$cluster)    # clusterwise info
              1        2         3
murder 11.60000 11.11429  3.433333
rape   41.01667 24.14286 11.600000
```

[프로그램 11.4] 범죄자료에 대한 모형기반 군집분석

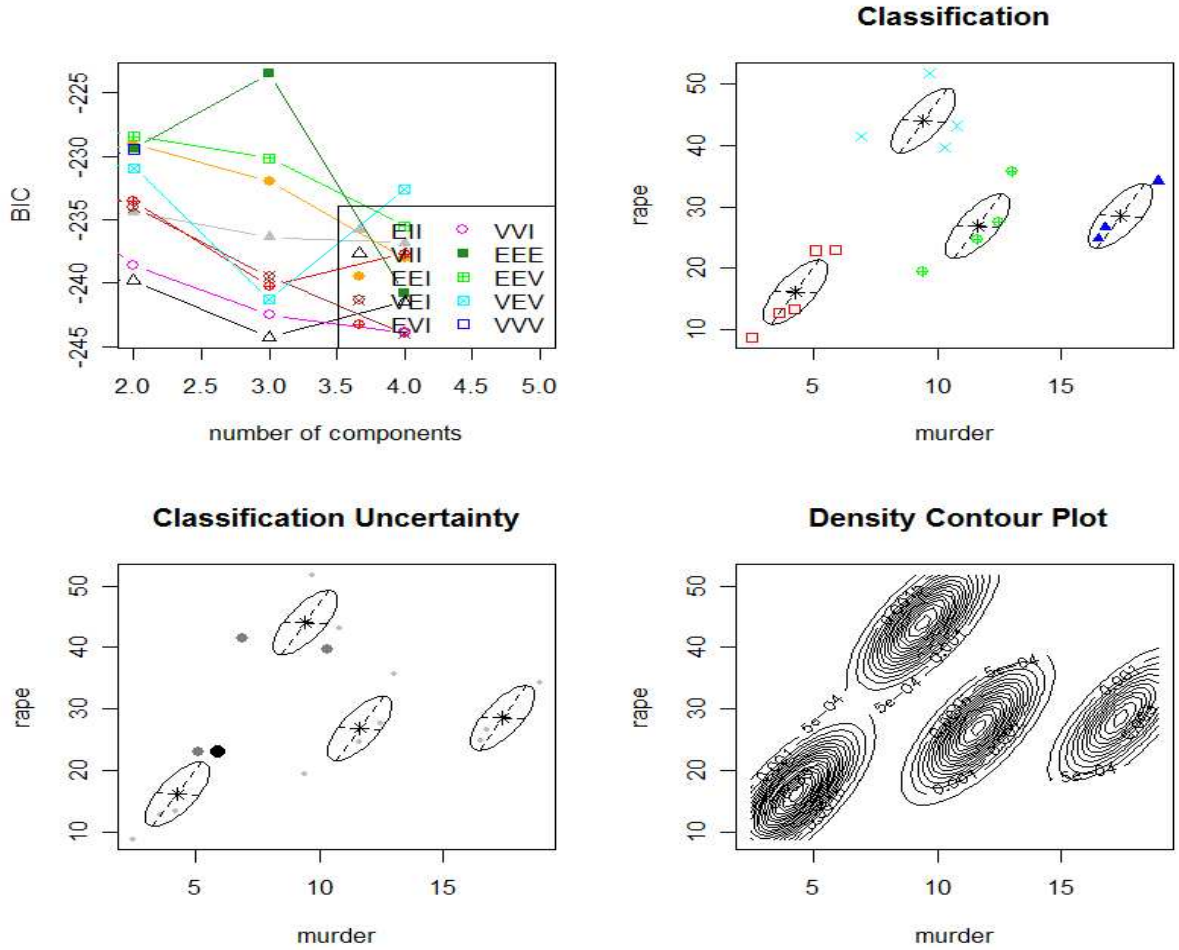
```
##############################
#          model-based  clustering                  #
##############################
library(mclust)
crime_mc=Mclust(x, 2:5)  # 군집개수 2~5개 사이 적절한 군집 결정
crime_mc
attributes(crime_mc)
crime_mc$classification  # 군집번호

mc=table(crime_mc$classificatio# number of each cluster
mc
plot(x, pch=crime_mc$classification,col
main="model-based clustering")
 text(x,labels=city, adj=0, cex=0.5)     # 그림 11.4

par(mfrow=c(2,2))
plot(crime_mc, data=crime[, 3:4])        # 그림 11.5
```

[결과 11.4] 모형기반 군집분석 수행결과

```
> crime_mc = Mclust(x, 2:5)
> crime_mc
 best model: EEE with 4 components
> crime_mc$classification
 [1] 1 2 3 1 4 3 2 2 1 4 4 4 3 2 2 3
> ccent(x, crime_mc$classification)
                1     2      3      4
murder 17.40000  4.26 11.625  9.425
rape   28.53333 16.14 26.850 44.050
> mc=table(crime_mc$classificatic[redacted]# number of each cluster
> mc
1 2 3 4
3 5 4 4
```



[그림 11.4] 모형기반 방법 이용 결과 군집



[그림 11.5] crime 데이터의 Mclust 적용후 군집 관련 그림

왼쪽 위: 공분산행렬 형태에 따른 BIC, 오른쪽 위: 군집 형성 그림

왼쪽 아래:공분산행렬에 따른 uncertainty, 오른쪽 아래:확률밀도함수 등고선