Mini-Project 2 Final Checkpoint

ECE/CS 498DS

Spring 2020

Akhilesh Somani (somani4)
Gowtham Kuntumalla (gowtham4)
Manan Mehta (mananm2)

Task 1 - Question 0

1. Why do biologists need multiple samples to identify microbes with significantly altered abundance?

Biologists need multiple samples to be sure that the data is statistically significant. Hypothesis needs to be backed by data. This helps them to conclude, with a greater confidence, which microbes are present in more numbers than usual.

2. Number of samples analyzed (in context of HE0): 764 samples
3. Number of microbes identified: 149

Task 1 – Question 1

- a. Factorization of joint probability distribution:

T = storage temp, M = collection method, C = contamination, L = Lab time, Q = quality

$$P(joint) = P(Q, C, L, M, T) = P(Q|C, L, M, T) * P(C|M, T, L) * P(M|T, L) * P(T|L) * P(L)$$
$$P(Q, C, L, M, T) = P(Q|C, L) * P(C|M, T) * P(M) * P(T) * P(L)$$

- b. Number of parameters needed to define conditional probability distribution:

Number of values taken by: C=2, L=2, T=2, M=2, Q=2

For the CPTs:

P(Quality | Contamination, Lab Time): 2*2 = 4 parameters

P(Contamination | Storage Temp, Collection Method): 2*2 = 4 parameters

P(Storage Temp): 1 parameter

P(Collection Method): 1 parameter

P(Lab Time): 1 parameter

Thus, we need (4 + 4 + 1 + 1 + 1) = **11 parameters**

Task 1 – Question 1 (continued)

- C. Conditional probability tables:

P(Contamination | Storage Temp, Collection Method)

	strtmp	coll	cont = low	cont = high
0	cold	nurse	0.956017	0.043983
1	cold	patient	0.923423	0.076577
2	cool	nurse	0.911565	0.088435
3	cool	patient	0.161765	0.838235

P(Quality | Contamination, Lab Time)

	cont	labtime	qual = good	qual = bad
0	low	short	0.957093	0.042907
1	low	long	0.919003	0.080997
2	high	short	0.935743	0.064257
3	high	long	0.033898	0.966102

P(Storage Temp)	{'cold': 0.8982, 'cool': 0.1018}
P(Collection Method)	{'nurse': 0.8976, 'patient': 0.1024}
P(Lab Time)	{'short': 0.7956, 'long': 0.2044}

Task 1 – Question 1 (continued)

- d. Table of P(Quality|Storage Temp, Collection Method, Lab Time)

	strtmp	coll	labtime	qual = good	qual = bad
0	cold	nurse	short	0.955112	0.044888
1	cold	nurse	long	0.887962	0.112038
2	cold	patient	short	0.943978	0.056022
3	cold	patient	long	0.862069	0.137931
4	cool	nurse	short	0.972376	0.027624
5	cool	nurse	long	0.822785	0.177215
6	cool	patient	short	0.960784	0.039216
7	cool	patient	long	0.117647	0.882353

- e. Total number of samples dropped: 65 (for HE0) + 65 (for HE1) = 130 samples

Task 1 – Question 2

- 1. Number of samples removed: 0
- 2. What are the benefits and drawbacks to using relative abundance data? Is there information that we lose when the normalization is performed?

While using relative abundance data, we have scaled the variance of the data and hence, we give equal emphasis to the variation for each bacteria. This normalization gives us a constrained snapshot of the relative distributions of microbes in a specific sample. There is a problem in doing this. We do not know the exact number of the bacteria present, which may be important to know rather than just the relative abundance. For e.g. - A relative abundance of 0.5:0.5 might mean 100:100 bacteria or 100k:100k bacteria. If there is a constraint on the number of bacteria to do some analysis, then this information is lost by scaling it.

Task 1 – Question 3

- Heatmaps (HE0 on top and HE1 on bottom) (Microbes as rows):

Task 1 – Question 3 (continued)

- Summarize your observations

The heatmaps help in visualize at a glance the trend between the relative abundance of different bacteria in all the samples. The darker zones refer to low abundance and lighter zones correspond to higher abundance. A preliminary glance at the heatmaps tell us that the trend for relative abundance for all bacteria is same for both HE0 patients and HE1 patients. The heatmaps also show that the relative abundance of a particular bacteria among different samples is also same (which is expected because of data cleaning).

- Which aspects of the data are the heatmaps good at highlighting? What types of things are heatmaps less suitable for?

A heatmap is a graphical representation of data where values are depicted by color. Heatmaps make it easy to visualize complex data and understand it at a glance. The problem is that when we perceive shading, our brains tend to think in terms of relativities. That is, it notices sharp contrasts between adjacent bits of an image. However, we are poor at comparing shading in non-adjacent regions of a visualization.

Task 2 – Question 1

- b. What is the null hypothesis of the KS test in our context? Use one microbe as an example to explain your answer.

Ho for the KS Test is that the 2 samples tested are drawn from the same underlying distribution. In our context, it can be interpreted as no significantly altered expression of a particular microbe in the stool samples from HE0 and HE1 patients.

- c. Count the number of microbes with significantly altered expression at alpha=0.1, 0.05, 0.01, 0.005 and 0.001 level? Summarize your answers in a table below:

Alpha Level	Number of bacteria with altered expressions
0.1	50
0.05	37
0.01	27
0.005	26
0.001	21

Task 2 – Question 2

- a. What does a p-value of 0.05 represent in our context?

P-value, in general, is the probability of observing the test statistic or a more extreme value assuming $H0$ is true. In our context, a p-value of 0.05 represents a 5% probability of observing the KS test statistic (D-statistic), given that there is no significantly altered expression of the microbe in the HE0 and HE1 samples. In simple words, P-value of 0.05 represents 5% probability of rejecting H0 falsely. In our context, H0: for a microbe, both HE0 and HE1 sample follow same distribution.

- b. If the null hypothesis is true, what distribution will the p-values follow?

If the null hypothesis is true, the p-values will follow a unifom distribution. The reason is how we define αα as the probability of erroneously rejecting $H0$H0. We reject $H0$H0 when p-value < αα and the only way this holds for any value of αα is when p is uniformly distributed.

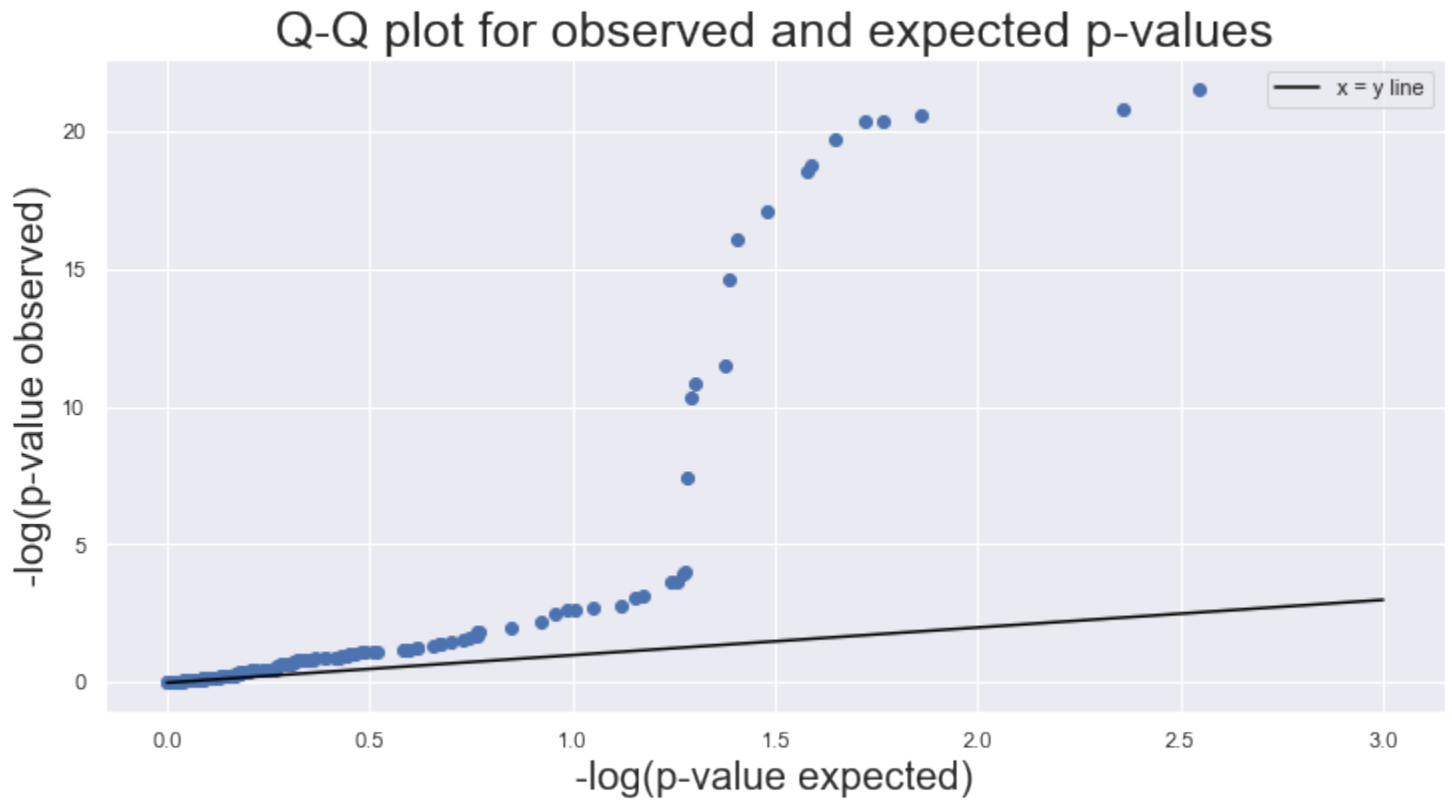
- c. If no microbe's abundance was altered, how many significant p-values does one expect to see at alpha=0.1, 0.05, 0.01, 0.005 and 0.001 level? Compare your answers with your results in Task 2.1.c. Show the comparison in a table below:

If no microbe's abundance was altered, which is to say that $H0$ is true, the significant p-values will be uniformly distributed. Thus, for an α value of 0.1, we expect to see 10% of the total number of samples (and so on). (We round the number of microbes to 150 instead of 149 here)

Alpha Level	# of Significant p-values if H_0 true	# observed from data in Task 2.1.c
0.1	15	50
0.05	8	37
0.01	2	27
0.005	1	26
0.001	0	21

Task 2 – Question 2 (continued)

- d. Q-Q plot:



Task 2 – Question 2 (continued)

- e.i. How does taking the -log10() of the p-values help you visualize the p-value distribution?

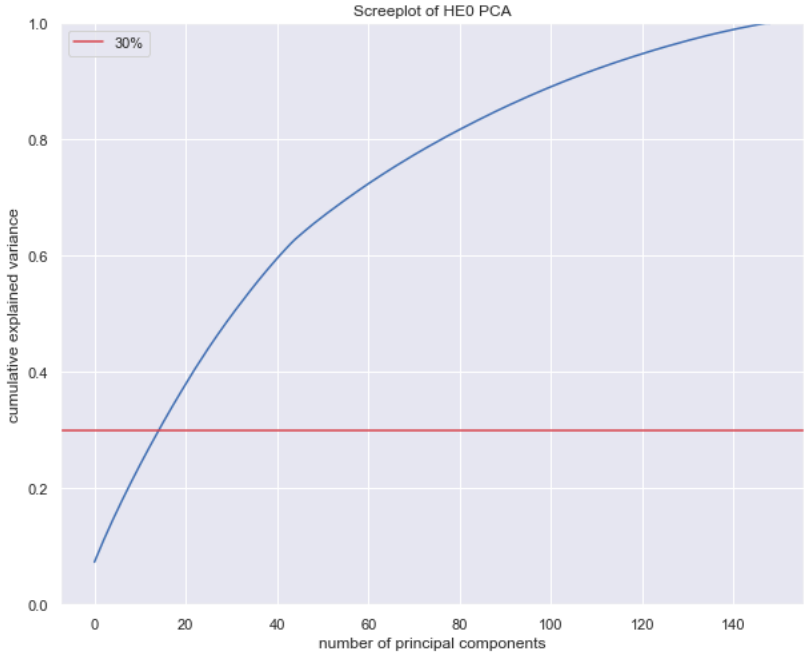
Function -log10 blows up the p-values closer to 0. For example -log(0.001) = 3 and -log(0.01) = 2. Data above 0.1 is less emphasized. This helps us focus more on the lower numerical values of p_value which are critical when making decision on elimination of H0

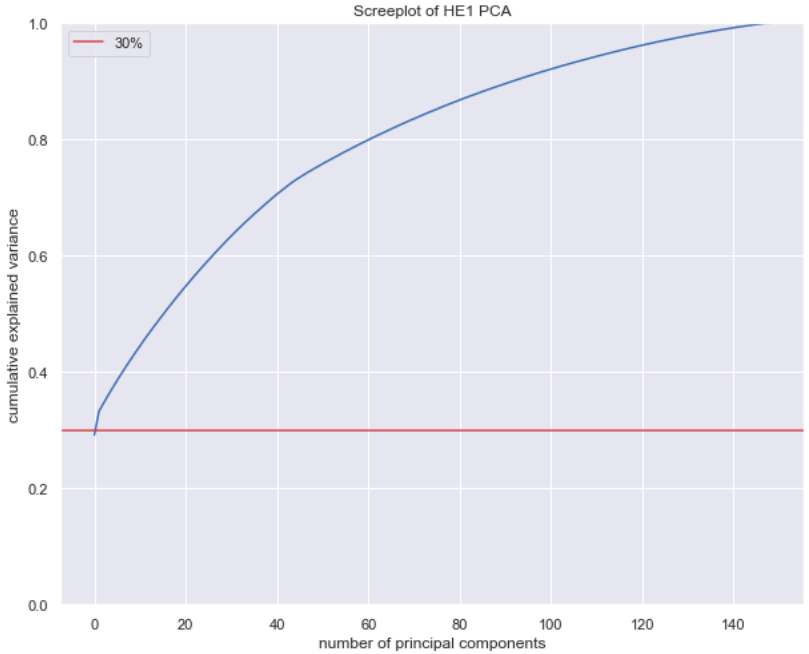
- e.ii. What can you conclude from the Q-Q plot?

Q-Q doesn't align with the x=y line hence the distributions are quite different, we can say expected and observed p-values follow different distributions. Assumption “H0 = True” is probably false. There is a difference between HE0 and HE1 samples and this difference is explained

Task 3 – Question 1

- b. Scree plots:

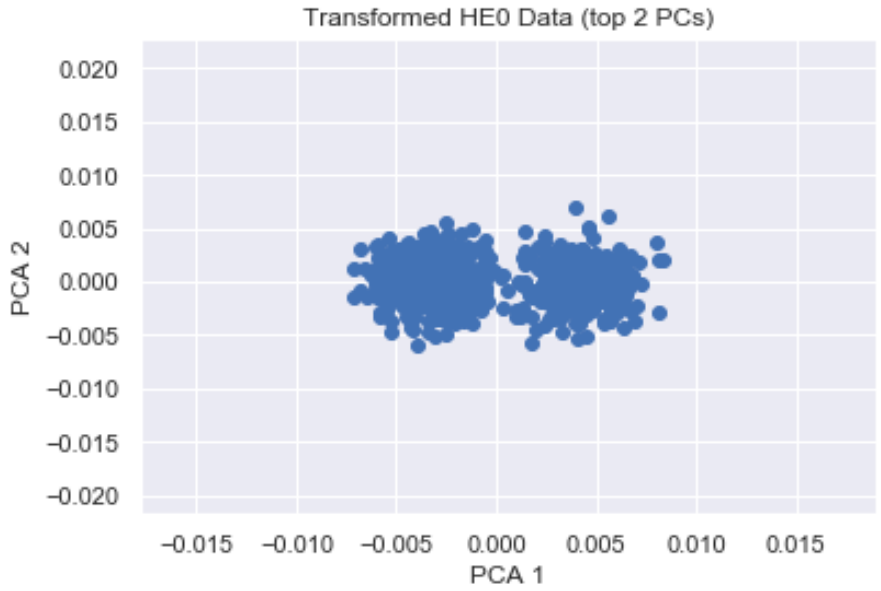


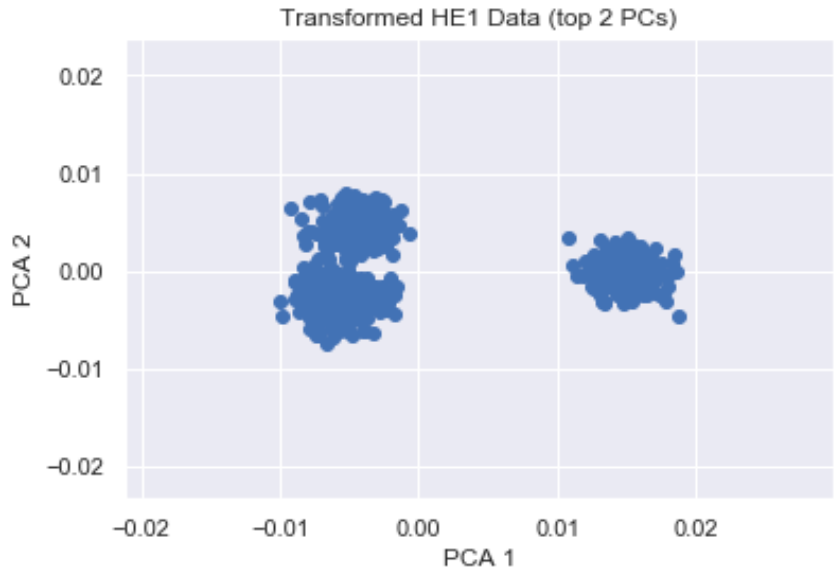


- Number of principal components needed to explain 30% of the total variance (HE0 and HE1):
  - HE0 =16
  - HE1 = 2

Task 3 – Question 1 (continued)

- c. Plots:

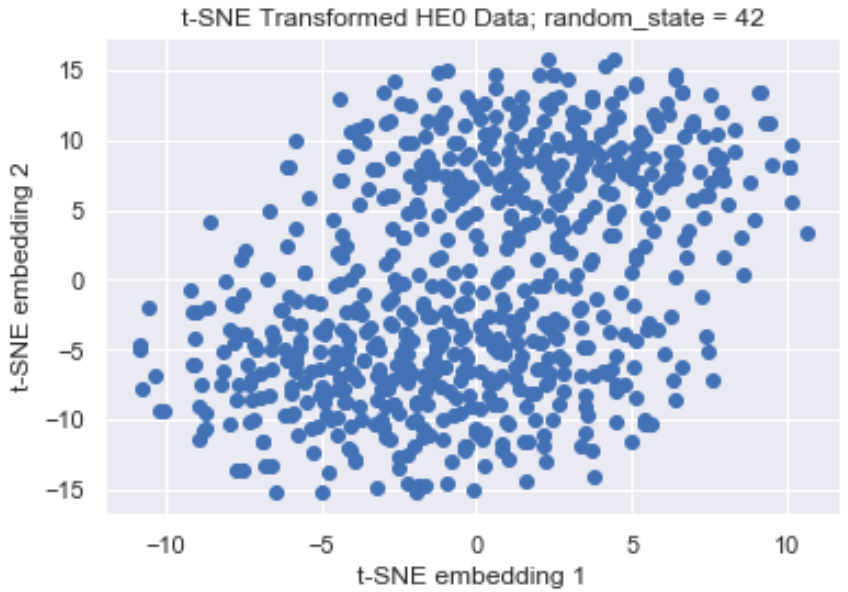




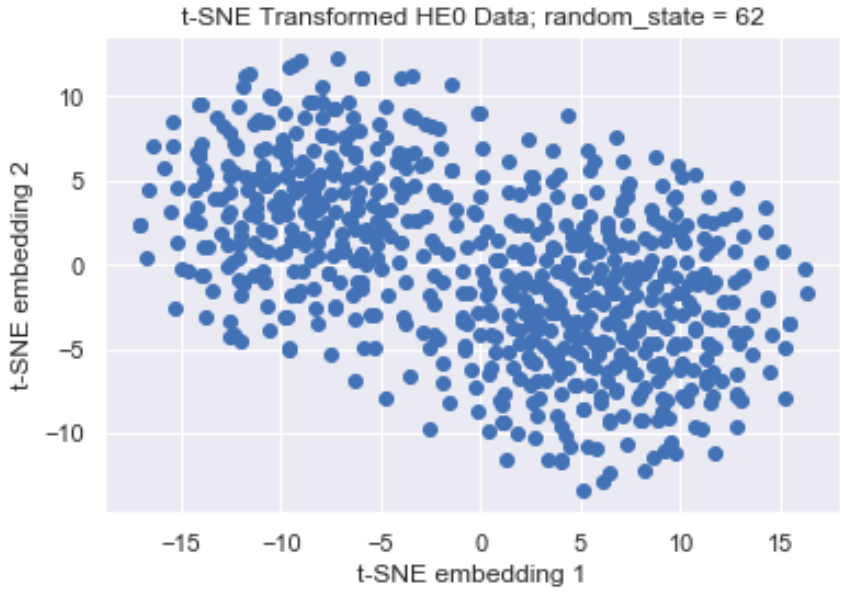
- Observations:
  - We see that tiny clustering is happening. It looks like there are 2 clusters in HE0 and 3 clusters in HE1.though it doesn't fully explain the variance in the data.
  - These sets of PCAs explain only about 30% of total variance

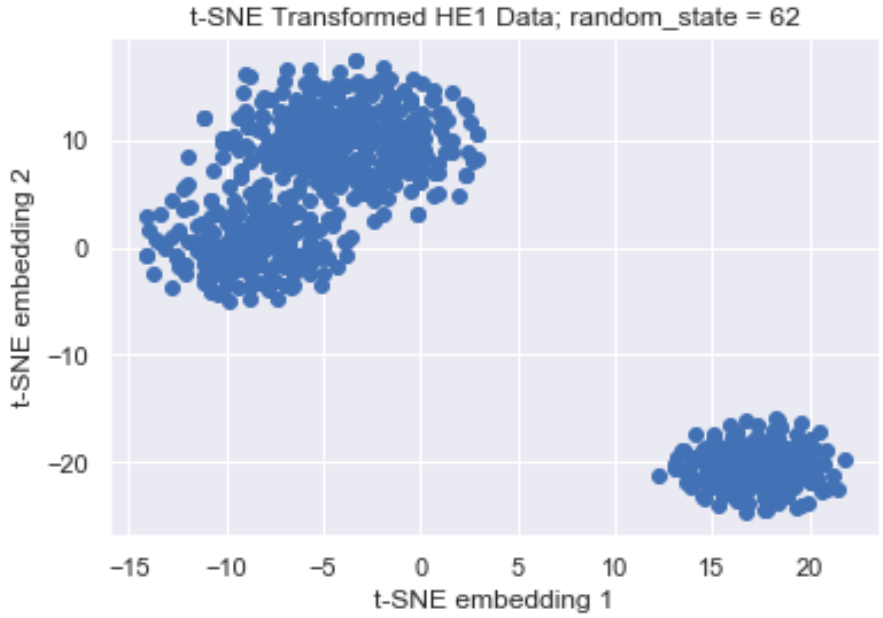
Task 3 – Question 2

- c. Plots (random_state=42):



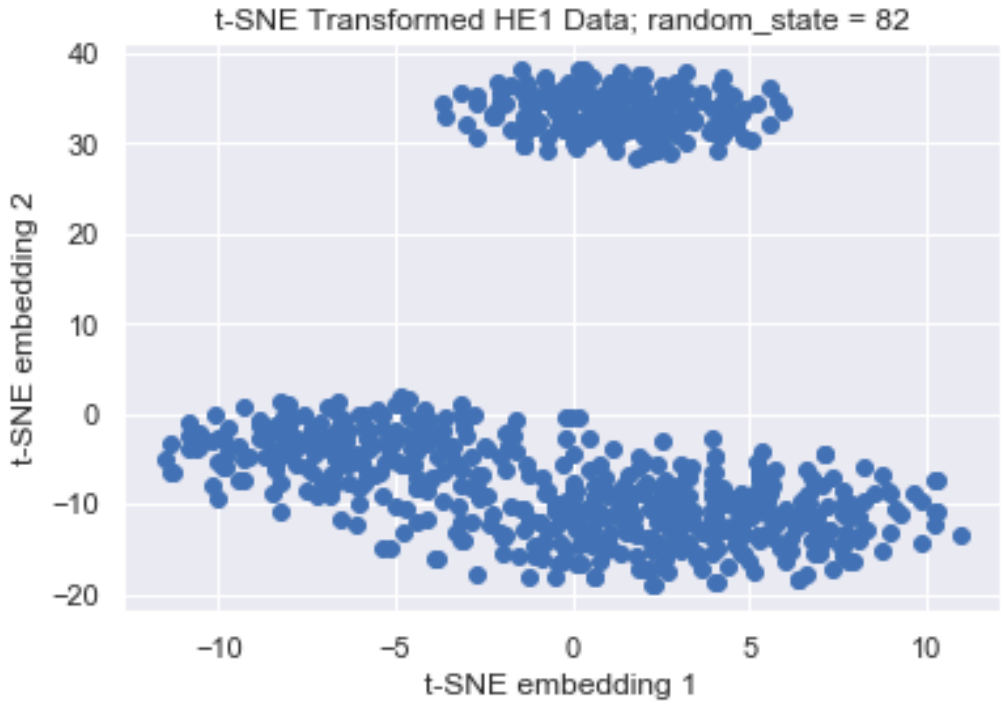
- Plots (random_state=62):





Task 3 – Question 2 (continued)

- c. Plots (random_state=82):



- Observations:

Different random states in the TSNE function generate seemingly different output graphs. But they are all somewhat similar in the sense, they are in transformed axes and number of clusters seems to be 2 and 3 respectively.

Task 3 – Question 2 (continued)

- d. Discussion of similarities and differences between PCA and t-SNE results:

The very first we observe is the speed of computation. PCA is way faster than TSNE.

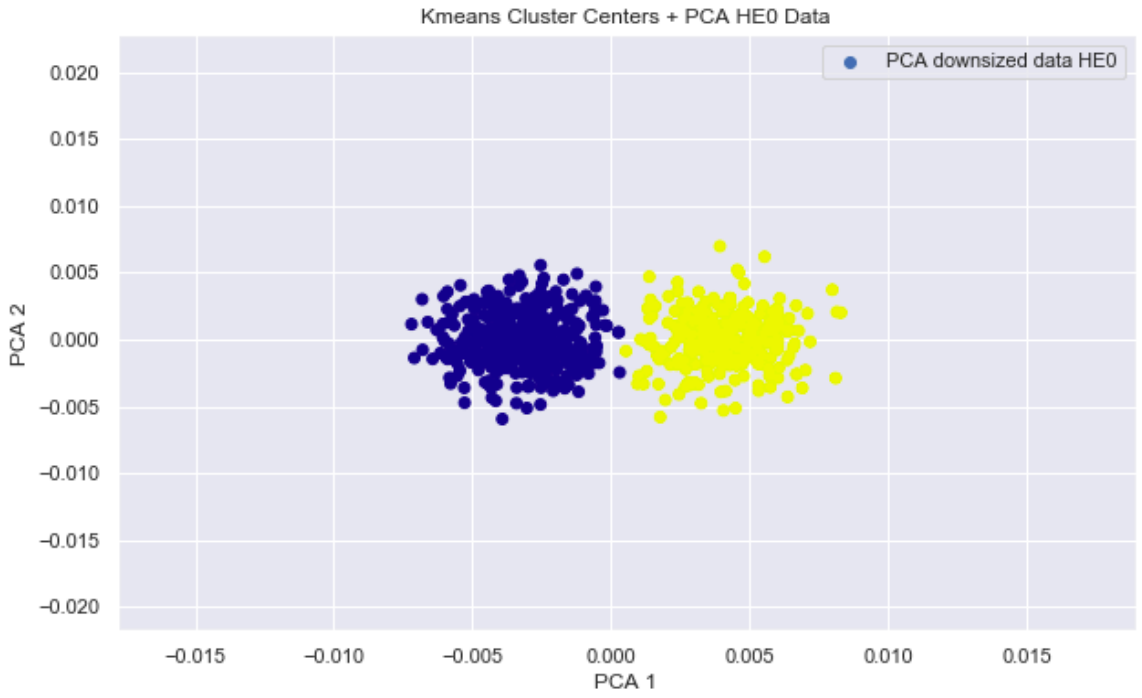
By the very nature of it's computation , results of TSNE may vary with different random states. But TSNE is a **good starting point** to see how many different clusters are there in the system. From this exercise we found that there are around 2 and 3 clusters respectively in HE0 and HE1 datasets.

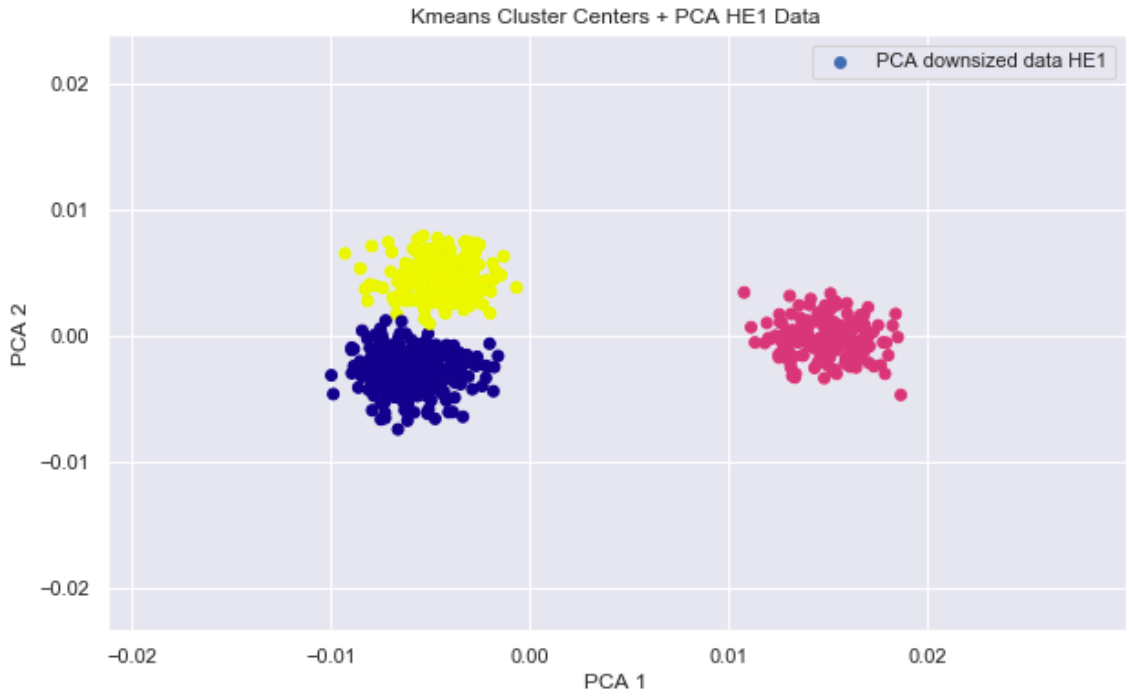
It is somewhat similar to PCA. But the distinction between clusters is more visible in the case of t-SNE. but it suffers from non-standard (random) initialization issues, slowness. The SciPy website suggests using PCA for features >50 which is true in this case.

Hence we are probably better off using PCA in this situation.

Task 3 – Question 3

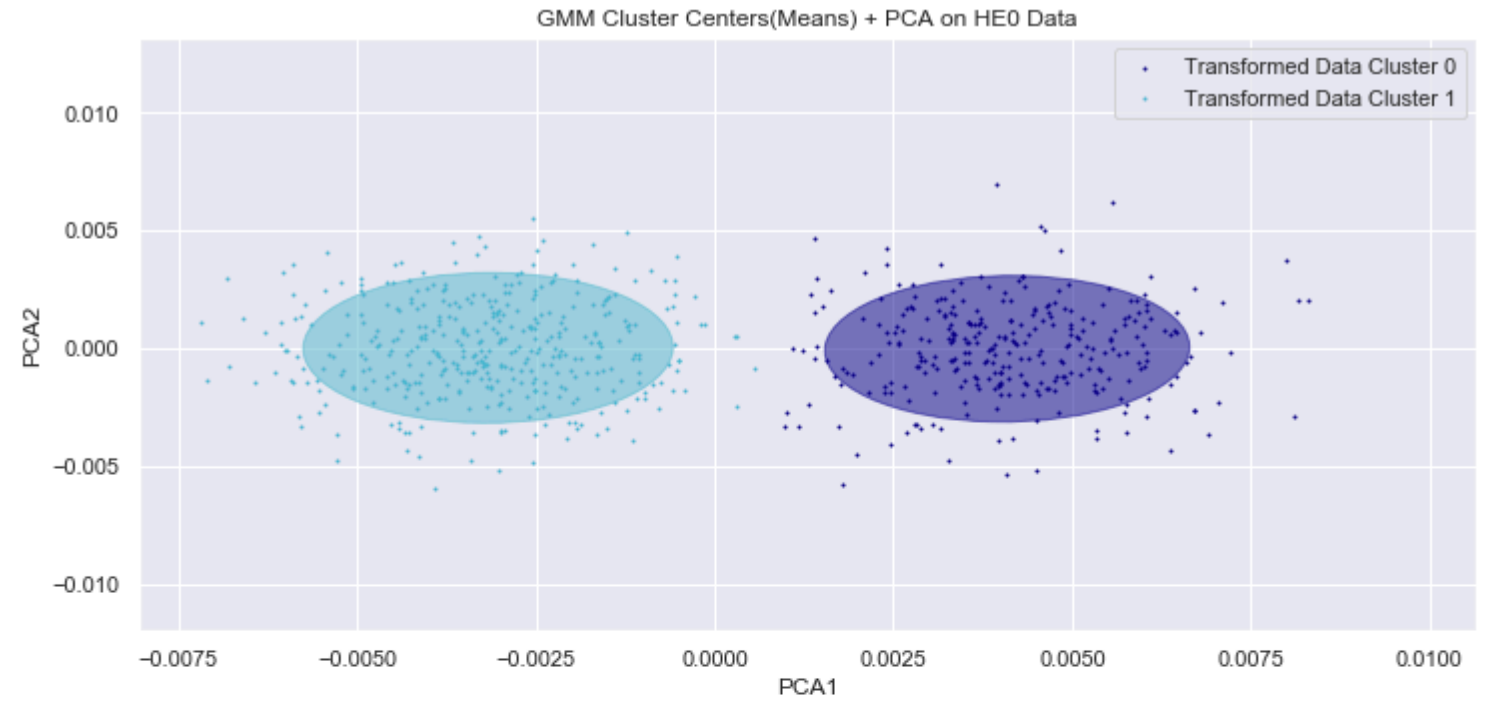
- a. K-means:

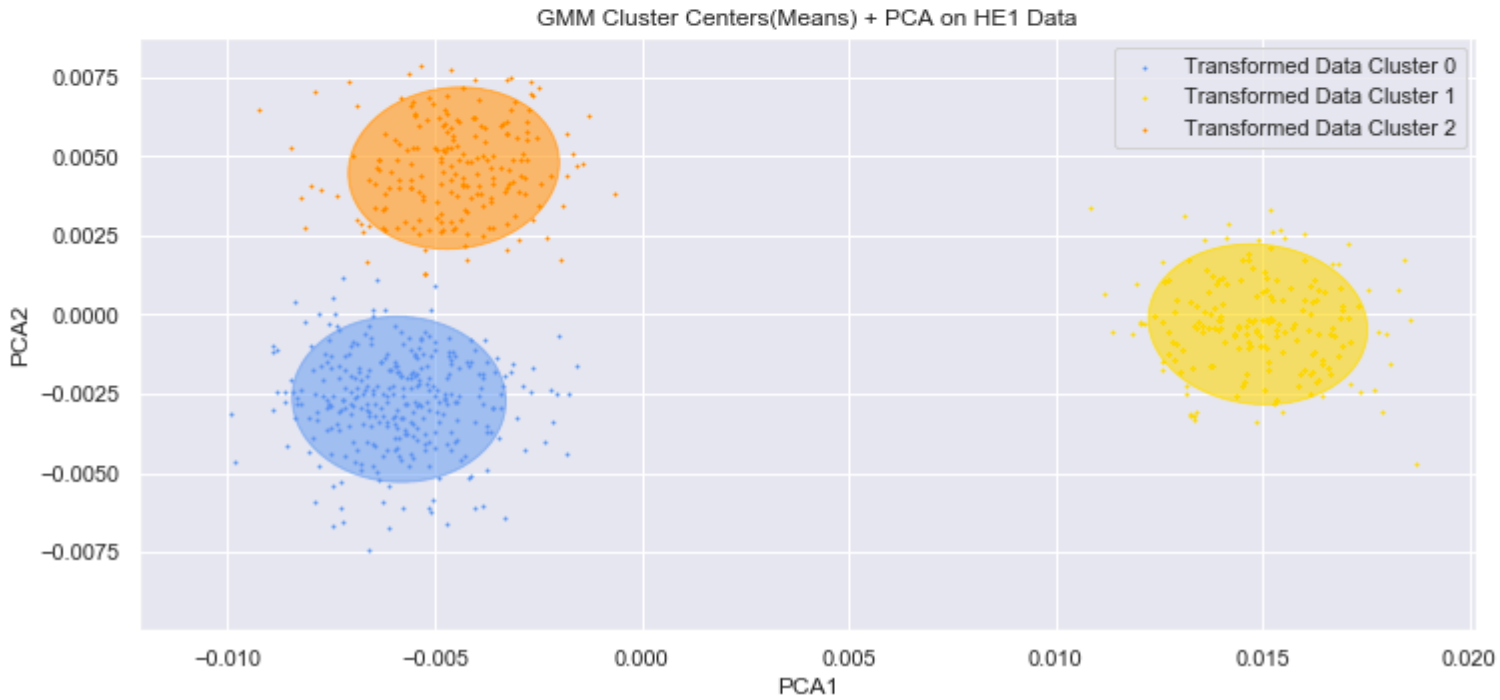




Task 3 – Question 3 (continued)

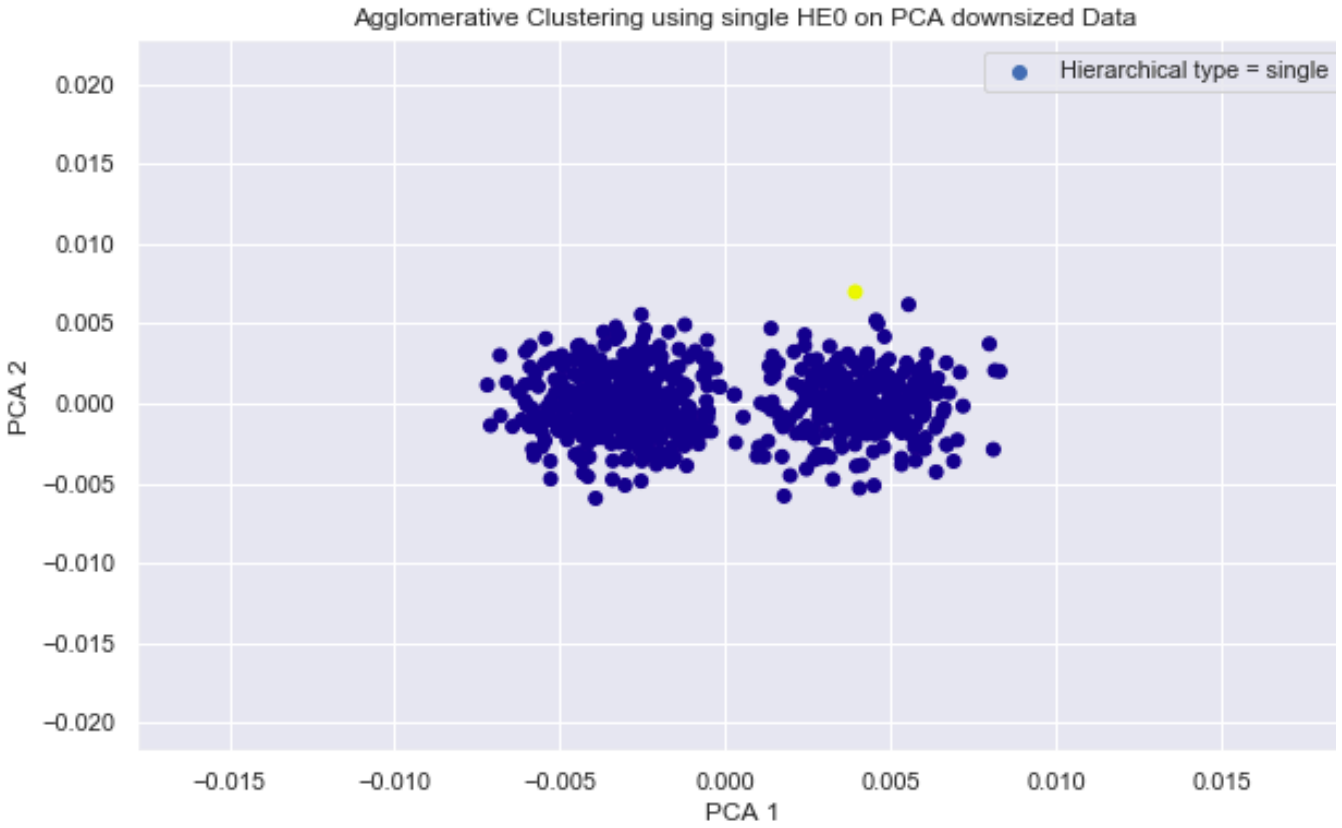
- b. Gaussian mixture model:

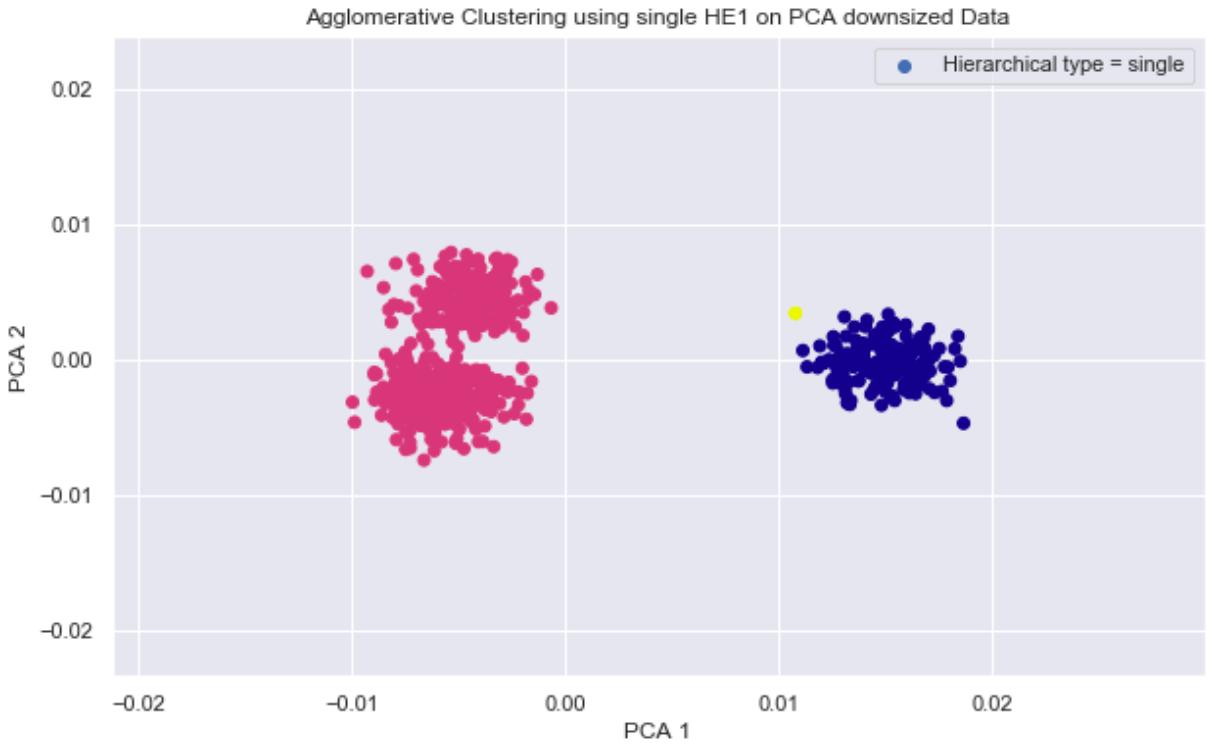




Task 3 – Question 3 (continued)

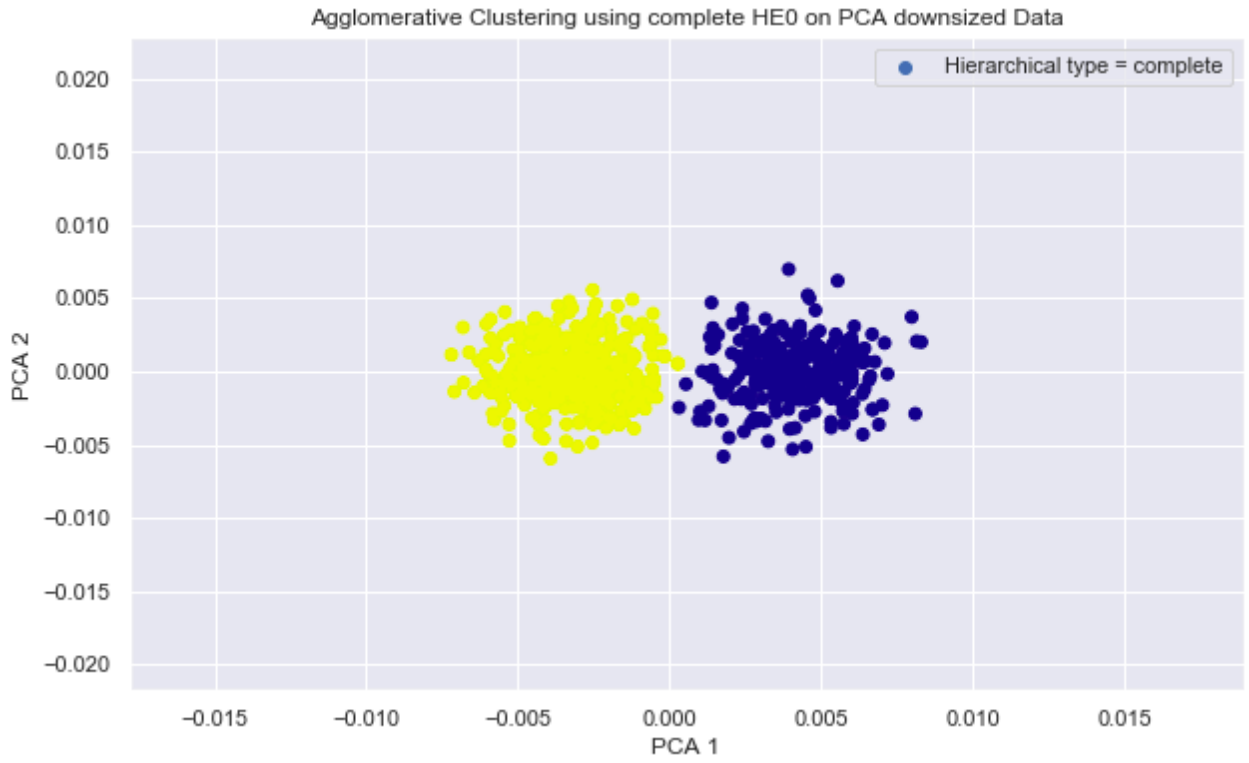
- c. single linkage hierarchical: **Note the point in Yellow**

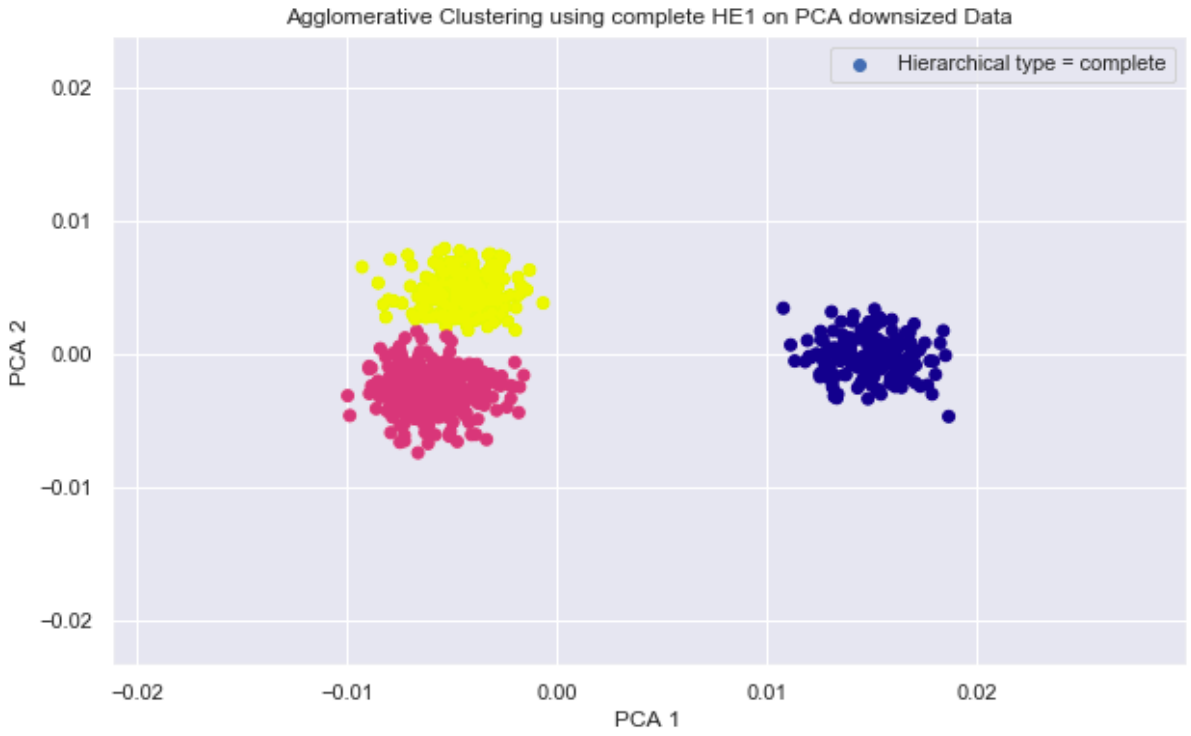




Task 3 – Question 3 (continued)

- c. complete linkage hierarchical:





Task 3 – Question 3 (continued)

- d. Discussion on single vs. complete linkage hierarchical methods:

    1. only one point in in bright color cluster in singly linked HE0 data

    2. We see that complete linkage works best for HE0 data but single linkage failed utterly due to it's implementation. It is well known that these algorithms are greedy and are not guaranteed to converge to the global optima.

    3. Compared to single link, complete link and average link performed similarly in HE1 data.

- e. Interpretation and comparison of the different methods:

1. Overall K-means and GMM performed better than Hierarchical clustering which is riddled with issues.
2. GMM is good but it is somewhat mathematically intensive. But it is on par with K-means
3. If we had to choose one, then K-means is much easier to compute.

Task 3 – Question 3 (continued)

- f. In context, what do the clusters you have found represent? What are some factors which could account for this type of clustering pattern?

  In the current context, the clusters represent closeness based on a linear combination of relative abundance of microbiome We can call these subpopulations of the respective HE dataset Different PCAs cover certain linear combination of microbiome along which the variance of the data is maximum.

  One possibility is the dependence on the way in which these samples are collected (we checked, it doesn't depend on that). Other possibility is some other condition (say condition X) that follows with liver cirrhosis in a particular number of samples (like HE).

- g. Based on your process for deciding the number of clusters to partition the data into, what situations or factors might result in your decision being inaccurate?

  Most of the current clustering happened to aid visual interpretation of the data. We only chose PCAs which cover 30% of the overall variance. In the case of HE1 data, most of the variance seems to be explained well. But in n-dimensions (n>2) we may potentially see different clustering action.

  if the clusters are being formed due to some external factor after the samples are being taken (for instance change in some microbe abundance when the sample comes in contact with atmosphere), the clusters may tend to be inaccurate. 3D visualization may show more info about data.

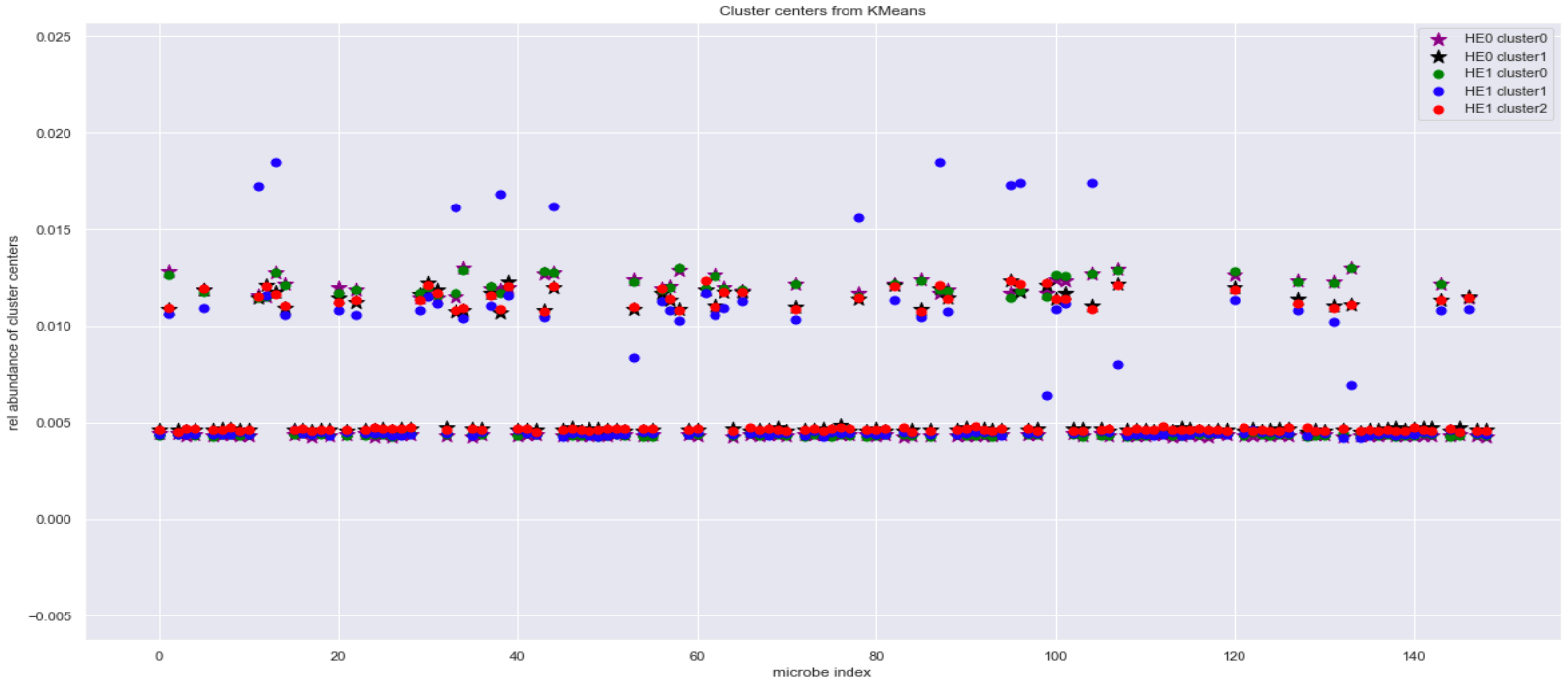
Task 4 – Question 1

Please view the submitted ipynb notebook for more details

- a. Determining which HE1 subpopulations had a significantly different microbiome than the HE0 samples. Explain your decision process and provide evidence supporting your conclusions.

Analysis procedure:

1. Initially we would like to see the various clusters that we plotted before and identify the PCAs which helped plot the figures.
2. Then we will compare the makeup of clusters centers (in PCA vectors coordinates).
3. To compare the difference we need to convert them back to same coordinates (dim = 149x1) This transformation occurs via addition of mean of original data (dim = 149x1)
4. Plot all the cluster centers on the same plot and observe the relations between different line scatter plots(colored)



Task 4 – Question 1 (continued)

Please view the submitted ipynb notebook for more details

- b. Determining the HE0subpopulation most similar to each HE1 subpopulation with a significantly different microbiome. Explain the decision process and provide evidence to support your conclusions.

Data is first transformed into the original coordinates. In this section, we extend our analysis from previous part. We calculate euclidean distances b/w different cluster centers and then make a decision based on the numerical results. Closer clusters are assumed to be related better. Detailed Results are shown in the iPYNB notebook.

1. cluster number in HE1 0

dist with HE0 cluster center 0 = 0.00107112767156627 49

dist with HE0 cluster center 1 = 0.0072761619797684 43

2. cluster number in HE1 1 (blue)

dist with HE0 cluster center 0 = 0.020916719735153 07

dist with HE0 cluster center 1 = 0.020243594797982 44

Cluster Relations: for colors refer to picture(lastpage)

black(HE0-1) -> red(HE1-2) - close match

Purple(HE0-0) -> green(HE1-0) - close match

blue (HE1)- significantly different

3. cluster number in HE1 2

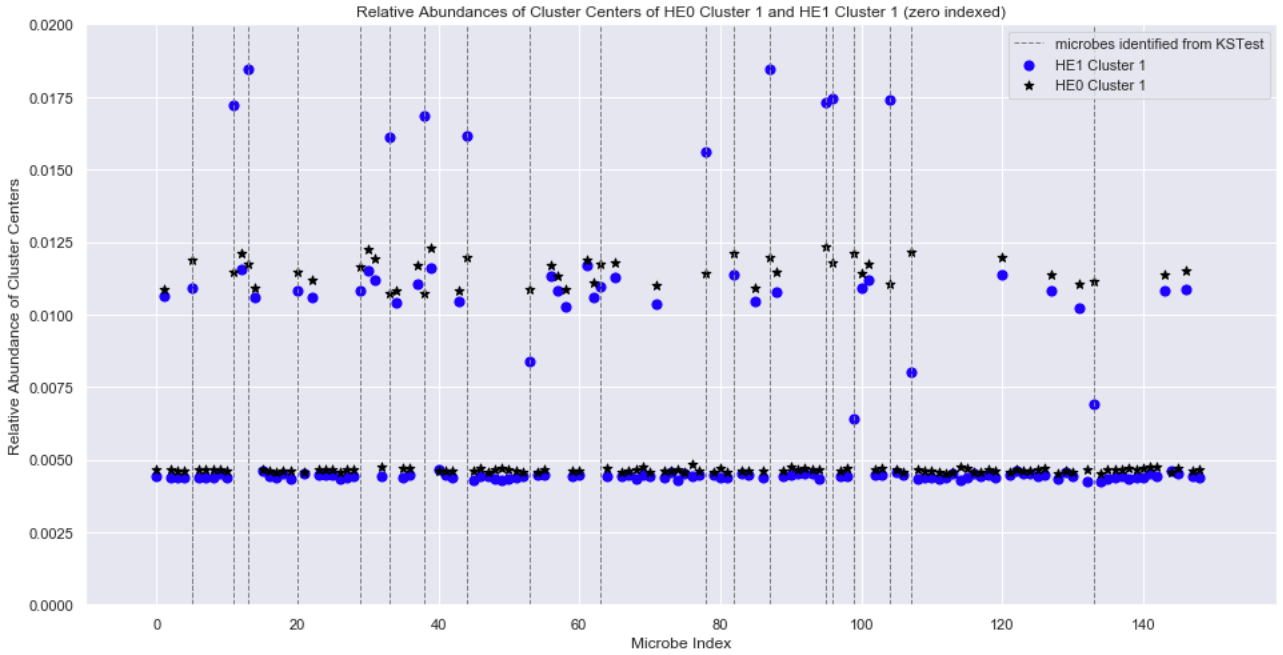
dist with HE0 cluster center 0 = 0.007446862643627164

dist with HE0 cluster center 1 = 0.001394484868496 5705

Task 4 – Question 1 (continued)

- c. Microbes with significantly altered abundance based on KS test

- 'Actinobacteria_Actinobacteria_Actinomycetales_Corynebacteriaceae', 'Actinobacteria_Actinobacteria_Actinomycetales_Nakamurellaceae', 'Actinobacteria_Actinobacteria_Actinomycetales_Propionibacteriaceae', 'Bacteroidetes_Bacteroidia_Bacteroidales_Bacteroidales_incertae_sedis', 'Bacteroidetes_Flavobacteriia_Flavobacteriales_Cryomorphaceae', 'Bacteroidetes_Sphingobacteriia_Sphingobacteriales_Sphingobacteriaceae', 'Chrysiogenetes_Chrysiogenetes_Chrysiogenales_Chrysiogenaceae', 'Firmicutes_Bacilli_Bacillales_Bacillales_Incertae Sedis XI', 'Firmicutes_Bacilli_Lactobacillales_Lactobacillaceae', 'Firmicutes_Clostridia_Clostridiales_Clostridiales_Incertae Sedis XIII', 'Firmicutes_Clostridia_Halanaerobiales_Halanaerobiaceae', 'Firmicutes_Negativicutes_Selenomonadales_Veillonellaceae', 'Parvarchaeota_Candidatus Parvarchaeum_Candidatus Parvarchaeum_Candidatus Parvarchaeum', 'Proteobacteria_Alphaproteobacteria_Rhizobiales_Brucellaceae', 'Proteobacteria_Alphaproteobacteria_Rhizobiales_Hyphomicrobiaceae', 'Proteobacteria_Alphaproteobacteria_Rhizobiales_Rhizobiaceae', 'Proteobacteria_Alphaproteobacteria_SAR11_SAR11', 'Proteobacteria_Betaproteobacteria_Burkholderiales_Burkholderiaceae', 'Proteobacteria_Gammaproteobacteria_Orbales_Orbaceae']



Task 4 – Question 2

- a. Which of the microbes that you identified show an increase of relative abundance in the HE1 sample? Do any show a decrease?

  Mean Microbe Abundances:

  - Microbe 11: Actinobacteria_Actinobacteria_Actinomycetales_Nakamurellaceae has higher microbe concentration
  - Microbe 13: Actinobacteria_Actinobacteria_Actinomycetales_Propionibacteriaceae has higher microbe concentration
  - Microbe 33: Bacteroidetes_Sphingobacteriia_Sphingobacteriales_Sphingobacteriaceae has higher microbe concentration
  - Microbe 38: Chrysiogenetes_Chrysiogenetes_Chrysiogenales_Chrysiogenaceae has higher microbe concentration
  - Microbe 44: Firmicutes_Bacilli_Bacillales_Bacillales_Incertae Sedis XI has higher microbe concentration
  - Microbe 53: Firmicutes_Bacilli_Lactobacillales_Lactobacillaceae has lower microbe concentration
  - Microbe 78: Firmicutes_Clostridia_Halanaerobiales_Halanaerobiaceae has higher microbe concentration
  - Microbe 87: Parvarchaeota_Candidatus Parvarchaeum_Candidatus Parvarchaeum_Candidatus Parvarchaeum has higher microbe concentration Microbe 95: Proteobacteria_Alphaproteobacteria_Rhizobiales_Brucellaceae has higher microbe concentration
  - Microbe 96: Proteobacteria_Alphaproteobacteria_Rhizobiales_Hyphomicrobiaceae has higher microbe concentration
  - Microbe 99: Proteobacteria_Alphaproteobacteria_Rhizobiales_Rhizobiaceae has lower microbe concentration
  - Microbe 104: Proteobacteria_Alphaproteobacteria_SAR11_SAR11 has higher microbe concentration
  - Microbe 107: Proteobacteria_Betaproteobacteria_Burkholderiales_Burkholderiaceae has lower microbe concentration
  - Microbe 133: Proteobacteria_Gammaproteobacteria_Orbales_Orbaceae has lower microbe concentration:

- Taxonomical relationships and groups among microbes with altered abundance:

We observe that the microbe names are grouped under certain criterion. For example look at: Proteobacteria family of bacteria, their name starts with that term. Taxonomical defn: Left to Right approach is followed. Where left part of the name is the parent or family type. Actual delineation of the microbe name is at the right end.