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., 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. 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. 3. (30 pts.) Compare the feature extraction results between the two methods for the two datasets, respectively, and report your comparison conclusion.
4. 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.