Bo Zhang bzhang73@binghamton.edu B-00766142

CS 535
Fall 2019
Assignment 1
Assigned on 9 September 2019
Due on 13 October 2019

Total Points: 150

This assignment is for both grad and undergrad students. The total number of points is 150. Below Problems 1 to 3 are for both grad and undergrad students while Problem 4 is for grad students only. An undergrad student who completes the problem for grad students gets extra credit.

You can use whatever programming language you are comfortable with.

This assignment is about statistical analysis of a data collection as well as different data reduction methods, and in particular, dimensionality reduction through feature extraction. You are given two datasets, each containing a data table of 1000 vector with 100 attributes (i.e., dimensions) in two files with 500 samples for each file. Each dataset is given by two tables of 500 samples each. Both datasets are given as text table files where each dataset is represented as a 1000 x 100 matrix where each row of the matrix is a vector. You are further told that for each dataset, for all the samples (i.e.,A vectors) the component values of each vector follow the same distribution. For all the datasets, the only possible distributions are either Gaussian or uniform.

1. (20 pts.) Determine the distributions of the vector component values for both datasets. For each dataset, randomly pick up 10 samples and report the distribution parameters for each of the 10 samples.

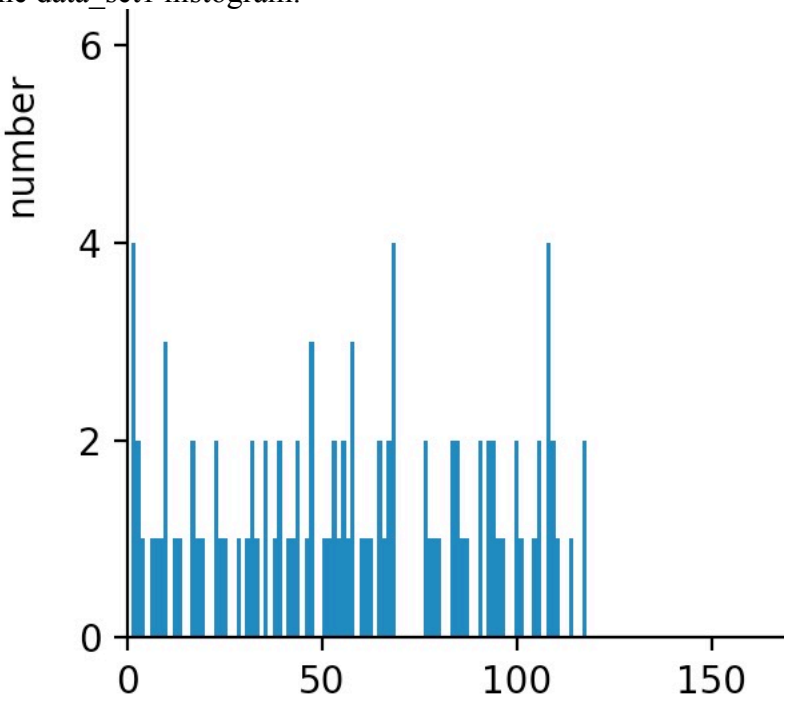
2. (50 pts.) Implement PCA and DCT methods and apply them for feature extraction to the two datasets, respectively. Report the principle you have proposed to truncate the dimensionality and the reduced dimensionalities for the two datasets after the feature extraction for PCA and DCT, respectively.

3. (30 pts.) Compare the feature extraction results between the two methods for the two datasets, respectively, and report your comparison conclusion.

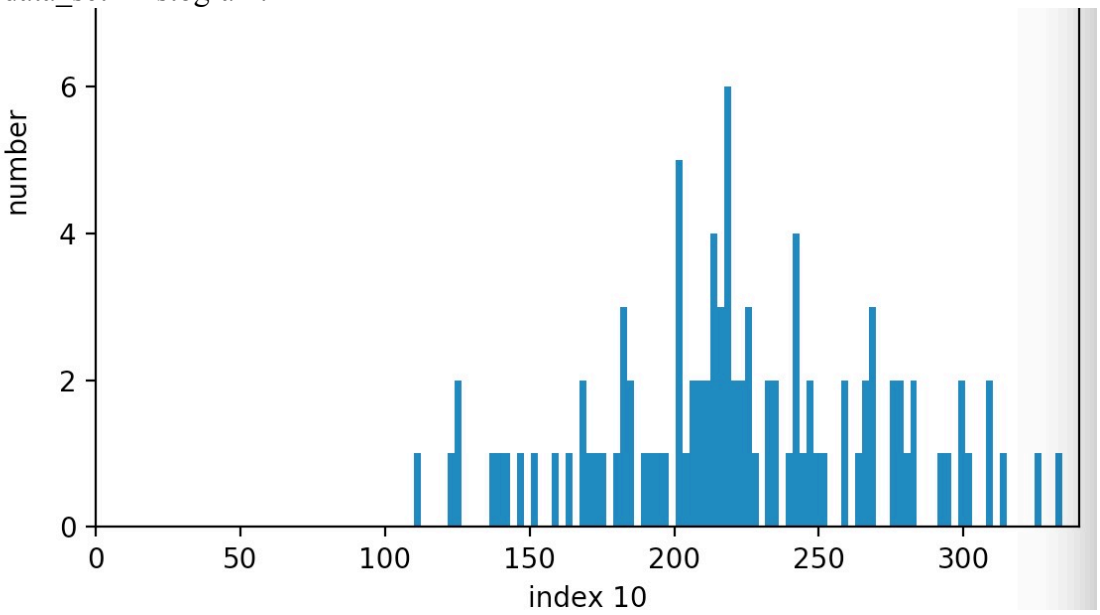
4. (50 pts.) Read the literature on Independent Component Analysis (ICA) and implement ICA. Then apply ICA to the two datasets, respectively. Report your comparison studies on the two datasets between PCA and ICA on feature extraction.

Bo Zhang bzhang73@binghamton.edu B-00766142

1.Answer: The data_set1 is uniform distributions. The data_set2 is Gaussian distributions.
The data_set1 histogram:



The data_set2 histogram:



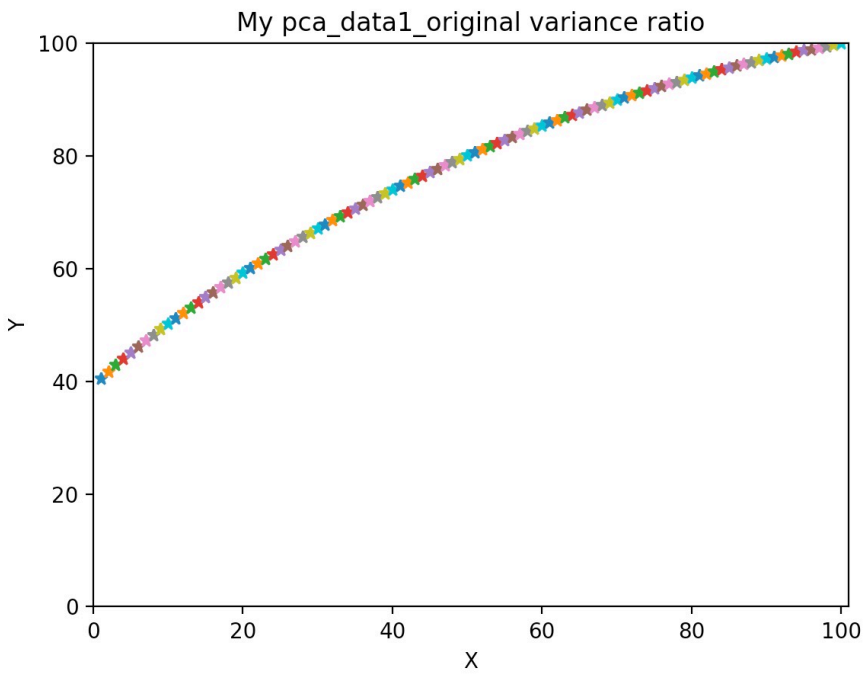
And using KSTEST to test the two sample:

```
The Data1 average kstest of Gaussian : 0.319010
The Data2 average kstest of Gaussian : 0.722464
The Data1 set is uniform distribution
The Data2 set is Gaussian distribution
```

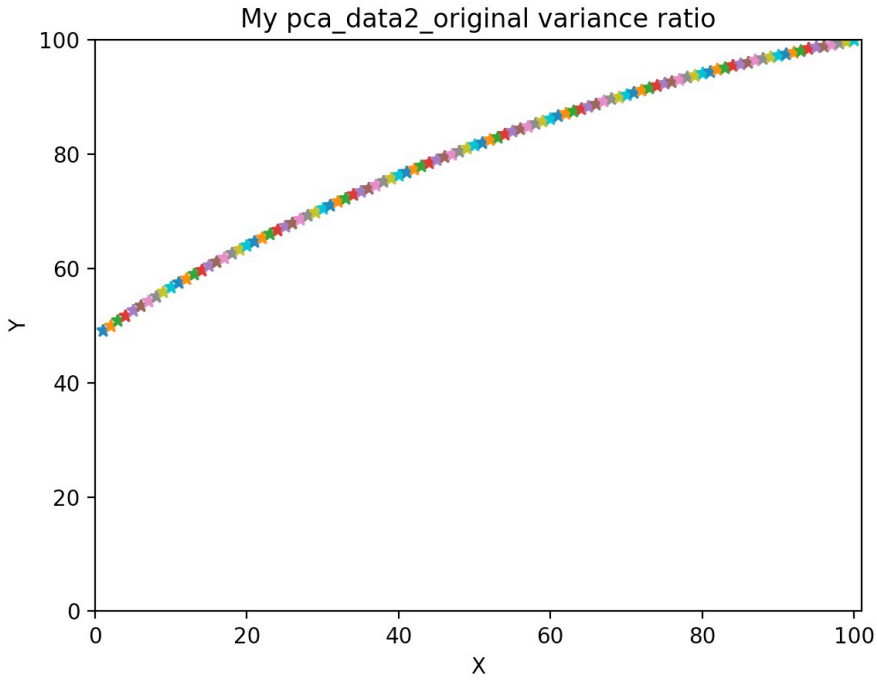
Bo Zhang bzhang73@binghamton.edu B-00766142

2.Answer:

Data_set1 PCA After compute the PCA variance ratio, sum of the k largest variance ratio:



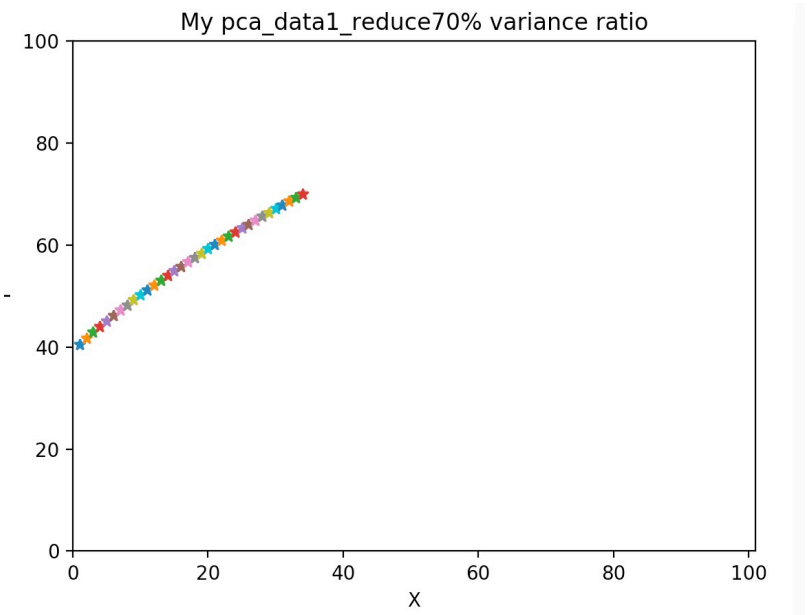
Data_set1 PCA After compute the PCA variance ratio, sum of the k largest variance ratio:



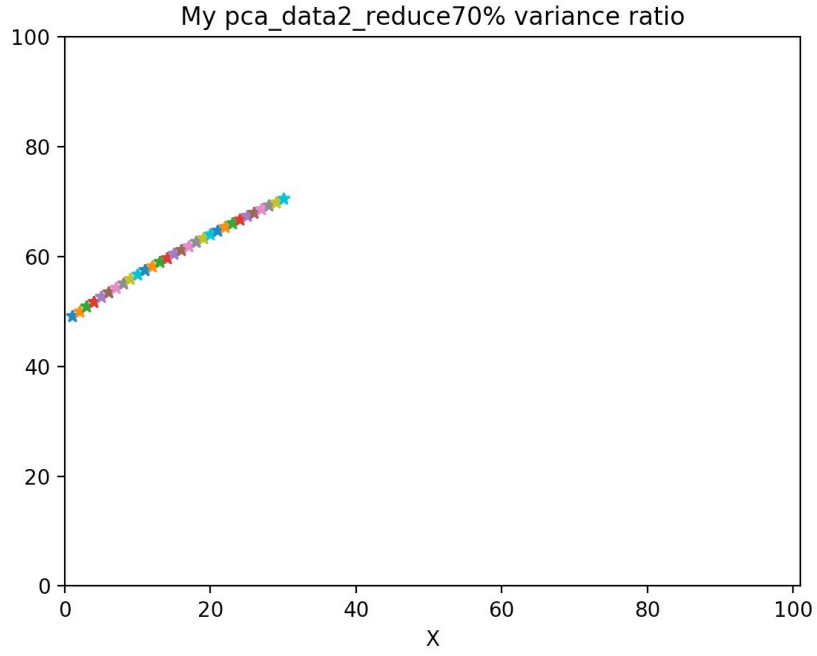
Bo Zhang bzhang73@binghamton.edu B-00766142

Truncate the dimensionality and the reduced dimensionalities for the two datasets after the feature extraction for PCA. My principle is to reduce the sum largest ratio to 70, when it comes to 70, remove the rest features. After transfer:

The data_set1 feature length is 34 and the result is:

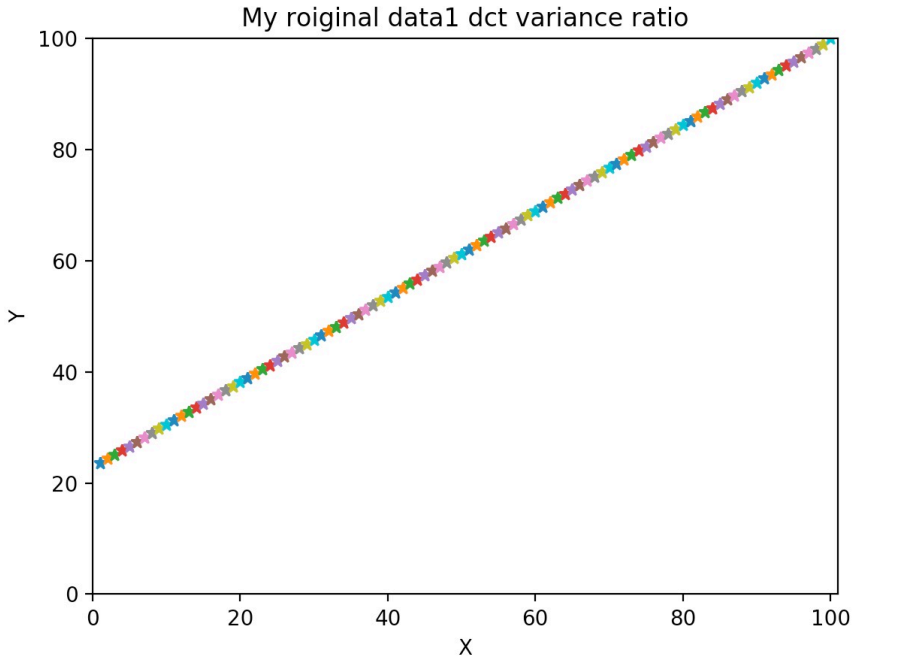


The data_set2 feature length is 30 and the result is:

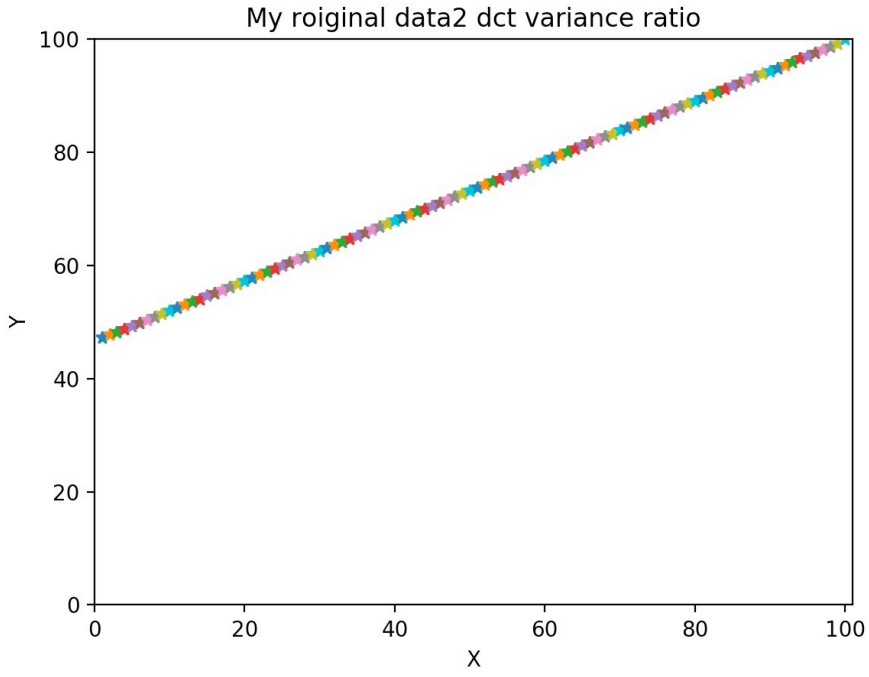


Bo Zhang bzhang73@binghamton.edu B-00766142

Data_set1 DCT: After using DCT function, calculate the average variance ratio of each feature, sum of the k largest variance ratio:



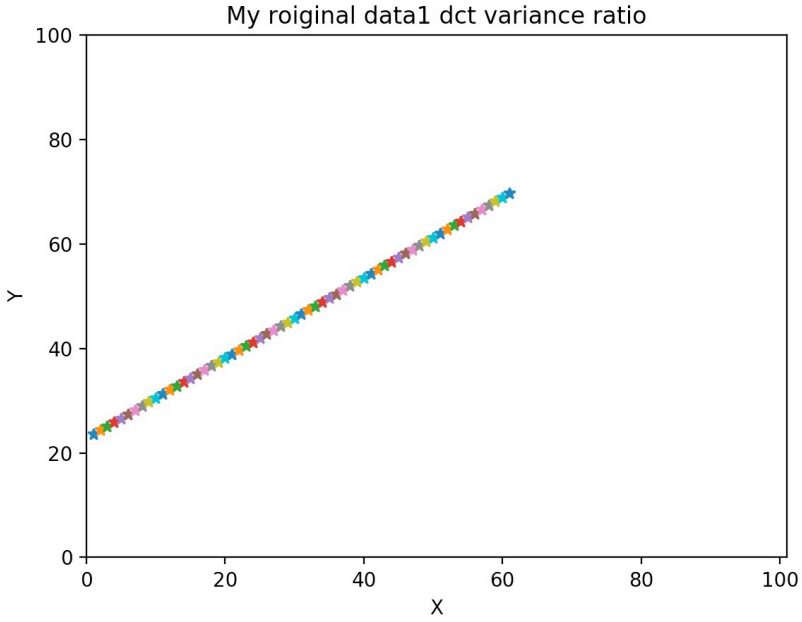
Data_set2 DCT: After using DCT function, calculate the average variance ratio of each feature, sum of the k largest variance ratio:



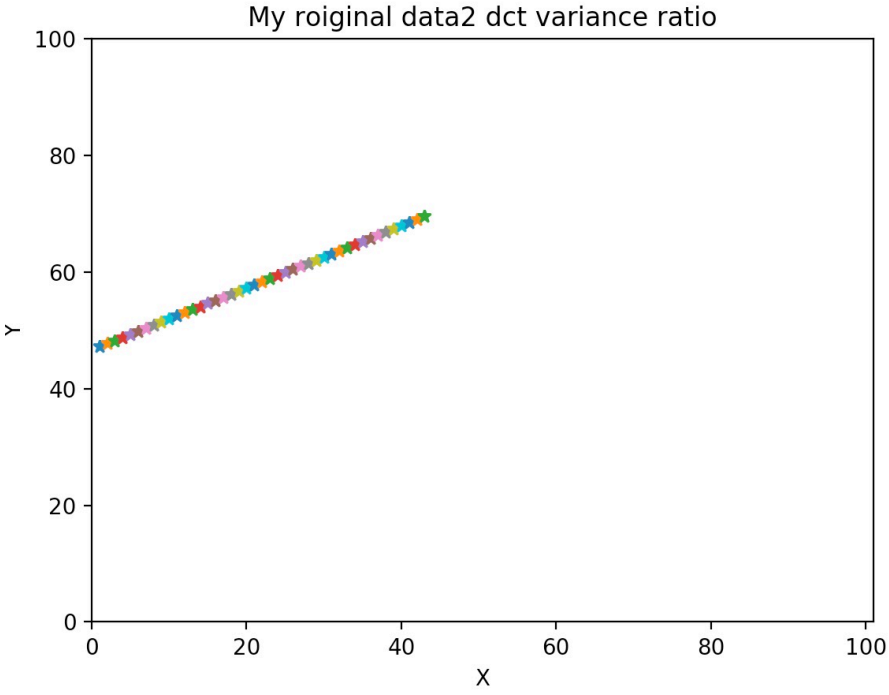
Bo Zhang bzhang73@binghamton.edu B-00766142

Truncate the dimensionality and the reduced dimensionalities for the two datasets after the feature extraction for DCT. My principle is to reduce the sum largest ratio to 70, when it comes to 70, remove the rest features. After transfer:

The data_set1 feature length is 61 and the result is:

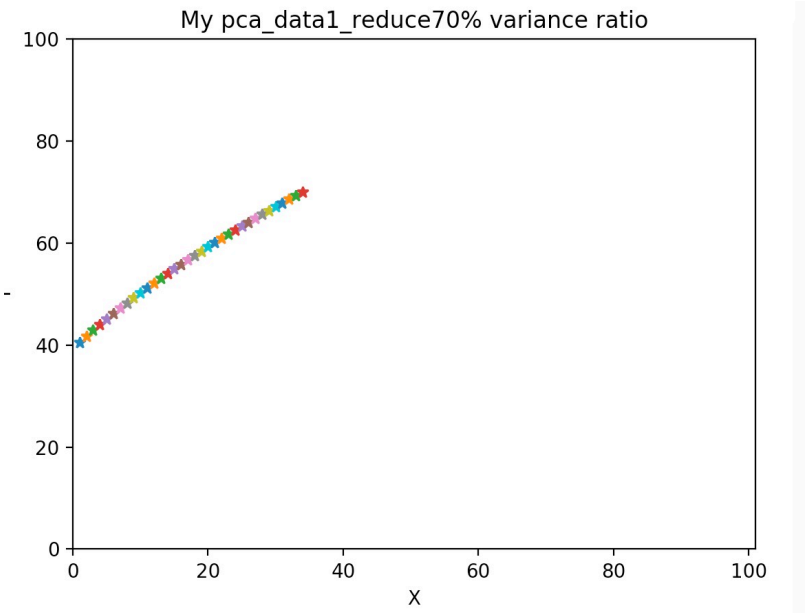


The data_set2 feature length is 43 and the result is:

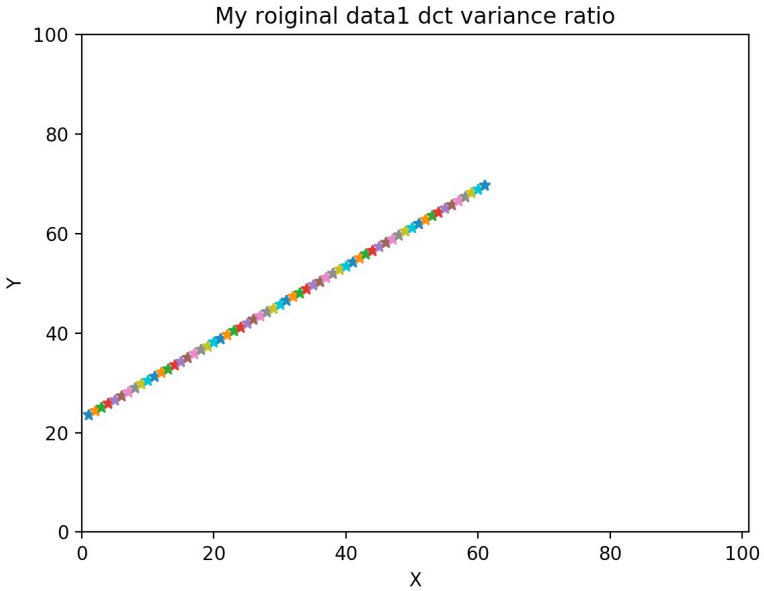


3.Answer: Compare the feature extraction results between the two methods for the two datasets, respectively:((PCA compare with DCT)

For data_Set1:
The data_set1 feature length is 34 and the result is:

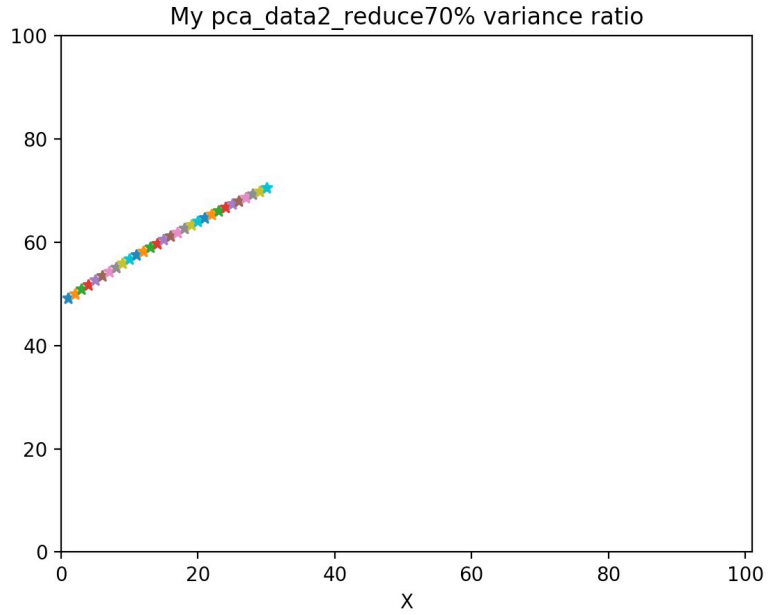


The data_set1 feature length is 61 and the result is:

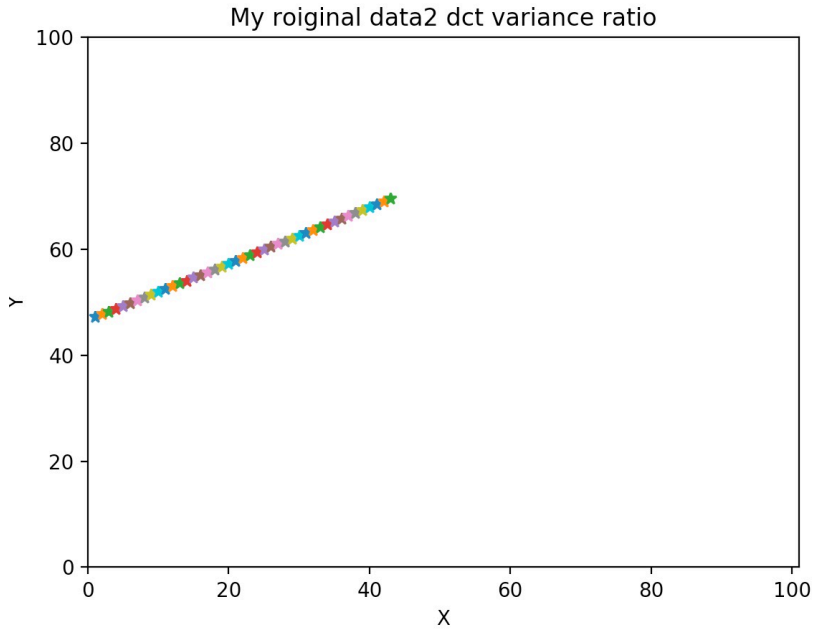


In conclusion of data_set1, the PCA is more useful when both methods are used method reduce to 70%, but PCA have smaller size of feature, which means the PCA can calculate more information feature than DCT method. And for PCA the first feature has more information than DCT first feature.

For data_Set2:
The data_set2 feature length is 30 and the result is:



The data_set2 feature length is 43 and the result is:

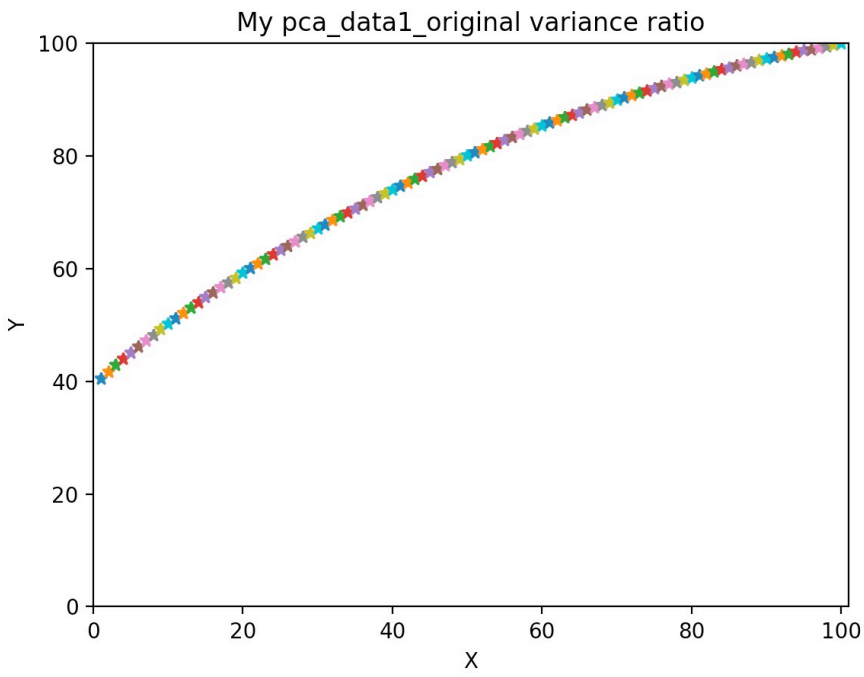


In conclusion of data_set2, the PCA is more useful when both methods are used method reduce to 70%, but PCA have smaller size of feature, which means the PCA can calculate more information feature than DCT method. And for PCA the first feature has same information as DCT first feature, but the rest features contain less information.

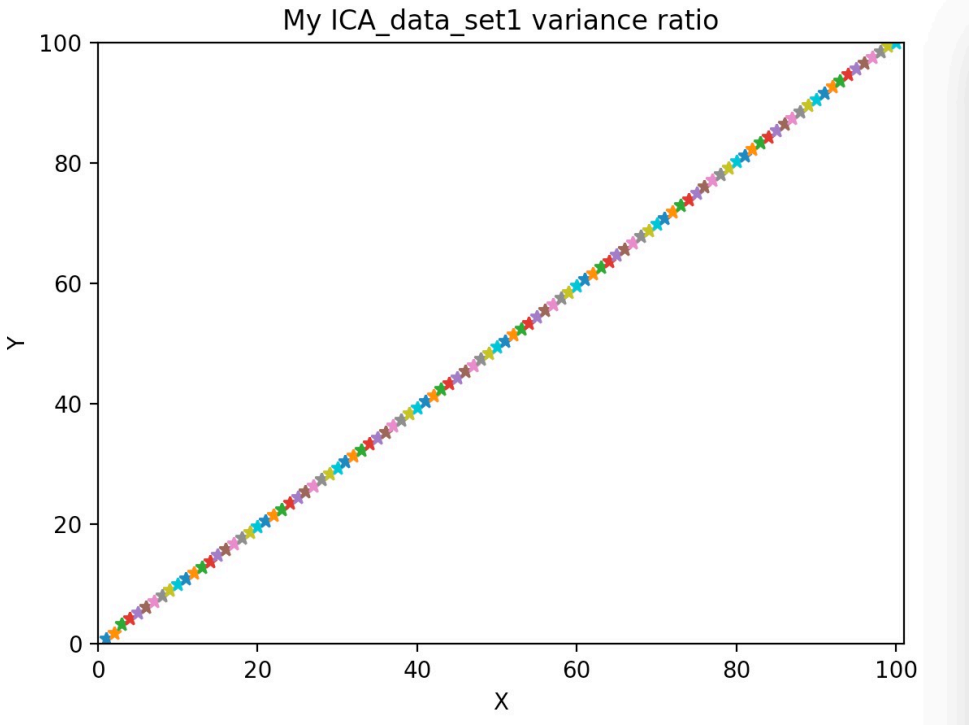
Therefore, the PCA is better than DCT method in the way to reduce same percentage, PCA has less feature, which means each feature contain more information than DCT.

4. Answer:

Data_set1 PCA After compute the PCA variance ratio, sum of the k largest variance ratio:



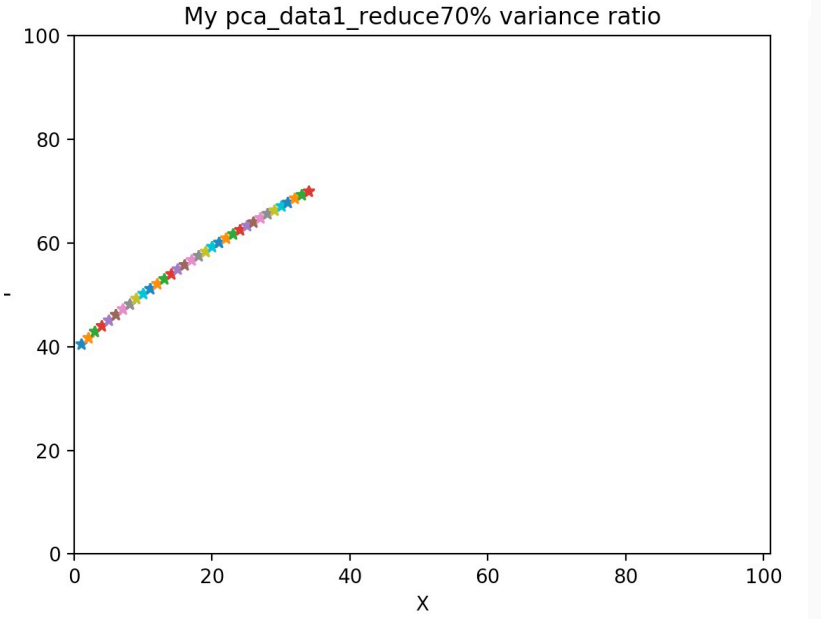
Data_set1 ICA After compute the ICA variance ratio, sum of the k largest variance ratio:



Bo Zhang bzhang73@binghamton.edu B-00766142

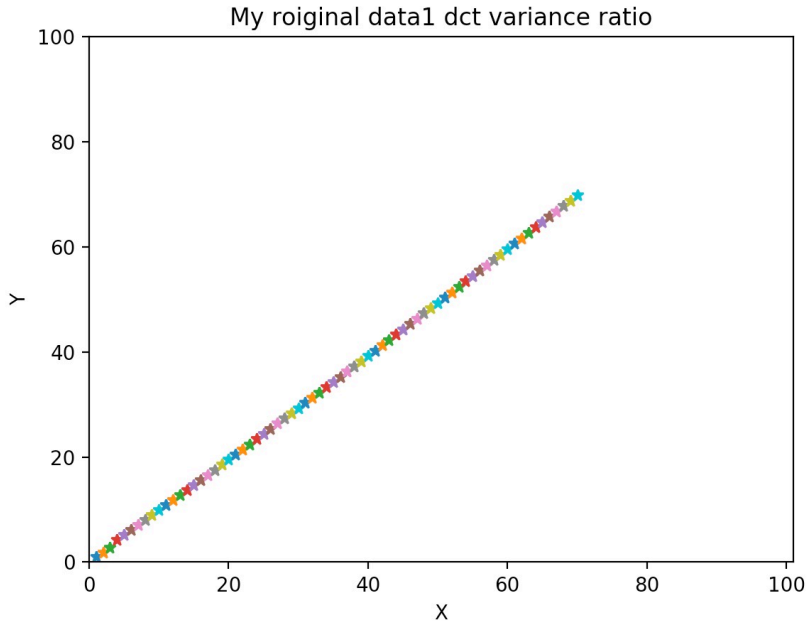
Truncate the dimensionality and the reduced dimensionalities for the two datasets after the feature extraction for PCA. My principle is to reduce the sum largest ratio to 70, when it comes to 70, remove the rest features. After transfer:

The data_set1 feature length is 34 and the result is:

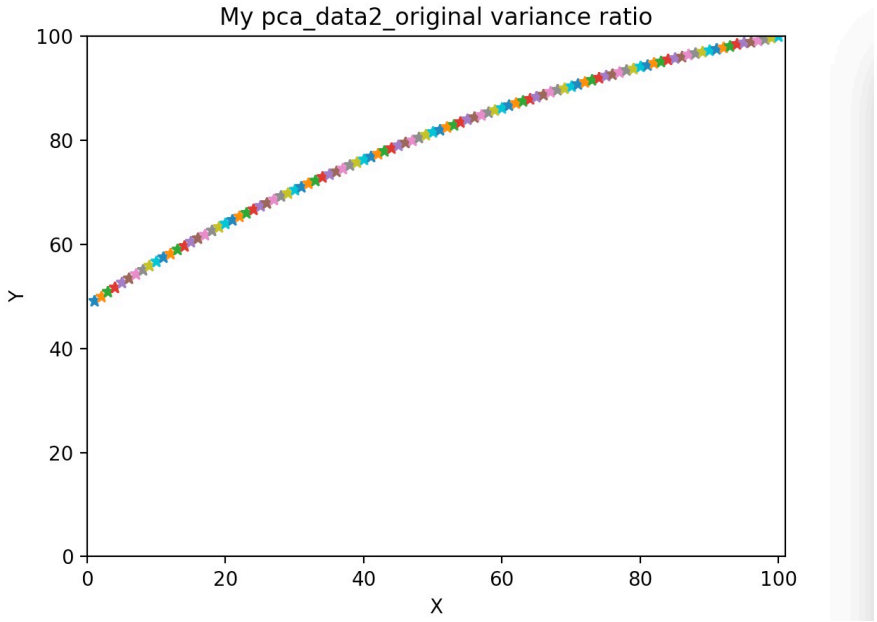


Truncate the dimensionality and the reduced dimensionalities for the two datasets after the feature extraction for ICA. My principle is to reduce the sum largest ratio to 70, when it comes to 70, remove the rest features. After transfer:

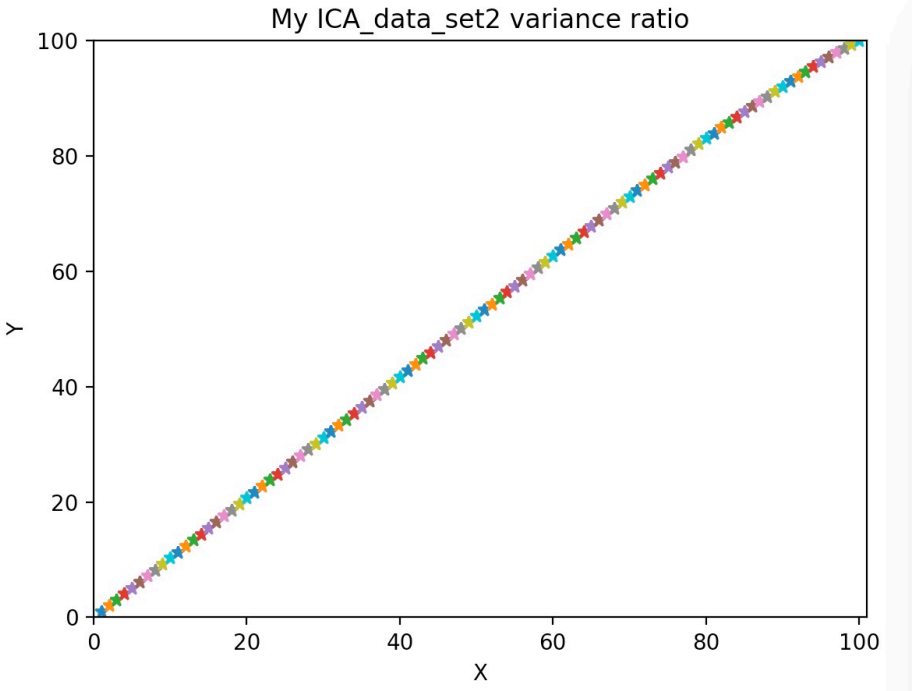
The data_set1 feature length is 70 and the result is:



Data_set2 PCA: After using PCA function, calculate the average variance ratio of each feature, sum of the k largest variance ratio:

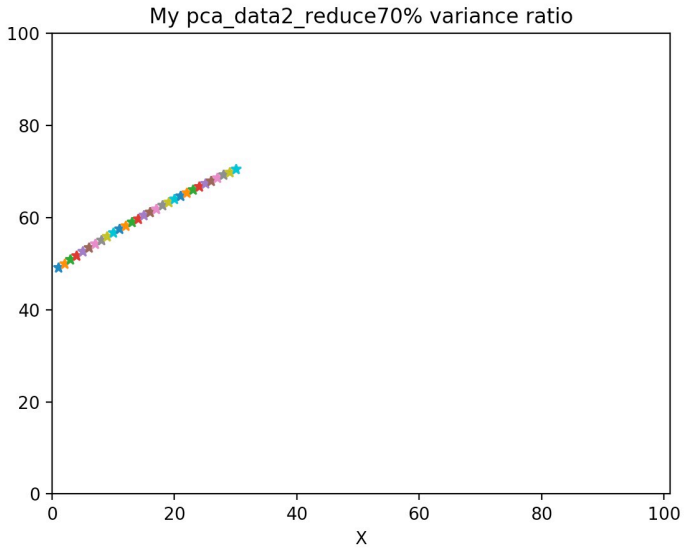


Data_set2 ICA: After using ICA function, calculate the average variance ratio of each feature, sum of the k largest variance ratio:

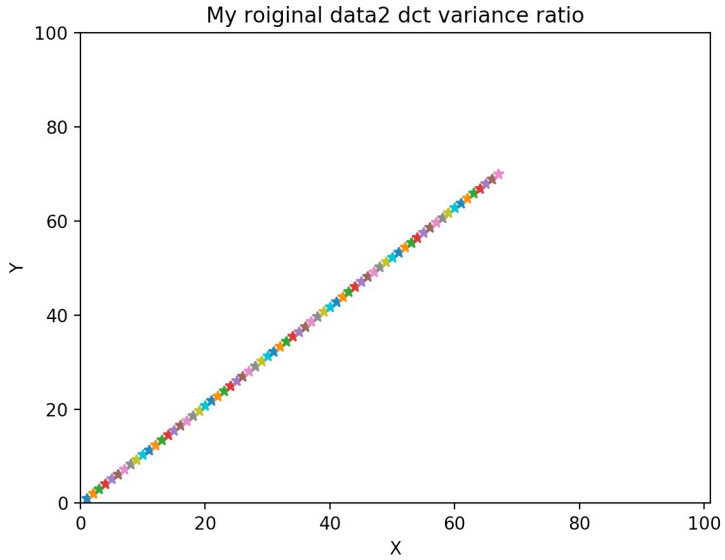


Truncate the dimensionality and the reduced dimensionalities for the two datasets after the feature extraction for PCA. My principle is to reduce the sum largest ratio to 70, when it comes to 70, remove the rest features. After transfer:

The data_set2 feature length is 30 and the result is:



The data_set2 feature length is 67 and the result is:



In conclusion of data_set1 and data_set2, the PCA is more useful when both methods are used method reduce to 70%, but PCA have smaller size of feature, which means the PCA can calculate more information feature than ICA method. And for PCA the first feature has larger information as ICA first feature, but the rest features contain less information.

Therefore, the PCA is better than ICA method in the way to reduce same percentage, PCA has less features, which means each feature contain more information than ICA.