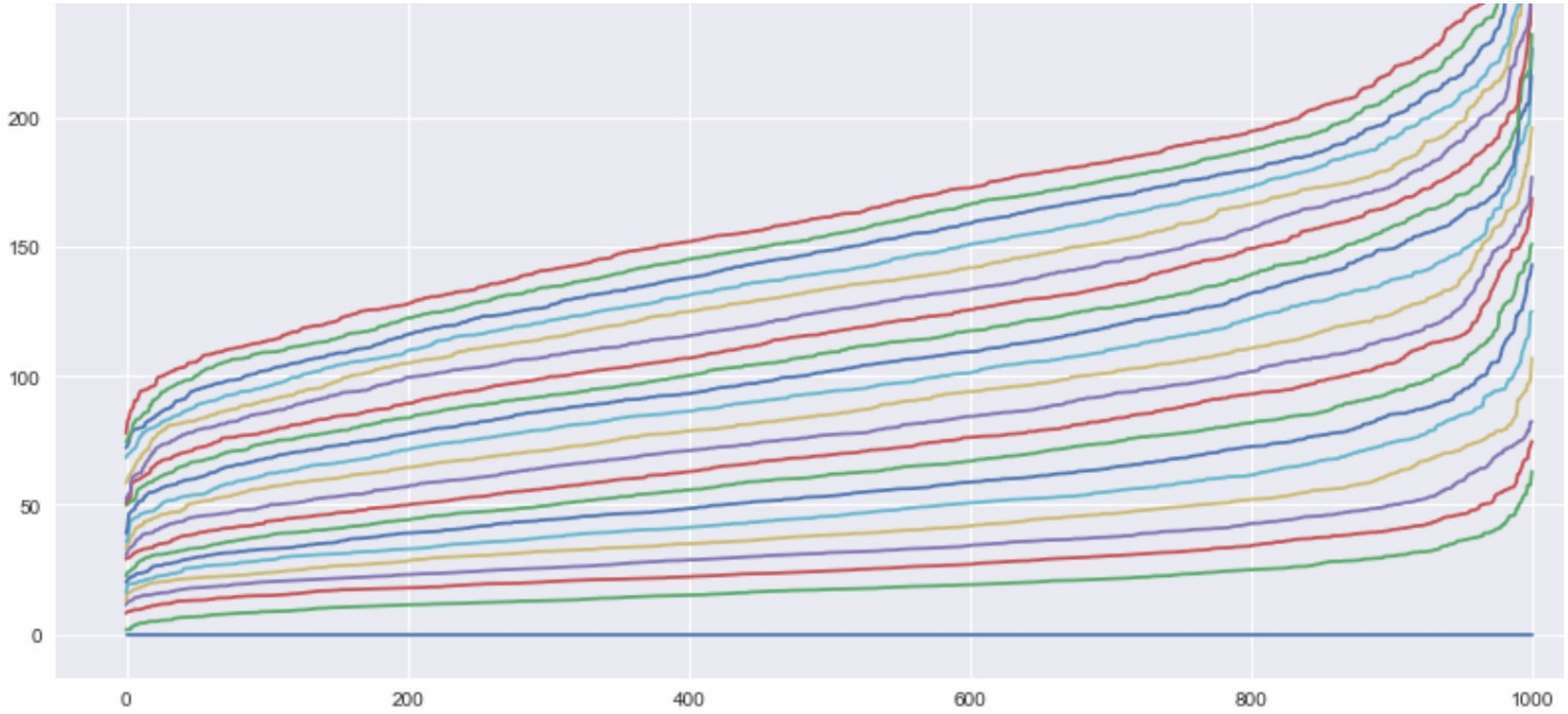
Data visualization
& clustering

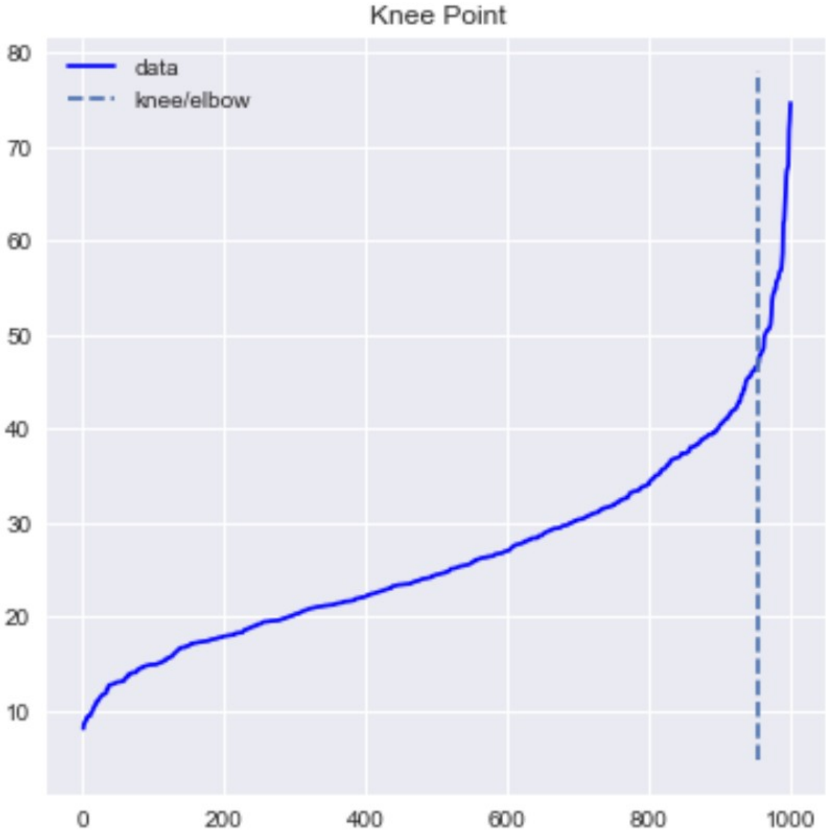
DBSCAN: find *eps*

2219

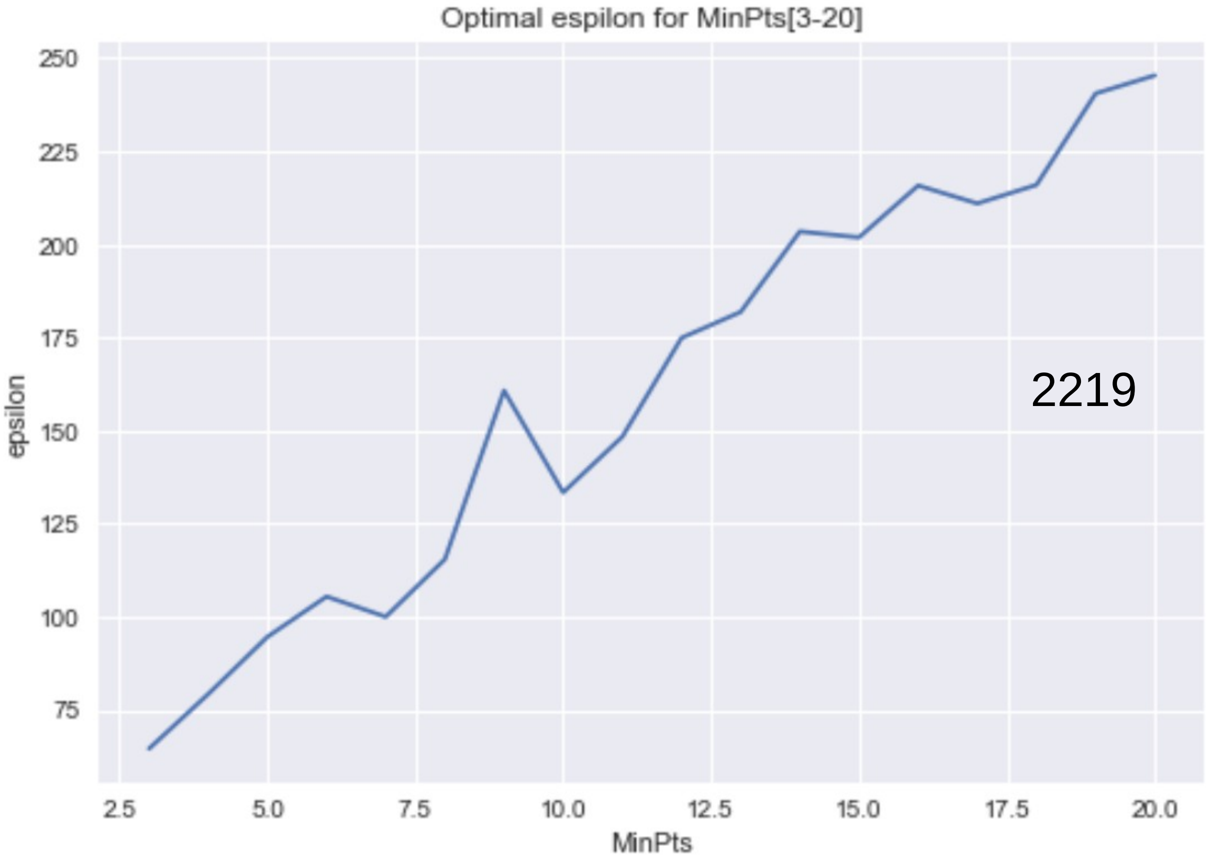


To find the threshold for the knee or elbow, the best approach we found was to use the &kneed$ python package. Given x and y arrays, kneed attempts to identify the knee/elbow point of a line fit to the data. The knee/elbow is defined as the point of the line with maximum curvature.

DBSCAN: find *eps*



Knee Point

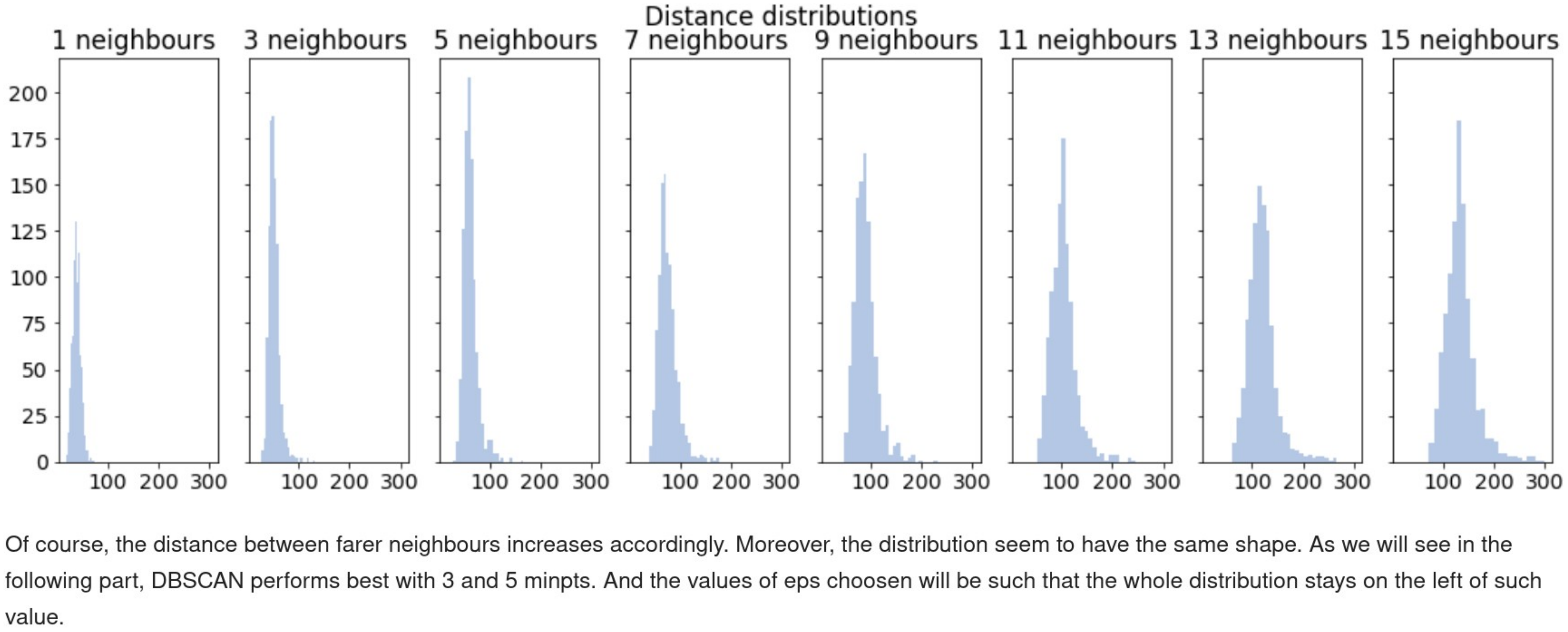


Optimal espilon for MinPts[3-20]

2219

DBSCAN: find *eps*

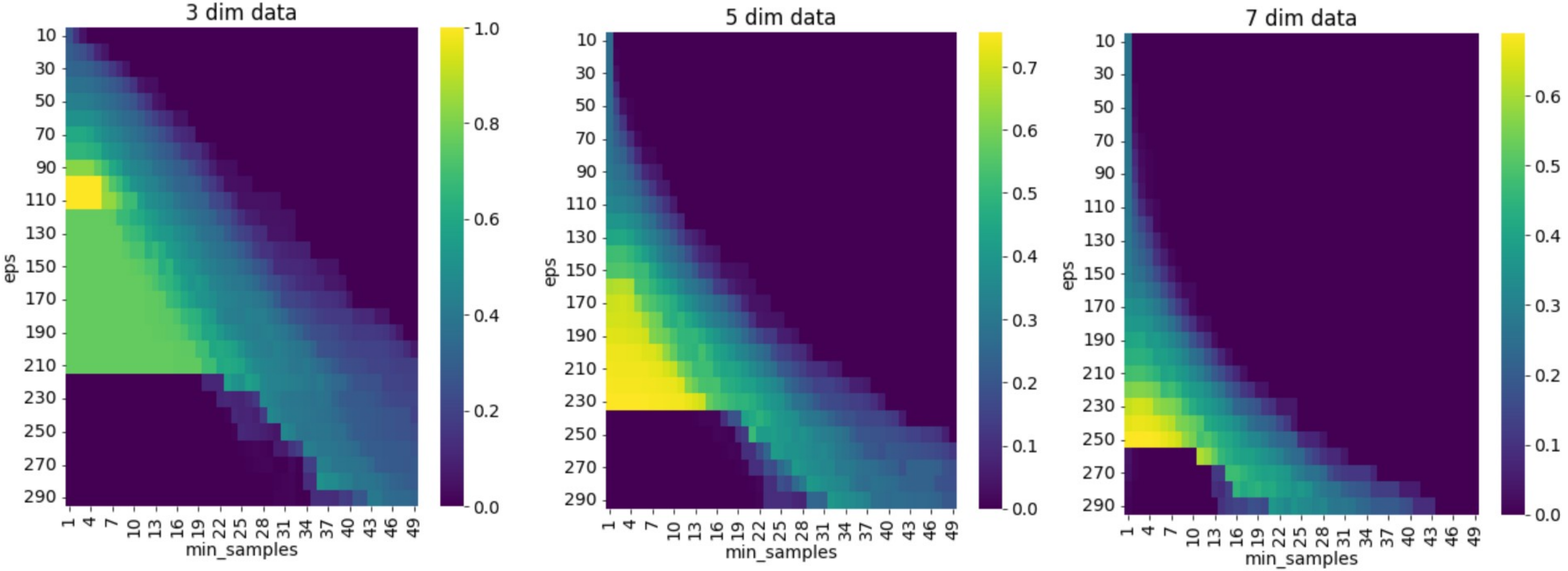
2203



Of course, the distance between farer neighbours increases accordingly. Moreover, the distribution seem to have the same shape. As we will see in the following part, DBSCAN performs best with 3 and 5 minpts. And the values of eps choosen will be such that the whole distribution stays on the left of such value.

DBSCAN: eps & minPts

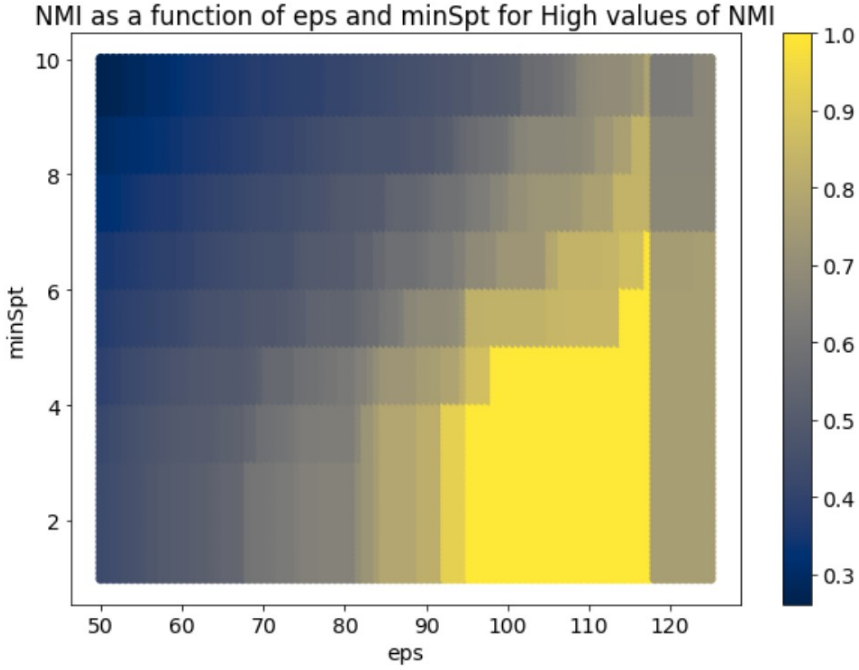
2216



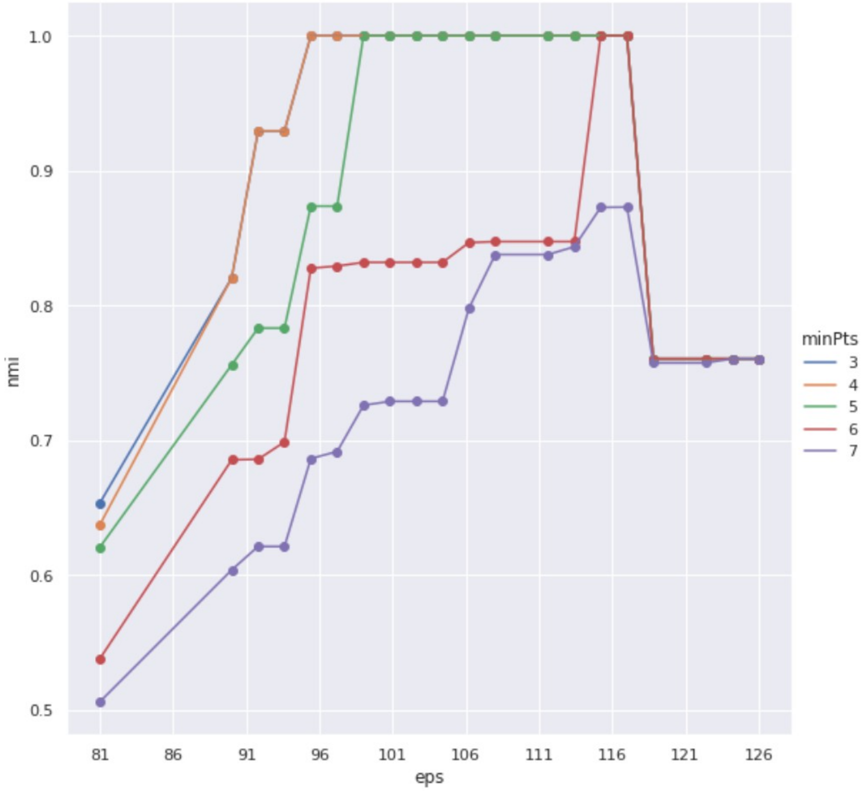
Usually, as a rule of thumb, minPts is often set to 2 times the dimensionality of the data.

DBSCAN: eps & minPts

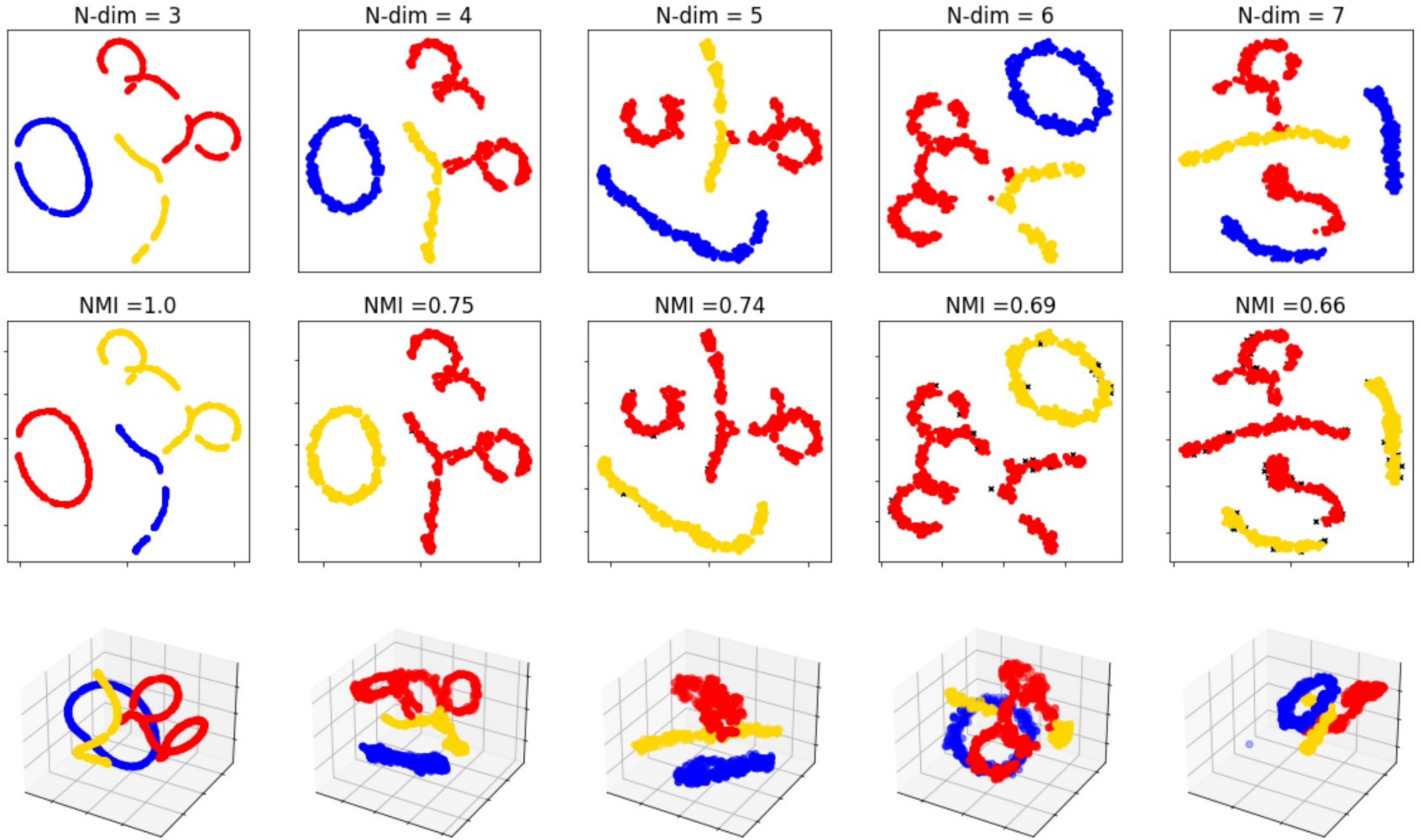
2211



2201



t-SNE and dimensions (L)



Affiliation Calculation

2220

If two points are strongly affiliated then they are more likly to be in the same cluster. If two points are closer together after T-SNE is applied then they are said to be more strongly affiliated. If they are closer together for many stochasitc itterations of t-SNE then they are also more strongly affiliated. The affiliation of point p between all other points P is calculated as shown below.

$$Affiliation = 1 - norm\left(\sum_i dist(p, P)\right)$$

The distance between two points, $p1 = (x_1, y_1)$ and $P_n = (x_2, y_2)$, is calculated as,

$$dist(p, P_n) = \sqrt((x_1 - x_2)^2 + (y_1 - y_2)^2).$$

dist(p,P) outputs an array of distances between p and every other point P_n. These distances are summed for every iteration of t-SNE, i. At this point p can be said to be most affiliated with the points where this sum is the lowest. Then they will have been closer together for many iterations of t-SNE. This sum is largly dependant on the number of itterations. To remove this dependancy the result is normalised as shown.

dist(p,P) outputs an array of distances between p and every other point P_n. These distances are summed for every iteration of t-SNE, i. At this point p can be said to be most affiliated with the points where this sum is the lowest. Then they will have been closer together for many iterations of t-SNE. This sum is largly dependant on the number of itterations. To remove this dependancy the result is normalised as shown.

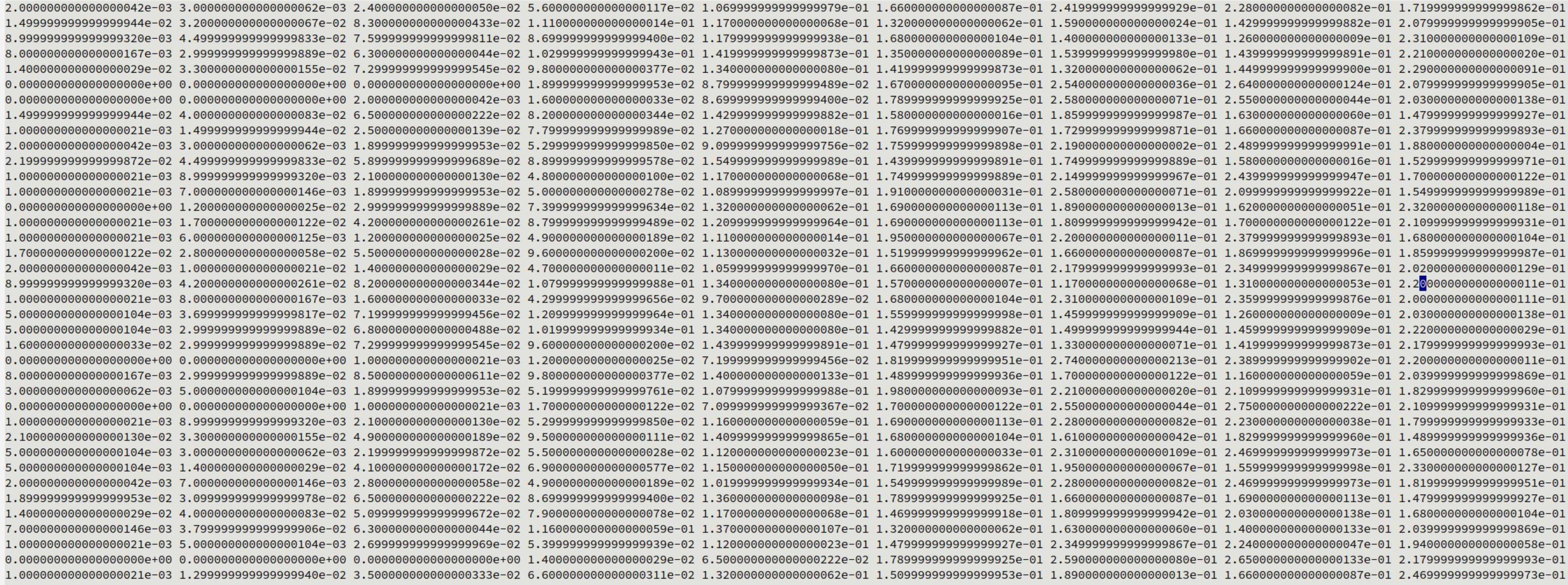
$$norm(data) = \frac{data - min(data)}{range(data)}$$

Where data is an array of numbers. Each element is the sum of the distances of p with P_n over many iterations of t-SNE. After normalisation each element in the array represents the affiliation of p with P_n, where numbers close to 0 indicate a strong affiliation and numbers close to 1 indicate a low affiliation. This is reversed by subtracting from 1. This makes affiliation easier to represent graphically.

Thus the affiliation of p with another point P_n, has been quantised as a number ranging from 0 to 1. Where a value of 1 means they are strongly affiliated and have been placed close together in many iterations of t-SNE and 0 means they are weakly affiliated and have been placed far apart in many iterations of t-SNE.

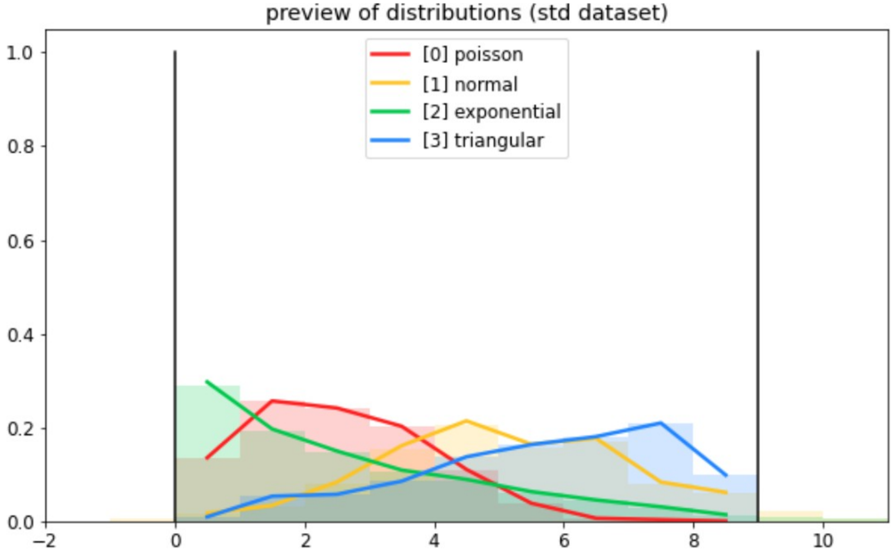
• Ex.5B: probability distributions

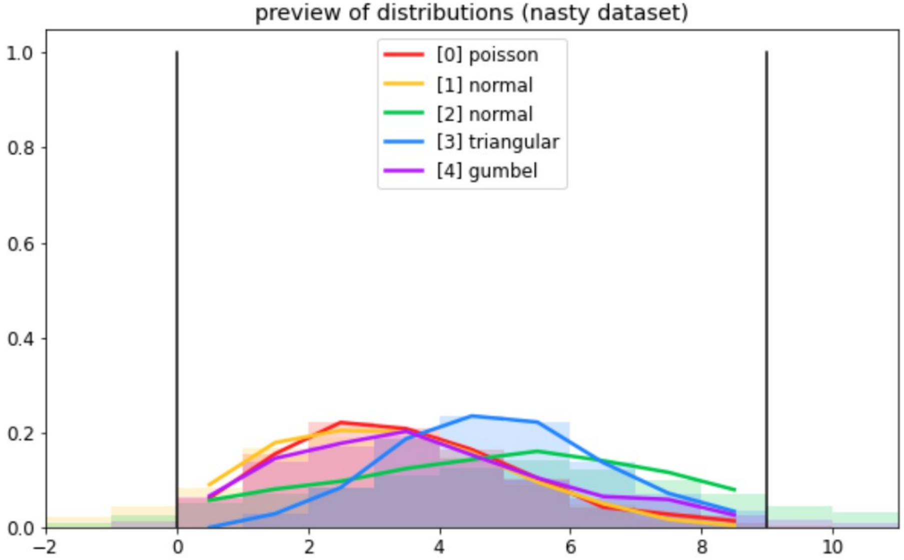
2202

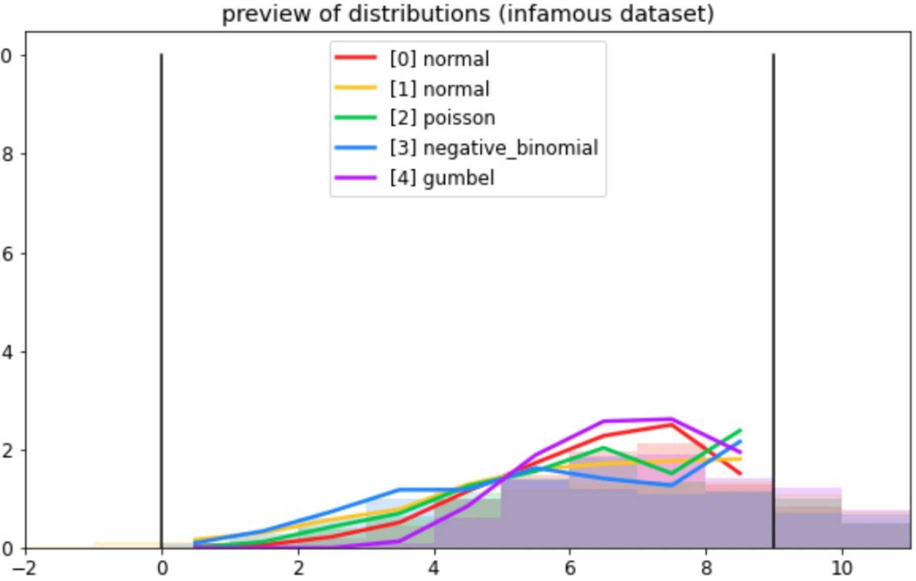


```
2.000000000000000042e-03 3.000000000000000062e-03 2.400000000000000050e-02 5.600000000000000117e-02 1.069999999999999979e-01 1.660000000000000087e-01 2.419999999999999929e-01 2.280000000000000082e-01 1.719999999999999862e-01
1.499999999999999944e-02 3.200000000000000067e-02 8.300000000000000433e-02 1.110000000000000014e-01 1.170000000000000068e-01 1.320000000000000062e-01 1.590000000000000024e-01 1.429999999999999882e-01 2.079999999999999905e-01
8.999999999999999320e-03 4.499999999999999833e-02 7.599999999999999811e-02 8.699999999999999400e-02 1.179999999999999938e-01 1.680000000000000104e-01 1.400000000000000133e-01 1.260000000000000009e-01 2.310000000000000109e-01
8.000000000000000167e-03 2.999999999999999889e-02 6.300000000000000044e-02 1.029999999999999943e-01 1.419999999999999873e-01 1.350000000000000089e-01 1.539999999999999980e-01 1.439999999999999891e-01 2.210000000000000020e-01
1.400000000000000029e-02 3.300000000000000155e-02 7.299999999999999545e-02 9.800000000000000377e-02 1.340000000000000080e-01 1.419999999999999873e-01 1.320000000000000062e-01 1.449999999999999900e-01 2.290000000000000091e-01
0.000000000000000000e+00 0.000000000000000000e+00 0.000000000000000000e+00 1.899999999999999953e-02 8.799999999999999489e-02 1.670000000000000095e-01 2.540000000000000036e-01 2.640000000000000124e-01 2.079999999999999905e-01
0.000000000000000000e+00 0.000000000000000000e+00 2.000000000000000042e-03 1.600000000000000033e-02 8.699999999999999400e-02 1.789999999999999925e-01 2.580000000000000071e-01 2.550000000000000044e-01 2.030000000000000138e-01
1.499999999999999944e-02 4.000000000000000083e-02 6.500000000000000222e-02 8.200000000000000344e-02 1.429999999999999882e-01 1.580000000000000016e-01 1.859999999999999987e-01 1.630000000000000060e-01 1.479999999999999927e-01
1.000000000000000021e-03 1.499999999999999944e-02 2.500000000000000139e-02 7.799999999999999989e-02 1.270000000000000018e-01 1.769999999999999907e-01 1.729999999999999871e-01 1.660000000000000087e-01 2.379999999999999893e-01
2.000000000000000042e-03 3.000000000000000062e-03 1.899999999999999953e-02 5.299999999999999850e-02 9.099999999999999756e-02 1.759999999999999898e-01 2.190000000000000002e-01 2.489999999999999991e-01 1.880000000000000004e-01
2.199999999999999872e-02 4.499999999999999833e-02 5.899999999999999689e-02 8.899999999999999578e-02 1.549999999999999989e-01 1.439999999999999891e-01 1.749999999999999889e-01 1.580000000000000016e-01 1.529999999999999971e-01
1.000000000000000021e-03 8.999999999999999320e-03 2.100000000000000130e-02 4.800000000000000100e-02 1.170000000000000068e-01 1.749999999999999889e-01 2.149999999999999967e-01 2.439999999999999947e-01 1.700000000000000122e-01
1.000000000000000021e-03 7.000000000000000146e-03 1.899999999999999953e-02 5.000000000000000278e-02 1.089999999999999997e-01 1.910000000000000031e-01 2.580000000000000071e-01 2.099999999999999922e-01 1.549999999999999989e-01
0.000000000000000000e+00 1.200000000000000025e-02 2.999999999999999889e-02 7.399999999999999634e-02 1.320000000000000062e-01 1.690000000000000113e-01 1.890000000000000013e-01 1.620000000000000051e-01 2.320000000000000118e-01
1.000000000000000021e-03 1.700000000000000122e-02 4.200000000000000261e-02 8.799999999999999489e-02 1.209999999999999964e-01 1.690000000000000113e-01 1.809999999999999942e-01 1.700000000000000122e-01 2.109999999999999931e-01
1.000000000000000021e-03 6.000000000000000125e-03 1.200000000000000025e-02 4.900000000000000189e-02 1.110000000000000014e-01 1.950000000000000067e-01 2.200000000000000011e-01 2.379999999999999893e-01 1.680000000000000104e-01
1.700000000000000122e-02 2.800000000000000058e-02 5.500000000000000028e-02 9.600000000000000200e-02 1.130000000000000032e-01 1.519999999999999962e-01 1.660000000000000087e-01 1.869999999999999996e-01 1.859999999999999987e-01
2.000000000000000042e-03 1.000000000000000021e-02 1.400000000000000029e-02 4.700000000000000011e-02 1.059999999999999970e-01 1.660000000000000087e-01 2.179999999999999993e-01 2.349999999999999867e-01 2.020000000000000129e-01
8.999999999999999320e-03 4.200000000000000261e-02 8.200000000000000344e-02 1.079999999999999988e-01 1.340000000000000080e-01 1.570000000000000007e-01 1.170000000000000068e-01 1.310000000000000053e-01 2.200000000000000011e-01
1.000000000000000021e-03 8.000000000000000167e-03 1.600000000000000033e-02 4.299999999999999656e-02 9.700000000000000289e-02 1.680000000000000104e-01 2.310000000000000109e-01 2.359999999999999876e-01 2.000000000000000111e-01
5.000000000000000104e-03 3.699999999999999817e-02 7.199999999999999456e-02 1.209999999999999964e-01 1.340000000000000080e-01 1.559999999999999998e-01 1.459999999999999909e-01 1.260000000000000009e-01 2.030000000000000138e-01
5.000000000000000104e-03 2.999999999999999889e-02 6.800000000000000488e-02 1.019999999999999934e-01 1.340000000000000080e-01 1.429999999999999882e-01 1.499999999999999944e-01 1.459999999999999909e-01 2.220000000000000029e-01
1.600000000000000033e-02 2.999999999999999889e-02 7.299999999999999545e-02 9.600000000000000200e-02 1.439999999999999891e-01 1.479999999999999927e-01 1.330000000000000071e-01 1.419999999999999873e-01 2.179999999999999993e-01
0.000000000000000000e+00 0.000000000000000000e+00 1.000000000000000021e-03 1.200000000000000025e-02 7.199999999999999456e-02 1.819999999999999951e-01 2.740000000000000213e-01 2.389999999999999902e-01 2.200000000000000011e-01
8.000000000000000167e-03 2.999999999999999889e-02 8.500000000000000611e-02 9.800000000000000377e-02 1.400000000000000133e-01 1.489999999999999936e-01 1.700000000000000122e-01 1.160000000000000059e-01 2.039999999999999869e-01
3.000000000000000062e-03 5.000000000000000104e-03 1.899999999999999953e-02 5.199999999999999761e-02 1.079999999999999988e-01 1.980000000000000093e-01 2.210000000000000020e-01 2.109999999999999931e-01 1.829999999999999960e-01
0.000000000000000000e+00 0.000000000000000000e+00 1.000000000000000021e-03 1.700000000000000122e-02 7.099999999999999367e-02 1.700000000000000122e-01 2.550000000000000044e-01 2.750000000000000222e-01 2.109999999999999931e-01
1.000000000000000021e-03 8.999999999999999320e-03 2.100000000000000130e-02 5.299999999999999850e-02 1.160000000000000059e-01 1.690000000000000113e-01 2.280000000000000082e-01 2.230000000000000038e-01 1.799999999999999933e-01
2.100000000000000130e-02 3.300000000000000155e-02 4.900000000000000189e-02 9.500000000000000111e-02 1.409999999999999865e-01 1.680000000000000104e-01 1.610000000000000042e-01 1.829999999999999960e-01 1.489999999999999936e-01
5.000000000000000104e-03 3.000000000000000062e-03 2.199999999999999872e-02 5.500000000000000028e-02 1.120000000000000023e-01 1.600000000000000033e-01 2.310000000000000109e-01 2.469999999999999973e-01 1.650000000000000078e-01
5.000000000000000104e-03 1.400000000000000029e-02 4.100000000000000172e-02 6.900000000000000577e-02 1.150000000000000050e-01 1.719999999999999862e-01 1.950000000000000067e-01 1.559999999999999998e-01 2.330000000000000127e-01
2.000000000000000042e-03 7.000000000000000146e-03 2.800000000000000058e-02 4.900000000000000189e-02 1.019999999999999934e-01 1.549999999999999989e-01 2.280000000000000082e-01 2.469999999999999973e-01 1.819999999999999951e-01
1.899999999999999953e-02 3.099999999999999978e-02 6.500000000000000222e-02 8.699999999999999400e-02 1.360000000000000098e-01 1.789999999999999925e-01 1.660000000000000087e-01 1.690000000000000113e-01 1.479999999999999927e-01
1.400000000000000029e-02 4.000000000000000083e-02 5.099999999999999672e-02 7.900000000000000078e-02 1.170000000000000068e-01 1.469999999999999918e-01 1.809999999999999942e-01 2.030000000000000138e-01 1.680000000000000104e-01
7.000000000000000146e-03 3.799999999999999906e-02 6.300000000000000044e-02 1.160000000000000059e-01 1.370000000000000107e-01 1.320000000000000062e-01 1.630000000000000060e-01 1.400000000000000133e-01 2.039999999999999869e-01
1.000000000000000021e-03 5.000000000000000104e-03 2.699999999999999969e-02 5.399999999999999939e-02 1.120000000000000023e-01 1.479999999999999927e-01 2.349999999999999867e-01 2.240000000000000047e-01 1.940000000000000058e-01
0.000000000000000000e+00 0.000000000000000000e+00 0.000000000000000000e+00 1.400000000000000029e-02 6.500000000000000222e-02 1.789999999999999925e-01 2.590000000000000080e-01 2.650000000000000133e-01 2.179999999999999993e-01
1.000000000000000021e-03 1.299999999999999940e-02 3.500000000000000333e-02 6.600000000000000311e-02 1.320000000000000062e-01 1.509999999999999953e-01 1.890000000000000013e-01 1.660000000000000087e-01 2.469999999999999973e-01
```

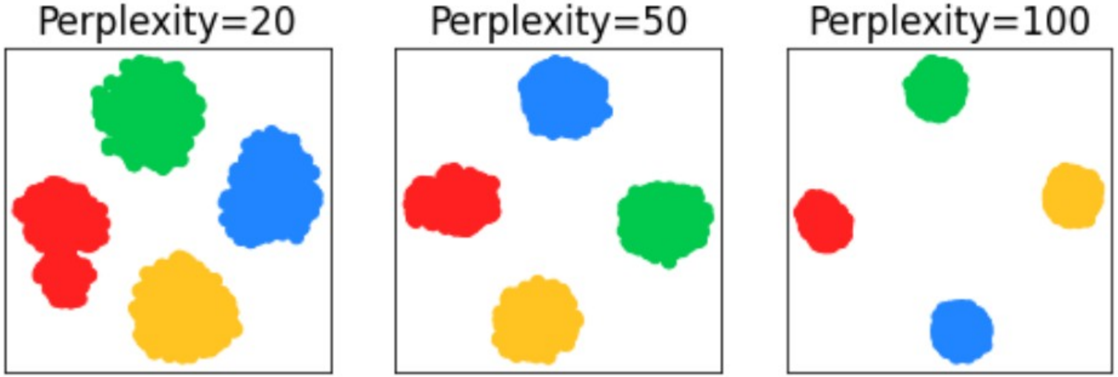
Datasets

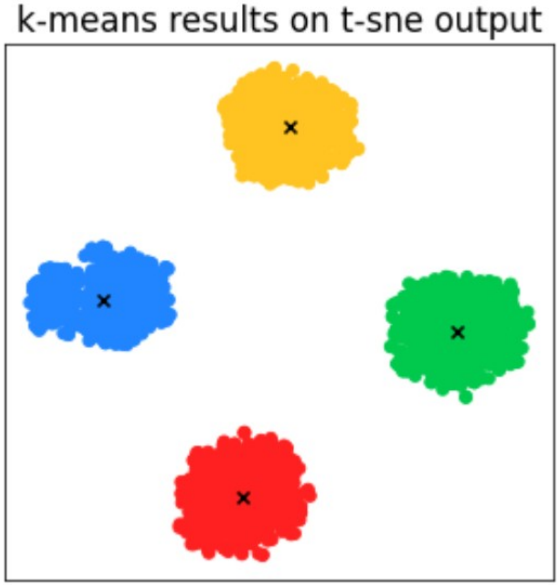

preview of distributions (std dataset)
[0] poisson
[1] normal
[2] exponential
[3] triangular
1.0
0.8
0.6
0.4
0.2
0.0
−2
0
2
4
6
8
10


preview of distributions (nasty dataset)
[0] poisson
[1] normal
[2] normal
[3] triangular
[4] gumbel
1.0
0.8
0.6
0.4
0.2
0.0
−2
0
2
4
6
8
10

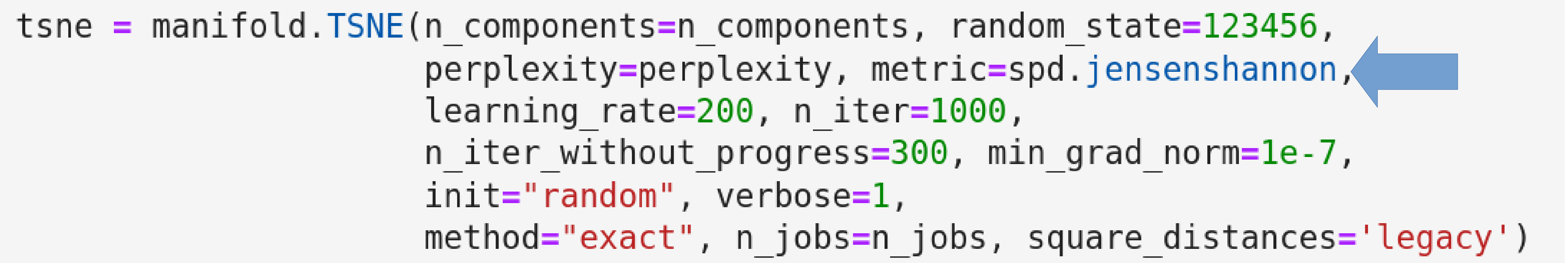

preview of distributions (infamous dataset)
[0] normal
[1] normal
[2] poisson
[3] negative_binomial
[4] gumbel
1.0
0.8
0.6
0.4
0.2
0.0
−2
0
2
4
6
8
10

T-SNE (Std. dataset)




k-means results on t-sne output

```python
import scipy.spatial.distance as spd
```

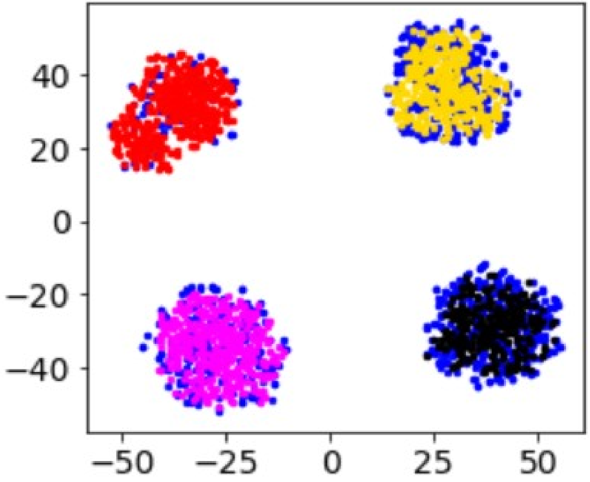


```python
tsne = manifold.TSNE(n_components=n_components, random_state=123456,
                     perplexity=perplexity, metric=spd.jensenshannon,
                     learning_rate=200, n_iter=1000,
                     n_iter_without_progress=300, min_grad_norm=1e-7,
                     init="random", verbose=1,
                     method="exact", n_jobs=n_jobs, square_distances='legacy')
```

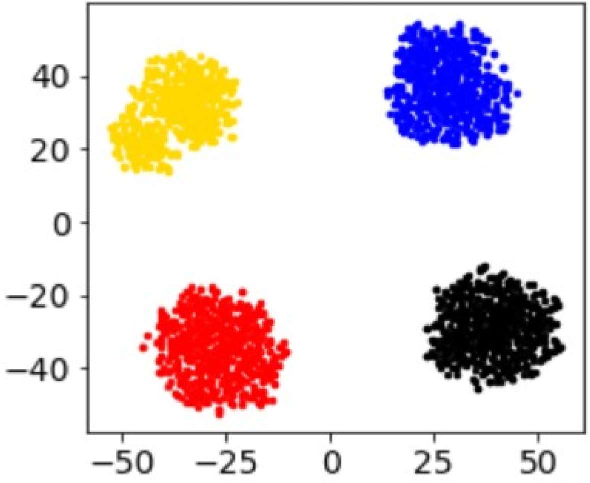
DBSCAN

Increasing eps

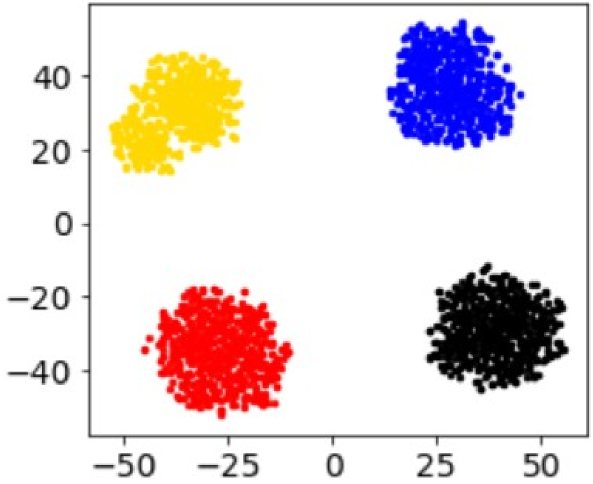
eps=0.02, minPts=10
nmi=0.67



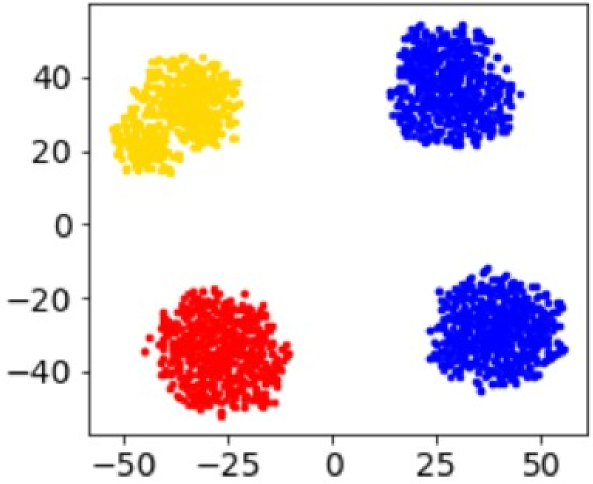
eps=0.03, minPts=10
nmi=1.00



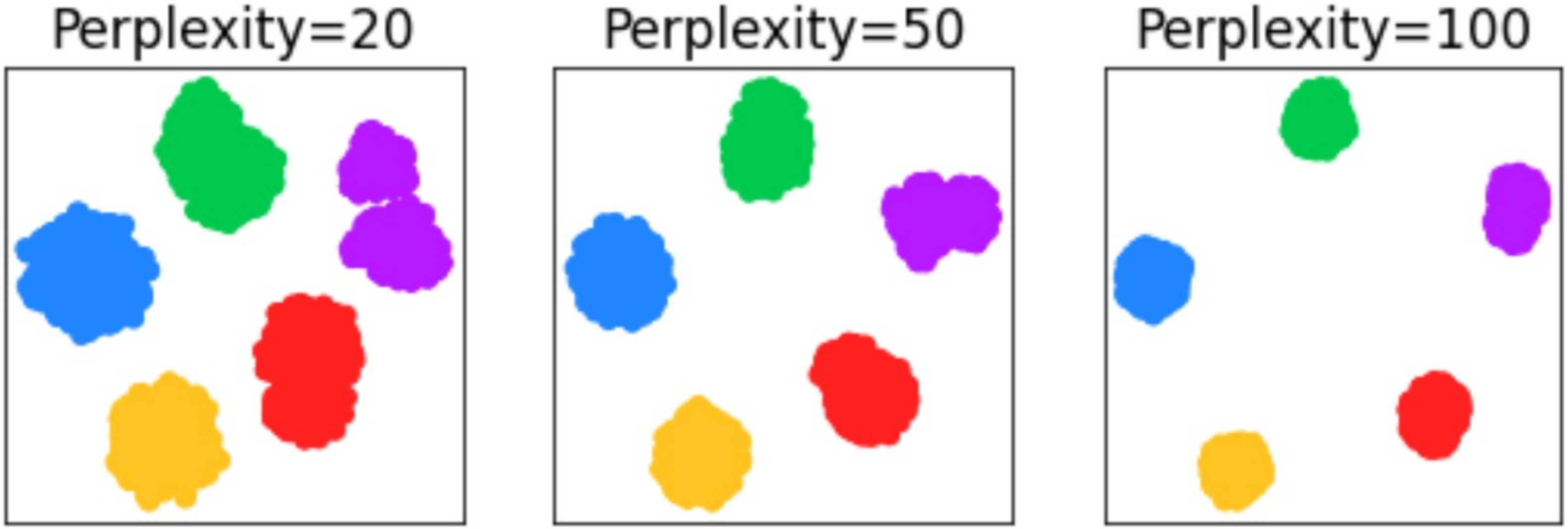
eps=0.07, minPts=10
nmi=1.00



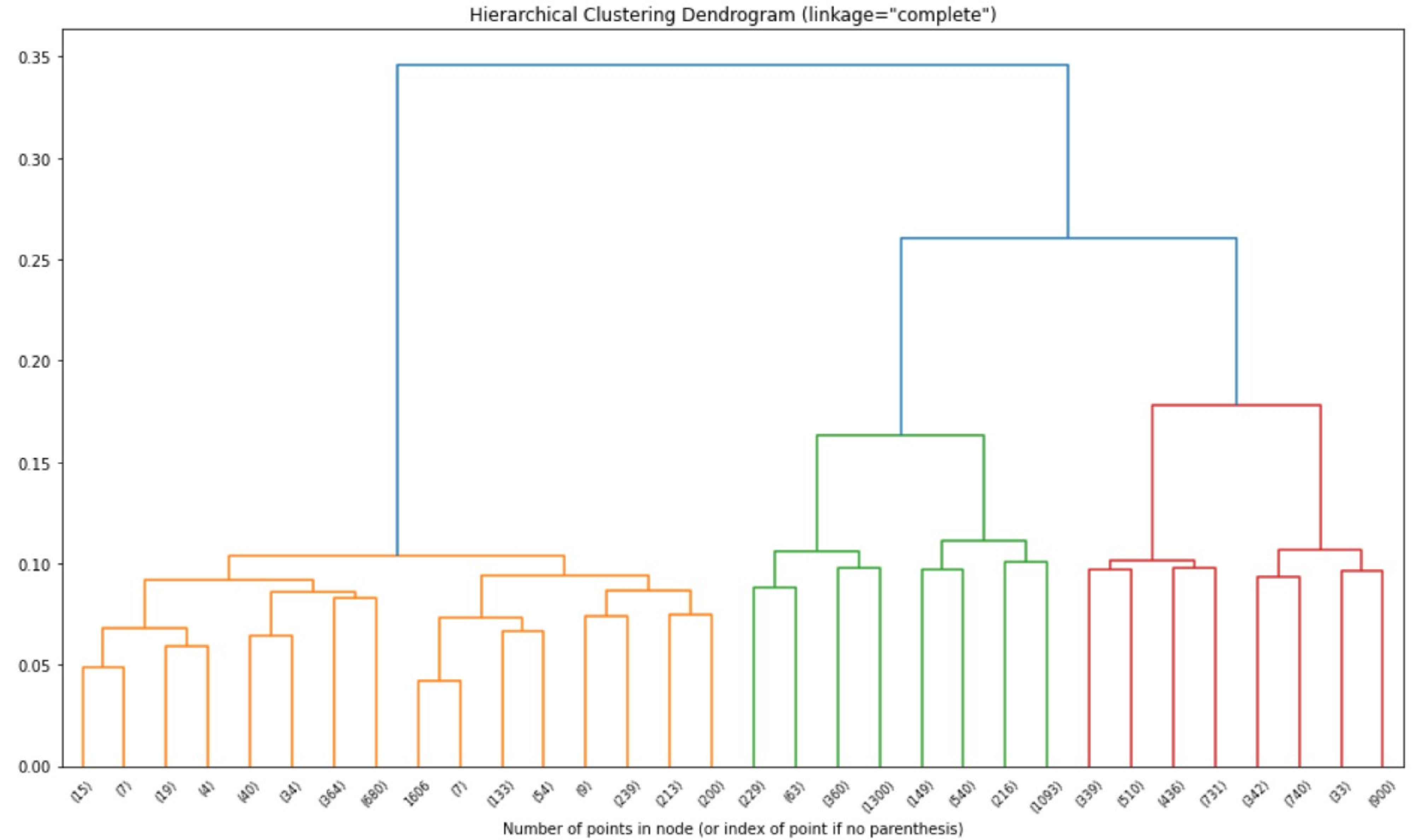
eps=0.14, minPts=10
nmi=0.85

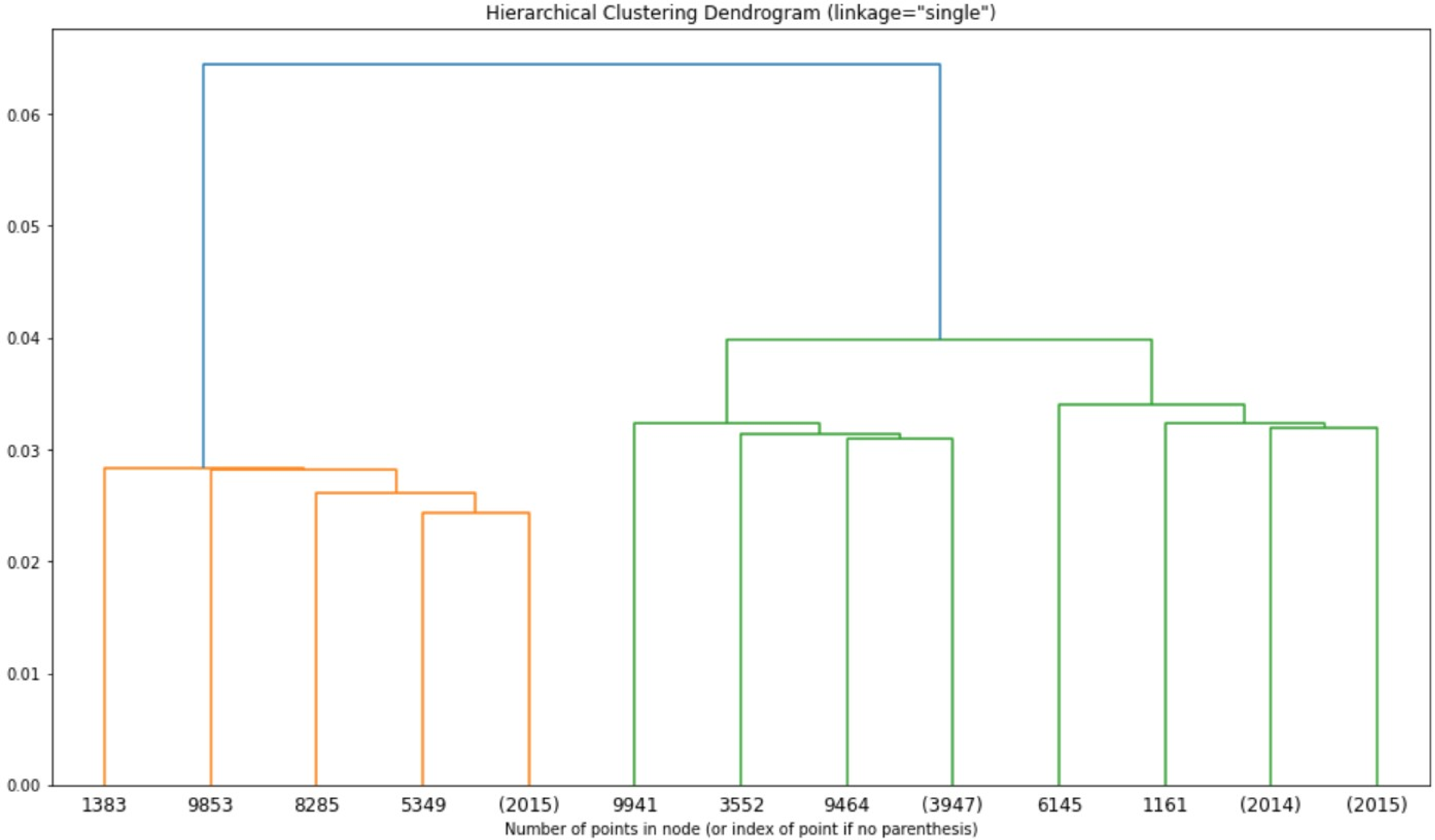


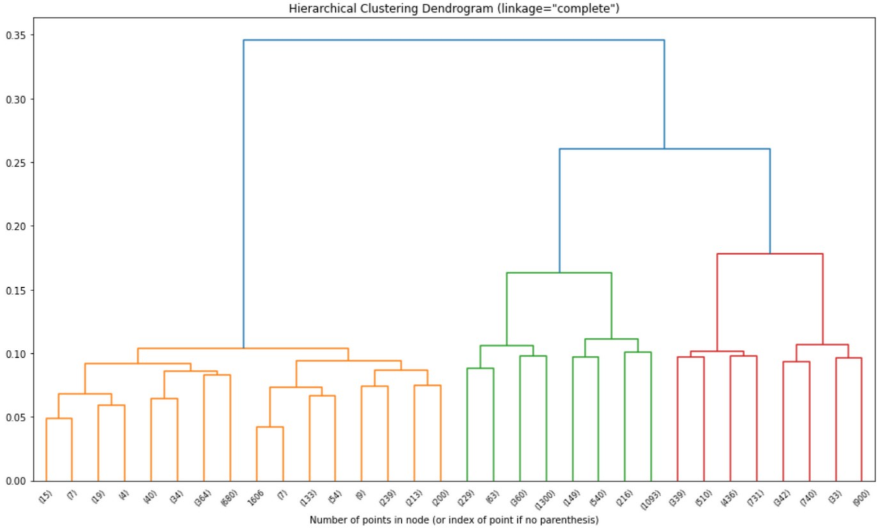
T-SNE (“infamous” dataset)

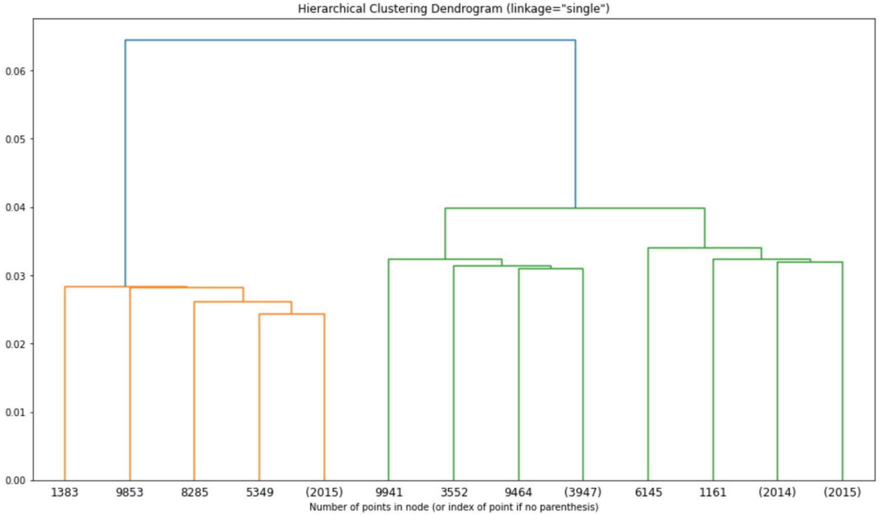


From the t-SNE results, we see that the algorithm is able to distinguish all the 5 distributions, even if the dataset has been generated with almost overlapping distributions.


Hierarchical Clustering Dendrogram (linkage="complete")
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
(15)
(7)
(19)
(4)
(40)
(34)
(364)
(680)
1606
(7)
(133)
(54)
(9)
(239)
(213)
(200)
(229)
(63)
(360)
(1300)
(149)
(540)
(216)
(1093)
(339)
(510)
(436)
(731)
(342)
(740)
(33)
(900)
Number of points in node (or index of point if no parenthesis)


Hierarchical Clustering Dendrogram (linkage="single")
0.06
0.05
0.04
0.03
0.02
0.01
0.00
1383
9853
8285
5349
(2015)
9941
3552
9464
(3947)
6145
1161
(2014)
(2015)
Number of points in node (or index of point if no parenthesis)





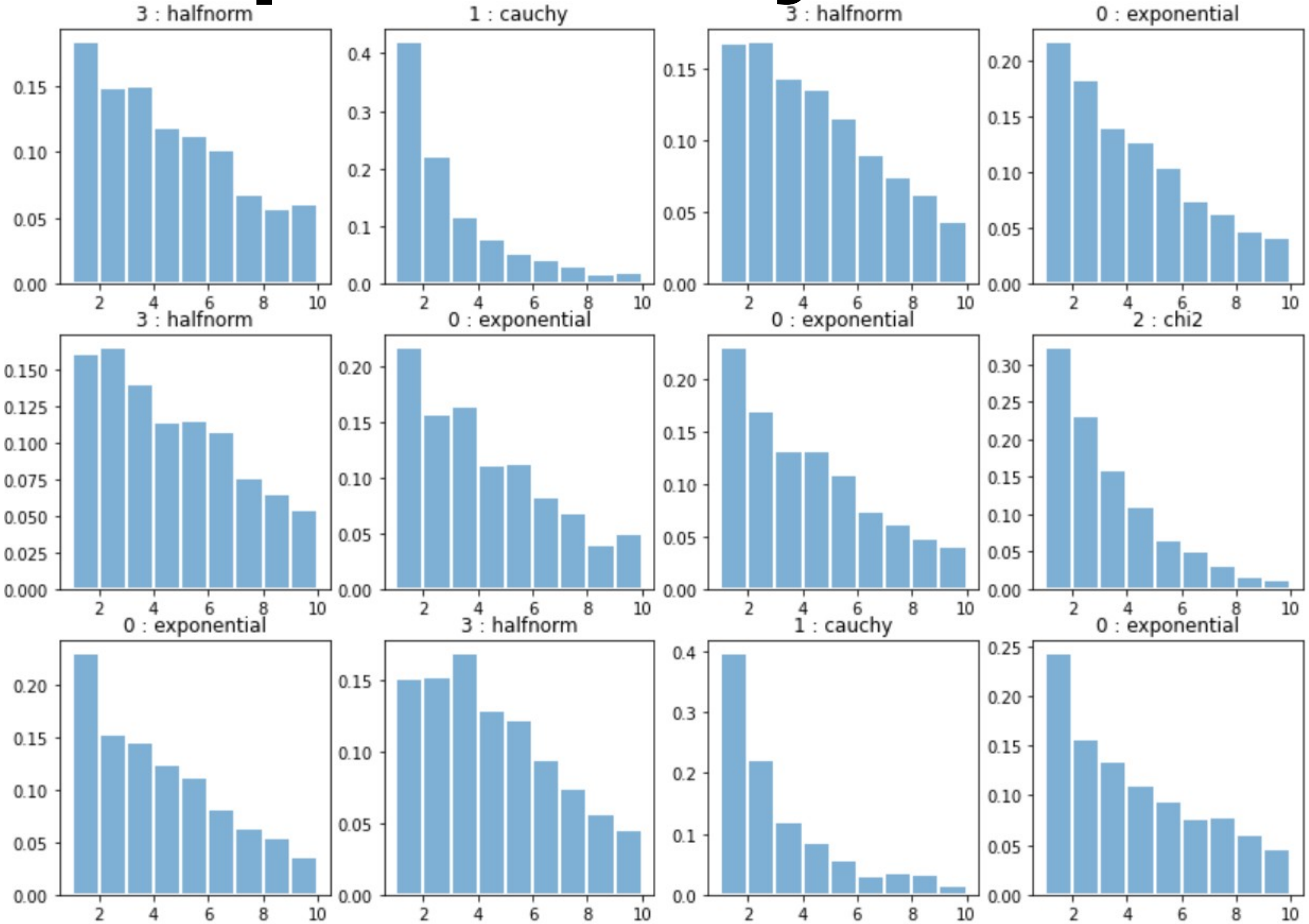
Linkage type	Adjusted Rand index score	Fowlkes-Mallows index score
complete	0.9988	0.9990
average	0.9988	0.9990
single	0.6166	0.7459

The results are similar to those obtained on the **nasty** dataset. In particular, the two scoring metrics agree that:

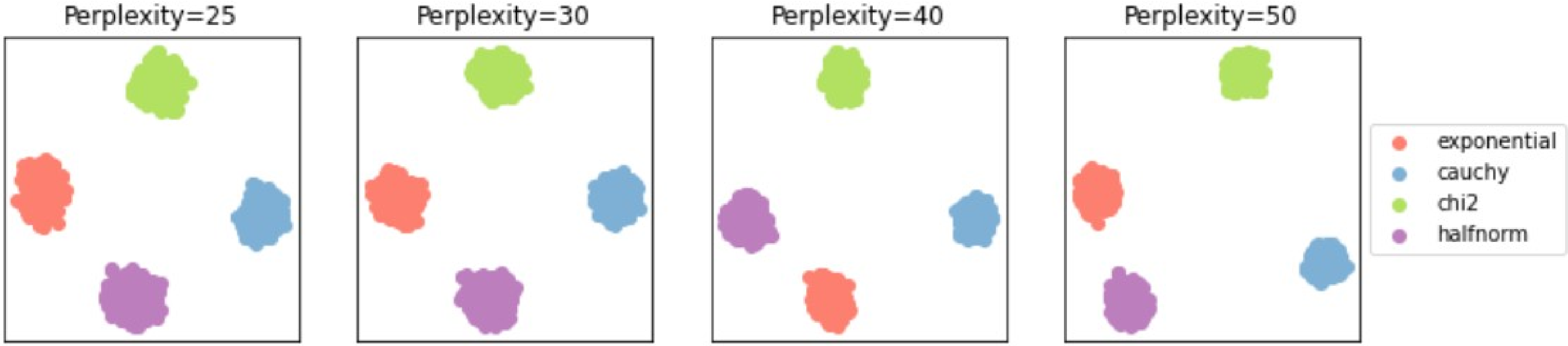
- both *complete* and *average* lead to significantly better results than those obtained with *single*
- both *complete* and *average* lead to an almost perfect clustering of the **infamous** dataset

- **Ex.5B: probability distributions**

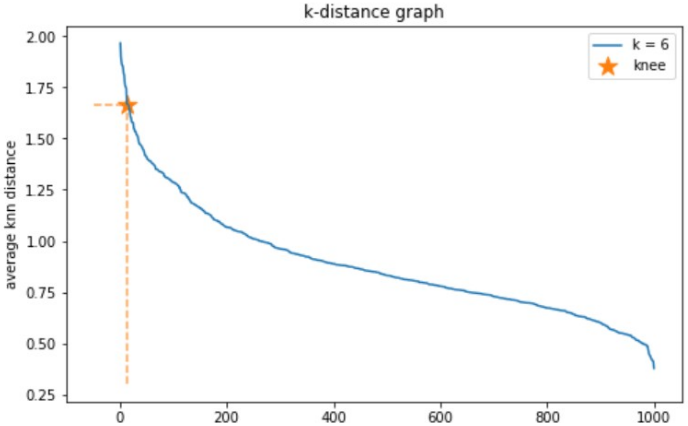
2207

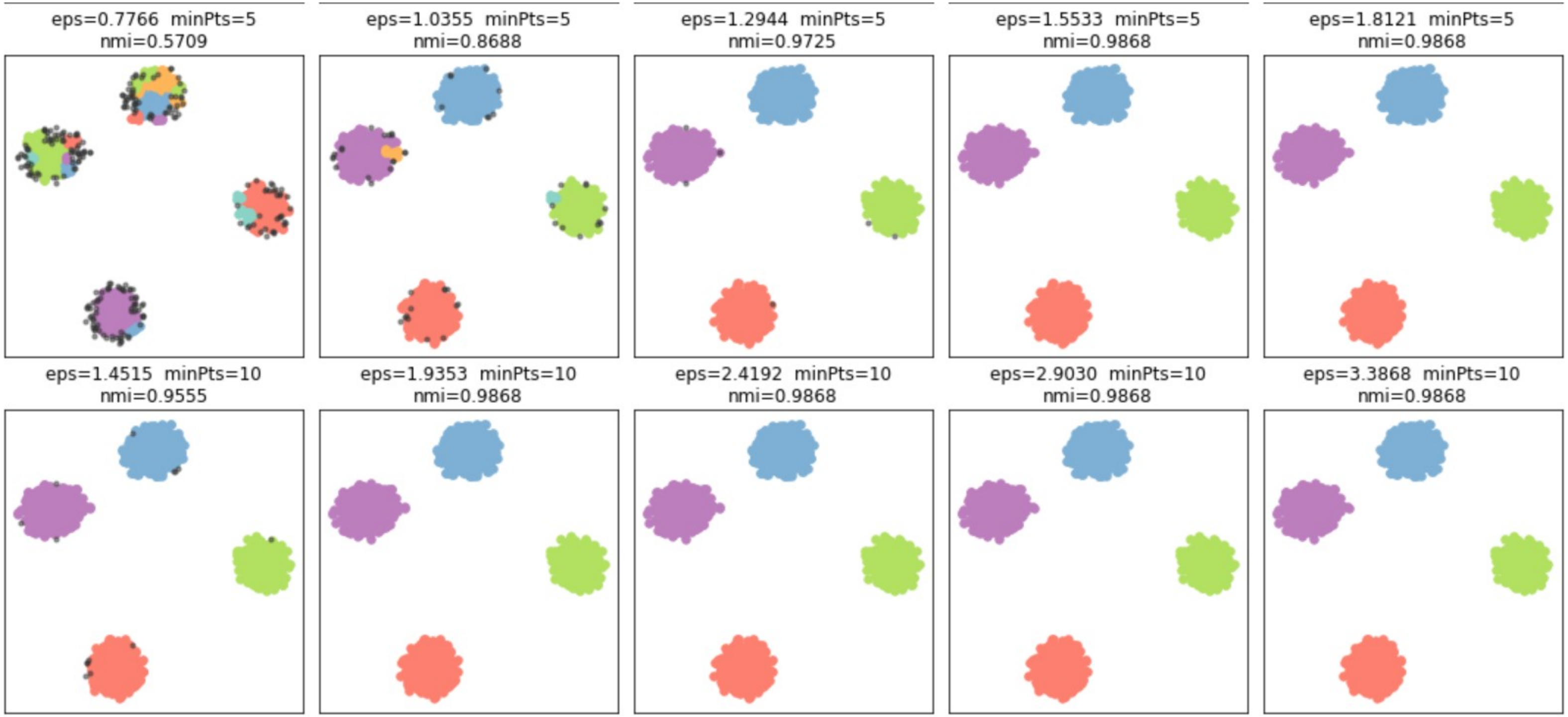


t-SNE


Perplexity=25
Perplexity=30
Perplexity=40
Perplexity=50
exponential
cauchy
chi2
halfnorm

DBSCAN





eps=0.7766 minPts=5
nmi=0.5709

eps=1.0355 minPts=5
nmi=0.8688

eps=1.2944 minPts=5
nmi=0.9725

eps=1.5533 minPts=5
nmi=0.9868

eps=1.8121 minPts=5
nmi=0.9868

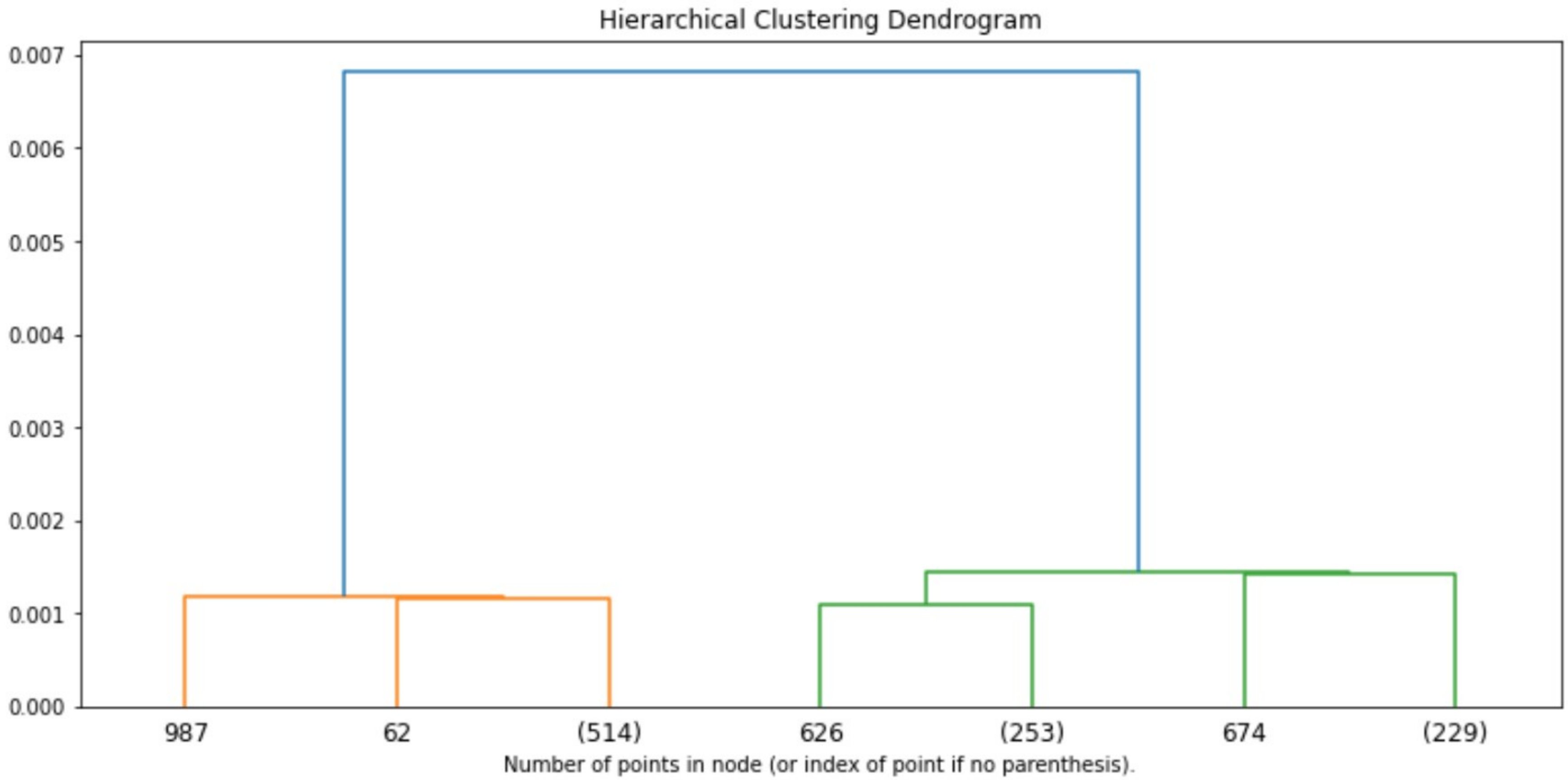
eps=1.4515 minPts=10
nmi=0.9555

eps=1.9353 minPts=10
nmi=0.9868

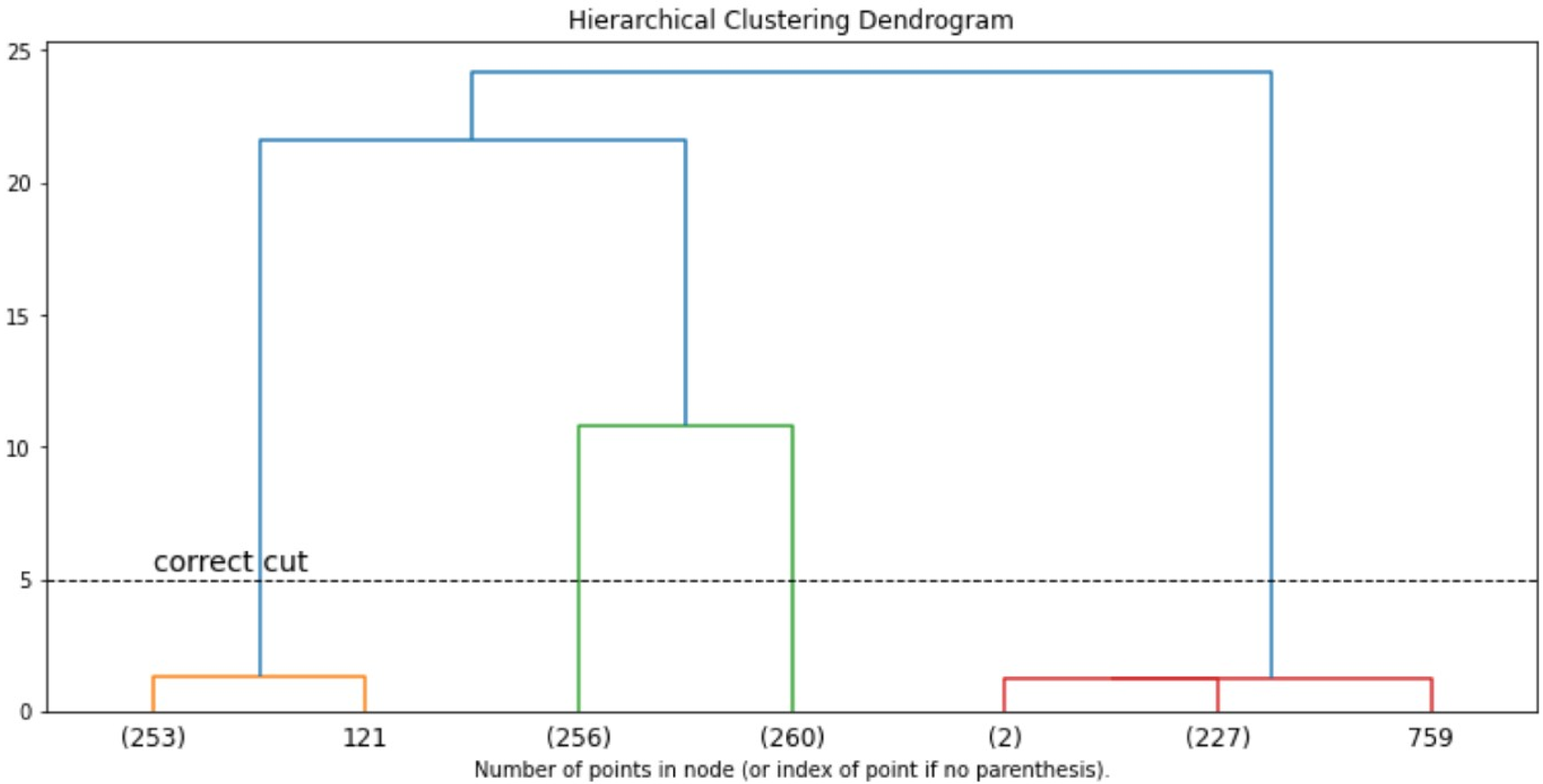
eps=2.4192 minPts=10
nmi=0.9868

eps=2.9030 minPts=10
nmi=0.9868

eps=3.3868 minPts=10
nmi=0.9868



Running the hierarchical clustering over our dataset it is evident what we have noticed from the previous analysis: there are two clusters that are difficult to distinguish with the used metric. As we can notice from the dendogram only three clusters are detected and the situation doesn't change increasing the number of showed levels. So, what we can do is again to use the projected 2D dataset from t-SNE and perfrom the hierarchical clustering using the euclidian metric.


Hierarchical Clustering Dendrogram
25
20
15
10
5
0
correct cut
(253)
121
(256)
(260)
(2)
(227)
759
Number of points in node (or index of point if no parenthesis).