A Simulation Study to Compare Two Clustering Methods with Non-normal Data

Tianshu Liu; Jiayi Shi; Jiong Ma

Contents

Objective

Clustering is a powerful unsupervised learning approach to discover intrinsic groups in an unstructured dataset. Many clustering methods have been proposed in the literature. Among them, k-Means and Latent Class Analysis (LCA) are the two best-known methods applied widely in medical applications. Both methods are proven successful and well-tested when the normally or elliptically distributed data. In real-world applications, data often exhibit non-normal features, including asymmetry/skewed, multimodality, heavy-tails, and the presence of outliers. The objective of this simulation study is to comprehensively assess and compare the performance of K-means and LCA when the data are non-normal.

Statistical methods

K-Means

K-means is a simple nonparametric approach grouping observations based on their similarities and spatial locations. It is a method used to partition a dataset into K clusters. The objective is to minimize the sum of squared distances between the data points and their assigned cluster centroids. The algorithm starts with an initial random assignment of the centroids and iteratively reassigns the data points to the nearest centroid until convergence.

There are several k-means algorithms available. The standard algorithm is the Hartigan-Wong algorithm, which provides a variation of k-means algorithm which progresses towards a local minimum of the minimum sum-of-squares problem with different solution updates[1]. The method is a local search that iteratively attempts to relocate a sample into a different cluster as long as this process improves the objective function. When no sample can be relocated into a different cluster with an improvement of the objective, the method stops (in a local minimum). In a similar way as the classical k-means, the approach remains a heuristic since it does not necessarily guarantee that the final solution is globally optimum.

The individual cost of the centroid is defined as:

$$\varphi(S_i) = \sum_{x \in S_j} (x - \mu_j)^2$$

x_i design a data point belonging to the cluster S_j; μ_j is the mean value of the points assigned to the cluster.

Assignment step: Hartigan and Wong's method starts by partitioning the points into random clusters $\{S_j\}_{j \in \{1,\ldots,k\}}$.

Update step: Next it determines the $n, m \in \{1, \ldots k\}$ and $x \in S_n$ for which the following function reaches a maximum

$$\Delta(m, n, x) = \varphi(S_n) + \varphi(S_m) - \varphi(S_n \backslash \{x\}) - \varphi(S_n \cup \{x\})$$

For the x, n, mthat reach this maximum, x moves from the cluster S_n to the cluster S_m.

Termination: The algorithm terminates once $\Delta(m, n, x)$ is less than zero for all x, n, m.

Different move acceptance strategies can be used. In a first-improvement strategy, any improving relocation can be applied, whereas in a best-improvement strategy, all possible relocations are iteratively tested and only the best is applied at each iteration. The former approach favors speed, whether the latter approach generally favors solution quality at the expense of additional computational time. The function Δ used to calculate the result of a relocation can also be efficiently evaluated by using equality

$$\Delta(x, n, m) = \frac{\mid S_n \mid}{\mid S_n \mid -1} \cdot \mid\mid \mu_n - x \mid\mid^2 - \frac{\mid S_m \mid}{\mid S_m \mid -1} \cdot \mid\mid \mu_m - x \mid\mid^2$$

The standard **R** function for k-means clustering is kmeans() in package stats[2]. fviz_nbclust() in factoextra package provides a convenient solution to estimate the optimal number of clusters. All the processes of K-means clustering in our study are implemented with help of these functions.

LCA

LCA is a probabilistic model-based method that assumes Gaussian Mixture distributions[3]. It relates a set of observed multivariate variables from a heterogeneous population to a set of underlying categorical latent variables. The assumption underlying LCA is that membership in unobserved groups (or classes) can be explained by patterns of scores across survey questions, assessment indicators, or scales. The LCA model:

$$\Pr(x) = \sum_{k=1}^{K} w_k \prod_{j=1}^{p} \pi_{jk}$$

LCA makes the conditional independence assumption:

$$\Pr(\mathbf{x} \mid \mathrm{c} = \mathrm{k}) = \prod_{j=1}^{p} \Pr\left(x_j \mid c = k\right)$$

where $w_k = \Pr(c = k)$ and $\pi_{jk} = \Pr\left(x_j \mid c = k\right), \mathrm{j} = 1 \ldots \mathrm{p}, \mathrm{k} = 1..\mathrm{K}$.

In **R**, mclust is a contributed R package for model-based clustering, classification, and density estimation based on finite normal mixture modelling.[4]. All the processes of LCA clustering in our study are implemented with help of this package.

Non-normal data generation

The Acceptance-Rejection Method, also called as rejection sampling, is used as a Monte Carlo Method in this study to generate 2-dimensional random variables from a specific target joint Probability Density Function(PDF) with non-normal features. The Acceptance-Rejection Algorithm is simple and efficient to implement for any distribution in $\mathbb{R}^m$ with a density, especially when the inverse function of target Cumulative Density Function(CDF) is difficult to obtain.

In this simulation study, the bivariate normal distribution is chosen as the convenient distribution:

$$g_{X,Y}(x,y) = \frac{1}{2\pi\sigma_1\sigma_2} exp\{-\frac{1}{2(1-\rho^2)}[(\frac{x-\mu_1}{\sigma_1})^2 + (\frac{y-\mu_2}{\sigma_2})^2 - 2\rho(\frac{x-\mu_1}{\sigma_1})(\frac{y-\mu_2}{\sigma_2})]\}$$

This is because the bivariate normal distribution covers the support of all target PDFs in this simulation study. A constant M is determined, such that $M = \sup_{x,y\in(-\infty,\infty)} \frac{f_{X,Y}(x,y)}{g_{X,Y}(x,y)}$. The method involves following steps:

1. A bivariate data point (X, Y) is generated from the convenient PDF. In this simulation study, function rmnorm() in **R** helps generate data points following bivariate normal distribution.
2. A random number U is generated, where $U \sim Unif(0,1)$. Accept (X, Y) when $U \leq \frac{f_{X,Y}(x,y)}{M g_{X,Y}(x,y)}$; otherwise, reject (X, Y) and go back to step 1.

Skewness

Skewed bivariate data refers to a type of data that exhibits skewness in both the X and Y variables when plotted on a scatter plot. Skewness refers to the extent to which a distribution of data is asymmetric, with one tail being longer or more stretched out than the other.

Skewed bivariate data can occur in various real-world applications. For example, the relationship between the size and weight of organisms can be skewed due to factors such as species differences and growth rates.

To include the case of dependency between the two random variables in a bivariate distribution, we derive a skewed joint distribution by log-transforming one of the variables in a bivariate normal distribution which introduces ρ as the correlation coefficient between two variables.

After the log-transformation, the target density function of skewed bivariate distribution is chosen as:

$$f_{X,Y}(x,y) = \frac{1}{2\pi\sigma_1\sigma_2 y} exp\{-\frac{1}{2(1-\rho^2)}[(\frac{x-\mu_1}{\sigma_1})^2 + (\frac{logy-\mu_2}{\sigma_2})^2 - 2\rho(\frac{x-\mu_1}{\sigma_1})(\frac{logy-\mu_2}{\sigma_2})]\}$$

where the support of X is $(-\infty, \infty)$, the support of Y is $(0, \infty)$.

The bivariate normal distribution is selected as the convenient distribution:

$$g_{X,Y}(x,y) = \frac{1}{2\pi\sigma_1\sigma_2} exp\{-\frac{1}{2(1-\rho^2)}[(\frac{x-\mu_1}{\sigma_1})^2 + (\frac{y-\mu_2}{\sigma_2})^2 - 2\rho(\frac{x-\mu_1}{\sigma_1})(\frac{y-\mu_2}{\sigma_2})]\}$$

Here the μ_i, $i = 1, 2$, represent the location of the distributions. Different values of μ are controlled using inter-cluster distance d. The skewness of the target distribution is affected by the shape parameter σ_i, $i = 1, 2$. The value of σ is positively correlated with the skewness, the larger the sigma, the more the skewness. The impact of different shape parameters values σ on clustering performance are also discussed.

Figure 1 shows the 3D plot and contour plot of a skewed distribution.

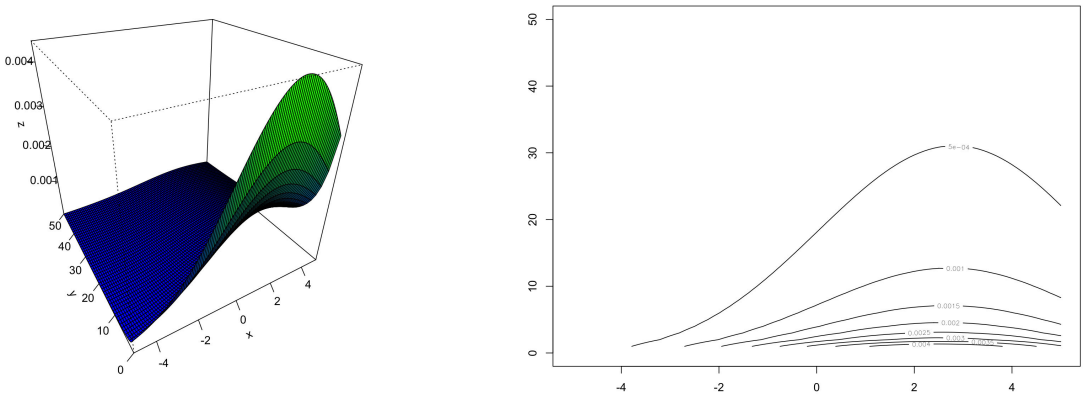

Figure 1: Skewed distribution

Multimodality

The mixture of two bivariate normal distribution with two different μ's is a distribution with two separate peaks when the distance of μ_{1i} and μ_{2i}, $i = 1, 2$ is large enough. Thus, we can get the density of a bimodal joint distribution of (X, Y) by assigning λ and $1 - \lambda$ as constants to two bivariate normal distributions using the formula below:

$$\begin{aligned} &f_{X,Y}(x,y) = \lambda N_1(x,y|\mu_1,\Sigma_1) + (1-\lambda)N_2(x,y|\mu_2,\Sigma_2) \\ &= \lambda\frac{1}{2\pi\sigma_{11}\sigma_{12}} exp\{-\frac{1}{2(1-\rho_1^2)}[(\frac{x-\mu_{11}}{\sigma_{11}})^2 + (\frac{y-\mu_{12}}{\sigma_{12}})^2 - 2\rho_1(\frac{x-\mu_{11}}{\sigma_{11}})(\frac{y-\mu_{12}}{\sigma_{12}})]\}+ \\ &(1-\lambda)\frac{1}{2\pi\sigma_{21}\sigma_{22}} exp\{-\frac{1}{2(1-\rho_2^2)}[(\frac{x-\mu_{21}}{\sigma_{21}})^2 + (\frac{y-\mu_{22}}{\sigma_{22}})^2 - 2\rho_2(\frac{x-\mu_{21}}{\sigma_{21}})(\frac{y-\mu_{22}}{\sigma_{22}})]\} \end{aligned}$$

where the support of X and Y are both $(-\infty, \infty)$.

The bivariate normal distribution can be used as convenient distribution in the process of acceptance and rejection.

The intra-cluster distance for cluster k is measured by the difference between μ_{ik} and μ_{2k}.the inter-cluster distance between is measured by the Euclidean distances between cluster centers. The parameter λ, taking

values from 0 to 1, controls the proportion of two peaks within one cluster. For example, a small value of λ produces two peaks extremely asymmetric in size. When $\lambda = 0.5$, two peaks are equal in size.

Figure 2 shows the 3D plot and contour plot of bimodal normal distribution.

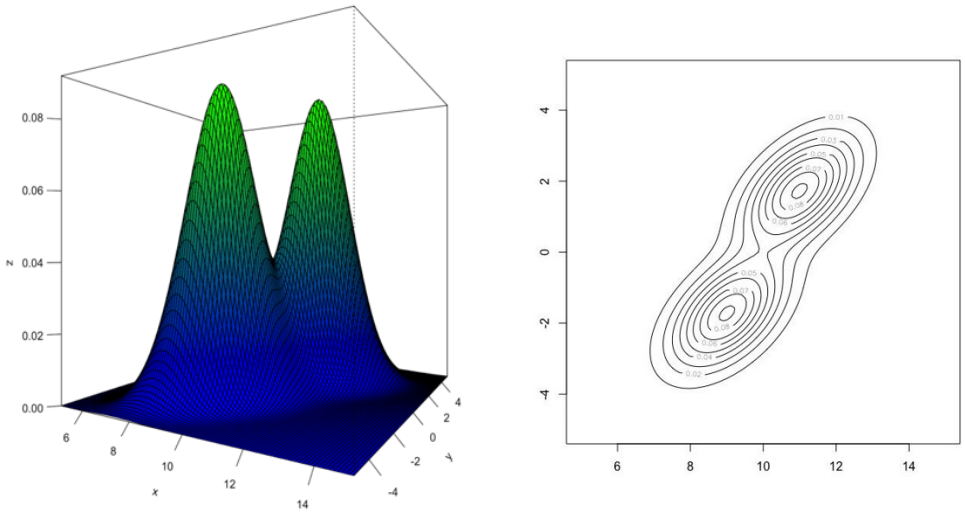

Figure 2: Bimodal normal distribution

Heavy-tails

We use bivariate t-distribution to generate data with heavy-tails distribution with joint pdf like:

$$f(x,y) = \frac{\Gamma[(\nu+2)/2]}{\Gamma(\nu/2)\nu\pi|\mathbf{\Sigma}|^{1/2}}\left[1+\frac{1}{\nu}(\mathbf{x}-\boldsymbol{\mu})^T\mathbf{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right]^{-(\nu+2)/2}$$

where Σ is the scale matrix, ν is the degrees of freedom. The larger the parameter ν, the heavier the tail is. The support of X and Y are both $(-\infty, \infty)$.

The bivariate normal distribution can be used as convenient distribution in the process of acceptance and rejection.

Figure 3 shows the 3D plot and contour plot of bivariate t-distribution.

Design of simulation setting

We conducted simulation studies to assess the performance of the two clustering methods under different scenarios. We fix the true number of clusters at 3. The initial location of the three cluster centers is fixed at (0,0), (1,0), (1/2, $\sqrt{3}/2$) respectively. For all three types of non-normal data, we increased the inter-cluster distance by adding a multiplier d on the initial coordinates of three cluster centers. d takes values in 1, 2, 4, 8, 16. We changed the number of data points within one cluster, defined as cluster size n, to be 10, 50, 200 for each distribution. To control the level of skewness, we changed coefficient of sigma to be 1, 10, 100. To control the proportion of peaks, we changed the lambda to be 0.1, 0.2, 0.3, 0.4, 0.5. We changed the degrees of freedom ν to be 2, 4, 16, 32, 64 to control how heavy the tail is. For each set of parameters, we repeat the simulation process 10 times. All the simulation processes were performed in R.

parameter	meaning	value
d	inter-cluster distance	1, 2, 4, 8, 16
n	cluster size	10, 50, 200

parameter	meaning	value
σ	level of skewness	1, 10, 100
λ	proportion of peaks	0.1, 0.2, 0.3, 0.4, 0.5
ν	degrees of freedom	2, 4, 16, 32, 64

Table 1: Parameters for simulation

Performance measure

We use misclassification rate to measure the performance of each clustering method. The value can be calculated by:

$$MisclassificationRate = \frac{IncorrectPredictions}{TotalPredictions}$$

which is calculated with confusion matrix. A lower value of misclassification rate suggests a better performance.

Results

Skewness

In skewed data, we explore how does the distance of cluster centers, cluster size and the degree of skewness affect the performance of two clustering methods.

Figure 4 shows that both K-means and LCA have a high misclassification rate which indicates a bad performance, when the center of clusters are too close to each other.But with the increase of the distance, the performances of both methods get better, especially we can see that K-means has a rapid improvement in performance and the misclassification rate almost approaches 0 when the distance is large enough. When the n gets larger, K-mean performs almost the same, but LCA has an obviously worse performance, since we can see that the misclassification rate remains pretty high even if the distance is large enough.

By fixing the inter-cluster distance, we change the coefficient of sigma matrix to control the degree of skewness. Since the greater the coefficient, the more skewness, it's easy to find out that the degree of skewness has a negative impact on performance of both methods in Figure 5. This negative impact becomes slight after the misclassification rate has already reached a very high level. When n gets larger, K-means has a better performance, while LCA has a worse performance.

Notably, LCA performs the worst in all degree of skewness at n = 200. It is interesting that LCA performs the worst with the largest cluster size in all distance value and sigma coefficient. We will discuss this in the later part.

Overall, we can conclude that K-means performs better than LCA in skewed data.

Multimodality

In multimodal data, we focus on the effect of inter-cluster distance, cluster size and lambda which is the proportion of two peaks within each cluster.

By changing the inter-cluster distance, we can find in Figure 6 that only when the inter-cluster distance exceeds the intra-cluster distance, the performance of both methods starts to improve. And after that, K-means always has a lower misclassification rate than LCA. When n gets larger, K-means performs better and becomes 100% correct at a smaller inter-cluster distance, while LCA has a worse performance with the increase of n. By changing the value of lambda which is the proportion of two peaks within one cluster, we can find that the asymmetry of peaks doesn't influence much on such trends.

Then, we fix the inter-cluster distance and try different lambda values ranging from 0.1 to 0.5. The result can be seen in Figure 7. When the inter-cluster distance is small, LCA has a better performance when the

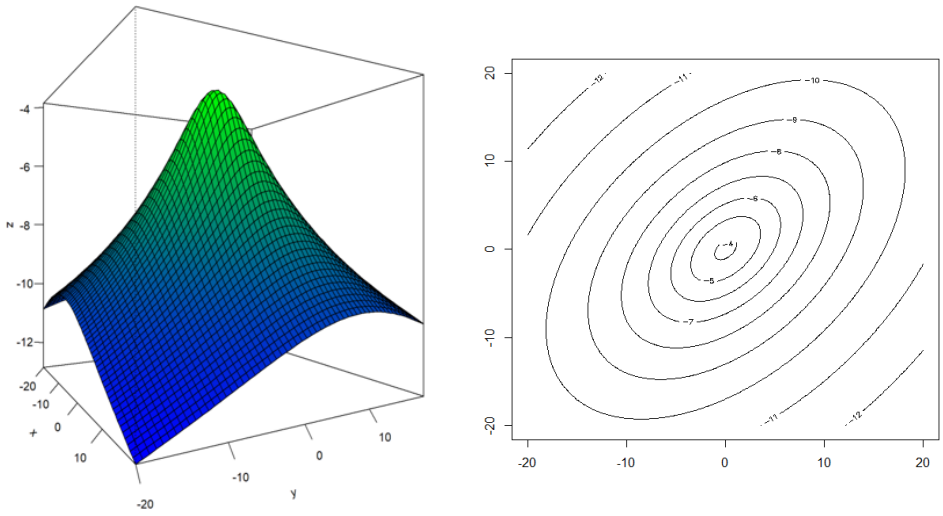

Figure 3: Bivariate t-distribution distribution

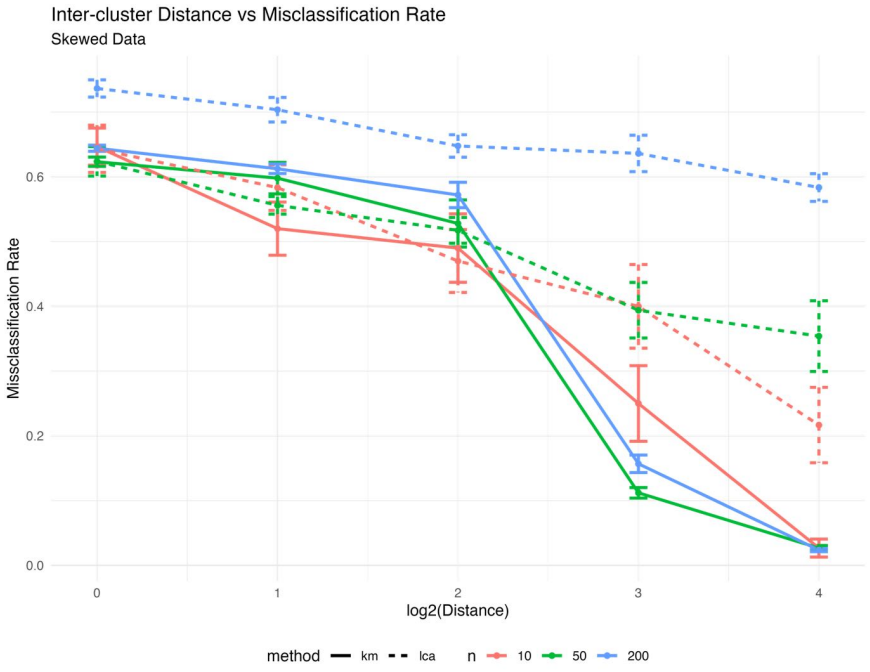

Figure 4: Skewness: Inter-cluster Distance vs Misclassification Rate

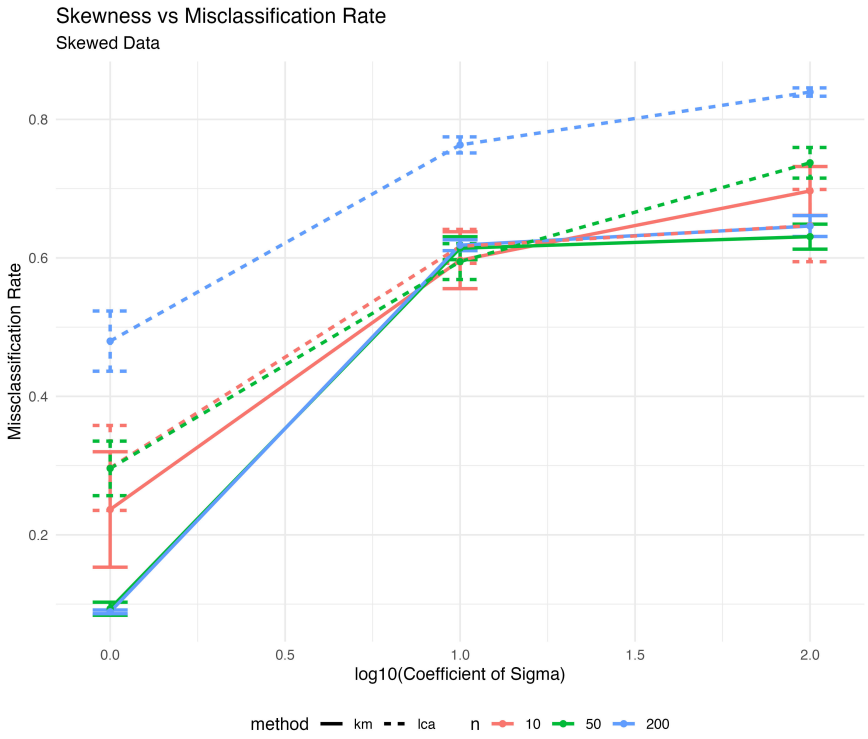

Figure 5: Skewness: Skewness vs Misclassification Rate

proportion of the two peaks are extremely asymmetric, but both methods perform bad when the proportion of two peaks approach almost the same. When the inter-cluster distance is large enough, K-means can always remain 100% correct after reaching a certain cluster size, while LCA performs worse with the increase of lambda and n.

Still, we can conclude that K-means performs better than LCA on multimodal data.

Heavy-tails

In heavy tails, we also change the inter-cluster distance, cluster size and degrees of freedom (df) of joint t distribution to explore their influence on the performance of clustering.

Figure 8 shows that both methods have a bad performance when the center of clusters are too close to each other but gradually get better with the increase of distance especially after reaching a threshold. When n gets larger, both K-means and LCA have a better performance. When the inter-cluster distance is large enough, K-means always performs better than LCA on relatively small cluster size.

By changing the value of df, Figure 9 shows that both methods have a bad performance when df is extremely small but get better and remain relatively stable when df > 2. When n gets larger, both K-means and LCA have a better performance, but K-means can reach a very low misclassification rate value at a smaller df than LCA.

Overall, K-means performs better than LCA in heavy-tails data.

Conclusion and Discussion

In conclusion, K-means performs better than LCA in different non-normal features, including skewness, multimodality and heavy-tails. But in fact, both methods perform bad in extreme overlapping and non-normal

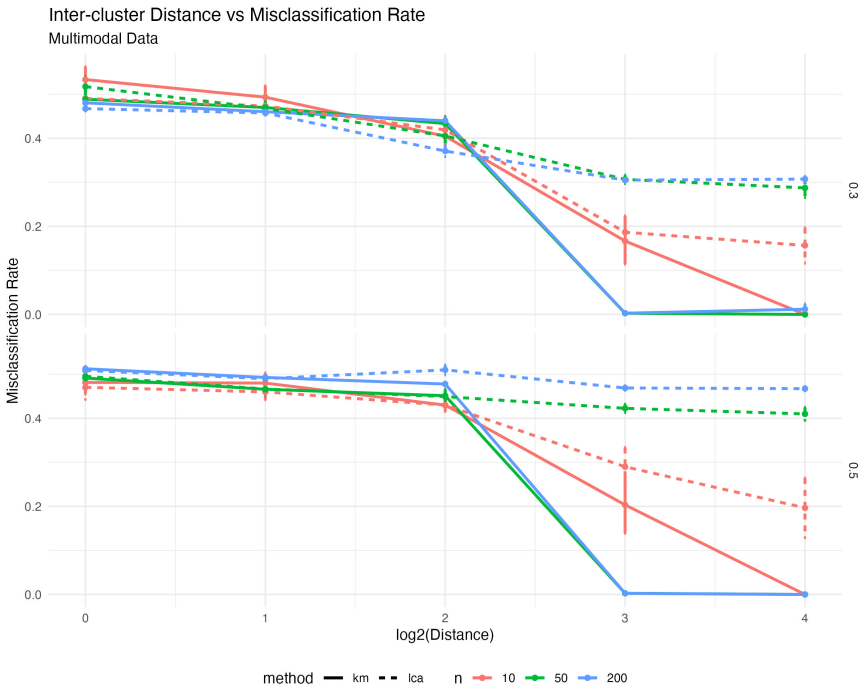

Figure 6: Multimodality: Inter-cluster Distance vs Misclassification Rate

cases and get better when the distance between clusters is large enough and the underlying distribution tends to be normal.

In most scenarios, the performances of the two clustering methods improve when the cluster size increases. This always works in the K-means method as the trend of improvement with the increase of n is shown in all three non-normal cases. The improvement from n = 10 to 50 is always more explicit than that from n = 50 to 200. However, when we continue increasing the value of n to 500, the improvement of performances in both methods are almost invisible compared to n = 200, which proves that the increase of cluster size cannot infinitely improve the performance.

Interestingly, we find that LCA sometimes performs worse on a large cluster size of non-normal data as LCA incorrectly separates data into more clusters. This may be because LCA assumes Gaussian Mixture distributions. When the sample size is large enough to show non-normal features, LCA may try to decompose the non-normal distribution into several normal distributions with different parameters and thus produce more clusters. But this rarely happens when the cluster size is too small to show the influence of non-normal property.

Acknowledgement

All members contributed equally to the project.

Reference

[1] Hartigan J A, Wong M A. Algorithm AS 136: A k-means clustering algorithm[J]. Journal of the royal statistical society. series c (applied statistics), 1979, 28(1): 100-108.

[2] K-Means Clustering in R: Algorithm and Practical Examples. https://www.datanovia.com/en/lessons/k-means-clustering-in-r-algorith-and-practical-examples/.

[3] Estiri H. Building and household X-factors and energy consumption at the residential sector: A structural equation analysis of the effects of household and building characteristics on the annual energy consumption

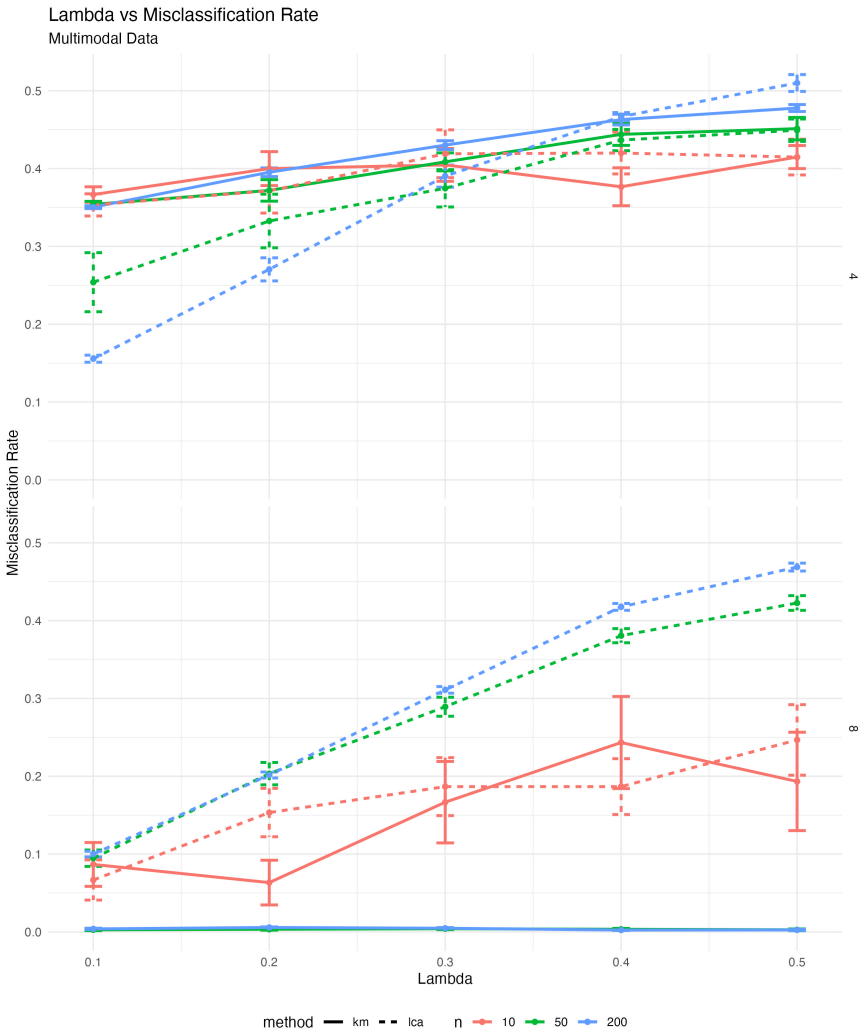

Figure 7: Multimodality: Lambda vs Misclassification Rate

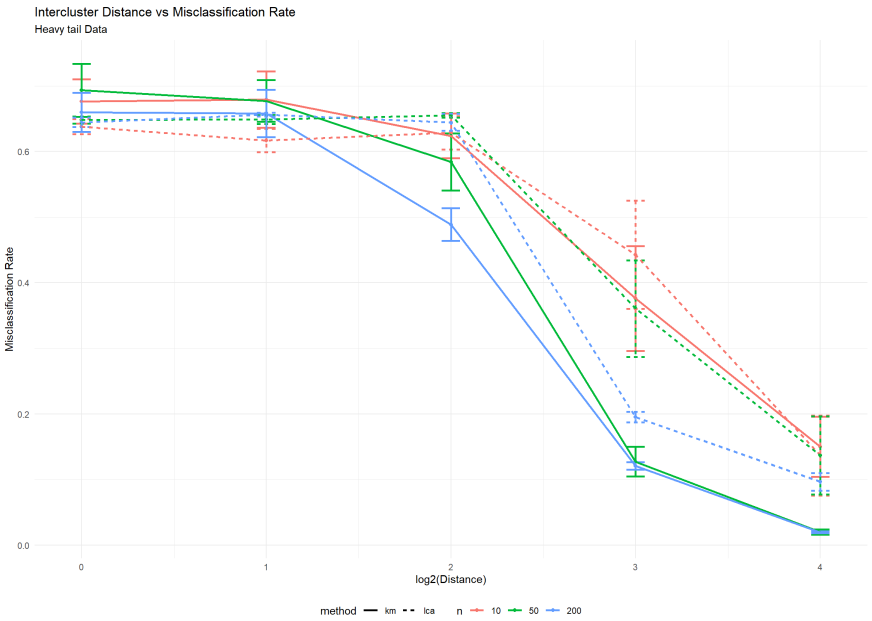

Figure 8: Heavy-tails: Inter-cluster Distance vs Misclassification Rate

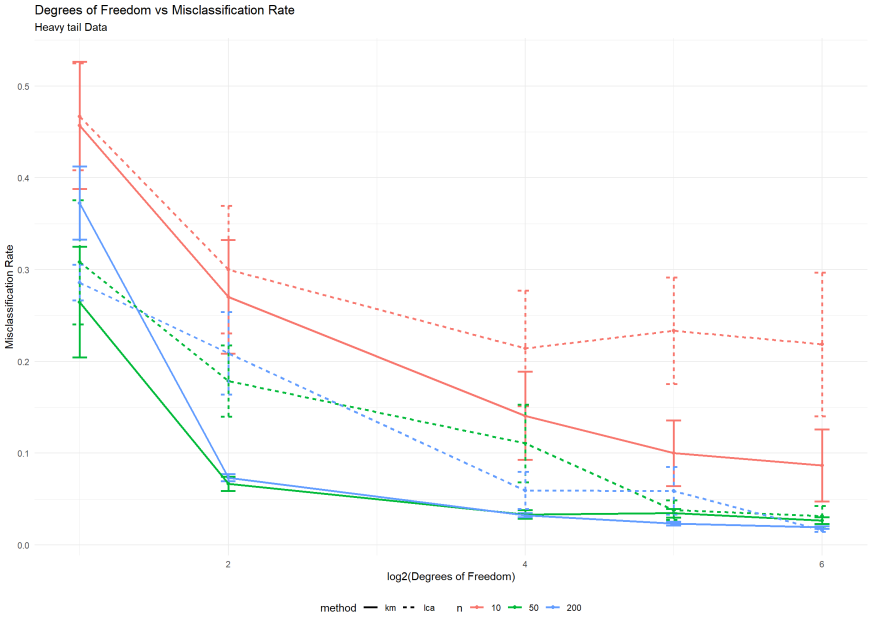

Figure 9: Heavy-tails: Degrees of Freedom vs Misclassification Rate

of US residential buildings[J]. Energy Economics, 2014, 43: 178-184.

[4] A quick tour of mclust. https://cran.r-project.org/web/packages/mclust/vignettes/mclust.html#clustering.

Supplementary materials

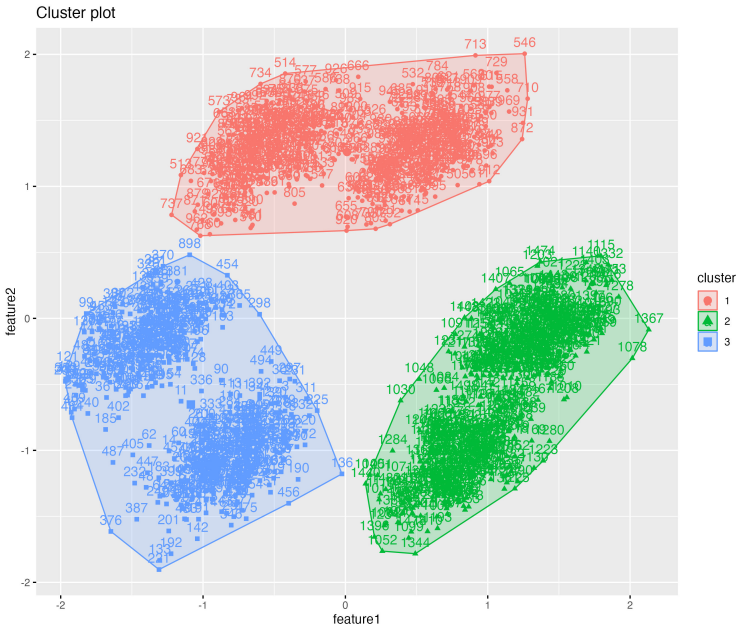

Figure 10: A sample result of K-means clustering

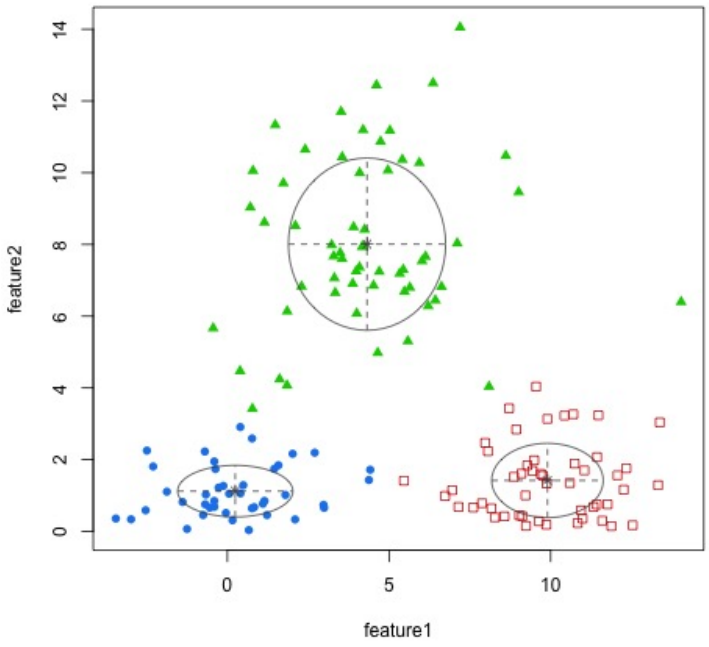

Figure 11: A sample result of LCA clustering

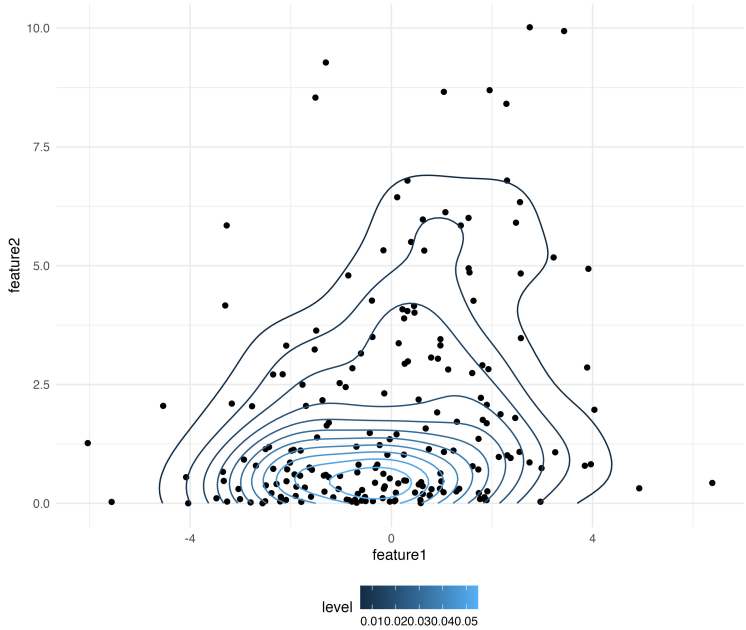

Figure 12: A sample of generated skewed data

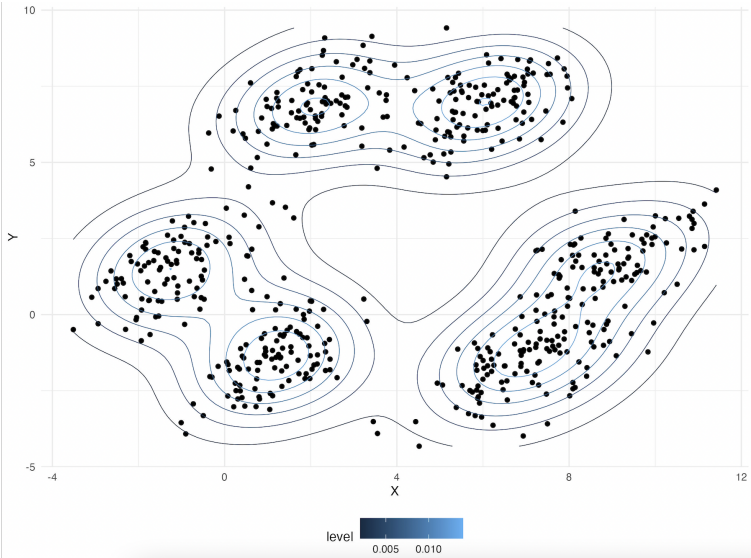

Figure 13: A sample of generated multimodal data

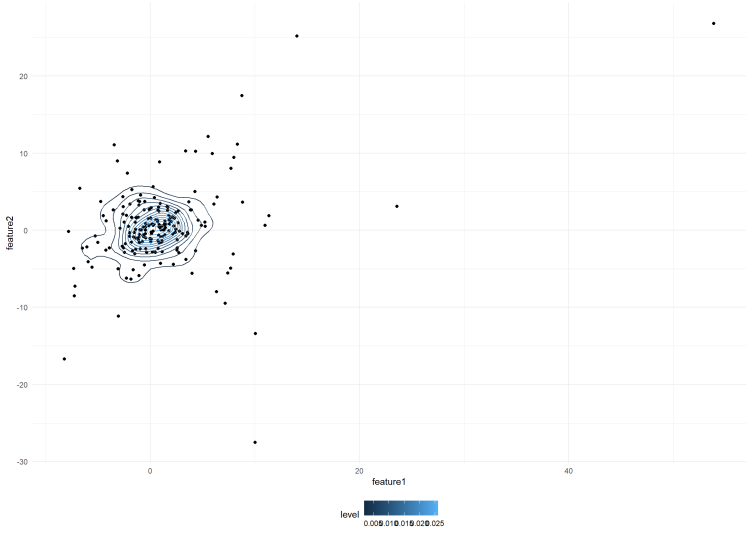

Figure 14: A sample of generated heavy-tails data