Zahra Jalil Pour

Email: h19zahja@du.se

Home Exercise 3

Statistical Learning (AMI22T)

Problem 1:

a:

In this part of problem1, we have a simple linear model, $y_i=\beta_0+\beta_1 x_i+\epsilon_i$

By having β_0= -1 , β_1= 2 , x_i ~N(0,1) and ϵ_i ~N(0,1) , we can simulate 100 observations from this model. The output of this linear model is:

y_i= -1+2 x_i+rnorm(100)

```
lm(formula = y ~ x)

Residuals:
     Min       1Q   Median       3Q      Max
-2.05195 -0.43265 -0.07854  0.48583  1.93858

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.98855    0.07929  -12.47   <2e-16 ***
x            1.89463    0.07807   24.27   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7929 on 98 degrees of freedom
Multiple R-squared:  0.8573, Adjusted R-squared:  0.8559
F-statistic: 588.9 on 1 and 98 DF,  p-value: < 2.2e-16
```

The slope of this model equals to 1.89463 .
Now we want to run another model by variables around their means.

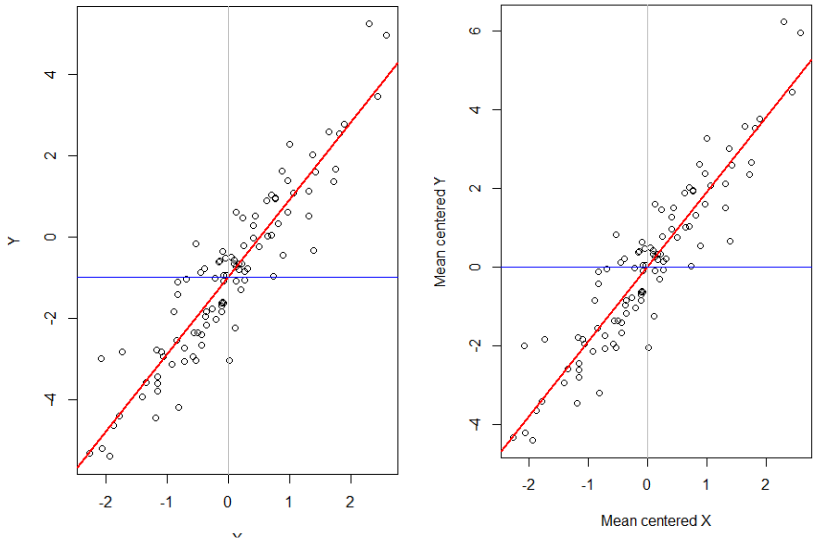
```
lm(formula = v ~ u)

Residuals:
     Min       1Q   Median       3Q      Max
-2.05195 -0.43265 -0.07854  0.48583  1.93858

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 4.207e-17  7.929e-02    0.00        1
u           1.895e+00  7.807e-02   24.27   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7929 on 98 degrees of freedom
Multiple R-squared:  0.8573, Adjusted R-squared:  0.8559
F-statistic: 588.9 on 1 and 98 DF,  p-value: < 2.2e-16
```

The slope in the centered version of the model equals to 1.895e+00 .In his case the model has the centered version. Centering simply means subtracting a constant from every value of a variable. What it does is redefine the 0 point for that predictor to be whatever value you subtracted. It shifts the scale over, but retains the units. The effect is that the slope between that predictor and the response variable doesn't change at all. But the interpretation of the intercept does. Figure below shows the observed data and regression line(red line). The blue line is mean of Y, and the gray line is mean of X. if we subtract the mean from X, the regression line will shift toward left.



And if we subtract the mean from Y, the regression line will shift down. The slope will keep the same. Sometimes it is useful to measure the independent variable around its mean. The centered version of the model will be as below:

$$y_1 = \beta_0 + \beta_1(x_i - \bar{x}) + \beta_1\bar{x} + \epsilon$$
$$= \beta^*_0 + \beta_1(x_i - \bar{x}) + \epsilon_i \quad , \quad \beta^*_0 = \beta_0 + \beta_1\bar{x}$$

The sum of the squares due to error is given by:

$$S(\beta^*_0, \beta_1) = \Sigma\epsilon_i^2 = \Sigma[y_i - \beta^*_0 - \beta_1(x_i - \bar{x})]^2$$

Now solving:

$$\partial S(\beta^*_0, \beta_1)/\partial\beta^*_0 = 0$$
$$\partial S(\beta^*_0, \beta_1)/\partial\beta_1 = 0$$

we get the direct regression least squares estimates of β^*_0 and β_1 as:

$$\beta^*_0 = \bar{y} \quad \text{and} \quad \beta_1 = S_{xy}/S_{xx}$$

respectively. Thus the form of the estimate of slope parameter β_1 remains the same in the usual and centered models whereas the form of the estimate of intercept term changes in the usual and centered models.

Problem1:

b:

The intercept parameter in the regression of v on u is 0. We run the lm(v~u) and lm(u~v). In both of them we see the t-statistic value in both of them are equal.

lm(v~u):

```
lm(formula = v ~ u)

Residuals:
     Min       1Q   Median       3Q      Max
-2.05195 -0.43265 -0.07854  0.48583  1.93858

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 4.207e-17  7.929e-02    0.00        1
u           1.895e+00  7.807e-02   24.27   <2e-16 ***
---
```

lm(u~v):

```
lm(formula = u ~ v)

Residuals:
     Min       1Q   Median       3Q      Max
-1.17358 -0.22120  0.02454  0.21603  1.10483

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -2.220e-17  3.875e-02    0.00        1
v            4.525e-01  1.865e-02   24.27   <2e-16 ***
```

This is not coincidence and we expect this equality:

$$\hat{\beta} = (\Sigma U_i V_i) \,/\, (\Sigma U_i{}'^2) \qquad (1)$$

$$t = \hat{\beta}/(SE(\hat{\beta})) \qquad (2)$$

$$SE(\beta) = \sqrt{(\Sigma (V_i - U_i\beta)^2/((n-1)\Sigma U_i{}'^2)} \qquad (3)$$

Here we put the equation (3) in (2):

$$t = \hat{\beta}/(\sqrt{(\Sigma (V_i - U_i\hat{\beta})^2/((n-1)\Sigma U_i{}'^2)}) \qquad (4)$$

Now we square both sides of (4):

$$t^2 = \hat{\beta}^2/((\Sigma (V_i - U_i\hat{\beta})^2/((n-1)\Sigma U_i{}'^2)) \rightarrow t^2 = (n-1)((\Sigma U_i{}'^2)\hat{\beta}^2/(\Sigma (V_i - U_i\hat{\beta})^2))$$

$$\rightarrow t^2 = (n-1)((\Sigma U_i{}'^2)\hat{\beta}^{2)}/\Sigma(V_i^2 - 2V_iU_i\hat{\beta} + U_i^2\hat{\beta}^2) = (n-1)\hat{\beta}^2(\Sigma U_i{}'^2)/(\Sigma V_i^2$$

$$-2\hat{\beta}\Sigma V_iU_i + \hat{\beta}^2\Sigma U_i^2)$$

$$t^2=(n-1)\hat{\beta}^2(\Sigma U_{i'}{}^2)/(\Sigma V_i{}^2+\hat{\beta}(\hat{\beta}\Sigma U_i{}^2-2\Sigma V_i U_i)$$

In the above equation, we insert (1):

$$t^2=(n-1)\hat{\beta}^2(\Sigma U_{i'}{}^2)/(\Sigma V_i{}^2+\hat{\beta}((\Sigma U_i V_i)\ /\ (\Sigma U_{i'}{}^2)).\Sigma U_i{}^2-2\Sigma V_i U_i)$$

$$t^2=(n-1)\hat{\beta}^2(\Sigma U_{i'}{}^2)/(\Sigma V_i{}^2-\hat{\beta}(\Sigma U_i V_i))$$

Again we insert (1) in the above equation:

$$t^2=(n-1)\hat{\beta}^2(\Sigma U_{i'}{}^2)/(\Sigma V_i{}^2-((\Sigma U_i V_i)\ /\ (\Sigma U_{i'}{}^2)).(\Sigma U_i V_i))=$$

$$= (n-1)\hat{\beta}^2(\Sigma U_{i'}{}^2)^2/(\Sigma V_i{}^2 U_{i'}{}^2-(\Sigma U_i V_i)^2)$$

In the upper side of the above equation we insert (1) again:

$$= (n-1)((\Sigma U_i V_i)\ /\ (\Sigma U_{i'}{}^2)^2)(\Sigma U_{i'}{}^2)^2/(\Sigma V_i{}^2 U_{i'}{}^2-(\Sigma U_i V_i)^2)$$

$$= (n-1)((\Sigma U_i V_i)^2)/(\Sigma V_i{}^2 U_{i'}{}^2-(\Sigma U_i V_i)^2)$$

$$\rightarrow t= \sqrt{(n-1)}\ (\Sigma U_i V_i)/(\sqrt{\Sigma V_i{}^2 U_{i'}{}^2}-(\Sigma U_i V_i)^2)$$

Now if we do regression onto U on V , we will have the same t-statistic:

$$t= \sqrt{(n-1)}\ (\Sigma V_i U_i/(\sqrt{\Sigma U_{i'}{}^2 V_i{}^2}-(\Sigma V_i U_i)^2)$$

Problem2:
Big_five_personality_traits

Introduction

In psychological trait theory, the Big Five personality traits, also known as the five-factor model (FFM) and the OCEAN model, is a suggested taxonomy, or grouping, for personality traits,[1] developed from the 1980s onwards. This theory uses descriptors of common language and suggests five broad dimensions commonly used to describe the human personality and psyche.[2][3]

The theory identifies five factors:

- openness to experience (inventive/curious vs. consistent/cautious)
- conscientiousness (efficient/organized vs. extravagant/careless)
- extraversion (outgoing/energetic vs. solitary/reserved)
- agreeableness (friendly/compassionate vs. challenging/callous)
- neuroticism (sensitive/nervous vs. resilient/confident)

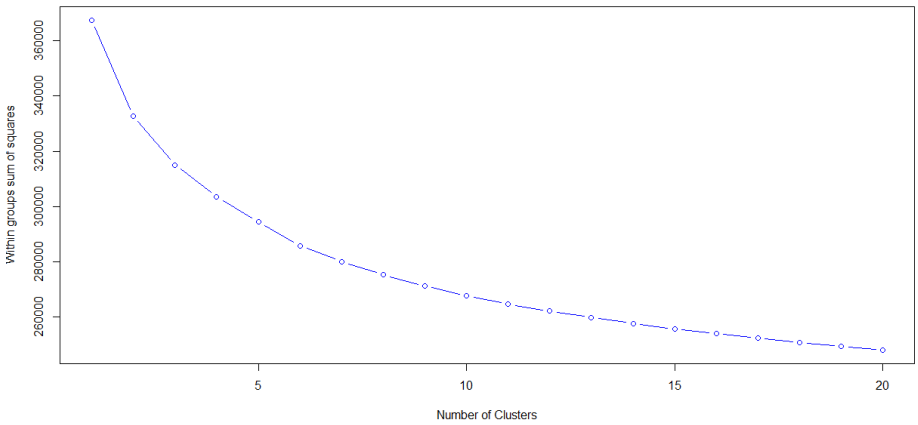
In this survey more than 1000000 individuals took an online personality test consisting of 50 statements. The aim of this project is classification of people based on their responses to the online test.

Data description

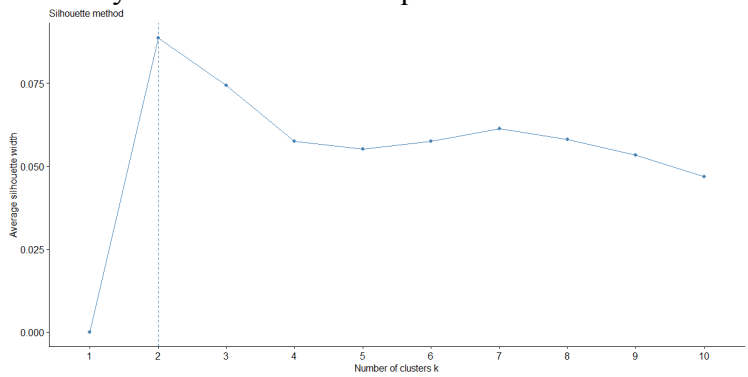
Dataset consists of 1015338 observations with 50 features. Part of the dataset is shown as below:

	EXT1	EXT2	EXT3	EXT4	EXT5	EXT6	EXT7	EXT8	EXT9	EXT10	EST1	EST2	EST3	EST4
1	4	1	5	2	5	1	5	2	4	1	1	4	4	2
2	3	5	3	4	3	3	2	5	1	5	2	3	4	1
3	2	3	4	4	3	2	1	3	2	5	4	4	4	2
4	2	2	2	3	4	2	2	4	1	4	3	3	3	2
5	3	3	3	3	5	3	3	5	3	4	1	5	5	3
6	3	3	4	2	4	2	2	3	3	4	3	4	3	2
7	4	3	4	3	3	3	5	3	4	3	2	4	4	2
8	3	1	5	2	5	2	5	2	3	2	2	4	2	4
9	2	2	3	3	4	2	2	2	4	4	3	4	4	1
10	1	5	3	5	2	3	2	4	5	4	3	3	3	3

Cleaning the data:
At first we should remove unnecessary features. Here we need all features from EXT1 to OPN10 ,then we should omit missing values in the dataset. As the main data set is quite large, I do sampling 5000 observations. For classification, it is better to normalize our data.
We can find the number of clusters by different method, like Elbow method, Silhouette method and Gap statistic



As you can see from elbow method,3 to 6 clusters looks optimum for the data set and we already know this research is to identify 5 different personalities. The optimal number of cluster by Silhouette method equals 2.

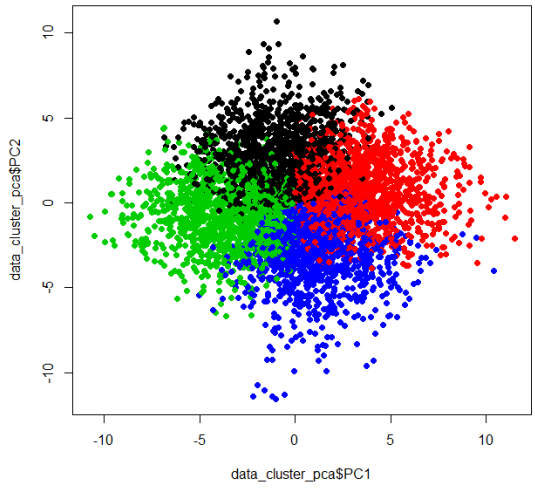


In the next step, I apply k-means clustering by the number of cluster=4 based on the elbow method.

Size of each cluster in 5000 samples is as below:

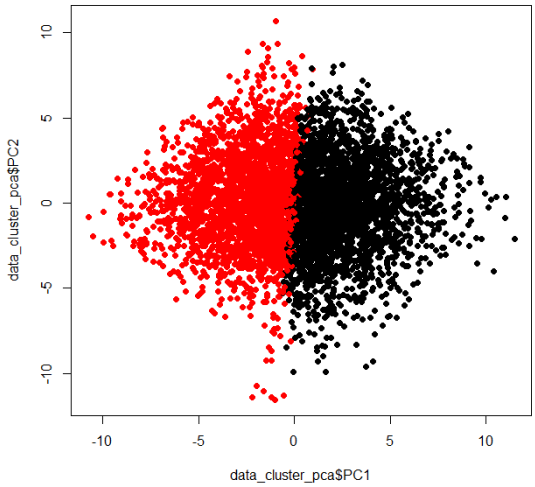
Cluster1	Cluster2	Cluster3	Cluster4
1219	1381	1291	1109

As the number of features are more than 2, we use PCA and plot the first two principal components score vectors. We will see the below figure by number of cluster=4.



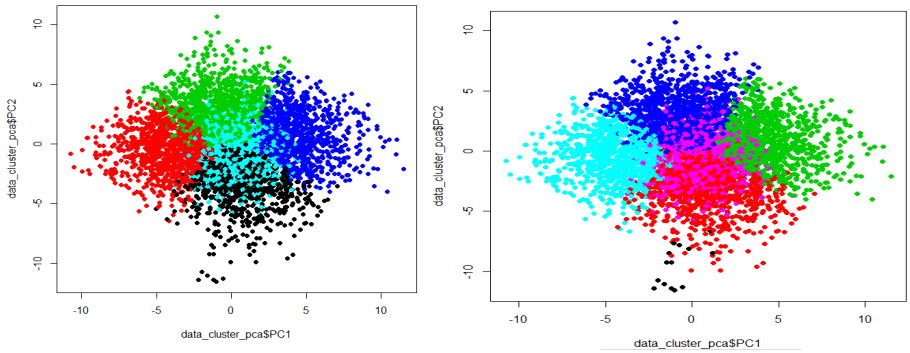
Now we consider the number of cluster=2, as Silhouette method:

Cluster1	Cluster2
2537	2463



In elbow method, we see that the number of K is between 4 to 6, hence we examine the number of cluster=5 and 6.

K=5 k=6



As we see, when the number of cluster is 5 or 6, the clusters overlap with each other. Hence 2 clusters or 4 clusters are optimal number of clusters.

For validating the clusters we should assign appropriate labels and add new column as label. Here we copy our data frame and create new data set. Based on the concept of the question we score each question and summaries the columns to EXT, EST, AGR, CSN, OPN, then add new column named cluster as below(head of new data frame):

```
       EXT EST AGR CSN OPN Cluster
970947  16   3  23  -8  20       2
188997  10  29  22  -1  16       2
134058   3  19  22  15  17       2
124022   0   1  14  -3  13       2
687065  14   8  25  10  20       2
227071  11  13  23   1  13       2
```

Now we calculate the means of each features:

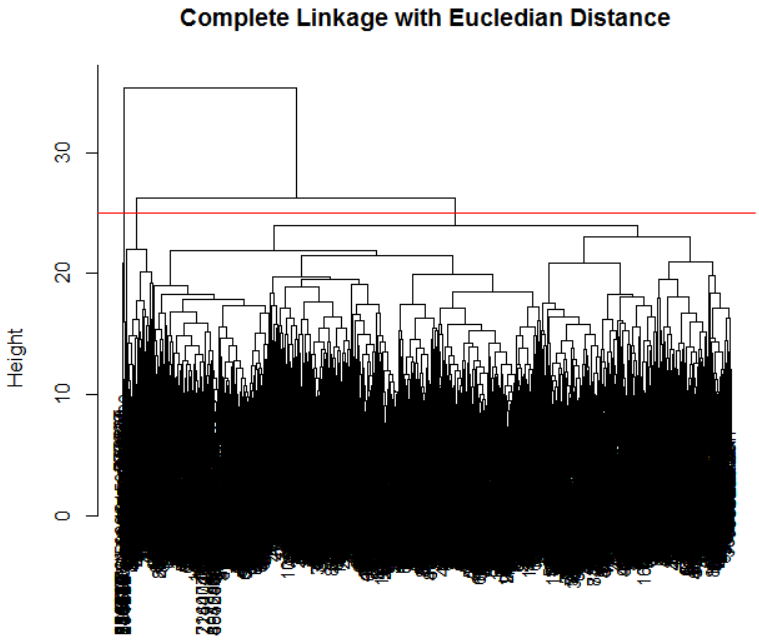
	Cluster	EXT	EST	AGR	CSN	OPN
1	1	-6.553015	22.49113	15.97635	7.657864	13.30390
2	2	5.927324	14.73975	20.33374	11.289484	15.38124

We apply this procedure for k=4. The mean of each feature is as below:

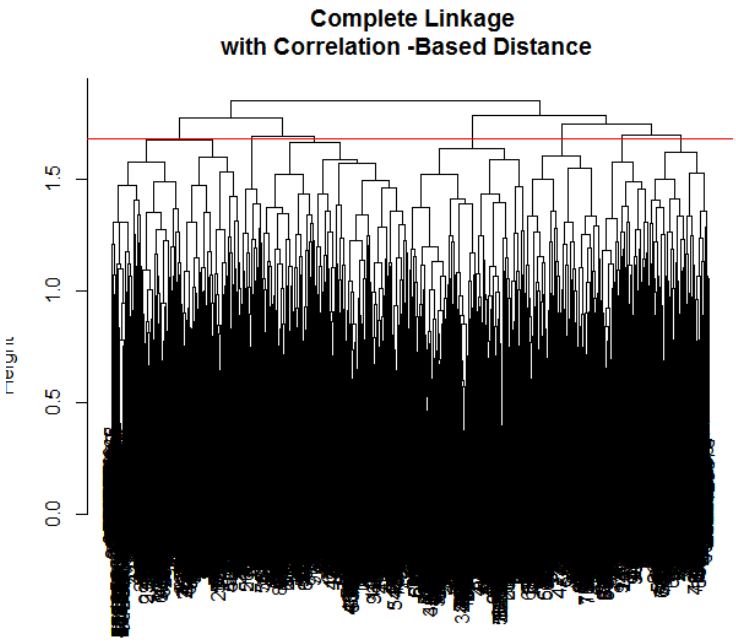
	Cluster	EXT	EST	AGR	CSN	OPN
1	1	7.412633	21.96473	19.91879	5.049221	15.51518
2	2	-7.940623	26.32875	17.42578	7.886314	13.14699
3	3	4.730442	12.31294	22.01936	15.649109	15.79009
4	4	-5.593327	12.92426	12.48061	9.003607	12.78810

Hierarchical clustering:

This algorithm produces a hierarchy for the observations of the dataset. For assessing the similarity between observations, a distance matrix is required to quantify closeness. Referring the dendrogram, a suitable height shall be chosen where a horizontal line shall be plotted and the number of instances where the horizontal line cut the branches of dendrogram are clusters of our data. The number of cluster by considering method= complete , equals 3.



In the dendrogram displayed above, each leaf corresponds to one observation. As we move up the tree, observations that are similar to each other are combined into branches, which are themselves fused at a higher height. The height of the fusion, provided on the vertical axis, indicates the (dist)similarity between two observations. The higher the height of the fusion, the less similar the observations are. Now we instead of using Euclidean distance, I'd like to use correlation_based distance. The number of cluster will be 7.



If we consider method= average, the number of cluster equals 3, that is near the number of cluster in k_means clustering.

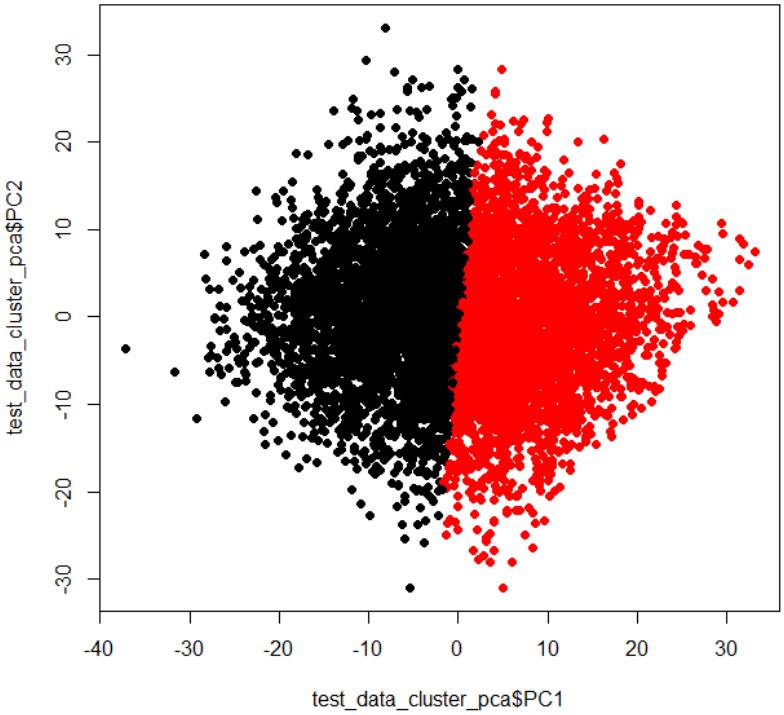
Now we should make a different subset to make prediction the cluster of new observation. We make a new test subset with 7000 observations.

Then by considering k=2, we perform clustering:

```
K-means clustering with 2 clusters of sizes 3580, 3420

Cluster means:
        EXT      OPN        CSN      AGR      EST
1 -6.475419 13.47765  8.089385 16.13464 22.34860
2  6.399708 15.42661 11.335088 20.10585 14.40439
```

Like previous procedure, as the number of features are more than 2, for plotting we apply PCA.



When we assign new column for test new dataset, the classification is:

	EXT	OPN	CSN	AGR	EST	Cluster
781757	-13	14	19	9	21	1
902630	-20	13	9	1	24	1
411239	-3	6	6	18	24	1
115829	0	23	23	17	12	2
814303	4	14	12	18	10	2
885330	-3	17	21	19	15	2
627435	0	4	9	19	22	1
228116	-14	17	-3	15	16	1
272904	4	19	2	25	19	2
193803	-4	12	17	31	11	2
279408	15	20	13	23	21	2

And the means of each feature is as below:

	Cluster	EXT	OPN	CSN	AGR	EST
1	1	-6.475419	13.47765	8.089385	16.13464	22.34860
2	2	6.399708	15.42661	11.335088	20.10585	14.40439

If we consider the number of cluster equals 4 , the means of each feature in new test data set is as below:

	Cluster	EXT	OPN	CSN	AGR	EST
1	1	8.267399	16.17766	14.482906	21.30342	11.28510
2	2	-5.763756	13.90191	11.014952	15.27811	11.88218
3	3	-8.972823	13.44759	8.066001	16.16694	25.98558
4	4	5.817700	14.31902	5.852146	19.57340	23.35294