STAT 501 Homework 8

Multinomial

April 27, 2018

First we standardized the data and use the standardized data in the clustering.

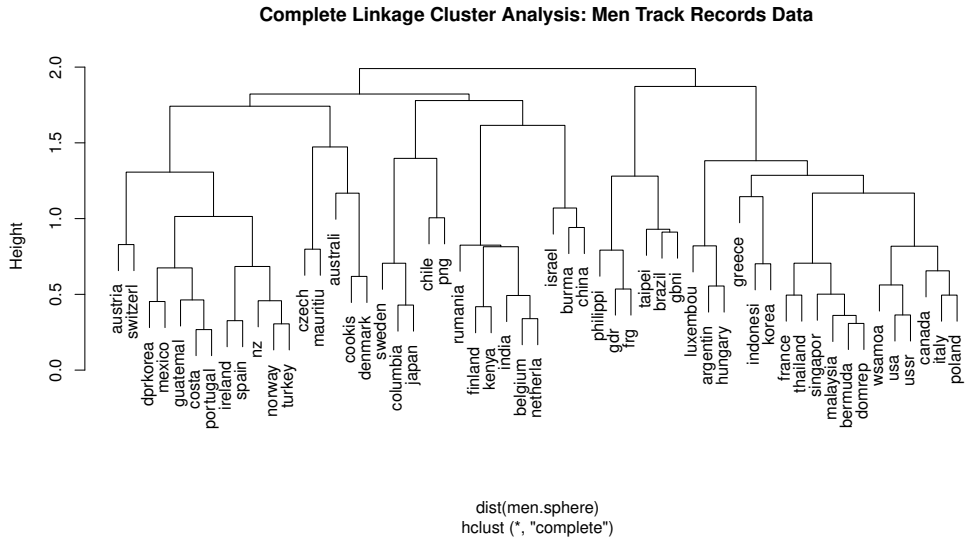
```
women <- read.table("women-track-records.dat",
                    header=F, col.names=c("x1", "x2", "x3", "x4", "x5","x6","x7",
                     ↪  "Country"))

men <- read.table("men-track-records.dat",
                    header=F, col.names=c("x1", "x2", "x3", "x4", "x5","x6","x7", "x8",
                     ↪  "Country"))

women.country <- women$Country
men.country <- men$Country

women <- women[,-ncol(women)]
men <- men[,-ncol(men)]

# standardize the data
women <- scale(women)
men <- scale(men)
```

- Hierarchical Clustering

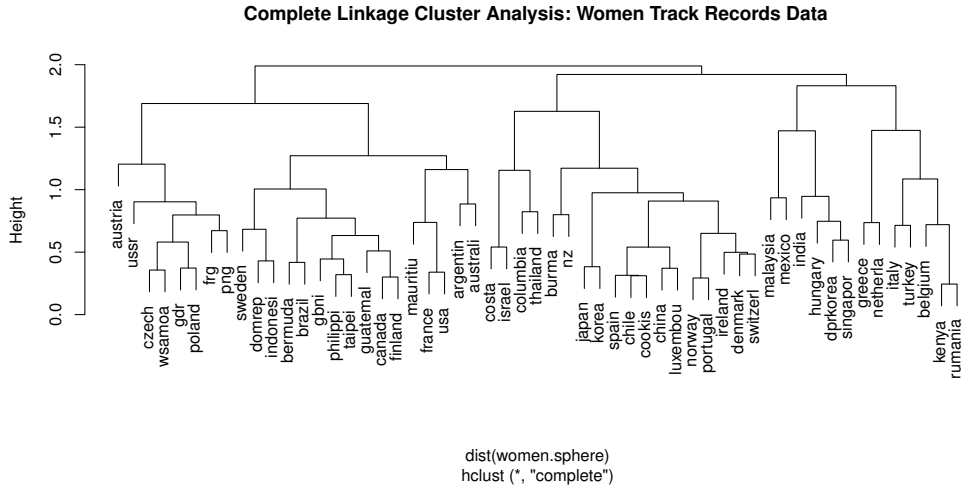
  Before doing hierarchical clustering, we mapped the data to the sphere. In this way the Euclidian distance is just the correlation similarity distance for the standardized the data. We adopted complete linkage in hierarchical clustering for men and women. The hierarchical tree is shown in Figure 1 and we display the clustering result with 3 clusters in Figure 2 using the 2 LDs by LDA.

```
#  Sphere the data for correlation similarity distance
men.sphere<-t(apply(men,1,FUN= function(x) (x - mean(x))/sqrt(sum((x - mean(x))^2))))
women.sphere<-t(apply(women,1,FUN= function(x) (x - mean(x))/sqrt(sum((x -
 ↪  mean(x))^2))))

#  Use complete linkage, this is also the default method
hc.men <- hclust(dist(men.sphere),method="complete")
plot(hc.men, label = men.country,
main = "Complete Linkage Cluster Analysis: Men Track Records Data")

hc.women <- hclust(dist(women.sphere),method="complete")
plot(hc.women, label = women.country,
     main = "Complete Linkage Cluster Analysis: Women Track Records Data")

#  Use canonical discriminants to display the clusters.
#  The first function computes linear canonical discriminants and
#  the second function is used to plot the computed scores.
library(MASS)
#  Compute canonical discriminant scores and display 2-dimensional
#  projections of the clusters

hc.men.lda <- lda(men, cutree(hc.men, 3))
plot(hc.men.lda)
```

```
23
24 hc.women.lda <- lda(women, cutree(hc.women, 3))
25 plot(hc.women.lda)
```

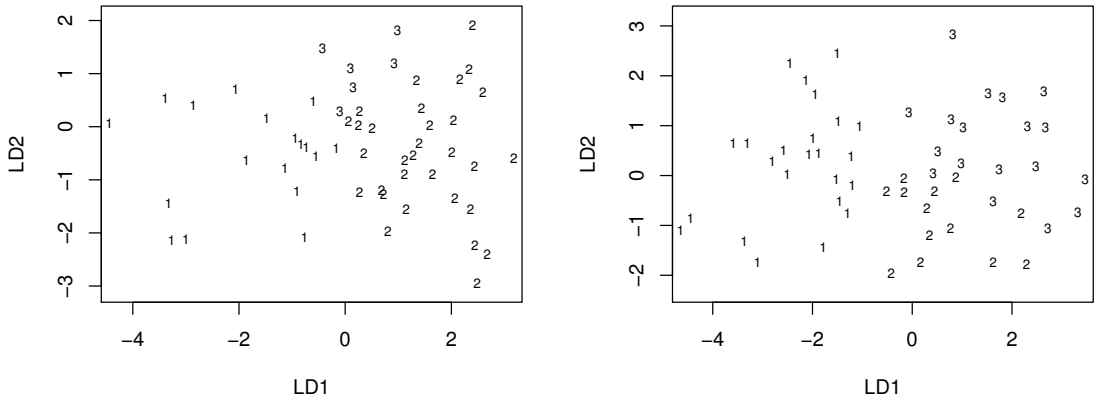


(a)



(b)

Figure 1: Hierarchical clustering tree



(a) men (b) women

Figure 2: Hierarchical clustering results

- k-means

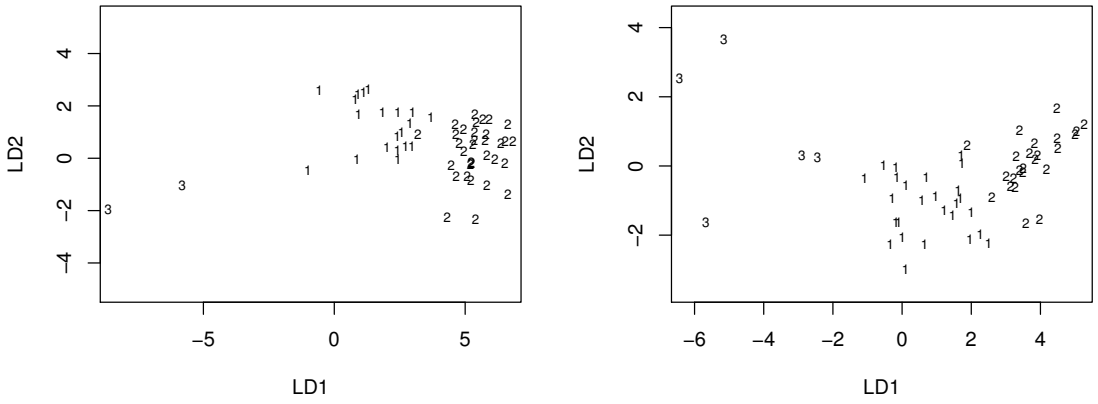
  We use the results of hierarchical clustering with ward linkage as our initial for k-means. Then we display the results from k-means with 3 clusters using the 2 LDs from LDA in Figure 3.

```
#  Compute K-means cluster analysis starting with results from hclust
kmnsinithcl <- function(x.data, nclus, ncut = nclus, hcl.tree)
{
  x.hcl <- hcl.tree
  x.cl <- cutree(x.hcl, k = ncut)
  data.x <- data.frame(x.data, cl = x.cl)
  means <- aggregate(. ~ cl, data = data.x, FUN = mean)
  return(kmeans(x.data,centers= means[, -1]))
}

hc.men2 <- hclust(dist(men.sphere),method="ward.D2")
km.men <- kmnsinithcl(men, nclus = 3, ncut = 3, hcl.tree = hc.men2)
km.men.lda <- lda(men, km.men$cluster)
plot(km.men.lda)

hc.women2 <- hclust(dist(women.sphere),method="ward.D2")
km.women <- kmnsinithcl(women, nclus = 3, ncut = 3, hcl.tree = hc.women2)
km.women.lda <- lda(women, km.women$cluster)
plot(km.women.lda)
```

- Model based clustering

  For model based clustering, we use BIC to pick the model and number of clusters. In Figure 4 we can see, for men, model VEE and 2 clusters is the best, while model VEV and 2 clusters for women is the best. The display of clustering is shown in Figure 5.

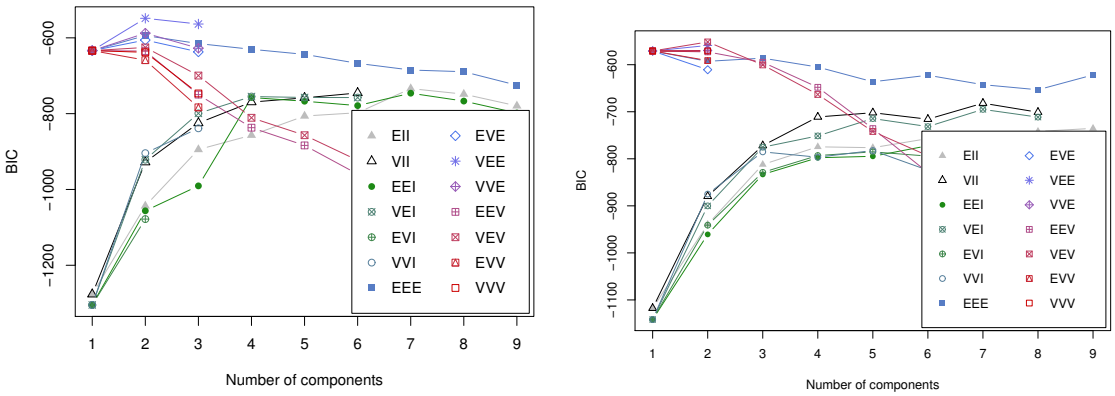
```
#  Model based clustering based on the assumption that
#  the data were sampled from a mixture of multivariate normal
#  distributions with common covariance matrix

library(mclust)

```



(a) men (b) women

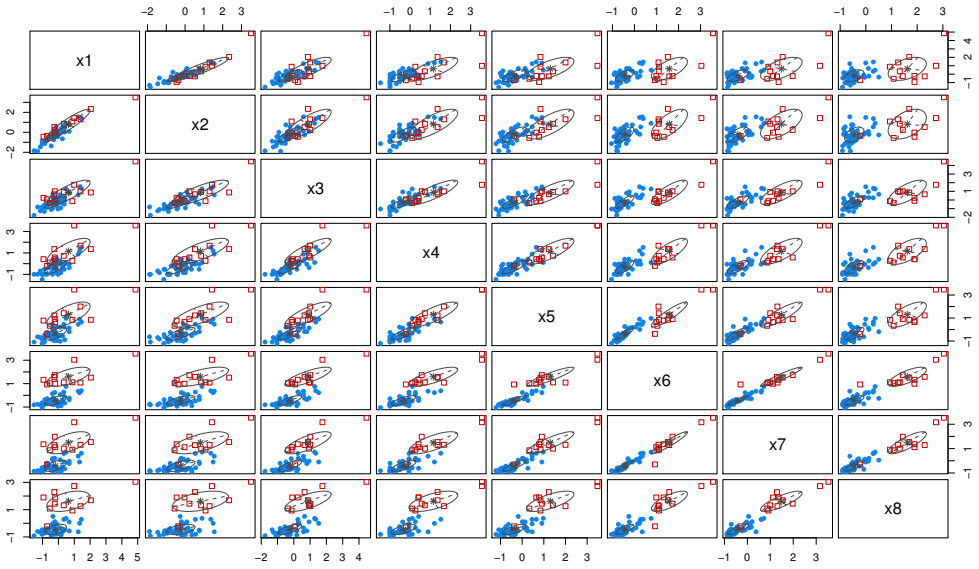
Figure 3: k-means clustering results

```
mcl.men <- Mclust(men)
plot(mcl.men$BIC)
plot.Mclust(mcl.men, what = "classification")

mcl.women <- Mclust(women)
plot(mcl.women$BIC)
plot.Mclust(mcl.women, what = "classification")
```

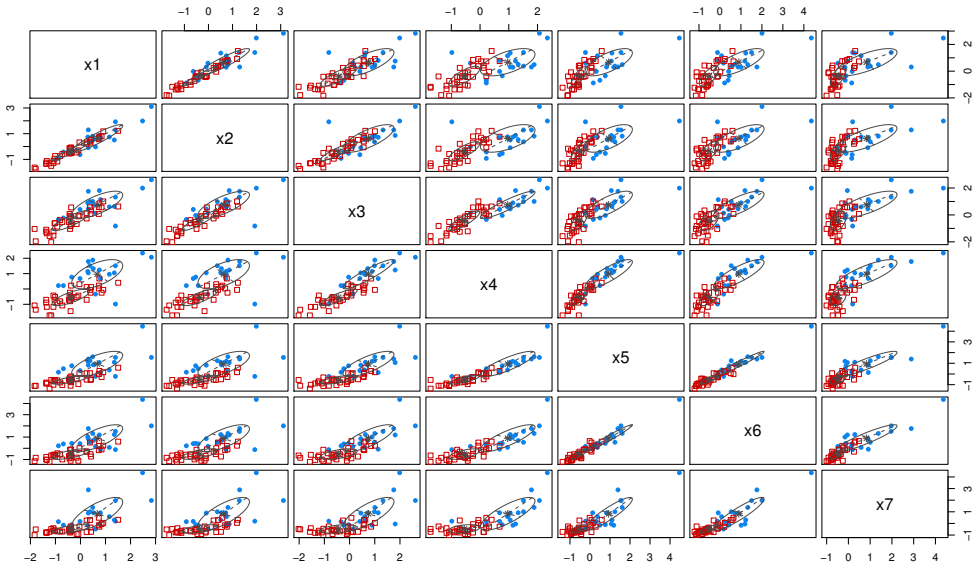


(a) men (b) women

Figure 4: BIC for model selection



(a) men



(b) women

Figure 5: Model based clustering results

Comment:

The clustering results using 3 methods shows some differences in the clusters patterns between men and women. However we can see from the results using 3 methods that there might be 2 clusters for both women and men.