Assignment 6: Principal Component Analysis and Cluster Analysis

Mats Hansson

May 18, 2015

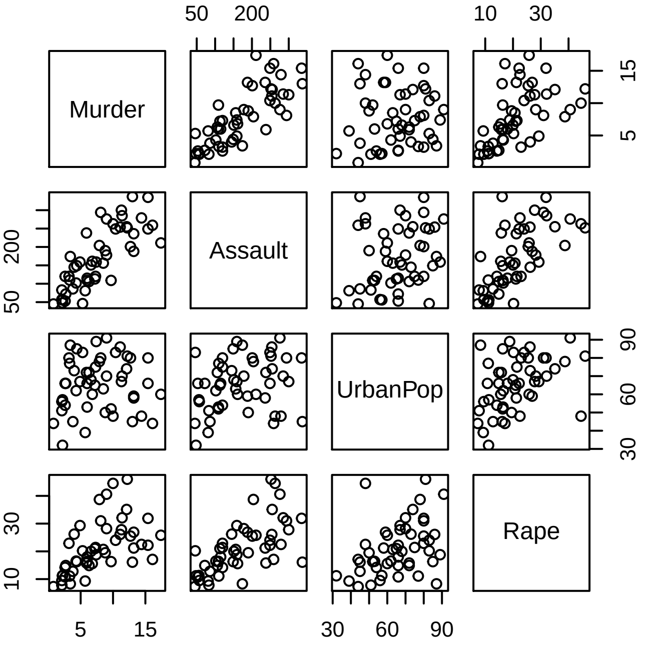
Problem 1: Hierarchical clustering

The data is loaded from ISLR packages, the file includes 50 row and 4 columns and there are no missing value. The summary of the data is:

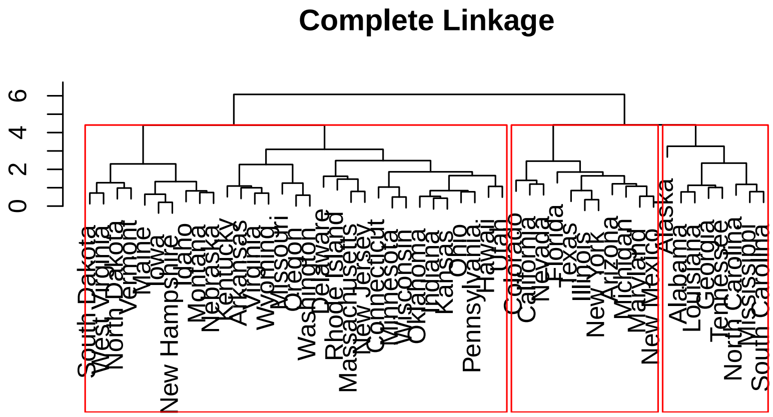
	Mean	Variance	lowest	highest
Murder	7.79	18.97	0.80	17.40
Assault	170.76	6945.17	45.00	337.00
UrbanPop	65.54	209.52	32.00	91.00
Rape	21.23	87.73	7.30	46.00

We can see that the variables have different mean and variance, therefore it should been standardized. The Correlation matrix and the scatterplot are:

	Murder	Assault	UrbanPop	Rape
Murder	1.00	0.80	0.07	0.56
Assault	0.80	1.00	0.26	0.67
UrbanPop	0.07	0.26	1.00	0.41
Rape	0.56	0.67	0.41	1.00



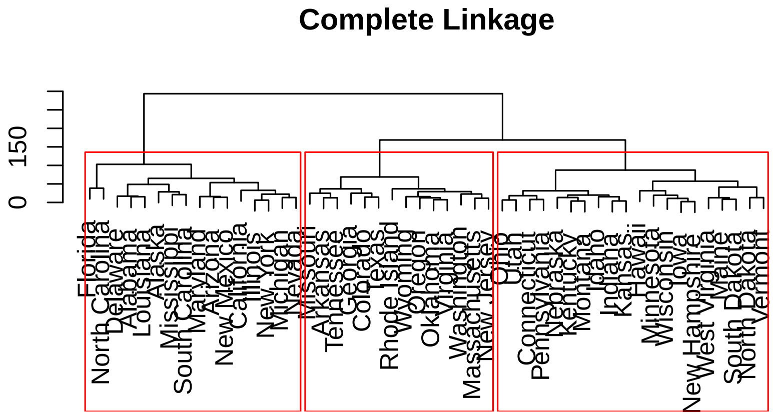
There is some correlation between Murder, assault and rape, there are also some correlation between assault and UrbanPop. We can also see in the scatterplot that the data is heteroscedastic. The dendrogram for the standardized data is:



	number
1	8
2	11
3	31

In the picture above we can see that each leaf represent one of the 50 observations, when we move up in the tree some of the leaf begin to fuse into branches. For example Mississippi and South Carolina seems to be similar, because it is lower in the tree and when the branches is higher in the tree than the observations do not have to be similar. Therefore we should only analysis the tree by looking at vertical and not horizontal. The first group in the dendrogram is to the right, the next in the middle and the third to the left.

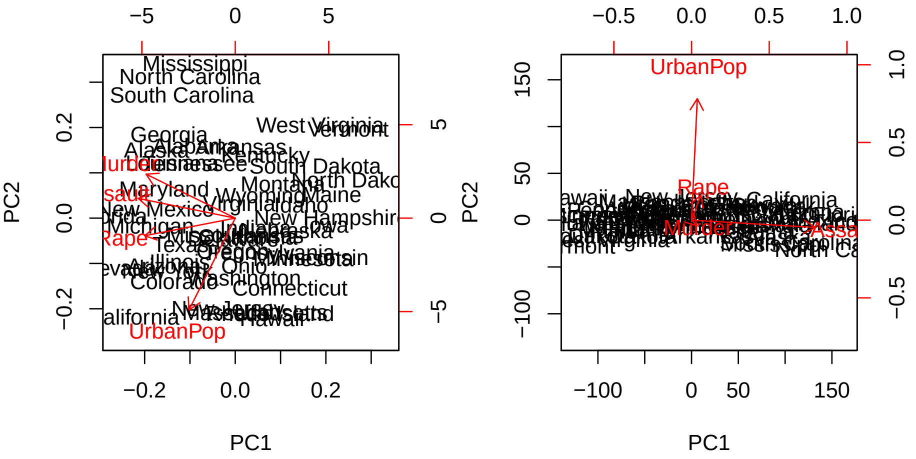
Dendogram for the not standardized data:



	number
1	16
2	14
3	20

There are 3 groups in the dendrogram if we cut the tree at 150. Than the first group is to the right, the next in the middle and the third to the left.

We can see that these two dendrogram is not similar. Therefore one of these dendrogram is misleading. The biplot for the standardized data and the not standardized data are:



The picture to the left is the biplot for the standardized data, where we can see that Rape, Assult and Murder have more loadings for PC 1 compare to PC2. This can also been interpreted as there is more correlation between these variables. The others picture is for the not standardized data, where UrbanPop and Assault have higher loadings compare to Rape and Murder and this is because the variance for UrbanPop and Assault are greater than Rape and Murder. Therefore it's better to standardized the data.

Problem 2: Principal Component Analysis and K-means

The files includes 25 rows and 10 columns and there are no missing value. The summary of the data is:

	Mean	Variance	lowest	highest
Country				
RedMeat	9.83	11.20	4.40	18.00
WhiteMeat	7.90	13.65	1.40	14.00
Eggs	2.94	1.25	0.50	4.70
Milk	17.11	50.49	4.90	33.70
Fish	4.28	11.58	0.20	14.20
Cereals	32.25	120.45	18.60	56.70
Starch	4.28	2.67	0.60	6.50
Nuts	3.07	3.94	0.70	7.80
Fr.Veg	4.14	3.25	1.40	7.90

We can see in the table that the mean and the variance is different between the variables. I choose to transform the data so the row names is the country name. The correlation matrix for the data is:

	RedMeat	WhiteMeat	Eggs	Milk	Fish	Cereals	Starch	Nuts	Fr.Veg
RedMeat	1.00	0.15	0.59	0.50	0.06	-0.50	0.14	-0.35	-0.07
WhiteMeat	0.15	1.00	0.62	0.28	-0.23	-0.41	0.31	-0.63	-0.06
Eggs	0.59	0.62	1.00	0.58	0.07	-0.71	0.45	-0.56	-0.05
Milk	0.50	0.28	0.58	1.00	0.14	-0.59	0.22	-0.62	-0.41
Fish	0.06	-0.23	0.07	0.14	1.00	-0.52	0.40	-0.15	0.27
Cereals	-0.50	-0.41	-0.71	-0.59	-0.52	1.00	-0.53	0.65	0.05
Starch	0.14	0.31	0.45	0.22	0.40	-0.53	1.00	-0.47	0.08
Nuts	-0.35	-0.63	-0.56	-0.62	-0.15	0.65	-0.47	1.00	0.37
Fr.Veg	-0.07	-0.06	-0.05	-0.41	0.27	0.05	0.08	0.37	1.00

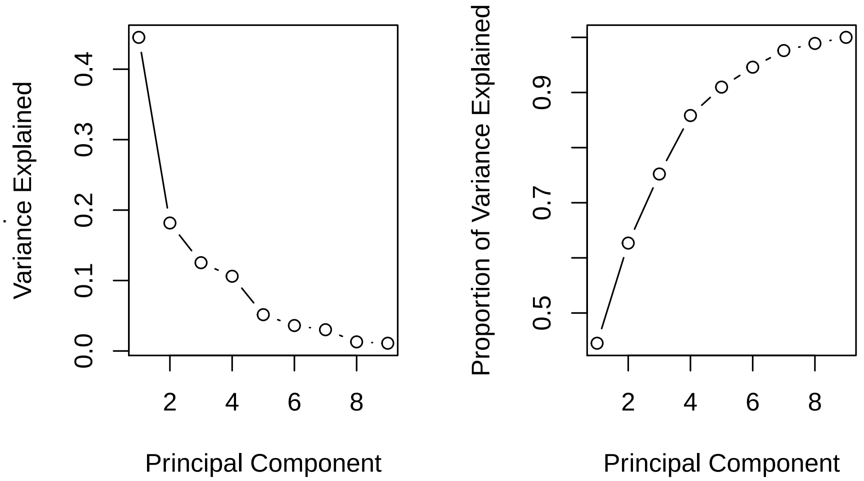
There are some correlation between the variables. I choose to standardized the data before I estimate the eigenvalue and eigenvector. The result is:

	1	2	3	4	5	6	7	8	9
PC	PC1	PC2	PC3	PC4	PC5,	PC6	PC7	PC8	PC9
	2.002	1.279	1.062	0.977	0.681	0.57	0.521	0.341	0.315

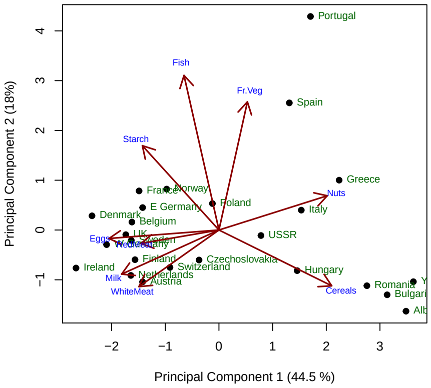
The first value is the first pricipal component and the second value is the second principal component. We can also see that the three first principal component seems to explain the most of the variation. The eigen vector is:

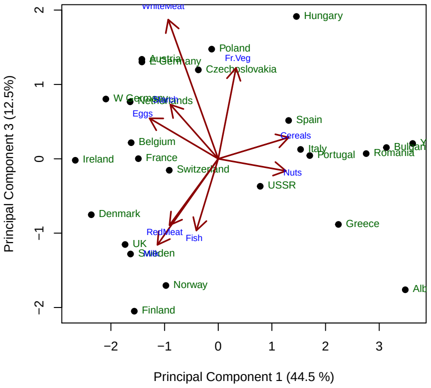
	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9
RedMeat	-0.30	-0.06	-0.30	-0.65	0.32	-0.46	0.15	-0.02	0.25
WhiteMeat	-0.31	-0.24	0.62	0.04	-0.30	-0.12	-0.02	-0.03	0.59
Eggs	-0.43	-0.04	0.18	-0.31	0.08	0.36	-0.44	-0.49	-0.33
Milk	-0.38	-0.18	-0.39	0.00	-0.20	0.62	0.46	0.08	0.18
Fish	-0.14	0.65	-0.32	0.22	-0.29	-0.14	-0.11	-0.45	0.31
Cereals	0.44	-0.23	0.10	0.01	0.24	0.08	0.40	-0.70	0.15
Starch	-0.30	0.35	0.24	0.34	0.74	0.15	0.15	0.12	0.12
Nuts	0.42	0.14	-0.05	-0.33	0.15	0.45	-0.41	0.18	0.52
Fr.Veg	0.11	0.54	0.41	-0.46	-0.23	0.12	0.45	0.09	-0.20

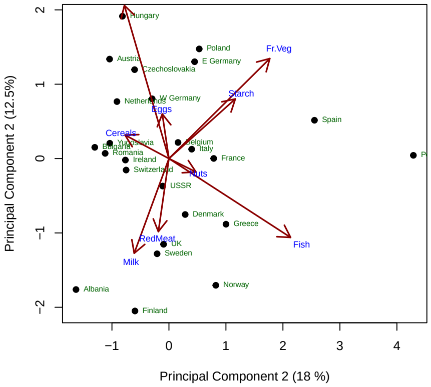
The Eigen vector is the loadings for each variables. Therefore we can see that Read meat and Eggs have approximate the same loadings and White meat and Eggs are also approximate the same. The loadings are calculated by the covariance matrix and therefore we can also say that Read meat and Eggs are correlated. The Scree plot for the eigen value is:



The picture to the left is the variance explain for each component, where the first component explain 44.5 % and the second component explain 18 %. The picture to the right is the proportions of the variance explain. There is many way to decide how many principal component that should include in the analysis, one way is to look in picture for the "elbow", Where we can see that there is an elbow at third component, this three component can explain 75 % of the variation. The score plots for the component are:







We can see in the picture on the top left that Romania, Bulgaria, Albania, Yugoslavia is cluster together. There seems also to be some cluster between Portugal, Spain, and Greek and the third group is the other countries.

The arrows is the loadings (eigen vector), where we can see in the first picture that Eggs, Red meat, Milk, Cereals and Nuts have more loadings for the first component. The second component seems to explain Fr.Veg and Fish. This indicates that the Red meat, Milk, Cereals and Nuts are more correlated with each other's compare to the others.

An example how the biplot works, the table is for the 4 largest values for Cereals:

	RedMeat	WhiteMeat	Eggs	Milk	Fish	Cereals	Starch	Nuts	Fr.Veg
Bulgaria	7.80	6.00	1.60	8.30	1.20	56.70	1.10	3.70	4.20
Romania	6.20	6.30	1.50	11.10	1.00	49.60	3.10	5.30	2.80
USSR	9.30	4.60	2.10	16.60	3.00	43.60	6.40	3.40	2.90
Yugoslavia	4.40	5.00	1.20	9.50	0.60	55.90	3.00	5.70	3.20

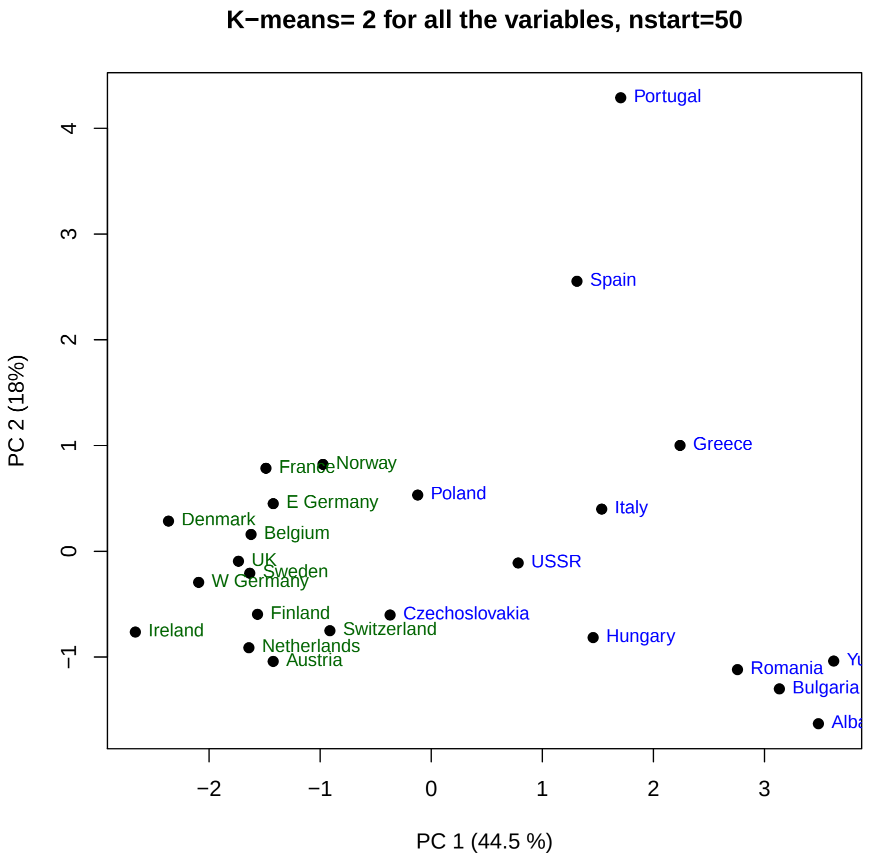
These countries are Romania, Bulgaria, Albania and Yugoslavia. Therefore the Cereals arrows is also pointing at which country do have the largest value.

K-means clustering:

K-means clustering for the original data, where k=2, and nstart=50. The cluster means is:

	1	2
RedMeat	7.68	11.81
WhiteMeat	6.04	9.61
Eggs	2.10	3.71
Milk	11.72	22.08
Fish	3.48	5.02
Cereals	41.11	24.07
Starch	3.75	4.76
Nuts	4.57	1.69
Fr.Veg	4.83	3.50

The cluster mean is different between the variables, where there are 9 variables. Each of the observations is cluster to the closest cluster mean (where is defined using the Euclidean distance), therefore it's harder to visualize the data because we have more than 2 dimension. Therefore I choose to vizulize my result by using the two first principal components score. The result is:



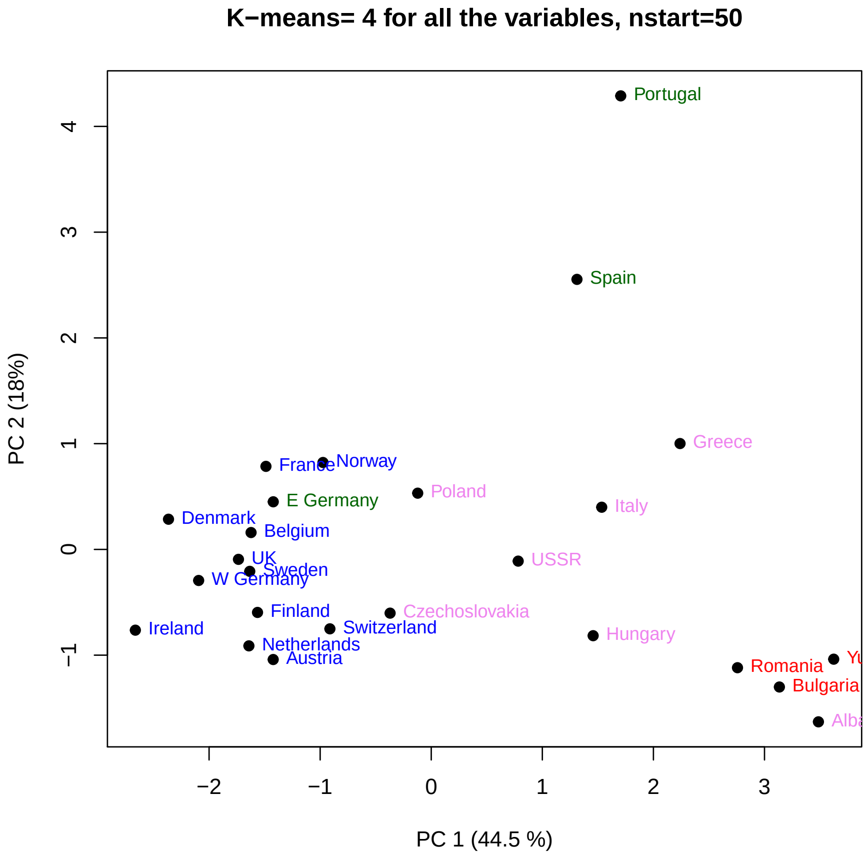
We can see in the picture above that K-means cluster with 9 variables can separate two groups. I choose also to plot the countries names because it's giving me a better hint how the countries is separated.

The variance for the right group is 1668.0933 and the left is 808.6554, where we can see that the variance is smaller for the blue group compare to the green. Therefore my two first components seem to capture all the information to classify two groups. This model can also explain 52.8 % of the variation, where the variance within is 2476.749 (each group has 808.6554 and 1668.0933) and the total variance is 5243.414.

For K=4, nstart=50:
The cluster means is:

	1	2	3	4
RedMeat	12.09	7.23	6.13	8.64
WhiteMeat	9.44	6.23	5.77	6.87
Eggs	3.71	2.63	1.43	2.39
Milk	23.00	8.20	9.63	14.04
Fish	4.99	8.87	0.93	2.54
Cereals	24.03	26.93	54.07	39.27
Starch	4.62	6.03	2.40	3.74
Nuts	1.77	3.80	4.90	4.21
Fr.Veg	3.49	6.23	3.40	4.66

There are 4 cluster means for each of the variables. Each of the 25 observations is cluster whose centroid is the closest. I choose to visualize my result by using my two first component, the result is:



We can see in the picture that there are 4 clusters (green, blue, red and pink). The total variance is the same as before (5243.414) and the total variance within is 1269.05 (each groups have 47.0000, 656.4517, 148.6067 and 416.9914). Therefore my new model can capture more of the variation, where this model can explain 75.8 % of the variation.

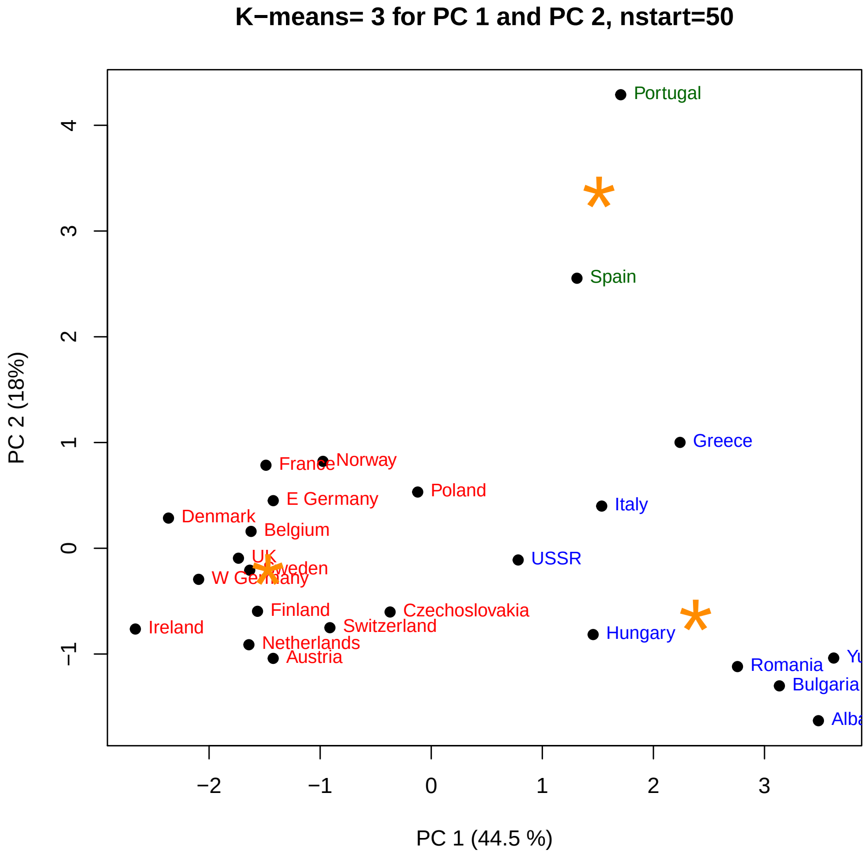
The group to the left is approximate the same as when k=2 and the variance for this group is 656.4517. The difference is that E Germany is now excluded from this group.

It seems that my two first components do not capture all of the information to explain my prediction, where we can see that E Germany is far away from Spain and Portugal.

K-means cluster for two principal component:
The group mean is:

	1	2	3
PC1	2.38	1.51	-1.47
PC2	-0.58	3.42	-0.15

There are two group mean for each of the component, where we explain 62.5 % of the variation. The advantage with this method is easier to visualize (2 dimensions). The plot is:



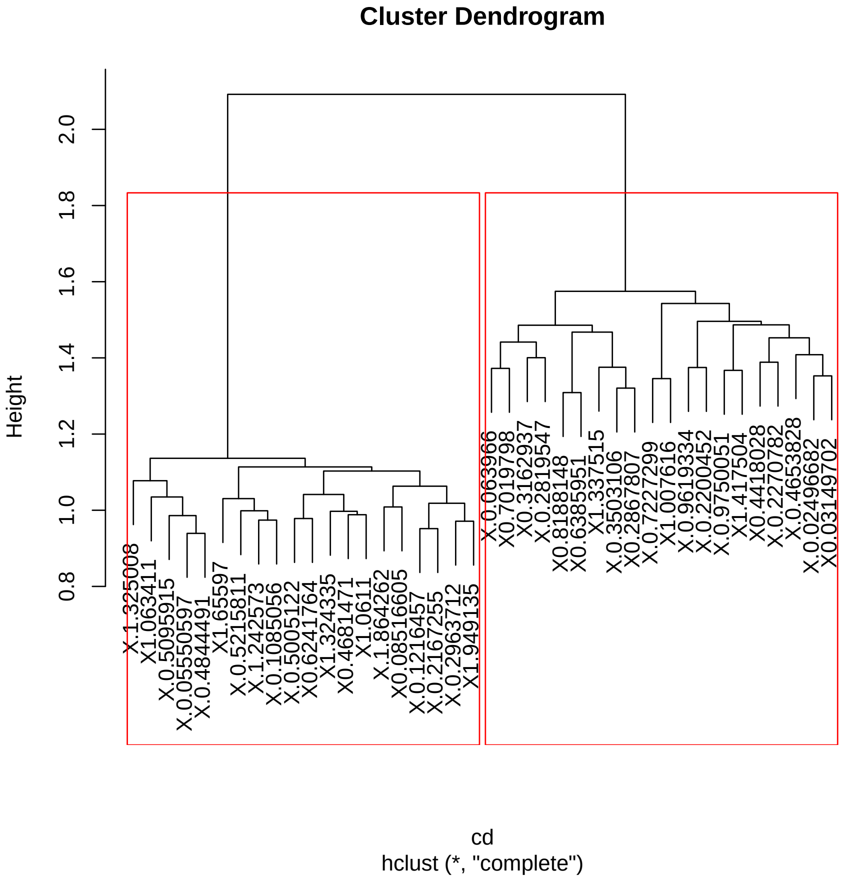
We can see that the countries are cluster in three groups. The star in the plot is the cluster mean, where it seems to be in the center of each groups. The total variance is 135.3945 and the total variance within is 26.89083 (each group have 1.58346, 11.83520, 13.47216), the model explain 80.1 % of the variation.

conclusion

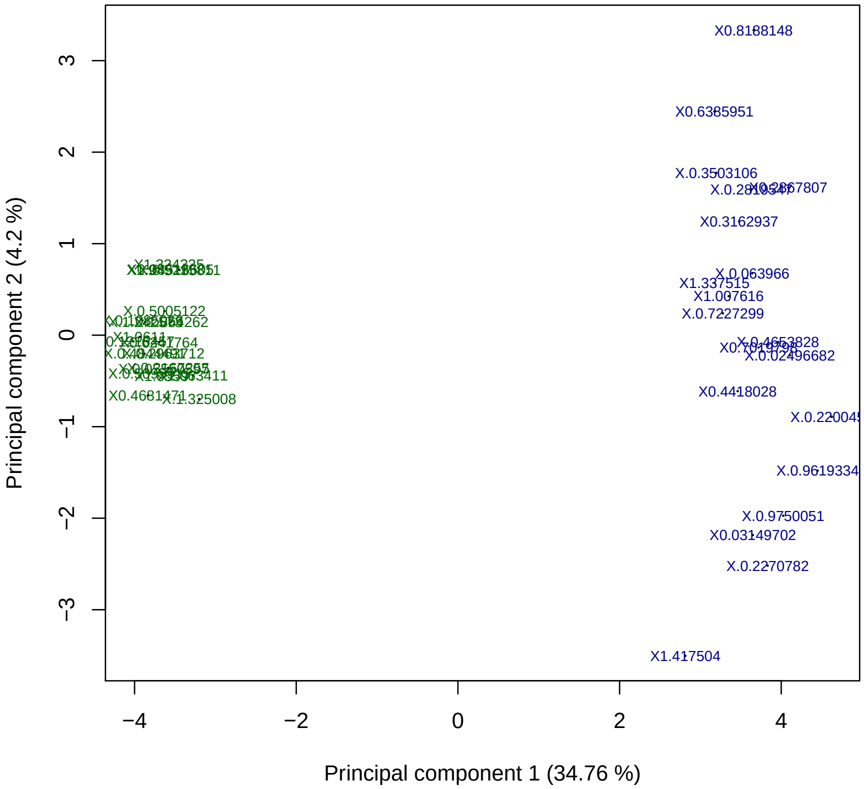
The different between K-means cluster on the original data or the principal component, is that the two first component explain 62.5 % of the variation. Therefore we can't explain all of the variation with the principal component. The advantage with the principal component it's easy to visualize and also to get a hint how the countries are separated. The disadvantage to classify my original data than it's hard to visualize. In my opinion the principal component it a better tool to classify the groups, because an image explains so much more than just some table.

Problem 3: Hierarchical Cluster

The data is loaded and the file includes 40 columns and 999 rows. The dendrogram for the correlation matrix is:

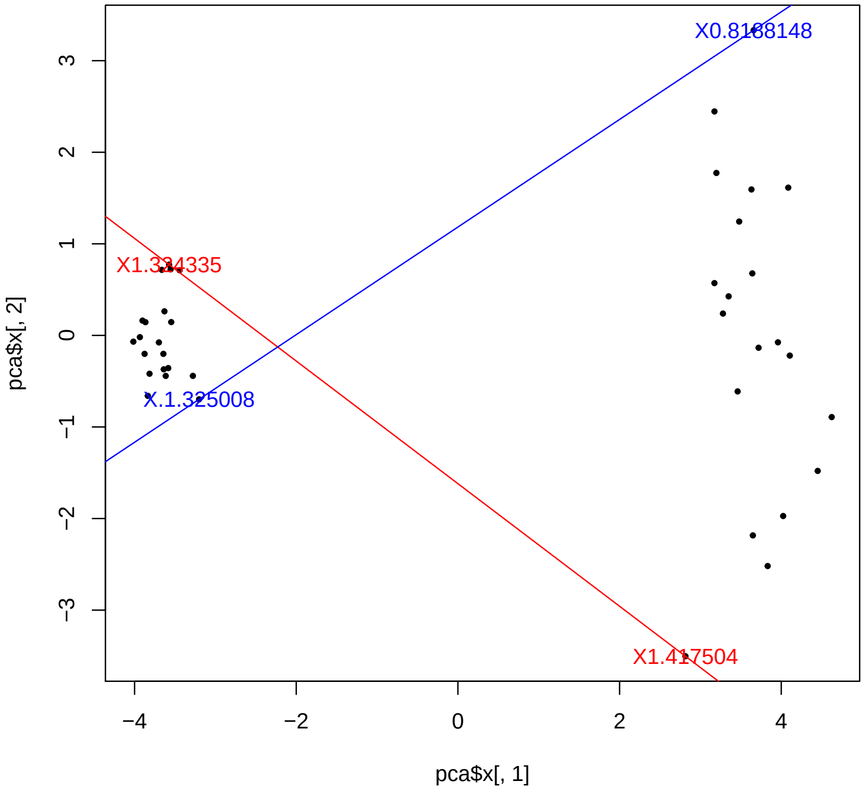


It seems to be some different between the protein where we can see in the dendrogram that there are 2 clear group at the height 1.8. This indicates that the genes to the left is more correlated to the genes in the same cluster group. I use my two first principal component to plot my score point to see which genes that differ most across the two group, the plot is:



The two first principal components explain 38.96% of the variation. These indicate this is enough to separate two groups, where the left group (green) has a lower variance compare to the right (blue).

We can also see that gene X0.8188148 and X.417504 is far away in the score plot, and this can we not see in the dendrogram, where both of these genes are in the same cluster group. The left group seems also to be more correlated to each other's compare to the right group, where this group has a lower variance. The gene that differs most across the group is:



The blue and the red genes are the genes that differ most across the group, where the blue line and the red line are the distance between the observations.

One other way to see which gene that differ most across the groups is to sum up the weights of each loadings and search for the gene that has the highest loadings. The 10 genes that differ most across the groups are:

```
## X.0.02496682  X.0.7227299   X.1.242573      X1.0611  X.0.4653828
##     2.242866      2.195542     2.103835     1.697906     1.600286
##  X.0.9750051  X.0.3503106  X.0.5095915  X.0.2167255      X1.65597
##     1.364476      1.354418     1.321750     1.228446     1.201213
```

We can see that X.0.02496682 has the most weight and the second is X.0.7227299, this indicate these are the 10 most different genes across the two groups.

conclusion

We can see in the dendrogram and the principal component score that we can see some unique pattern for the genes, where it can cluster in two groups. It seems also that the left group has a lower variance compare to the right. But the dendrogram seems to have some problem to visualize the variance in each group.

There are two way to see which gene that differ most across the groups, the first way is the principal component score and search for the largest distance between the genes. The second alternative is to analysis the loadings and search for the highest weight for the gene.