Northeastern University
College of Science

Module 10 – Homework

Problem 1: (40 points)
Clustering analysis on the "CCND3 Cyclin D3" gene expression values of the Golub et al. (1999) data.

(a) Conduct hierarchical clustering using single linkage and Ward linkage. Plot the cluster dendrogram for both fit. Get two clusters from each of the methods. Use function table () to compare the clusters with the two patient groups ALL/AML. Which linkage function seems to work better here?

Answer)

```
> #(a)
> grep("CCND3 Cyclin D3",golub.gnames[,2])
[1] 1042
> data <- data.frame(golub[1042,])
> gol.fac <- factor(golub.cl,levels=0:1, labels= c("ALL","AML"))
> single_link <- hclust(dist(data, method="euclidian"), method = "single")
> plot(single_link, main = "Single linkage dendrogram", labels=gol.fac)
> ward <- hclust(dist(data, method = "euclidian"), method = "ward.D2")
> plot(ward, main = "Ward linkage dendrogram", labels=gol.fac)
> single_cluster <- cutree(single_link, k=2)
> table(single_cluster, gol.fac)
              gol.fac
single_cluster ALL AML
             1  27  10
             2   0   1
> ward_cluster <- cutree(ward, k=2)
> table(ward_cluster, gol.fac)
            gol.fac
ward_cluster ALL AML
           1  21   0
           2   6  11
```

Conclusion:

From above output, we can see that the ward.D2 method works better, as we see that cluster generated by this method

# have a more balanced distribution of the patients.

(b) Use *k*-means cluster analysis to get two clusters. Use table () to compare the two clusters with the two patient groups ALL/AML.

Answer)

```
> data <- data.frame(golub[1042,])
> gol.fac <- factor(golub.cl,levels=0:1, labels= c("ALL","AML"))
> K_CCND3 <- kmeans(data, centers = 2, nstart=10)
> table(K_CCND3$cluster, gol.fac)
   gol.fac
    ALL AML
  1  22    1
  2   5   10
```

(c) Which clustering approach (hierarchical versus k-means) produce the best matches to the two diagnose groups ALL/AML?

Answer)

By comparing with patients groups, k-means clustering produces best matches to the groups ALL and AML.

(d) Find the two cluster means from the k-means cluster analysis. Perform a bootstrap on the cluster means. Do the confidence intervals for the cluster means overlap? Which of these two-cluster means is estimated more accurately?

Answer)

```
> #(d)
> initial <-K_CCND3$centers
> n <- dim(data)[1]; nboot<-1000
> boot.cl <- matrix(NA,nrow=nboot,ncol = 4)
> for (i in 1:nboot){
+   dat.star <- data[sample(1:n,replace=TRUE),]
+   cl <- kmeans(dat.star, initial, nstart = 10)
+   boot.cl[i,] <- c(cl$centers[1], cl$centers[2])
+ }
> apply(boot.cl,2,mean)
[1] 2.0320243 0.7052078 2.0320243
[4] 0.7052078
> quantile(boot.cl[,1],c(0.025, 0.975))
    2.5%    97.5%
1.844151 2.199546
> quantile(boot.cl[,2],c(0.025, 0.975))
     2.5%     97.5%
0.2600586 1.0743602
```

```
> quantile(boot.cl[,3],c(0.025, 0.975))
    2.5%    97.5%
1.844151 2.199546
> quantile(boot.cl[,4],c(0.025, 0.975))
     2.5%     97.5%
0.2600586 1.0743602
```

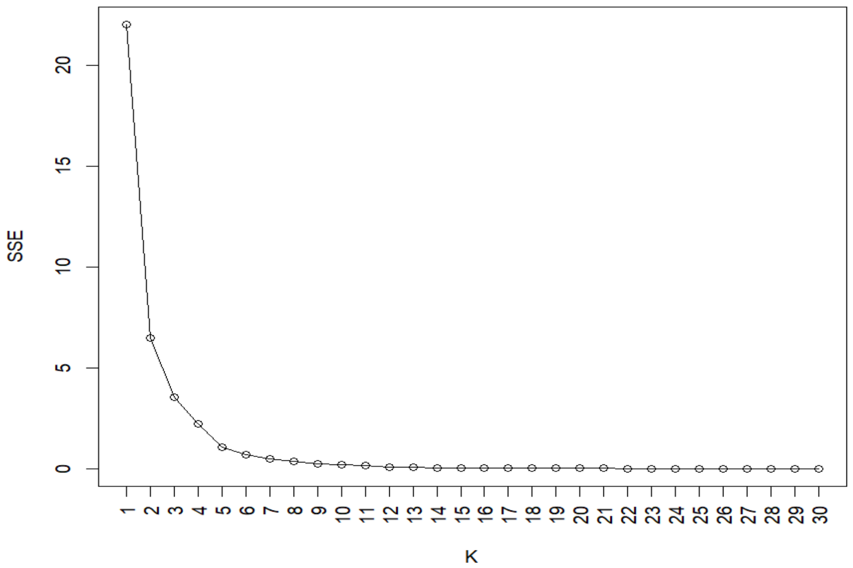
Conclusion:

(e) Produce a plot of K versus SSE, for K=1, …, 30. How many clusters does this plot suggest?

Answer)

```
> #(E)
>
> K<-(1:30); SSE<-rep(NA,length(K))
> for (k in K) {
+   SSE[k]<-kmeans(data, centers=k,nstart = 10)$tot.withinss
+ }
> plot(K, SSE, type='o', xaxt='n'); axis(1, at = K, las=2)
```

Plot:



Conclusion:
SSE shows a significant decline between K=1 and K=2. SSE continues to decline until K = 4. After that, the SSE decline begins to level out. The plot suggests that three or four clusters work best.

Problem 2 (30 points):
Cluster analysis on part of Golub data.

(a) Select the oncogenes and antigens from the Golub data. (Hint: Use grep()).

Answer)

```
> cancer <- grep("oncogene", golub.gnames[,2])
> cancer
 [1]  501  502  503  587  758  766  775  805  817  819  938 1067 1090 1111 1211 1268 1542
[18] 1596 1615 1735 1747 1750 1788 1818 1820 1837 1839 2004 2291 2302 2488 2517 2661 2681
[35] 2692 2703 2714 2715 2892 2981 2990 2993
> anti <- grep("antigen",golub.gnames[,2])
> anti
 [1]  166  313  388  497  504  514  527  540  548  614  646  664  685  763  808  826  832
[18]  833  834  872  885  890  892  893  926  936  947 1008 1010 1075 1087 1208 1258 1279
[35] 1287 1412 1422 1467 1531 1616 1645 1719 1748 1752 1756 1760 1781 1789 1798 1806 1808
[52] 1827 1852 1863 1882 1893 1908 1911 1964 2007 2170 2171 2231 2371 2546 2581 2613 2653
[69] 2672 2749 2761 2855 2989 3026 3047
```

(b) On the selected data, do clustering analysis for the genes (not for the patients). Using K-means and K-medoids with K=2 to cluster the genes. Use table () to compare the resulting two clusters with the two gene groups oncogenes and antigens for each of the two-clustering analysis.

Answer)

```
> #(b)
> data_clus<-rbind(golub[cancer,], golub[anti,])
> names <-rep(c("oncogene","antigen"), c(length(cancer), length(anti)))
> k_means <- kmeans(data_clus, centers = 2)
> k_medoids <-pam(data_clus, k=2)
> table(k_means$cluster, names)
   names
    antigen oncogene
  1      41       22
  2      34       20
> table(k_medoids$cluster, names)
   names
    antigen oncogene
  1      49       29
  2      26       13
```

(c) Use appropriate tests (from previous modules) to test the marginal independence in the two-by-two tables in (b). Which clustering method provides clusters related to the two gene groups?

Answer)

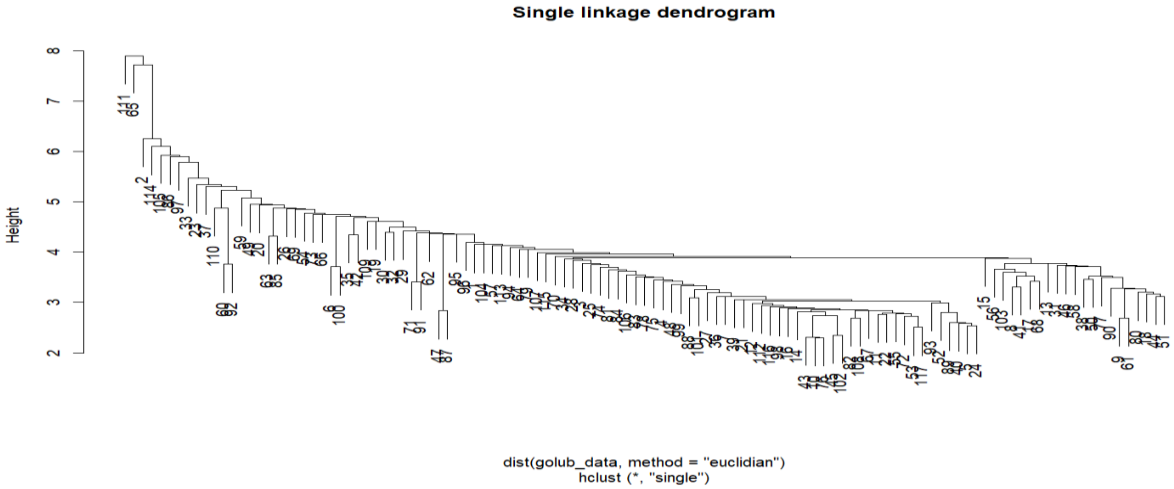
```
> #(c)
> chisq.test(table(k_means$cluster, names))

        Pearson's Chi-squared test with
        Yates' continuity correction

data:  table(k_means$cluster, names)
X-squared = 0.0019898, df = 1,
p-value = 0.9644

> chisq.test(table(k_medoids$cluster, names))

        Pearson's Chi-squared test with
        Yates' continuity correction

data:  table(k_medoids$cluster, names)
X-squared = 0.041786, df = 1,
p-value = 0.838
```

Conlusion:

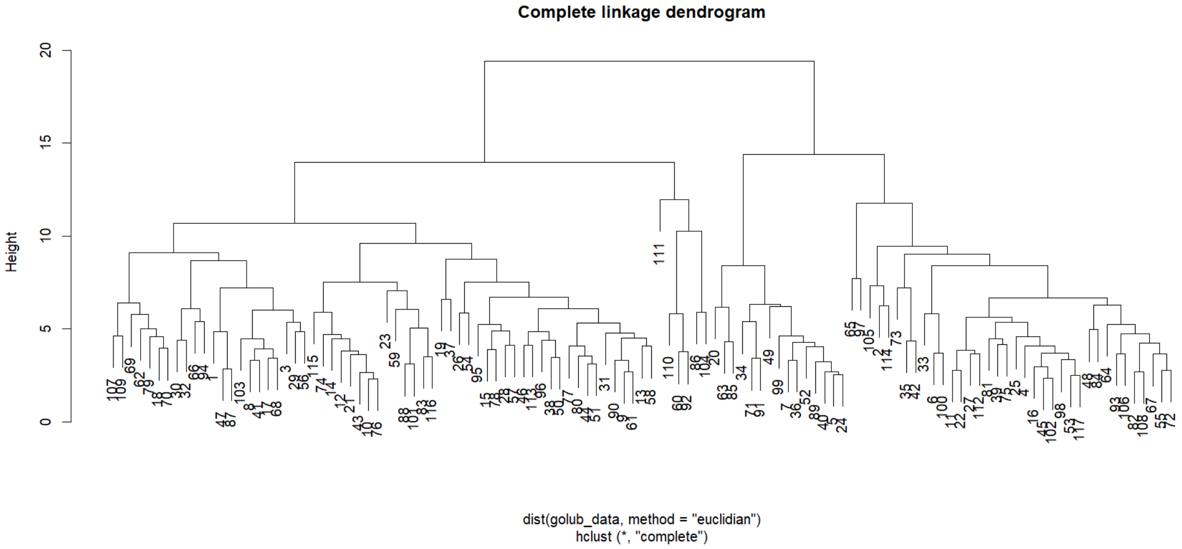
Two clustering method does nothing, so both of them are bad.

(d) Plot the cluster dendrograms for this part of golub data with single linkage and complete linkage, using Euclidean distance.

Answer) single linkage



Complete linkage



Problem 3 (30 points):
Clustering analysis on NCI60 cancer cell line microarray data (Ross et al. 2000)

We use the data set in package ISLR from r-project (Not Bioconductor). You can use the following commands to load the data set.

install.packages('ISLR')
library(ISLR)
ncidata<-NCI60$data
ncilabs<-NCI60$labs

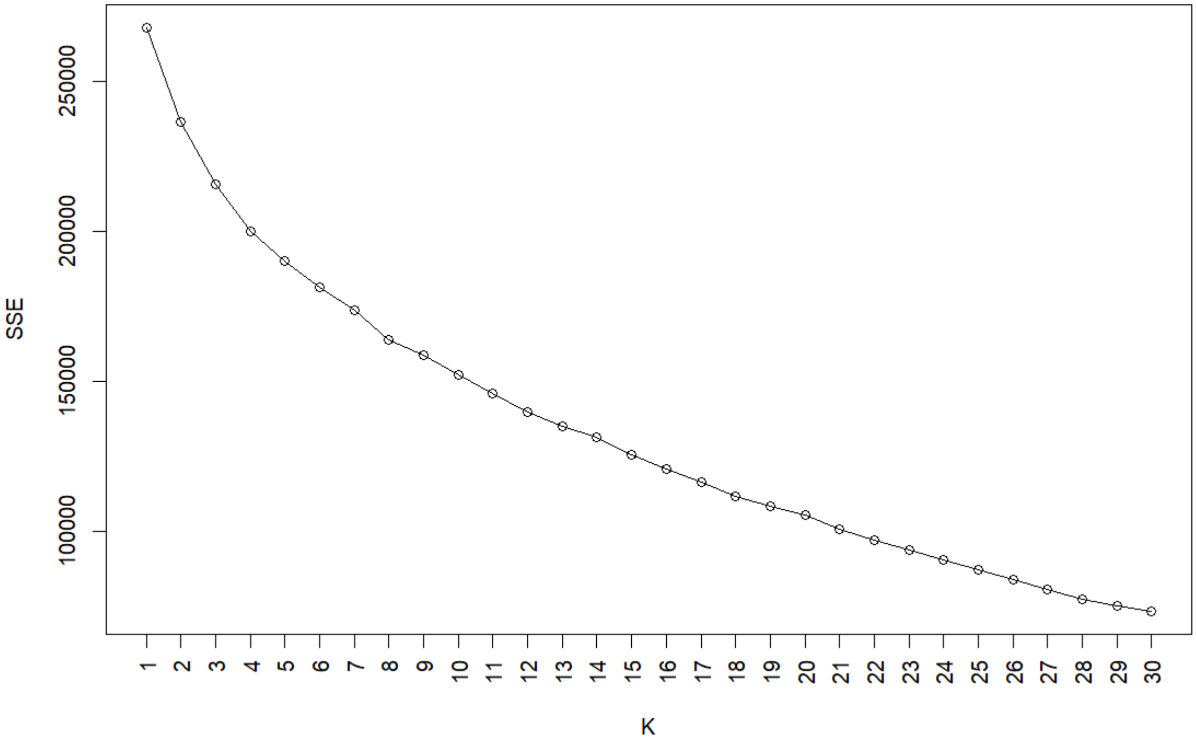
The ncidata (64 by 6830 matrix) contains 6830 gene expression measurements on 64 cancer cell lines. The cancer cell lines labels are contained in ncilabs. We do clustering analysis on the 64 cell lines (the rows).

(a) Using k-means clustering, produce a plot of K versus SSE, for K=1,…, 30. How many clusters appear to be there?

Answer)

```
> #(a)
> K<-(1:30); SSE<-rep(NA,length(K))
> for (k in K) {
+   SSE[k]<-kmeans(nci.data, centers=k,nstart = 10)$tot.withinss
+ }
> plot(K, SSE, type='o', xaxt='n'); axis(1, at = K, las=2)
> #(a)
> K<-(1:30); SSE<-rep(NA,length(K))
> for (k in K) {
+   SSE[k]<-kmeans(nci.data, centers=k,nstart = 10)$tot.withinss
+ }
> plot(K, SSE, type='o', xaxt='n'); axis(1, at = K, las=2)
```

Plot:



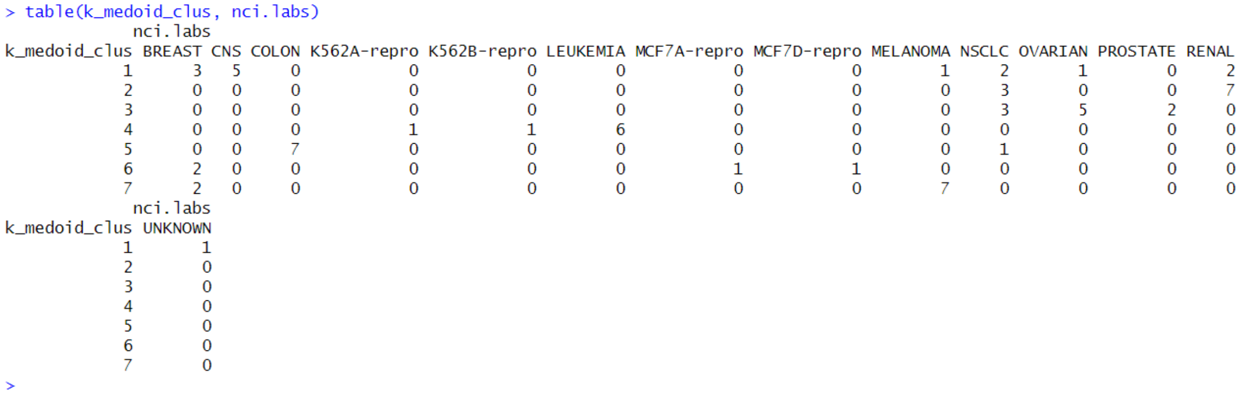
Conclusion:

The plot shows that the SSE rapidly declines as K goes from 1 to about 4-6, then levels out, showing that adding more clusters beyond this point doesn't significantly improve the quality of clustering. Hence, it appears that 4-6 clusters would be suitable for this data set.

(b) Do K-medoids clustering (K=7) with 1-correlation as the dissimilarity measure on the data. Compare the clusters with the cell lines. Which type of cancer is well identified in a cluster? Which type of cancer is not grouped into a cluster? According to the clustering results, which types of cancer are most similar to ovarian cancer?

For (b) make sure you show the table in the output file based on which you are making these conclusions.

```
> #(b)
> k_medoid <- pam(dist(1-cor(t(nci.data))), k=7)
> k_medoid_clus <- k_medoid$cluster
> table(k_medoid_clus, nci.labs)
```



```
> table(k_medoid_clus, nci.labs)
            nci.labs
k_medoid_clus BREAST CNS COLON K562A-repro K562B-repro LEUKEMIA MCF7A-repro MCF7D-repro MELANOMA NSCLC OVARIAN PROSTATE RENAL
            1      3   5     0           0           0        0           0           0        1     2       1        0     2
            2      0   0     0           0           0        0           0           0        0     3       0        0     7
            3      0   0     0           0           0        0           0           0        0     3       5        2     0
            4      0   0     0           1           1        6           0           0        0     0       0        0     0
            5      0   0     7           0           0        0           0           0        0     1       0        0     0
            6      2   0     0           0           0        0           1           1        0     0       0        0     0
            7      2   0     0           0           0        0           0           0        7     0       0        0     0
             nci.labs
k_medoid_clus UNKNOWN
            1       1
            2       0
            3       0
            4       0
            5       0
            6       0
            7       0
> 
```

Conclusion:

By seeing the clustering analysis, we can conclude that colon and melanoma is well clustered. NSCLC is very scattered and closest to ovarian is NLSLC.