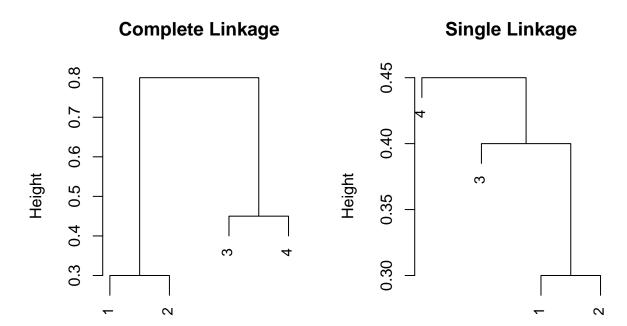
# STATS 415 hw11 solution

Weijing Tang April 7, 2018

## Question 1

## (a) & (b)



(c)

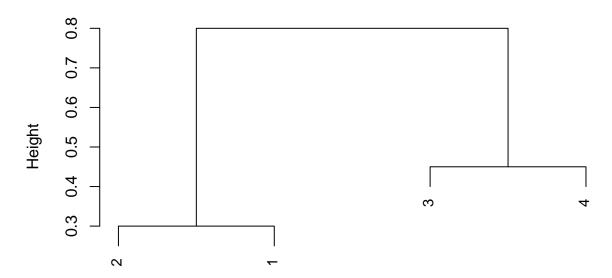
Cluster1: 1,2; Cluster2:3,4.

(d)

Cluster1:4; Cluster2:2,3,4.

(e)

## **Complete Linkage**



## Question 2

(a)

We load the package ISLR so to access the dataset USArrests and then use the hclust() function to perform hierarchical clustering.

```
library(ISLR)

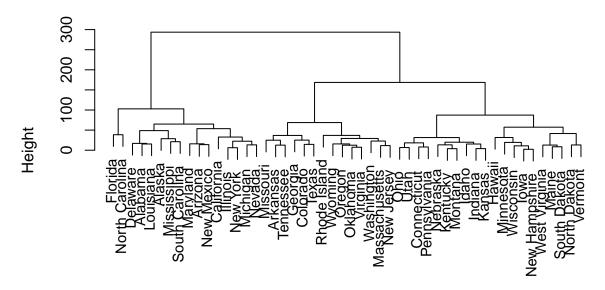
## Warning: package 'ISLR' was built under R version 3.3.3

library(cluster)

## Warning: package 'cluster' was built under R version 3.3.3

data("USArrests")
hc.complete=hclust(dist(USArrests),method='complete')
plot(hc.complete,main="Complete Linkage", xlab="", sub="", cex=.9)
```

# **Complete Linkage**

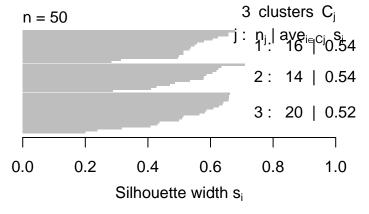


(b)

| cut | ree(hc.complete | ,3)         |                |               |                |
|-----|-----------------|-------------|----------------|---------------|----------------|
|     |                 |             |                |               |                |
| ##  | Alabama         | Alaska      | Arizona        | Arkansas      | California     |
| ##  | 1               | 1           | 1              | 2             | 1              |
| ##  | Colorado        | Connecticut | Delaware       | Florida       | Georgia        |
| ##  | 2               | 3           | 1              | 1             | 2              |
| ##  | Hawaii          | Idaho       | Illinois       | Indiana       | Iowa           |
| ##  | 3               | 3           | 1              | 3             | 3              |
| ##  | Kansas          | Kentucky    | Louisiana      | Maine         | Maryland       |
| ##  | 3               | 3           | 1              | 3             | 1              |
| ##  | Massachusetts   | Michigan    | Minnesota      | Mississippi   | Missouri       |
| ##  | 2               | 1           | 3              | 1             | 2              |
| ##  | Montana         | Nebraska    | Nevada         | New Hampshire | New Jersey     |
| ##  | 3               | 3           | 1              | 3             | 2              |
| ##  | New Mexico      | New York    | North Carolina | North Dakota  | Ohio           |
| ##  | 1               | 1           | 1              | 3             | 3              |
| ##  | Oklahoma        | Oregon      | Pennsylvania   | Rhode Island  | South Carolina |
| ##  | 2               | 2           | 3              | 2             | 1              |
| ##  | South Dakota    | Tennessee   | Texas          | Utah          | Vermont        |
| ##  | 3               | 2           | 2              | 3             | 3              |
| ##  | Virginia        | Washington  | West Virginia  | Wisconsin     | Wyoming        |
| ##  | 2               | 2           | 3              | 3             | 2              |

```
names(which(cutree(hc.complete,1)==1))
    [1] "Alabama"
                           "Alaska"
                                             "Arizona"
                                                               "Arkansas"
##
                                             "Connecticut"
##
    [5] "California"
                          "Colorado"
                                                               "Delaware"
                                                               "Idaho"
   [9] "Florida"
                           "Georgia"
                                             "Hawaii"
                                                               "Kansas"
## [13] "Illinois"
                           "Indiana"
                                             "Iowa"
## [17]
       "Kentucky"
                           "Louisiana"
                                             "Maine"
                                                               "Maryland"
## [21] "Massachusetts"
                                                               "Mississippi"
                          "Michigan"
                                             "Minnesota"
## [25] "Missouri"
                          "Montana"
                                             "Nebraska"
                                                               "Nevada"
                                             "New Mexico"
                                                               "New York"
## [29] "New Hampshire"
                          "New Jersey"
## [33]
       "North Carolina"
                          "North Dakota"
                                             "Ohio"
                                                               "Oklahoma"
## [37]
        "Oregon"
                          "Pennsylvania"
                                             "Rhode Island"
                                                               "South Carolina"
## [41] "South Dakota"
                           "Tennessee"
                                             "Texas"
                                                               "Utah"
## [45] "Vermont"
                           "Virginia"
                                             "Washington"
                                                               "West Virginia"
## [49] "Wisconsin"
                          "Wyoming"
names(which(cutree(hc.complete,2)==2))
    [1] "Arkansas"
                         "Colorado"
                                           "Connecticut"
                                                            "Georgia"
    [5] "Hawaii"
                         "Idaho"
                                          "Indiana"
                                                            "Iowa"
##
##
   [9]
        "Kansas"
                         "Kentucky"
                                          "Maine"
                                                            "Massachusetts"
                                                            "Nebraska"
## [13] "Minnesota"
                         "Missouri"
                                          "Montana"
## [17] "New Hampshire"
                         "New Jersey"
                                          "North Dakota"
                                                            "Ohio"
  [21] "Oklahoma"
                         "Oregon"
                                          "Pennsylvania"
                                                            "Rhode Island"
  [25]
                                                            "Utah"
        "South Dakota"
                         "Tennessee"
                                          "Texas"
## [29] "Vermont"
                         "Virginia"
                                          "Washington"
                                                            "West Virginia"
## [33] "Wisconsin"
                         "Wyoming"
names(which(cutree(hc.complete,3)==3))
                         "Hawaii"
                                           "Idaho"
                                                            "Indiana"
##
    [1] "Connecticut"
##
    [5] "Iowa"
                         "Kansas"
                                          "Kentucky"
                                                            "Maine"
   [9] "Minnesota"
                         "Montana"
                                           "Nebraska"
                                                            "New Hampshire"
## [13] "North Dakota"
                         "Ohio"
                                                            "South Dakota"
                                           "Pennsylvania"
## [17] "Utah"
                         "Vermont"
                                          "West Virginia"
                                                           "Wisconsin"
plot(silhouette(cutree(hc.complete,3),dist(USArrests)))
```

### Silhouette plot of (x = cutree(hc.complet

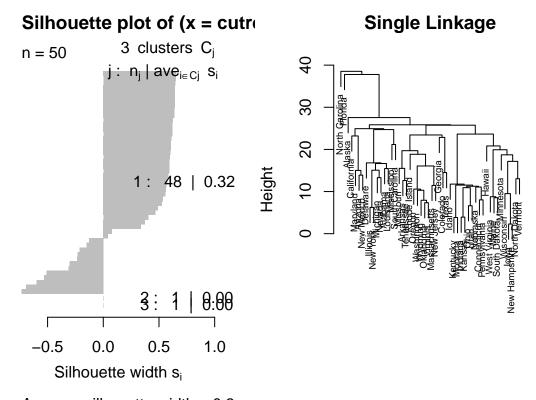


Average silhouette width: 0.53

The average silhouette width is 0.53. There is no observation with nagetive silhouette score which indicates a good clustering result.

(c)

```
hc.single=hclust(dist(USArrests),method='single')
par(mfrow=c(1,2))
plot(silhouette(cutree(hc.single,3),dist(USArrests)))
plot(hc.single,main="Single Linkage", xlab="", sub="", cex=.6)
```



### Average silhouette width: 0.3

"Wyoming"

names(which(cutree(hc.single,1)==1))

## [49] "Wisconsin"

### [1] "Alabama" "Alaska" "Arizona" "Arkansas" ## [5] "California" "Colorado" "Connecticut" "Delaware" [9] "Florida" "Idaho" ## "Georgia" "Hawaii" ## [13] "Illinois" "Indiana" "Iowa" "Kansas" [17] "Kentucky" "Maine" "Maryland" ## "Louisiana" [21] "Massachusetts" "Michigan" "Minnesota" "Mississippi" ## "Nevada" ## [25] "Missouri" "Montana" "Nebraska" ## [29] "New Hampshire" "New Jersey" "New Mexico" "New York" "Ohio" [33] "North Carolina" "North Dakota" "Oklahoma" [37] "Oregon" "Pennsylvania" "Rhode Island" "South Carolina" ## "Utah" "Tennessee" "Texas" [41] "South Dakota" "Vermont" "Virginia" "West Virginia" [45] "Washington"

```
names(which(cutree(hc.single,2)==2))
## [1] "North Carolina"
names(which(cutree(hc.single,3)==3))
```

### ## [1] "North Carolina"

There are 48 states are clustered in one group. Compared with complete linkage, single linkage tends to yield trailing clusters. The average silhouette width is 0.3 which is much poorer than complete linkage.

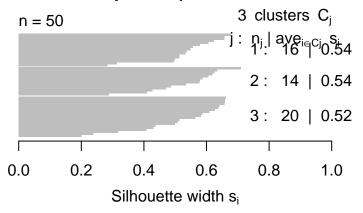
(d)

```
set.seed(1111)
km.out=kmeans(USArrests,3,nstart=20)
km.out$cluster
```

| ##                         | Alabama                                | Alaska                        | Arizona                                       | Arkansas                                    | California                            |
|----------------------------|--|-------------------------------|---|---|---------------------------------------|
| ##                         | 1                                      | 1                             | 1   | 2   | 1                                     |
| ##                         | Colorado                               | Connecticut                   | Delaware                                      | Florida                                     | Georgia                               |
| ##                         | 2                                      | 3                             | 1   | 1   | 2                                     |
| ##                         | Hawaii                                 | Idaho                         | Illinois                                      | Indiana                                     | Iowa                                  |
| ##                         | 3                                      | 3                             | 1   | 3   | 3                                     |
| ##                         | Kansas                                 | Kentucky                      | Louisiana                                     | Maine                                       | Maryland                              |
| ##                         | 3                                      | 3                             | 1   | 3   | 1                                     |
| ##                         | Massachusetts                          | Michigan                      | Minnesota                                     | Mississippi                                 | Missouri                              |
| ##                         | 2                                      | 1                             | 3   | 1   | 2                                     |
|                            | 34 .                                   | Nahaa alaa                    | Nevada  | New Hampshire                               | New Jersey                            |
| ##                         | Montana                                | Nebraska                      | Nevaua  | Mew Hampshire                               | New Jersey                            |
| ##<br>##                   | Montana<br>3                           | Nebraska<br>3                 | nevada<br>1                                   | New Hampshille                              | New Jersey                            |
|                            |  | 3                             | North Carolina                                | North Dakota                                | New Jersey<br>2<br>Ohio               |
| ##                         | 3                                      | 3                             | 1   | 3   | 2                                     |
| ##                         | 3                                      | 3                             | North Carolina 1                              | 3<br>North Dakota<br>3                      | 2                                     |
| ##<br>##<br>##             | 3<br>New Mexico<br>1                   | 3<br>New York<br>1            | North Carolina<br>1                           | 3<br>North Dakota<br>3                      | 2<br>Ohio<br>3                        |
| ##<br>##<br>##             | 3<br>New Mexico<br>1                   | 3<br>New York<br>1            | North Carolina<br>1                           | 3<br>North Dakota<br>3                      | 2<br>Ohio<br>3                        |
| ##<br>##<br>##<br>##       | 3<br>New Mexico<br>1<br>Oklahoma<br>2  | New York 1 Oregon 2           | 1<br>North Carolina<br>1<br>Pennsylvania<br>3 | 3<br>North Dakota<br>3<br>Rhode Island<br>2 | 2<br>Ohio<br>3<br>South Carolina<br>1 |
| ##<br>##<br>##<br>##<br>## | 3 New Mexico 1 Oklahoma 2 South Dakota | New York 1 Oregon 2 Tennessee | 1<br>North Carolina<br>1<br>Pennsylvania<br>3 | 3<br>North Dakota<br>3<br>Rhode Island<br>2 | 2<br>Ohio<br>3<br>South Carolina<br>1 |

```
par(mfrow=c(1,1))
plot(silhouette(km.out$cluster,dist(USArrests)))
```

### Silhouette plot of (x = km.out cluster, di



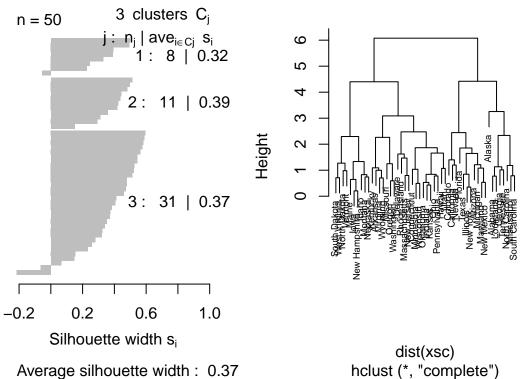
Average silhouette width: 0.53

We set 20 random initial cluster assignments and report the best result. Kmeans gives us a similar clustering result as hierarchical clustering with complete linkage.

(e)

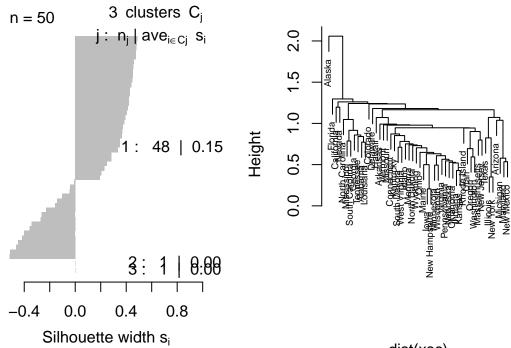
```
# hierarchical clustering
xsc=scale(USArrests)
hc.complete.sd=hclust(dist(xsc),method="complete")
hc.single.sd=hclust(dist(xsc),method='single')
par(mfrow=c(1,2))
plot(silhouette(cutree(hc.complete.sd,3),dist(xsc)))
plot(hc.complete.sd,main="Scaled Features and Complete Linkage",cex = 0.6)
```

## Silhouette plot of (x = cutr\Scaled Features and Complete Linl



plot(silhouette(cutree(hc.single.sd,3),dist(xsc)))
plot(hc.single.sd,main="Scaled Features and Single Linkage", cex=.6)

## Silhouette plot of (x = cutre Scaled Features and Single Linka



Average silhouette width: 0.15 dist(xsc) hclust (\*, "single")

| cut | ree(hc.complete. | sd,3)       |                |               |                |  |
|-----|------------------|-------------|----------------|---------------|----------------|--|
| ##  | Alabama          | Alaska      | Arizona        | Arkansas      | California     |  |
|     | ALADAIIA         | HIASKA      | AI 12011a      | Arkansas<br>3 | Calliolila     |  |
| ##  | 1                | 1           | _              | · ·           | 2              |  |
| ##  |                  | Connecticut | Delaware       |               | Georgia        |  |
| ##  | 2                | 3           | 3              | 2             | 1              |  |
| ##  | Hawaii           | Idaho       | Illinois       | Indiana       |                |  |
| ##  | 3                | 3           | 2              | 3             | 3              |  |
| ##  | Kansas           | Kentucky    | Louisiana      | Maine         | Maryland       |  |
| ##  | 3                | 3           | 1              | 3             | 2              |  |
| ##  | Massachusetts    | Michigan    | Minnesota      | Mississippi   | Missouri       |  |
| ##  | 3                | 2           | 3              | 1             | 3              |  |
| ##  | Montana          | Nebraska    | Nevada         | New Hampshire | New Jersey     |  |
| ##  | 3                | 3           | 2              | 3             | 3              |  |
| ##  | New Mexico       | New York    | North Carolina | North Dakota  | Ohio           |  |
| ##  | 2                | 2           | 1              | 3             | 3              |  |
| ##  | Oklahoma         | Oregon      | Pennsylvania   | Rhode Island  | South Carolina |  |
| ##  | 3                | 3           | 3              | 3             | 1              |  |
| ##  | South Dakota     | Tennessee   | •              | Utah          | Vermont        |  |
| ##  | 30utii Dakota    | 16111162266 | 1exas          | 0 tan<br>3    | vermont 3      |  |
|     | •                | I           | =              | · ·           | •              |  |
| ##  | •                | wasnington  | West Virginia  |               |                |  |
| ##  | 3                | 3           | 3              | 3             | 3              |  |
| cut | ree(hc.single.sd | 1,3)        |                |               |                |  |
| ##  | Alabama          | Alaska      | Arizona        | Arkansas      | California     |  |
|     |                  |             | Arizona        |               | Caillornia     |  |
| ##  | 1                | 2           | 1              | 1             | 1              |  |

```
##
         Colorado
                      Connecticut
                                        Delaware
                                                         Florida
                                                                        Georgia
##
                                                                               1
                1
##
           Hawaii
                            Idaho
                                        Illinois
                                                         Indiana
                                                                            Iowa
##
                                                                               1
           Kansas
##
                         Kentucky
                                       Louisiana
                                                           Maine
                                                                       Maryland
##
                                1
##
    Massachusetts
                        Michigan
                                       Minnesota
                                                     Mississippi
                                                                        Missouri
##
                                1
                                               1
##
          Montana
                        Nebraska
                                          Nevada
                                                   New Hampshire
                                                                      New Jersey
##
                1
                               1
                                               1
                                                                               1
                         New York North Carolina
##
       New Mexico
                                                    North Dakota
                                                                            Ohio
##
                               1
##
         Oklahoma
                           Oregon
                                    Pennsylvania
                                                    Rhode Island South Carolina
##
                                                               1
##
     South Dakota
                        Tennessee
                                           Texas
                                                            Utah
                                                                         Vermont
##
##
         Virginia
                       Washington West Virginia
                                                       Wisconsin
                                                                         Wyoming
##
```

### # kmeans

set.seed(1111)

km.out.sd=kmeans(xsc,3,nstart=20)

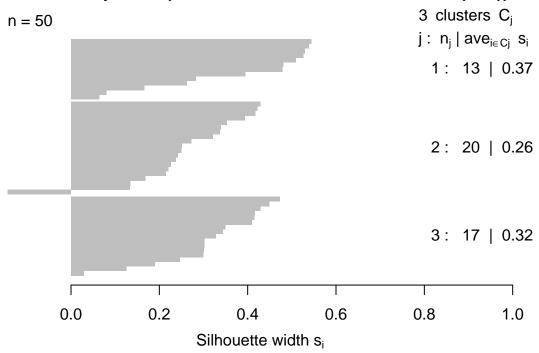
km.out.sd\$cluster

| ##       | Alabama                       | Alaska                        | Arizona                         | Arkansas                       | California       |
|----------|-------------------------------|-------------------------------|---------------------------------|--------------------------------|------------------|
| ##       | 2                             | 2                             | 2                               | 3                              | 2                |
| ##       | Colorado                      | Connecticut                   | Delaware                        | Florida                        | Georgia          |
| ##       | 2                             | 3                             | 3                               | 2                              | 2                |
| ##       | Hawaii                        | Idaho                         | Illinois                        | Indiana                        | Iowa             |
| ##       | 3                             | 1                             | 2                               | 3                              | 1                |
| ##       | Kansas                        | Kentucky                      | Louisiana                       | Maine                          | Maryland         |
| ##       | 3                             | 1                             | 2                               | 1                              | 2                |
| ##       | Massachusetts                 | Michigan                      | Minnesota                       | Mississippi                    | Missouri         |
| ##       | 3                             | 2                             | 1                               | 2                              | 2                |
| ##       | Montana                       | Nebraska                      | Nevada                          | New Hampshire                  | New Jersey       |
| ##       | 1                             | 1                             | 2                               | 1                              | 3                |
| ##       | New Mexico                    | New York                      | North Carolina                  | North Dakota                   | Ohio             |
| ##       | •                             |                               |                                 |                                |                  |
|          | 2                             | 2                             | 2                               | 1                              | 3                |
| ##       | 2<br>Oklahoma                 | 2<br>Oregon                   | 2<br>Pennsylvania               | 1<br>Rhode Island              | 3 South Carolina |
| ##<br>## | 0klahoma<br>3                 | 2<br>Oregon<br>3              | 2<br>Pennsylvania<br>3          | 1<br>Rhode Island<br>3         | J                |
|          | 0klahoma<br>3<br>South Dakota | 2<br>Oregon<br>3<br>Tennessee | 2<br>Pennsylvania<br>3<br>Texas | 1<br>Rhode Island<br>3<br>Utah | South Carolina 2 |
| ##       | 3                             | 3                             | 3                               | 3                              | South Carolina 2 |
| ##<br>## | 3                             | 3<br>Tennessee<br>2           | 3                               | 3                              | South Carolina 2 |

```
par(mfrow=c(1,1))
```

plot(silhouette(km.out.sd\$cluster,dist(xsc)))

## Silhouette plot of (x = km.out.sd\$cluster, dist = dist(xsc))



Average silhouette width: 0.31

The average silhouette width of scaled data are smaller in all three cases than those of unscaled data.

(f)

##

3 13 0 7

We use the table() function to compare the results

```
# compare HC
hc = cutree(hc.complete,3)
hc.sd = cutree(hc.complete.sd,3)
table(hc,hc.sd)
##
     hc.sd
## hc
        1
          2 3
##
       6 9 1
     1
       2 2 10
     3 0 0 20
##
km = km.out$cluster
km.sd = km.out.sd$cluster
table(km,km.sd)
##
      km.sd
          2
## km
        1
             3
##
       0 15
              1
     2 0
##
          5 9
```

We see that the three clusters obtained using original variables and scaled variables are somewhat different. For hierarchical clustering, cluster 3 obtained using scaled variables mainly combines the cluster 2 and 3 obtained using original variables. Regardless of permutation of cluster number, the difference between k-means clustering results is smaller than hierarchical clustering.

Note that the variables "Murder", "Assault" and "Rape" are the number of corresponding arrests per 100,000 while the variable "UrbanPop" is the percentage of urban population. These variables have different units but all represent a ratio. The range of them don't differ too much. The results from unscaled data also give better clustering result. Therefore, we prefer not to scaling data.