H.F.R.I call "Basic Research Funding (Horizontal Support for all Sciences)"
National Recovery and Resilience Plan
(Greece 2.0)

Project Name: Counterfactuals for Clustering: Explainability, Fairness and Quality

Acronym: FairXCluster

Project No: 15940

Deliverable D3.1: State-of-the-art on Clustering Quality Indices

The research project is implemented in the framework of H.F.R.I call "Basic Research Funding (Horizontal Support for all Sciences)" under the National Recovery and Resilience Plan "Greece2.0", funded by the European Union – NextGenerationEU (H.F.R.I project Number: 15940)

State-of-the-art on Clustering Quality Indices

Georgios Vardakas and Aristidis Likas

Department of Computer Science & Engineering
University of Ioannina, Greece

1 Introduction

Clustering is a fundamental unsupervised learning task, grouping data points based on their similarity. Clustering quality indices help measure how well clusters capture the data structure. These indices vary across clustering methods, including traditional clustering, graph-based, kernel-based, agglomerative, Gaussian Mixture Models (GMM), and Bayesian approaches. This report summarizes common indices for each of these clustering techniques.

2 External Indices

External indices use ground truth labels to measure clustering accuracy. These indices are typically used when known labels are available for evaluation.

2.1 Rand Index (RI)

The Rand Index (RI) [1] measures clustering quality by evaluating the agreement between the predicted clustering and the ground truth labels. For a dataset X with N data points, let $\mathcal{P}$ represent the predicted partition of the data into clusters and $\mathcal{G}$ represent the ground truth partition. The Rand Index is computed as:

$$\mathrm{RI} = \frac{a+b}{\binom{N}{2}}, \tag{1}$$

where a is the number of pairs of data points that are correctly assigned to the same cluster in both $\mathcal{P}$ and $\mathcal{G}$, and b is the number of pairs correctly assigned to different clusters in both partitions. The denominator $\binom{N}{2}$ is the number of possible pairs of data points.

The Rand Index ranges from 0 to 1, with higher values indicating greater agreement between the predicted clustering and the ground truth. A value of 1 corresponds to perfect clustering alignment.

2.2 Adjusted Rand Index (ARI)

The Adjusted Rand Index (ARI) [2] improves upon the Rand Index by accounting for the chance grouping of points into clusters. For a dataset X with N data points, the ARI is computed as:

$$\mathrm{ARI} = \frac{\mathrm{RI} - \mathbb{E}[\mathrm{RI}]}{\max(\mathrm{RI}) - \mathbb{E}[\mathrm{RI}]}, \tag{2}$$

where RI is the Rand Index, $\mathbb{E}[\mathrm{RI}]$ is the expected Rand Index under random clustering, and $\max(\mathrm{RI}) = 1$. The adjustment ensures that the ARI equals 0 when clustering results are no better than random assignments and 1 for perfect clustering alignment. Typically, probabilities are computed as normalized frequencies.

The ARI ranges from -1 to 1, where higher values indicate better clustering quality. Unlike the Rand Index, ARI is particularly useful when comparing clustering solutions with different numbers of clusters, as it corrects for chance alignments.

2.3 Mutual Information (MI)

Mutual Information (MI) measures the shared information between the predicted clustering and the ground truth labels. For a dataset X, let $\mathcal{P}$ be the predicted partition and $\mathcal{G}$ the ground truth partition. MI quantifies the dependency between $\mathcal{P}$ and $\mathcal{G}$:

$$\mathrm{MI}(\mathcal{P}, \mathcal{G}) = \sum_{C \in \mathcal{P}} \sum_{G \in \mathcal{G}} P(C, G) \log \frac{P(C, G)}{P(C)P(G)}, \tag{3}$$

where $P(C, G)$ is the joint probability of a data point belonging to both cluster C and group G, while $P(C)$ and $P(G)$ are the marginal probabilities. MI values range from 0 (no mutual information) to $\log(K)$ for perfect alignment, where K is the number of clusters.

2.4 Normalized Mutual Information (NMI)

Normalized Mutual Information (NMI) extends Mutual Information (MI) by scaling it to account for differences in the sizes of the predicted and ground truth partitions. The general definition of NMI is:

$$\mathrm{NMI}(\mathcal{P}, \mathcal{G}) = \frac{\mathrm{MI}(\mathcal{P}, \mathcal{G})}{\sqrt{H(\mathcal{P})H(\mathcal{G})}}, \tag{4}$$

where $H(\mathcal{P})$ and $H(\mathcal{G})$ denote the entropies of the predicted clustering $\mathcal{P}$ and the ground truth partition $\mathcal{G}$, respectively:

$$H(\mathcal{P}) = -\sum_{C \in \mathcal{P}} P(C) \log P(C), \quad H(\mathcal{G}) = -\sum_{G \in \mathcal{G}} P(G) \log P(G). \tag{5}$$

Several other normalized MI scores, each with different normalization scheme, are also used in practice:

$$\mathrm{NMI}(\mathcal{P}, \mathcal{G}) = \frac{\mathrm{MI}(\mathcal{P}, \mathcal{G})}{\min(H(\mathcal{P}), H(\mathcal{G}))}, \tag{6}$$

$$\mathrm{NMI}(\mathcal{P}, \mathcal{G}) = \frac{\mathrm{MI}(\mathcal{P}, \mathcal{G})}{\max(H(\mathcal{P}), H(\mathcal{G}))}, \tag{7}$$

and

$$\mathrm{NMI}(\mathcal{P}, \mathcal{G}) = \frac{\mathrm{MI}(\mathcal{P}, \mathcal{G})}{\frac{1}{2}\big(H(\mathcal{P}) + H(\mathcal{G})\big)}. \tag{8}$$

NMI values range from 0 to 1, where 0 indicates no correlation between the predicted and ground truth clusters, and 1 indicates perfect alignment. Each normalization scheme is symmetric with respect to $\mathcal{P}$ and $\mathcal{G}$, making NMI robust to differences in cluster sizes and suitable for comparing clustering solutions.

2.5 Adjusted Mutual Information (AMI)

Adjusted Mutual Information (AMI) [3] further refines MI by correcting for the overlap expected under random clustering. It is computed as:

$$\mathrm{AMI}(\mathcal{P}, \mathcal{G}) = \frac{\mathrm{MI}(\mathcal{P}, \mathcal{G}) - \mathbb{E}[\mathrm{MI}(\mathcal{P}, \mathcal{G})]}{\max(H(\mathcal{P}), H(\mathcal{G})) - \mathbb{E}[\mathrm{MI}(\mathcal{P}, \mathcal{G})]}, \tag{9}$$

where $\mathbb{E}[\mathrm{MI}(\mathcal{P}, \mathcal{G})]$ is the expected $\mathrm{MI}(\mathcal{P}, \mathcal{G})$ under a random model. AMI adjusts for both randomness and cluster size imbalance, making it particularly useful for comparing clustering solutions with varying numbers of clusters. AMI values range from -1 to 1, where 0 indicates results no better than random and 1 signifies perfect clustering alignment.

2.6 Fowlkes-Mallows Index (FMI)

The Fowlkes-Mallows Index (FMI) [4] evaluates clustering quality by measuring the geometric mean of precision and recall based on pairwise cluster comparisons. It is defined as:

$$\text{FMI} = \sqrt{\frac{\text{TP}}{\text{TP}+\text{FP}} \cdot \frac{\text{TP}}{\text{TP}+\text{FN}}}, \tag{10}$$

where:

- TP (true positives): The number of data point pairs that are correctly assigned to the same cluster in both the predicted clustering and the ground truth.
- FP (false positives): The number of pairs assigned to the same cluster in the predicted clustering but to different clusters in the ground truth.
- FN (false negatives): The number of pairs assigned to different clusters in the predicted clustering but to the same cluster in the ground truth.

The FMI ranges from 0 to 1, where higher values indicate better clustering quality. An FMI of 1 signifies perfect agreement between the clustering solution and the ground truth.

3 Internal Indices

Internal indices assess clustering quality using intra-cluster similarity (how close points within the same cluster are) and inter-cluster separation (how distinct different clusters are).

3.1 Silhouette Coefficient

The silhouette score [5] evaluates clustering quality by measuring compactness and separation in the clustering solution. Let $d(x_i, x_j)$ denote the distance between data points x_i and x_j. For a data point x_i, its silhouette score $s(x_i)$ combines two components: the average intra-cluster distance $a(x_i)$, reflecting cohesion, and the minimum average inter-cluster distance $b(x_i)$, capturing separation. These are defined as:

$$a(x_i) = \frac{1}{|C_I| - 1} \sum_{x_j \in C_I, i \neq j} d(x_i, x_j), \tag{11}$$

$$b(x_i) = \min_{J \neq I} \frac{1}{|C_J|} \sum_{x_j \in C_J} d(x_i, x_j), \tag{12}$$

where $|C_I|$ and $|C_J|$ are cluster sizes. The silhouette score for x_i is given by:

$$s(x_i) = \frac{b(x_i) - a(x_i)}{\max\{a(x_i), b(x_i)\}}, \quad -1 \leq s(x_i) \leq 1. \tag{13}$$

Higher $s(x_i)$ values indicate better cluster assignments, while lower or negative values indicate bad clustering solution. The overall silhouette score $S(X)$ for dataset X is given using micro-averaging strategy:

$$S(X) = \frac{1}{N} \sum_{i=1}^{N} s(x_i), \tag{14}$$

or macro-averaging strategy:

$$S(X) = \frac{1}{K} \sum_{k=1}^{K} \frac{1}{|C_i|} \sum_{x_i \in C_i} s(x_i). \tag{15}$$

In the overall silhouette score computation, micro-averaging assumes equal weight between data-points, while macro-averaging assumes equal weight between clusters [6].

3.2 Soft Silhouette

The soft silhouette score [7] extends the traditional silhouette score to probabilistic cluster assignments. For a dataset $X = \{x_1, \ldots, x_N\}$ partitioned into K clusters $C = \{C_1, \ldots, C_K\}$, with $P_{C_I}(x_i)$ denoting the probability of x_i belonging to cluster C_I ($\sum_{I=1}^{K} P_{C_I}(x_i) = 1$), the intra-cluster distance $a_{C_I}(x_i)$ is defined as the weighted average distance of x_i to all other points in C_I:

$$a_{C_I}(x_i) = \frac{\sum_{j=1}^{N} P_{C_I}(x_j) d(x_i, x_j)}{\sum_{j=1, j\neq i}^{N} P_{C_I}(x_j)}. \tag{16}$$

The inter-cluster distance $b_{C_I}(x_i)$ is the minimum weighted average distance of x_i to other clusters, computed as:

$$b_{C_I}(x_i) = \min_{J \neq I} \frac{\sum_{j=1}^{N} P_{C_J}(x_j) d(x_i, x_j)}{\sum_{j=1, j\neq i}^{N} P_{C_J}(x_j)} = \min_{J \neq I} a_{C_J}(x_i). \tag{17}$$

Using these, the silhouette value $s_{C_I}(x_i)$ for x_i within cluster C_I is given by:

$$s_{C_I}(x_i) = \frac{b_{C_I}(x_i) - a_{C_I}(x_i)}{\max\{a_{C_I}(x_i), b_{C_I}(x_i)\}}. \tag{18}$$

The soft silhouette score $sf(x_i)$ of a data point x_i is then the expected value of $s_{C_I}(x_i)$ over its cluster probabilities:

$$sf(x_i) = \sum_{I=1}^{K} P_{C_I}(x_i) s_{C_I}(x_i), \tag{19}$$

and the overall soft silhouette score for the dataset is:

$$Sf(X) = \frac{1}{N} \sum_{i=1}^{N} sf(x_i). \tag{20}$$

When cluster probabilities are one-hot vectors, corresponding to hard clustering, the soft silhouette score reduces to the typical silhouette score.

3.3 Davies-Bouldin Index (DBI)

The Davies-Bouldin Index (DBI) [8] assesses clustering quality by measuring the average similarity between each cluster and its most similar cluster. For K clusters $C = \{C_1, \ldots, C_K\}$, DBI is computed as:

$$\text{DBI} = \frac{1}{K} \sum_{i=1}^{K} \max_{j \neq i} \frac{\sigma_i + \sigma_j}{d_{ij}}, \tag{21}$$

where σ_i is the average distance between points in cluster C_i and its centroid, and d_{ij} is the distance between the centroids of clusters C_i and C_j.

Lower DBI values indicate better clustering quality, with smaller intra-cluster distances (σ_i) and larger inter-cluster separations (d_{ij}). The range of DBI is $[0, \infty)$, and a well-separated clustering solution achieves a lower DBI.

3.4 Dunn Index

The Dunn Index [9] evaluates clustering quality by comparing the minimum inter-cluster distance to the maximum intra-cluster distance, emphasizing well-separated and compact clusters. For K clusters $C = \{C_1, \ldots, C_K\}$, the Dunn Index is defined as:

$$D = \frac{\min\limits_{1 \leq i \neq j \leq K} d(C_i, C_j)}{\max\limits_{1 \leq k \leq K} \delta(C_k)}, \tag{22}$$

where $d(C_i, C_j)$ is the distance between clusters C_i and C_j, and $\delta(C_k)$ is the maximum distance between points within cluster C_k.

Higher Dunn Index values indicate better clustering, reflecting larger cluster separations and tighter intra-cluster cohesion. The index ranges from $[0, \infty)$, with higher values signaling well-defined clusters.

3.5 Calinski-Harabasz Index (Variance Ratio Criterion)

The Calinski-Harabasz Index [10] (also known as the Variance Ratio Criterion) evaluates clustering quality by comparing the between-cluster dispersion to the within-cluster dispersion. For a dataset X with N points partitioned into K clusters $C = \{C_1, \ldots, C_K\}$, the index is computed as:

$$\mathrm{CH} = \frac{\mathrm{trace}(S_B)}{\mathrm{trace}(S_W)} \cdot \frac{N - K}{K - 1}. \tag{23}$$

Here, S_B is the between-cluster scatter matrix and is defined as:

$$S_B = \sum_{k=1}^{K} |C_k|(\mu_k - \mu)(\mu_k - \mu)^\top, \tag{24}$$

where $|C_k|$ is the size of cluster C_k, μ_k is the centroid of C_k, and μ is the overall mean of the dataset.

The within-cluster scatter matrix S_W is defined as:

$$S_W = \sum_{k=1}^{K} \sum_{x_i \in C_k} (x_i - \mu_k)(x_i - \mu_k)^\top. \tag{25}$$

Higher values of the Calinski-Harabasz Index indicate better clustering, as they reflect greater between-cluster variability relative to within-cluster compactness. There is no upper limit, and higher values signify superior clustering solutions.

3.6 Gap statistic

The Gap Statistic [11] is an internal clustering validation method that determines the optimal number of clusters by comparing the clustering result to a random reference distribution. It assesses how much better a clustering structure is compared to a dataset with no apparent structure. Specifically, for a given number of clusters K, the Gap Statistic is defined as:

$$\mathrm{Gap}(K) = \mathbb{E}[\log W_K] - \log W_K, \tag{26}$$

where W_K is the within-cluster dispersion, computed as:

$$W_K = \sum_{k=1}^{K} \sum_{x_i \in C_k} \|x_i - \mu_k\|^2. \tag{27}$$

Here, μ_k is the center of cluster C_k, and the expectation $\mathbb{E}[\log W_K]$ represents the average value of $\log W_K$ over multiple datasets generated from a reference distribution under the null hypothesis (e.g., a uniform distribution). This reference distribution assumes no inherent clustering structure, allowing for a comparison between the observed clustering result and what would be expected by chance.

Given clustering solutions for several values of K, the best number of clusters is chosen as the smallest K for which:

$$\mathrm{Gap}(K) \geq \mathrm{Gap}(K + 1) - s_{K+1}, \tag{28}$$

where s_K is the standard deviation of $\log W_K$ from the reference distribution. It should be noted that higher Gap values indicate stronger clustering structures, as they suggest that the observed clustering deviates significantly from a random distribution.

3.7 Stability

Stability is an important criterion for assessing the reliability of a clustering solution [12]. When applied to slightly perturbed dataset versions, a stable clustering algorithm produces consistent results. This indicates robustness to noise and sampling variations. Stability analysis helps determine whether the identified clusters represent true underlying structures or are sensitive to minor changes in the data.

Formally, let C be the clustering obtained by applying a method on a dataset X, and let C' be the clustering obtained using that method on a perturbed version of X (e.g., through resampling, noise addition, or small data shifts). A stability measure $S(C, C')$ quantifies the similarity between these two clusterings. Common approaches for measuring stability include:

$$S(C, C') = \text{Adjusted Rand Index (ARI)}, \quad S(C, C') = \text{Normalized Mutual Information (NMI)}, \tag{29}$$

where higher values indicate greater consistency between the clusterings.

A well-clustered dataset should exhibit high stability across different perturbations, implying that the clusters are meaningful and not artifacts of the specific sample. Conversely, unstable clustering solutions suggest that the algorithm may be overly sensitive to small variations, potentially leading to unreliable results.

Stability analysis is particularly useful for model selection, helping to choose the optimal number of clusters and evaluate different clustering methods. A clustering method that produces solutions with high stability while maintaining good performance on internal or external clustering indices is generally preferred.

3.8 Unimodality

Unimodality [13] is a statistical property that characterizes a probability density function, indicating whether the distribution has a single peak. A univariate density function is considered unimodal if there exists a mode m such that the density is non-decreasing for values smaller than m and non-increasing for values greater than m. This property ensures that the data exhibit a single region of high density without statistically significant gaps. In contrast, multimodal distributions contain multiple local density maxima, suggesting the presence of multiple clusters.

In clustering, unimodality plays a crucial role in determining whether a dataset naturally contains a single-cluster structure or more. Traditional unimodality tests, such as Hartigan's Dip Test [14] and Silverman's Bandwidth Test [15], are commonly used to assess whether a dataset follows a unimodal or multimodal distribution. These tests provide statistical evidence on whether data should be separated into multiple clusters, making them valuable tools for guiding clustering algorithms. Several clustering approaches leverage unimodality as a criterion for defining clusters [16–18]. By integrating unimodality into clustering, methods have been developed that improve cluster definition and automatically estimate the number of clusters.

4 Simalarity-Based and Graph Clustering Internal Quality Indices

Graph clustering focuses on grouping nodes in a graph based on connectivity patterns.

4.1 Modularity

Modularity [19] is a widely used quality measure for evaluating clustering results, particularly in graph-based clustering and community detection. It quantifies how well a given partition separates a network into densely connected communities while minimizing inter-community connections. A higher modularity value indicates that the clustering structure captures meaningful relationships between nodes. For a network represented as a graph $G = (V, E)$ with N nodes and M edges, modularity is defined as:

$$Q = \frac{1}{2M} \sum_{i,j} \left(A_{ij} - \frac{k_i k_j}{2M} \right) \delta(C_i, C_j), \tag{30}$$

where A_{ij} is the adjacency matrix, k_i and k_j are the degrees of nodes i and j, $2M$ is the total number of edges in the graph and $\delta(C_i, C_j)$ is an indicator function.

Modularity measures how much better the observed clustering is compared to random graph partitioning. It ranges from -1 to 1, where values close to 1 indicate a strong community structure, while values near zero or negative suggest a weak or no clustering structure.

Modularity is widely used in community detection algorithms such as Louvain [20] and spectral clustering [21], helping to identify natural divisions in networks. However, it has limitations, such as a resolution limit, where it may fail to detect small communities in large networks. Despite these limitations, modularity remains a fundamental measure in graph-based clustering, balancing intra-cluster density and inter-cluster separation.

4.2 Inclusion

The Inclusion Criterion [22] is a similarity-based clustering quality measure that evaluates the balance between intra-cluster density and inter-cluster separation. Initially introduced for community detection in unweighted graphs, it has been extended to general clustering problems using arbitrary similarity matrices. For a given dataset represented as a similarity graph, inclusion quantifies how well a data point fits within its assigned cluster while being distinct from other clusters. The measure is computed per node v and is defined as:

$$I_v = \frac{1}{2}\left(\frac{E_v^{in}}{d_v} + \frac{E_v^{out}+1}{N-d_v}\right), \tag{31}$$

where E_v^{in} represents the number of edges connecting node v to other nodes within its cluster, while d_v is the degree of node v, representing the total number of edges. The term E_v^{out} refers to the number of missing edges between v and nodes outside its cluster, and N is the total number of nodes in the dataset.

The total inclusion score of a clustering partition is obtained by averaging the individual inclusion values across all nodes. This measure is designed to maximize both internal connectivity, ensuring that points within the same cluster are well-connected, and external separation, ensuring that points are distinct from other clusters.

Unlike modularity, inclusion explicitly accounts for missing edges to measure inter-cluster separation, making it particularly useful in similarity-based and graph-based clustering problems.

5 GMM Clustering Internal Quality Indices

In GMM clustering, a Gaussian Mixture Model is first trained on the dataset. Then we associate each mixture component with a cluster. For each data point, the posterior probability that it has been generated by each mixture component is computed. Finally, each data point is assigned to the cluster with maximum posterior probability [23].

The Akaike Information Criterion (AIC) [24] and the Bayesian Information Criterion (BIC) [25] are widely used model selection criteria that assess the goodness of fit of a clustering model while penalizing the complexity of the model. These criteria help to determine the optimal number of clusters by balancing model accuracy and simplicity.

The Akaike Information Criterion (AIC) is defined as:

$$\mathrm{AIC} = -2\log L + 2k, \tag{32}$$

where L is the likelihood of the model given the data, and k is the number of parameters. A lower AIC value indicates a better trade-off between model fit and complexity. AIC aims to minimize the information loss by favoring models that explain the data well while avoiding excessive parameters.

The Bayesian Information Criterion (BIC) is similar but introduces a stronger penalty for model complexity:

$$\mathrm{BIC} = -2\log L + k\log N, \tag{33}$$

where N is the number of data points. BIC penalizes models with more parameters more heavily than AIC, making it more conservative in selecting models with additional complexity.

In clustering, AIC and BIC are often used to compare models with different numbers of clusters, particularly in probabilistic clustering methods such as Gaussian Mixture Models (GMMs). The model with the lowest AIC or BIC is preferred, though BIC tends to favor simpler models due to its stronger penalty term. Although both criteria are useful for selecting the optimal number of clusters, BIC is generally preferred when the true number of clusters is assumed to be finite, whereas AIC is more flexible and less biased toward smaller models.

These criteria provide an objective and statistical approach to model selection, reducing the reliance on arbitrary clustering validation metrics. However, their effectiveness depends on the underlying model assumptions, and they may not perform well in non-probabilistic clustering settings.

6 MML

Minimum Message Length [26] (MML) is a Bayesian information-theoretic criterion used for model selection, including clustering evaluation. It is based on the principle that the best model is the one that enables the most compact encoding of both the data and the model itself. MML provides a formal balance between model complexity and goodness of fit by minimizing the total message length required to describe the data and the clustering model.

For a given clustering model M with parameters θ, MML estimates the total encoding cost as:

$$\text{MML} = L(M) + L(X|M), \tag{34}$$

where $L(M)$ is the length of encoding the model parameters, and $L(X|M)$ is the length of encoding the dataset given the model. The first term acts as a complexity penalty, discouraging overly complex models, while the second term represents the data fit, favoring models that describe the dataset efficiently.

In clustering, MML is commonly applied to Gaussian Mixture Models (GMMs) and other probabilistic clustering approaches, where it helps determine the optimal number of clusters by penalizing models that overfit the data. MML provides a rigorous, information-theoretic approach to selecting the best clustering model, unlike heuristic methods like the elbow method.

MML is closely related to other model selection criteria, such as the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC). However, it differs because it is derived from information theory rather than asymptotic approximations. As a Bayesian method, MML integrates prior knowledge and naturally accounts for uncertainty in parameter estimation.

By minimizing message length, MML ensures that the clustering model is both parsimonious and statistically robust, making it a powerful tool for unsupervised learning and clustering validation.

7 Predictive likelihood

Predictive likelihood [27] is a model evaluation criterion used to assess the generalizability of a clustering model by measuring how well it predicts new, unseen data. It is particularly useful in probabilistic clustering methods, such as Gaussian Mixture Models (GMMs) and Bayesian clustering approaches, where clusters are modeled using probability distributions.

Given a dataset $X = \{x_1, x_2, \ldots, x_N\}$ and a clustering model M trained on X, the predictive likelihood evaluates the likelihood of new test data X' under the learned model parameters θ:

$$\log P(X'|\theta) = \sum_{x_i \in X'} \log P(x_i|\theta). \tag{35}$$

A higher predictive likelihood indicates that the clustering model generalizes well to new data, suggesting that the discovered clusters effectively capture the underlying data structure.

In clustering, predictive likelihood is often used for model selection, particularly in Bayesian frameworks, where it is computed using techniques like cross-validation or marginal likelihood estimation. Unlike criteria such as AIC and BIC, which balance model fit and complexity, predictive

likelihood directly measures a model's performance on unseen data, making it a valuable tool for evaluating clustering stability and robustness.

However, computing predictive likelihood can be computationally expensive, especially for complex models. Approximate methods, such as leave-one-out likelihood estimation or Bayesian posterior predictive checks, are often used to make the computation more feasible. Despite these challenges, predictive likelihood remains a decisive criterion for assessing clustering quality, particularly in probabilistic and generative clustering models.

8 Marginal Likelihood (Bayesian GMM)

The marginal likelihood, also known as the model evidence, is a key quantity in Bayesian clustering that evaluates how well a probabilistic model explains the observed data while integrating over all possible parameter values. It is beneficial for model selection, as it balances model fit and complexity without requiring additional penalty terms like those in AIC or BIC.

Given a dataset X and a clustering model M with parameters θ, the marginal likelihood is defined as:

$$P(X|M) = \int P(X|\theta, M)P(\theta|M)d\theta. \tag{36}$$

This integral marginalizes the parameters θ using their prior distribution $P(\theta|M)$, ensuring that models with excessive complexity are automatically penalized by Bayesian Occam's razor. Models with too many parameters tend to spread their probability mass too thinly, leading to lower marginal likelihood values. In contrast, models that appropriately balance complexity and data fit achieve higher marginal likelihoods.

In clustering, marginal likelihood is commonly used to determine the optimal number of clusters in Bayesian methods, such as Gaussian Mixture Models (GMMs). Since direct computation of the marginal likelihood is often intractable, techniques like Laplace approximation, Bayesian Information Criterion (BIC) approximation, and Markov Chain Monte Carlo (MCMC) sampling are frequently used to estimate it.

Compared to predictive likelihood, which assesses a model's ability to generalize to new data, marginal likelihood evaluates how well a model fits the observed data while accounting for uncertainty in parameter estimation. This makes it a powerful tool for Bayesian model comparison in clustering applications.

9 Conclusions

This report provides a comprehensive review of clustering quality indices, covering both external and internal evaluation metrics. External indices, such as the Rand Index, Adjusted Mutual Information, and Fowlkes-Mallows Index, measure clustering performance based on ground truth labels. These indices are essential for benchmarking clustering algorithms when labeled data is available.

Internal indices, including the Silhouette Score, Davies-Bouldin Index, and Calinski-Harabasz Index, assess clustering quality without requiring ground truth labels. These measures evaluate intra-cluster cohesion and inter-cluster separation, helping to compare different clustering solutions. Additionally, stability and unimodality tests ensure that clustering structures are both consistent and statistically meaningful.

For similarity-based and graph clustering approaches, modularity and inclusion criteria provide specialized measures tailored to network data. In probabilistic clustering models, such as Gaussian Mixture Models (GMMs), statistical model selection criteria, including AIC, BIC, Minimum Message Length (MML), and Predictive Likelihood, offer robust approaches for determining the optimal number of clusters while balancing model complexity and data fit.

Overall, the selection of clustering quality indices depends on the specific characteristics of the dataset and the clustering method used. No single metric is universally superior; instead, a combination of indices often provides a more reliable assessment and this constitutes a commonly used approach in practice. For example in the case where supervised measures are used both NMI and ARI are typically provided to assess the quality of clustering results.

References

[1] W. M. Rand, "Objective criteria for the evaluation of clustering methods," *Journal of the American Statistical association*, vol. 66, no. 336, pp. 846–850, 1971.

[2] L. Hubert and P. Arabie, "Comparing partitions," *Journal of classification*, vol. 2, pp. 193–218, 1985.

[3] N. X. Vinh, J. Epps, and J. Bailey, "Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance," *Journal of Machine Learning Research*, vol. 11, no. 95, pp. 2837–2854, 2010.

[4] E. B. Fowlkes and C. L. Mallows, "A method for comparing two hierarchical clusterings," *Journal of the American statistical association*, vol. 78, no. 383, pp. 553–569, 1983.

[5] P. J. Rousseeuw, "Silhouettes: a graphical aid to the interpretation and validation of cluster analysis," *Journal of computational and applied mathematics*, vol. 20, pp. 53–65, 1987.

[6] J. Pavlopoulos, G. Vardakas, and A. Likas, "Revisiting silhouette aggregation," in *Discovery Science* (D. Pedreschi, A. Monreale, R. Guidotti, R. Pellungrini, and F. Naretto, eds.), (Cham), pp. 354–368, Springer Nature Switzerland, 2025.

[7] G. Vardakas, I. Papakostas, and A. Likas, "Deep clustering using the soft silhouette score: Towards compact and well-separated clusters," *arXiv preprint arXiv:2402.00608*, 2024.

[8] D. L. Davies and D. W. Bouldin, "A cluster separation measure," *IEEE transactions on pattern analysis and machine intelligence*, no. 2, pp. 224–227, 1979.

[9] J. C. Dunn, "Well-separated clusters and optimal fuzzy partitions," *Journal of cybernetics*, vol. 4, no. 1, pp. 95–104, 1974.

[10] T. Caliński and J. Harabasz, "A dendrite method for cluster analysis," *Communications in Statistics-theory and Methods*, vol. 3, no. 1, pp. 1–27, 1974.

[11] R. Tibshirani, G. Walther, and T. Hastie, "Estimating the number of clusters in a data set via the gap statistic," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 63, no. 2, pp. 411–423, 2001.

[12] U. Von Luxburg *et al.*, "Clustering stability: an overview," *Foundations and Trends® in Machine Learning*, vol. 2, no. 3, pp. 235–274, 2010.

[13] S. Dharmadhikari and K. Joag-Dev, *Unimodality, convexity, and applications*. Elsevier, 1988.

[14] J. A. Hartigan and P. M. Hartigan, "The dip test of unimodality," *The annals of Statistics*, pp. 70–84, 1985.

[15] B. W. Silverman, "Using kernel density estimates to investigate multimodality," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 43, no. 1, pp. 97–99, 1981.

[16] A. Kalogeratos and A. Likas, "Dip-means: an incremental clustering method for estimating the number of clusters," *Advances in neural information processing systems*, vol. 25, 2012.

[17] C. Leiber, L. G. Bauer, B. Schelling, C. Böhm, and C. Plant, "Dip-based deep embedded clustering with k-estimation," in *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pp. 903–913, 2021.

[18] G. Vardakas, A. Kalogeratos, and A. Likas, "Uniforce: The unimodality forest method for clustering and estimation of the number of clusters," *arXiv preprint arXiv:2312.11323*, 2023.

[19] M. E. Newman, "Modularity and community structure in networks," *Proceedings of the national academy of sciences*, vol. 103, no. 23, pp. 8577–8582, 2006.

[20] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, "Fast unfolding of communities in large networks," *Journal of statistical mechanics: theory and experiment*, vol. 2008, no. 10, p. P10008, 2008.

[21] U. Von Luxburg, "A tutorial on spectral clustering," *Statistics and computing*, vol. 17, pp. 395–416, 2007.

[22] N. Kornelakis and A. Likas, "The inclusion criterion for data clustering quality," in *Proceedings of the 13th Hellenic Conference on Artificial Intelligence*, pp. 1–4, 2024.

[23] G. McLachlan and D. Peel, *Finite Mixture Models*. Wiley series in probability and statistics: Applied probability and statistics, Wiley, 2004.

[24] H. Akaike, "A new look at the statistical model identification," *IEEE transactions on automatic control*, vol. 19, no. 6, pp. 716–723, 1974.

[25] G. Schwarz, "Estimating the dimension of a model," *The annals of statistics*, pp. 461–464, 1978.

[26] C. S. Wallace, *Statistical and inductive inference by minimum message length*. Springer Science & Business Media, 2005.

[27] J. F. Bjørnstad, "Predictive likelihood: A review," *Statistical Science*, pp. 242–254, 1990.