Evaluating Clustering Techniques on Non-elliptical Data

Safiya Sirota, Yijin Wang, Bin Yang

1 Background and Objective

Latent Class Analysis and K-means are two successful clustering algorithms that partition the given data into distinct subgroups where observations in each group are very similar. They are popular choice of method in solving medical research problems such as x,y,z. One important assumption that both of these algorithms require is the normality of the given data.

In our project, we focuse on the unsupervised learning scenario and we investigate how well these two algorithms perform when the normality assumption is violated via simulations. We designed 4 simulation settings as violation of the normality assumption : 1) skewed data, 2) data with heavy tail, 3) data with outliers, 4) multimodal data. In the end, we compare their performance with Rand Index in each settings with varying sample sizes.

2 Statistical Method

2.1 K Means

2.2 Latent Class Analysis

3 Performance Metric

4 Simulations Settings and Results

In simulations, we consider bivariate data with 2 true clusters. We generated data from different non-elliptical distributions in each setting.

For each simulation setting, we create 100 runs using randomly generated data sets in size of 500, 1000, 5000 and evaluate the Rand index of each method.

4.1 Skewed Data - Data Generation

In this setting, we want to test the performance of K-means and LCA when all features follow a skewed distribution. We assume the features X_1, X_2 are independent.

All data in the first feature $X_{11}, X_{12}, ..X_{1N}$ are random draws from a mixture model of Weibull distributions with density $0.5f_1(x|1,5) + 0.5f_2(x|12,14)$ where $f_1 \sim Weibull(1,5)$ and $f_2 \sim Weibull(12,14)$.

Similarly, all data in the second feature $X_{21}, X_{22}, ..X_{2N}$ are also random draws from another mixture model of Weilbull distributions with density $0.5f_1(x|1,3) + 0.5f_2(x|1,4)$ where $f_1 \sim Weibull(1,3)$ and $f_2 \sim Weibull(1,4)$.

For each feature, we generated samples with total size being 500, 1000 and 5000. An example of a sample with size 5000 is shown in Fig.x

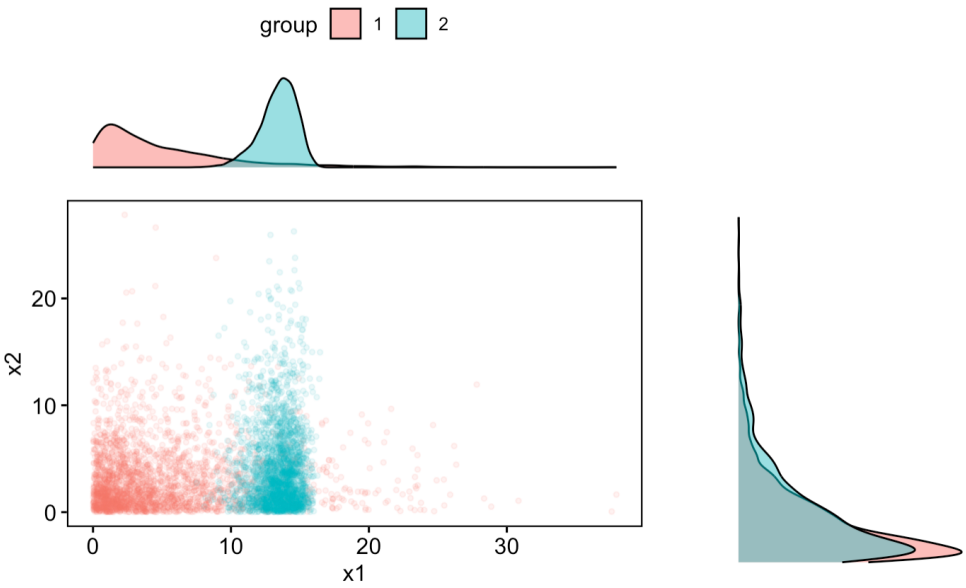

Figure 1: Fig.x skewed sample

4.2 Skewed Data - Results

5 Conclusion and Discussion