Mini-Project 2

ECE/CS 498DS

Spring 2020

Chuhao Feng (chuhaof2), Boyang Zhou (boyangz3), Mengxuan Yu (my13)

All Registered

Task 1 - Question 0

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

According to law of large numbers, the average of the results obtained from a large number of trials should be close to the expected value and will tend to become closer to the expected value as more trials are performed. Therefore, a large number of samples are more likely to represent the real situation accurately than less samples.

2. Number of samples analyzed: 764
3. Number of microbes identified: 149

Task 1 – Question 1

- a. Factorization of joint probability distribution:

$$P(S, CM, CO, L, Q) = P(Q|CO, L, S, CM)P(L|CO, S, CM)P(CO|S, CM)P(S|CM)P(CM)$$
$$= P(Q|CO, L)P(L)P(CO|S, CM)P(CM)P(S)$$

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

$$\#\text{ of parameters needed} = \underbrace{4}_{P(Q|CO,L)} + \underbrace{1}_{P(L)} + \underbrace{4}_{P(CO|S,CM)} + \underbrace{1}_{P(CM)} + \underbrace{1}_{P(S)} = 11$$

- c. Conditional probability tables:

Table of P(Quality | Contamination, Lab Time)

Lab Time	Contamination	Good	Bad
long	high	0.033898	0.966102
long	low	0.919003	0.080997
short	high	0.935743	0.064257
short	low	0.957093	0.042907

Table of P(Contamination | Store Temperature, Collection Method)

Storage Temp	Collection Method	High	Low
cold	nurse	0.043983	0.956017
cold	patient	0.076577	0.923423
cool	nurse	0.088435	0.911565
cool	patient	0.838235	0.161765

Table of P(Store Temperature)

Category	cold	cool
Store Temperature	0.8982	0.1018

Table of P(Collection Method)

Category	nurse	patient
Collection Method	0.8976	0.1024

Table of P(Lab Time)

Category	long	short
Lab Time	0.2044	0.7956

Task 1 – Question 1 (continued)

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

Storage Temp	Collection Method	Lab Time	Good	Bad
cold	nurse	long	0.887962	0.112038
cold	nurse	short	0.955112	0.044888
cold	patient	long	0.862069	0.137931
cold	patient	short	0.943978	0.056022
cool	nurse	long	0.822785	0.177215
cool	nurse	short	0.972376	0.027624
cool	patient	long	0.117647	0.882353
cool	patient	short	0.960784	0.039216

- e. Total number of samples dropped: HE0: 65, HE1: 65, In total: 130

Task 1 – Question 2

- 1. Number of samples removed: HE0: 101, HE1: 103, In total: 204
- 2. What are the benefits and drawbacks to using relative abundance data? Is there information that we lose when the normalization is performed?

Benefits:

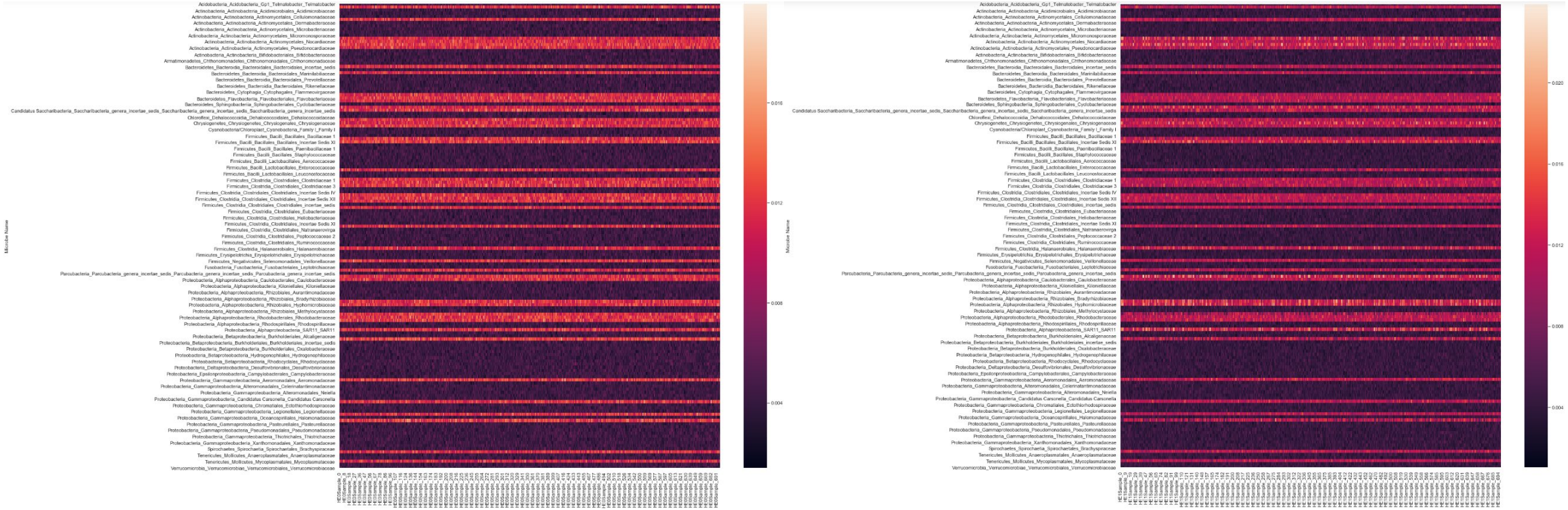
relative abundance data can show the proportion that each microbe takes up clearly, which is easy for researchers to identify the relationship between the composition of microbes and Hepatic Encephalopathy. Besides, relative abundance data are real numbers in [0, 1], which are easy to compute and free of units, so the computation is fast, and researchers do not have to deal with unit issues.

Drawback:

Using relative abundance data, the information of the number of each microbe as well as the total number of microbes in a sample is lost. As a result, researchers cannot analyze the relationship between the numbers of microbes and Hepatic Encephalopathy. Besides, once the number of a kind of microbe is adjusted in a sample, the entire relative abundance data needs to be recomputed, which requires extra computation compared with absolute abundance data.

Task 1 – Question 3

- Heatmaps (HE0 on left HE1 on right):



- Summarize your observations

To be honest, these two heatmaps seem similar to each other for human eyes, except that the heatmap for HE0 samples seems brighter than that for HE1 samples, or, in other words, that color comparison is stronger in the heatmap for HE0 samples than in that for HE1 samples. That's to say, the relatively abundant microbes in HE0 samples take up larger proportions respectively than their counterparts in HE1 samples do, and the relatively less microbes in HE0 samples take up smaller proportions respectively than their counterparts in HE1 samples do.

Task 1 – Question 3 (continued)

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

  The heatmaps are good at highlighting the data that have relatively large values and the difference between the data that have different values. In our context, the heatmaps are good at highlighting the relatively abundant microbes in HE0 or HE1 samples.

  The heatmaps are less suitable for analyzing networks or hierarchical relationships, because these types of things are irrelevant to the values of data.

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.
  Null hypothesis:
  The distribution of the relative abundance of a certain microbe in HE0 samples is the same as the distribution of the relative abundance of that kind of microbe in HE1 samples.

  Example: Acidobacteria_Acidobacteria_Gp1_Telmatobacter_Telmatobacter
  The null hypothesis is that the distribution of the relative abundance of Acidobacteria_Acidobacteria_Gp1_Telmatobacter_Telmatobacter in HE0 samples is the same as the distribution of the relative abundance of Acidobacteria_Acidobacteria_Gp1_Telmatobacter_Telmatobacter in HE1 samples.

- 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:

α	# of microbes with altered expression
0.100	42
0.050	36
0.010	27
0.005	26
0.001	21

Task 2 – Question 2

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

  A p-value of 0.05 in our context represents that the probability of the observed or less similarity between the distribution of the relative abundance of a certain microbe in HE0 samples and the distribution of the relative abundance of that kind of microbe in HE1 samples is 0.05, given that the null hypothesis is true.

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

  The p-values follow uniform distribution in [0,1].

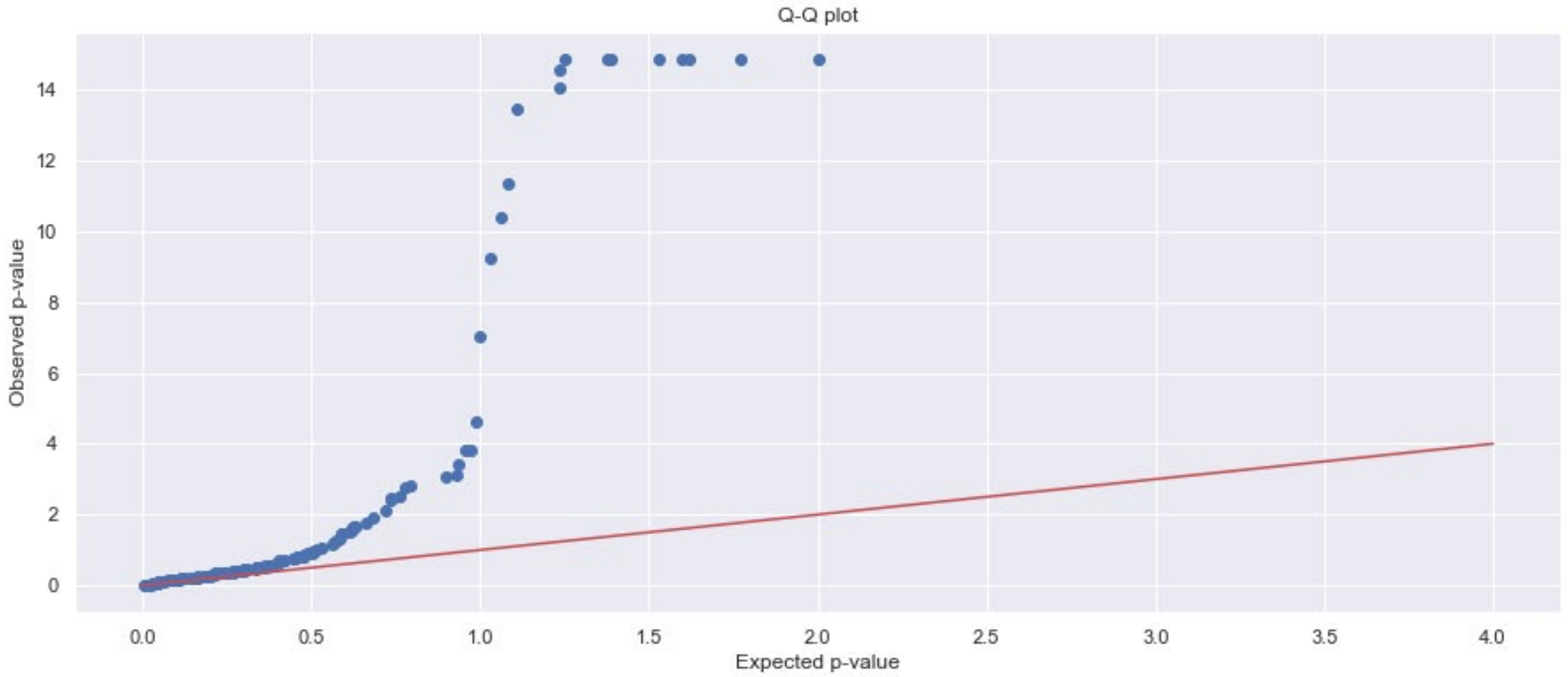
- c. If no microbe's abundance was altered, how many significant p-values does one expect to see at alpha=0.1, 0.05, 0.01, 0.005 and 0.001 level? Compare your answers with your results in Task 2.1.c. Show the comparison in a table below:

α	Observed # of microbes with altered expression	Expected # of microbes with altered expression
0.100	42	15
0.050	36	7
0.010	27	1
0.005	26	1
0.001	21	0

From the comparision table, we find that as α goes larger the ratio between the observed number of microbes with altered expression and the expected number of microbes with altered expression becomes closer to 1 and that both the observed and expected numbers increase as α increases.

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?

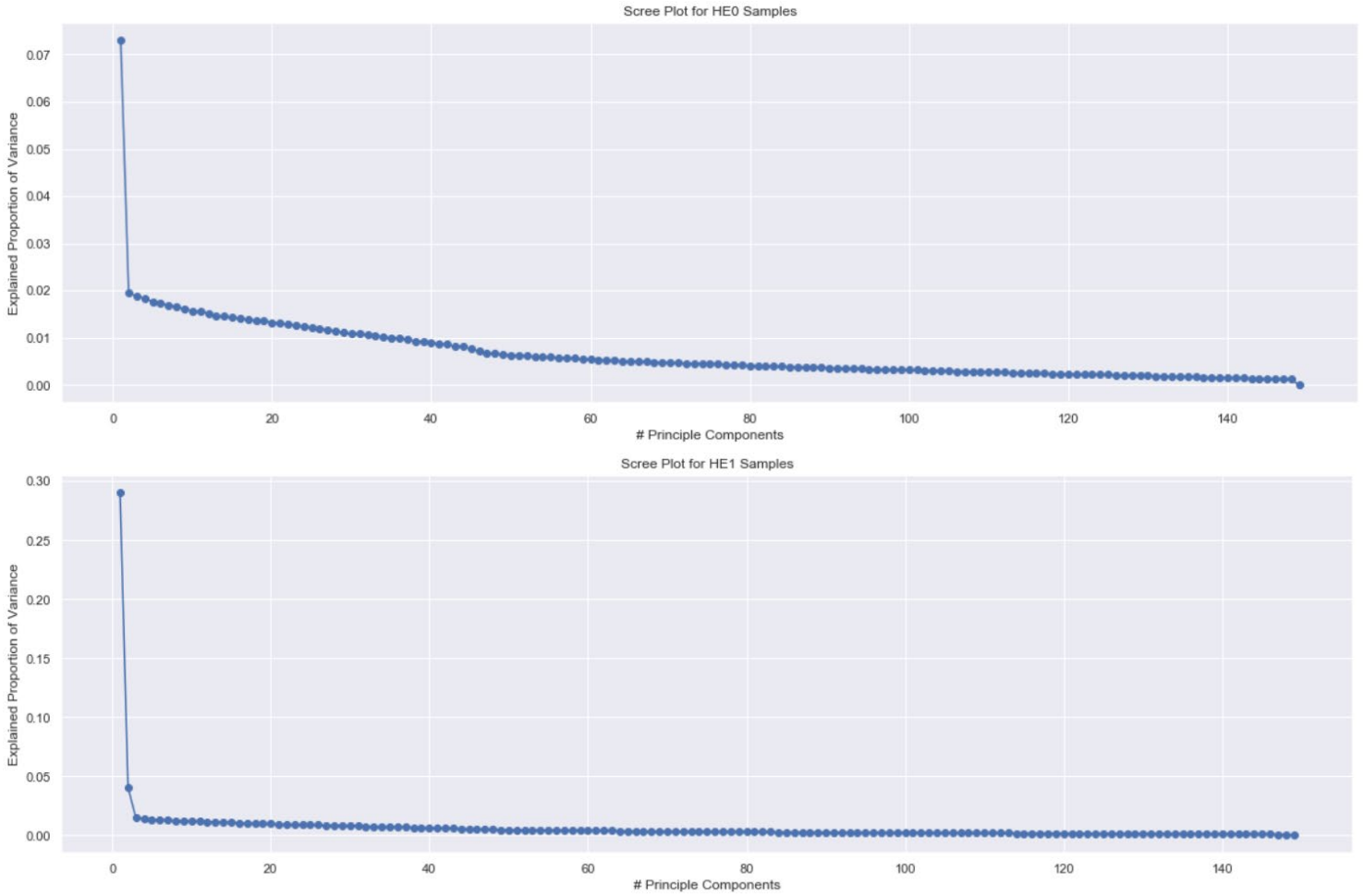
  The –log10() maps the original p-values from a smaller range [0,1] to a much larger range [0, +∞], which enlarges the difference between p-values, helping us visualize the distribution of dots in the graph and see the deviation between the dots and the y=x line.

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

  From the Q-Q plot, it is easy to observe that expected p-values are really close to observed p-values when p-values are larger than approximately $10^{-0.4}$, which is 0.3981071705534972. Besides, observed p-values are smaller than expected p-values when p-values are smaller than approximately $10^{-0.4}$, which is 0.3981071705534972, and the gap between observed p-values and expected p-values increases as p-values become smaller. That's to say, observed p-values have similar tendency as expected p-values have to fail to reject the null hypothesis when they are large, since they are close to each other. However, observed p-values have larger tendency than expected p-values have to reject the null hypothesis when they are small, because the observed p-values are smaller than expected p-values and the gap between them increases as they become smaller. To conclude, the observed p-value distribution has larger overall tendency to reject the null hypothesis than the expected p-value distribution has.

Task 3 – Question 1

- b. Scree plots:

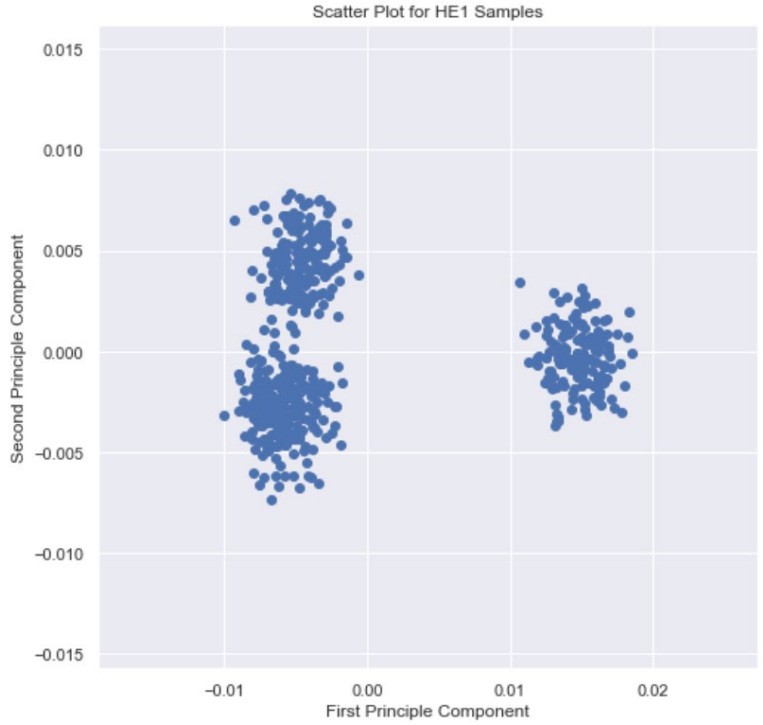


- Number of principal components needed to explain 30% of the total variance (HE0 and HE1):

  15 for HE0, 2 for HE1

Task 3 – Question 1 (continued)

- c. Plots:

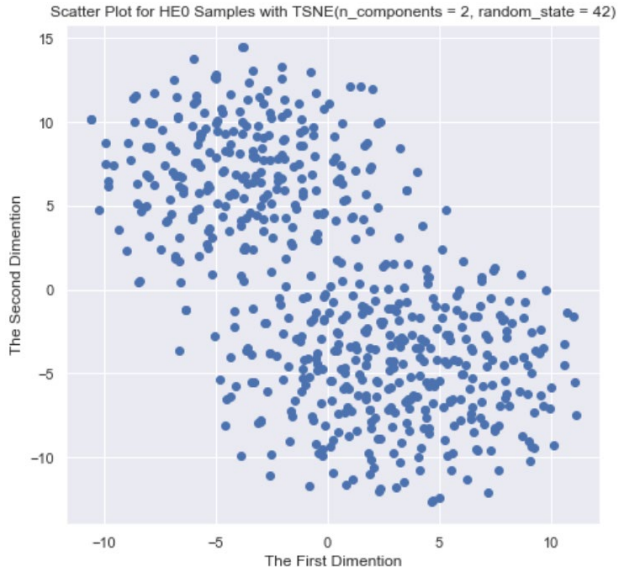


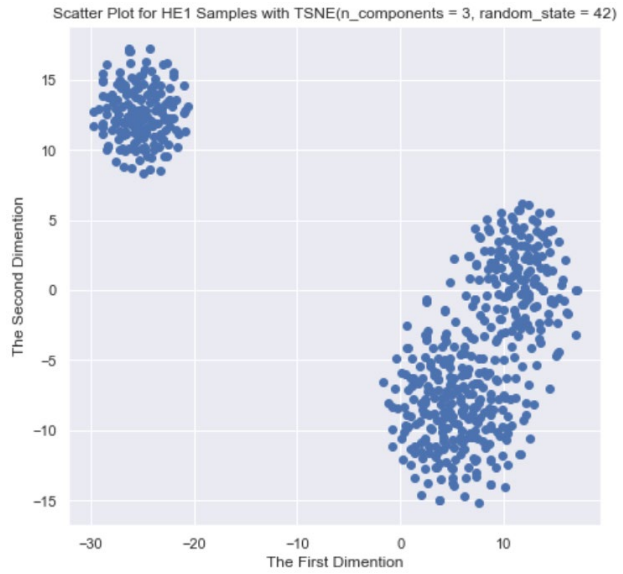
- Observations:

  It seems that there are two clusters in the HE0 samples while three clusters in the HE1 samples. Besides, the distance between the right cluster and the left cluster/clusters along the first principle component in HE1 samples is larger than that in HE0 samples.

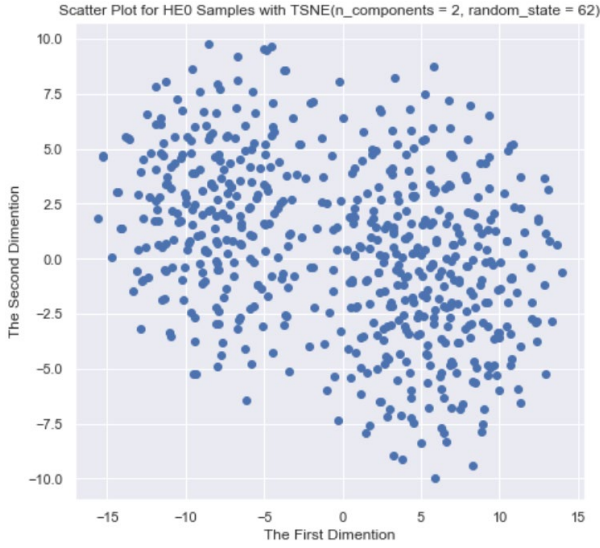
Task 3 – Question 2

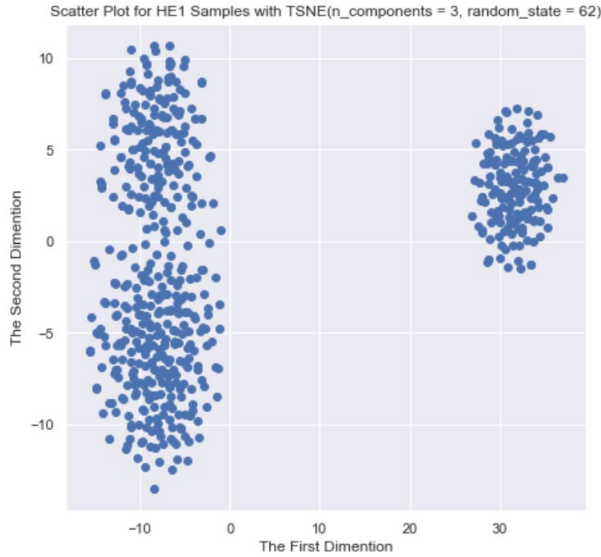
- c. Plots (random_state=42):





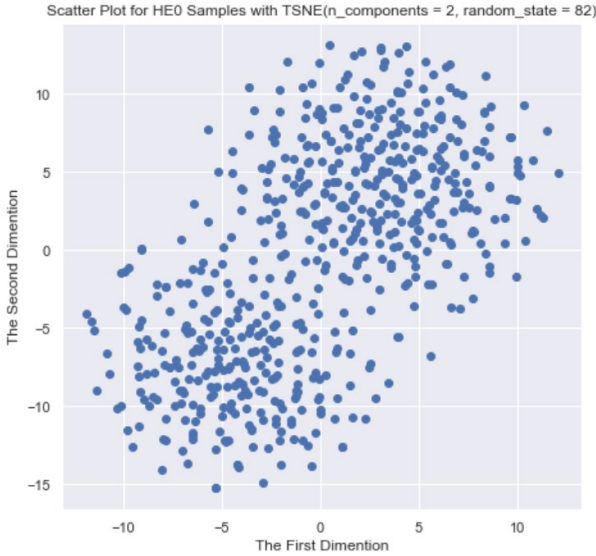
- Plots (random_state=62):

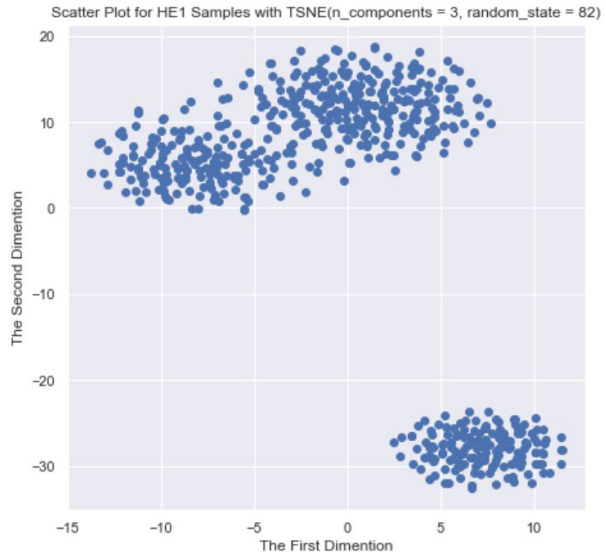




Task 3 – Question 2 (continued)

- c. Plots (random_state=82):





- Observations:

No mater what the value of the random_state parameter is, there tends to be two clusters in the result derived by the t-SNE method for the HE0 samples, which is consistent with the result derived by the PCA method for the HE0 samples. By contrast, no mater what the value of the random_state parameter is, there tends to be three clusters in the result derived by the t-SNE method for the HE1 samples, which is also consistent with the result derived by the PCA method for the HE1 samples. In addition, the results derived by the PCA method are independent of the value of the random_state parameter, no matter for the HE0 or HE1 samples. However, the results derived by the t-SNE method vary with the value of the random_state parameter, no matter for the HE0 or HE1 samples.

Task 3 – Question 2 (continued)

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

Similarity:

Both the two methods can reduce the high dimensional data to 2D and keep the inner relationship of the data. Besides, both the two methods tend to divide the HE0 samples into two clusters while the HE1 samples into three clusters.

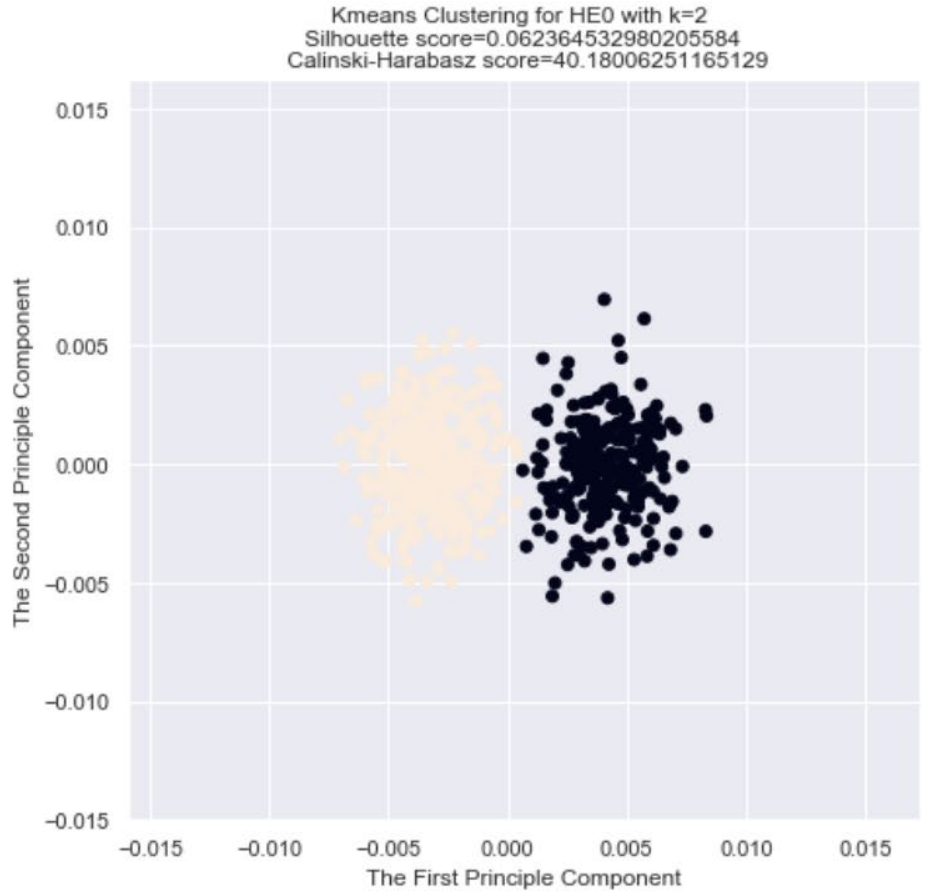
Difference:

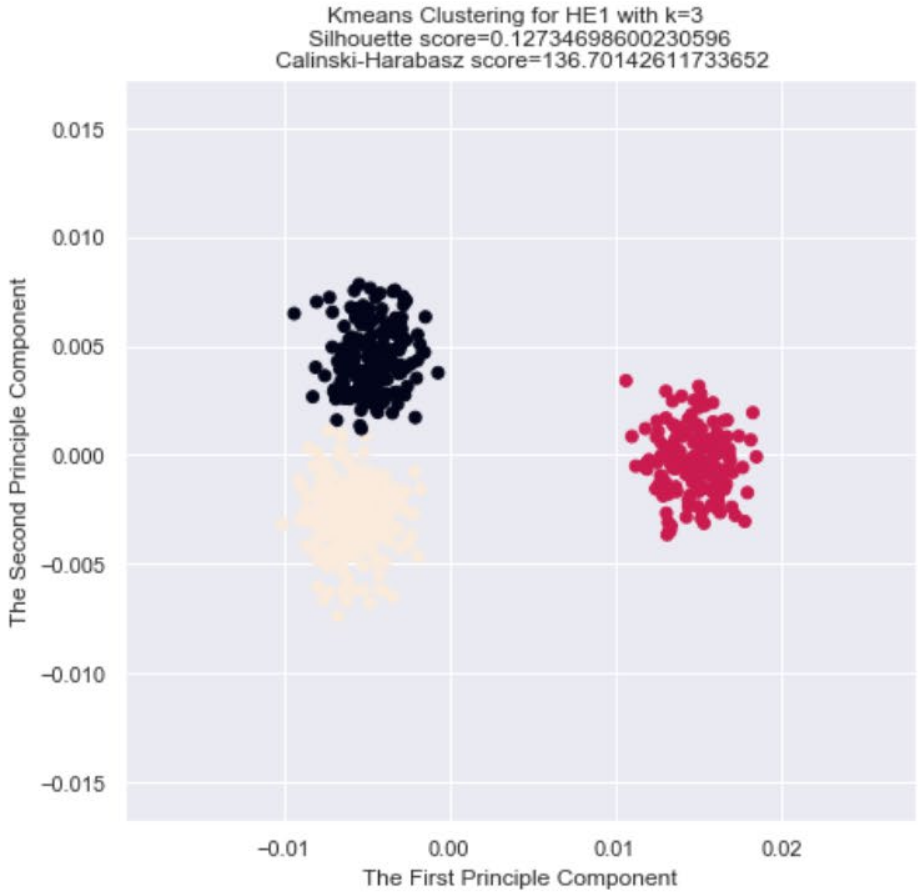
1) The results from the PCA method are not affected by the value of the random_state parameter while the results from the t-SNE method are affected by the value of the random_state parameter.

2) Points in each cluster in the results from the PCA method are much closer to each other than those in each cluster in the results from the t-SNE method, which may come from that the results from the PCA method still use the same magnitude as the original data do while the t-SNE method uses larger magnitude to present its results. Remaining the magnitude of the original data, the PCA may cause the Crowding Problem, because it maps so many high dimentional data points to a 2D space limited by small magnitude. By contrast, the t-SNE method somehow increases the magnitude, so it enlarges the distance between the data points projected onto the 2D space, which may avoid the Crowding Problem.

Task 3 – Question 3

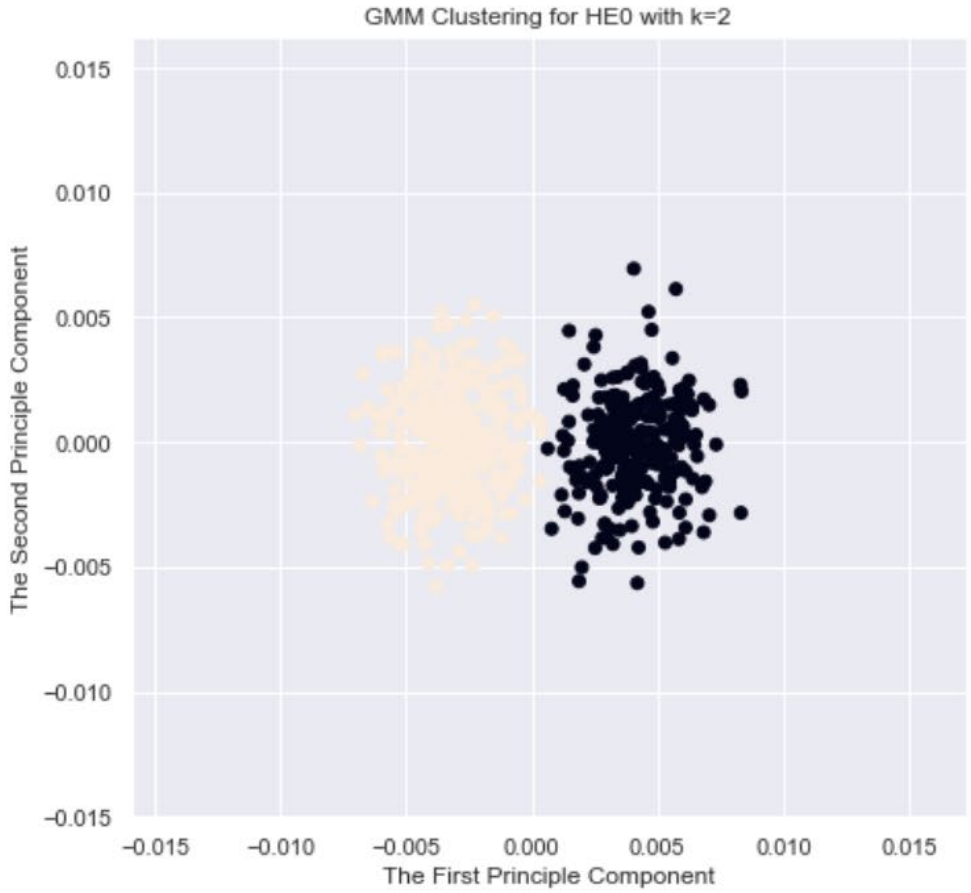
- a. K-means: Please refer to the notebook for how we decide the number of clusters.





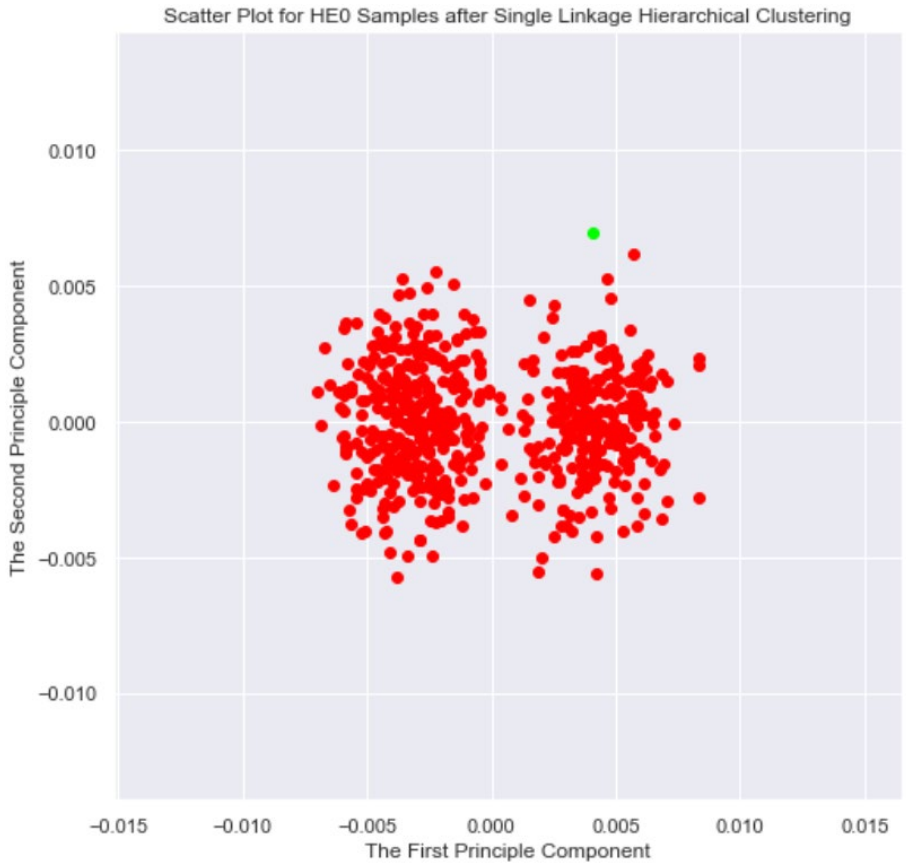
Task 3 – Question 3 (continued)

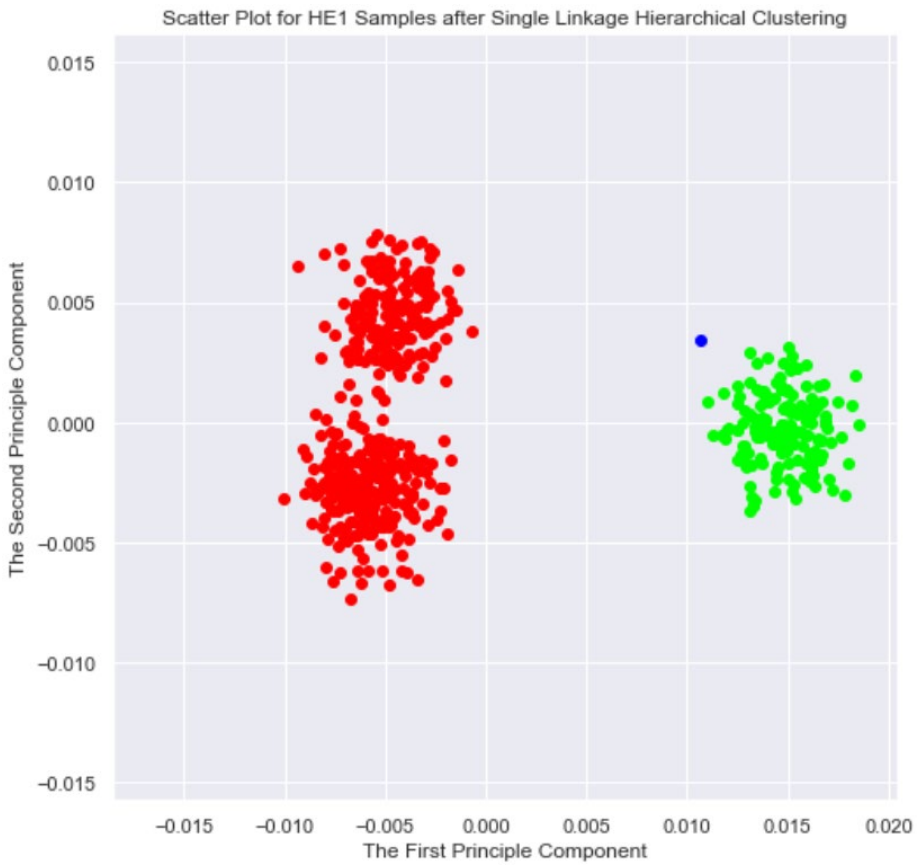
- b. Gaussian mixture model: Please refer to the notebook for how we decide the number of clusters.



Task 3 – Question 3 (continued)

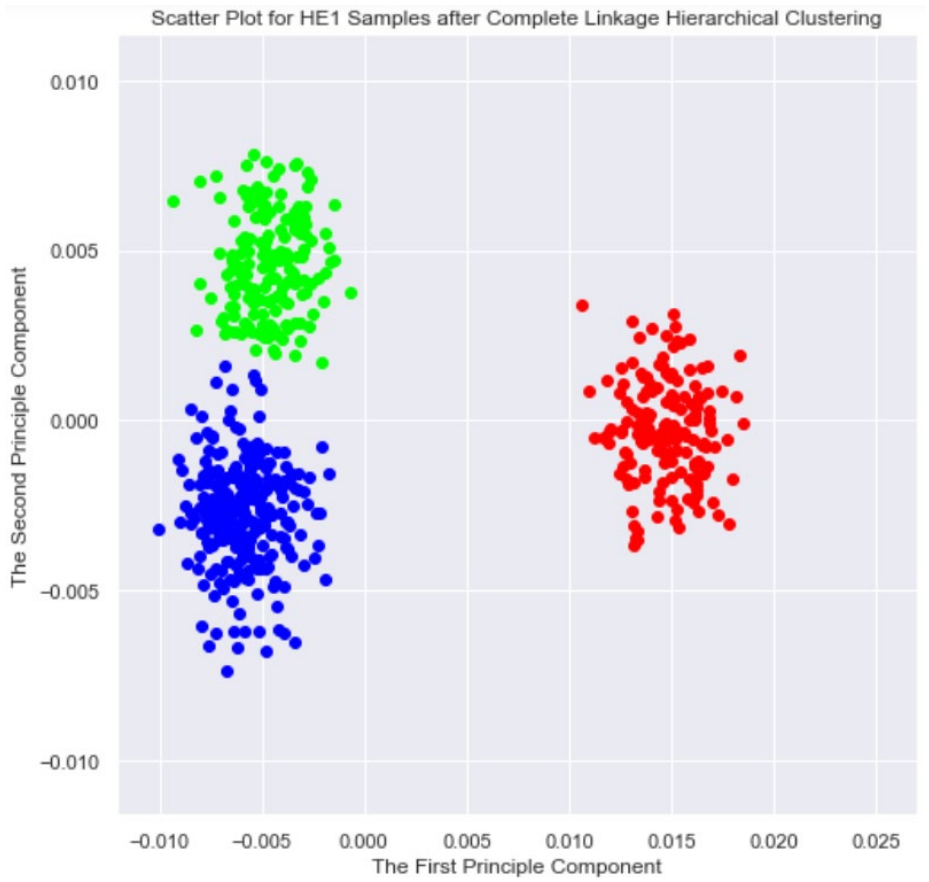
- c. single linkage hierarchical: Please refer to the notebook for how we decide the number of clusters.





Task 3 – Question 3 (continued)

- c. complete linkage hierarchical: Please refer to the notebook for how we decide the number of clusters.



Task 3 – Question 3 (continued)

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

  The difference between the single and complete linkage hierarchical clustering methods is that their ways to define the distance between two clusters are different. Even though these two methods both start from treating each data point as a cluster and merge two clusters which are most similar (closest) to each other, the single linkage method defines the distance between two clusters as the distance of the closest pair of data objects belonging to different clusters while the complete linkage method defines the distance between two clusters as the distance of the farthest pair of data objects belonging to different clusters.

  Yes, major differences can be seen from the generated clusters. Specifically, the result from the single linkage method is quite different from the result from the complete linkage method for HE0 data. Besides, the result from the single linkage method is also quite different from the result from the complete linkage method for HE1 data. We think that it is the outliers in the data that cause the differences. Since the single linkage method is sensitive to outliers and suffers the chaining effect, the results from the single linkage method will be heavily affected by the outliers in our data. Based on observation, there should be at least one outlier in both the HE0 and HE1 samples, so the performance of the single linkage method is really bad for either the HE0 data or the HE1 data. By contrast, the complete linkage method is robust to outliers and prefers spherical clusters, so it performs pretty well for both the HE0 data and the HE1 data.

- e. Interpretation and comparison of the different methods:

  Intuitively, there should be two clusters in the HE0 data while three clusters in the HE1 data. Except the single linkage hierarchical clustering method, all other methods give really reasonable results. Obtaining reasonable results is expected, because the clusters in the HE0 and the HE1 data are obvious. By contrast, it is the outliers in the HE0 and the HE1 data that cause the unreasonable result from the single linkage hierarchical clustering method. The single linkage hierarchical clustering method is sensitive to outliers, so the outliers in the HE0 and the HE1 data can heavily affect the performance of the single linkage hierarchical clustering method. We select the GMM results for the following analyses.

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?

  Generally speaking, the clusters I have found represent subpopulations in HE0 and HE1 samples, respectively. Specifically, the two clusters I have found in HE0 samples represent the two subpopulations that have significantly different microbial relative abundance in HE0 samples. Similarly, the three clusters I have found in HE1 samples represent the three subpopulations that have significantly different microbial relative abundance in HE1 samples.

  Factors:

  • Maybe whether having HE or not is mainly determined by microbial composition.

  In HE0 samples, there may be two different types of microbial composition that contribute to not having HE. Similarly, there may be three different types of microbial composition that contribute to having HE in HE1 samples.

  • Maybe whether having HE or not is mainly determined by certain types of microbes.

  Maybe whether having HE or not is mainly determined by several certain types of microbes that help divide HE0 samples into two subpopulations while HE1 samples into three subpopulations.

- 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?

  Factors:

  • There is oo much noise.

  If there is too much noise, the noise will affect the evaluation of the cluster algorithm, which may suggest an incorrect cluster number.

  • Cluster centroids are too close to each other.

  If cluster centroids are too close to each other, the evaluation of the cluster algorithm tends to suggest that some clusters should be merged together while the suggested merging is against our observations.

  • The data reveals weak clustering patterns.

  If the data reveals really weak clustering patterns by observation, we can only depend on the the evaluation of the cluster algorithm to decide the number of clusters, where it is really easy to use an incorrect cluster number, because we cannot verify the suggested cluster number by our observation.

Task 4 – Question 1

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

  Since the decision process is really long, we cannot put it here. Therefore, please refer to the notebook for our decision process.

  Based on our analysis and the criteria we define, we obtain the following information.

  - HE1_a subpopulation has no significantly different microbiome than the HE0_a or HE0_b subpopulations.
  - HE1_b subpopulation has a significantly different microbiome than both the HE0_a and HE0_b subpopulations.
  - HE1_c subpopulation has a significantly different microbiome only than the HE0_b subpopulation.

Task 4 – Question 1 (continued)

- 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.

  For this problem, we can directly make use of the results from Task 4.1.a, because the decision process is the same.

  - For HE1_b subpopulation, it is most similar to HE0_a subpopulation, because there are 14 microbes that have a significantly altered abundance between HE1_b and HE0_a subpopulations while 22 microbes that have a significantly altered abundance between HE1_b and HE0_b subpopulations.
  - For HE1_c subpopulation, it is most similar to HE0_a subpopulation, because there are no microbes that have a significantly altered abundance between HE1_c and HE0_a subpopulations while 8 microbes that have a significantly altered abundance between HE1_c and HE0_b subpopulations.

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
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_Betaproteobacteria_Rhodocyclales_Rhodocyclaceae
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?

Microbes that show an increase

Actinobacteria_Actinobacteria_Actinomycetales_Nakamurellaceae
Actinobacteria_Actinobacteria_Actinomycetales_Propionibacteriaceae
Bacteroidetes_Sphingobacteriia_Sphingobacteriales_Sphingobacteriaceae
Chrysiogenetes_Chrysiogenetes_Chrysiogenales_Chrysiogenaceae
Firmicutes_Bacilli_Bacillales_Bacillales_Incertae Sedis XI
Firmicutes_Clostridia_Halanaerobiales_Halanaerobiaceae
Parvarchaeota_Candidatus Parvarchaeum_Candidatus Parvarchaeum_Candidatus Parvarchaeum
Proteobacteria_Alphaproteobacteria_Rhizobiales_Brucellaceae
Proteobacteria_Alphaproteobacteria_Rhizobiales_Hyphomicrobiaceae
Proteobacteria_Alphaproteobacteria_SAR11_SAR11

Microbes that show an decrease

Actinobacteria_Actinobacteria_Actinomycetales_Corynebacteriaceae
Bacteroidetes_Bacteroidia_Bacteroidales_Bacteroidales_incertae_sedis
Bacteroidetes_Flavobacteriia_Flavobacteriales_Cryomorphaceae
Firmicutes_Bacilli_Lactobacillales_Lactobacillaceae
Firmicutes_Clostridia_Clostridiales_Clostridiales_Incertae Sedis XIII
Proteobacteria_Alphaproteobacteria_Rhizobiales_Rhizobiaceae
Proteobacteria_Betaproteobacteria_Burkholderiales_Burkholderiaceae
Proteobacteria_Betaproteobacteria_Rhodocyclales_Rhodocyclaceae
Proteobacteria_Gammaproteobacteria_Orbales_Orbaceae

Taxonomical relationships and groups among microbes with altered abundance:

Taxonomical groups:

Actinobacteria:
Actinobacteria_Actinobacteria_Actinomycetales_Corynebacteriaceae
Actinobacteria_Actinobacteria_Actinomycetales_Nakamurellaceae
Actinobacteria_Actinobacteria_Actinomycetales_Propionibacteriaceae

Bacteroidetes:
Bacteroidetes_Bacteroidia_Bacteroidales_Bacteroidales_incertae_sedis
Bacteroidetes_Flavobacteriia_Flavobacteriales_Cryomorphaceae
Bacteroidetes_Sphingobacteriia_Sphingobacteriales_Sphingobacteriaceae

Firmicutes:
Firmicutes_Bacilli_Bacillales_Bacillales_Incertae Sedis XI
Firmicutes_Bacilli_Lactobacillales_Lactobacillaceae
Firmicutes_Clostridia_Clostridiales_Clostridiales_Incertae Sedis XIII
Firmicutes_Clostridia_Halanaerobiales_Halanaerobiaceae

Proteobacteria:
Proteobacteria_Alphaproteobacteria_Rhizobiales_Brucellaceae
Proteobacteria_Alphaproteobacteria_Rhizobiales_Hyphomicrobiaceae
Proteobacteria_Alphaproteobacteria_Rhizobiales_Rhizobiaceae
Proteobacteria_Alphaproteobacteria_SAR11_SAR11
Proteobacteria_Betaproteobacteria_Burkholderiales_Burkholderiaceae
Proteobacteria_Betaproteobacteria_Rhodocyclales_Rhodocyclaceae
Proteobacteria_Gammaproteobacteria_Orbales_Orbaceae