Machine Learning in Computational Biology - Second assignment

Mastoraki Aikaterina – 7115152100031

On this assignment, we were asked to create a complete pipeline for dimensionality reduction, classification and visualization of 5 different datasets. To achieve the requested task I used Python programming language and jupyter notebook. The notebook that contains the code for this assignment is enclosed to the assignment's folder along with this report. First of all my code consists of 3 main parts:

- Imports
- Functions
- Main loop

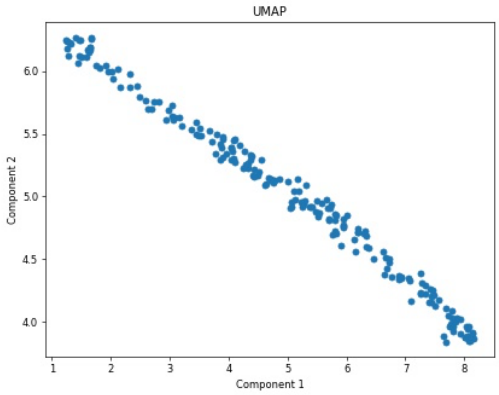
I have defined 5 functions. UMAP_pipeline, PCA_pipeline, TSNE_pipeline, GMM_clustering and BIC_calculation.

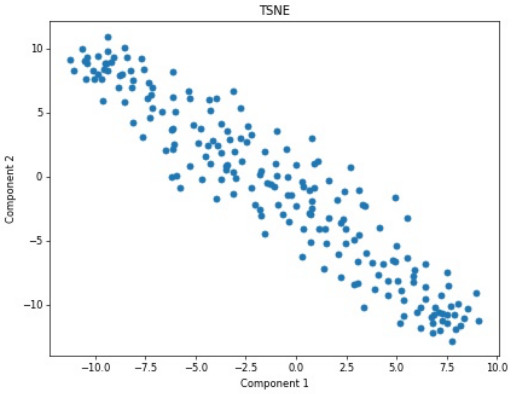
The main loop is a for loop that reads each of the 5 different single-cell synthetic datasets (the datasets should be in a folder named Assignment2_Datasets) , normalize their entries and calls the three dimensionality reduction algorithms and the clustering algorithm. UMAP is the first algorithm used for dimensionality reduction. UMAP is called with 13 neighbors and the minimum value between 50 and the number of samples divided by 8 as components. Then, PCA algorithm is used through PCA_pipeline function. The function calculates the number of components needed for some specific percentage of data variance to be maintained after the reduction (default set to 0.9) and does the reduction accordingly. The last algorithm used for dimensionality reduction is the TSNE. It has been proved that TSNE gives better results if the data are already somehow reduced by PCA. Thus, TSNE performs dimensionality reduction on PCA's output data and bring their dimension to the 2D space.

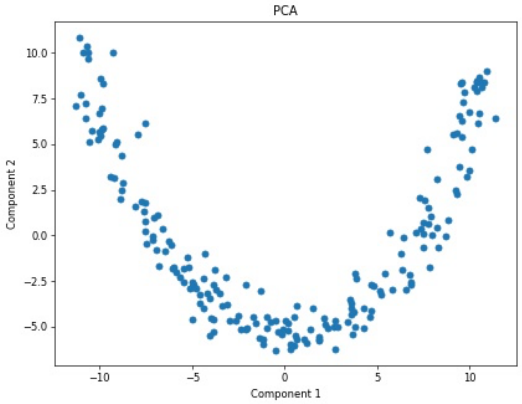
Once the dimensionality reduction step is over, the GMM_clustering function is called for the resulted reduced data of both PCA and TSNE. This function calls the BIC_calculation function to calculate the different BIC values for different number of clusters and different gaussian modes. The best BIC value is selected and the clustering is performed for the specific number of clusters and the specific gaussian mode.

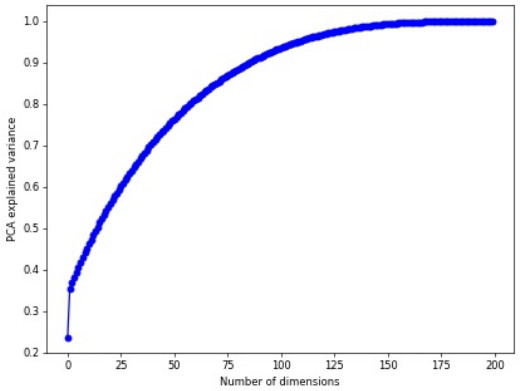
First dataset

- Dimensionality reduction:



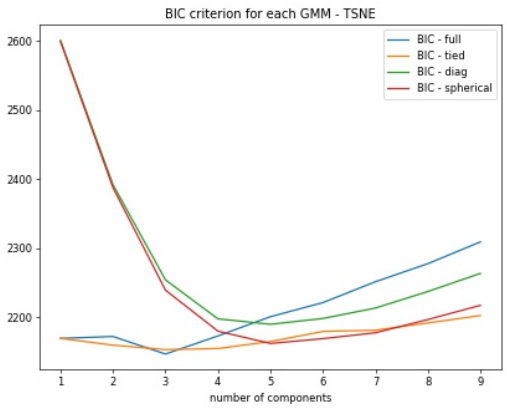


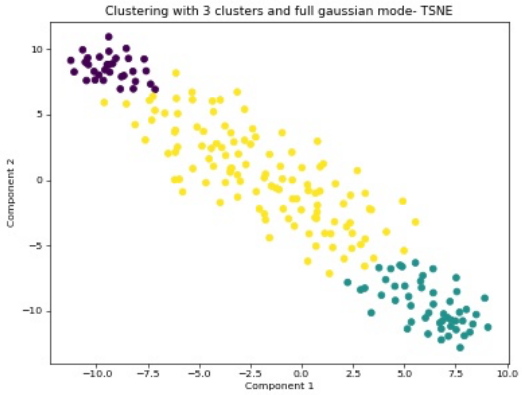


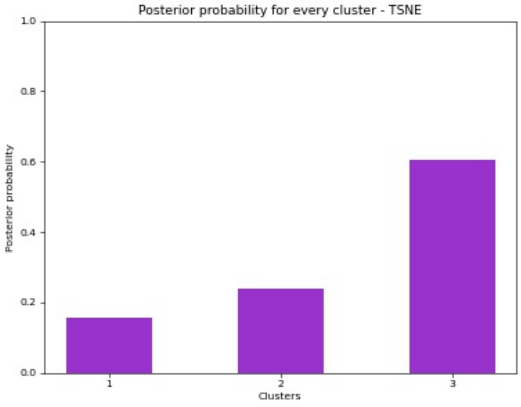



```
PCA needs 86 components to reach 90.0% of the total variance
```

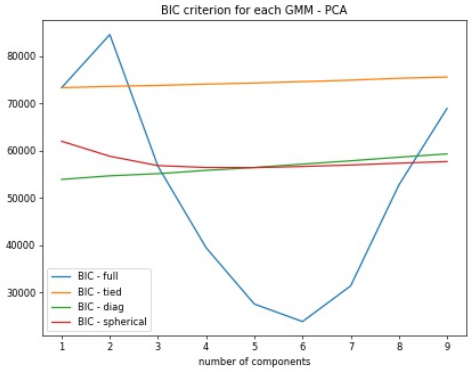
- GMM clustering -TSNE:

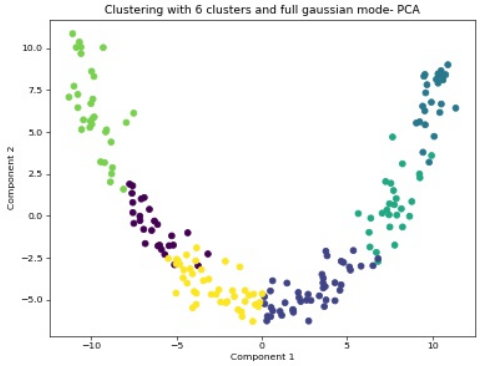


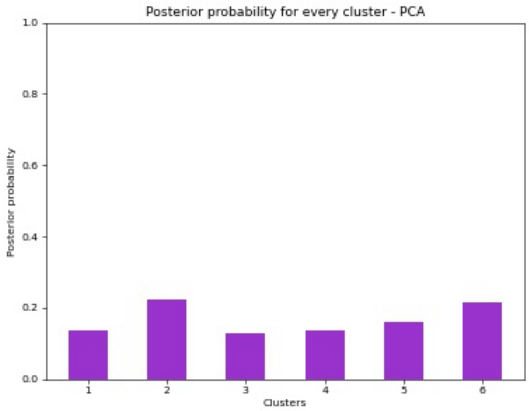




- GMM clustering -PCA:

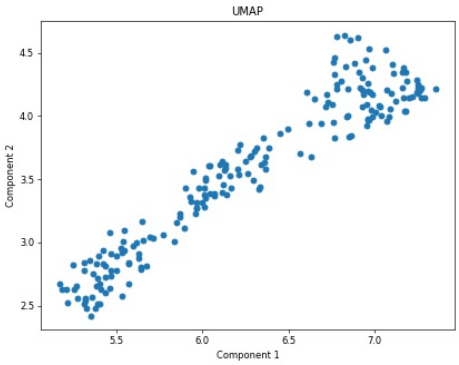


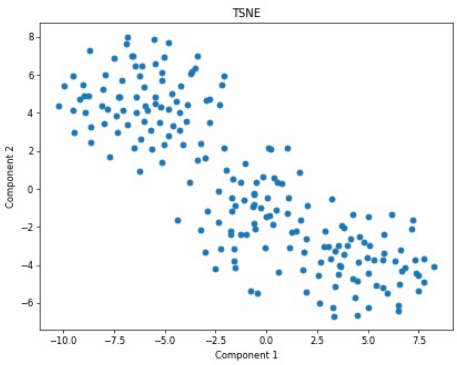


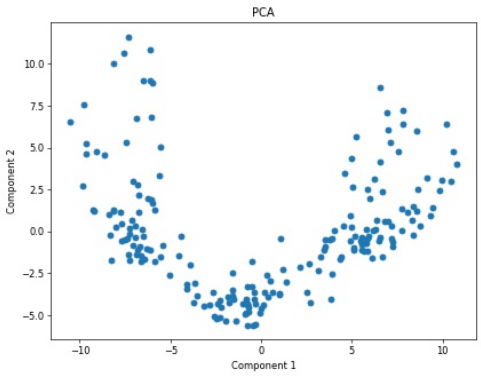


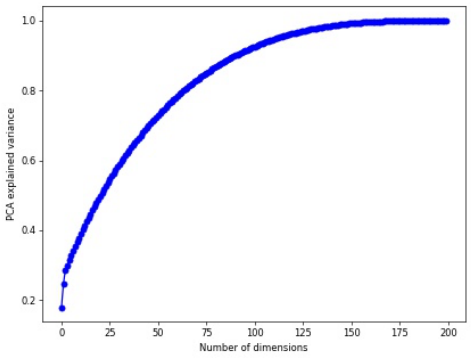
Second dataset

- Dimensionality reduction:



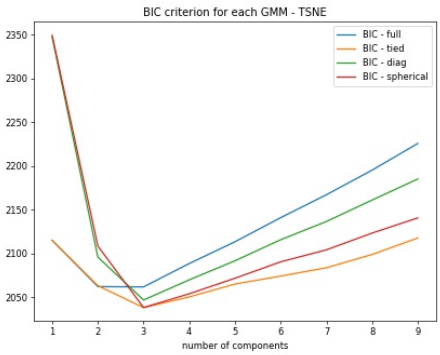


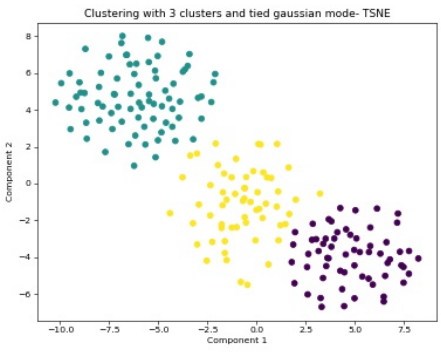


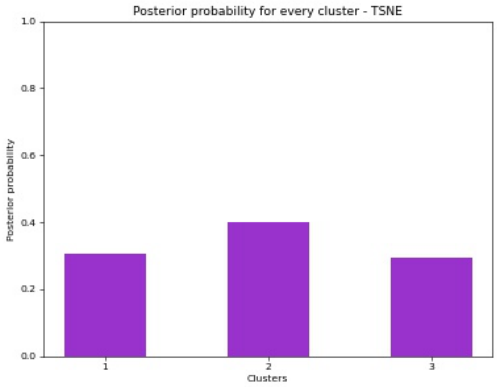



```
PCA needs 90 components to reach 90.0% of the total variance
```

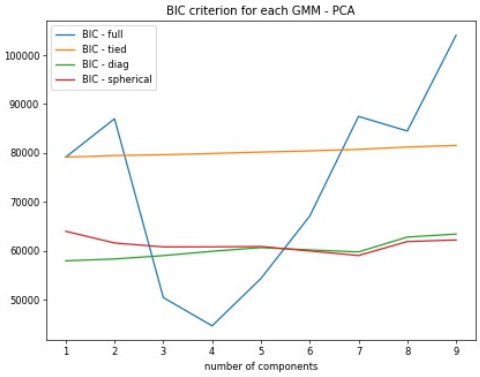
- GMM clustering -TSNE:

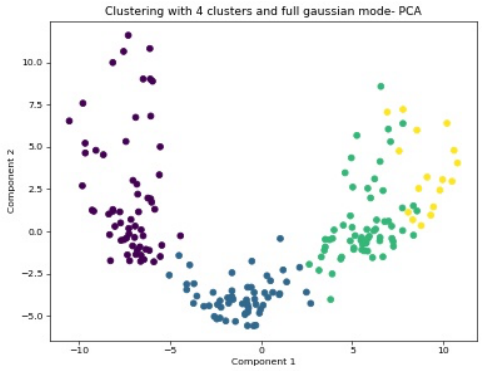


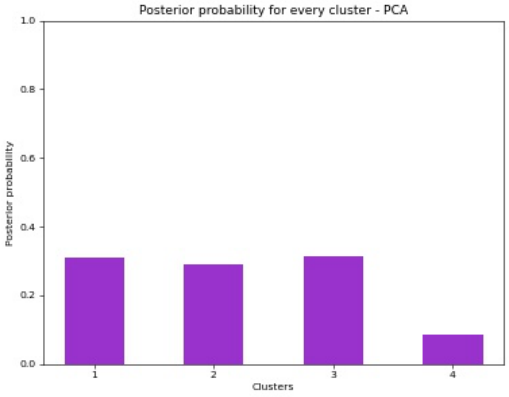




- GMM clustering -PCA:

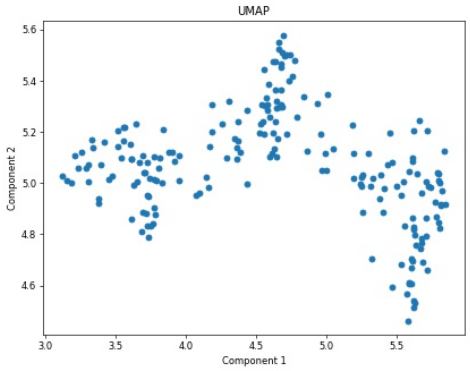


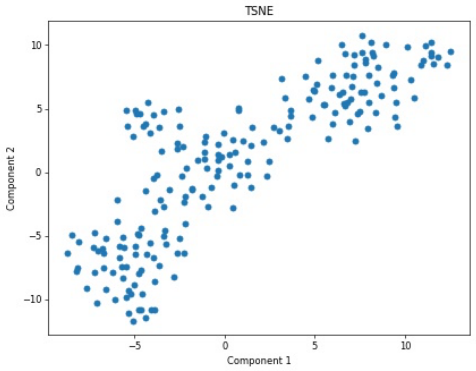


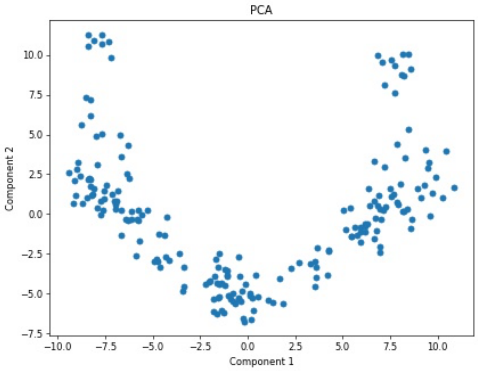


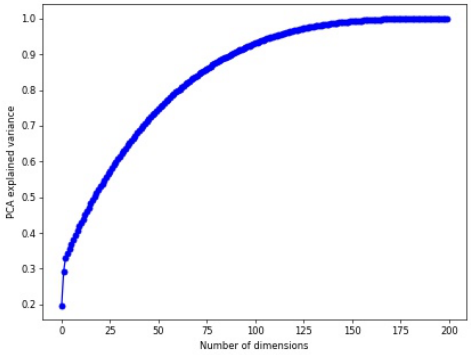
Third dataset

- Dimensionality reduction:



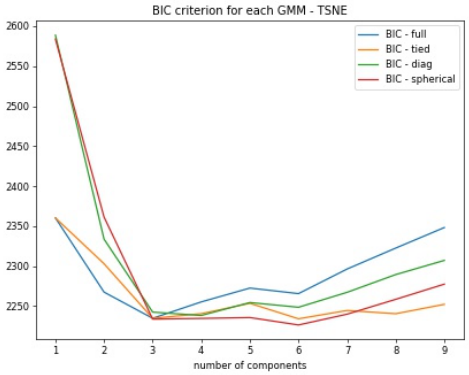


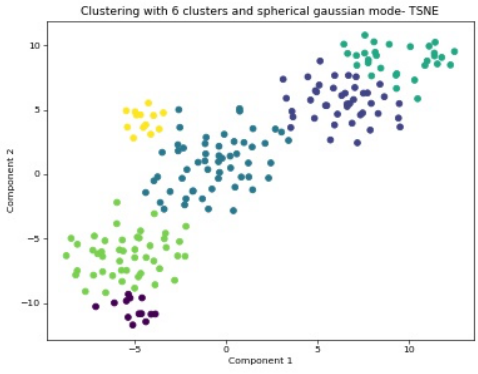


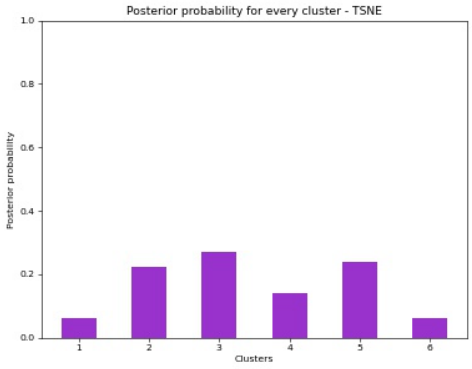



```
PCA needs 88 components to reach 90.0% of the total variance
```

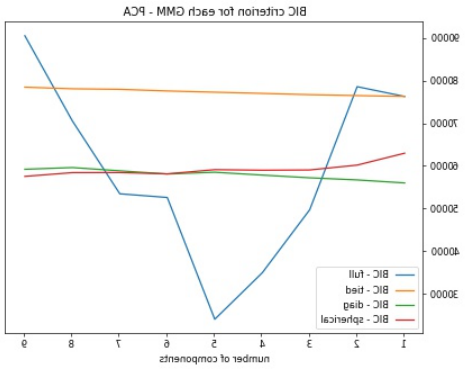
- GMM clustering -TSNE:

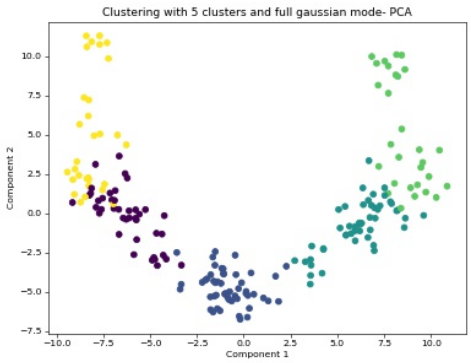


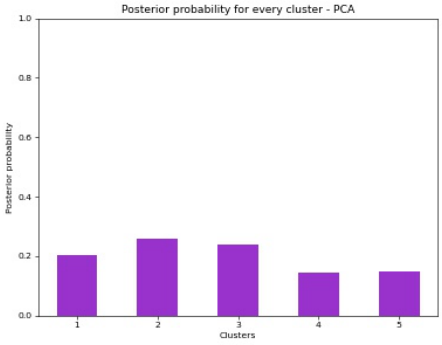




- GMM clustering -PCA:

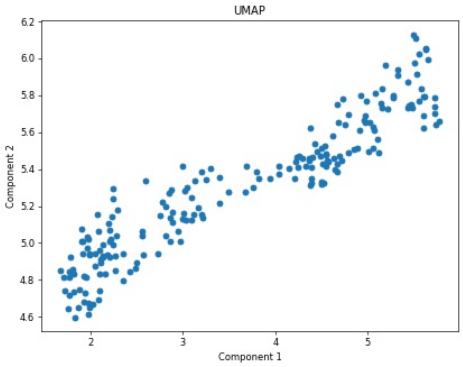


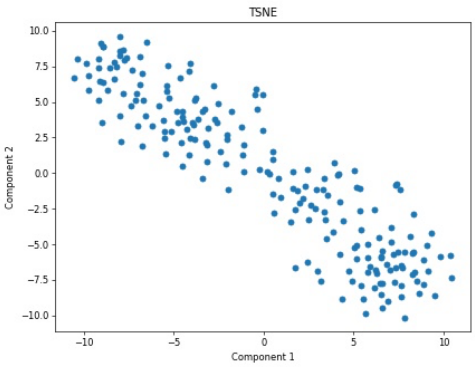


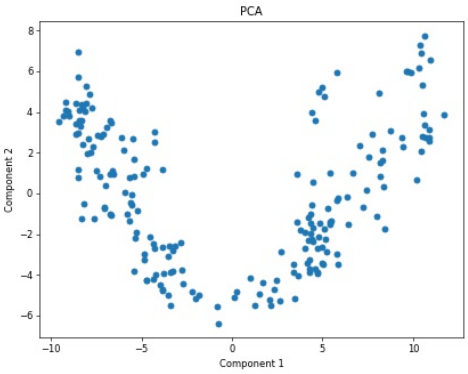


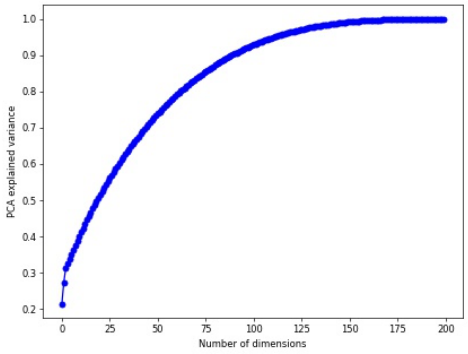
Fourth dataset

- Dimensionality reduction:



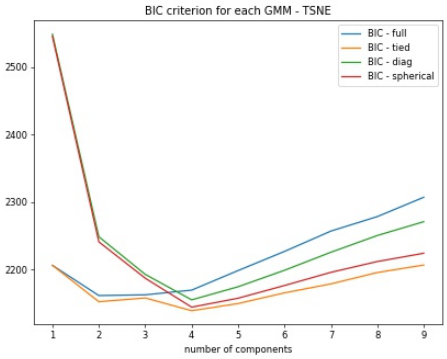


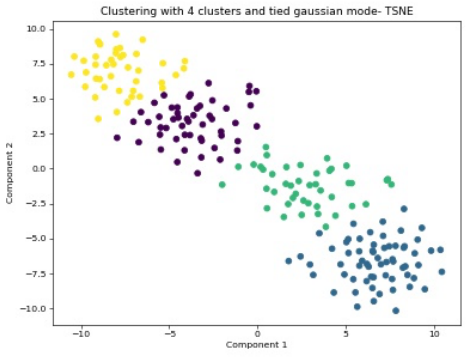


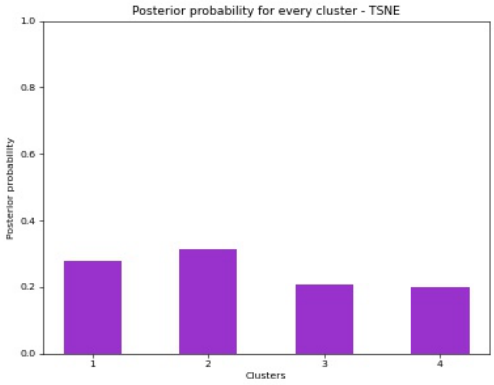



```
PCA needs 89 components to reach 90.0% of the total variance
```

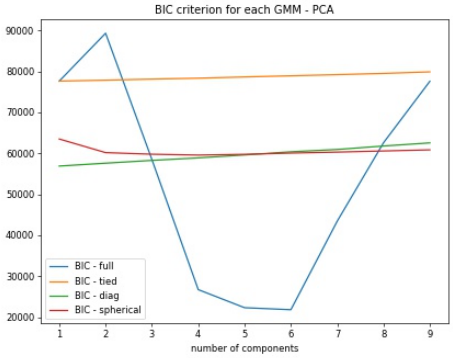
- GMM clustering -TSNE:

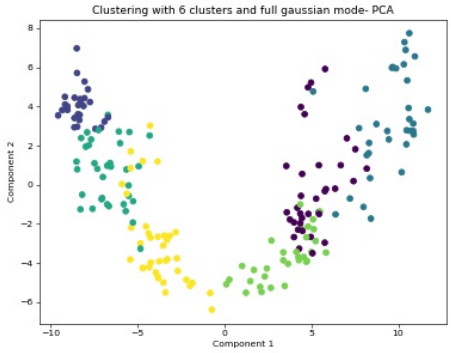


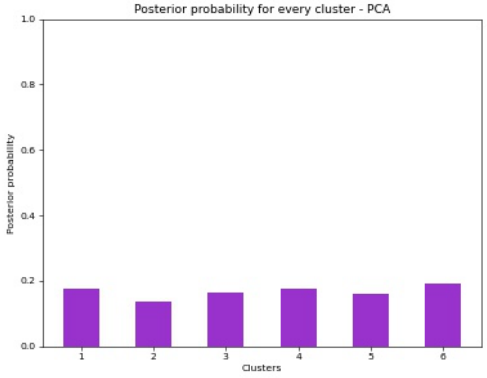




- GMM clustering -PCA:

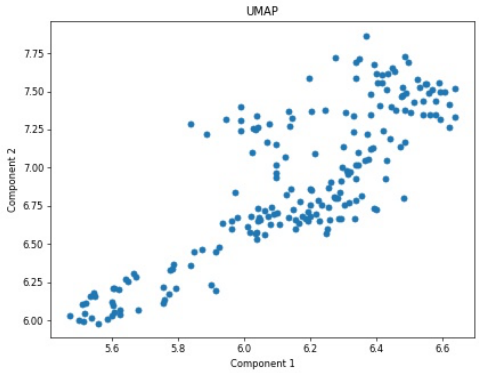


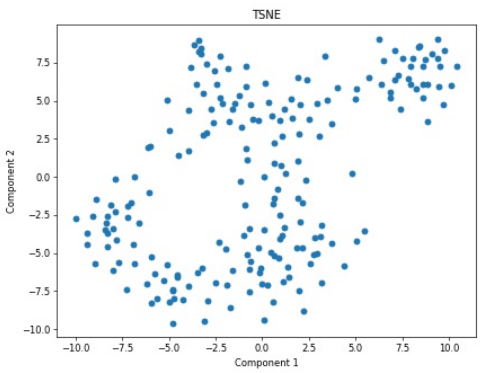


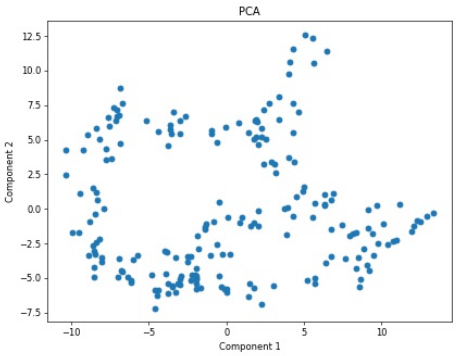


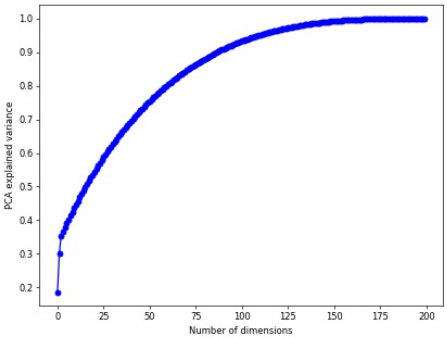
Fifth dataset

- Dimensionality reduction:



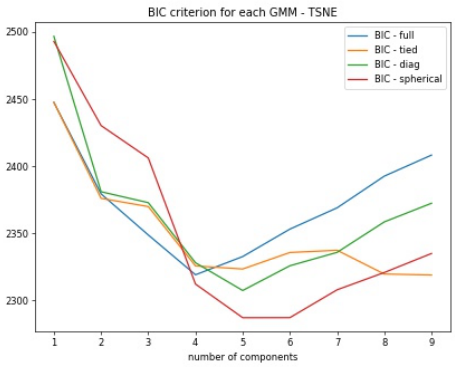


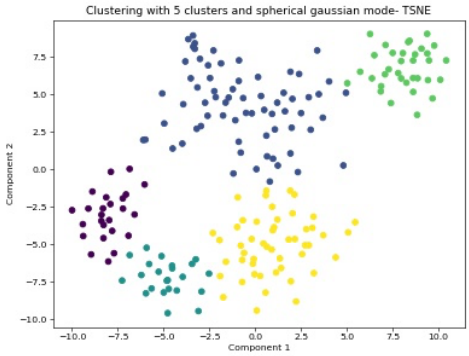


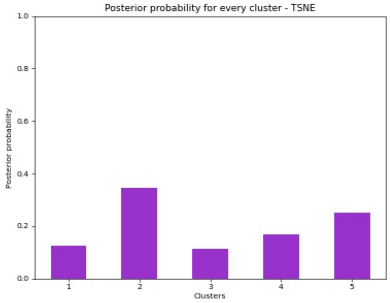



```
PCA needs 87 components to reach 90.0% of the total variance
```

- GMM clustering -TSNE:







- GMM clustering -PCA: