

Student id: 200110688

Name: Akshay vasant Nayak

R. Bonus Question (R implementation) (4% of grade) (Please note that this is completely optional; use your time wisely as the implementation may take time). (You can use any 2-d data, real or simulated for implementation; test data will be provided later to answer part b of this question)

(a) Implement G-Means (paper is provided under additional resources) (Algorithm 1, listed on page 3). (submit code as separate file; make single zip file)

(b) Generate 2-d plots (scatter plots and draw ellipsoids) (data will be provided later), include these plots as part of h/w solution)

PART B

NOTE: 2d plots for test data has been plotted using significance level:1

(the six plots correspond to the clusters detected during the corresponding iterations of the algorithm)

The plots for this model has been attached in the zip file.

NOTE: THE CODE GENERATES PLOTS AND STORES IT IN A DIRECTORY SPECIFIED BY THE FOLLOWING COMMAND

mypath <-

file.path("C:", "Users", "AkshayN", "Desktop", "fds", "r_codes", "hw5_papers", "cluster_plots", paste("cluster_generated_during_iteration_0.jpg", sep = ""))

PLEASE UNCOMMENT THESE LINES AND SPECIFY THE DIRECTORY TO STORE PLOTS IF YOU WANT TO VIEW THEM .

THEY HAVE BEEN MARKED IN 2 SECTIONS AS SHOWN BELOW

LINES OF CODE

Screenshots for sample 2d data:

```

> sigma1<-matrix(c(0.8,0,0,0.8),2,2)
> a<-mvrnorm(n=100,mu=c(20,20),sigma=sigma1)
>
> sigma2<-matrix(c(0.8,0,0,0.8),2,2)
> b<-mvrnorm(n=100,mu=c(30,35),sigma=sigma2)
>
> sigma1<-matrix(c(0.8,0,0,0.8),2,2)
> c<-mvrnorm(n=100,mu=c(28,25),sigma=sigma1)
>
> sigma2<-matrix(c(0.8,0,0,0.8),2,2)
> d<-mvrnorm(n=100,mu=c(22,35),sigma=sigma2)
>
> sigma2<-matrix(c(0.8,0,0,0.8),2,2)
> e<-mvrnorm(n=100,mu=c(25,30),sigma=sigma2)
>
> plot(rbind(a,b,c,d,e))
> x<-rbind(a,b,c,d,e)
> center<-colMeans(x)
> center<-matrix(center,ncol = ncol(x))
> prin_comp<-eigen(cov(x))
> s=prin_comp$vectors[,1]
> m=s*sqrt(2*prin_comp$values[1]/pi)
>
> c1<-center+m
> c2<-center-m
> cat("proposed centers(c1,c2) for cluster\n",c1,"\n",c2)
proposed centers(c1,c2) for cluster
26.7183 33.66191
23.21179 24.38043

```

```

> KMC = kmeans(X, centers=matrix(rbind(c1,c2), ncol = ncol(X)))
> cat("centers found by using kmeans \n",KMC$centers[1,],"\n",KMC$centers[2,])
centers found by using kmeans
25.66202 33.36844
23.91958 22.50026
> v<-KMC$centers[1,]-KMC$centers[2,]
> normv<-sqrt(sum(v^2))
> xdash<-X%%cbind(v)/normv
> scaled_xdash<-scale(xdash)
> zxdash<-ecdf(scaled_xdash)
> library(nortest)
> test_result<-ad.test(zxdash(scaled_xdash))
> cat("the test statistic is",test_result$statistic)
the test statistic is 5.525998
> if(test_result$statistic>1.8692)
+ {
+   newcenters=rbind(KMC$centers[1,],KMC$centers[2,])
+   cat("test statistic is greater than cv(1.869) for alpha 0.00
01 and so we reject H0 and accept the split")
+ }else
+ {
+   print("the dataset just has one cluster")
+   return
+ }
test statistic is greater than cv(1.869) for alpha 0.0001 and so
we reject H0 and accept the split

```

```

> iteration<-1
> while(nrow(newcenters)>nrow(center)){
+ cat("\n-----Next Iteration-----",iteration)
+ center<-newcenters
+ newcenters<-{}
+ KMC<-kmeans(X,centers = center)
+ mypath <- file.path("C:", "Users", "AkshayN", "Desktop", "fds", "r_codes", "hw5_papers", "cluster_plots", paste("cluster_generated_during_iteration", iteration, ".jpg", sep = ""))
+
+ jpeg(file=mypath)
+ mytitle = paste("clusters detected during iteration:",iteration)
+ plot(X,col=KMC$cluster,xlab = "dim1",ylab="dim2",main=mytitle)
+ points(KMC$centers,col="orange",pch=11,lwd=3)
+ dev.off()
+ for(i in 1:nrow(KMC$centers))
+ {
+   cat("\n\n-----considering cluster-----",i)
+   ci<-KMC$centers[i,]
+   xi<-subset(X,KMC$cluster==i)
+   prin_comp<-eigen(cov(xi))
+   s=prin_comp$vectors[,1]
+   m=s*sqrt(2*prin_comp$values[1]/pi)
+   c1<-ci+m
+   c2<-ci-m
+   cat("\n\n proposed centers(c1,c2) for this cluster\n",c1,"\n",c2)
+   kmcxi<-kmeans(xi,centers=matrix(c(c1,c2),ncol=2))
+   cat("\n\n centers found by using kmeans \n",kmcxi$centers[1,],"\n",kmcxi$centers[2,])
+
+   v=kmcxi$centers[1,]-kmcxi$centers[2,]
+
+   normv<-sqrt(sum(v^2))
+   xdash<-xi%%cbind(v)/normv
+   scaled_xdash<-scale(xdash)
+   zxdash<-ecdf(scaled_xdash)
+   library(nortest)
+   test_result<-ad.test(zxdash(scaled_xdash))
+   cat("\n A*(Z)",test_result$statistic)
+   if(test_result$statistic>1.8692)
+   {
+     newcenters=rbind(newcenters,kmcxi$centers[1,],kmcxi$centers[2,])
+     cat("\n\n since test statistic is greater than the critical value 1.8692(alpha=0.0001) we reject H0 and keep {c1,c2}")
+   }else
+   {
+     newcenters=rbind(newcenters,ci)
+     cat("\n\n since test statistic is smaller than cv we accept H0 and discard{c1,c2}")
+   }
+ }
+ iteration<-iteration+1
+ }

-----Next Iteration----- 1

-----considering cluster----- 1
proposed centers(c1,c2) for this cluster
23.00295 32.87052
28.3211 33.86637
centers found by using kmeans
23.55167 32.53606
29.88273 35.03321
A*(Z) 3.303036
since test statistic is greater than the critical value 1.8692(alpha=0.0001) we reject H0 and keep{c1,c2}

```

```

-----considering cluster----- 2
proposed centers(c1,c2) for this cluster
20.66201 20.41893
27.17714 24.58159
centers found by using kmeans
27.9285 25.06372
19.91065 19.93681
A*(Z) 2.192238
since test statistic is greater than the critical value 1.8692(alpha=0.0001) we reject H0 and keep{c1,c2}
-----Next Iteration----- 2

-----considering cluster----- 1
proposed centers(c1,c2) for this cluster
22.19489 34.739
24.65438 30.75819
centers found by using kmeans
22.04328 35.01154
24.95947 30.23421
A*(Z) 2.081224
since test statistic is greater than the critical value 1.8692(alpha=0.0001) we reject H0 and keep{c1,c2}

-----considering cluster----- 2
proposed centers(c1,c2) for this cluster
29.47951 35.69267
30.28595 34.37375
centers found by using kmeans
30.26463 34.34509
29.50084 35.72133
A*(Z) 1.083709
since test statistic is smaller than cv we accept H0 and discard{c1,c2}

-----considering cluster----- 3
proposed centers(c1,c2) for this cluster
27.08468 26.353
28.4154 24.39883
centers found by using kmeans
26.05248 28.17156
27.97755 25.00124
A*(Z) 1.194285
since test statistic is smaller than cv we accept H0 and discard{c1,c2}

-----considering cluster----- 4
proposed centers(c1,c2) for this cluster
19.77333 20.67439
20.04796 19.19923
centers found by using kmeans
19.52941 20.64182
20.24872 19.31161
A*(Z) 1.083709
since test statistic is smaller than cv we accept H0 and discard{c1,c2}
-----Next Iteration----- 3

-----considering cluster----- 1
proposed centers(c1,c2) for this cluster
21.2787 34.78379
22.80785 35.23929
centers found by using kmeans
21.33502 34.67438
22.84195 35.39174
A*(Z) 1.083709
since test statistic is smaller than cv we accept H0 and discard{c1,c2}

```

```

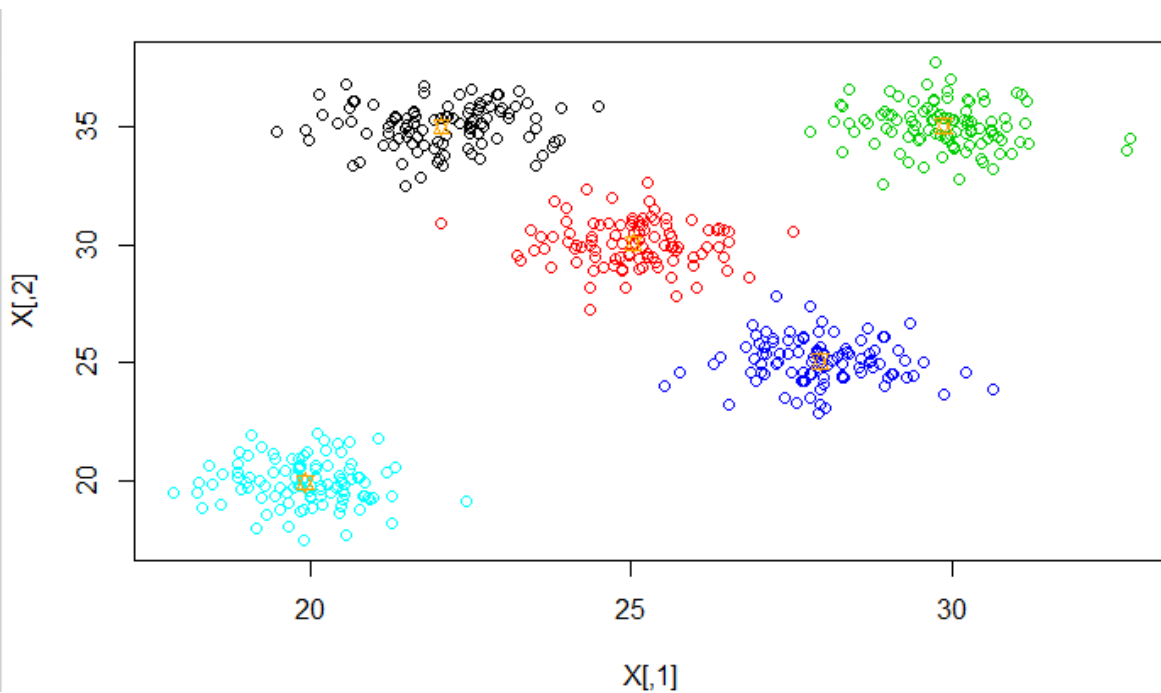
-----considering cluster----- 2
proposed centers(c1,c2) for this cluster
24.7347 30.77451
25.32708 29.33527
centers found by using kmeans
24.82043 29.46329
25.33373 30.90621
A*(Z) 1.083709
since test statistic is smaller than cv we accept H0 and discard{c1,c2}

-----considering cluster----- 3
proposed centers(c1,c2) for this cluster
29.47951 35.69267
30.28595 34.37375
centers found by using kmeans
30.26463 34.34509
29.50084 35.72133
A*(Z) 1.083709
since test statistic is smaller than cv we accept H0 and discard{c1,c2}

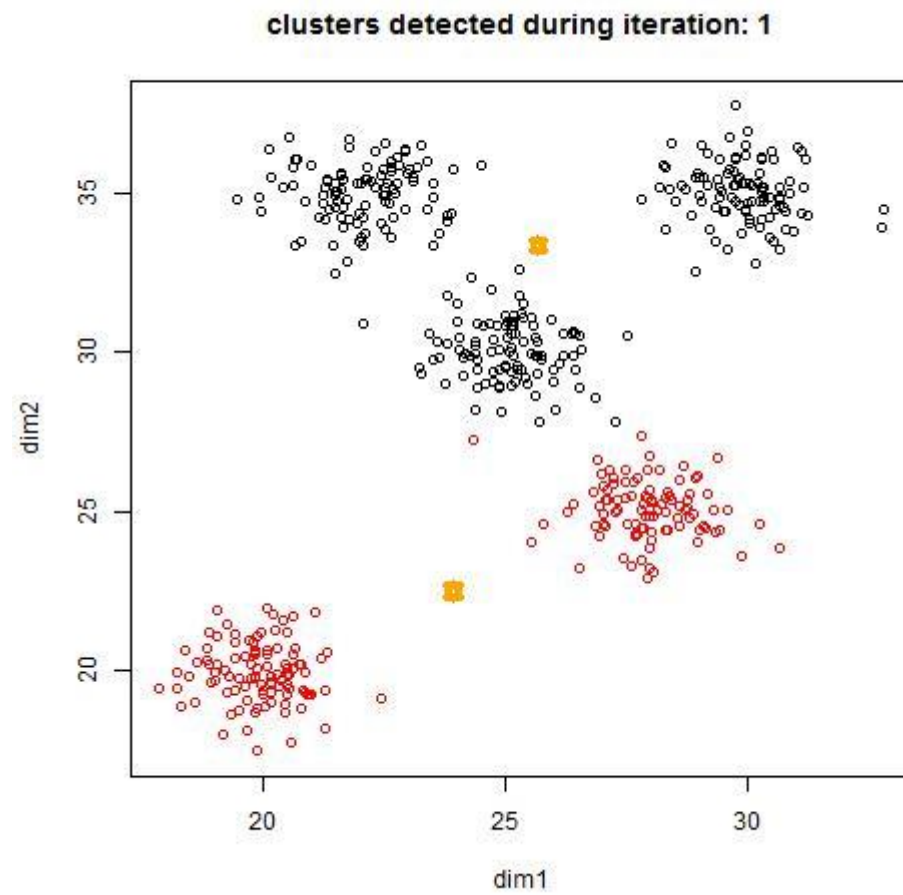
-----considering cluster----- 4
proposed centers(c1,c2) for this cluster
27.52722 25.68143
28.38815 24.45738
centers found by using kmeans
27.52135 25.67137
28.39402 24.46744
A*(Z) 1.083709
since test statistic is smaller than cv we accept H0 and discard{c1,c2}

-----considering cluster----- 5
proposed centers(c1,c2) for this cluster
19.77333 20.67439
20.04796 19.19923
centers found by using kmeans
19.52941 20.64182
20.24872 19.31161
A*(Z) 1.083709
since test statistic is smaller than cv we accept H0 and discard{c1,c2}
> plot(x,col=KMC$cluster)
> points(KMC$centers,col="orange",pch=11)
>

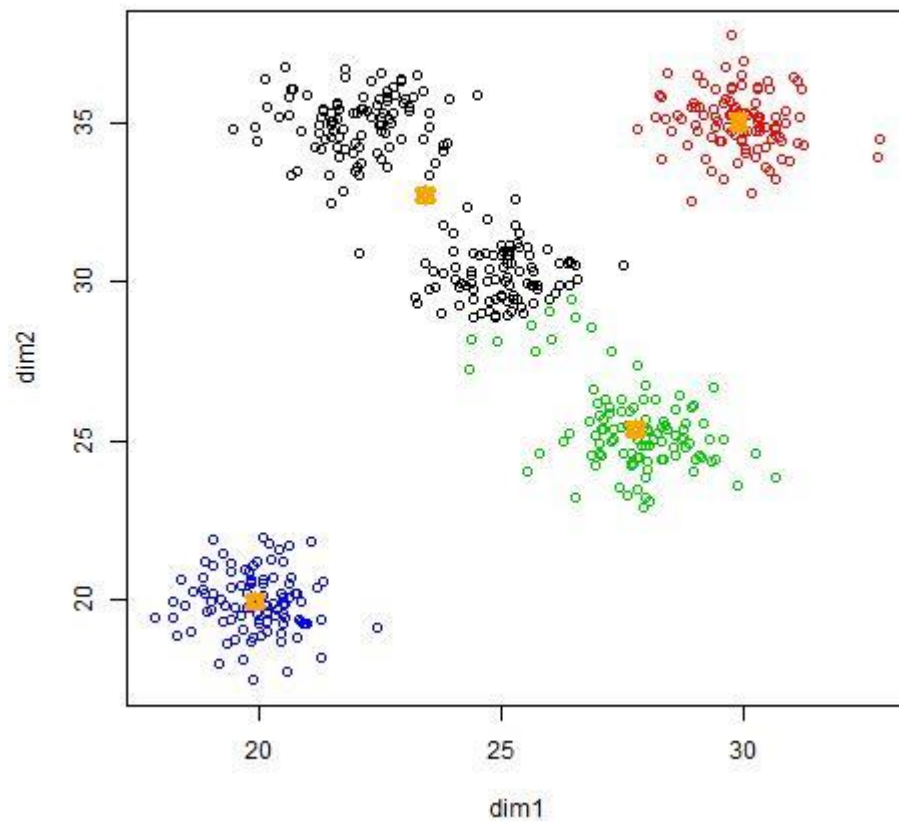
```

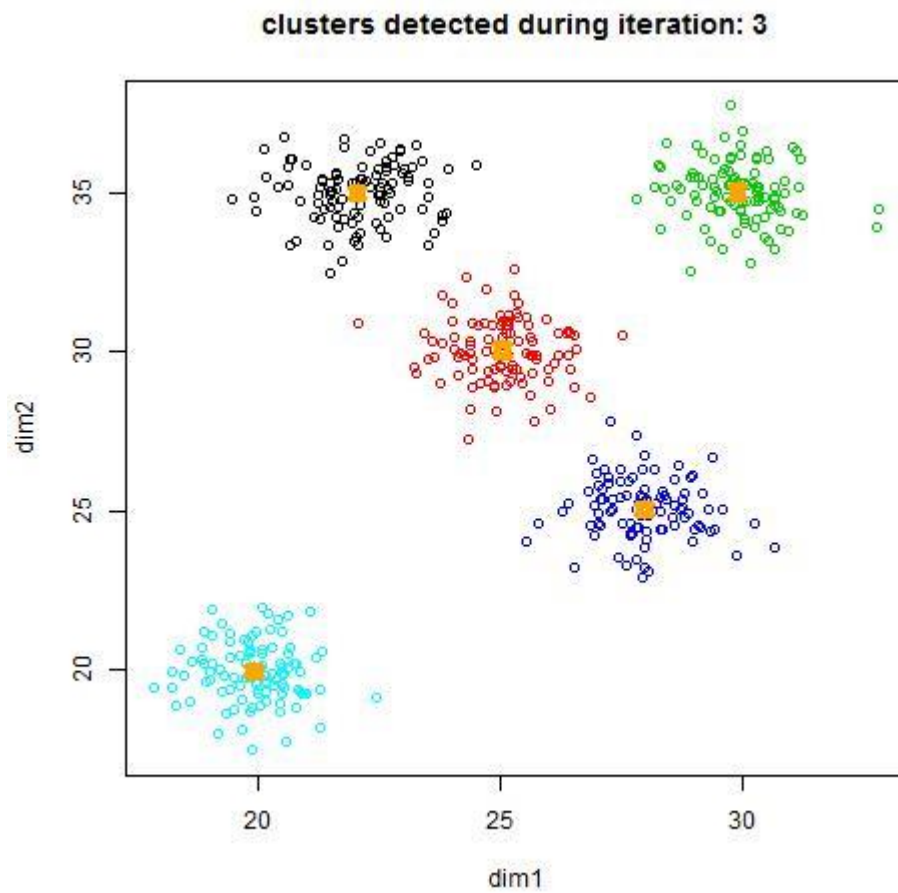


Plots generated during the iteration



clusters detected during iteration: 2





Gmeans algorithm gives 12 clusters on the test data when we select 1% significance level. The critical value for which is 1.035.

Screenshots for the sample output for significance level:1


```

+ J
> gmeans(significance_level = 1)

critical value for this significance_level(in %) is: 1.035
Proposed centers(c1,c2) for cluster
101.2998 112.2597 103.7811
166.3196 165.2737 152.58
centers found by using kmeans
101.8897 112.8513 103.9564
171.5333 169.3939 156.8091
the test statistic is 7.967638
A*(Z) 8.011519
test statistic is greater than critical value we accept the split
-----Next Iteration----- 1

-----considering cluster----- 1
Current center for this cluster 101.8897 112.8513 103.9564
proposed centers(c1,c2) for this cluster
121.6958 125.5885 115.2509
82.08368 100.114 92.66192
centers found by using kmeans
118.7068 124.0201 113.8795
72.19149 93.12766 86.43262
A(Z) 4.308759
A*(Z) 4.332489
since test statistic is greater than the critical value we reject H0 and keep{c1,c2}

-----considering cluster----- 2
Current center for this cluster 171.5333 169.3939 156.8091
proposed centers(c1,c2) for this cluster
184.9815 185.0372 170.693
158.0851 153.7507 142.9252
centers found by using kmeans
193.0865 193.5288 177.4327
161.615 158.2876 147.3186
A(Z) 3.646008
A*(Z) 3.666088
since test statistic is greater than the critical value we reject H0 and keep{c1,c2}
-----Next Iteration----- 2

```

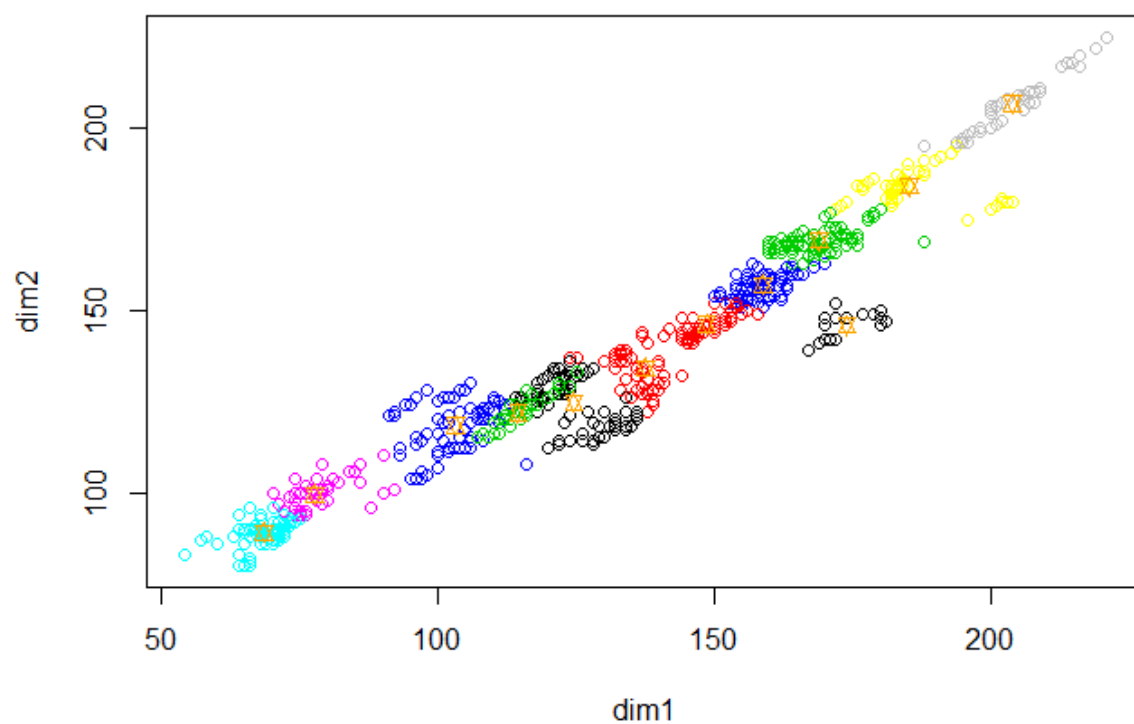
```

Console ~/
since test statistic is smaller than cv we do not reject H0 and discard{c1,c2}

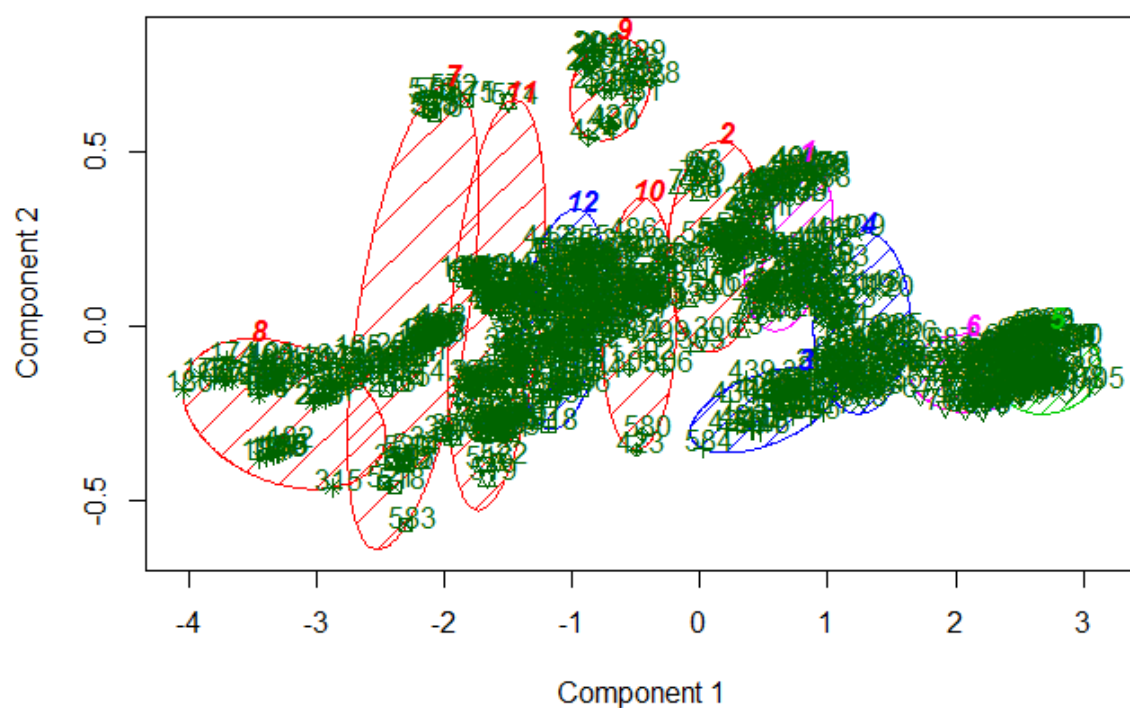
-----considering cluster----- 12
Current center for this cluster 158.8553 157.0132 146.6842
proposed centers(c1,c2) for this cluster
159.4043 159.2117 150.0112
158.3062 154.8146 143.3572
centers found by using kmeans
160.0909 159.2273 149.3636
157.1562 153.9688 143
A(Z) 0.8258728
A*(Z) 0.8304212
since test statistic is smaller than cv we do not reject H0 and discard{c1,c2}
-----Final centers for the clusters-----

124.6517 124.7528 112.1573
137.5417 134.1875 120.4375
114.5 122 124.8684
103.1538 118.5128 106.4231
68.47674 89.01163 83.0814
77.72222 99.37037 91.55556
185.4828 184.0862 168.4828
204.1364 206.75 189.5682
174.0556 145.8889 129.2778
148.6 146.2889 137.1556
168.9651 169.1977 158.5116
158.8553 157.0132 146.6842
> |

```



CLUSPLOT(X)



These two components explain 99.4 % of the point variability.

Sample output :2

Significance level 10%

```
> gmeans(significance_level = 10)

critical value for this significance_level(in %) is: 0.631
Proposed centers(c1,c2) for cluster
101.2998 112.2597 103.7811
166.3196 165.2737 152.58
centers found by using kmeans
101.8897 112.8513 103.9564
171.5333 169.3939 156.8091
the test statistic is 7.967638
A*(Z) 8.011519
test statistic is greater than critical value we accept the split
-----Next Iteration----- 1

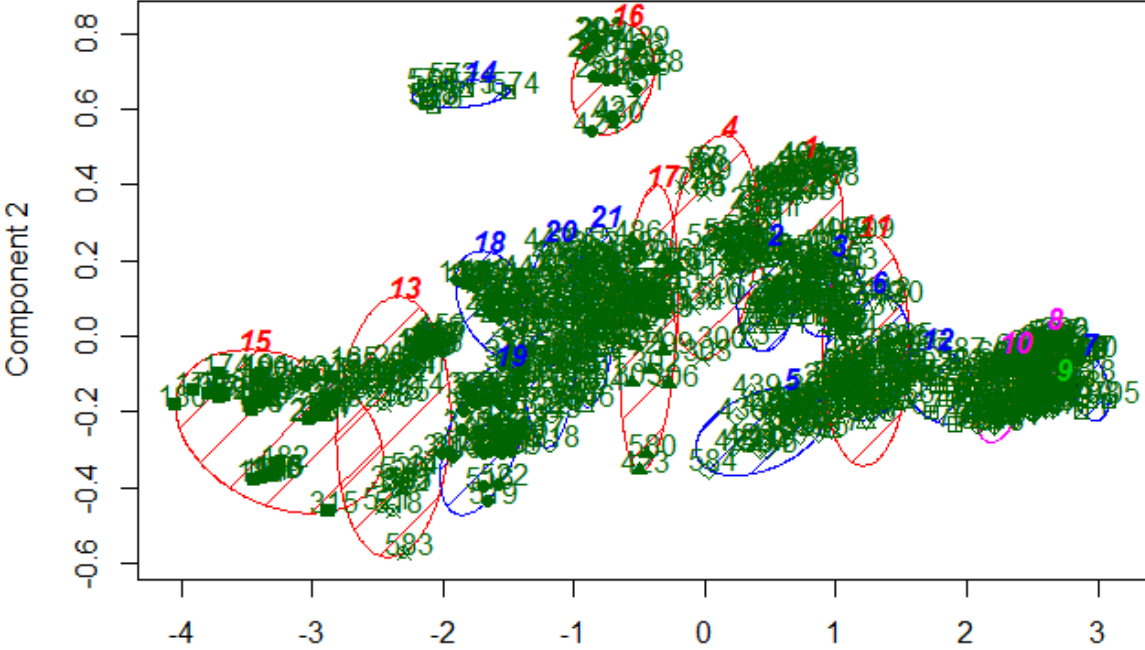
-----considering cluster----- 1
current center for this cluster 101.8897 112.8513 103.9564
proposed centers(c1,c2) for this cluster
121.6958 125.5885 115.2509
82.08368 100.114 92.66192
centers found by using kmeans
118.7068 124.0201 113.8795
72.19149 93.12766 86.43262
A(Z) 4.308759
A*(Z) 4.332489
since test statistic is greater than the critical value we reject H0 and keep{c
1,c2}
```

Console ~/

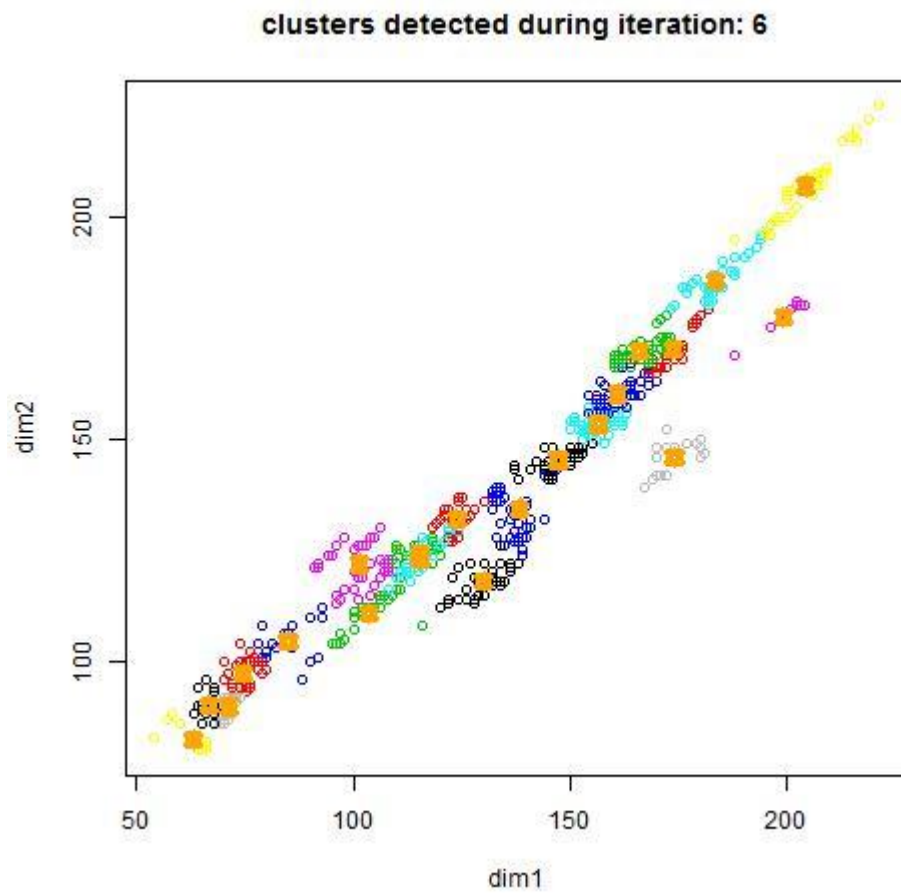
```
Current center for this cluster 156.4878 153.2683 142.439
proposed centers(c1,c2) for this cluster
159.2454 153.5738 141.2419
153.7302 152.9628 143.6362
centers found by using kmeans
159.1364 153.5909 141.3636
153.4211 152.8947 143.6842
A(Z) 0.437369
A*(Z) 0.439777
since test statistic is smaller than cv we do not reject H0 and discard{c1,c2}
-----Final centers for the clusters-----

130.0278 117.8889 111.75
123.9189 131.973 114.5676
115.1212 124.2727 108.1212
138.2955 134.2955 120.7273
115.3939 123.0303 125.8788
101.3953 121.7209 104.6744
63.14286 82.21429 78.78571
71.37838 89.78378 83.40541
66.65517 90.10345 84
74.68182 97.25 89.18182
103.56 110.84 112.48
84.88889 104.3333 96.66667
183.551 185.4694 171.9388
199.1111 177.4444 148.7778
204.3721 207 189.7907
174.0556 145.8889 129.2778
147.2222 145.1667 136.3333
173.6571 170.2571 154.4571
165.907 169.7674 163.6512
160.9245 160.1132 149.9811
156.4878 153.2683 142.439
> |
```

CLUSPLOT(X)




These two components explain 99.4 % of the point variability.



The clusters generated during the iterations have been attached in the zip file and have been labelled accordingly.

Checking other clustering algorithms :

```
> library(mclust)
 version 5.1
Type 'citation("mclust")' for citing this R package in publications.
> d_clust <- Mclust(as.matrix(X), G=1:20)
> m.best <- dim(d_clust$z)[2]
> cat("model-based optimal number of clusters:", m.best, "\n")
model-based optimal number of clusters: 12
> # 4 clusters
> plot(d_clust)
Model-based clustering plots:

1: BIC
2: classification
3: uncertainty
4: density

selection: 1
Model-based clustering plots:

1: BIC
2: classification
3: uncertainty
4: density

selection: 2
Model-based clustering plots:

1: BIC
2: classification
3: uncertainty
4: density
```

Optimal number of clusters suggested by model based clustering is 12.