Solutions to homework module 11

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?

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

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

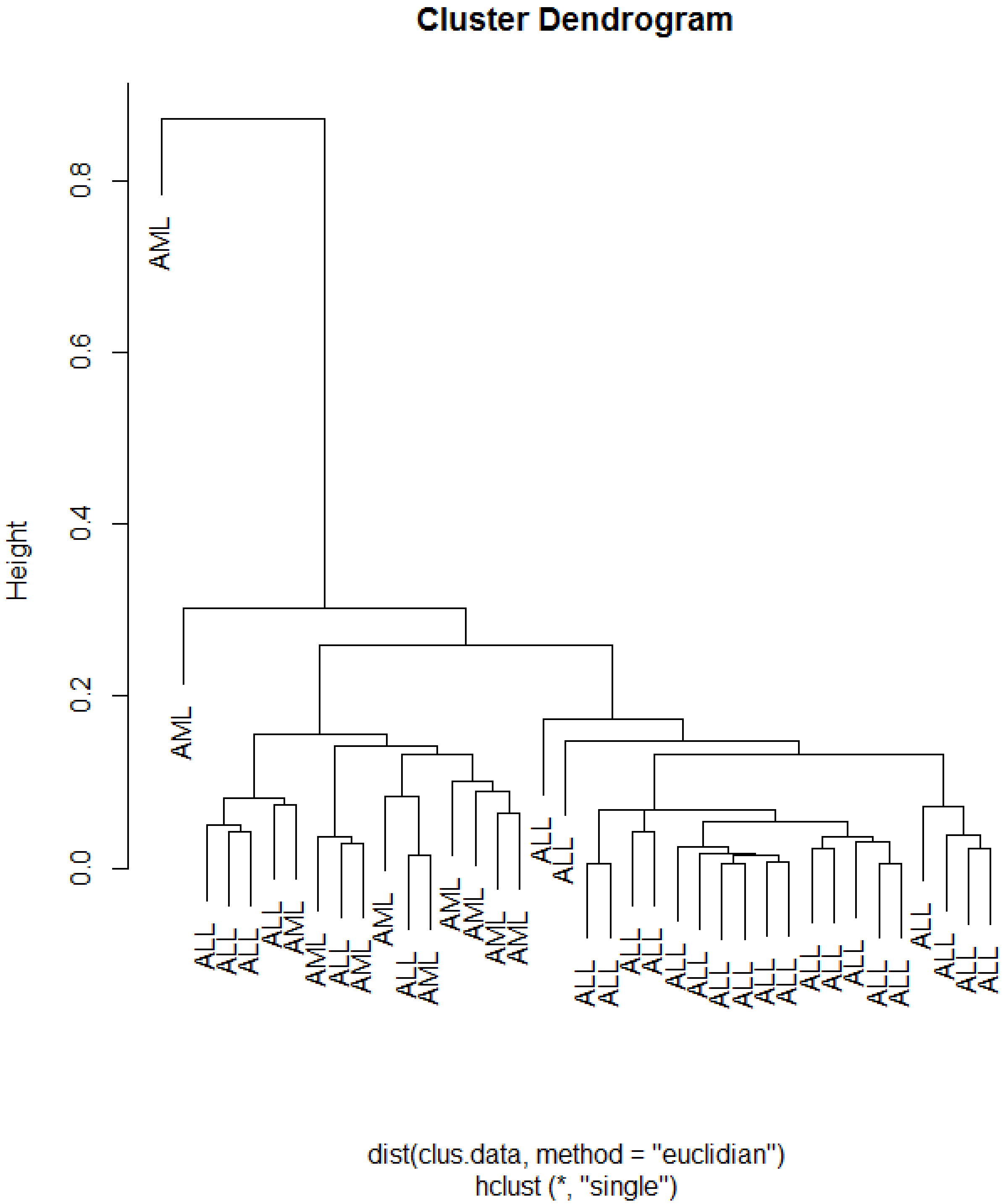
(d) Find the two cluster means from the k-means cluster analysis. Perform a bootstrap on the cluster means. Do the confidence intervals for the cluster means overlap? Which of these two cluster means is estimated more accurately?

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

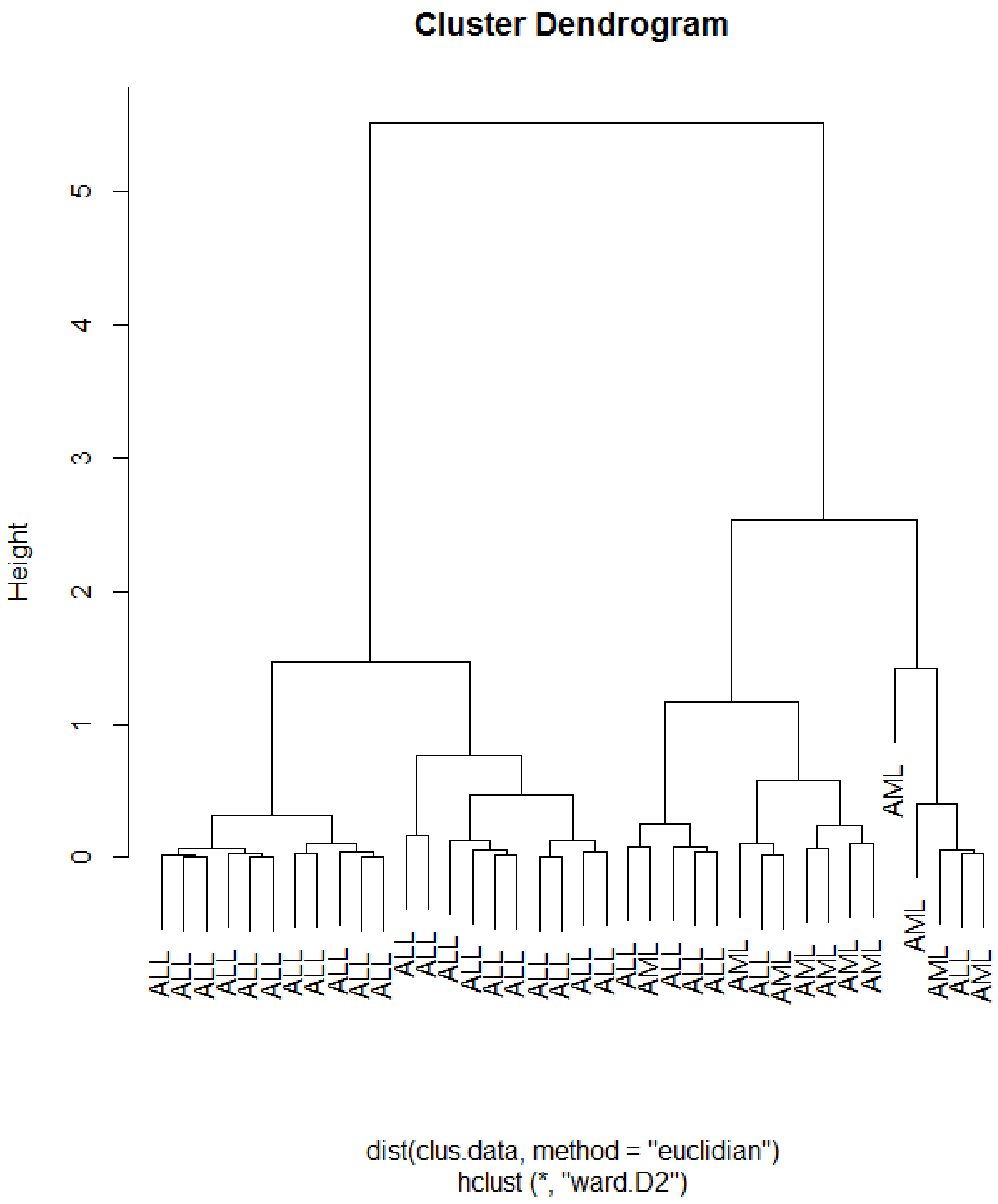
Solutions:

Solutions to homework module 11

1a)



Solutions to homework module 11



```
> table(gol.fac, clust.2single)
```

Solutions to homework module 11

```
       clust.2single
gol.fac  1  2
    ALL 27  0
    AML 10  1
```

```
> table(gol.fac, clust.2ward)
       clust.2ward
gol.fac  1  2
    ALL 21  6
    AML  0 11
```

We conclude from the above data Ward linkage function seems to be better than in Single linkage, all ALL and all except one AML are grouped under cluster 1, which is not accurate compared to ward linkage.

1b)

```
> table(gol.fac, clusters.kmean$cluster)

gol.fac  1  2
    ALL  5 22
    AML 10  1
```

1c)

K-mean produces better data (compared to single & ward) but when they i.e k-mean and ward, are compared , both seem to be almost the same from the above data .

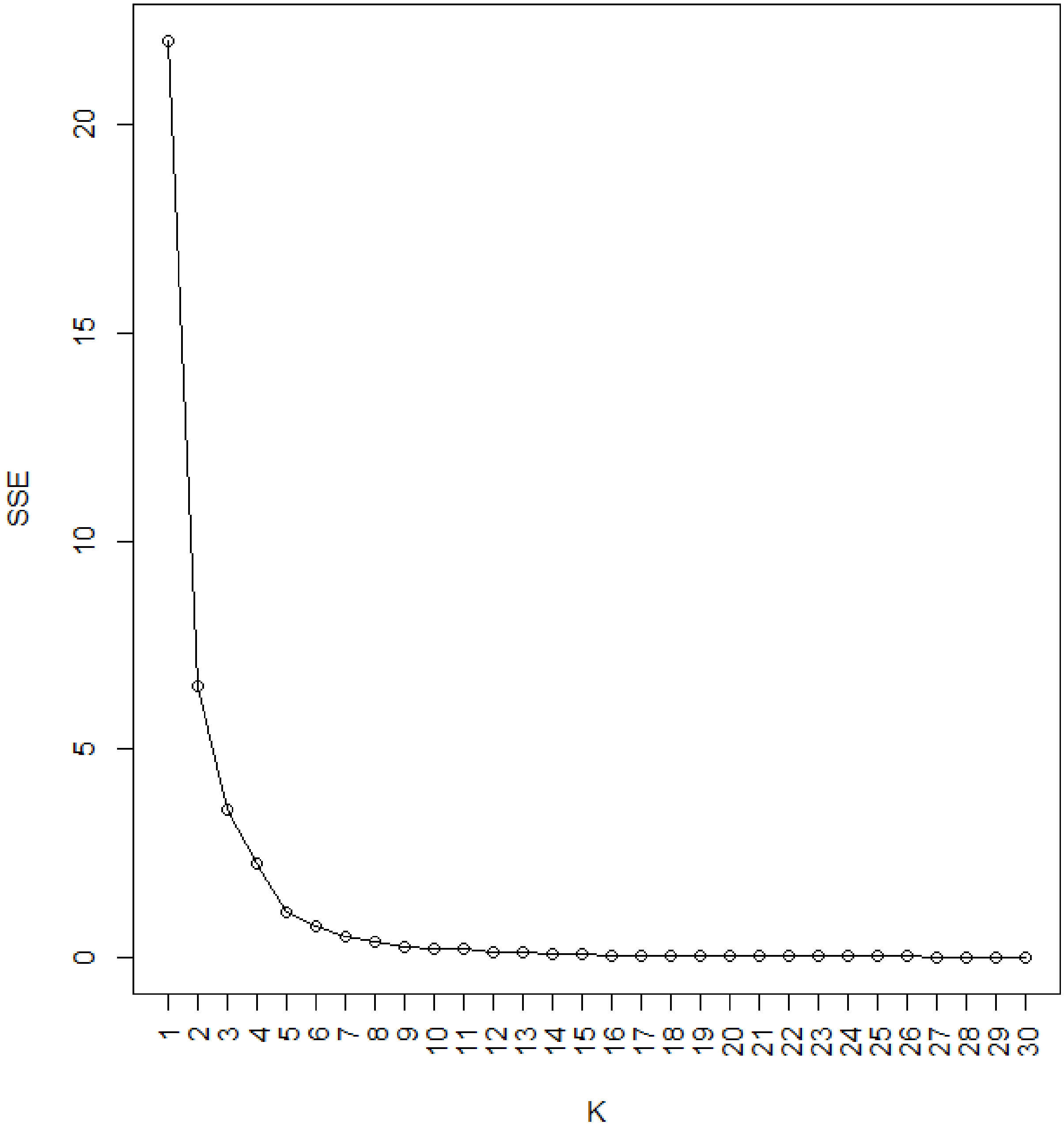
1d)

```
> clusters.kmean$centers
      [,1]
1 0.738366
2 2.045689
> apply(boot.cl,2,mean)
[1] 0.694473 2.033665
> quantile(boot.cl[,1],c(0.025,0.975))
     2.5%     97.5%
0.1660664 1.0646221
> quantile(boot.cl[,2],c(0.025,0.975))
    2.5%    97.5%
1.848128 2.195581
```

The cluster means are within the CI. The initial mean is estimated more accurately. They almost overlap each other.

Solutions to homework module 11

1e)



This plot suggests atleast 2 (big drop-off from k1 to k2) to 4 (lower drop-off) clusters after which , it decreases to level-off slowly

Solutions to homework module 11

Problem 2 (30 points):

Cluster analysis on part of Golub data.

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

(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.

(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?

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

Solutions:

2a)

```
> golub.oncogene <- agrep("^oncogene",golub.gnames[,2])
> golub.antigen <- agrep("^antigen",golub.gnames[,2])
> golub.oncoangn <- unique(c(golub.onco, golub.angn))
> gene.factor <- factor(c(rep("oncogene",length(golub.onco)),rep("antigen",le
ngth(golub.angn))))
```

2b)

```
> #oncogenes comparision
> oncocomp <- table(gene.factor, clustermea$cluster)
> #antigens comparision
> antigcomp <- table(gene.factor, clustermed$cluster)

> oncocomp

gene.factor  1  2
   antigen  41 34
   oncogene 22 20
> antigcomp

gene.factor  1  2
   antigen  49 26
   oncogene 29 13
```

Solutions to homework module 11

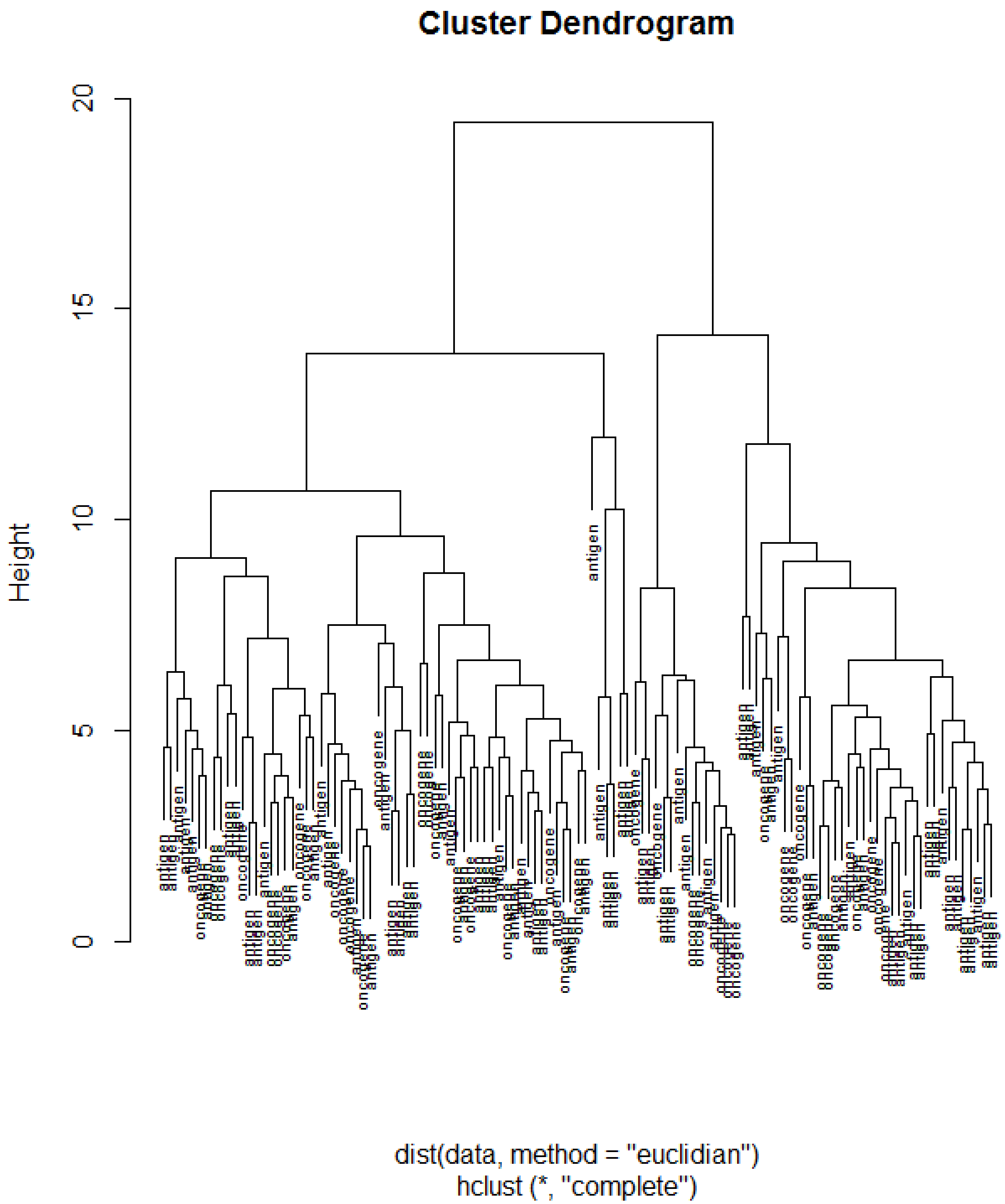
2c)

```
> fisher.test(oncocomp)

        Fisher's Exact Test for Count Data

data:  oncocomp
p-value = 0.8484
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
 0.4790382 2.5005814
sample estimates:
odds ratio
  1.095394

> fisher.test(antigcomp)

        Fisher's Exact Test for Count Data

data:  antigcomp
p-value = 0.8383
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
 0.3426507 2.0276730
sample estimates:
odds ratio
  0.846038
```

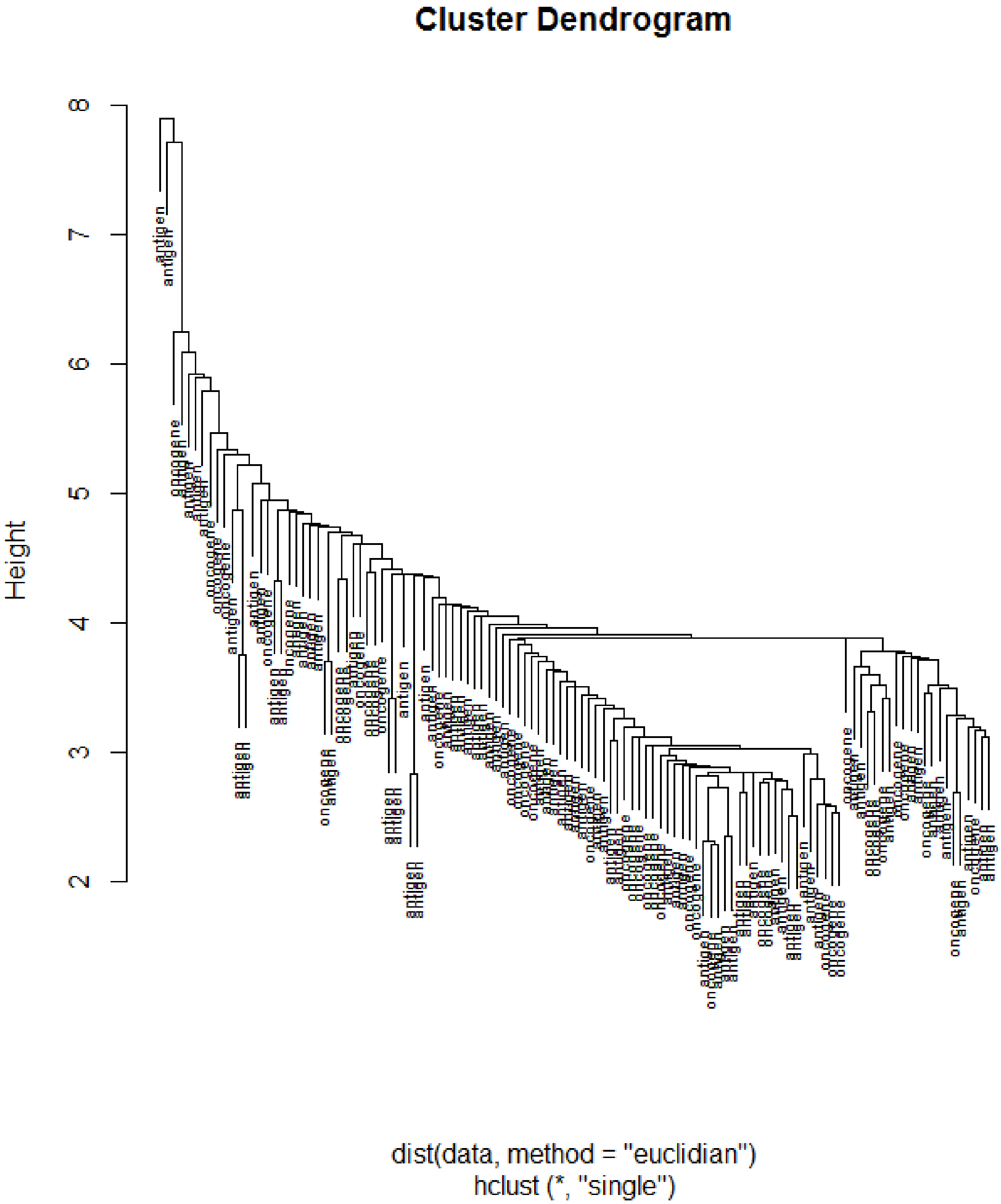
From the above data we can see that, as the p-value is greater than 0.05. Hence we accept the null hypothesis of independence. Hence, this method does not provides clusters to the two gene groups.

2d)

Solutions to homework module 11



Solutions to homework module 11



Solutions to homework module 11

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
```

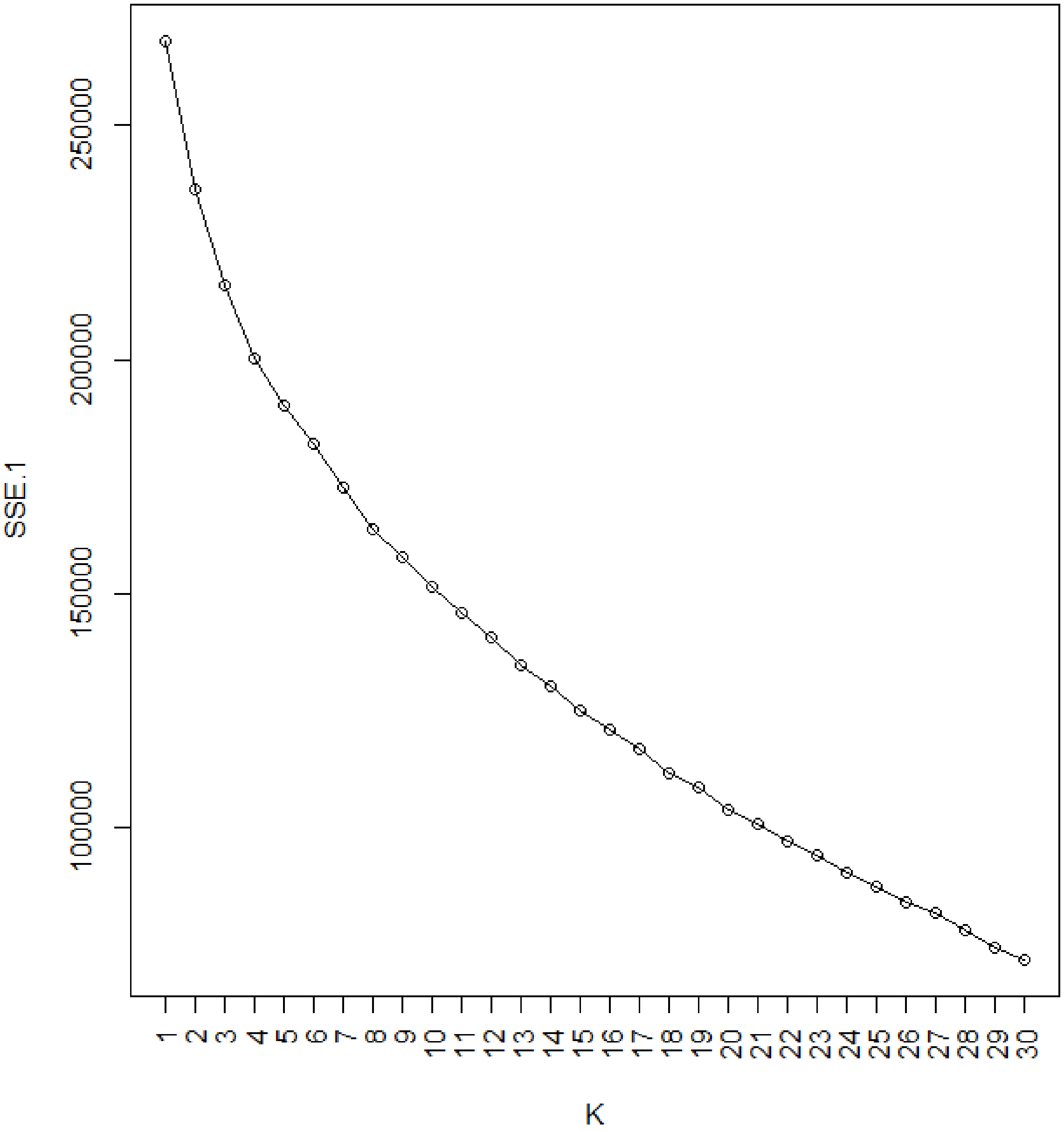
The ncidata (64 by 6830 matrix) contains 6830 gene expression measurements on 64 cancer cell lines. The cancer cell lines labels are contained in ncilabs. We do clustering analysis on the 64 cell lines (the rows).

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

(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?

Solutions to homework module 11

3a)



Solutions to homework module 11

3b)

```
> library(cluster)
> clusters.7km <- pam(as.dist(1-cor(t(ncidata))),k=7)
> table(factor(ncilabs), clusters.7km$cluster)
```

```
              1 2 3 4 5 6 7
  BREAST      0 3 0 0 2 0 2
  CNS         1 4 0 0 0 0 0
  COLON       0 0 0 7 0 0 0
  K562A-repro 0 0 0 0 0 1 0
  K562B-repro 0 0 0 0 0 1 0
  LEUKEMIA    0 0 0 0 0 6 0
  MCF7A-repro 0 0 0 0 1 0 0
  MCF7D-repro 0 0 0 0 1 0 0
  MELANOMA    0 1 0 0 0 0 7
  NSCLC       2 2 0 3 1 1 0
  OVARIAN     2 0 1 2 1 0 0
  PROSTATE    0 0 1 1 0 0 0
  RENAL       7 1 1 0 0 0 0
  UNKNOWN     0 0 1 0 0 0 0
```

Which type of cancer is well defined in a cluster?

Colon cancer appears to well clustered -> 7 in cluster 4. Renal and Melanoma too have 7 in a cluster but they have members in other cluster too.

Which type of cancer is not grouped into a cluster?

All the 64 are grouped, with 1 unknown type.

According to the clustering results, which types of cancer are most similar to ovarian cancer?

NSCLC is most similar to ovarian cancer