Journal of Data Acquisition and Processing
ISSN: 1004-9037
https://sjcjycl.cn/
DOI: 10.5281/zenodo.777533

SPECIFY UNDERLINING DISTRIBUTION FOR CLUSTERING LINEARLY SEPARABLE DATA: NORMAL AND UNIFORM DISTRIBUTION CASE

Farag Hamad[1*], Najiah Younus[2], Mohamad M.A. Muftah[2], and Mohamad Jaber[3]
[1]Department of Statistics, Faculty of Arts and Science/ Al Maraj, University of Benghazi, Libya
[2]Department of Mathematics, Faculty Arts and Science/ Al Abayar, University of Benghazi, Libya
[3]Department of Statistics, Faculty Science, University of Misurata, Libya

ABSTRACT

Clustering is one of the useful methods that we use to classify data. There are diverse statistical methods that can be used to divide the data into different groups. Cluster analysis is performed to discover distinct individuals that share the same common features within a large population. The observations within the same group have similar features from one to another and are different from observations in other groups. Eventually, clusters are classified by each group to determine which individual belongs to what group. In the past decades, clustering has been increasingly used in data analysis and data mining. Several clustering methods have been developed for grouping the data that share common features. k-means clustering: k-means, k-means++, and kernel k-means are the most important statistical tools used for clustering data. These methods are performed to classify the linearly separable data. Moreover, there are different ways to classify the data by assigning underline distributions to the data. In this paper, different distributions have been assigned as underline distributions for clustering the data. Clustering simulation data can be accomplished by assigning a normal or uniform distribution, as was done in this study. In order to see the improvement for each method, we assigned two different distributions (normal and uniform distributions) to classify linearly separable simulated data. The results were compared with the k-means method and with the ground truth of the data. The study found improvements in clustering when using a uniform density function. Moreover, a lower overlap percentage was found when we used the uniform density function for clustering the data. Using a significance test, there is no significant difference found between the estimated cluster mean and the cluster underlying mean. In addition, the proposed methods perform well when the sample size is larger.

Keywords: Probability density,k-means algorithm, linearly separable data, normal, and uniform distribution.

INTRODUCTION

Cluster analysis has been extensively used in a variety of areas, such as artificial intelligence, visual pattern recognition, web search, biology, and security[1]. Clustering is a technique used in business intelligence to categorize a large number of customers into groups where each group of customers has many characteristics with the others [2]. Moreover,

considering a consulting company that has several ongoing tasks, developing company plans for improving customer relationship management can be made by cluster analysis. Cluster analysis can be also used in the image process to discover clusters or subclasses. Clustering analysis can be used to improve project management by partitioning projects into categories or groups based on similarity [3].

Grouping the data into distinct clusters is one of the most important tools for learning. Moreover, cluster analysis studies the ways to find objects or observations which are similar. There are several algorithms for producing data separation in cluster analysis. Many of these algorithms are performed well under various conditions [4]. In cluster analysis, the number of clusters should be determined before starting the clustering procedure. There are several steps can be defined in order to start the clustering such as cluster center and cluster interval [5]. The data partition is the simplest and most fundamental version of cluster analysis. Cluster analysis is one of the most common unsupervised learning methods in machine learning which is used to discover underlying patterns or grouping data [6]. Several algorithms have been developed for clustering data such as k-means, Fuzzy c-means, mixture models, and spectral clustering [4].

The k-Means algorithm can be viewed as a gradient method that starts with an initial set of k cluster centers and iteratively updates it to reduce the error function. The k-means algorithm is also widest algorithm used for data partitioning [7]. In the k-means algorithm, we defined the number k of the clusters before selecting randomly centroid for each cluster. In the k-means procedure, the center for each cluster is updated in each iteration [8]. The k-means cluster method differs from the determination of the cluster's center and minimizes the distance from the entities to the mean of the assigned cluster. Despite the efficiency of the k-means cluster algorithm to extract information from huge data, the k-means cluster is sensitive to outliers [9]. The sum of squared error between all objects within the cluster can be used to demonstrate the quality of the cluster. There is no guarantee that the k-means method will converge to the global optimum and usually terminates at a local optimum [7]. The overlap between clusters means that the points belong to a certain cluster and are classified into different clusters [10]. The two clusters are overlapping when the data space is present in the same area. The overlapping was measured by the percentage of misclassification between two clusters partition [11], [12].

METHODOLOGY

Many different ways to cluster the data using underlying distributions, but one common approach is to use a probabilistic model for the data. One popular probabilistic model for clustering is the mixture model, which assumes that the data is generated from a mixture of different probability distributions (also called components or clusters). In these methods, we would like to cluster data points based on their probability density function. Our goal of clustering data using the proposed methods is to estimate the probabilities that each data point belongs to each cluster, given the current underline distribution.

Linearly separable data refers to a set of data points in a high-dimensional space that can be separated into two or more classes by a single straight line or hyperplane. A cluster of linearly separable data is a useful technique used for data visualization. There are several

methods for separating the data linearly such as Discriminant Analysis, logistic regression, and support vector machine. These methods work well for linearly separable data, as the clear boundary between classes makes it possible to accurately classify new samples. However, if the data is not linearly separable, more complex models, such as non-linear SVMs or neural networks are needed. Figure 1 demonstrates data visualization based on their classes.

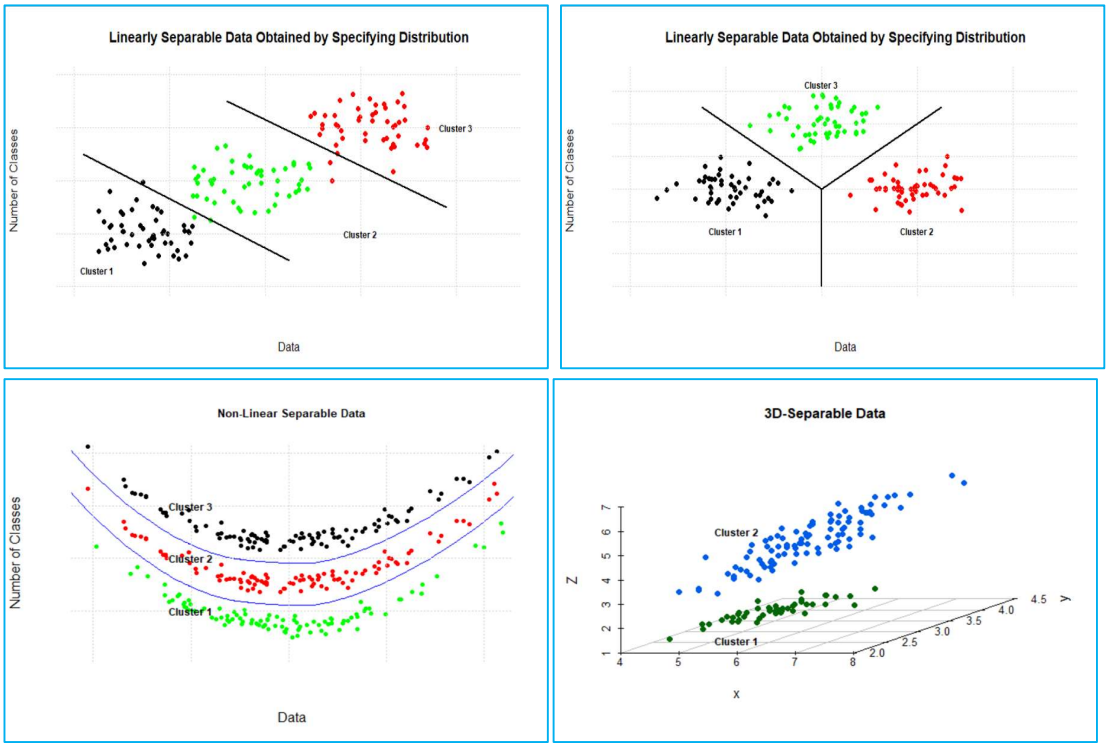

Figure 1: Data clustered into three groups using underlying distribution

Here are some of the most popular clustering methods. Each method has its strengths, weaknesses, and the suitability of each method depends on the structure of the data. More details and results are discussed next:

1- *K*-means clustering

k-means is a clustering algorithm that groups similar data points together by defining a prototype for each group, called a centroid, which is typically the mean of the data points in the group[3]. The observations within a group are likely to be similar to each other and dissimilar to observations in other groups. The k-means algorithm is known for its simplicity and ease of implementation. The k-means algorithm is given by:

$$\arg\min_{S} \sum_{i=1}^{k} \sum_{x_j \in S_i} \|X_j - \mu_i\|^2$$

where, μ_i is the mean of points in S_i. The k-mean cluster analysis algorithm is formally described below steps [4].

① Select k points as initial centroids.
② Repeat.
③ From the k cluster by assigning each point to its closest centroid.
④ Recompute the centroid of each cluster.

⑤ Until centroids do not change

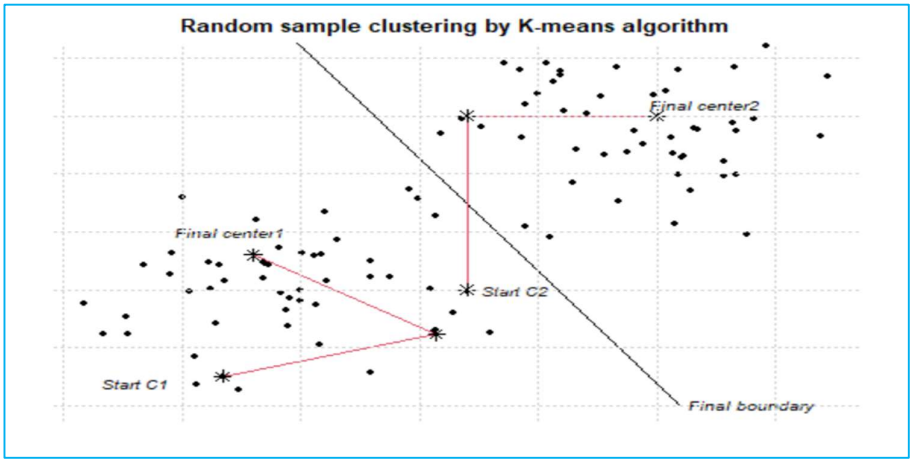

Figure 2: Finding two classes center using the k-means algorithm, c_1, and c_2 are randomly chosen.

2- Normal Distribution Clustering

The most important and most widely used distribution in Statistics is a normal distribution. Normal distributions can differ in their means and standard deviations [13]. The probability density function of a normal distribution lies from minus infinity to positive infinity $(-\infty, \infty)$ [14]. The normal distribution is called symmetric distribution when the distribution mean, median, and mode are equals [15]. In the symmetric distribution, there are 50% of the data lying to the left of the mean, and 50% lies to the right of the mean. The probability density function of normal distribution includes two parameters μ and σ are the mean and standard deviation, respectively [16]. A normal distribution with zero mean and one standard deviation is called a standard normal distribution. The PDF of the normal distribution is given by:

$$f(x) = \frac{1}{\sqrt{2\pi\,\delta^2}}\ \exp\left[(x-\mu)/\delta\right]^2, \quad -\infty \le x \le \infty ,$$

where, $\mu = \int_{-\infty}^{\infty} xf(x)dx$ and $\delta^2 = \int_{-\infty}^{\infty} x^2 f(x)dx - \left[\int_{-\infty}^{\infty} xf(x)dx\right]^2$

The normal PDF properties are below

① $f(x) \ge 0$

② $\int_{-\infty}^{\infty} f(x)\,dx = 1$

③ 68.4% from the area under pdf curve between $(\mu - \sigma,\ \mu + \delta)$.

④ 95.4% from the area under pdf curve between $(\mu - 2\sigma,\ \mu + 2\delta)$.

⑤ 99.7% from the area under pdf curve between. $(\mu - 3\sigma\ , \mu + 3\delta)$.

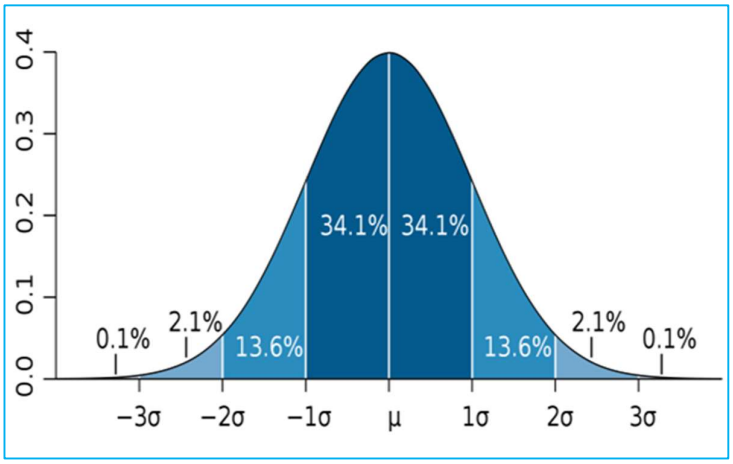

Figure 3: Areas Under the Normal Distribution

Clustering data using normal distribution can be done by constructing the boundaries of each class using probability density measurements. The density function of the normal distribution is used to find the likelihood of each data observation. By comparing the observation probability with the class boundaries, we can determine which class the observation belongs to [8]. The data classes and boundaries can be determined by estimating the parameters of the normal distribution (mean and standard deviation) for each class and using the probability density function to define the class boundaries. This method is also known as the Gaussian mixture model or Expectation-maximization algorithm. We determine the classes boundaries using normal distribution density by:

Class	Class boundary
1	$\int_{-\infty}^{c1} f(x)dx = \frac{1}{k} \rightarrow \int_{-\infty}^{c1} f(z)dz = \frac{1}{k} \rightarrow p(-\infty < z < c1) = \frac{1}{k}$
2	$\int_{c1}^{c2} f(x)dx = \frac{2}{k} \rightarrow \int_{c1}^{c2} f(z)dz = \frac{2}{k} \rightarrow p(c1 < z < c2) = \frac{2}{k}$
3	$\int_{c2}^{c3} f(x)dx = \frac{3}{k} \rightarrow \int_{c2}^{c3} f(z)dz = \frac{3}{k} \rightarrow p(c2 < z < c3) = \frac{3}{k}$
Where $c_i, i = 1,2,..,k$ Class boundary and k # of classes	

3- Uniform Distribution Clustering

Clustering using a uniform distribution refers to a method of grouping similar data points together based on a uniform distribution function. This means that each data point has an equal chance of being assigned to a cluster [17]. Moreover, the data points with an equal likelihood of occurring are uniformly distributed. The uniform distribution is also known as a rectangular distribution [18]. The uniform distribution is widely used for generating a random sample because the uniform distribution defines equal probability over a given range $[a, b]$ [14]. The density function of random variable X that restricted to a finite interval $[a, b]$ and its constant over the interval defined by:

$$f(x) = \begin{cases} \frac{1}{b-a} & , \quad \text{a} \leq \text{x} \leq \text{b} \\ 0 & o.w \end{cases}$$

$$\mu = \int_{-\infty}^{\infty} xf(x)dx = \frac{a+b}{2} \text{ and } \delta^2 = \int_{-\infty}^{\infty} x^2 f(x)dx - \left[\int_{-\infty}^{\infty} xf(x)dx\right]^2 = \frac{(b-a)^2}{12}$$

Class	Class boundary
1	$\int_a^{c1} f(x)dx = \frac{1}{k} \rightarrow \int_a^{c1} \frac{1}{b-a}dx = \frac{1}{k} \rightarrow \frac{c1-a}{b-a} = \frac{1}{k}$
2	$\int_{c1}^{c2} f(x)dx = \frac{2}{k} \rightarrow \int_{c1}^{c2} \frac{1}{b-a}dx = \frac{2}{k} \rightarrow \frac{c2-c1}{b-a} = \frac{2}{k}$
3	$\int_{c2}^{c3} f(x)dx = \frac{3}{k} \rightarrow \int_{c2}^{c3} \frac{1}{b-a}dx = \frac{3}{k} \rightarrow \frac{c3-c2}{b-a} = \frac{3}{k}$
Where $c_i, i = 1,2,..,k$ class boundary and k # of classes	

SIMULATION STUDY

A simulation study was conducted to illustrate the performance of the proposed methods. We evaluated the proposed methods by running simulation studies and comparing the results with k-means algorithm results. The boundaries of each class are estimated using proposed methods by assigning underline distribution such as normal and uniform distribution. We replicated the experiment by generating three groups of data, each group including 20, 50, and 200 observations that drew from the multivariate normal distribution. The dataset was generated from a multivariate normal distribution with three different centers. The variance-covariance matrix of the multivariate normal distribution has been chosen to make the data well linearly separable data. For each cluster, the distribution of the data is assigned first i.e., for the observation i we compute the probability that the observation belongs to class j using the probability density function of the assigned distribution. For each group, we compute the class boundary and the probabilities of the observations based on the density function of the normal distribution and uniform distribution.

RESULTS AND DISCUSSION

In this section, we use the simulation to demonstrate the performance of each method. By comparing the results of simulations, we can discover the strengths and weaknesses of each method and we can make informed decisions on which method is performed for a particular use case. Numerical results based on the proposed methods are demonstrated in the next tables:

Table 1: Random sample with size 20 observations clustered by density distribution.

Method	K-means			Normal			Uniform		
Cluster	1	2	3	1	2	3	1	2	3

1	17	3	0	20	0	0	20	0	0
2	0	20	0	0	16	4	0	18	2
3	0	1	19	0	0	20	0	0	20
Total	17	24	19	20	16	24	20	18	22
Overlap	6.67%			6.67%			3.33%		

Table 1 shows clustering simulation data using the proposed. Three different methods are used to cluster the random data including 20 observations into three different clusters. The cells in the table contain numbers, which represent the number of data points that were assigned to each cluster under each method. From the results, we can see that the data clustering by k-means and underline normal distribution are comparable; based on the parentage of the overlap. Moreover, the clustering by assigning underlined uniform distribution providesless overlapping between the clusters.

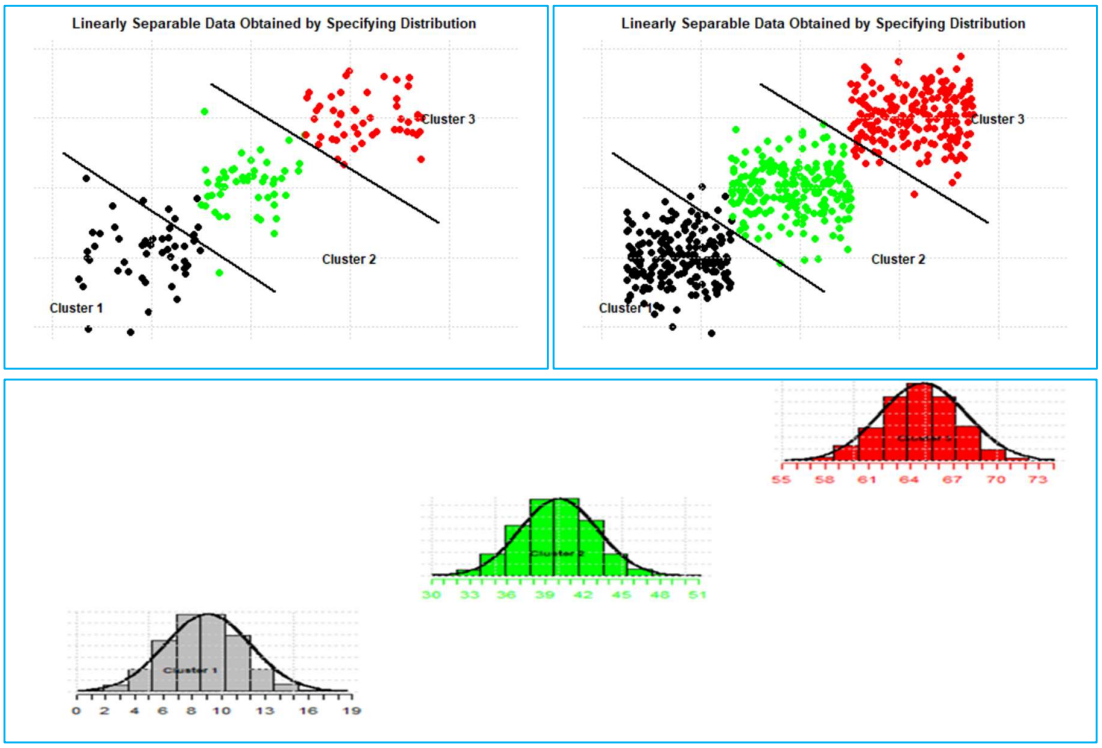

Figure 4: Random linearly separable data with different sample sizes (50 top left and 200 top right), the bottom is underlined normal distribution for three different centers (mean).

Table 2: Random sample with a size of 50 observations clustered by density distribution.

Method	K-means			Normal			Uniform		
Cluster	1	2	3	1	2	3	1	2	3
1	50	0	0	50	0	0	50	0	0
2	7	43	0	6	36	8	0	47	3
3	0	6	44	0	0	50	0	0	50
Total	57	49	44	56	36	58	50	47	53
Overlap	8.67%			9.33%			2.00%		

In Table 2, we replicated the experiment when the sample size is 50 observations. We applied the proposed methods to cluster the data into three groups. From the table results, we observe that the overlap percentage using both methods (k-means and normal) is increased compared with the previous experiment. We can also see that the clustering based on the uniform distribution density function provides less overlapping between the groups compared with other methods.

Table 3: Random sample with a size of 200 observations clustered by density distribution.

Method	K-means			Normal			Uniform		
Cluster	1	2	3	1	2	3	1	2	3
1	199	1	0	200	0	0	200	0	0
2	0	200	0	30	141	29	0	196	4
3	0	6	194	0	0	200	0	0	200
Total	199	207	194	230	141	229	200	196	204
Overlap	1.17%			9.83%			0.67%		

From above Table 3, we can see that the cluster method by specifying underlined uniform distribution for the dataset is improved when the sample size is increased. We can also observe that the method using underlined normal distribution comes after the cluster using the density of uniform distribution and the k-mean method. The improvement of the cluster using the density of the uniform distribution might come from the likelihood, each data point has an equal chance of being assigned to a cluster.

Table 4: Statistical difference between the estimated mean for each cluster with ground mean.

Method	Cluster	n	Mean	Sd	T	P-value
K-means	1	193	20.50	6.01	1.20	0.23
	2	202	40.07	5.44	-3.49	0.08
	3	205	60.91	6.34	-2.40	0.02
Normal	1	220	21.26	6.44	2.90	0.00
	2	159	40.01	4.01	-6.26	0.00
	3	221	59.43	7.09	-5.39	0.00
Uniform	1	211	20.81	6.18	1.90	0.06
	2	214	41.85	5.84	-0.37	0.71
	3	175	61.86	5.87	-0.31	0.75

Table 4 shows the comparison between the estimated mean, standard deviation for each group with the ground mean and standard deviation. From the results, we can observe that the proposed methods (normal and uniform) are provided clusters for the data equivalent to the clusters that done by k-means method. The variation inside the clusters shows that the uniform method is classified the data similar to clusters that done by k-means method. Moreover, using

the significant test, we can accept the hypothesis that there is no significant difference between the estimated clusters mean and the cluster ground mean specially using uniform method.

CONCLUSION

Clustering-based density is a method for dividing a dataset into clusters based on the probability density of each data point. These algorithms find densely populated regions in the data to identify clusters and consider points outside of these regions as noise. The algorithms require specifying density parameters and may also need other parameters, such as minimum cluster size or maximum distance between points. Linearly separable data can be clustered by assuming the underlying distribution and using algorithms like k-means or Gaussian mixture model to group similar data points. This work aims to compare the clustering using the k-means algorithm and the clustering using the density function. We used the density of normal distribution and uniform distribution to cluster the dataset after the assigned underlined distribution. The results show that the clustering using uniform density is improved better, by assuming the better method that provides less overlap percentage (since the data are linearly separable). Using a significance test, we can accept the assumption that there is no significant difference between the estimated cluster mean and the cluster underlying mean. Moreover, the method performance is increased when the sample size is increased.

References

[1] A. K. Jain, M. N. Murty, and P. J. Flynn, "Data clustering: a review," *ACM Comput. Surv. CSUR*, vol. 31, no. 3, pp. 264–323, 1999.

[2] P. Fränti and S. Sieranoja, "How much can k-means be improved by using better initialization and repeats?," *Pattern Recognit.*, vol. 93, pp. 95–112, 2019.

[3] P.-N. Tan, M. Steinbach, and V. Kumar, "Data mining cluster analysis: basic concepts and algorithms," *Introd. Data Min.*, vol. 487, p. 533, 2013.

[4] L. Kaufman and P. J. Rousseeuw, *Finding groups in data: an introduction to cluster analysis*. John Wiley & Sons, 2009.

[5] L. Rokach and O. Maimon, "Clustering methods," in *Data mining and knowledge discovery handbook*, Springer, 2005, pp. 321–352.

[6] M. Lux and S. Rinderle-Ma, "DDCAL: Evenly Distributing Data into Low Variance Clusters Based on Iterative Feature Scaling," *J. Classif.*, pp. 1–39, 2023.

[7] W. I. D. Mining, "Data mining: Concepts and techniques," *Morgan Kaufinann*, vol. 10, pp. 559–569, 2006.

[8] D. Ho-Kieu, T. Vo-Van, and T. Nguyen-Trang, "Clustering for probability density functions by new-medoids method," *Sci. Program.*, vol. 2018, 2018.

[9] W. Fu and P. O. Perry, "Estimating the number of clusters using cross-validation," *J. Comput. Graph. Stat.*, vol. 29, no. 1, pp. 162–173, 2020.

[10] A. Adam and H. Blockeel, "Dealing with overlapping clustering: A constraint-based approach to algorithm selection," presented at the Meta-learning and Algorithm Selection workshop-ECMLPKDD2015, 2015, vol. 1, no. 1, pp. 43–54.

[11] F. Bonchi, A. Gionis, and A. Ukkonen, "Overlapping correlation clustering," *Knowl. Inf. Syst.*, vol. 35, pp. 1–32, 2013.

[12] F. Hamad and N. N. Kachouie, "A hybrid method to estimate the full parametric hazard model," *Commun. Stat.-Theory Methods*, vol. 48, no. 22, pp. 5477–5491, 2019.

[13] D. Lane, D. Scott, M. Hebl, R. Guerra, D. Osherson, and H. Zimmer, *Introduction to statistics*. Citeseer, 2003.

[14] J. L. Devore, *Probability and Statistics for Engineering and the Sciences*. Cengage learning, 2011.

[15] C. Kraaikamp and H. Meester, "A modern introduction to probability and statistics," 2005.

[16] N. L. Johnson, S. Kotz, and N. Balakrishnan, *Continuous univariate distributions, volume 2*, vol. 289. John wiley & sons, 1995.

[17] T. S. Madhulatha, "An overview on clustering methods," *ArXiv Prepr. ArXiv12051117*, 2012.

[18] F. Hamad, S. Abdulkarim, and A. Hamad, "Mixture method to estimate baseline hazard for non-arbitrary function of the Cox proportional model," *IJSBAR*, vol. 62, no. 2, pp. 235–248, May 2022.