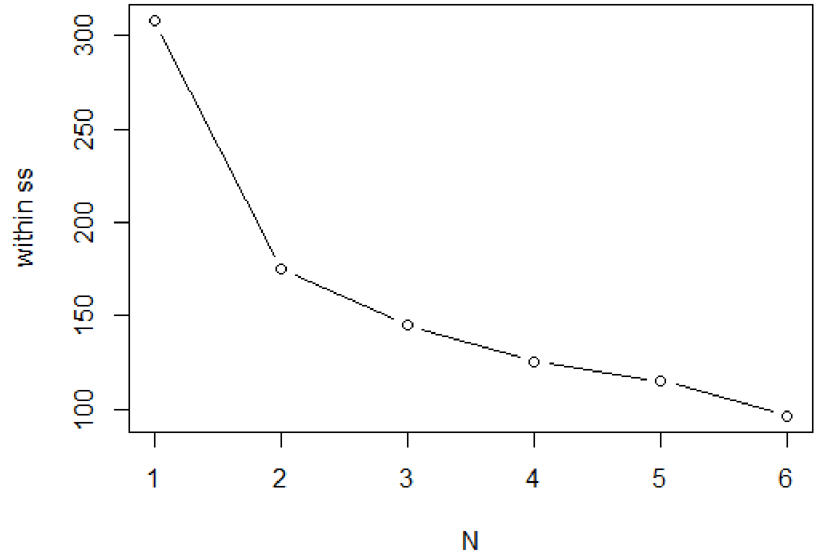
Problem(1)

Open the data:

```
crime <- as.data.frame(crime)
#rm the outlier#
pairs(crime)
which(crime$Murder>15)
rm.crime=crime[-24,]
sd=sapply(rm.crime,sd)
crime.s=sweep(rm.crime,2,sd,FUN="/")
#find the proper K#
n=nrow(crime.s)
wss=rep(0,6)
wss[1]=(n-1)*sum(sapply(crime.s,var))
for(i in 2:6){
  wss[i]=sum(kmeans(crime.s,centers=i)$withinss)
}
plot(1:6,wss,type="b",xlab="N",ylab="within  ss")
final=kmeans(crime.s,centers=3)
names(final)
result=list(names(final$cluster[final$cluster==1]),names(final$cluster[final$cluster==2]),names(final$cluster[final$cluster==3]))
#show the result:
names(result)=c("group1","group2","group3")
result
```



3 centers would be much better. And let K=3, we get:

```
> result
$group1
 [1] "ME" "NH" "VT" "PA" "IN" "WI" "MN" "IA" "ND" "SD" "NE" "KS" "VA" "WV" "K
Y" "MS" "AR" "MT" "ID"
[20] "WY" "UT" "HI"

$group2
 [1] "MA" "RI" "CT" "NY" "NJ" "OH" "IL" "MO" "DE" "MD" "NC" "SC" "GA" "TN" "A
L" "OK"

$group3
 [1] "MI" "FL" "LA" "TX" "CO" "NM" "AZ" "NV" "WA" "OR" "CA" "AK"
```

The result is different from the textbook, where it uses ranges to standardize variables.

Problem (2)

```
library(MASS)
```

```
library("KernSmooth")
```

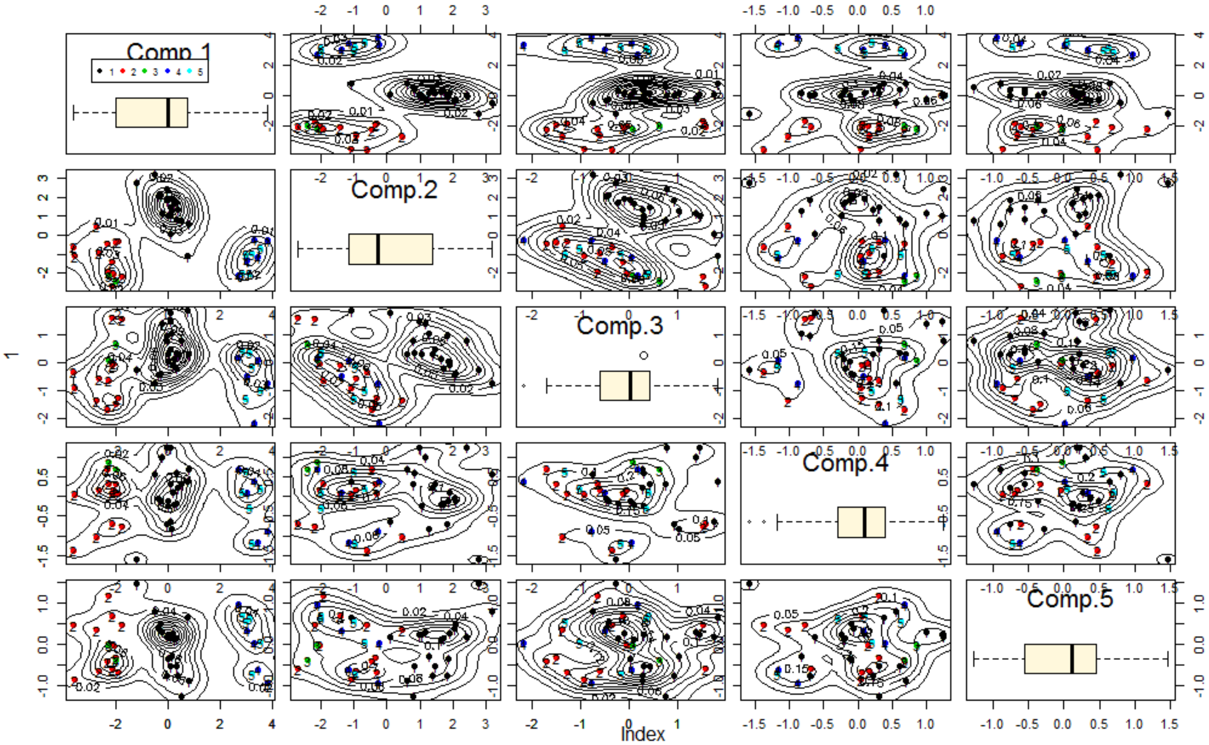
```
data(pottery)
```

```
fix(pottery)
```

```
names(pottery)
```

```
pr.out=princomp(pottery[,1:9],cor=T)
```

```
summary(pr.out,loadings=T)
names(pr.out)
score=as.data.frame(pr.out$scores[,1:5])
score$kiln=pottery$kiln
library(SciViews)#  use the function "panel.boxplot"
#draw the scatterplot matrix
pairs(score[,1:5],
    diag.panel =panel.boxplot,
    panel = function (x,y) {
     data <- data.frame(cbind(x,y))
     par(new = TRUE)
     plot(x,y,pch=16,col=score$kiln)
     text(x,y,labels=score$kiln)
     den <- bkde2D(data, bandwidth = sapply(data, dpik))
     contour(x = den$x1, y = den$x2,
          z = den$fhat, axes = FALSE,add=T)
    }
)
#to add a legend#
par(new=T)
plot(1,axes=F)
legend("topleft",pch=16,col=c(1,2,3,4,5),legend=c(1,2,3,4,5),ncol=5,cex=0.5,inset=0)
```



Problem(3)

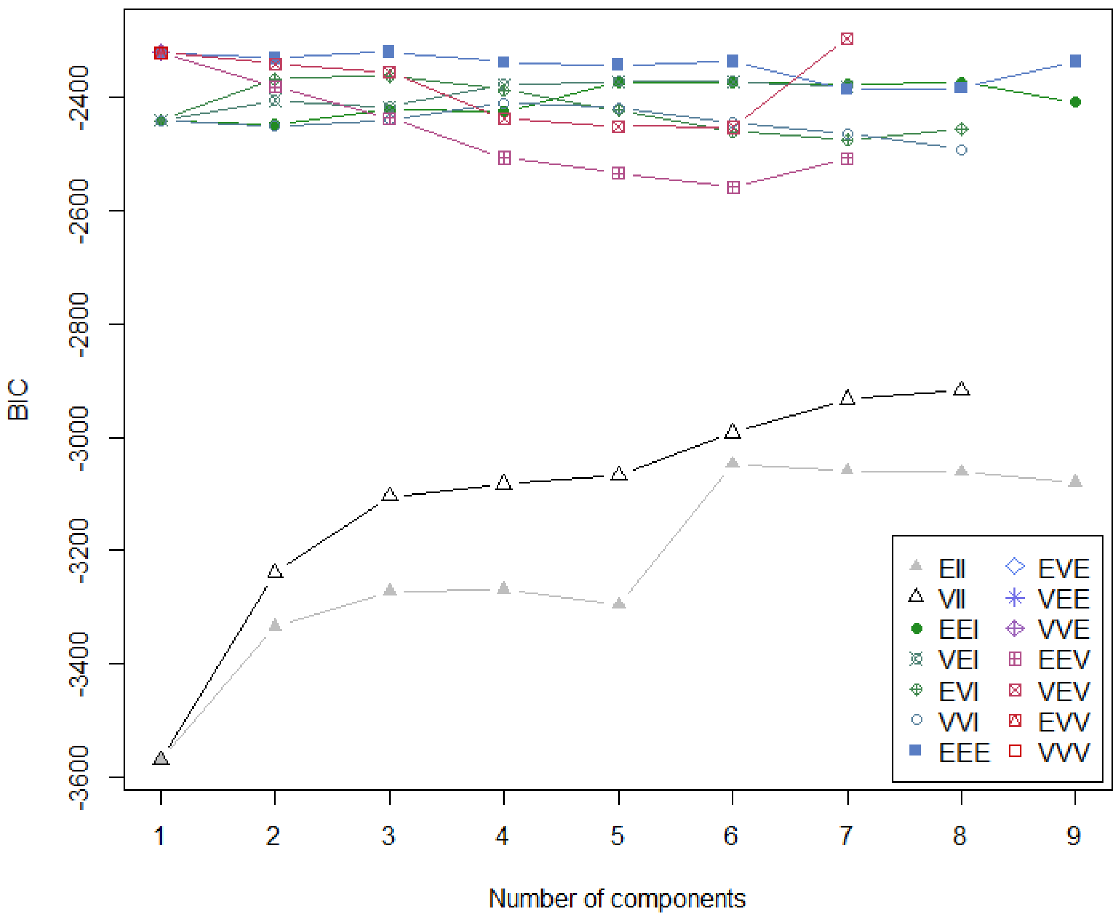
```
data("USairpollution")
data=USairpollution
r.data=data[,-1]
library(mclust)
#do model based cluster analysis by using Mclust() function:
r.data.m=Mclust(r.data)
#show the result of cluster
summary(r.data.m,parameters  = T)
# BIC for model selection and boxplot to show the distribution  of so2 in each cluster
plot(r.data.m)
names(r.data.m)
result=as.data.frame(r.data.m$classification)
data$group=result[,1]
```

```
names(data)
```

```
boxplot(data$SO2~data$group,ylab="level  of so2",xlab="group")
```

the result of cluster analysis by finite mixtures method:

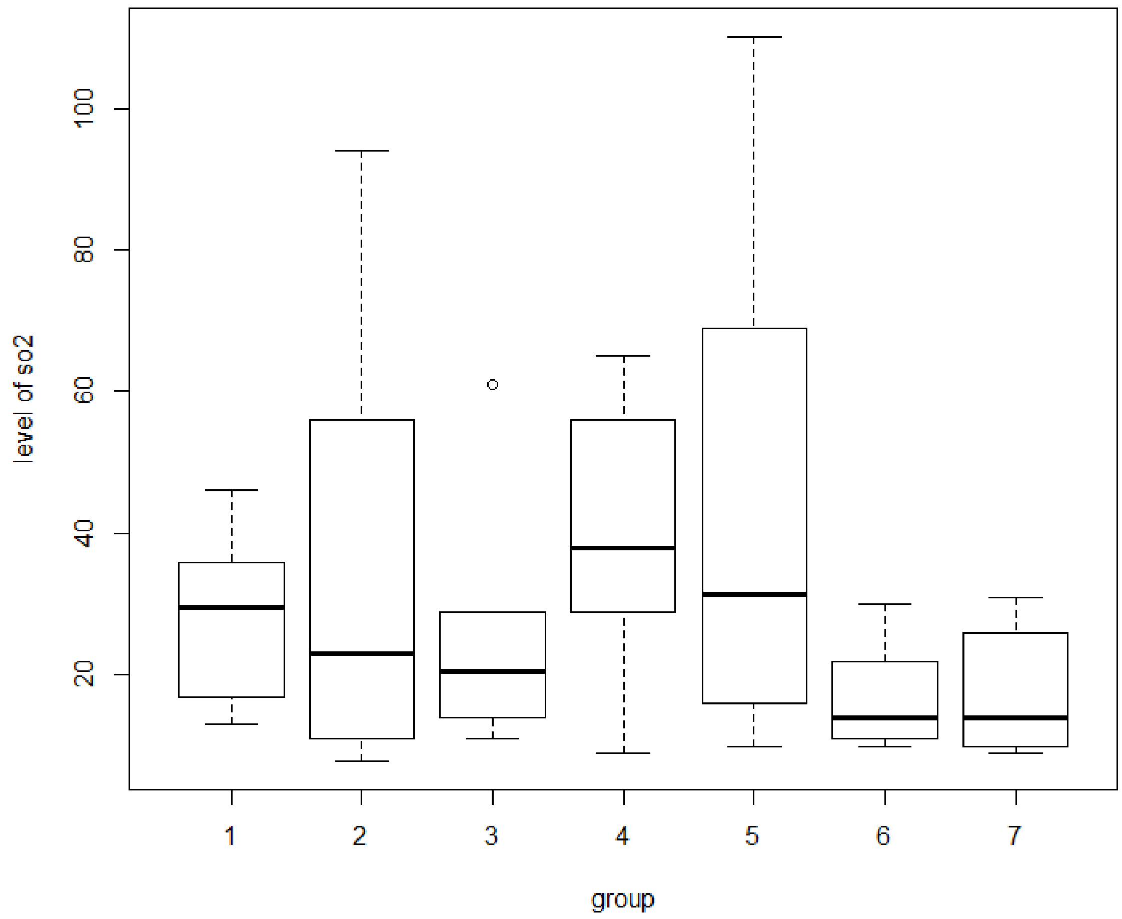
```
> summary(r.data.m,parameters = F)
----------------------------------------------------
Gaussian finite mixture model fitted by EM algorithm
----------------------------------------------------

Mclust VEV (ellipsoidal, equal shape) model with 7 components:

 log.likelihood  n  df       BIC       ICL
      -841.0466 41 165 -2294.833 -2294.833

Clustering table:
1 2 3 4 5 6 7
6 5 6 6 6 7 5
```



Model selection process induced by BIC criterion:

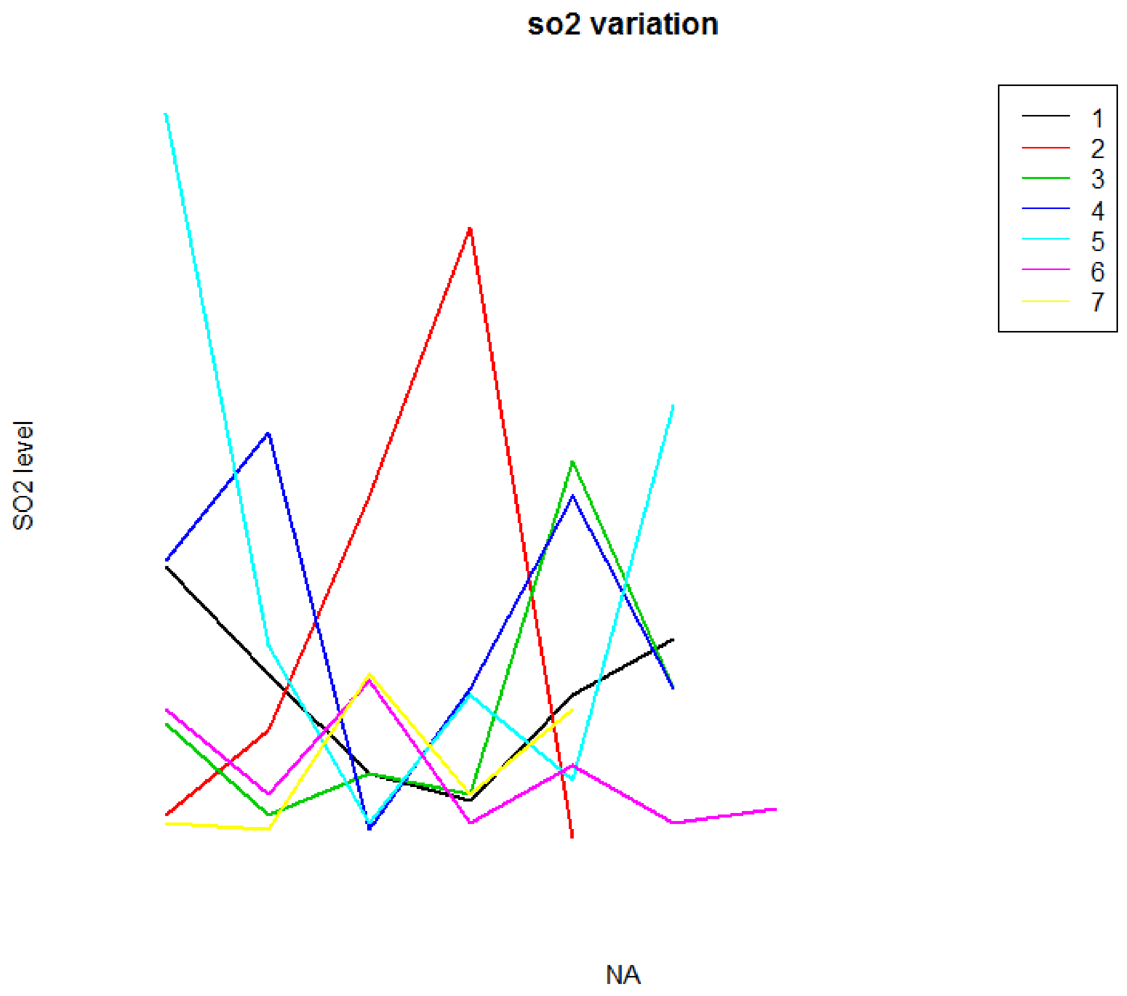
The BIC criterion selects model VEV and 7 clusters as the optimal solution.

How Sulphur dioxide concentration varies in the clusters?



The variation of so2 in each cluster:

```
plot(1,ylim=c(8,110),xlim=c(1,10),main="so2  variation",xlab="NA",ylab="SO2  level",axes=F)
for(i in unique(data$group)){
 ind=data[data$group==i,]
 lines(seq(length(ind$SO2)),ind$SO2,col=i,lwd=2,type="l")
}
legend("topright",col=unique(data$group),lty=1,legend=unique(data$group))
```



I don’t think the so2 variation in each cluster is a good indication of how well the cluster analysis goes.