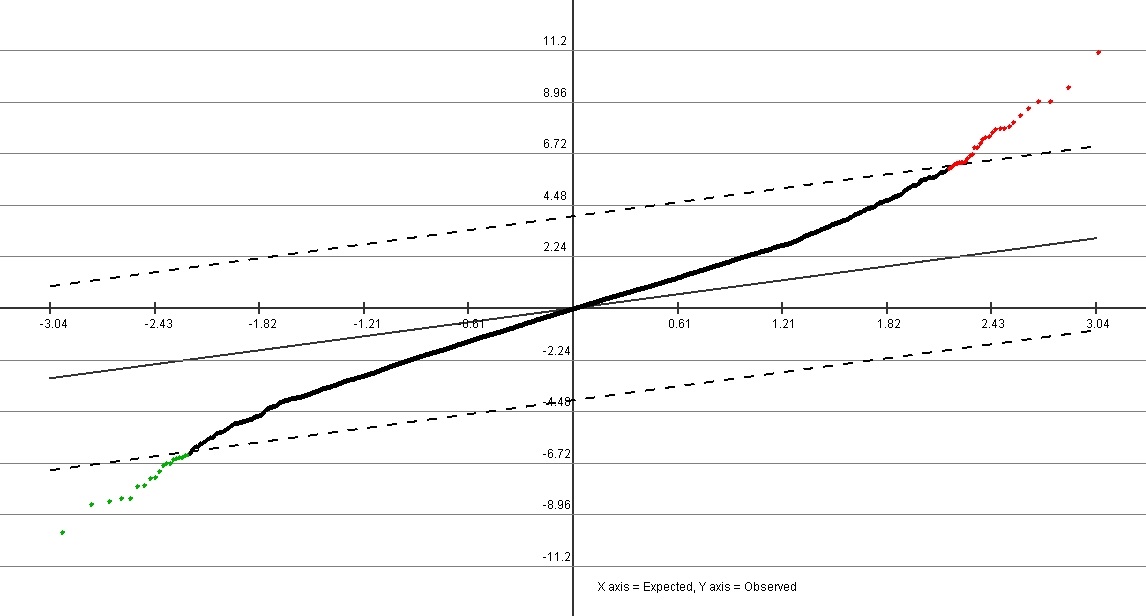
**Home Work-5**

**Mahmood Zaman (E01407909)**

**Part-1**

Select the 50 genes from golub test data using SAM and used for all the analysis.



**Part 2: HC with 3 linkages and 2 distances:**

**% average linkage & correlation distance**

>> tr50 = load('golub\_test50.txt');

>> labels = zeros(34, 1);

>> dist = pdist(tr50, 'correlation');

>> tree = linkage(dist, 'average');

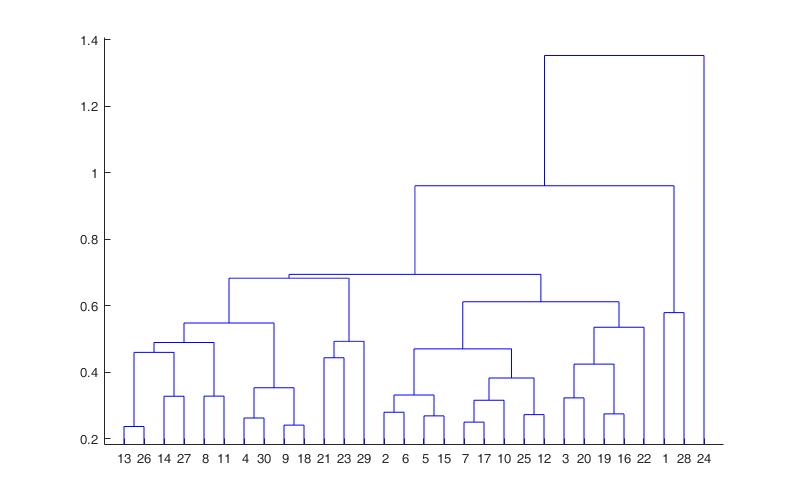
>> clusters = cluster(tree, 'maxclust', 2);

>> labels(1:20, 1) = 2;

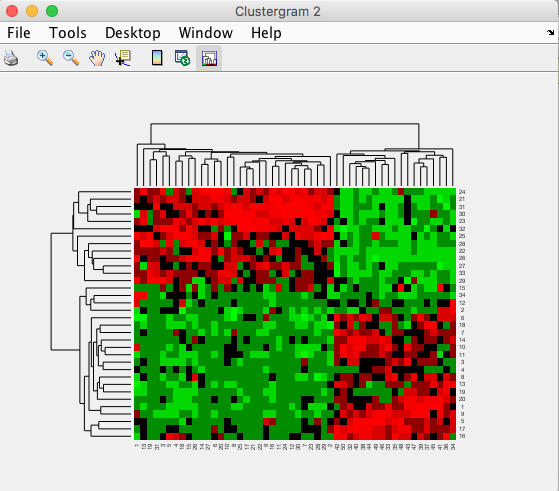
labels(21:34, 1) = 1;

>> error = nnz(labels - clusters)/length(labels)\*100

error = 2.9412

>> dendrogram(tree); 

>>clustergram(tr50, 'linkage', 'average', 'standardize', 'col');



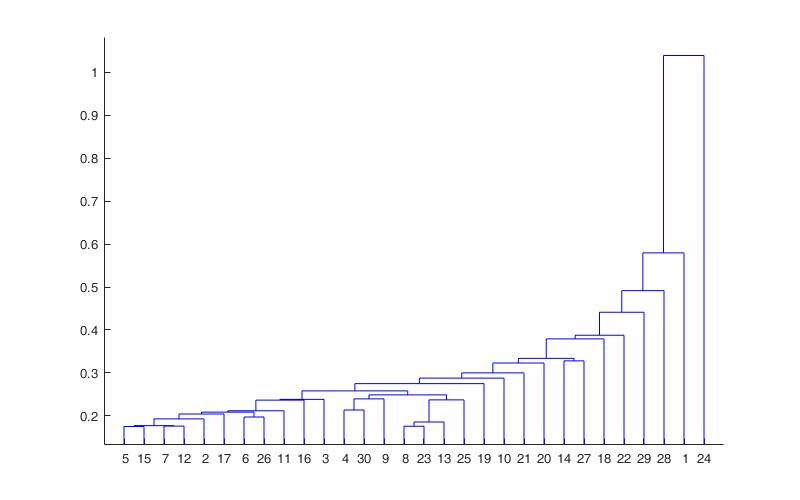
**% single linkage correlation**

>> tree = linkage(dist, 'single');

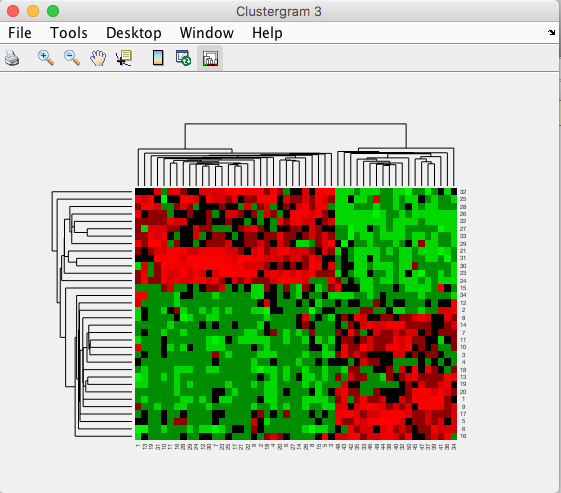
>> clusters = cluster(tree, 'maxclust', 2);

error = nnz(labels - clusters)/length(labels)\*100

error = 38.2353 % v high

>> dendrogram(tree); 

>> clustergram(tr50, 'linkage', 'sigle', 'standardize', 'col'); >>



**% complete linkage correlation**

>> tree = linkage(dist, 'complete');

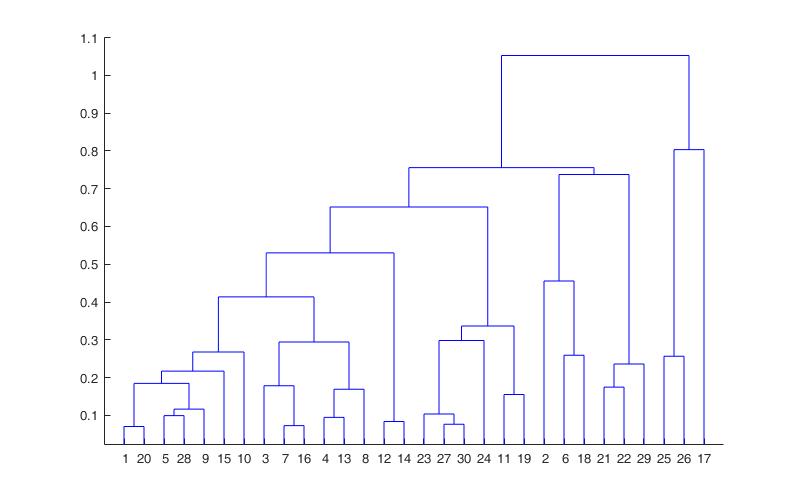
>> clusters = cluster(tree, 'maxclust', 2);

>> error = nnz(labels - clusters)/length(labels)\*100

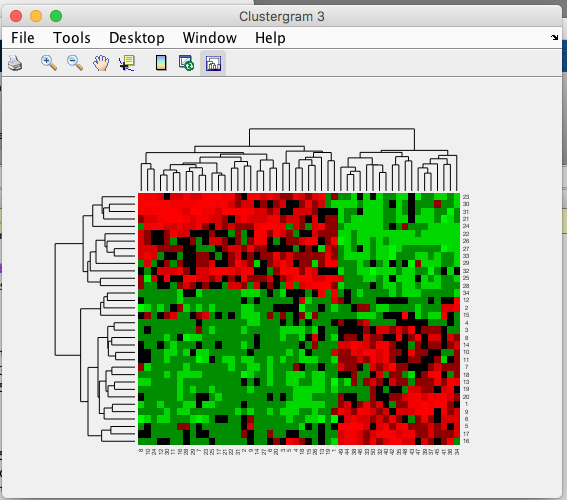
error =

2.9412

>> dendrogram(tree);



>> clustergram(tr50, 'linkage', 'complete', 'standardize', 'col');



**% single linkage euclidean**

>> dist = pdist(tr50, 'euclidean');

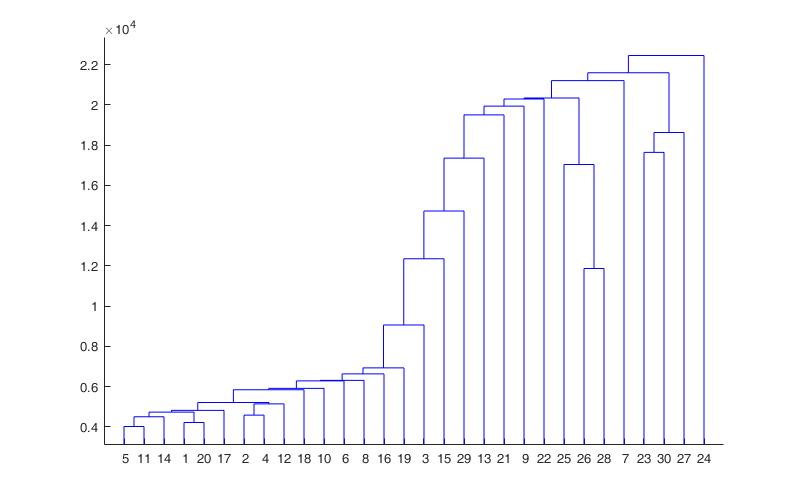
>> tree = linkage(dist, 'single');

>> clusters = cluster(tree, 'maxclust', 2);

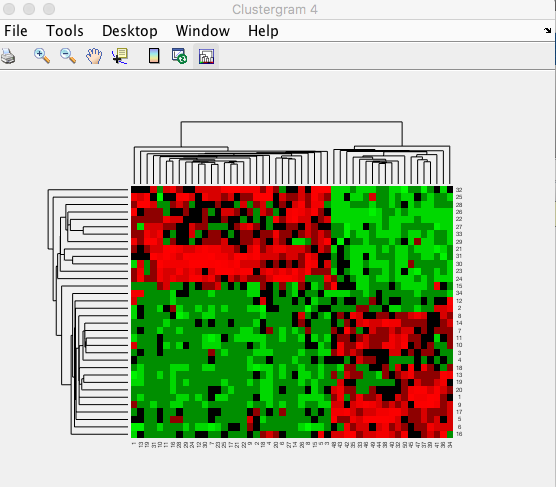
>> error = nnz(labels - clusters)/length(labels)\*100

error =

38.2353

>> dendrogram(tree); 

clustergram(tr50, 'linkage', single, 'standardize', 'col');



**% average linkage euclidean**

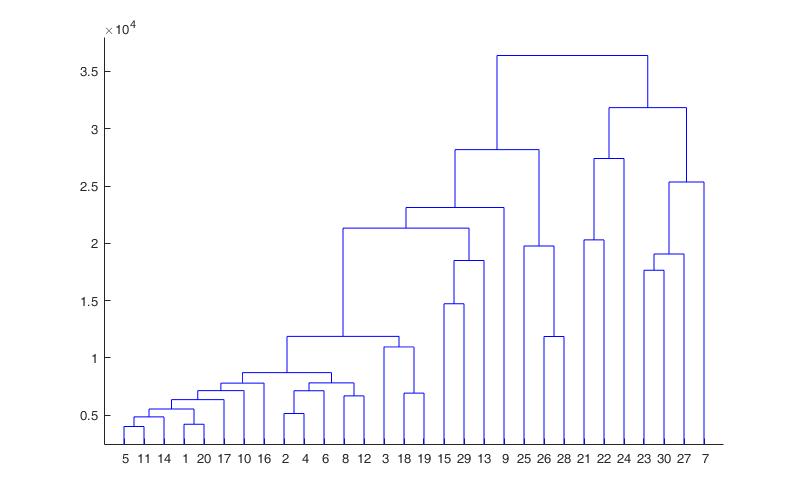
>> tree = linkage(dist, 'average');

>> clusters = cluster(tree, 'maxclust', 2);

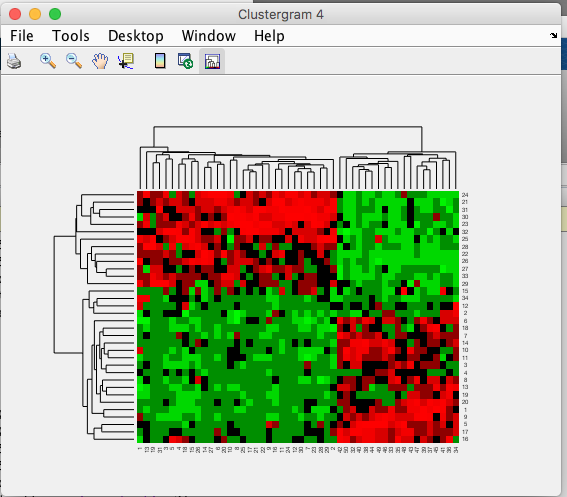
>> error = nnz(labels - clusters)/length(labels)\*100

error =

5.8824

>> dendrogram(tree); 

clustergram(tr50, 'linkage', 'average', 'standardize', 'col');



**% complete linkage euclidean**

>> tree = linkage(dist, 'complete');

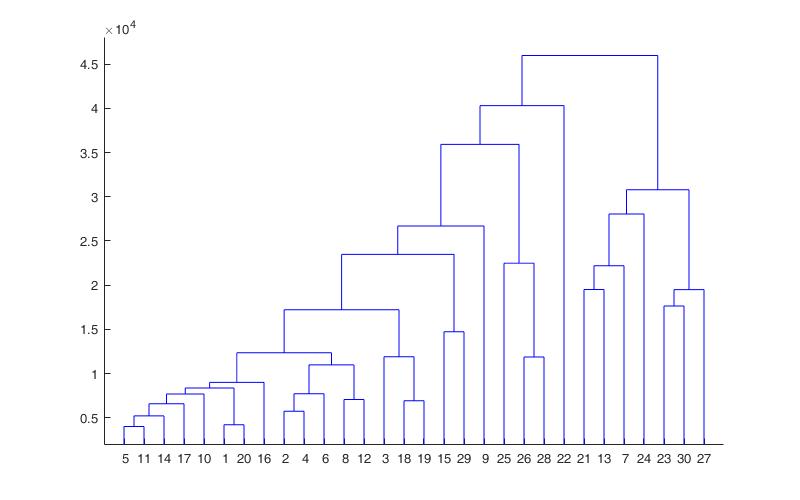
>> clusters = cluster(tree, 'maxclust', 2);

>> error = nnz(labels - clusters)/length(labels)\*100

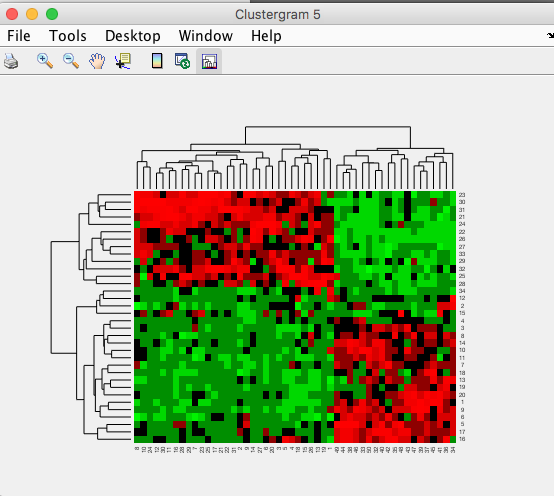
error =

5.8824

>> dendrogram(tree);



clustergram(tr50, 'linkage', complete, 'standardize', 'col');



**part3:K Mean Clustering**

**% taking distance as correlation**

>> ts50 = load('golub\_test50.txt');

>> labels = zeros(34, 1);

>> clusters = kmeans(ts50, 2, 'dist', 'correlation');

>> [clusters, centers] = kmeans(ts50, 2, 'dist', 'correlation');

>> labels(1:20, 1) = 1;

>> labels(21:34, 1) = 2;

>> error = nnz(labels - clusters)/length(labels)\*100

error =

2.941

>> [clusters, centers] = kmeans(ts50, 2, 'dist', 'correlation');

>> error = nnz(labels - clusters)/length(labels)\*100

error =

2.941

>> [clusters, centers] = kmeans(ts50, 2, 'dist', 'correlation');

>> labels(1:20, 1) = 2;

>> labels(21:34, 1) = 1;

>> error = nnz(labels - clusters)/length(labels)\*100

error = 2.941

**% taking distance as sqeuclidean**

>> [clusters, centers] = kmeans(ts50, 2, 'dist', 'sqeuclidean');

>> labels(1:20, 1) = 2;

>> labels(21:34, 1) = 1;

>> error = nnz(labels - clusters)/length(labels)\*100

error =2.941

>> [clusters, centers] = kmeans(ts50, 2, 'dist', 'sqeuclidean');

>> error = nnz(labels - clusters)/length(labels)\*100

error = 2.941

>> [clusters, centers] = kmeans(ts50, 2, 'dist', 'sqeuclidean');

>> error = nnz(labels - clusters)/length(labels)\*100

error = 2.941

**Part 4: Apply PCA to get 2D scatter plot**

**%apply on test data with origional clusters(1-20,21-34)**

>> PC = princomp(ts50);

>> ts502D = ts50 \* PC(:, 1:2);

>> ts502Dclass = [zeros(34,1) ts502D];

>> ts502Dclass(1:20,1)=1;

>> ts502Dclass(21:34,1)=2;

>> plotClass2D(ts502Dclass, 1, 10);

**% Applying to HC to get scattter plot**

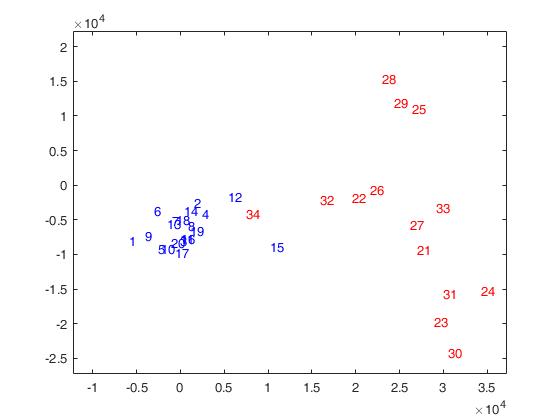
>>ts502Dclass(1:34,1)=clusters;

>> plotClass2D(ts502Dclass, 1, 10);

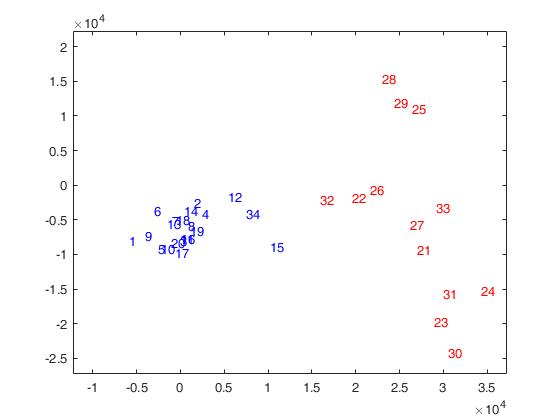
**% Applying to KMC to get scattter plot**

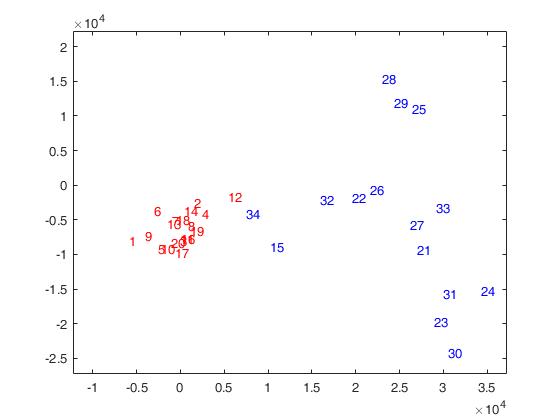
>>ts502Dclass(1:34,1)=clusters;

>> plotClass2D(ts502Dclass, 1, 10);

**%original clusters**

**% Applying to HC to get scattter plot(corr,avg)**



**% Applying to KMC to get scattter plot(corr, k=2)**

**Note:** We can see from scatter plots points 12, 34 and 15 are swing points that are changing cluters on applying HC and KMC othewise both results are pretty much same.

Below is the summary of all results:

|  |  |  |
| --- | --- | --- |
| **Method of analysis** | **Distance** | **Error %** |
| HC single linkage | correlation | 38.23 |
| HC avg linkage | Correlation | 2.941 |
| HC complete Linkage | Correlation | 2.941 |
| HC single linkage | equclidean | 38.23 |
| HC avg linkage | equclidean | 2.941 |
| HC complete Linkage | equclidean | 2.941 |
| KMC with k=2 | Correlation | 2.941 |
| KMC with k=2 | Correlation | 2.941 |
| KMC with k=2 | Correlation | 2.941 |
| KMC with k=2 | sqeuclidean | 2.941 |
| KMC with k=2 | sqeuclidean | 2.941 |
| KMC with k=2 | sqeuclidean | 2.941 |

Only single linkage for both method HC and KMC increased error otherwise rest of the result are really close

**Part 5(SOM):**

>> Q=ts50;

>> Q=Q';

>> net = newsom([min(Q,[],2) max(Q,[],2)], [2 1], 'hextop');

>> net.trainParam.epochs = 100;

>> net = train(net, Q);

>> clusters = vec2ind(net(Q));

>> clusters =  clusters'

>> labels(21:34, 1) = 1;

>> labels(21:34, 1) = 1;

>> labels(21:34, 1) = 2;

>> error = nnz(labels - clusters)/length(labels)\*100

**error = 2.9412**

>> net = train(net, Q);

>> clusters = vec2ind(net(Q));

>> clusters =  clusters';

>> error = nnz(labels - clusters)/length(labels)\*100

**error =2.9412**

>> net.trainParam.epochs = 10;

>> net = train(net, Q);

>> clusters = vec2ind(net(Q));

>> clusters =  clusters';

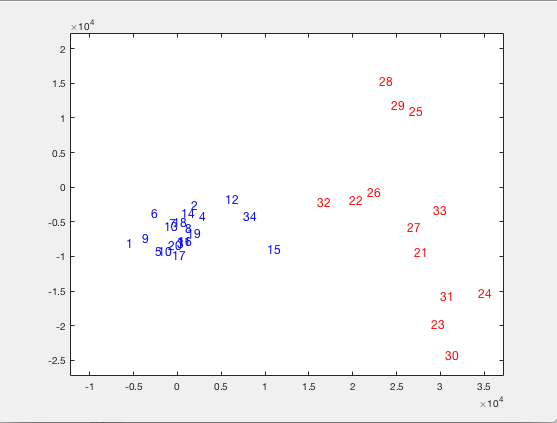
>> labels(1:21, 1) = 2;

>> labels(21:34, 1) = 1;

>> error = nnz(labels - clusters)/length(labels)\*100

**error = 5.8824**

Note: SOM results for first two test are same as HC and KMC and we get the error of 2.94 but notice that with low no of iteration we getting more error. Below scatter plot for SOM shows all swing poits(12,34,15) go to blue cluster.



**Conclusion:**

In all this exercise we apply different clustering algorithms on our significant 50 genes and we find out that there are some swing points(genes) that change their cluster position but most of our results are supporting each other for their correctness because we did all this exercise on the data that we already know about. More over not a single analysis can be relied unless we verify our result with some other analusis technique.For example in this exaample for apply three method with different paramets and verify our result.