PILLA SATISH

STUDENT NO-210472095

Newcastle University

MAS8404 Summative Assignment

Cluster Analysis

The ISLR gene expression dataset (Ch10Ex11.csv) consists of 40 tissue samples with measurements on 1,000 genes. The first 20 samples are from healthy patients, while the second 20 are from a diseased group.

Reading the dataset

```
filename="C:/Users/pilla/Desktop/8404_Assignment/Ch10Ex11.csv"
gexpr=read.csv(filename,header = FALSE)
df=gexpr
dim(df)
```

While checking the data we observed that the gene expression data is stored in wrong way around the rows representing genes and columns representing tissue samples. We need to do transpose to get the data matrix.

```
g=t(gexpr)
dim(g)
```

Hierarchical clustering

1a. Apply hierarchical clustering with single-linkage using correlation-based distance.

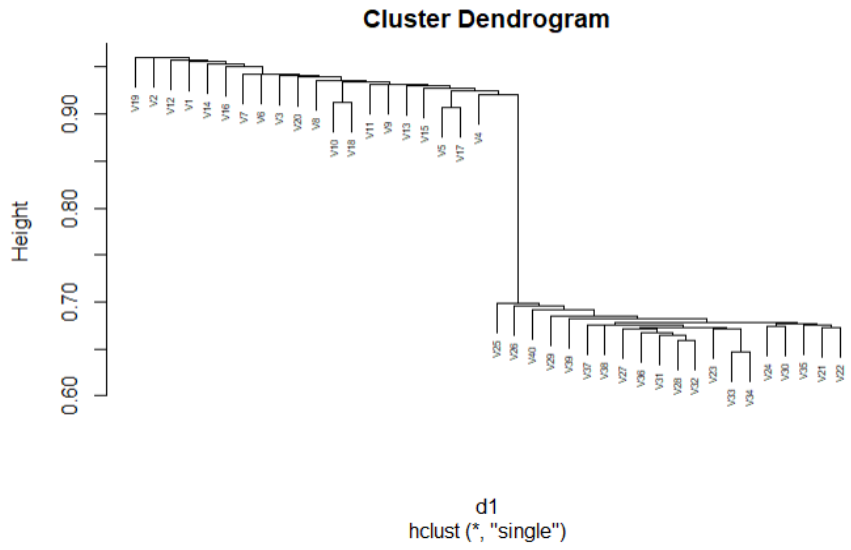

Fig 1: Dendrogram from applying agglomerative hierarchical clustering with single-linkage using correlation-based distance to ISLR gene expression data

By observing the dendrogram we can clearly see that the nearby observations are fused one-by-one leading to long, thin cluster. By using cutree function when we cut the tree at height 0.90 it is producing k=21 clusters. We can say that gene is not separating the samples into two groups.

1b. Complete linkage and average linkage

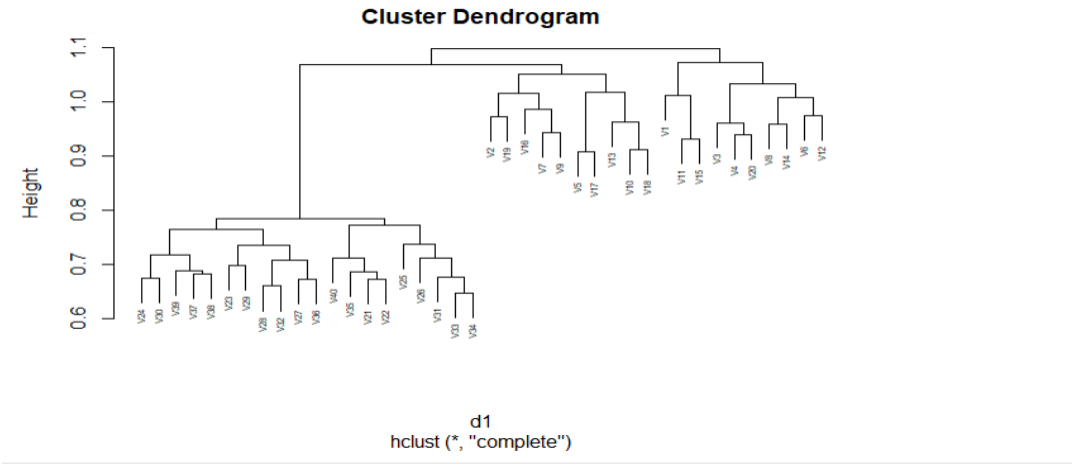

Fig 2: Dendrogram from applying agglomerative hierarchical clustering with complete-linkage using correlation-based distance to ISLR gene expression data

The dendrogram of complete linkage is balanced when compared to single-linkage and average-linkage but when it is cut at 0.9 it produces same results as single linkage that is k=21 clusters.

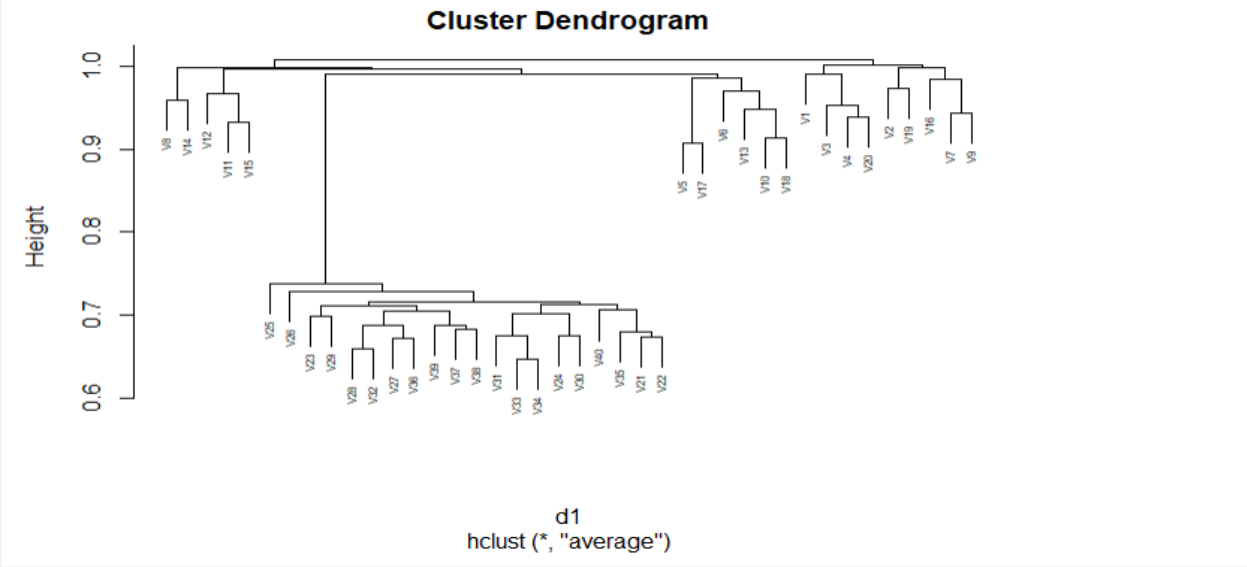

Fig 3: Dendrogram from applying agglomerative hierarchical clustering with average-linkage using correlation-based distance to ISLR gene expression data

The dendrogram for average-linkage gives k=21 when the dendrogram is cut a height of 0.9. For all the three linkage when the dendrogram is cut a height of 0.9 it is producing k=21 clusters.

1c. Single-linkage using Euclidean distance

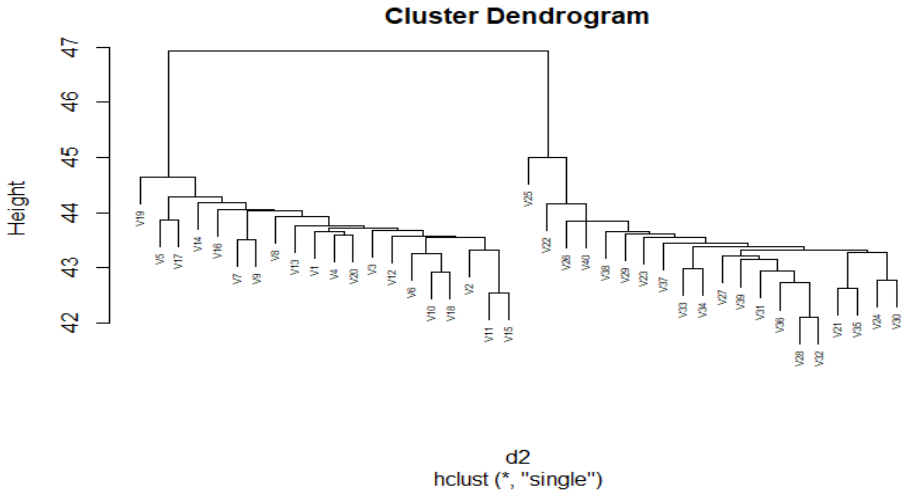

Fig 4: Dendrogram from applying agglomerative hierarchical clustering with Single-linkage using Euclidean distance to ISLR gene expression data.

From the above dendrogram we can say that when the dendrogram is cut at height=45.5, it gives exactly two clusters. By comparing the above, we can say that numbers of clusters produced depends on the distance metric used, when we used correlation-based distance we are getting higher cluster values(k=21) when Euclidean distance is used we are getting k=2 clusters. From above we say that the using Euclidean distance gives us right number of clusters and it's is best method to determine the cluster for ISLR gene expression data.

K-means clustering

2a. Applying the K-means algorithm for a range of values of K

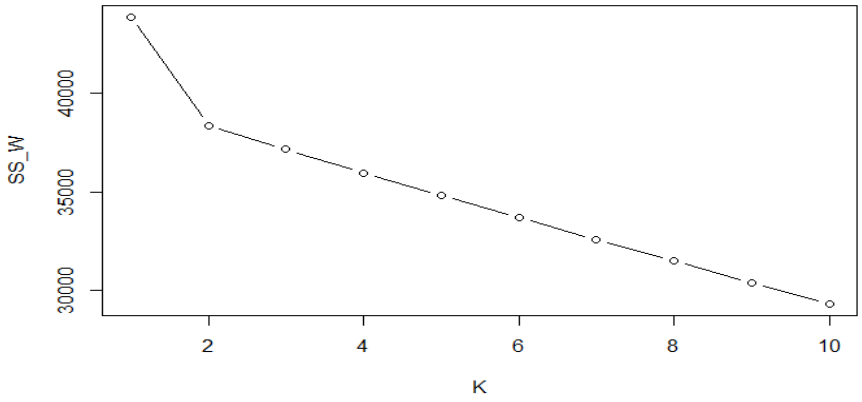

Fig 5: SSw against K for K-means clustering of the ISLR gene expression dataset

From the above plot we can see there is an elbow at K=2, which suggest it might be appropriate to use 2 cluster.

```
round(SS_B)
```

```
[1]     0  5531  6748  7939  9094 10219 11320 12434 13524 14596
```

We can also see that difference between-cluster sum of squares (SSb) values after K=2 are approximately equal. If we divided the cluster more than 2 the data will be fused near to each other and it will be hard to detect any clusters, which we can see clearly when the cluster is divided into 4 cluster.

2b. visual display of the four clusters in a two-dimensional plot.

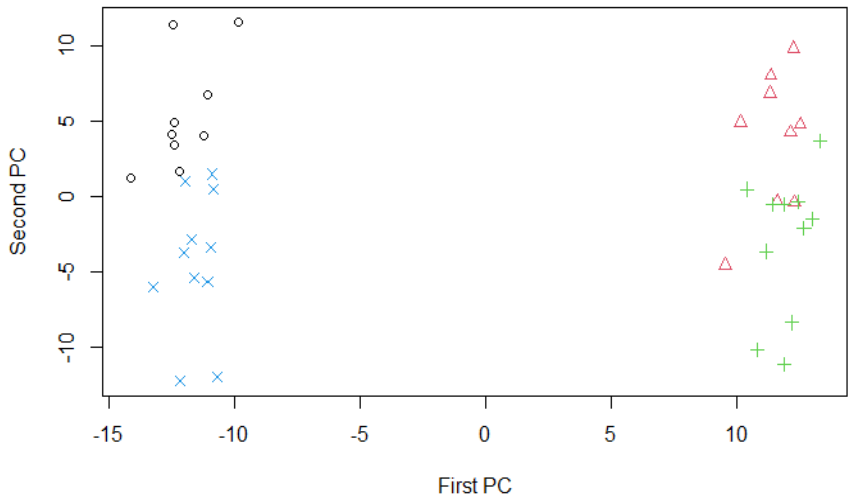

Fig 6: Cluster solutions in the ISLR gene expression data in the space of two PCs. Clusters were identified by K-means clustering with K=4

From the above we see that in the low-dimensional representation of the ISLR gene expression data, the cluster does not appear to be well separated when K=4.

Linear Regression

3b. Test error from the full data set (applying least squares model) is 2842.256.

```
[1] 2842.256
```

3c. Test error from fitting a least squares model to the training data using the 6 predictors is 2800.964.

```
[1] 2800.964
```

3di(I). Optimal Lambda value is 0.011

```
[1] 0.01353048
[1] 86
```

3di (II). Fixing the optimal value and fitting the model to all the training data and computing the test error over the validation data, Test error is 2908.782

```
[1] 2913.103
```

3dii Based on the full data applying cross-validation to identify an optimal value for the tuning parameter. Optimal value is 0.001

```
[1] 0.001
[1] 100
```

3dii. Plot showing how the estimates of the regression coefficients change as the tuning parameter is increased

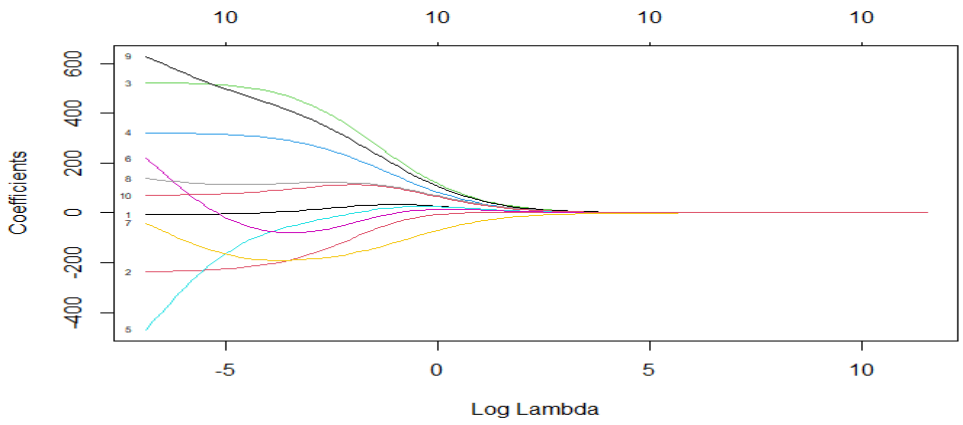

Fig 7: Graphical illustration of the effect of varying the tuning parameter when applying ridge regression to the diabetes dataset.

Each line represents the regression coefficient for a different variable. We can see from the plot that by the time log lambda is around 4(i.e. lambda is around e^5=57), all regression coefficients are essentially equal to zero. Ridge regression includes all the variables in the model. Ridge minimizes the residual sum of squares plus a shrinkage penalty of lambda multiplied by the sum of squares of the coefficients. As lambda increases, the coefficients approach zero.

```
11 x 1 sparse Matrix of class "dgCMatrix"
                      s1
(Intercept)  152.133484
age           -8.036848
sex         -236.348533
bmi          521.220228
map          321.762558
tc          -468.699837
ldl          220.457336
hdl          -40.968299
tch          139.235652
ltg          627.587343
glu           69.898524
```

```
lsq_fit=lm(dis~.,data=diabetes)
coef(lsq_fit)
```

```
(Intercept)         age         sex         bmi         map          tc         ldl         hdl         tch
  152.13348   -10.01220  -239.81909   519.83979   324.39043  -792.18416   476.74584   101.04457   177.06418
        ltg         glu
  751.27932    67.62539
```

Fig 8: estimated regression coefficients for ridge regression and estimated coefficients in the full model fitted by least squares.

From above we can see when lambda is 0.001 we are getting a solution very close to the least square estimates. Only tc, ldl, and hdl coefficients values are different when compared to least square estimates.

3e. Comparing the three test errors:

Description: df [3 x 2]

methods <chr>	Test_errors <dbl>
least square Method	2842.256
least square on 6 best variables	2800.964
Ridge Method	2913.103

3 rows

Fig 9: Comparing the test errors for least square, least square on 6 best variables, ridge method

Test error on the least square using 6 variables is best method as it produces least test error value among other two methods. Using only 6 variables we are able to build a model which helps in predicting the response variable with low test error among 3 methods.