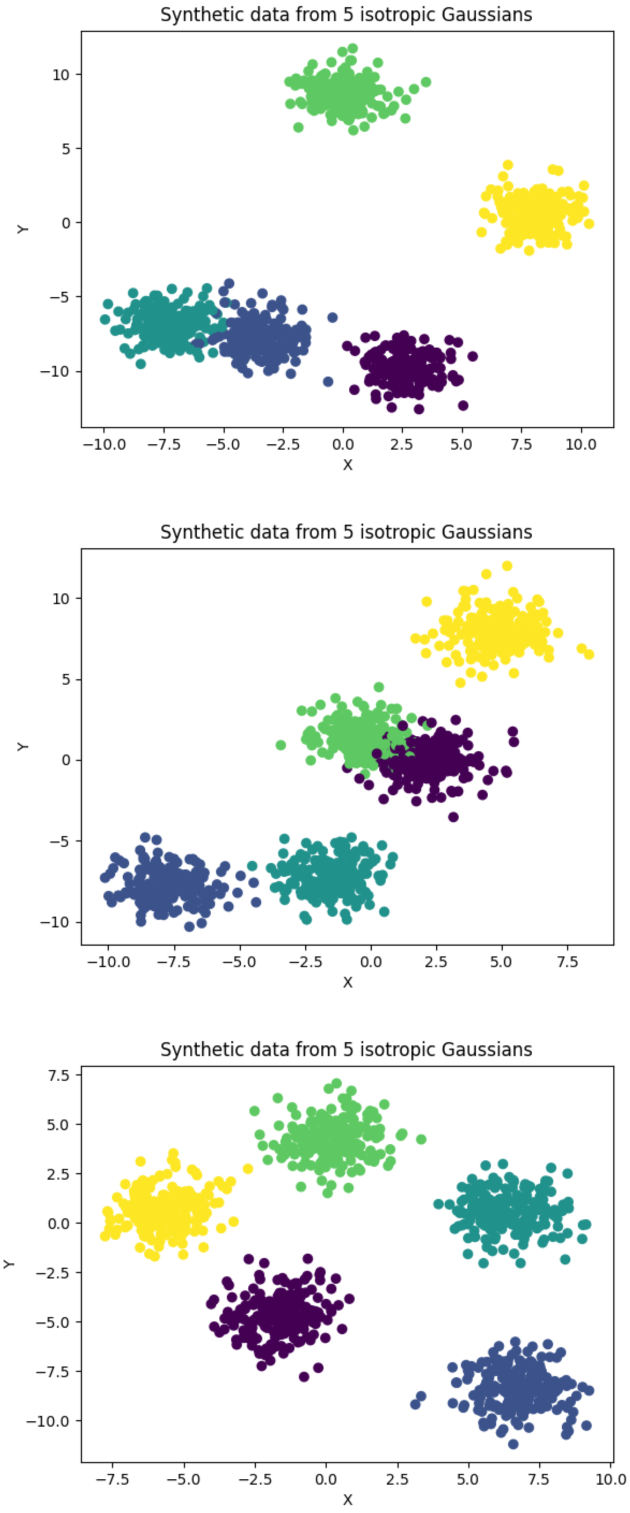
Topics in Unsupervised Learning

Amit Ezer, Tzvi Greenfeld

April 2023

Contents

1 Synthetic data tests



Synthetic data from 5 isotropic Gaussians

Synthetic data from 5 isotropic Gaussians

Synthetic data from 5 isotropic Gaussians

Figure 1: Different initialization results

1.1 K-means results (For the first initialization)

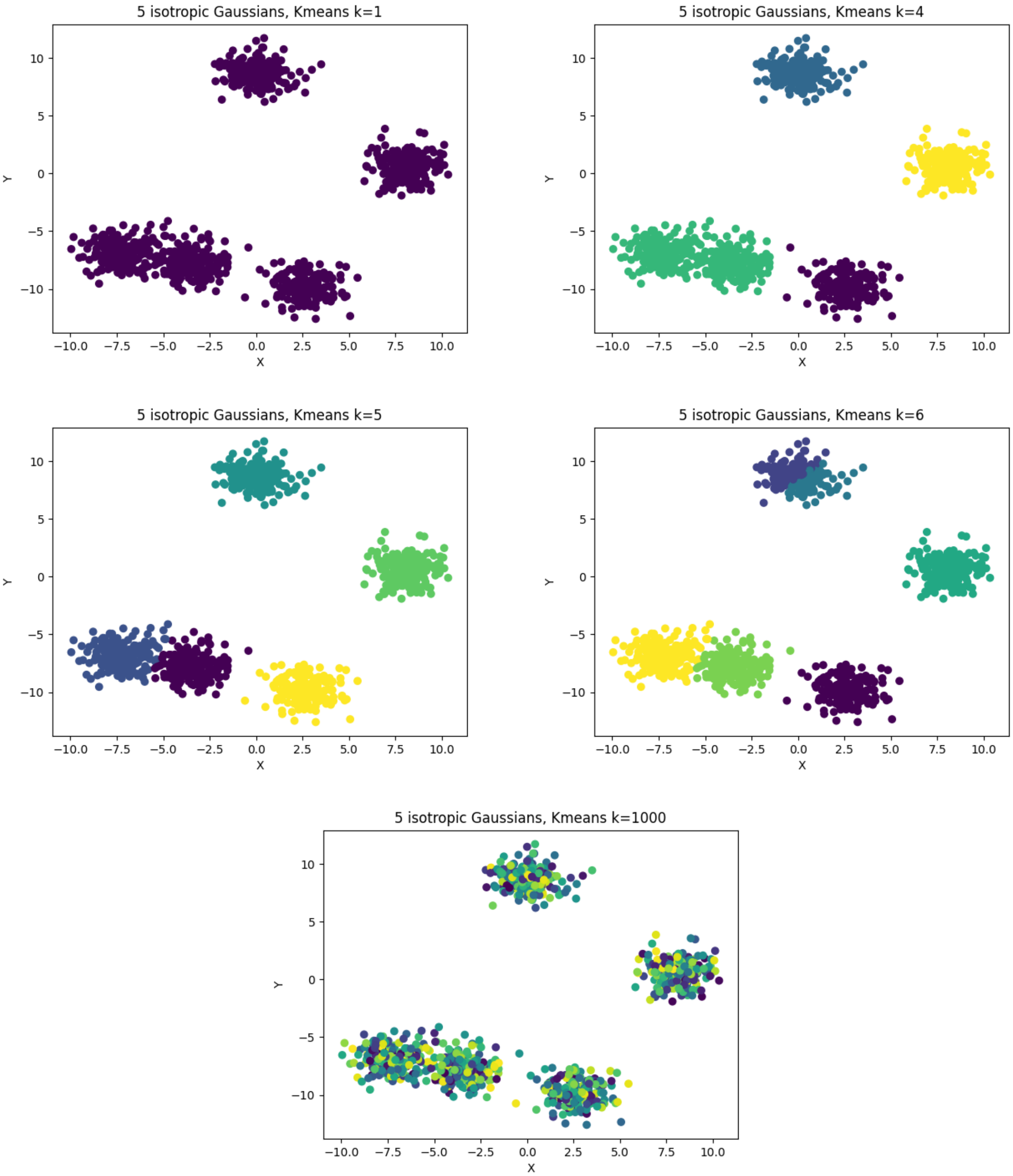
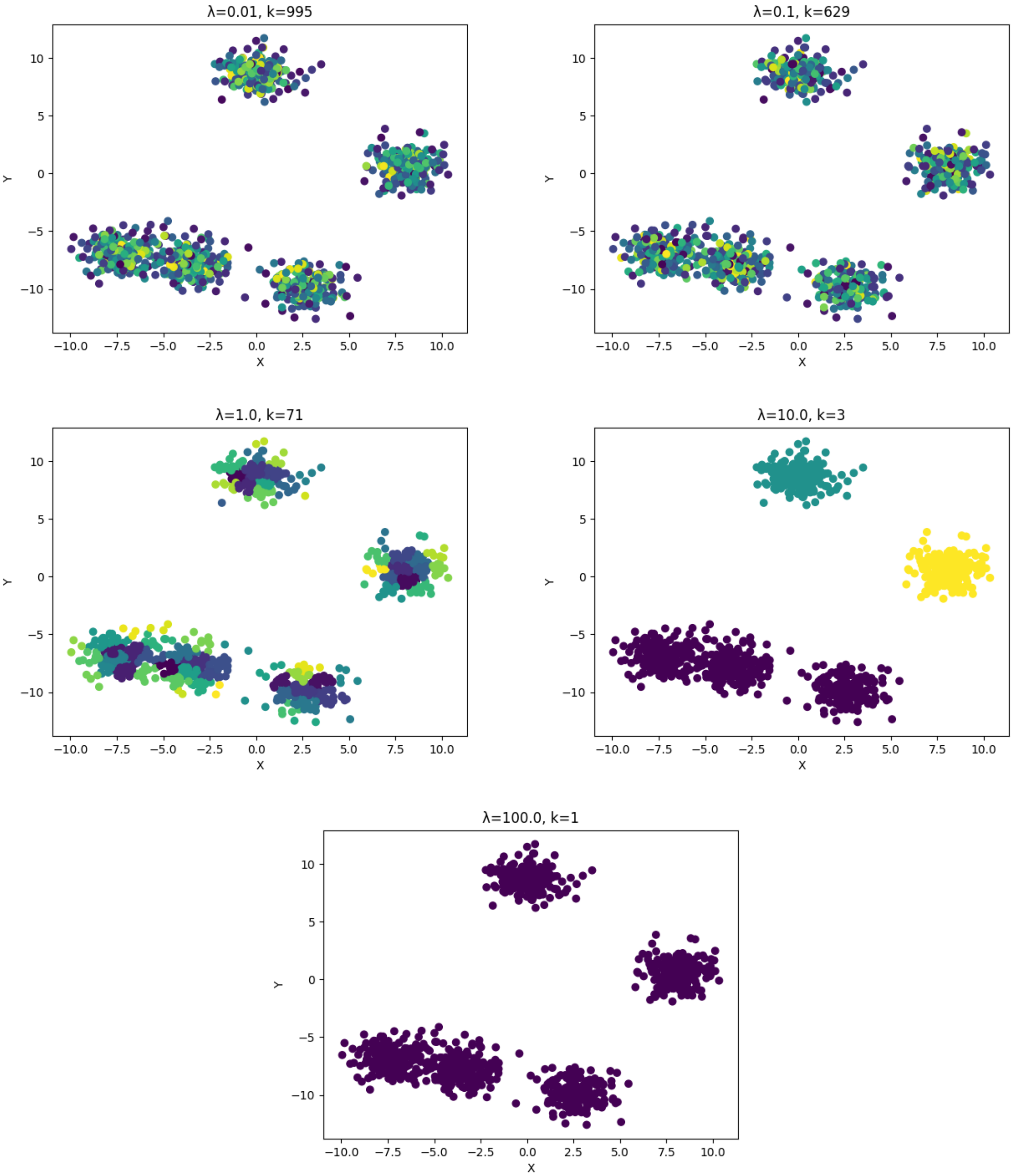


Figure 2: K-means clustering for synthetic data with different k values

As shown by the plots above, we can clearly see that as k gets closer to the number of isotropic Gaussians used to generate the data, the results are clustered better. As expected, for $k = 1$ we only got 1 cluster and for $k = n$ there are n clusters with single point in each cluster.

1.2 PDC-DP results (For the first initialization)



λ=0.01, k=995

λ=0.1, k=629

λ=1.0, k=71

λ=10.0, k=3

λ=100.0, k=1

Figure 3: PDC-DP clustering for synthetic data with different λ values

When clustering with PDC-DP-means, we can see the relation between λ and the space populated by our data. Here we have limited the data to be in $\{[-10, 10], [-10, 10]\}$ and the best results were obtained by setting $\lambda = 10$.

2 Mandrill

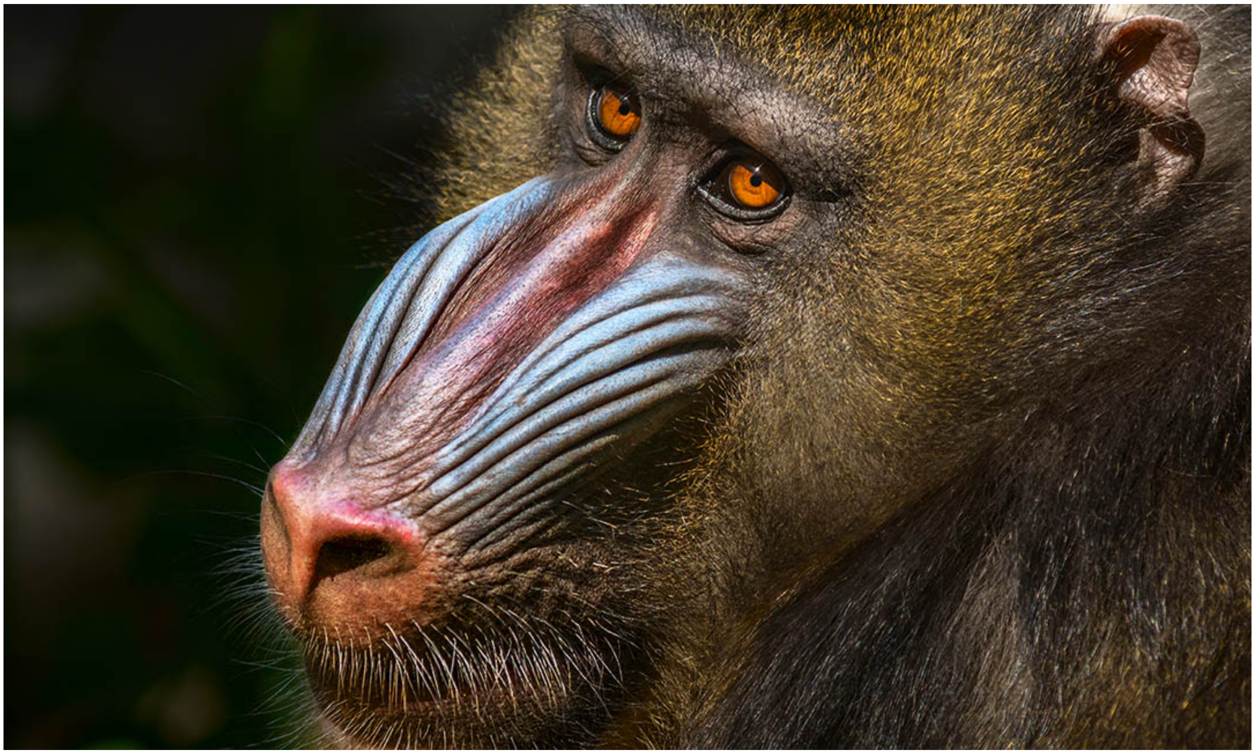
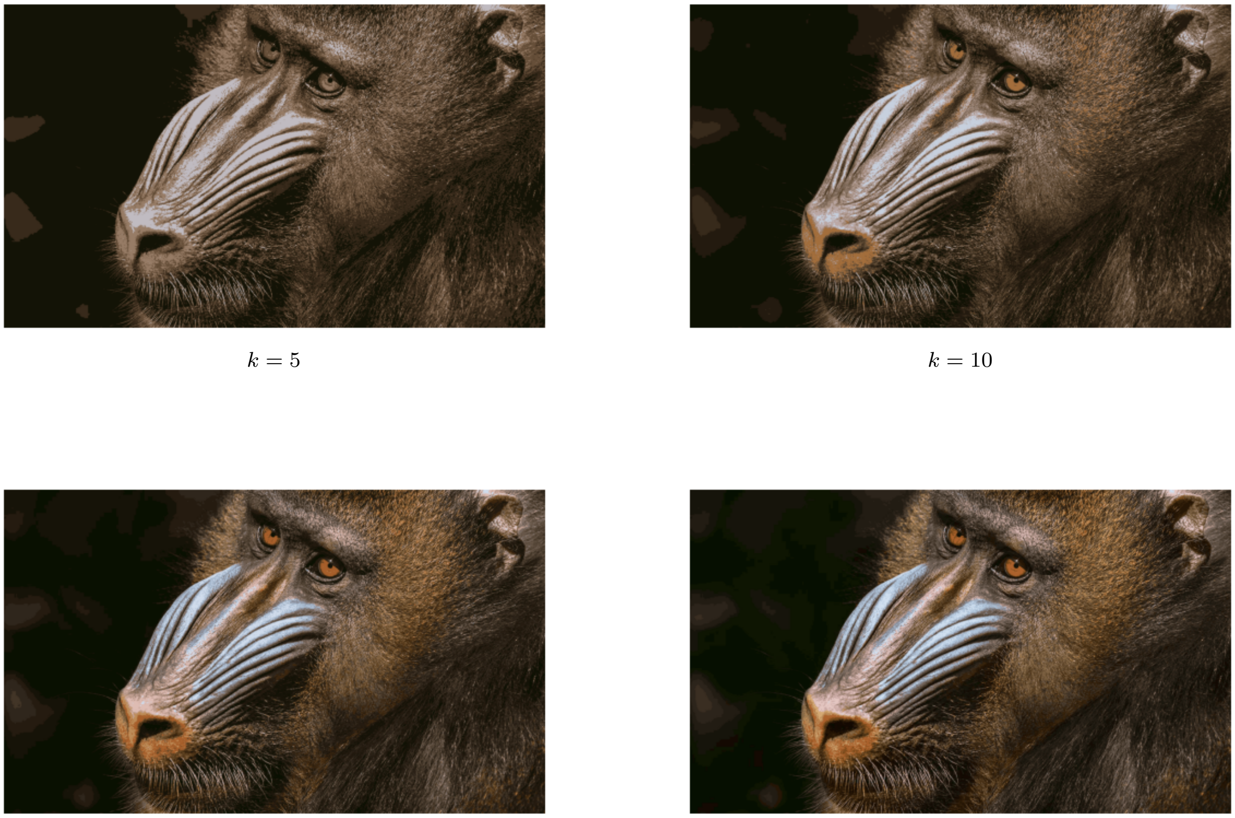


Figure 4: Original image of a mandrill

The image is 500 × 800 pixels, therefore the data is $n = 40,000, data \subset R^3$

2.1 K-means



$k = 5$

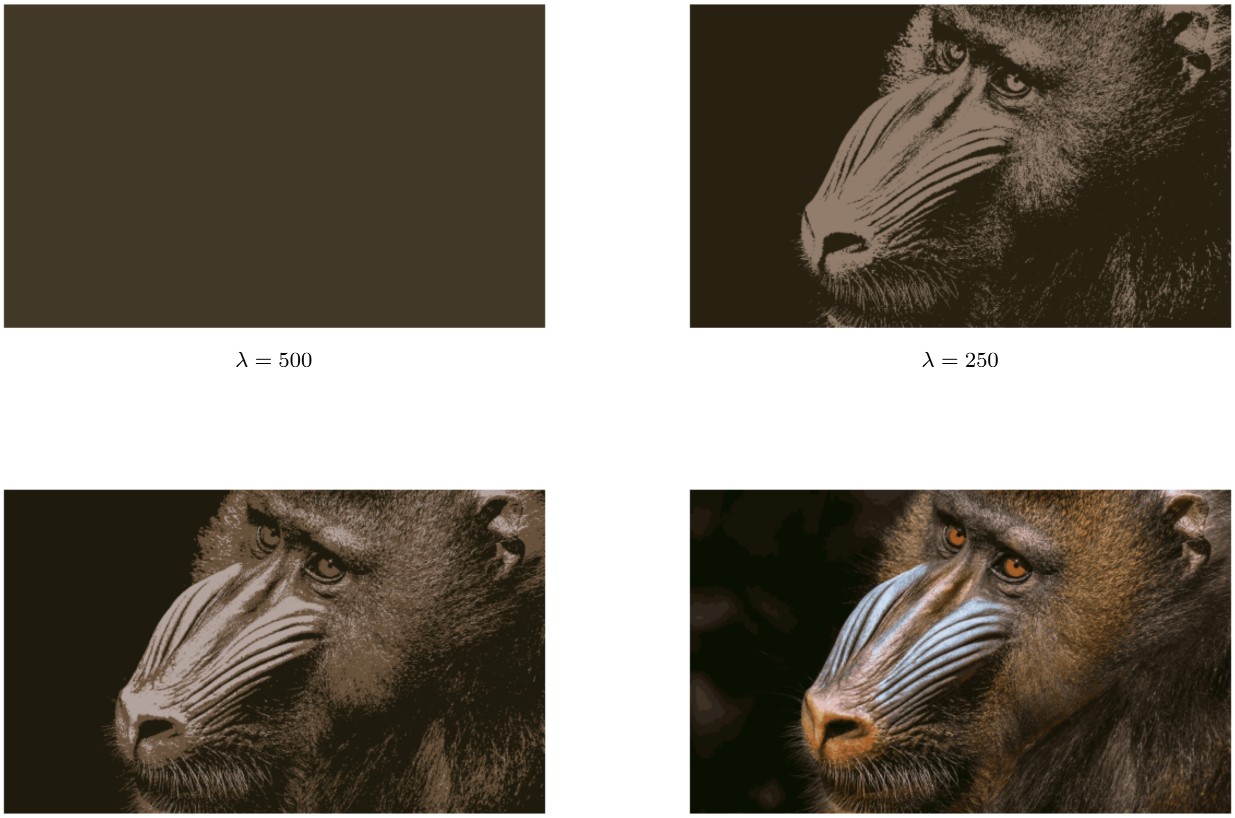
$k = 10$

$k = 20$

$k = 30$

Figure 5: K-means RGB clustering results

2.2 PDC-DP



$\lambda = 500$

$\lambda = 250$

$\lambda = 180$

$\lambda = 100$

Figure 6: PDC-DP RGB clustering results

As λ gets smaller, the number of clusters grows. For $\lambda = 500$ we got 3 clusters, and for $\lambda = 100$ we got 64 clusters, leading to the output image being closer to the input data.

3 Final project statement

We chose to Extend the (P)DC-DP-means Algorithm to Streaming data. (Project #1)