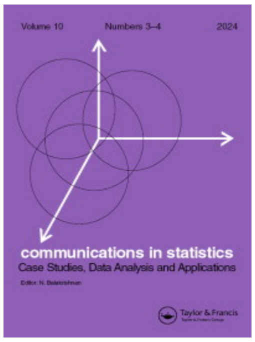

Volume 10
Numbers 3–4
2024
communications in statistics
Case Studies, Data Analysis and Applications
Taylor & Francis

Communications in Statistics: Case Studies, Data Analysis and Applications

ISSN: (Print) (Online) Journal homepage: www.tandfonline.com/journals/ucas20

Some density-based silhouette diagnostics for soft clustering algorithms

Shrikrishna Bhat Kapu & Kiruthika

COMMUNICATIONS IN STATISTICS: CASE STUDIES, DATA ANALYSIS AND APPLICATIONS
2024, VOL. 10, NOS. 3–4, 221–238
https://doi.org/10.1080/23737484.2024.2408534

Taylor & Francis
Taylor & Francis Group

Some density-based silhouette diagnostics for soft clustering algorithms

Shrikrishna Bhat Kapu and Kiruthika

Department of Statistics, Pondicherry University, Puducherry, India

ABSTRACT

One of the main objectives of cluster analysis is to determine the most effective clustering algorithm. With the wide variety of algorithms available, assessing which one performs better is important. The performance of different clustering methods is typically measured using the Adjusted Rand Index (ARI), which relies on knowledge of the original class labels. However, this study introduces flexible modified alternatives of density-based silhouette methods for evaluating cluster performance. These proposed Density-based silhouettes can be applied to any soft clustering algorithms and do not require the original class labels. Instead, they rely on posterior probabilities. In this study, eight different soft clustering algorithms were evaluated using real and simulated data sets. The goal is to compare their effectiveness and performance using existing and proposed measures based on silhouette and the ARI.

ARTICLE HISTORY
Received 31 July 2023
Accepted 20 September 2024

KEYWORDS
Model-based clustering; Soft clustering algorithms; Density-based silhouettes; Cluster diagnostics

1. Introduction

Cluster analysis is a statistical technique that groups similar observations to identify patterns in data. It is widely used for exploratory data analysis and can uncover hidden relationships within the data. Different methods exist for cluster analysis, including hierarchical, K-means, and model-based clustering. Model-based clustering assumes that the data is generated from a mixture of probability distributions, allowing for complex data structures and prior knowledge incorporation.

Cluster validity measures assess the quality of clustering results and aid in determining the appropriate number of clusters. Internal measures, such as silhouette-based methods (Rousseeuw 1987), evaluate clustering based on intrinsic properties and the clustering structure. The silhouette coefficient measures an observation's similarity to its cluster compared to neighboring clusters, with higher values indicating better matching. Silhouette-based methods can

CONTACT Shrikrishna Bhat Kapu ✉ skbhat.in@pondiuni.ac.in Department of Statistics, Pondicherry University, Puducherry, India.

determine the optimal number of clusters by comparing silhouette coefficient values, offering easy calculation, intuitive interpretation, and compatibility with different algorithms and data types. In this paper, new and modified density-based silhouette measures are proposed to evaluate soft clustering algorithms' performance and diagnostic ability.

The rest of the paper is organized as follows: Sec. 2 reviews cluster methods, assignments, and validity. In Sec. 3, different silhouette formulations for cluster validity and performance are proposed. In Sec. 4, the diagnostic ability of these formulations is evaluated on simulated and real data sets. Finally, Sec. 5 presents some concluding remarks.

2. Review of clustering and validity measures

2.1. Cluster assignment and model-based clustering

In the context of a $n \times p$ data matrix $\mathbf{X} = [x_i]_{i=1}^{n} \in R^p$, the goal is to assign each observation x_i to a cluster C_k with $k = 1, \ldots, K$. There are two ways to assign observations to clusters: soft or fuzzy clustering and hard clustering. In soft clustering, the assignment of x_i to a cluster C_k is determined using posterior probabilities $P(C_k|x_i)$, where $\sum_{k=1}^{K} P(C_k|x_i) = 1$. Soft clustering methods such as Fuzzy C-means (Bezdek, Ehrlich, and Full 1984), Probabilistic distance clustering (Ben-Israel and Iyigun 2018; Iyigun and Ben-Israel 2008; Tortora, McNicholas, and Palumbo 2020), and Model-based clustering (Fraley and Raftery 2002, 2007) are commonly used for this purpose.

On the other hand, in hard clustering, the observation x_i is assigned to a cluster C_k without considering posterior probabilities. The most popular hard clustering methods are K-means (Lloyd 1957,1982; Forgy 1965; MacQueen 1965; Hartigan and Wong 1979) and Partition around medoids (PAM) (Kaufman and Rousseeuw 1990). It is also possible to define hard clustering in terms of posterior probabilities, i.e.,

$$\text{If } x_i \in C_k \text{ then } P(C_k|x_i) = 1 \;\&\; P(C_{k'}|x_i) = 0 \;\forall\, k^{'} \neq k = 1, \ldots, K. \tag{1}$$

Model-based clustering is a sophisticated soft clustering algorithm built using finite mixture models. The probability density function of a K-component finite mixture distribution (see McNicholas 2016) of a p-dimensional random vector X is given by

$$f(x; \Theta) = \sum_{k=1}^{K} \pi_k f_k(x; \theta_k), \tag{2}$$

where π_k is mixing proportion such that $\pi_k \geq 0$ and $\sum_{k=1}^{K} \pi_k = 1$ and $f_k(x; \theta_k)$ are component densities with parameters θ_k. $\Theta = \{\pi_k, \theta_k\}_{k=1}^{K}$ is parametric space that will be estimated using the Expectation Maximisation (EM) (Dempster, Laird, and Rubin 1977) algorithm. Estimates of posterior probability $P(C_k|x_i) = \gamma_{ik}$ and π_k are given as

$$\gamma_{ik} = \frac{\pi_k f_k(x_i;\theta_k)}{\sum_{k=1}^{K} \pi_k f_k(x_i;\theta_k)}, \tag{3}$$

$$\pi_k = \frac{\sum_{i=1}^{n} \gamma_{ik}}{n}. \tag{4}$$

The hard clustering variation of EM for model-based clustering is called Classification EM (CEM) algorithm (Celeux and Govaert 1992). In this paper, CEM is utilized to build density-based silhouette indices.

2.2. Cluster validity and silhouette index

In both hard and soft clustering algorithms, different validity measures are used. For example, BIC (Schwarz 1978) and Integrated completed likelihood (ICL) (Biernacki, Celeux, and Govaert 2000) used as validity measures for model-based clustering but it cannot be used for other clustering algorithms. This limitation of the above-said measures is that these cannot be generalized to other cluster algorithms. However, there are classical validity measures like Dunn's Index (Dunn 1974), Silhouette Index (Rousseeuw 1987; Van der Laan, Pollard, and Bryan 2003) and Davies Bouldin Index (Davies and Bouldin 1979) which are mainly developed to validate across different distance-based clustering. A brief review of silhouette based index measures are discussed below.

The silhouette index is defined as for $x_i \in C_k$, if $a(x_i)$ is the average dissimilarity from point x_i to all other objects of C_k and $b(x_i)$ is the average dissimilarity from point x_i to all other objects of nearest cluster other than C_k to x_i. Then silhouette index (Rousseeuw 1987) (using dissimilarity proximity measure) of x_i is given by

$$S_{\mathrm{R87D}}(x_i) := \frac{b(x_i) - a(x_i)}{\max\{b(x_i), a(x_i)\}}.$$

above equation modifies for similarity proximity measure (Rousseeuw 1987) as

$$S_{\mathrm{R87S}}(x_i) := \frac{a(x_i) - b(x_i)}{\max\{a(x_i), b(x_i)\}}.$$

Van der Laan, Pollard, and Bryan (2003) simplified the above silhouette index for PAM and referred to it as medoid silhouette. For $x_i \in C_k$ the medoid silhouette for dissimilarity proximity measure is given by

$$S_{\mathrm{V03D}}(x_i) := \frac{\left(\min_{k' \neq k} \delta_{ik'}\right) - \delta_{ik}}{\max\left(\left(\min_{k' \neq k} \delta_{ik'}\right), \delta_{ik}\right)}, \tag{5}$$

where δ_{ik} is distance between x_i to cluster medoid of cluster C_k and $\max_{k' \neq k} \delta_{ik'}$ is minimum distance between x_i to cluster medoid of cluster other than C_k. For

$x_i \in C_k$ the medoid silhouette for similarity proximity measure is given by

$$S_{\mathrm{V03S}}(x_i) := \frac{\delta'_{ik} - \left(\max_{k' \neq k} \delta'_{ik'}\right)}{\max\left(\delta'_{ik}, \left(\max_{k' \neq k} \delta'_{ik'}\right)\right)}, \tag{6}$$

where δ'_{ik} is similarity between x_i to cluster medoid of cluster C_k and $\max_{k' \neq k} \delta_{ik'}$ is maximum similarity between x_i to cluster medoid of cluster other than C_k.

The Overall clustering silhouette index is calculated using the *Average silhouette width* (ASW), and it is given by

$$CS = \frac{1}{n} \sum_{i=1}^{n} S\left(x_i\right). \tag{7}$$

Since Eq. (7) is calculated for hard clustering algorithms like PAM, ASW here referred to as *Crisp silhouette* (CS). For Fuzzy clustering, (Campello and Hruschka 2006) proposed ASW calculation using posterior probabilities called *Fuzzy silhouette* (FS), which is given by

$$FS = \frac{\sum_{i=1}^{n} \left(\gamma_{ik} - \max_{k' \neq k} \gamma_{ik'}\right)^{\alpha} S\left(x_i\right)}{\sum_{i=1}^{n} \left(\gamma_{ik} - \max_{k' \neq k} \gamma_{ik'}\right)^{\alpha}}, \tag{8}$$

here $\alpha \geq 0$ is a weighting coefficient. The calculating FS (Fuzzy Silhouettes) for soft clustering algorithms involves two steps:

(1) Identify the proximity measure in the algorithm and calculate $S(x_i)$ using Eqs. (5) or (6), based on the nature of the proximity measure.
(2) Calculate FS for the soft clustering algorithm using the membership probabilities γ_{ik} and the $S(x_i)$ values from the previous step.

The EM algorithm, used in model-based clustering, relies on cluster membership probabilities, making it a fuzzy clustering method. However, the hard clustering variant of the EM algorithm can be used to determine proximity measures, as hard clustering algorithms are based on proximity rather than membership probabilities. This logic is applied in calculating proximity measure-based fuzzy silhouettes in Sec. 3.2

Under model-based clustering, Menardi (2011) proposed a density-based silhouette (DBS) measure, and it is given by

$$S_{\mathrm{DBS}}(x_i) := \frac{\ln \frac{\gamma_{ik}}{\max_{k' \neq k} \gamma_{ik'}}}{\max_{i=1,2,\ldots,n} |\ln \frac{\gamma_{ik}}{\max_{k' \neq k} \gamma_{ik'}}|}, \tag{9}$$

the numerator of Eq. (9) is obtained from substituting $\delta_{ik} = -\ln \gamma_{ik}$ in Eq. (5). But the denominator of Eqs. (9) and (5) are obtained differently. DBS doesn't replicate the formula of the original silhouette index. Along with ASW of DBS, Menardi (2011) used the median of ASW for the overall cluster silhouette Index.

This ASW is plotted along with different values of k. The possible optimal $K = k$ is the local maximum points in the plot (Local maxima criteria).

Raymaekers and Rousseeuw (2022) proposed silhouette based on the probability of the alternative class (PACS) evaluating the classification performance from different classifiers, and it is given by

$$S_{\text{PACS}}(x_i) := 1 - 2\frac{\max_{k' \neq k} \gamma_{ik'}}{\max_{k' \neq k} \gamma_{ik'} + \gamma_{ik}} = \frac{\gamma_{ik} - \max_{k' \neq k} \gamma_{ik'}}{\gamma_{ik} + \max_{k' \neq k} \gamma_{ik'}}. \tag{10}$$

Equation (10) is simplified PACS. PACS satisfies all properties of the original silhouette index proposed by Rousseeuw (1987) but it is not derived from the original silhouettes . In terms of $a(x_i)$ and $b(x_i)$ notations of Rousseeuw (1987) PACS-based silhouette for dissimilarity and similarity proximity measures, respectively can be written as

$$S_{\text{PACSD}}(x_i) := \frac{b(x_i) - a(x_i)}{b(x_i) + a(x_i)}, \tag{11}$$

$$S_{\text{PACSS}}(x_i) := \frac{a(x_i) - b(x_i)}{a(x_i) + b(x_i)}. \tag{12}$$

PACS denominator is greater than denominator of $S_{\text{R87D}}(x_i)$ and $S_{\text{R87S}}(x_i)$ which offers more penalization when misclassification happens. The equation Raymaekers and Rousseeuw (2022) used PACS to evaluate the classification performance of different classifiers. However, the silhouette index assesses the number of clusters in the data in cluster analysis. This paper checked the possibility of existing and proposed novel posterior probability-based silhouette indices for assessing the performance of different soft clustering algorithms. Compared cluster performance evaluated from Silhouette indices with ARI (Rand 1971; Hubert and Arabie 1985). It also examined the diagnostic ability of the proposed silhouette indices for evaluating the number of clusters in real and simulated data sets.

3. Flexible alternatives of density based silhouette measures

In this section we proposed and discusses three new posterior probability-based and two new proximity-measure-based silhouette measures.

3.1. Posterior probability based silhouette measures

The posterior probability γ_{ik} is commonly used in all soft clustering algorithms. The posterior probability is often believed to be a similarity measure by its nature. However, there is no evidence that the posterior probability satisfies all properties of a similarity proximity measure. The following silhouette indices are highlights based on the assumption that the posterior probability is a similarity proximity measure. Considering this and substituting $\delta'_{ik} = \gamma_{ik}$ in

Eq. (6). For $x_i \in C_k$, the result is obtained as.

$$S_{\text{PPS}}(x_i) := \frac{\gamma_{ik} - \max_{k' \neq k} \gamma_{ik'}}{\max\left(\gamma_{ik}, \max_{k' \neq k} \gamma_{ik'}\right)} = 1 - \frac{\max_{k' \neq k} \gamma_{ik'}}{\gamma_{ik}}. \tag{13}$$

$S_1(x_i)$ in Eq. (13) is termed as *Posterior Probability Silhouette* (PPS). PPS is a less penalized variation of PACS.

The posterior probability, which is a measure of the likelihood of an observation belonging to a certain class, falls between 0 and 1. On the other hand, Proximity measures typically range from 0 to ∞, where a value of 0 represents identical or very close instances, and larger values indicate greater dissimilarity. To transform the posterior probability into a proximity measure, one can take the negative logarithm of the posterior probability which ranges from 0 to ∞. From this by substituting $\delta_{ik} = -\ln \gamma_{ik}$ in Eq. (5), where x_i belongs to class C_k, the equation is transformed as

$$S_{\text{NLPPS}}(x_i) := \frac{\left(\min_{k' \neq k} - \ln \gamma_{ik'}\right) - (-\ln \gamma_{ik})}{\max\left(\min_{k' \neq k} - \ln \gamma_{ik'}, -\ln \gamma_{ik}\right)} = 1 - \frac{\ln \gamma_{ik}}{\min_{k' \neq k} \ln \gamma_{ik'}}. \tag{14}$$

(14) is called *Negative Log of Posterior Probability Silhouette* (NLPPS). The numerators of NLPPS and DBS are the same. DBS is a more penalized silhouette variation of NLPPS.

Apart from the previous statement, the maximum posterior probability of observation exhibits similar properties to the silhouette index. It lies within the range of 0 and 1, with a value of 1 indicating high certainty that a data point belongs to its assigned cluster and a value close to 0 suggesting ambiguity or uncertainty when assigning the data point to a particular cluster. This maximum posterior probability silhouette is called the *Certainty Silhouette* (CeS).

$$S_{\text{CeS}}(x_i) := \max_k \gamma_{ik}, \tag{15}$$

it is named as Certainty Index because it is the counterpart of Uncertainty (Bouveyron et al. 2019, p. 29) used in model-based clustering. To calculate ASW of PPS, NLPPS, and CeS, use the CS given in Eq. (7).

3.2. Proximity measure-based silhouette measures

The logic of fuzzy silhouette calculation proposed by Campello and Hruschka (2006) is utilized to build a proximity-measure-based silhouette index for model-based clustering. First, the proximity measure embedded in the EM algorithm used for model-based clustering is identified. Then, the fuzzy average silhouette width is calculated using FS. In model-based clustering, finite mixture distributions are used to identify clusters where a proximity measure is not used.

To identify the proximity measure in model based clustering, CEM proposed by Celeux and Govaert (1992) is utilized. CEM algorithm is a hard cluster-limiting case of EM algorithm i.e., For $x_i \in C_k$, if $P\,(C_k|x_i) \to 1$ & $P\left(C_{k'}\middle|x_i\right) \to 0\ \forall\, k \neq k^{'} = 1, \ldots, K$. EM algorithm becomes CEM algorithm. The simplified log-likelihood and likelihood of CEM algorithm is given in Eqs. (16) and (17), respectively.

$$l(\Theta|X) = \sum_{k=1}^{K} \sum_{x_i \in C_k} \ln f_k\,(x_i; \theta_k), \tag{16}$$

$$L\,(\Theta|X) = \prod_{k=1}^{K} \prod_{x_i \in C_k} f_k\,(x_i; \theta_k). \tag{17}$$

(16) and (17) likelihood functions are similar to the hard clustering objective function of PAM

$$Q'\,(\Theta|X) = \sum_{k=1}^{K} \sum_{x_i \in C_k} \delta_{ik}, \tag{18}$$

here δ_{ik} is the dissimilarity function or within the sum of squares of kth cluster, and θ_k is its medoid. On observing the reduced optimization function of CEM in Eqs. (16) and (18), one can easily assess that $\ln f_k\,(x_i; \theta_k)$ works as a similarity measure that maximizes the likelihood of CEM. But $\ln f_k\,(x_i; \theta_k)$ is not always positive, which violates one of the properties of similarity measure (Theodoridis and Koutroumbas 2009, p. 602). Tortora, McNicholas, and Palumbo (2020) proposed a dissimilarity measure for model-based clustering, which is given as

$$\delta_{ik} = \left\{\ln f_k\,(M_k; \theta_k) - \ln f_k\,(x_i; \theta_k)\right\}, \tag{19}$$

M_k represents the mode parameter estimate of density $f_k\,(M_k; \theta_k)$. In this context, M_k remains constant for a given density, and the only value that varies for a given x_i is $\ln f_k\,(x_i; \theta_k)$. Equation (16) can be rewritten based on this analogy as

$$l'(\Theta|X) = \sum_{k=1}^{K} \sum_{x_i \in C_k} \left\{\ln f_k\,(M_k; \theta_k) - \ln f_k\,(x_i; \theta_k)\right\}. \tag{20}$$

Using Eq. (19) in (5), the *Density-Based Probabilistic-distance Silhouette* (DBPS) is obtained as follows: for $x_i \in C_k$, the DBPS is calculated as

$$S_{\mathrm{DBPS}}(x_i) := \frac{\min_{k' \neq k}\left\{\ln f_{k'}\left(M_{k'}; \theta_{k'}\right) - \ln f_{k'}\left(x_i; \theta_{k'}\right)\right\} - \left\{\ln f_k\left(M_k; \theta_k\right) - \ln f_k\left(x_i; \theta_k\right)\right\}}{\max\left(\min_{k' \neq k}\left\{\ln f_{k'}\left(M_{k'}; \theta_{k'}\right) - \ln f_{k'}\left(x_i; \theta_{k'}\right)\right\}, \left\{\ln f_k\left(M_k; \theta_k\right) - \ln f_k\left(x_i; \theta_k\right)\right\}\right)}. \tag{21}$$

Since Model-based clustering uses soft assignment, ASW of DBPS can be extended to Model-based clustering using FS given in Eq. (8).

It was demonstrated that Eqs. (17) and (18) can be considered comparable, and $f_k\,(x_i; \theta_k)$ functions as a similarity proximity measure. Any unimodal density function demonstrates a characteristic in which the probability decreases as the distance from the mode increases, showing an inverse relationship to the distance, indicating density as similarity proximity measure. Symmetric probability density function also satisfies three properties of similarity proximity measure (Theodoridis and Koutroumbas 2009, p. 602). Building upon this reasoning and considering it as evidence for the general probability density function as a similarity proximity measure, the *Probability Density Silhouette* (PDS) was proposed. In the case of $x_i \in C_k$, the PDS can be obtained as

$$S_{\text{PDS}}(x_i) := \frac{f_k(x_i, \theta_k) - \max_{k' \neq k} f_{k'}(x_i, \theta_{k'})}{\max\left(f_k(x_i, \theta_k), \max_{k' \neq k} f_{k'}(x_i, \theta_{k'})\right)}. \tag{22}$$

Using Eqs. (2)–(4), rewriting (22) in terms of posterior probabilities as

$$S_{\text{PDS}}(x_i) := \frac{\frac{\gamma_{ik}}{\pi_k} - \max_{k' \neq k} \frac{\gamma_{ik'}}{\pi_{k'}}}{\max\left(\frac{\gamma_{ik}}{\pi_k}, \max_{k' \neq k} \frac{\gamma_{ik'}}{\pi_{k'}}\right)}, \tag{23}$$

and for cluster silhouettes, one can use CS and FS. Since the modification in (23) is based on the Bayes theorem, (23) can be generally used for soft clustering methods if the posterior or cluster membership probabilities is known. All posterior probability-based silhouette measures proposed in Sec. 3.1 and PDS are purely based on posterior or cluster membership probabilities, these measures can be extended to any soft clustering algorithm. By following this logic, one can compare the performance of any soft clustering algorithm using soft cluster silhouettes.

In Sec. 4, a comparison is carried out to evaluate the performance of various soft clustering algorithms using proposed silhouette measures and the ARI. A higher ARI value indicates better cluster performance, suggesting that the algorithm's clusters are more similar to the true cluster labels. Similarly, a higher silhouette value also indicates better cluster performance, suggesting that the data points within each cluster are more tightly grouped and well-separated from other clusters.

To investigate the relationship between the performance given by ARI and the silhouette measures for each algorithm, Spearman rank correlation between ARI and each silhouette measure is computed. This correlation analysis helps us understand the degree of association between ARI and each silhouette measure.

It is worth noting that while silhouette measures are typically used for predicting the number of clusters, they are not primarily intended for performance evaluation. Nevertheless, the proposed silhouette measure can also be employed to predict the number of clusters in the dataset by plotting the ASW over different values of k (here k is used instead of K to denote plotting for varying cluster sizes). Usually, the graph for the proposed silhouette measure shows a reverse elbow shape, and the value of k where the primary reverse elbow drop forms is considered the optimal number of clusters for the given data (Reverse elbow drop criteria).

4. Results and discussions

The proposed silhouette measures are evaluated on simulated and real datasets. The simulated datasets have been used to illustrate the ability of silhouette measures in predicting the number of clusters in the datasets and also how well they work as a performance measure when compared to ARI. Real Datasets are used to check how well silhouette measures perform in evaluation and diagnostics of real life problems.

Eight soft clustering algorithms, namely Gaussian Mixture Models (GMM) (Scrucca et al. 2016), T Mixture Models (TMM) (Andrews et al. 2018), Gaussian Probabilistic Distance Clustering (GPDC), T Probabilistic Distance Clustering (TPDC), Probabilistic Distance Clustering (PDC), Probabilistic Distance Clustering adjusted for cluster Size (PDQ) (Tortora et al. 2022), Fuzzy C Means (Fuzzy), and Fuzzy C Means 2 (Fuzzy2) (Cebeci 2019), are considered for performance evaluation using silhouette measures and ARI in simulation study and real data analysis.

4.1. Simulation study

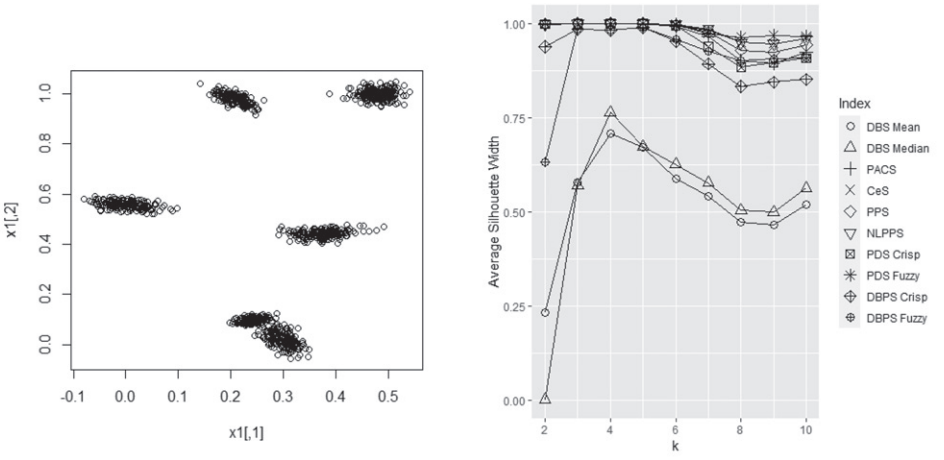
The simulation methodology proposed by Maitra and Melnykov (2010) allows parametrizing different degree of overlapping in Gaussian mixture densities. This simulation method helps to check the performance of proposed methods in different overlapping scenarios instead of fixing some predefined parameters. For the simulation study, three datasets were generated with 1,000 observations from a 6-component bivariate Gaussian mixture density using “`MixSim`” R Package (Melnykov, Chen, and Maitra 2012) which uses the methodology of Maitra and Melnykov (2010) with non-spherical covariance matrix structure with overlapping, moderately overlapping and non-overlapping components. The nature of 3 datasets is differentiable with `BarOmega` argument of `MixSim()` function i.e., `BarOmega = 0.001` (Well separated clusters), `BarOmega = 0.01` (Moderately overlapped clusters) and `BarOmega = 0.1` (Completely overlapped clusters). A detailed setup of arguments for silmulation is given in Table 1.

The number of clusters in each dataset is predicted by calculating mixture parameters and cluster membership values of observations using Gaussian Mixture Models with the ”mclust” R Package (Scrucca et al. 2016) for various values of $k = 2, \ldots, 10$. The resulting outputs for different k are then utilized to calculate and plot existing and proposed silhouette measures across the different values of k for three datasets. These plots are presented in Figs. 1–3 for the respective datasets. To determine the optimal number of

Table 1. Arguments considered in `MixSim()` function for 3 simulated datasets.

Arguments	Explanation	Values
`BarOmega`	Value of desired average overlap.	0.001, 0.01, 0.1
`MaxOmega`	Value of desired maximum overlap.	`NULL`[a]
`K`	Number of components.	6
`p`	Number of dimensions.	2
`sph`	Covariance matrix structure (`FALSE` - non-spherical, `TRUE` - spherical)	`FALSE`
`hom`	Heterogeneous or homogeneous clusters (`FALSE` - heterogeneous, `TRUE` - homogeneous).	`FALSE`
`ecc`	Maximum eccentricity.	0.9
`PiLow`	Value of the smallest mixing proportion	0.167[b]

[a] `MaxOmega = NULL` means Maximum overlap is considered randomly which adjusts with given `BarOmega`.
[b] Equal sized clusters.



(a) Simulated Data Set of 2 Dimensions 6 Component Gaussian Mixture Data with `BarOmega = 0.001`.

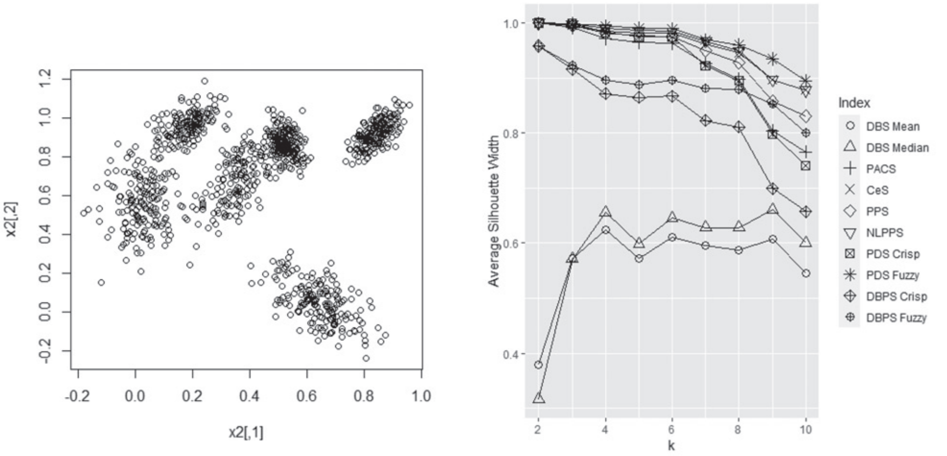
(b) k versus ASW of simulated data using posterior probabilities obtained from GMM.

Figure 1. Well separated clusters.

clusters for the simulated datasets, the local maxima criteria for DBS silhouette measure and the reverse elbow drop criteria for the PACS and proposed silhouette measures are employed. The summarized results of the optimal number of clusters for each datasets from different silhouette measures are presented in Table 2.

When clusters are well separated and moderately overlapped, the PACS and proposed silhouette measures accurately predict the number of clusters compared to the DBS Mean and DBS Median silhouettes. When clusters are completely overlapped, the PDS Fuzzy silhouette index accurately predicts the number of clusters, while other existing and proposed silhouette measures are far away from the actual cluster number K.

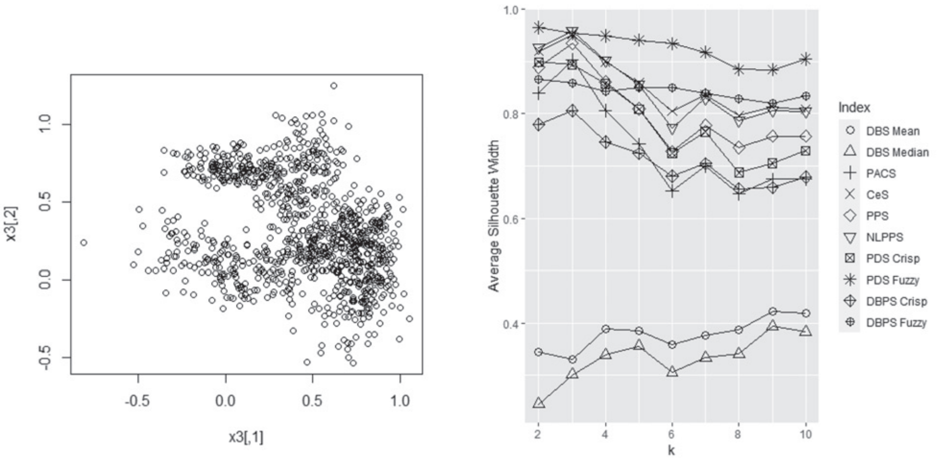
The performance evaluation of eight previously considered soft clustering algorithms on simulated datasets involved calculating the ARI and silhouette indices for each simulated dataset with $K = 6$. The results can be found in Tables 3–5. In Table 3, GMM and TMM algorithms display the highest ARI



(a) Simulated Data Set of 2 Dimensions 6 Component Gaussian Mixture Data with `BarOmega = 0.01`.

(b) k versus ASW of simulated data using posterior probabilities obtained from GMM.

Figure 2. Moderately overlapped clusters.



(a) Simulated Data Set of 2 Dimensions 6 Component Gaussian Mixture Data with `BarOmega = 0.1`.

(b) k versus ASW of simulated data using posterior probabilities obtained from GMM.

Figure 3. Completely overlapped clusters.

values, suggesting their suitability for well-separated cluster data. Concerning ASW, GMM exhibits high scores for PACS, CeS, PPS, NLPPS, and PDS Crisp and Fuzzy silhouettes, whereas TMM demonstrates high ASW scores for DBS Mean and Median compared to other algorithms.

Moving to Table 4, the TMM algorithm achieves the highest ARI value, indicating its appropriateness for moderately overlapped cluster data. In terms of ASW, GMM shows high scores for PACS, CeS, PPS, NLPPS, and PDS Crisp and Fuzzy silhouettes, while TMM exhibits high ASW scores for DBS Mean and Median compared to other algorithms.

Table 2. Optimal number of clusters that has been predicted using k versus ASW plots.

	Optimal Number of Clusters									
Dataset (BarOmega)	DBS Mean[a]	DBS Median[a]	PACS[b]	CeS[b]	PPS[b]	NLPPS[b]	PDS Crisp[b]	PDS Fuzzy[b]	DBPS Crisp[b]	DBPS Fuzzy[b]
0.001	4	4	6	6	6	6	6	6	5	5
0.01	4,6,9	4,6,9	6	6	6	6	6	6	6	6
0.1	5,9	5,9	3	3	3	3	3	6	3	3

[a] Local maxima criteria.
[b] Reverse elbow drop criteria.

Table 3. Well separated clusters: Comparison of silhouette indices and ARI for soft clustering algorithms.

	Average silhouette width								
Clustering algorithm	DBS Mean	DBS Median	PACS	CeS	PPS	NLPPS	PDS Crisp	PDS Fuzzy	ARI
GMM	0.5882	0.6259	0.9929[a]	0.9964[a]	0.9956[a]	0.9976[a]	0.9958[a]	0.9980[a]	0.9976[a]
TMM	0.7713[a]	0.8359[a]	0.9926	0.9963	0.9954	0.9975	0.9956	0.9979	0.9976[a]
GPDC	0.4915	0.4848	0.9125	0.9297	0.9427	0.9551	0.9433	0.9774	0.9516
TPDC	0.4834	0.4734	0.9071	0.9185	0.9394	0.9500	0.9400	0.9756	0.9495
PDC	0.4489	0.4228	0.7470	0.7389	0.8360	0.8344	0.7995	0.8816	0.7843
PDQ	0.4944	0.5037	0.7624	0.7681	0.8487	0.8574	0.8519	0.9059	0.9684
Fuzzy	0.4724	0.4624	0.9245	0.9509	0.9494	0.9650	0.9512	0.9796	0.9578
Fuzzy2	0.4150	0.3541	0.0000	0.1667	0.0000	0.0000	0.0000	0.0000	0.3331

[a] Best performance algorithm according to respective silhouette index and ARI.

Table 4. Moderately overlapped clusters: Comparison of silhouette indices and ARI for soft clustering algorithms.

	Average silhouette width								
Clustering algorithm	DBS Mean	DBS Median	PACS	CeS	PPS	NLPPS	PDS Crisp	PDS Fuzzy	ARI
GMM	0.6112	0.6453	0.9626[a]	0.9813[a]	0.9752[a]	0.9848[a]	0.9742[a]	0.9896[a]	0.9437
TMM	0.6772[a]	0.7371[a]	0.9594	0.9797	0.9738	0.9847	0.9729	0.9885	0.9546[a]
GPDC	0.4461	0.4339	0.8155	0.8278	0.8722	0.8750	0.8560	0.9493	0.8102
TPDC	0.4461	0.4327	0.8021	0.8028	0.8642	0.8614	0.8507	0.9447	0.8215
PDC	0.3056	0.2522	0.4494	0.4966	0.5621	0.5247	0.5704	0.8472	0.7183
PDQ	0.4176	0.4007	0.5861	0.5670	0.7100	0.6657	0.7097	0.8515	0.8789
Fuzzy	0.3940	0.3675	0.7500	0.8023	0.8257	0.8436	0.8092	0.9297	0.6997
Fuzzy2	0.1038	0.0677	0.0000	0.1667	0.0000	0.0000	0.0000	0.0000	0.2671

[a] Best performance algorithm according to respective silhouette index and ARI.

Finally, Table 5 reveals that the GMM algorithm obtains the highest ARI value, making it suitable for completely overlapped cluster data. In terms of ASW, GMM demonstrates high scores for PACS, CeS, PPS, and PDS Crisp and Fuzzy silhouettes, whereas TMM shows high ASW scores for NLPPS, DBS Mean, and Median compared to other algorithms.

For Tables 3–5, Spearman rank correlation is calculated between ARI and each silhouette ASW values and it is presented in Table 6.

From Table 6, it is observed that for both well-separated and moderately overlapped datasets, the DBS Mean silhouette index shows a slightly higher rank correlation with the Adjusted Rand Index (ARI) when compared to all other

Table 5. Completely overlapped clusters: Comparison of silhouette indices and ARI for soft clustering algorithms.

	Average silhouette width								
Clustering algorithm	DBS Mean	DBS Median	PACS	CeS	PPS	NLPPS	PDS Crisp	PDS Fuzzy	ARI
GMM	0.3580	0.3057	0.6525[a]	0.8051[a]	0.7277[a]	0.7733	0.7251[a]	0.9348[a]	0.5552[a]
TMM	0.403[a]	0.3662[a]	0.6427	0.8032	0.7329	0.7908[a]	0.7123	0.9049	0.5239
GPDC	0.3708	0.3316	0.6339	0.7058	0.7268	0.7349	0.7147	0.9013	0.5007
TPDC	0.3685	0.3216	0.6167	0.6703	0.7136	0.7089	0.7082	0.8990	0.5165
PDC	0.2932	0.2319	0.3824	0.4352	0.5154	0.4587	0.5237	0.7772	0.5055
PDQ	0.2999	0.2785	0.4180	0.4571	0.5499	0.4947	0.5511	0.7950	0.4535
Fuzzy	0.3723	0.3405	0.6415	0.6990	0.7384	0.7423	0.7361	0.9062	0.4574
Fuzzy2	0.0639	0.0308	0.0000	0.1667	0.0000	0.0000	0.0000	0.0000	0.2460

[a] Best performance algorithm according to respective silhouette index and ARI.

Table 6. Spearman rank correlation of different cluster silhouettes with ARI of each data set for eight clustering methods performed on simulated data using `BarOmega`.

	Rank correlation between ARI and ASW							
Dataset (`BarOmega`)	DBS Mean	DBS Median	PACS	CeS	PPS	NLPPS	PDS Crisp	PDS Fuzzy
0.001	0.9222[a]	0.9222[a]	0.8503	0.8503	0.8503	0.8503	0.8503	0.8503
0.01	0.9048[a]	0.8810	0.7619	0.7619	0.7619	0.7619	0.7619	0.7619
0.1	0.4286	0.4285	0.7142	0.7381[a]	0.4762	0.6905	0.4286	0.6190

[a] High correlated ASW and ARI ranking.

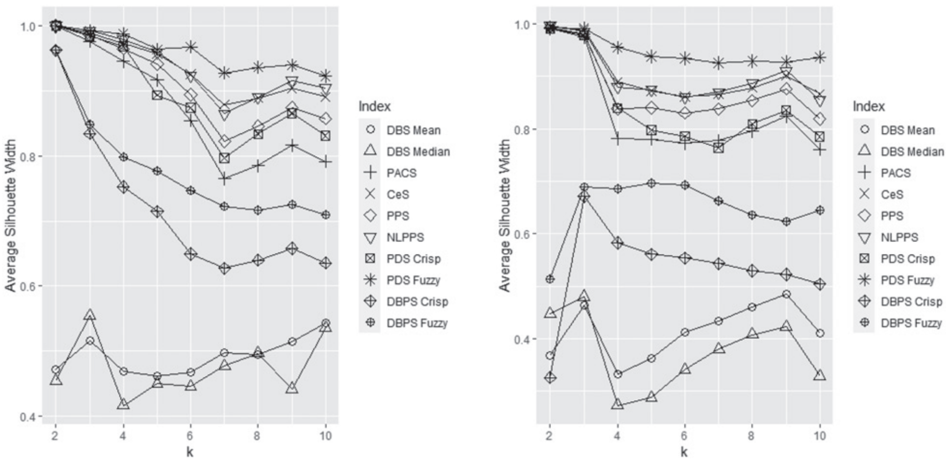
silhouette indices. Notably, the rank correlations of all silhouette indices with ARI are reasonably high, exceeding 0.7.

This suggests that when dealing with well-separated and moderately overlapped datasets, all the existing and proposed silhouette-based methods can serve as reliable performance indicators for clustering quality assessment, which is typically used to predict the appropriate number of clusters K. In the context of a completely overlapped cluster dataset, CeS exhibits a higher rank correlation with ARI when compared to other silhouette measures.

Since this correlation varies from data to data, and all existing and proposed silhouette measures are calculated without knowledge of existing class labels, one can calculate all silhouette measures to assess the clustering algorithm's performance more effectively when class labels are unknown.

4.2. Real data analysis

For Real Data Analysis, Iris Data Fisher (1936) and Thyroid Data Coomans et al. (1983) which has three classes in each are used. The number of clusters in two real datasets was predicted using the proposed and existing silhouette measures by computing ASW for varying values of k, ranging from 1 to 10. The relationship between the number of clusters (k) and the corresponding ASW values was plotted in Fig. 4. The prediction of the number of clusters in Iris and Thyroid datasets is performed by computing ASW of both proposed and



(a) Iris Data.

(b) Thyroid Data.

Figure 4. ASW versus *k* plots for real-life datasets.

Table 7. Iris data: Comparison of silhouette indices and ARI for soft clustering algorithms.

	Average silhouette width								
Clustering algorithm	DBS Mean	DBS Median	PACS	CeS	PPS	NLPPS	PDS Crisp	PDS Fuzzy	ARI
GMM	0.5154	0.5547	0.9762[a]	0.9881[a]	0.9855[a]	0.9925[a]	0.9847[a]	0.9926	0.9039[a]
TMM	0.6120[a]	0.6785[a]	0.9715	0.9857	0.9800	0.9865	0.9807	0.9935[a]	0.9038
GPDC	0.4963	0.4792	0.6590	0.7789	0.7608	0.8076	0.6292	0.8226	0.4868
TPDC	0.4848	0.4692	0.6981	0.8136	0.7857	0.8326	0.6641	0.8311	0.5418
PDC	0.4785	0.4808	0.5016	0.6795	0.6252	0.6740	0.6247	0.8116	0.7860
PDQ	0.3882	0.3722	0.5161	0.6871	0.6461	0.6984	0.6342	0.8080	0.7437
Fuzzy	0.4894	0.4828	0.7541	0.8572	0.8288	0.8750	0.8187	0.9280	0.7294
Fuzzy2	0.5242	0.4856	0.2936	0.5655	0.4148	0.4593	0.3981	0.6864	0.5619

[a] Best performance algorithm according to respective silhouette index and ARI.

existing silhouette measures using posterior probabilities obtained from GMM for $k = 2, \ldots, 10$. k versus ASW plots for Iris and Thyroid datasets are ploted in Fig. 4.

Figure 4(a), PACS and proposed silhouette measures reveals reverse elbow drop around 2 and 3, while the DBS silhouette exhibits a local maximum at 3. This suggests the presence of either 2 or 3 clusters in the Iris dataset. Figure 4(b), PACS and proposed silhouette measures reveals reverse elbow drop around 3, while the DBS silhouette exhibits a local maximum at 3. This suggests the presence of either 3 clusters in the Thyroid dataset.

The performance evaluation of eight previously considered soft clustering algorithms on Iris and Thyroid datasets involved calculating the ARI and silhouette indices for each dataset with $K = 3$. The results can be found in Tables 7 and 8.

From Table 7, GMM algorithms display the highest ARI values, suggesting their suitability for Iris data. Concerning ASW, GMM exhibits high scores for

Table 8. Thyroid data: Comparison of silhouette indices and ARI for soft clustering algorithms.

	Average silhouette width								
Clustering algorithm	DBS Mean	DBS Median	PACS	CeS	PPS	NLPPS	PDS Crisp	PDS Fuzzy	ARI
GMM	0.4631	0.4801	0.9748[a]	0.9874[a]	0.9806[a]	0.9850	0.9777[a]	0.9911[a]	0.8771
TMM	0.5554[a]	0.5867[a]	0.9741	0.9870	0.9804	0.9853[a]	0.9731	0.9911[a]	0.8917[a]
GPDC	0.3939	0.3706	0.3921	0.6094	0.5099	0.5540	0.5128	0.7504	0.2464
TPDC	0.3406	0.3169	0.4370	0.6384	0.5567	0.6014	0.5036	0.7690	0.1909
PDC	0.2962	0.2282	0.2870	0.5302	0.4075	0.4351	0.4336	0.6937	0.0407
PDQ	0.3619	0.3340	0.3503	0.6356	0.4801	0.5559	0.4211	0.6244	0.0659
Fuzzy	0.3864	0.3657	0.5952	0.7537	0.7007	0.7547	0.6388	0.8136	0.4413
Fuzzy2	0.3277	0.2893	0.3597	0.5881	0.4945	0.5397	0.4820	0.7009	0.0230

[a] Best performance algorithm according to respective silhouette and ARI.

Table 9. Spearman rank correlation of different cluster silhouettes with ARI of each data set for eight clustering methods performed.

	Rank correlation between ARI and ASW							
Dataset	DBS Mean	DBS Median	PACS	CeS	PPS	NLPPS	PDS Crisp	PDS Fuzzy
Iris	0.1667	0.5714[a]	0.4524	0.4524	0.4524	0.4524	0.5476	0.4524
Thyroid	0.9286[a]	0.9286[a]	0.8810	0.8810	0.8810	0.9048	0.8810	0.8571

[a] High correlated ASW and ARI ranking.

PACS, CeS, PPS, NLPPS, and PDS Crisp silhouettes, whereas TMM demonstrates high ASW scores for PDS Fuzzy, DBS Mean and Median compared to other algorithms. For Thyroid dataset Table 8, the TMM algorithm achieves the highest ARI value, indicating its appropriateness for Thyroid cluster data. In terms of ASW, GMM shows high scores for PACS, CeS, PPS, PDS Crisp, and PDS Fuzzy silhouettes, while TMM exhibits high ASW scores for PDS Fuzzy DBS Mean and Median compared to other algorithms.

For Tables 7 and 8, Spearman rank correlation is calculated between ARI and each silhouette ASW values and it is presented in Table 6.

The results from the correlation analysis in Table 9 demonstrate that DBS Median silhouette exhibits a notably strong correlation with ARI compared to other silhouette measures. The proposed silhouettes also exhibit higher correlations, which makes them suitable for scenarios where class labels are unknown in the two datasets.

5. Conclusion

In conclusion, this paper introduces innovative diagnostic tools for evaluating the performance and validity of a partition obtained through a soft clustering algorithm. The proposed method, akin to a density-based silhouette index, offers the added benefit of assessing various soft clustering techniques. By estimating the posterior probabilities of data points belonging to the identified clusters, the approach provides valuable insights into the quality of the clustering results.

Furthermore, the study extends the concept of evaluating classifier performance to clustering through the Average Silhouette Width, as suggested by Raymaekers and Rousseeuw (2022). The newly introduced silhouette measures consistently correlate with the classical ARI across all real and simulated datasets examined in this research. Moreover, the silhouette measures proposed in this article outperform the existing DBS method proposed by Menardi (2011) and excel in identifying the number of components within the datasets.

These findings highlight the efficacy and potential of the diagnostic tools presented in this paper, offering researchers and practitioners valuable means to assess and compare the performance of soft clustering algorithms. These tools advance the field and facilitate better decision-making in various domains by enhancing our understanding of clustering results.

References

Andrews, J. L., J. R. Wickins, N. M. Boers, and P. D. McNicholas. 2018. “teigen: An R Package for Model-Based Clustering and Classification via the Multivariate *t* Distribution.” *Journal of Statistical Software* 83 (7):1–32. doi:10.18637/jss.v083.i07.

Ben-Israel, A., and C. Iyigun. 2008. “Probabilistic d-clustering.” *Journal of Classification* 25:5–26. doi:10.1007/s00357-008-9002-z.

Bezdek, J. C., R. Ehrlich, and W. Full. 1984. “FCM: The Fuzzy C-Means Clustering Algorithm.” *Computers & Geosciences* 10 (2–3):191–203. doi:10.1016/0098-3004(84)90020-7.

Biernacki, C., G. Celeux, and G. Govaert. 2000. “Assessing a Mixture Model for Clustering with the Integrated Completed Likelihood.” *IEEE Transactions on Pattern Analysis and Machine Intelligence* 22 (7):719–725. doi:10.1109/34.865189.

Bouveyron, C., G. Celeux, T. Brendan Murphy, and A. E. Raftery. 2019. “Model-Based Clustering: Basic Ideas.” In *In Model-Based Clustering and Classification for Data Science: With Applications in R*. Cambridge Series in Statistical and Probabilistic Mathematics, 15–78. Cambridge: Cambridge University Press. doi:10.1017/9781108644181.003.

Campello, R. J., and E. R. Hruschka. 2006. “A Fuzzy Extension of the Silhouette Width Criterion for Cluster Analysis.” *Fuzzy Sets and Systems* 157 (21):2858–2875. doi:10.1016/j.fss.2006.07.006.

Cebeci, Z. 2019. “Comparison of Internal Validity Indices for Fuzzy Clustering.” *Journal of Agricultural Informatics* 10 (2):1–14. doi:10.17700/jai.2019.10.2.537.

Celeux, G., and G. Govaert. 1992. “A Classification EM Algorithm for Clustering and Two Stochastic Versions.” *Computational Statistics & Data Analysis* 14 (3):315–332. doi:10.1016/0167-9473(92)90042-E.

Coomans, D., I. Broeckaert, M. Jonckheer, and D. L. Massart. 1983. “Comparison of Multivariate Discrimination Techniques for Clinical Data—Application to the Thyroid Functional State.” *Methods of Information in Medicine* 22 (2):93–101. doi:10.1055/s-0038-1635425.

Davies, D. L., and D. W. Bouldin. 1983. “A Cluster Separation Measure.” *IEEE Transactions on Pattern Analysis and Machine Intelligence* 1: 224–227. doi:10.1109/TPAMI.1979.4766909.

Dempster, A. P., N. M. Laird, and D. B. Rubin. 1977. “Maximum Likelihood from Incomplete Data Via the *EM* Algorithm.” *Journal of the Royal Statistical Society: Series B (Methodological).*” 39 (1):1–22. doi:10.1111/j.2517-6161.1977.tb01600.x.

Dunn, J. C. 1974. “Well-Separated Clusters and Optimal Fuzzy Partitions.” *Journal of Cybernetics* 4 (1):95–104. doi:10.1080/01969727408546059.

Fisher, R. A. 1936. “The Use of Multiple Measurements in Taxonomic Problems.” *Annals of Eugenics* 7 (2):179–188. doi:10.1111/j.1469-1809.1936.tb02137.x.

Forgy, E. W. 1965. “Cluster Analysis of Multivariate Data: Efficiency vs Interpretability of Classifications.” *Biometrics* 21:768–769.

Fraley, C., and A. E. Raftery. 2002. “Model-Based Clustering, Discriminant Analysis and Density Estimation.” *Journal of the American Statistical Association* 97 (458):611–631. doi:10.1198/016214502760047131.

Fraley, C., and A. E. Raftery. 2002. “Bayesian Regularization for Normal Mixture Estimation and Model-Based Clustering.” *Journal of Classification* 24: 155–181. doi:10.1007/s00357-007-0004-5.

Hartigan, J. A., and M. A. Wong. 1979. “Algorithm AS 136: A K-means Clustering Algorithm.” *Applied Statistics* 28:100–108. doi:10.2307/2346830.

Hubert, L., and P. Arabie. 1985. “Comparing Partitions.” *Journal of Classification* 2 (1):193–218. doi:10.1007/BF01908075.

Iyigun, C., and A. Ben-Israel. 2008. “Probabilistic Distance Clustering Adjusted for Cluster Size.” *Probability in the Engineering and Informational Sciences* 22 (4):603–621. doi:10.1017/S0269964808000351.

Kaufman, L., and P. J. Rousseeuw. 1990. *Finding Groups in Data: An Introduction to Cluster Analysis*. Hoboken, NJ: Wiley. doi:10.1002/9780470316801.

Lloyd, S. 1957, 1982. “Least Squares Quantization in PCM.” *IEEE Transactions on Information Theory* 28 (2):129–137 doi:10.1109/TIT.1982.1056489.

MacQueen, J. 1965. “Some Methods for Classification and Analysis of Multivariate Observations.” *Berkeley Symposium on Mathematical Statistics and Probability*, Vol. 1, pp. 281–297.

Maitra, R., and V. Melnykov. 2010. “Simulating Data to Study Performance of Finite Mixture Modeling and Clustering Algorithms.” *The Journal of Computational and Graphical Statistics* 2 (19):354–376. doi:10.1198/jcgs.2009.08054

Melnykov, V., W.-C. Chen, and R. Maitra. 2012. “MixSim: An R Package for Simulating Data to Study Performance of Clustering Algorithms.” *Journal of Statistical Software* 51 (12):1–25. doi:10.18637/jss.v051.i12.

Menardi, G. 2011. “Density-based Silhouette Diagnostics for Clustering Methods.” *Statistics and Computing* 21: 295–308. doi:10.1007/s11222-010-9169-0.

McNicholas, P. D. 2016. “Model-based Clustering.” *Journal of Classification* 33 (3):331–373. doi:10.1007/s00357-016-9211-9.

Rand, W. M. 1971. “Objective Criteria for the Evaluation of Clustering Methods.” *Journal of the American Statistical Association* 66: 846–850. doi:10.1080/01621459.1971.10482356.

Raymaekers, J., and P. J. Rousseeuw. 2022. “Silhouettes and Quasi Residual Plots for Neural Nets and Tree-based Classifiers.” *Journal of Computational and Graphical Statistics* 31 (4):1332–1343. doi:10.1080/10618600.2022.2050249.

Rousseeuw, P. J. 1987. “Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis.” *Journal of Computational and Applied Mathematics* 20:53–65. doi:10.1016/0377-0427(87)90125-7.

Schwarz, G. 1978. “Estimating the Dimension of a Model.” *Annals of statistics* 6 (2):461–464. doi:10.1214/aos/1176344136.

Scrucca, L., M. Fop, T. B. Murphy, and A. E. Raftery. 2016. “mclust 5: Clustering, Classification and Density Estimation Using Gaussian Finite Mixture Models.” *The R Journal* 8 (1):289–317. doi:10.32614/RJ-2016-021.

Theodoridis, S., and K. Koutroumbas. 2009. *Pattern Recognition*, 4th ed. New York: Academic Press.

Tortora, C., P. D. McNicholas, and F. Palumbo. 2020. “A Probabilistic Distance Clustering Algorithm Using Gaussian and Student-t Multivariate Density Distributions.” *SN Computer Science* 1:1–22. doi:10.1007/s42979-020-0067-z.

Tortora, C., N. Vidales, F. Palumbo, T. Kalra, and P. D. McNicholas. 2020. "FPDclustering: PD-Clustering and Factor PD-Clustering." R package version 2.2 https://CRAN.R-project.org/package=FPDclustering.

Van der Laan, M., K. Pollard, and J. Bryan. 2003. "A New Partitioning Around Medoids Algorithm." *Journal of Statistical Computation and Simulation* 73 (8):575–584. doi:10.1080/0094965031000136012.