Discriminative ordering through ensemble consensus

Louis Ohl[1] **Fredrik Lindsten**[1]

[1]The Division of Statistics and Machine Learning, Department of Computer and Information Science, Linköping University, Linköping, Sweden

Abstract

Evaluating the performance of clustering models is a challenging task where the outcome depends on the definition of what constitutes a cluster. Due to this design, current existing metrics rarely handle multiple clustering models with diverse cluster definitions, nor do they comply with the integration of constraints when available. In this work, we take inspiration from consensus clustering and assume that a set of clustering models is able to uncover hidden structures in the data. We propose to construct a discriminative ordering through ensemble consensus based on the distance between the connectivity of a clustering model and the consensus matrix. We first validate the proposed method with synthetic scenarios, highlighting that the proposed score ranks the models that best match the consensus first. We then show that this simple ranking score significantly outperforms other scoring methods when comparing sets of different clustering algorithms that are not restricted to a fixed number of clusters and is compatible with clustering constraints.

1 INTRODUCTION

Clustering is an essential task in data analysis where one seeks to partition the observations of a dataset into K clusters. Due to its ill-posed nature, the design of a clustering algorithm requires hypotheses about what defines good clusters. Different hypotheses may lead to different clusters. In other words, cluster definition and methodology must be adapted to the context in which they are applied (Hennig, 2015).

Evaluating the quality of a clustering model is a complex problem that requires appropriate metrics. In an experimental setting, synthetic data can be generated according to hypotheses about the definition of clusters, allowing verification that a clustering algorithm recovers the expected partition. This verification can be done using *external* metrics such as the (unsupervised) accuracy, the normalised mutual information (NMI), or the adjusted Rand index (ARI, Hubert and Arabie, 1985). Conversely, in an exploratory context, *i.e.* when no labels are available, we rely on *internal* metrics that depend solely on the data observations and the model predictions, *e.g.* the variance-ratio criterion (Caliński and Harabasz, 1974), the silhouette score (Rousseeuw, 1987), or the integrated complete likelihood (Biernacki et al., 2000). Internal metrics are often built with a specific view on clustering hypotheses, and therefore must be used with algorithms that match those hypotheses.

Despite the large number of clustering metrics (Desgraupes, 2013; Charrad et al., 2014), there are few metrics that are suitable for comparing clustering models with different clustering hypotheses. In addition, and to the best of our knowledge, there is no clustering metric that can integrate constraints when targets are partially observed.

To compare different clustering algorithms independently of their clustering hypotheses, we take inspiration from consensus clustering (Strehl and Ghosh, 2002) and rank clustering algorithms according to their proximity to a consensus matrix. Our underlying hypothesis is that a diverse set of clustering algorithms will shed light on clusters whose observations are more frequently connected. Our contributions are:

- The proposal of a simple-to-compute and fast score for ranking clustering algorithms based on consensus clustering that is compatible with pairwise constraints regularisations.
- The first exploration of clustering ensembles as a mean of performing model selection, both for our metric and some baselines.
- An extensive benchmark including synthetic and real data for several internal metrics showing strong performances in favour of our metric.

2 RELATED WORKS

Evaluating the quality of a clustering model is a challenging task. In fact, the absence of a formal definition of what a cluster is leads to the absence of a definition of what quality is, and finding an objective measure of quality that allows comparison of different algorithms is challenging (Boley et al., 1999). For example, Han et al. (2012) define quality as "*[s]ome methods [that] measure how well the clusters fit the data set, while others measure how well the clusters match the ground truth*". These two categories can be referred to as internal metrics and external metrics.

2.1 EXTERNAL INDICES

In the presence of ground truth, *i.e.* targets, we can evaluate clustering models using external indices. Common evaluation metrics include the (unsupervised) accuracy, the NMI (Strehl and Ghosh, 2002) or the ARI (Hubert and Arabie, 1985). It is worth noting that NMI and ARI are preferable to unsupervised accuracy when the number of clusters in a model differs from the number of clusters in the targets, which is unknown in practice.

When we partially observe the targets, we can still use them as constraints, as in semi-supervised clustering (Bair, 2013; Cai et al., 2023). Two types of constraints can be distinguished: labels and must-link or cannot-link constraints. The former explicitly assign samples to a cluster, whereas the latter only indicate whether samples should be together or in different clusters, regardless of the cluster membership. Labels imply must-link and cannot-link constraints, but the reverse is not true. For example, we can evaluate the model using the pairwise recall, precision and F-measure (Basu et al., 2004), or the constrained Rand index (CRI, Klein et al., 2002). For both measures, the evaluation is restricted to the set of samples (or set of sample pairs) that are not affected by constraints. These metrics are external and require ground truth. Consequently, they cannot be used if we do not have access to labels other than those used to constrain the clustering algorithm. In the absence of such information, the approximate measure of informativeness (Davidson et al., 2006) could be preferred: it is simply the average number of constraints not satisfied by a clustering algorithm.

2.2 INTERNAL INDICES

We can distinguish two types of internal indices: those that integrate the clustering hypotheses from the model, and those that carry their own hypotheses about the definition of what makes good clusters.

If a model defines a tractable likelihood, we can use this value to reflect the fit between the model and the data. For example, the Akaike information criterion (AIC, Akaike, 1974, 1998) penalises the likelihood by the complexity of the model, expressed in terms of the number of free parameters. The Bayesian information criterion (BIC; Schwarz, 1978) weights this parameter penalty by the logarithm of the number of training samples. The integrated complete likelihood (ICL, Biernacki et al., 2000) extends the BIC in model-based clustering by distinguishing between model components in model-based clustering and their correspondence to clusters using an entropy penalty term.

If a model does not define a tractable likelihood, we may not have access to a fitness measure from the model and have to construct it post hoc. This is notably the case for discriminative clustering models, *e.g.* KMeans or DBSCAN (Ester et al., 1996). In this sense, the most well-known internal metric is the within-group sum of squares (WGSS, Edwards and Cavalli-Sforza, 1965), also known as the KMeans score. This score is efficient for clusters that are assumed to be concentrated around a centroid. However, ensuring that samples are concentrated around a centroid is not enough; a clear separation between clusters is also a desirable property. From this desire come criteria such as the variance-ratio criterion (Caliński and Harabasz, 1974), the Dunn index (Dunn, 1974) and its generalisations (Bezdek and Pal, 1998), which include the Davies-Bouldin index (Davies and Bouldin, 1979), the Silhouette score (Rousseeuw, 1987) or the PBM index (Pakhira et al., 2004). Internal metrics comparing the coherence of pairwise clustering have also been proposed. For instance, the Gamma index (Baker and Hubert, 1975) and the G+ index (Rohlf, 1974) are based on the notion of discordant and concordant pairs. A pair of samples from a similar cluster is concordant with another pair of samples from different clusters when their distance is shorter than the second pair. The two pairs are discordant if the distance is greater for dissimilar clusters than for similar clusters. To alleviate the requirement on the choice of distances, some of these scores were adapted for connectivity matrices (Saha and Bandyopadhyay, 2012) derived from relative neighbourhood graphs (Toussaint, 1980).

2.3 RANKING MODELS IN CONSENSUS CLUSTERING

If we restrict the goal of a clustering metric to comparing models, then the most relevant property is the ability of a metric to *rank* algorithms well. In a ranking context, we necessarily have several models: this allows us to use ensemble methods. Consensus clustering is an unsupervised ensemble clustering method that stems from classification ensembles (Strehl and Ghosh, 2002). The goal is to use several clustering algorithms, called base clusterings, and to combine their results into a single final clustering using a consensus function. Combining the results thus increases the quality of the clustering, in the sense of an evaluation using external labels.

Several works then developed some filtering criteria on

the base clusterings to improve the quality of consensus clustering. This field, sometimes called *ensemble clustering selection* (Golalipour et al., 2021), focuses on selecting a subset of base clusterings based on the belief that some of the base models hinder the global quality and should be discarded. The goal is to keep clusterings of quality while maintaining some diversity (Kuncheva and Hadjitodorov, 2004; Hadjitodorov et al., 2006; Fern and Lin, 2008). This selection can be done by keeping the base clusterings that are closest to the consensus result (Hong et al., 2009; Azimi and Fern, 2009; Jia et al., 2011). Although this introduces ordering between models, this ordering is non-deterministic as it relies on the outcome of the consensus which can be stochastic. Selection can also be achieved by solving a K-vertex subgraph problem on a graph, where edges are the similarities between pairs of base clusterings (Fern and Lin, 2008; Yang et al., 2017). However, such an approach does not introduce an order between models.

In some cases, this selection is made thanks to a ranking. Often, this ranking is done by interpolating between the quality and the diversity of each clustering algorithm (Fern and Lin, 2008; Naldi et al., 2013; Wang and Liu, 2018). The ranking then depends simultaneously on the definition of what is the quality of a base clustering and what diversity represents, in an internal metric sense, and on the interpolation coefficient. Thus, metrics for ranking clustering algorithms are not new, *but their purpose is different*. Therefore, and to the best of our knowledge, ranking clustering algorithms through consensus has never been used as a metric for selecting clustering models.

3 THE DISCOTEC

We seek to build a score for ranking clustering algorithms that simultaneously take into account the results of all algorithms and comply with must-link and cannot-link constraints. We start with the general unconstrained score, then detail how it can be simplified thanks to consensus binarising and finish with the addition of constraints.

3.1 THE UNCONSTRAINED SCORE

We assume that we have a set of $T \geq 3$ clustering models and a dataset of n unlabelled samples $\mathcal{D} = \{x_i\}_{i=1}^n$. Each clustering model t defines a partition of this dataset into K^t clusters: $\pi^t \in \{1, \dots, K^t\}^n$. Note that we only consider hard clusterings here, so that our method is compatible with any clustering algorithm, since soft clusterings can always be converted to hard ones.

We construct for each partition its respective connectivity matrix. Its entries are binary values indicating whether two samples were in the same cluster:

$$\boldsymbol{A}^t = \left[\mathbb{1}[\pi_i^t = \pi_j^t]\right]. \tag{1}$$

Table 1: Examples of formula for the distance D between connectiviy and consensus with different statistical distances.

	$D(\mathcal{B}(0)\|\mathcal{B}(\boldsymbol{C}_{ij}))$	$D(\mathcal{B}(1)\|\mathcal{B}(\boldsymbol{C}_{ij}))$
KL	$-\log(1-\boldsymbol{C}_{ij})$	$-\log \boldsymbol{C}_{ij}$
TV	$\boldsymbol{C}_{ij}$	$1-\boldsymbol{C}_{ij}$
H^2	$1-\sqrt{1-\boldsymbol{C}_{ij}}$	$1-\sqrt{\boldsymbol{C}_{ij}}$

We can then build the consensus matrix (Monti et al., 2003) by averaging all connectivity matrices:

$$\boldsymbol{C} = T^{-1} \sum_{t=1}^{T} \boldsymbol{A}^t. \tag{2}$$

The entries of the consensus matrix can be interpreted as parameters of Bernoulli distributions: they describe the probability that two samples end up in the same cluster according to the ensemble of models. The more often a pair of observations end up in the same cluster, the higher their consensus value. Consequently, we would like to identify a clustering that respects this trend. Conversely, when the consensus value is close to zero, we would like to select a clustering that did not link the two observations.

To order the clustering algorithms, we propose to measure the distance between their respective connectivity matrix and the consensus matrix. The smaller the distance, the better. We expect that the model with the lowest distance corresponds to a partition that best matches the consensus established by the ensemble. For an arbitrary distance or divergence D, *e.g.* the total variation distance or the KL divergence:

$$\mathcal{S}(\pi^t) = \sum_{i,j} D(\boldsymbol{A}_{ij}^t \| \boldsymbol{C}_{ij}). \tag{3}$$

Note that some combinations of inputs are impossible when computing D. We cannot have 0 (resp. 1) for connectivity $\boldsymbol{A}_{ij}$ and 1 (resp. 0) for the consensus $\boldsymbol{C}_{ij}$ because the consensus is an average. When both values are 0, or 1, the distance is necessarily 0. Therefore, the only distances we compute correspond to the cases where $\boldsymbol{C}_{ij} \in]0, 1[$. We summarise three examples of distances for this case, which we will use in experiments, in Table 1.

It is possible that this score favours solutions with too few or too many clusters. In fact, when clustering models tend to connect most of the samples together through large clusters, the ranking would favour solutions with few clusters because they minimise the number of terms $D(\mathcal{B}(0)\|\mathcal{B}(\boldsymbol{C}_{ij}))$, which incur a large penalty. Conversely, when most clustering models have a large number of clusters, the consensus matrix may become sparse or filled with very low values and the score would favour solutions with many clusters because they minimise the number of terms $D(\mathcal{B}(1)\|\mathcal{B}(\boldsymbol{C}_{ij}))$.

Algorithm 1 The binarised DISCOTEC

Require: A set of partitions $\pi^t \in \{1,\dots,K^t\}^n; t \in \{1,\dots,T\}$.
Require: Must-link constraints $\mathcal{C}_{\text{ML}} = \{(a_i, b_i)\}_{i=1}^{n_{\text{ML}}}$
Require: Cannot-link constraints $\mathcal{C}_{\text{CL}} = \{(a_i, b_i)\}_{i=1}^{n_{\text{CL}}}$
  for $t \in \{1,\dots,T\}$ **do**
    $\boldsymbol{A}^t \leftarrow \left[\mathbb{1}[\pi_i^t = \pi_j^t]\right]$ ▷ Connectivity matrices
  end for
  $\boldsymbol{C} \leftarrow T^{-1}\sum_{t=1}^{T}\boldsymbol{A}^t$ ▷ Consensus matrix
  $\mu \leftarrow n^{-2}\sum_{i,j}^{n}\boldsymbol{C}_{ij}$
  $\boldsymbol{Q} \leftarrow [\mathbb{1}[\boldsymbol{C}_{ij} \geq \mu]]$ ▷ Binarise the consensus
  for $t \in \{1,\dots,T\}$ **do**
    $\mathcal{S}^t \leftarrow n^{-2}\sum_{i,j}^{n}\left|\boldsymbol{Q}_{ij} - \boldsymbol{A}_{ij}^t\right|$ ▷ Score of model t
    $\mathcal{R}^t \leftarrow 0$ ▷ Regularisation by constraints
    for $(i,j) \in \mathcal{C}_{\text{ML}}$ **do**
      $\mathcal{R}^t \leftarrow \mathcal{R}^t + (1 - \boldsymbol{A}_{ij}^t)$
    end for
    for $(i,j) \in \mathcal{C}_{\text{CL}}$ **do**
      $\mathcal{R}^t \leftarrow \mathcal{R}^t + \boldsymbol{A}_{ij}^t$
    end for
    $\mathcal{S}^t \leftarrow \mathcal{S}^t + \frac{\mathcal{R}^t}{n_{\text{ML}}+n_{\text{CL}}}$ ▷ Regularised DISCOTEC
  end for

When the number of clusters varies from both extremes in the pool of clustering models, then the behaviour of the score would be in favour of solutions with many clusters because the consensus matrix gets low values.

In order to alleviate the limitation of having only high values or only low values in the consensus matrix, we propose to binarise it with respect to its mean:

$$\boldsymbol{Q} = \left[\mathbb{1}\left[\boldsymbol{C}_{ij} \geq n^{-2}\sum_{i'j'}\boldsymbol{C}_{i'j'}\right]\right]. \tag{4}$$

In this variant, we measure only the absolute differences between zeros and ones from both the connectivity and the consensus matrices. While this binarised consensus matrix is not compatible with the original perspective of statistical distances between two matrices, it can be interpreted as the ratio of mismatching connectivities between observations: the lower the better.

3.2 ADDING REGULARISATIONS

An important feature of the proposed score is its compatibility with the approximate measure of informativeness (Davidson et al., 2006), *i.e.* the average number of violated constraints. Given a set of n_{ML} must-link constraints $\mathcal{C}_{n_{\text{ML}}} = \{(a_i, b_i)\}_{i=1}^{n_{\text{ML}}}$, and a set of n_{CL} cannot-link constraints $\mathcal{C}_{n_{\text{CL}}} = \{(a_i, b_i)\}_{i=1}^{n_{\text{CL}}}$, this regularisation is:

$$\mathcal{R}(\pi^t) = \frac{\sum_{(a,b)\in\mathcal{C}_{\text{ML}}} D(\boldsymbol{A}_{ab}^t\|1) + \sum_{(a,b)\in\mathcal{C}_{\text{CL}}} D(\boldsymbol{A}_{ab}^t\|0)}{n_{\text{ML}} + n_{\text{CL}}}. \tag{5}$$

Both the regularisation and our score are contained in [0,1] and correspond to the sum of distances between a connectivity and a target value. Thus, both measures are compatible according to dimensional analysis.

We summarise the binarised version of the discriminative ordering through ensemble consensus (DISCOTEC) in Algorithm 1. We evaluate the computational complexity of this algorithm to $\mathcal{O}(T(n^2 + n_{\text{ML}} + n_{\text{CL}}))$ for T models and n observations.

We may note that the DISCOTEC scales linearly with the number of models. In comparison, the average NMI (ANMI, Strehl and Ghosh, 2002) and the average ARI, which were used for clustering ensemble selection (Fern and Lin, 2008), scale quadratically. These metrics consist in the average of a score between a partition and all other partitions, *e.g.* for ANMI:

$$\text{ANMI}(\pi^t) = \frac{1}{T-1}\sum_{t'\neq t}\text{NMI}(\pi^t, \pi^{t'}). \tag{6}$$

Consequently, evaluating the AARI or ANMI requires $T(T-1)/2$ pairwise computations, which becomes expensive when the number of models is large.

4 EXPERIMENTS

For our experiments, we incrementally moved from synthetic partitions to datasets and constraints integration. We first show that the DISCOTEC and other ensemble baselines perform on par on synthetic cases. We then introduce datasets and test both clustering algorithms with a fixed number of clusters or an unrestricted number of clusters. We highlight that the DISCOTEC has strong performances for the latter. Finally, we show that constraints can enhance the performance of DISCOTEC on real datasets, even with few constraints.

4.1 GENERAL PROTOCOL

To evaluate the DISCOTEC, we have borrowed the methodology of Vendramin et al. (2010, section 4). We first select a pool of $N_{\mathcal{D}}$ datasets, and for each dataset we apply T clustering algorithms. We then evaluate the correlation between an internal metric of interest and an external metric that describes how well a model matches some targets. A higher correlation value indicates that the ranking proposed by the internal metric is efficient in identifying the most relevant clustering. We report the average correlations over the $N_{\mathcal{D}}$ datasets. Note that we have negated all scores that should be minimised, so a positive correlation means good performance. We chose to show the Kendall's tau correlation (Kendall, 1945) in the paper because it measures how well two rankings compare. For extended results, including

the Pearson correlation as originally proposed by Vendramin et al. (2010), see Appendix D.

We distinguish two types of baselines: internal metrics that are also based on ensemble clustering and internal metrics that evaluate models individually using distances between observations. For the former, we use the ANMI and AARI, and emphasise that this is the first time the ranking properties of these metrics is studied. For the latter, we used clustering metrics available in the permetrics Python library (Thieu, 2024) and implemented ourselves some from the clusterCrit package (Desgraupes, 2013). For the sake of clarity, we have restricted all figures and tables to the ensemble metrics and the top-performing distance-based metrics where relevant. Extended tables with the 20 baselines can be found in Appendix D. Code can be found at: `https://github.com/oshillou/Discotec`.

4.2 SYNTHETIC PARTITIONS

We started by evaluating the DISCOTEC with synthetic partitions, which allowed us to control the difficulty of the consensus. We started by generating a ground truth of n observations and K clusters, then generated T different partitions trying to imitate the ground truth with some controlled accuracy. To that end, we sample for each observation a conservation indicator according to some probability ρ. If the observation is conserved, it keeps the same cluster as the ground truth. Otherwise, it is assigned to a different cluster than the ground truth.

We tested two synthetic scenarios: one with a uniform distribution of accuracies to the ground truth and one with unbalanced accuracies. For the first scenario, we uniformly sample a conservation threshold $\rho^t \in [0.1, \rho_{\text{max}}]$ for each model. This ensures that the models have a minimum accuracy of 10%, and an average maximum accuracy of ρ_{max}. For the second scenario, we first sample two partitions called *hubs*: one with $\rho = 0.2$ and one with $\rho = 0.9$. Then, we sample a fraction αT of the models with an accuracy in the range $[0.2, 0.9]$ to the first hub, and the remaining $(1-\alpha)T$ models with identical accuracy range to the second hub.

Since both of these scenarios do not have any underlying data samples on which we can measure distances, we can only evaluate the AARI, the ANMI and the DISCOTEC. We ran both scenarios with $n = 200$ samples and $K = 10$ clusters. In the first scenario, we varied T from 5 to 50 models. In the second scenario, we fixed $T = 50$. Each simulation was repeated 50 times. The results of the first scenario are shown in Figure 1 and of the second scenario in Figure 2. For the sake of readability, we only report the DISCOTEC with KL divergence and with binarised consensus in the figures, as the rankings using total variation distance and squared Hellinger distance followed the KL curve perfectly.

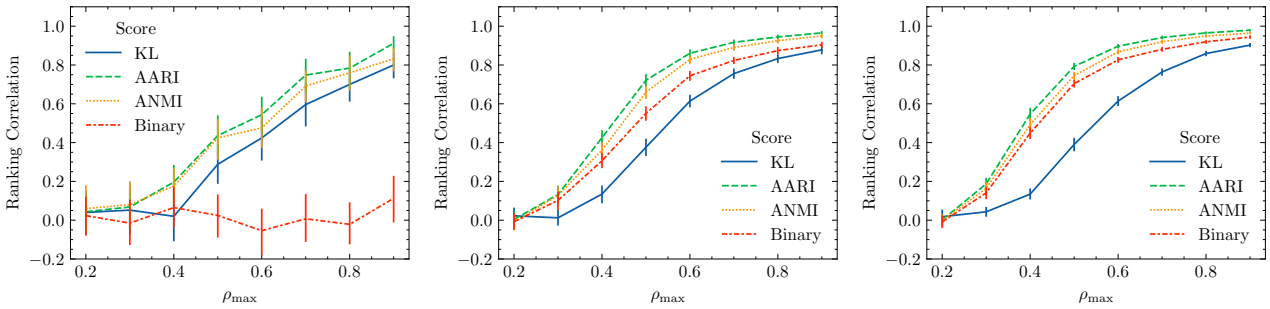
From the first scenario, we observe that increasing the maximum possible accuracy with ρ_{max} increases the correlation of the ranking. Indeed, when the maximum accuracy is low, most of the synthetic partitions tend to disagree with each other, resulting in a noisy consensus. Consequently, no pattern can emerge from the consensus matrix, and the DISCOTEC fails to correctly identify the correct clusterings. In contrast, if the maximum accuracy is high, a pattern can be seen in the consensus matrix, and the ranking can be coherent with this pattern. For completeness, we have included examples of such matrices in Appendix C. We can note in Figure 1 that the number of models is crucial to improve the performance of both baselines and DISCOTEC. Indeed, the correlation between the ARI of the partitions on the targets and the ranking of each method increases, and its standard deviation decreases from 5 to 50 models. This effect is even stronger for the binarised DISCOTEC. We further discuss and experiment with scaling within this scenario in Appendix B.

The success of the first scenario is due to the uniform distribution in terms of accuracy of all sampled models, but does not transfer to the second scenario. The second scenario highlights that both the DISCOTEC and the baselines are attracted to dominant hubs in terms of clustering solutions. Indeed, we can see in Figure 2 that when the partitions are close to a solution with high accuracy, *i.e.* $\alpha \approx 0$, then the ranking has a high correlation with the ARI. Conversely, increasing the number of models that are similar to a poor solution with very low accuracy, *i.e.* $\alpha = 1$, decreases the correlation for the same reason of noisy patterns as described above.

In summary, we have shown with these synthetic scenarios that ranking according to the relationships between models, both in baselines and the DISCOTEC, depends on two main factors: (i) the number of models and (ii) the distribution of the clustering ARIs. The number of models should preferably be large enough. However, too large a number of models is detrimental to the AARI and ANMI, which scale quadratically while the DISCOTEC scales linearly. Regarding the distribution of the clustering ARIs, we can expect better performance if it is more concentrated on solutions that are close enough to the ground truth and uniform. In other words: the diversity of base clusterings matters, in the sense of different cluster definitions.

4.3 SYNTHETIC AND REAL DATASETS CLUSTERING

To simulate more complex distributions of ARI with respect to targets, we now turn to different combinations of clustering models and datasets. We considered two different categories of datasets for our experiment: the fundamental clustering problem suite (FCPS, Thrun and Stier, 2021) and real datasets from the UCI repository, summarised in Ap-



(a) 5 models (b) 25 models (c) 50 models

Figure 1: Evolution of the average Kendall's tau (↑) correlation between ranking metrics and ARI with targets of synthetic partition when the label preservation rate $\rho_{\max}$ increases.

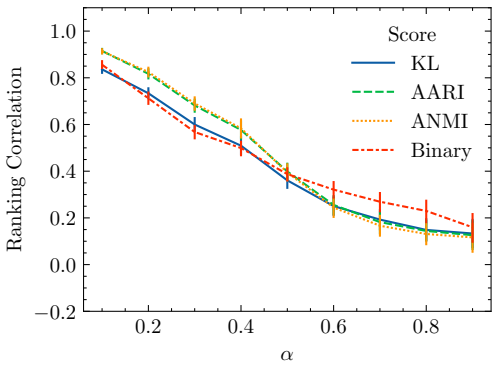


Figure 2: Evolution of Kendall's tau correlation (↑) of the selected models per ranking method as the interpolation α varies between highly accurate models ($\alpha = 0$) and non-accurate models ($\alpha = 1$).

Table 2: Average Kendall tau correlation$_{\text{std}}$ (↑) of the ranking metrics when the base clusterings are restricted to as many clusters as targets. KL and Binary respectively stand for the DISCOTEC with KL divergence, and consensus binarisation.

	FCPS		UCI	
Model	Agg.	Kmeans	Agg.	Kmeans
AARI	$0.48_{0.47}$	$0.06_{0.79}$	$-0.35_{0.60}$	$0.05_{0.37}$
ANMI	$0.57_{0.46}$	$0.04_{0.78}$	$-0.31_{0.54}$	$0.10_{0.46}$
CHI	$0.36_{0.51}$	$0.50_{0.54}$	$\mathbf{0.95_{0.11}}$	$-0.02_{0.40}$
SI	$0.07_{0.69}$	$\mathbf{0.68_{0.40}}$	$-0.45_{0.36}$	$0.25_{0.56}$
KL	$0.38_{0.63}$	$0.05_{0.79}$	$-0.52_{0.44}$	$0.05_{0.37}$
Binary	$\mathbf{0.73_{0.24}}$	$0.02_{0.91}$	$0.44_{0.65}$	$0.06_{0.46}$

pendix A. The FCPS consists of different simulated datasets in two or three dimensions, so that the definition of clusters is consensual to the naked eye. In contrast, the UCI datasets are intended for classification, which means that the classes and their number may not reflect the clusters and their number. Therefore, we must be careful in our interpretation of the ARI depending on the category of the datasets.

4.3.1 Restriction to a fixed number of clusters

Similarly to the synthetic scenarios, we restrict our experiments to clustering models that must find as many clusters as the number of clusters (resp. classes) indicated by the targets of the FCPS (resp. UCI) datasets. We try two different algorithms: KMeans and agglomerative clustering. We run KMeans 50 times. Since agglomerative clustering deterministically produces the same clustering, we vary its parameters using single, average, complete, and Ward linkage, and also Euclidean or Manhattan distance. This results in 7 models, because the Manhattan distance and the Ward linkage are incompatible.

Following the general protocol, we report the correlation in Table 2. Since the average correlation can be high due to some lucky runs, we extend our results by also reporting the regret score on the ARI of the top-ranked model for all methods. We define the regret score as the difference between the best performance of all methods and the performance of one method, which we average over all datasets. A lower regret score is better, and a regret score of 0 indicates that the method always had the best performance. We report the regret scores in Table 3. Regret scores on the correlation can be found in Appendix D.

These results complement the observations made previously in our second synthetic scenario. Indeed, we can see in Table 3 that the clusterings proposed by the agglomerative algorithms are more diverse than for KMeans since the ARI regret is up to 38% behind for some scores. This diversity leads to higher correlations compared to KMeans algorithms in Table 2. In contrast, the KMeans algorithms were attracted to specific clusterings that had a low ARI with respect to the targets for some datasets, leading to lower correlations. Furthermore, the lack of diversity between the

Table 3: Regret on the ARI of the model selected$_{\text{std}}$ (↓) by each ranking metric when the base clusterings are restricted to as many clusters as targets.

	FCPS		UCI	
Model	Agg.	Kmeans	Agg.	Kmeans
AARI	$0.16_{0.28}$	$0.06_{0.10}$	$0.31_{0.22}$	$0.05_{0.05}$
ANMI	$0.16_{0.28}$	$0.06_{0.10}$	$0.31_{0.22}$	$0.03_{0.03}$
CHI	$0.24_{0.33}$	$0.05_{0.09}$	$\mathbf{0.00_{0.00}}$	$0.06_{0.05}$
SI	$0.31_{0.37}$	$0.02_{0.04}$	$0.38_{0.25}$	$0.07_{0.08}$
KL	$0.26_{0.37}$	$0.06_{0.10}$	$0.32_{0.21}$	$0.05_{0.05}$
Binary	$0.15_{0.24}$	$0.05_{0.10}$	$0.17_{0.18}$	$0.05_{0.06}$

base clusterings and the regret on the selected model ARI is similar for all scores.

Among the compared baselines, we do not distinguish any score that offers a better ranking than any other for this experiment. We only mention both the silhouette index (SI, Rousseeuw, 1987) as an example and the strong success of the Calinski-Harabasz index (CHI, Caliński and Harabasz, 1974) for the UCI datasets with agglomerative clustering, highlighted by a high correlation in Table 2 and a regret score of 0 on the selected model ARI in Table 3.

4.3.2 Unrestricted pool of clustering models

We now extend the previous experiments by proposing a more diverse pool of clustering algorithms. We run KMeans clustering with K varying from 2 to 20 for each dataset, 5 times per value of K. We run agglomerative clustering with the same linkage parameters as before with Euclidean distances for 2 to 20 clusters. We then add DBSCAN models with parameter epsilon varying from the 1% quantile of the Euclidean distances of the dataset to the 25% quantile. We discard degenerate clusterings. Finally, we also evaluate the performance of the ranking methods when we merge all the models.

We observe in Table 4 that the average correlation is the highest for the DISCOTEC with binarised consensus matrix. Moreover, the ARI regret of the selected model is also the lowest in Table 5, which reveals better selection. In contrast, the DISCOTEC using the KL divergence did not perform better than the AARI baselines.

The performance of the DISCOTEC with KL divergence suffered precisely from the overclustering bias that motivated the introduction of the binarised consensus. The high proportion of models with a large number of clusters contributed to lowering the values of the consensus matrix, bringing them all close to 0. The KL ranking consequently favoured models with the largest number of clusters because they are less penalised when connecting as few observations as possible.

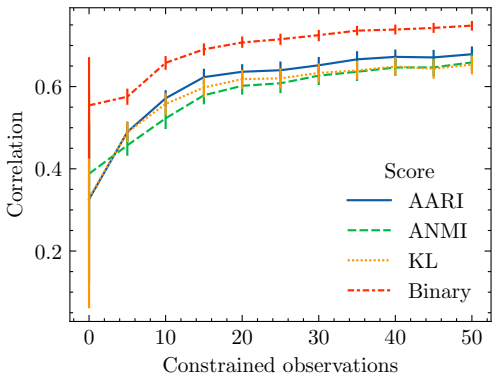


Figure 3: Kendall's tau correlation (↑) after adding must-link cannot-link constraints from an increasing number of random observations for the UCI datasets with mixed clustering models.

Finally, there is a notable difference between the FPCS and UCI datasets. For the former, we have the certainty that at least one of the proposed algorithms will achieve an ARI of 1 when merged together, because it matches the empirical definition of the clusters in a dataset. In contrast, the latter does not guarantee that the targets reflect clusters. Therefore, it is likely that the set of clustering algorithms will point to different clusters than the targets, sometimes with a different number of clusters compared to the number of classes. This accounts for the lower correlation values for the UCI datasets compared to the FCPS datasets in Table 4.

4.4 CONSTRAINT INTEGRATION

To make the most sense of the correlation between targets and clusters in the UCI datasets, we add constraints to the ranking. Thus, we simultaneously look for a model that captures clusters that correspond to what most models find, while also respecting the classes as much as possible.

We measured constraint satisfaction using the approximate measure of informativeness $\mathcal{R}$, and added it to both the DISCOTEC and the AARI/ANMI baselines. We chose not to add it to distance-based metrics because they do not correspond to the approximate measure of informativeness in terms of dimensional analysis. Moreover, the constraint regularisation is bounded in [0,1], whereas some other metrics are unbounded.

To assess the benefit of constraints, we report for each dataset the initial unconstrained correlation between targets and rankings using the same models from the previous experiment, and report the correlation after adding constraints. For each dataset, we randomly selected n observations and generated all must-link and cannot-link constraints they implied using the targets, and then evaluated the correlation of

Table 4: Average Kendall tau correlation$_{\text{std}}$ (↑) of the ranking metrics when the base clusterings seek different number of clusters.

	FCPS				UCI			
Model	Agg.	DBSCAN	Kmeans	Alltogether	Agg.	DBSCAN	Kmeans	Alltogether
AARI	$0.34_{0.32}$	$0.62_{0.48}$	$0.20_{0.36}$	$0.28_{0.42}$	$0.18_{0.33}$	$0.19_{0.55}$	$0.17_{0.30}$	$0.32_{0.30}$
ANMI	$0.17_{0.39}$	$0.63_{0.49}$	$0.03_{0.46}$	$0.14_{0.47}$	$0.30_{0.37}$	$0.16_{0.53}$	$-0.01_{0.38}$	$0.36_{0.22}$
CHI	$0.01_{0.36}$	$0.35_{0.77}$	$0.15_{0.51}$	$-0.04_{0.52}$	$0.39_{0.24}$	$0.40_{0.27}$	$0.33_{0.46}$	$0.36_{0.26}$
WGSS	$-0.41_{0.32}$	$-0.11_{0.93}$	$-0.68_{0.26}$	$-0.47_{0.32}$	$0.29_{0.54}$	$\mathbf{0.55_{0.36}}$	$-0.37_{0.49}$	$0.20_{0.38}$
KL	$0.25_{0.45}$	$0.57_{0.57}$	$-0.16_{0.46}$	$0.20_{0.50}$	$0.06_{0.40}$	$0.17_{0.51}$	$-0.11_{0.39}$	$0.31_{0.38}$
Binary	$\mathbf{0.79_{0.24}}$	$\mathbf{0.84_{0.24}}$	$\mathbf{0.82_{0.09}}$	$\mathbf{0.73_{0.36}}$	$\mathbf{0.63_{0.20}}$	$\mathbf{0.55_{0.30}}$	$\mathbf{0.47_{0.41}}$	$\mathbf{0.52_{0.21}}$

Table 5: Regret on the ARI of the model selected$_{\text{std}}$ (↓) by each ranking metric when the base clusterings seek different number of clusters.

	FCPS				UCI			
Model	Agg.	DBSCAN	Kmeans	Alltogether	Agg.	DBSCAN	Kmeans	Alltogether
AARI	$0.41_{0.32}$	$0.10_{0.17}$	$0.33_{0.27}$	$0.41_{0.32}$	$0.20_{0.24}$	$0.11_{0.13}$	$0.20_{0.19}$	$0.20_{0.25}$
ANMI	$0.41_{0.33}$	$0.10_{0.17}$	$0.33_{0.30}$	$0.40_{0.34}$	$0.17_{0.23}$	$0.12_{0.13}$	$0.19_{0.24}$	$0.14_{0.15}$
CHI	$0.24_{0.32}$	$0.07_{0.15}$	$0.16_{0.19}$	$0.27_{0.34}$	$0.13_{0.20}$	$0.06_{0.09}$	$0.15_{0.19}$	$0.12_{0.17}$
SI	$0.24_{0.34}$	$0.08_{0.17}$	$\mathbf{0.10_{0.14}}$	$0.31_{0.34}$	$0.40_{0.26}$	$0.19_{0.19}$	$0.14_{0.24}$	$0.41_{0.28}$
KL	$0.40_{0.32}$	$0.10_{0.17}$	$0.42_{0.28}$	$0.43_{0.31}$	$0.23_{0.25}$	$0.12_{0.13}$	$0.23_{0.23}$	$0.15_{0.21}$
Binary	$\mathbf{0.14_{0.20}}$	$\mathbf{0.06_{0.14}}$	$0.17_{0.21}$	$\mathbf{0.22_{0.24}}$	$\mathbf{0.11_{0.16}}$	$\mathbf{0.05_{0.06}}$	$\mathbf{0.09_{0.12}}$	$\mathbf{0.10_{0.14}}$

the regularised rankings. We repeated the constraint addition 50 times. Correlations before and after constraint addition can be found in Figure 3.

We observe that the addition of constraints is rarely detrimental to the correlation, as highlighted by the standard deviation decreasing from 0 constrained observations to 5 constrained observations. Moreover, increasing the number of constraints increases the correlation of the ranking with the targets. However, this increase is more substantial when introducing the first few constraints and tends to flatten afterwards. Nonetheless, we may note that 50 constrained observations is relatively small for some UCI datasets , *e.g.* Segmentation with 2310 observations.

5 CONCLUSION

We introduced a metric based on the distance between connectivity and consensus matrices to rank clustering algorithms, called the DISCOTEC. Overall, this metric works as intended and tends to select clustering models that are most similar to the consensus. We therefore suggest, as validated through experiments, that a diverse pool of clustering algorithms is required to get the most out of the DISCOTEC. In other words, the more the merrier when going to the disco.

We have shown experimentally that among several choices of distances, the most efficient is to binarise the consensus matrix with respect to its mean and compute its difference with the connectivity matrix. In general, the resulting performance is equal to or better than other ensemble clustering baselines such as the average ARI. The main difference with this baseline is that the DISCOTEC is faster to compute with respect to the number of models. Compared to other internal metrics, the advantage of the DISCOTEC is its tolerance to any type of clustering algorithm, *i.e.* definition of clusters. Consequently, the DISCOTEC shows better performance when the ranking a diverse set of clustering algorithms. In the case of a single clustering algorithm with limited parameters, a specialised internal metric may be preferred.

Finally, we have shown that the DISCOTEC can be regularised with must-link/cannot-link constraints thanks to the approximate measure of informativeness. Moreover, both methods are compatible from a dimensional analysis perspective because they average differences between edges of connectivity matrices.

In future work, it would be interesting to investigate how to further improve the performance of the DISCOTEC when the pool of base clusterings is not diverse. Additionally, it would be interesting to explore different approaches to the raw binarisation of the consensus matrix, *e.g.* a nonlinear bijection to obtain extreme values without being binary.

Acknowledgements

This research is financially supported by the Swedish Research Council via projects 2020-04122 and 2024-05011, the Knut and Alice Wallenberg Foundation via project KAW 2020.0033 and the Wallenberg AI, Autonomous Systems and Software Program (WASP), and the Excellence Center at Linköping–Lund in Information Technology (ELLIIT).

References

Hirotogu Akaike. Information Theory and an Extension of the Maximum Likelihood Principle. In *Selected Papers of Hirotugu Akaike*, pages 199–213. Springer, 1998.

Hirotugu Akaike. A New Look at the Statistical Model Identification. *Ieee Transactions on Automatic Control*, 19(6):716–723, 1974.

Javad Azimi and Xiaoli Fern. Adaptive Cluster Ensemble Selection. In *Proceedings of the 21st International Joint Conference on Artificial Intelligence*, IJCAI'09, pages 992–997, San Francisco, CA, USA, July 2009. Morgan Kaufmann Publishers Inc.

Eric Bair. Semi-supervised Clustering Methods. *Wiley Interdisciplinary Reviews: Computational Statistics*, 5(5): 349–361, 2013. Publisher: Wiley Online Library.

Frank B. Baker and Lawrence J. Hubert. Measuring the Power of Hierarchical Cluster Analysis. *Journal of the American Statistical Association*, 70(349):31–38, March 1975. ISSN 0162-1459. doi: 10.1080/01621459.1975.10480256. Publisher: ASA Website.

Geoffrey H Ball. Isodata, a Novel Method of Data Analysis and Pattern Classification. *Stanford Research Institute*, pages AD–699616, 1965.

Jeffrey D. Banfield and Adrian E. Raftery. Model-Based Gaussian and Non-Gaussian Clustering. *Biometrics*, 49 (3):803–821, 1993. ISSN 0006-341X. doi: 10.2307/2532201. Publisher: International Biometric Society.

Sugato Basu, Arindam Banerjee, and Raymond J. Mooney. Active Semi-Supervision for Pairwise Constrained Clustering. In *Proceedings of the 2004 SIAM International Conference on Data Mining (SDM)*, Proceedings, pages 333–344. Society for Industrial and Applied Mathematics, April 2004. ISBN 978-0-89871-568-2. doi: 10.1137/1.9781611972740.31.

J.C. Bezdek and N.R. Pal. Some New Indexes of Cluster Validity. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 28(3):301–315, June 1998. ISSN 1941-0492. doi: 10.1109/3477.678624. Conference Name: IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics).

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

Daniel Boley, Maria Gini, Robert Gross, Eui-Hong Sam Han, Kyle Hastings, George Karypis, Vipin Kumar, Bamshad Mobasher, and Jerome Moore. Partitioning-Based Clustering for Web Document Categorization. *Decision Support Systems*, 27(3):329–341, 1999. Publisher: Elsevier.

Jianghui Cai, Jing Hao, Haifeng Yang, Xujun Zhao, and Yuqing Yang. A Review on Semi-Supervised Clustering. *Information Sciences*, 632:164–200, 2023. ISSN 0020-0255. doi: 10.1016/j.ins.2023.02.088.

T. Caliński and J Harabasz. A Dendrite Method for Cluster Analysis. *Communications in Statistics*, 3(1): 1–27, January 1974. ISSN 0090-3272. doi: 10.1080/03610927408827101. Publisher: Taylor & Francis.

Malika Charrad, Nadia Ghazzali, Véronique Boiteau, and Azam Niknafs. NbClust: an R Package for Determining the Relevant Number of Clusters in a Data Set. *Journal of Statistical Software*, 61:1–36, 2014.

Ian Davidson, Kiri L Wagstaff, and Sugato Basu. Measuring Constraint-set Utility for Partitional Clustering Algorithms. In *European Conference on Principles of Data Mining and Knowledge Discovery*, pages 115–126. Springer, 2006.

David L. Davies and Donald W. Bouldin. A Cluster Separation Measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-1(2):224–227, April 1979. ISSN 1939-3539. doi: 10.1109/TPAMI.1979.4766909. Conference Name: IEEE Transactions on Pattern Analysis and Machine Intelligence.

Bernard Desgraupes. Clustering Indices. *University of Paris Ouest-Lab Modal'X*, 1(1):34, 2013. Publisher: Paris, France:.

Richard O. Duda and Peter E. Hart. Pattern Classification and Scene Analysis. In *A Wiley-Interscience publication*, 1974.

J. C. Dunn. Well-Separated Clusters and Optimal Fuzzy Partitions. *Journal of Cybernetics*, 4(1):95–104, January 1974. ISSN 0022-0280. doi: 10.1080/01969727408546059. Publisher: Taylor & Francis.

A. W. F. Edwards and L. L. Cavalli-Sforza. A Method for Cluster Analysis. *Biometrics*, 21(2):362–375, 1965. ISSN 0006-341X. doi: 10.2307/2528096. Publisher: Wiley, International Biometric Society.

Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. In *kdd*, volume 96, pages 226–231, 1996. Issue: 34.

Xiaoli Z Fern and Wei Lin. Cluster Ensemble Selection. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, 1(3):128–141, 2008. Publisher: Wiley Online Library.

Keyvan Golalipour, Ebrahim Akbari, Seyed Saeed Hamidi, Malrey Lee, and Rasul Enayatifar. From Clustering to Clustering Ensemble Selection: A Review. *Engineering Applications of Artificial Intelligence*, 104:104388, September 2021. ISSN 0952-1976. doi: 10.1016/j.engappai.2021.104388.

Stefan T. Hadjitodorov, Ludmila I. Kuncheva, and Ludmila P. Todorova. Moderate Diversity for Better Cluster Ensembles. *Information Fusion*, 7(3):264–275, 2006. ISSN 1566-2535. doi: 10.1016/j.inffus.2005.01.008.

Jiawei Han, Micheline Kamber, and Jian Pei. 10 - Cluster Analysis: Basic Concepts and Methods. In Jiawei Han, Micheline Kamber, and Jian Pei, editors, *Data Mining (Third Edition)*, The Morgan Kaufmann Series in Data Management Systems, pages 443–495. Morgan Kaufmann, Boston, third edition edition, 2012. ISBN 978-0-12-381479-1. doi: https://doi.org/10.1016/B978-0-12-381479-1.00010-1.

John A. Hartigan. *Clustering Algorithms*. Wiley, 1975. ISBN 978-0-471-35645-5. Google-Books-ID: cDnvAAAAMAAJ.

Christian Hennig. What are the true clusters? *Pattern Recognition Letters*, 64:53–62, 2015. ISSN 0167-8655. doi: 10.1016/j.patrec.2015.04.009.

Yi Hong, Sam Kwong, Hanli Wang, and Qingsheng Ren. Resampling-based Selective Clustering Ensembles. *Pattern Recognition Letters*, 30(3):298–305, February 2009. ISSN 0167-8655. doi: 10.1016/j.patrec.2008.10.007.

Lawrence Hubert and Phipps Arabie. Comparing Partitions. *Journal of Classification*, 2(1):193–218, 1985.

Lawrence Hubert and James Schultz. Quadratic Assignment as a General Data Analysis Strategy. *British Journal of Mathematical and Statistical Psychology*, 29(2):190–241, 1976. ISSN 2044-8317. doi: 10.1111/j.2044-8317.1976.tb00714.x.

Jianhua Jia, Xuan Xiao, Bingxiang Liu, and Licheng Jiao. Bagging-based Spectral Clustering Ensemble Selection. *Pattern Recognition Letters*, 32(10):1456–1467, July 2011. ISSN 0167-8655. doi: 10.1016/j.patrec.2011.04.008.

Maurice G Kendall. The Treatment of Ties in Ranking Problems. *Biometrika*, 33(3):239–251, 1945. Publisher: JSTOR.

Dan Klein, Sepandar D Kamvar, and Christopher D Manning. From Instance-Level Constraints to Space-Level Constraints: Making the Most of Prior Knowledge in Data Clustering. In *Proceedings of the Nineteenth International Conference on Machine Learning*, pages 307–314, 2002.

L.I. Kuncheva and S.T. Hadjitodorov. Using Diversity in Cluster Ensembles. In *2004 IEEE International Conference on Systems, Man and Cybernetics (IEEE Cat. No.04CH37583)*, volume 2, pages 1214–1219 vol.2, 2004. doi: 10.1109/ICSMC.2004.1399790.

F. H. C. Marriott. Practical Problems in a Method of Cluster Analysis. *Biometrics*, 27(3):501–514, 1971. ISSN 0006-341X. doi: 10.2307/2528592. Publisher: International Biometric Society.

John O. McClain and Vithala R. Rao. CLUSTISZ: A Program to Test for the Quality of Clustering of a Set of Objects. *Journal of Marketing Research*, 12(4):456–460, 1975. ISSN 0022-2437. Publisher: American Marketing Association.

Stefano Monti, Pablo Tamayo, Jill Mesirov, and Todd Golub. Consensus Clustering: A Resampling-Based Method for Class Discovery and Visualization of Gene Expression Microarray Data. *Machine Learning*, 52(1):91–118, July 2003. ISSN 1573-0565. doi: 10.1023/A:1023949509487.

M. C. Naldi, A. C. P. L. F. Carvalho, and R. J. G. B. Campello. Cluster Ensemble Selection Based on Relative Validity Indexes. *Data Mining and Knowledge Discovery*, 27(2):259–289, September 2013. ISSN 1384-5810, 1573-756X. doi: 10.1007/s10618-012-0290-x.

Malay K. Pakhira, Sanghamitra Bandyopadhyay, and Ujjwal Maulik. Validity Index for Crisp and Fuzzy Clusters. *Pattern Recognition*, 37(3):487–501, March 2004. ISSN 0031-3203. doi: 10.1016/j.patcog.2003.06.005.

F. James Rohlf. Methods of Comparing Classifications. *Annual Review of Ecology and Systematics*, 5:101–113, 1974. ISSN 0066-4162. Publisher: Annual Reviews.

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

Sriparna Saha and Sanghamitra Bandyopadhyay. Some Connectivity Based Cluster Validity Indices. *Applied Soft Computing*, 12(5):1555–1565, May 2012. ISSN 1568-4946. doi: 10.1016/j.asoc.2011.12.013.

Gideon Schwarz. Estimating the Dimension of a Model. *The Annals of Statistics*, 6(2):461 – 464, 1978. doi: 10.1214/aos/1176344136.

A. J. Scott and M. J. Symons. Clustering Methods Based on Likelihood Ratio Criteria. *Biometrics*, 27(2):387–397, 1971. ISSN 0006-341X. doi: 10.2307/2529003. Publisher: International Biometric Society.

Alexander Strehl and Joydeep Ghosh. Cluster Ensembles - a Knowledge Reuse Framework for Combining Multiple Partitions. *Journal of Machine Learning Research*, 3 (Dec):583–617, 2002.

Nguyen Van Thieu. PerMetrics: A Framework of Performance Metrics for Machine Learning Models. *Journal of Open Source Software*, 9(95):6143, March 2024. doi: 10.21105/joss.06143.

Michael C. Thrun and Quirin Stier. Fundamental Clustering Algorithms Suite. *SoftwareX*, 13:100642, 2021. ISSN 2352-7110. doi: 10.1016/j.softx.2020.100642.

Godfried T Toussaint. The Relative Neighbourhood Graph of a Finite Planar Set. *Pattern Recognition*, 12(4):261–268, 1980. Publisher: Elsevier.

Lucas Vendramin, Ricardo JGB Campello, and Eduardo R Hruschka. Relative Clustering Validity Criteria: A Comparative Overview. *Statistical Analysis and Data Mining: The Asa Data Science Journal*, 3(4):209–235, 2010. Publisher: Wiley Online Library.

Hongling Wang and Gang Liu. Two-Level-Oriented Selective Clustering Ensemble Based on Hybrid Multi-Modal Metrics. *IEEE Access*, 6:64159–64168, 2018. ISSN 2169-3536. doi: 10.1109/ACCESS.2018.2877666. Conference Name: IEEE Access.

Xuanli Lisa Xie and Gerardo Beni. A Validity Measure for Fuzzy Clustering. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 13(08):841–847, 1991. Publisher: IEEE Computer Society.

Fan Yang, Tao Li, Qifeng Zhou, and Han Xiao. Cluster Ensemble Selection with Constraints. *Neurocomputing*, 235:59–70, April 2017. ISSN 0925-2312. doi: 10.1016/j.neucom.2017.01.001.

Discriminative ordering through ensemble consensus (Supplementary Material)

Louis Ohl[1] **Fredrik Lindsten**[1]

[1]The Division of Statistics and Machine Learning, Department of Computer and Information Science, Linköping University, Linköping, Sweden

A SUPPLEMENTARY DETAILS FOR EXPERIMENTS

We used for our experiments two different types of datasets: synthetic datasets from the fundamental clustering problem suite (FCPS) and real datasets from the UCI repository. Their characteristics are summarised in Table 6.

Table 6: Datasets used in experiments

(a) FCPS datasets

Name	n	d	K (Clusters)
Atom	800	3	2
Chainlink	1000	3	2
EngyTime	4096	2	2
Hepta	212	3	7
LSun3D	404	3	4
Target	770	2	6
Tetra	400	3	4
TwoDiamonds	800	2	2
WingNut	1016	2	2

(b) UCI datasets

Name	n	d	K (Classes)
Dermatology	366	33	6
Digits	1797	64	10
Glass	214	9	6
Ionosphere	351	34	2
Iris	150	4	3
Lung	32	54	3
Segmentation	2310	19	7
WDBC	569	30	2
Wine	178	13	3

We used different metrics for evaluating the clustering models as baselines. We essentially took metrics available in the permetrics Python package (Thieu, 2024), and expanded with additional metrics described in some R packages (Desgraupes, 2013; Charrad et al., 2014), which we implemented ourselves. The name of the metrics thus mainly follow the permetrics name, and we propose a lexicon in Table 7.

B SCALING WITH THE NUMBER OF MODELS UNDER SYNTHETIC SCENARIOS

In this experiment, we further explore the relationship between the number models and the quality of the ranking. We keep the initial synthetic scenario from our first experiment in Section 4, where a ground truth is first generated and then T models are created by preserving between 10% and ρ_{max} of the labels. The resulting models have an accuracy bounded between 10% and ρ_{max} on average. For three specific thresholds $\rho_{\text{max}} \in \{0.2, 0.5, 0.9\}$, which correspond to a decreasing difficulty of consensus, we increase the number of models T from 5 models to 200. We report the average correlations for 50 runs per value of T and ρ_{max}. Figure 4 corresponds to the Pearson correlation and Figure 5 corresponds to Kendall's tau correlation coefficient. For clarity of both figures, we omitted the squared Hellinger and total variation distances because they perfectly followed the KL curve.

We observe from both figures that the scaling depends on the difficulty to reach a consensus. When a consensus is hard to

Table 7: Lexicon for the name of the metrics used in experiments. The accronyms are taken from the permetrics library (Thieu, 2024). Scores marked with an asterisk were re-implemented.

Short Name	Name	Reference
AARI	Average ARI	-
ANMI	Average NMI	Strehl and Ghosh, 2002
BHI	Ball-Hall index	Ball, 1965
BRI	Banfield-Raftery index	Banfield and Raftery, 1993
CHI	Calinski-Harabasz index	Caliński and Harabasz, 1974
CI*	C-index	Hubert and Schultz, 1976
DBI	Davies-Bouldin index	Davies and Bouldin, 1979
DHI	Duda-Hart index	Duda and Hart, 1974
DI*	Dunn index	Dunn, 1974
DRI	Det-ratio index	Scott and Symons, 1971
HI	Hartigan index	Hartigan, 1975
KDI	K squared determinant index	Marriott, 1971
LDRI	Log det ratio	Scott and Symons, 1971
LSRI	Log sum of squared error	-
McRao*	McClain-Rao index	McClain and Rao, 1975
PBM*	Pakhira-Bandyopadhyay-Maulik index	Pakhira et al., 2004
SI	Silhouette index	Rousseeuw, 1987
WGSS	Within-group sum of squares	Edwards and Cavalli-Sforza, 1965
XBI	Xie-Beni index	Xie and Beni, 1991
WG*	Wermmert-Gancarski index	Desgraupes, 2013

find, *i.e.* $\rho_{\max} = 0.2$, even 200 models is insufficient to establish a strong correlation between ranking and ARI with targets. In contrast, an easy scenario, *i.e.* $\rho_{\max} = 0.9$, requires few models to achieve excellent correlations as we are already close to 1 with 20 models. In a mitigated scenario, increasing the number of models increases steadily the correlations. It is notable that the binary DISCOTEC displays stronger performances even in a mitigated scenario compared to the DISCOTEC based on the KL distance.

We conclude that when the consensus is not clear-cut, adding models in the ensemble may be beneficial to the DISCOTEC ranking.

C CONSENSUS MATRIX VISUALISATION

We show in Figure 6 the examples of 3 consensus matrices from the first synthetic scenario in Section 4.

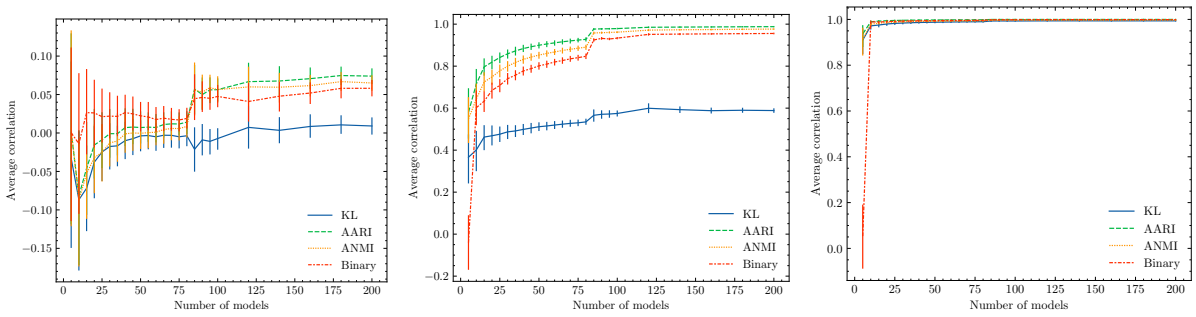
D EXTENDED BENCHMARK RESULTS

This section describes all extended tables of the experiments from Section 4. They concern the DISCOTEC performances and all baselines when testing different clustering algorithms on both the FCPS and UCI dataset.

D.1 COMPLETE TABLES WITH KENDALL'S TAU CORRELATION

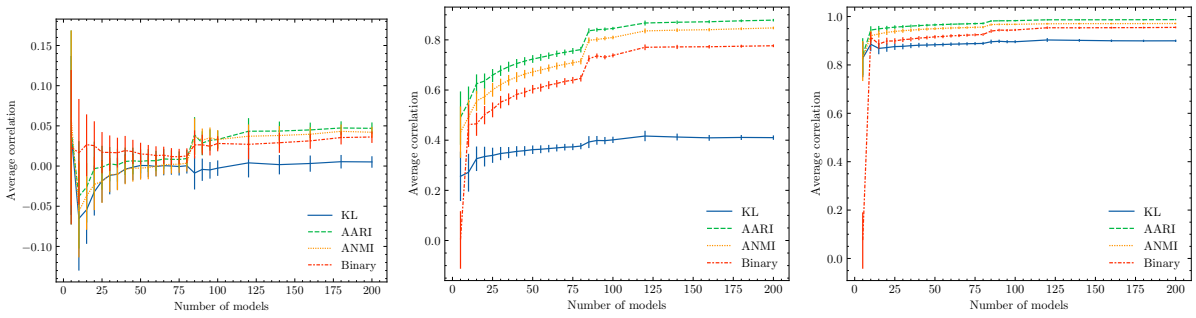
The tables 8 and 9 correspond to the performances of scoring methods when the base clusterings are restricted to as many clusters as the number specified by the targets of the dataset.

The tables 10 and 11 correspond to the extension of the experiment where clustering models cover different number of clusters.



(a) $\rho_{\max} = 0.2$ (b) $\rho_{\max} = 0.5$ (c) $\rho_{\max} = 0.9$

Figure 4: Pearson correlation between ranking metrics and ARI with ground truth as the number of models in the ensemble increase. Each model has an accuracy bounded between 10% and $\rho_{\max}$.



(a) $\rho_{\max} = 0.2$ (b) $\rho_{\max} = 0.5$ (c) $\rho_{\max} = 0.9$

Figure 5: Kendall's tau correlation between ranking metrics and ARI with ground truth as the number of models in the ensemble increase. Each model has an accuracy bounded between 10% and $\rho_{\max}$.

D.2 ADDITIONAL TABLES WITH PEARSON CORRELATION

To further complete the results, we show the exact same tables measuring the Pearson correlation instead of the Kendall's tau correlation. The tables 12 and 13 respectively correspond to the correlation and regret on correlation when the number of clusters is fixed. The tables 14 and 15 respectively correspond to the correlation and regret on correlation when the number of clusters varies between clustering algorithms.

D.3 COMPLETE TABLES FOR THE ARI REGRET OF SELECTED MODELS

We finally report the extended results for the regret on the ARI of the top-selected model per score in Table 17 when the number of clusters is restricted and Talbe 16 when the number of clusters can vary.

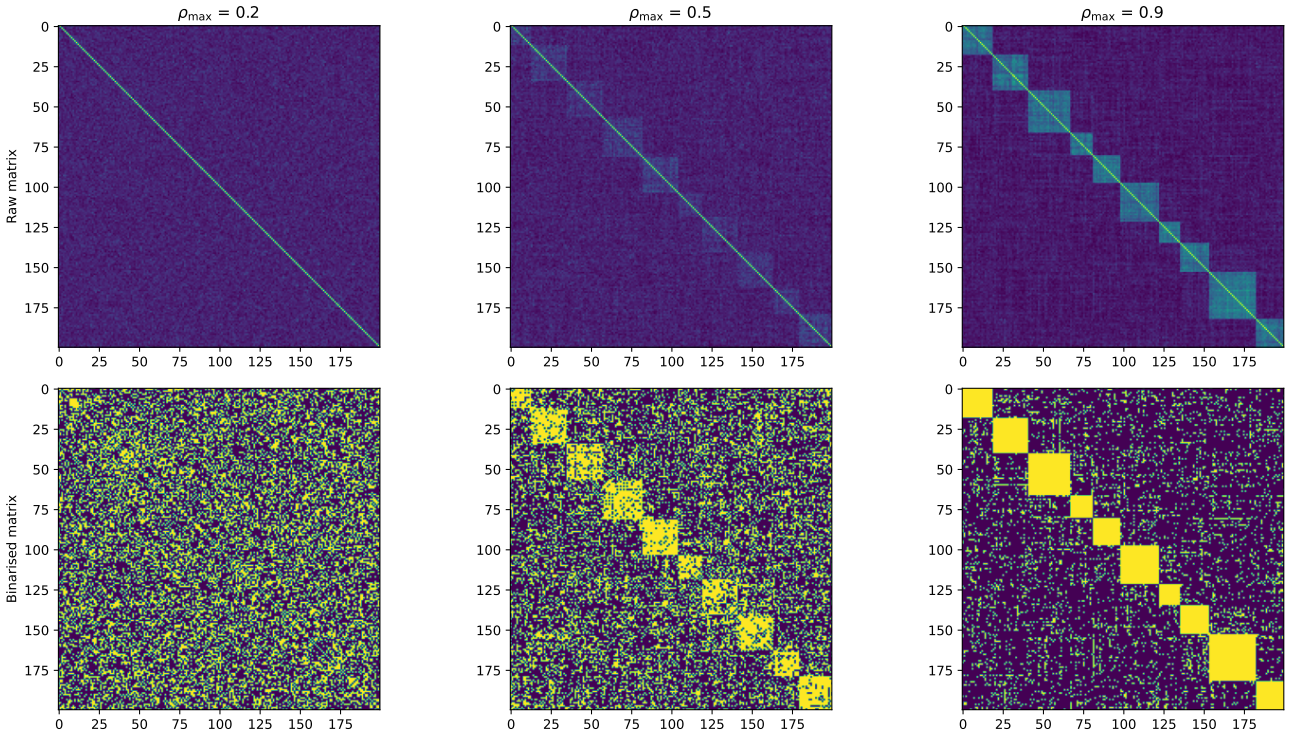


Figure 6: Example of consensus matrices from the first synthetic scenario when the upper bound on the label preservation rate $\rho_{\max}$ increases from 20% to 90% with the ground truth. The top row is shows the initial consensus matrix, the bottom row shows the same matrix after binarising with respect to its mean value.

Table 8: Extended results for the average Kendall tau correlation$_{\text{std}}$ (↑) of the ranking metrics when the base clusterings are restricted to as many clusters as targets.

	FCPS		UCI	
Model	Agglomerative	Kmeans	Agglomerative	Kmeans
AARI	$0.48_{0.47}$	$0.06_{0.79}$	$-0.35_{0.60}$	$0.05_{0.37}$
ANMI	$0.57_{0.46}$	$0.04_{0.78}$	$-0.31_{0.54}$	$0.10_{0.46}$
BHI	$0.36_{0.51}$	$0.47_{0.54}$	$\mathbf{0.95_{0.11}}$	$-0.02_{0.40}$
BRI	$0.63_{0.37}$	$0.41_{0.58}$	$0.86_{0.13}$	$0.21_{0.60}$
CHI	$0.36_{0.51}$	$0.50_{0.54}$	$\mathbf{0.95_{0.11}}$	$-0.02_{0.40}$
CI	$0.39_{0.37}$	$0.41_{0.52}$	$-0.35_{0.46}$	$-0.21_{0.45}$
DBI	$0.02_{0.77}$	$0.25_{0.69}$	$-0.80_{0.25}$	$0.27_{0.38}$
DHI	$-0.08_{0.54}$	$0.45_{0.56}$	$-0.80_{0.19}$	$0.16_{0.59}$
DI	$0.36_{0.58}$	$0.10_{0.73}$	$-0.68_{0.36}$	$0.08_{0.24}$
DRI	$0.21_{0.73}$	$0.42_{0.64}$	$0.55_{0.66}$	$0.29_{0.34}$
HI	$0.02_{0.79}$	$0.42_{0.55}$	$-0.83_{0.20}$	$0.33_{0.42}$
KDI	$-0.16_{0.56}$	$0.19_{0.62}$	$0.02_{0.56}$	$-0.01_{0.53}$
LDRI	$0.21_{0.73}$	$0.42_{0.64}$	$0.55_{0.66}$	$0.29_{0.34}$
LSRI	$0.36_{0.51}$	$0.52_{0.55}$	$\mathbf{0.95_{0.11}}$	$-0.02_{0.40}$
McRao	$0.26_{0.43}$	$0.38_{0.47}$	$-0.30_{0.56}$	$-0.18_{0.37}$
PBM	$-0.02_{0.63}$	$0.41_{0.69}$	$-0.32_{0.59}$	$0.09_{0.63}$
SI	$0.07_{0.69}$	$\mathbf{0.68_{0.40}}$	$-0.45_{0.36}$	$0.25_{0.56}$
WGSS	$0.36_{0.51}$	$0.51_{0.58}$	$\mathbf{0.95_{0.11}}$	$-0.02_{0.40}$
WG	$-0.25_{0.62}$	$0.39_{0.57}$	$-0.93_{0.12}$	$0.04_{0.47}$
XBI	$0.12_{0.67}$	$0.42_{0.60}$	$-0.61_{0.20}$	$0.14_{0.52}$
H	$0.42_{0.64}$	$0.05_{0.79}$	$-0.52_{0.44}$	$0.05_{0.37}$
KL	$0.38_{0.63}$	$0.05_{0.79}$	$-0.52_{0.44}$	$0.05_{0.37}$
TV	$0.42_{0.64}$	$0.05_{0.79}$	$-0.52_{0.44}$	$0.05_{0.37}$
Binary	$\mathbf{0.73_{0.24}}$	$0.02_{0.91}$	$0.44_{0.65}$	$0.06_{0.46}$

Table 9: Extended results for the regret on the Kendall tau correlation$_{\text{std}}$ (↓) of the ranking metrics when the base clusterings are restricted to as many clusters as targets.

	FCPS		UCI	
Model	Agglomerative	Kmeans	Agglomerative	Kmeans
AARI	$0.40_{0.48}$	$0.78_{0.84}$	$1.31_{0.65}$	$0.62_{0.50}$
ANMI	$0.30_{0.48}$	$0.81_{0.82}$	$1.28_{0.59}$	$0.57_{0.54}$
BHI	$0.52_{0.49}$	$0.37_{0.47}$	$\mathbf{0.01_{0.03}}$	$0.69_{0.53}$
BRI	$0.24_{0.28}$	$0.43_{0.54}$	$0.11_{0.13}$	$0.46_{0.60}$
CHI	$0.52_{0.49}$	$0.34_{0.46}$	$\mathbf{0.01_{0.03}}$	$0.69_{0.53}$
CI	$0.49_{0.33}$	$0.43_{0.45}$	$1.31_{0.52}$	$0.88_{0.66}$
DBI	$0.86_{0.81}$	$0.60_{0.69}$	$1.77_{0.33}$	$0.40_{0.22}$
DHI	$0.95_{0.59}$	$0.39_{0.55}$	$1.77_{0.24}$	$0.51_{0.45}$
DI	$0.52_{0.58}$	$0.74_{0.75}$	$1.65_{0.43}$	$0.59_{0.38}$
DRI	$0.67_{0.70}$	$0.42_{0.56}$	$0.39_{0.69}$	$0.47_{0.14}$
HI	$0.85_{0.84}$	$0.43_{0.53}$	$1.80_{0.25}$	$0.34_{0.19}$
KDI	$1.03_{0.54}$	$0.65_{0.63}$	$0.94_{0.52}$	$0.70_{0.66}$
LDRI	$0.67_{0.70}$	$0.42_{0.56}$	$0.39_{0.69}$	$0.47_{0.14}$
LSRI	$0.52_{0.49}$	$0.33_{0.46}$	$\mathbf{0.01_{0.03}}$	$0.69_{0.53}$
McRao	$0.61_{0.46}$	$0.46_{0.42}$	$1.27_{0.58}$	$0.85_{0.59}$
PBM	$0.89_{0.68}$	$0.44_{0.62}$	$1.29_{0.62}$	$0.59_{0.64}$
SI	$0.81_{0.73}$	$\mathbf{0.16_{0.23}}$	$1.42_{0.45}$	$0.42_{0.47}$
WGSS	$0.52_{0.49}$	$0.34_{0.50}$	$\mathbf{0.01_{0.03}}$	$0.69_{0.53}$
WG	$1.14_{0.64}$	$0.45_{0.57}$	$1.90_{0.21}$	$0.63_{0.43}$
XBI	$0.76_{0.72}$	$0.42_{0.51}$	$1.58_{0.26}$	$0.54_{0.59}$
H	$0.46_{0.70}$	$0.79_{0.83}$	$1.49_{0.52}$	$0.62_{0.50}$
KL	$0.50_{0.68}$	$0.79_{0.83}$	$1.49_{0.52}$	$0.62_{0.50}$
TV	$0.46_{0.70}$	$0.79_{0.83}$	$1.49_{0.52}$	$0.62_{0.50}$
Binary	$\mathbf{0.15_{0.21}}$	$0.82_{0.96}$	$0.52_{0.65}$	$0.62_{0.50}$

Table 10: Extended results for the average Kendall tau correlation$_{\text{std}}$ (↑) of the ranking metrics when the base clusterings seek different number of clusters.

	FCPS				UCI			
Model	Agg.	DBSCAN	Kmeans	Alltogether	Agg.	DBSCAN	Kmeans	Alltogether
AARI	$0.34_{0.32}$	$0.62_{0.48}$	$0.20_{0.36}$	$0.28_{0.42}$	$0.18_{0.33}$	$0.19_{0.55}$	$0.17_{0.30}$	$0.32_{0.30}$
ANMI	$0.17_{0.39}$	$0.63_{0.49}$	$0.03_{0.46}$	$0.14_{0.47}$	$0.30_{0.37}$	$0.16_{0.53}$	$-0.01_{0.38}$	$0.36_{0.22}$
BHI	$-0.34_{0.29}$	$-0.20_{0.83}$	$-0.68_{0.26}$	$-0.43_{0.30}$	$0.12_{0.45}$	$0.47_{0.39}$	$-0.37_{0.49}$	$0.10_{0.37}$
BRI	$-0.40_{0.31}$	$-0.19_{0.81}$	$-0.69_{0.25}$	$-0.43_{0.31}$	$0.19_{0.20}$	$0.33_{0.61}$	$-0.07_{0.26}$	$0.23_{0.23}$
CHI	$0.01_{0.36}$	$0.35_{0.77}$	$0.15_{0.51}$	$-0.04_{0.52}$	$0.39_{0.24}$	$0.40_{0.27}$	$0.33_{0.46}$	$0.36_{0.26}$
CI	$-0.06_{0.42}$	$-0.15_{0.89}$	$-0.35_{0.44}$	$-0.21_{0.48}$	$-0.12_{0.40}$	$0.07_{0.57}$	$-0.12_{0.43}$	$0.04_{0.29}$
DBI	$0.03_{0.38}$	$-0.23_{0.78}$	$0.02_{0.37}$	$-0.10_{0.40}$	$-0.38_{0.33}$	$-0.19_{0.34}$	$0.03_{0.29}$	$-0.08_{0.24}$
DHI	$-0.09_{0.26}$	$-0.35_{0.62}$	$-0.66_{0.26}$	$-0.29_{0.26}$	$-0.37_{0.15}$	$0.01_{0.43}$	$-0.31_{0.43}$	$-0.17_{0.23}$
DI	$0.11_{0.27}$	$0.30_{0.71}$	$-0.11_{0.21}$	$0.09_{0.27}$	$-0.48_{0.27}$	$-0.17_{0.57}$	$-0.01_{0.29}$	$-0.18_{0.31}$
DRI	$-0.42_{0.33}$	$-0.12_{0.92}$	$-0.68_{0.26}$	$-0.48_{0.33}$	$0.04_{0.50}$	$0.47_{0.30}$	$-0.49_{0.39}$	$0.08_{0.29}$
HI	$0.28_{0.36}$	$0.26_{0.65}$	$0.62_{0.28}$	$0.29_{0.34}$	$-0.29_{0.49}$	$-0.36_{0.48}$	$0.34_{0.53}$	$-0.19_{0.42}$
KDI	$0.16_{0.25}$	$0.29_{0.68}$	$0.63_{0.26}$	$0.30_{0.21}$	$-0.11_{0.37}$	$0.07_{0.51}$	$0.35_{0.31}$	$-0.03_{0.37}$
LDRI	$-0.42_{0.33}$	$-0.12_{0.92}$	$-0.68_{0.26}$	$-0.48_{0.33}$	$0.04_{0.50}$	$0.44_{0.38}$	$-0.49_{0.39}$	$0.05_{0.31}$
LSRI	$-0.41_{0.32}$	$-0.11_{0.93}$	$-0.68_{0.26}$	$-0.47_{0.32}$	$0.29_{0.54}$	$\mathbf{0.55_{0.36}}$	$-0.37_{0.49}$	$0.20_{0.38}$
McRao	$-0.39_{0.27}$	$-0.28_{0.81}$	$-0.66_{0.28}$	$-0.47_{0.30}$	$-0.28_{0.45}$	$-0.02_{0.63}$	$-0.30_{0.46}$	$-0.12_{0.38}$
PBM	$0.12_{0.35}$	$0.21_{0.74}$	$0.47_{0.47}$	$0.14_{0.47}$	$-0.01_{0.29}$	$-0.22_{0.63}$	$0.38_{0.26}$	$0.01_{0.29}$
SI	$0.07_{0.32}$	$0.17_{0.72}$	$0.32_{0.36}$	$0.01_{0.44}$	$-0.30_{0.37}$	$-0.18_{0.62}$	$0.23_{0.32}$	$-0.11_{0.26}$
WGSS	$-0.41_{0.32}$	$-0.11_{0.93}$	$-0.68_{0.26}$	$-0.47_{0.32}$	$0.29_{0.54}$	$\mathbf{0.55_{0.36}}$	$-0.37_{0.49}$	$0.20_{0.38}$
WG	$0.04_{0.24}$	$-0.38_{0.62}$	$-0.45_{0.32}$	$-0.18_{0.28}$	$-0.19_{0.18}$	$0.10_{0.46}$	$-0.21_{0.41}$	$-0.16_{0.22}$
XBI	$-0.12_{0.33}$	$0.14_{0.78}$	$0.09_{0.27}$	$-0.12_{0.34}$	$-0.29_{0.34}$	$-0.23_{0.60}$	$0.14_{0.26}$	$-0.16_{0.27}$
H	$0.24_{0.46}$	$0.57_{0.57}$	$-0.16_{0.46}$	$0.20_{0.50}$	$0.06_{0.41}$	$0.18_{0.51}$	$-0.11_{0.39}$	$0.31_{0.38}$
KL	$0.25_{0.45}$	$0.57_{0.57}$	$-0.16_{0.46}$	$0.20_{0.50}$	$0.06_{0.40}$	$0.17_{0.51}$	$-0.11_{0.39}$	$0.31_{0.38}$
TV	$0.24_{0.46}$	$0.57_{0.57}$	$-0.16_{0.46}$	$0.19_{0.50}$	$0.06_{0.41}$	$0.18_{0.51}$	$-0.11_{0.39}$	$0.31_{0.39}$
Binary	$\mathbf{0.79_{0.24}}$	$\mathbf{0.84_{0.24}}$	$\mathbf{0.82_{0.09}}$	$\mathbf{0.73_{0.36}}$	$\mathbf{0.63_{0.20}}$	$\mathbf{0.55_{0.30}}$	$\mathbf{0.47_{0.41}}$	$\mathbf{0.52_{0.21}}$

Table 11: Extended results for the regret on the Kendall tau correlation$_{\text{std}}$ (↓) of the ranking metrics when the base clusterings seek different number of clusters.

	FCPS				UCI			
Model	Agg.	DBSCAN	Kmeans	Alltogether	Agg.	DBSCAN	Kmeans	Alltogether
AARI	$0.47_{0.32}$	$0.26_{0.34}$	$0.65_{0.35}$	$0.55_{0.39}$	$0.55_{0.40}$	$0.64_{0.55}$	$0.46_{0.36}$	$0.31_{0.30}$
ANMI	$0.64_{0.42}$	$0.25_{0.36}$	$0.83_{0.46}$	$0.69_{0.48}$	$0.43_{0.40}$	$0.67_{0.54}$	$0.65_{0.45}$	$0.26_{0.13}$
BHI	$1.15_{0.41}$	$1.10_{0.86}$	$1.54_{0.30}$	$1.25_{0.39}$	$0.61_{0.43}$	$0.37_{0.40}$	$1.01_{0.65}$	$0.53_{0.33}$
BRI	$1.21_{0.43}$	$1.09_{0.86}$	$1.54_{0.29}$	$1.25_{0.42}$	$0.54_{0.23}$	$0.50_{0.62}$	$0.71_{0.48}$	$0.39_{0.24}$
CHI	$0.80_{0.35}$	$0.55_{0.74}$	$0.71_{0.47}$	$0.86_{0.51}$	$0.35_{0.25}$	$0.44_{0.34}$	$0.31_{0.42}$	$0.26_{0.26}$
CI	$0.87_{0.44}$	$1.05_{0.91}$	$1.21_{0.46}$	$1.03_{0.49}$	$0.85_{0.45}$	$0.76_{0.61}$	$0.76_{0.55}$	$0.58_{0.30}$
DBI	$0.78_{0.33}$	$1.13_{0.75}$	$0.83_{0.39}$	$0.93_{0.41}$	$1.12_{0.37}$	$1.03_{0.40}$	$0.61_{0.31}$	$0.70_{0.19}$
DHI	$0.90_{0.40}$	$1.25_{0.65}$	$1.51_{0.31}$	$1.11_{0.47}$	$1.11_{0.24}$	$0.82_{0.48}$	$0.94_{0.58}$	$0.79_{0.21}$
DI	$0.70_{0.37}$	$0.59_{0.62}$	$0.96_{0.25}$	$0.73_{0.27}$	$1.22_{0.31}$	$1.00_{0.60}$	$0.65_{0.43}$	$0.81_{0.40}$
DRI	$1.23_{0.43}$	$1.01_{0.95}$	$1.54_{0.31}$	$1.30_{0.40}$	$0.69_{0.56}$	$0.33_{0.28}$	$1.12_{0.72}$	$0.57_{0.31}$
HI	$0.53_{0.28}$	$0.63_{0.59}$	$0.24_{0.26}$	$0.53_{0.38}$	$1.02_{0.53}$	$1.19_{0.52}$	$0.29_{0.45}$	$0.81_{0.48}$
KDI	$0.65_{0.17}$	$0.61_{0.64}$	$0.23_{0.24}$	$0.53_{0.26}$	$0.88_{0.37}$	$0.77_{0.48}$	$0.34_{0.30}$	$0.68_{0.43}$
LDRI	$1.23_{0.43}$	$1.01_{0.95}$	$1.54_{0.31}$	$1.30_{0.40}$	$0.69_{0.56}$	$0.36_{0.36}$	$1.12_{0.72}$	$0.60_{0.29}$
LSRI	$1.22_{0.42}$	$1.01_{0.95}$	$1.53_{0.30}$	$1.29_{0.39}$	$0.44_{0.51}$	$\mathbf{0.28_{0.39}}$	$1.01_{0.65}$	$0.43_{0.35}$
McRao	$1.20_{0.38}$	$1.18_{0.81}$	$1.51_{0.34}$	$1.30_{0.39}$	$1.01_{0.49}$	$0.85_{0.68}$	$0.94_{0.62}$	$0.74_{0.39}$
PBM	$0.69_{0.27}$	$0.69_{0.72}$	$0.38_{0.45}$	$0.69_{0.52}$	$0.74_{0.34}$	$1.05_{0.68}$	$0.26_{0.16}$	$0.61_{0.33}$
SI	$0.74_{0.31}$	$0.72_{0.63}$	$0.54_{0.35}$	$0.81_{0.44}$	$1.04_{0.42}$	$1.01_{0.69}$	$0.40_{0.28}$	$0.74_{0.22}$
WGSS	$1.22_{0.42}$	$1.01_{0.95}$	$1.53_{0.30}$	$1.29_{0.39}$	$0.44_{0.51}$	$\mathbf{0.28_{0.39}}$	$1.01_{0.65}$	$0.43_{0.35}$
WG	$0.77_{0.36}$	$1.28_{0.64}$	$1.30_{0.34}$	$1.00_{0.42}$	$0.92_{0.21}$	$0.74_{0.48}$	$0.85_{0.53}$	$0.78_{0.20}$
XBI	$0.93_{0.36}$	$0.76_{0.71}$	$0.76_{0.33}$	$0.94_{0.38}$	$1.03_{0.37}$	$1.06_{0.66}$	$0.49_{0.25}$	$0.78_{0.27}$
H	$0.57_{0.48}$	$0.31_{0.44}$	$1.01_{0.49}$	$0.63_{0.51}$	$0.67_{0.49}$	$0.65_{0.52}$	$0.74_{0.52}$	$0.31_{0.39}$
KL	$0.56_{0.48}$	$0.31_{0.44}$	$1.01_{0.48}$	$0.62_{0.51}$	$0.67_{0.48}$	$0.66_{0.51}$	$0.74_{0.52}$	$0.31_{0.38}$
TV	$0.57_{0.48}$	$0.31_{0.44}$	$1.01_{0.49}$	$0.63_{0.51}$	$0.68_{0.49}$	$0.65_{0.52}$	$0.75_{0.52}$	$0.31_{0.39}$
Binary	$\mathbf{0.02_{0.07}}$	$\mathbf{0.05_{0.14}}$	$\mathbf{0.03_{0.05}}$	$\mathbf{0.09_{0.25}}$	$\mathbf{0.10_{0.20}}$	$0.29_{0.31}$	$\mathbf{0.17_{0.29}}$	$\mathbf{0.11_{0.13}}$

Table 12: Extended results for the average Pearson correlation$_{\text{std}}$ (↑) of the ranking metrics when the base clusterings are restricted to as many clusters as targets.

	FCPS		UCI	
Model	Agglomerative	Kmeans	Agglomerative	Kmeans
AARI	$0.72_{0.49}$	$0.03_{0.94}$	$-0.35_{0.78}$	$0.43_{0.54}$
ANMI	$0.80_{0.39}$	$-0.01_{0.86}$	$-0.29_{0.77}$	$0.49_{0.48}$
BHI	$0.41_{0.73}$	$0.57_{0.68}$	$0.92_{0.10}$	$0.37_{0.57}$
BRI	$0.63_{0.28}$	$0.33_{0.83}$	$0.87_{0.11}$	$0.30_{0.64}$
CHI	$0.41_{0.69}$	$0.56_{0.68}$	$\mathbf{0.93_{0.10}}$	$0.37_{0.57}$
CI	$0.39_{0.68}$	$0.57_{0.65}$	$-0.39_{0.53}$	$0.05_{0.67}$
DBI	$0.15_{0.79}$	$0.48_{0.69}$	$-0.78_{0.17}$	$0.02_{0.63}$
DHI	$0.05_{0.60}$	$0.56_{0.56}$	$-0.90_{0.11}$	$-0.15_{0.62}$
DI	$0.64_{0.45}$	$0.18_{0.76}$	$-0.84_{0.13}$	$-0.01_{0.39}$
DRI	$0.22_{0.85}$	$0.44_{0.86}$	$0.51_{0.84}$	$0.55_{0.41}$
HI	$0.13_{0.78}$	$0.47_{0.72}$	$-0.85_{0.13}$	$-0.03_{0.59}$
KDI	$-0.27_{0.69}$	$0.12_{0.81}$	$-0.08_{0.72}$	$-0.22_{0.52}$
LDRI	$0.22_{0.86}$	$0.44_{0.86}$	$0.55_{0.86}$	$\mathbf{0.61_{0.39}}$
LSRI	$0.35_{0.72}$	$0.58_{0.67}$	$0.92_{0.09}$	$0.37_{0.57}$
McRao	$0.33_{0.66}$	$0.56_{0.70}$	$-0.32_{0.74}$	$0.10_{0.63}$
PBM	$0.26_{0.74}$	$0.52_{0.61}$	$-0.38_{0.66}$	$0.06_{0.67}$
SI	$0.20_{0.79}$	$\mathbf{0.67_{0.56}}$	$-0.63_{0.37}$	$0.10_{0.72}$
WGSS	$0.41_{0.73}$	$0.57_{0.68}$	$0.92_{0.10}$	$0.37_{0.57}$
WG	$-0.13_{0.66}$	$0.48_{0.59}$	$-0.92_{0.06}$	$-0.08_{0.55}$
XBI	$0.23_{0.80}$	$0.59_{0.56}$	$-0.66_{0.25}$	$0.22_{0.60}$
H	$0.53_{0.71}$	$0.04_{0.92}$	$-0.51_{0.68}$	$0.47_{0.51}$
KL	$0.53_{0.70}$	$0.04_{0.91}$	$-0.51_{0.68}$	$0.47_{0.48}$
TV	$0.53_{0.72}$	$0.03_{0.92}$	$-0.51_{0.68}$	$0.46_{0.52}$
Binary	$\mathbf{0.85_{0.24}}$	$0.01_{0.97}$	$0.46_{0.66}$	$0.52_{0.56}$

Table 13: Extended results for the regret on the Pearson correlation$_{\text{std}}$ (↓) of the ranking metrics when the base clusterings are restricted to as many clusters as targets.

	FCPS		UCI	
Model	Agglomerative	Kmeans	Agglomerative	Kmeans
AARI	$0.25_{0.48}$	$0.91_{0.96}$	$1.33_{0.78}$	$0.33_{0.39}$
ANMI	$0.17_{0.39}$	$0.95_{0.87}$	$1.27_{0.77}$	$0.27_{0.37}$
BHI	$0.55_{0.71}$	$0.37_{0.67}$	$\mathbf{0.06_{0.09}}$	$0.39_{0.52}$
BRI	$0.34_{0.27}$	$0.61_{0.84}$	$0.11_{0.11}$	$0.46_{0.63}$
CHI	$0.55_{0.68}$	$0.38_{0.67}$	$\mathbf{0.06_{0.10}}$	$0.39_{0.52}$
CI	$0.57_{0.66}$	$0.37_{0.64}$	$1.38_{0.54}$	$0.71_{0.72}$
DBI	$0.82_{0.78}$	$0.46_{0.67}$	$1.77_{0.18}$	$0.73_{0.66}$
DHI	$0.92_{0.60}$	$0.38_{0.53}$	$1.88_{0.11}$	$0.91_{0.70}$
DI	$0.33_{0.44}$	$0.76_{0.74}$	$1.83_{0.14}$	$0.76_{0.48}$
DRI	$0.74_{0.83}$	$0.50_{0.86}$	$0.48_{0.85}$	$0.33_{0.29}$
HI	$0.84_{0.77}$	$0.47_{0.70}$	$1.83_{0.14}$	$0.78_{0.63}$
KDI	$1.24_{0.69}$	$0.82_{0.81}$	$1.07_{0.71}$	$1.00_{0.72}$
LDRI	$0.75_{0.84}$	$0.50_{0.86}$	$0.43_{0.87}$	$0.26_{0.28}$
LSRI	$0.61_{0.70}$	$0.36_{0.66}$	$0.07_{0.09}$	$0.39_{0.52}$
McRao	$0.63_{0.64}$	$0.38_{0.70}$	$1.31_{0.75}$	$0.66_{0.68}$
PBM	$0.70_{0.72}$	$0.42_{0.59}$	$1.36_{0.66}$	$0.69_{0.72}$
SI	$0.76_{0.77}$	$\mathbf{0.27_{0.54}}$	$1.61_{0.39}$	$0.66_{0.74}$
WGSS	$0.55_{0.71}$	$0.37_{0.67}$	$\mathbf{0.06_{0.09}}$	$0.39_{0.52}$
WG	$1.10_{0.66}$	$0.47_{0.57}$	$1.90_{0.06}$	$0.84_{0.64}$
XBI	$0.73_{0.79}$	$0.35_{0.53}$	$1.65_{0.25}$	$0.53_{0.59}$
H	$0.44_{0.72}$	$0.90_{0.93}$	$1.50_{0.68}$	$0.29_{0.38}$
KL	$0.44_{0.71}$	$0.90_{0.92}$	$1.50_{0.68}$	$0.29_{0.36}$
TV	$0.44_{0.72}$	$0.91_{0.94}$	$1.50_{0.68}$	$0.29_{0.40}$
Binary	$\mathbf{0.12_{0.24}}$	$0.93_{0.99}$	$0.52_{0.66}$	$\mathbf{0.24_{0.37}}$

Table 14: Extended results for the average Pearson correlation$_{\text{std}}$ (↑) of the ranking metrics when the base clusterings seek different number of clusters.

	FCPS				UCI			
Model	Agg.	DBSCAN	Kmeans	Alltogether	Agg.	DBSCAN	Kmeans	Alltogether
AARI	$0.40_{0.43}$	$0.59_{0.64}$	$0.00_{0.67}$	$0.33_{0.52}$	$0.20_{0.49}$	$0.21_{0.80}$	$0.13_{0.58}$	$0.43_{0.40}$
ANMI	$0.26_{0.48}$	$0.61_{0.65}$	$-0.08_{0.67}$	$0.20_{0.52}$	$0.26_{0.50}$	$0.26_{0.79}$	$-0.02_{0.65}$	$0.50_{0.31}$
BHI	$0.14_{0.41}$	$-0.08_{0.96}$	$-0.11_{0.59}$	$-0.05_{0.53}$	$0.17_{0.36}$	$0.54_{0.58}$	$-0.14_{0.62}$	$0.19_{0.55}$
BRI	$-0.40_{0.45}$	$-0.18_{0.89}$	$-0.41_{0.56}$	$-0.22_{0.28}$	$0.36_{0.26}$	$0.43_{0.66}$	$0.17_{0.33}$	$0.19_{0.23}$
CHI	$-0.05_{0.55}$	$0.43_{0.80}$	$0.27_{0.49}$	$-0.08_{0.62}$	$0.61_{0.33}$	$0.60_{0.37}$	$0.33_{0.60}$	$0.53_{0.35}$
CI	$0.02_{0.62}$	$-0.13_{0.92}$	$-0.03_{0.64}$	$-0.14_{0.65}$	$-0.07_{0.51}$	$0.30_{0.77}$	$0.00_{0.60}$	$0.28_{0.30}$
DBI	$0.06_{0.57}$	$-0.34_{0.78}$	$0.25_{0.45}$	$-0.16_{0.50}$	$-0.49_{0.44}$	$-0.23_{0.42}$	$0.16_{0.42}$	$-0.15_{0.36}$
DHI	$0.19_{0.41}$	$-0.37_{0.67}$	$-0.35_{0.52}$	$-0.09_{0.46}$	$-0.55_{0.23}$	$0.09_{0.61}$	$-0.27_{0.56}$	$-0.23_{0.30}$
DI	$0.20_{0.48}$	$0.23_{0.81}$	$0.07_{0.36}$	$0.26_{0.44}$	$-0.56_{0.29}$	$-0.10_{0.70}$	$0.06_{0.34}$	$-0.26_{0.28}$
DRI	$-0.36_{0.29}$	$-0.05_{0.94}$	$-0.53_{0.28}$	$-0.41_{0.23}$	$-0.14_{0.39}$	$0.62_{0.39}$	$-0.36_{0.34}$	$-0.12_{0.29}$
HI	$0.21_{0.44}$	$0.30_{0.63}$	$0.57_{0.33}$	$0.24_{0.42}$	$-0.33_{0.55}$	$-0.50_{0.48}$	$0.32_{0.62}$	$-0.16_{0.50}$
KDI	$0.05_{0.35}$	$0.29_{0.74}$	$0.54_{0.30}$	$0.23_{0.27}$	$-0.05_{0.45}$	$-0.01_{0.55}$	$0.26_{0.22}$	$0.06_{0.26}$
LDRI	$-0.20_{0.55}$	$-0.04_{1.01}$	$-0.45_{0.53}$	$-0.32_{0.54}$	$0.15_{0.66}$	$0.71_{0.48}$	$-0.46_{0.46}$	$0.09_{0.48}$
LSRI	$-0.06_{0.59}$	$-0.09_{0.96}$	$-0.35_{0.55}$	$-0.21_{0.60}$	$0.43_{0.63}$	$0.75_{0.57}$	$-0.24_{0.65}$	$0.32_{0.59}$
McRao	$-0.13_{0.52}$	$-0.26_{0.87}$	$-0.26_{0.59}$	$-0.26_{0.55}$	$-0.16_{0.61}$	$0.27_{0.73}$	$-0.19_{0.60}$	$0.08_{0.46}$
PBM	$0.00_{0.50}$	$0.26_{0.71}$	$0.45_{0.56}$	$0.07_{0.55}$	$-0.15_{0.27}$	$-0.08_{0.76}$	$0.47_{0.31}$	$-0.04_{0.31}$
SI	$-0.08_{0.56}$	$0.15_{0.77}$	$\mathbf{0.72_{0.36}}$	$-0.10_{0.63}$	$-0.34_{0.38}$	$-0.06_{0.69}$	$0.42_{0.45}$	$-0.12_{0.35}$
WGSS	$-0.03_{0.60}$	$-0.07_{0.99}$	$-0.23_{0.58}$	$-0.18_{0.62}$	$0.45_{0.62}$	$\mathbf{0.76_{0.58}}$	$-0.24_{0.64}$	$0.33_{0.57}$
WG	$0.38_{0.38}$	$-0.33_{0.70}$	$-0.32_{0.45}$	$0.10_{0.41}$	$-0.34_{0.20}$	$0.08_{0.60}$	$-0.23_{0.48}$	$-0.23_{0.19}$
XBI	$-0.34_{0.45}$	$0.12_{0.74}$	$0.31_{0.26}$	$-0.27_{0.42}$	$-0.38_{0.35}$	$-0.18_{0.67}$	$0.22_{0.40}$	$-0.10_{0.35}$
H	$0.26_{0.55}$	$0.51_{0.73}$	$-0.12_{0.69}$	$0.25_{0.58}$	$-0.12_{0.67}$	$0.13_{0.82}$	$-0.07_{0.63}$	$0.39_{0.57}$
KL	$0.30_{0.50}$	$0.50_{0.74}$	$-0.10_{0.68}$	$0.29_{0.54}$	$-0.12_{0.66}$	$0.13_{0.81}$	$-0.05_{0.61}$	$0.38_{0.56}$
TV	$0.24_{0.57}$	$0.51_{0.73}$	$-0.13_{0.69}$	$0.23_{0.60}$	$-0.13_{0.68}$	$0.13_{0.82}$	$-0.08_{0.63}$	$0.39_{0.57}$
Binary	$\mathbf{0.80_{0.25}}$	$\mathbf{0.90_{0.19}}$	$\mathbf{0.72_{0.21}}$	$\mathbf{0.70_{0.37}}$	$\mathbf{0.67_{0.39}}$	$0.48_{0.58}$	$\mathbf{0.54_{0.39}}$	$\mathbf{0.66_{0.25}}$

Table 15: Extended results for the regret on the Perason correlation$_{\text{std}}$ (↓) of the ranking metrics when the base clusterings seek different number of clusters.

	FCPS				UCI			
Model	Agg.	DBSCAN	Kmeans	Alltogether	Agg.	DBSCAN	Kmeans	Alltogether
AARI	$0.46_{0.39}$	$0.37_{0.60}$	$0.89_{0.65}$	$0.47_{0.50}$	$0.65_{0.51}$	$0.75_{0.79}$	$0.66_{0.62}$	$0.37_{0.40}$
ANMI	$0.60_{0.47}$	$0.35_{0.61}$	$0.97_{0.65}$	$0.60_{0.54}$	$0.59_{0.51}$	$0.71_{0.78}$	$0.81_{0.69}$	$0.30_{0.20}$
BHI	$0.72_{0.41}$	$1.04_{0.98}$	$1.01_{0.58}$	$0.86_{0.60}$	$0.67_{0.35}$	$0.43_{0.59}$	$0.92_{0.68}$	$0.61_{0.48}$
BRI	$1.26_{0.45}$	$1.14_{0.90}$	$1.30_{0.55}$	$1.02_{0.36}$	$0.49_{0.32}$	$0.53_{0.65}$	$0.62_{0.41}$	$0.62_{0.31}$
CHI	$0.91_{0.54}$	$0.53_{0.80}$	$0.62_{0.47}$	$0.89_{0.66}$	$0.24_{0.36}$	$0.37_{0.38}$	$0.45_{0.64}$	$0.27_{0.31}$
CI	$0.84_{0.58}$	$1.10_{0.94}$	$0.92_{0.62}$	$0.94_{0.71}$	$0.92_{0.50}$	$0.67_{0.78}$	$0.78_{0.67}$	$0.53_{0.31}$
DBI	$0.81_{0.48}$	$1.30_{0.77}$	$0.64_{0.42}$	$0.97_{0.55}$	$1.34_{0.41}$	$1.20_{0.44}$	$0.63_{0.36}$	$0.96_{0.32}$
DHI	$0.68_{0.42}$	$1.33_{0.67}$	$1.24_{0.51}$	$0.89_{0.58}$	$1.40_{0.24}$	$0.88_{0.62}$	$1.06_{0.64}$	$1.03_{0.28}$
DI	$0.66_{0.52}$	$0.73_{0.78}$	$0.83_{0.33}$	$0.55_{0.43}$	$1.41_{0.27}$	$1.07_{0.71}$	$0.73_{0.43}$	$1.07_{0.38}$
DRI	$1.22_{0.36}$	$1.01_{0.95}$	$1.43_{0.27}$	$1.21_{0.31}$	$1.04_{0.43}$	$0.34_{0.35}$	$1.10_{0.65}$	$0.93_{0.33}$
HI	$0.65_{0.40}$	$0.66_{0.61}$	$0.32_{0.34}$	$0.56_{0.47}$	$1.17_{0.55}$	$1.47_{0.51}$	$0.46_{0.59}$	$0.97_{0.57}$
KDI	$0.81_{0.37}$	$0.68_{0.72}$	$0.36_{0.32}$	$0.58_{0.23}$	$0.94_{0.41}$	$0.99_{0.54}$	$0.60_{0.25}$	$0.78_{0.30}$
LDRI	$1.07_{0.56}$	$1.00_{1.03}$	$1.34_{0.52}$	$1.13_{0.56}$	$0.75_{0.71}$	$0.24_{0.43}$	$1.20_{0.76}$	$0.72_{0.47}$
LSRI	$0.92_{0.58}$	$1.06_{0.97}$	$1.25_{0.54}$	$1.01_{0.65}$	$0.42_{0.60}$	$0.22_{0.58}$	$1.03_{0.72}$	$0.49_{0.52}$
McRao	$0.99_{0.50}$	$1.23_{0.87}$	$1.16_{0.57}$	$1.06_{0.60}$	$1.01_{0.62}$	$0.70_{0.73}$	$0.97_{0.69}$	$0.73_{0.48}$
PBM	$0.86_{0.46}$	$0.70_{0.69}$	$0.45_{0.55}$	$0.74_{0.64}$	$1.00_{0.27}$	$1.05_{0.78}$	$0.32_{0.17}$	$0.85_{0.32}$
SI	$0.95_{0.52}$	$0.82_{0.75}$	$\mathbf{0.18_{0.33}}$	$0.91_{0.65}$	$1.19_{0.34}$	$1.02_{0.71}$	$0.36_{0.33}$	$0.92_{0.31}$
WGSS	$0.89_{0.58}$	$1.03_{1.01}$	$1.13_{0.56}$	$0.99_{0.66}$	$0.40_{0.59}$	$\mathbf{0.20_{0.59}}$	$1.02_{0.71}$	$0.48_{0.49}$
WG	$0.48_{0.35}$	$1.30_{0.71}$	$1.21_{0.43}$	$0.71_{0.49}$	$1.19_{0.21}$	$0.88_{0.59}$	$1.02_{0.54}$	$1.03_{0.20}$
XBI	$1.20_{0.45}$	$0.84_{0.72}$	$0.58_{0.25}$	$1.07_{0.45}$	$1.23_{0.31}$	$1.14_{0.69}$	$0.56_{0.36}$	$0.91_{0.32}$
H	$0.60_{0.55}$	$0.45_{0.69}$	$1.02_{0.67}$	$0.56_{0.63}$	$0.97_{0.71}$	$0.83_{0.81}$	$0.86_{0.68}$	$0.42_{0.59}$
KL	$0.56_{0.50}$	$0.46_{0.70}$	$0.99_{0.66}$	$0.52_{0.60}$	$0.96_{0.70}$	$0.84_{0.81}$	$0.84_{0.67}$	$0.42_{0.59}$
TV	$0.62_{0.57}$	$0.45_{0.69}$	$1.03_{0.68}$	$0.58_{0.65}$	$0.97_{0.72}$	$0.83_{0.81}$	$0.87_{0.68}$	$0.42_{0.59}$
Binary	$\mathbf{0.06_{0.16}}$	$\mathbf{0.07_{0.16}}$	$\mathbf{0.18_{0.22}}$	$\mathbf{0.11_{0.24}}$	$\mathbf{0.18_{0.38}}$	$0.49_{0.57}$	$\mathbf{0.25_{0.24}}$	$\mathbf{0.14_{0.16}}$

Table 16: Extended results for the regret on the ARI of the model selected$_{\text{std}}$ (↓) by each ranking metric when the base clusterings seek different number of clusters.

	FCPS				UCI			
Model	Agg.	DBSCAN	Kmeans	Alltogether	Agg.	DBSCAN	Kmeans	Alltogether
AARI	$0.41_{0.32}$	$0.10_{0.17}$	$0.33_{0.27}$	$0.41_{0.32}$	$0.20_{0.24}$	$0.11_{0.13}$	$0.20_{0.19}$	$0.20_{0.25}$
ANMI	$0.41_{0.33}$	$0.10_{0.17}$	$0.33_{0.30}$	$0.40_{0.34}$	$0.17_{0.23}$	$0.12_{0.13}$	$0.19_{0.24}$	$0.14_{0.15}$
BHI	$0.66_{0.21}$	$0.25_{0.21}$	$0.54_{0.26}$	$0.69_{0.20}$	$0.23_{0.22}$	$0.14_{0.23}$	$0.30_{0.26}$	$0.26_{0.24}$
BRI	$0.67_{0.21}$	$0.24_{0.18}$	$0.58_{0.20}$	$0.72_{0.18}$	$0.21_{0.21}$	$0.14_{0.19}$	$0.29_{0.27}$	$0.26_{0.24}$
CHI	$0.24_{0.32}$	$0.07_{0.15}$	$0.16_{0.19}$	$0.27_{0.34}$	$0.13_{0.20}$	$0.06_{0.09}$	$0.15_{0.19}$	$0.12_{0.17}$
CI	$0.36_{0.37}$	$0.20_{0.20}$	$0.23_{0.28}$	$0.38_{0.36}$	$0.30_{0.23}$	$0.14_{0.15}$	$0.25_{0.22}$	$0.32_{0.22}$
DBI	$0.30_{0.33}$	$0.29_{0.33}$	$0.19_{0.19}$	$0.31_{0.35}$	$0.43_{0.27}$	$0.24_{0.24}$	$0.16_{0.24}$	$0.45_{0.29}$
DHI	$0.25_{0.25}$	$0.33_{0.27}$	$0.54_{0.26}$	$0.26_{0.28}$	$0.39_{0.26}$	$0.22_{0.23}$	$0.27_{0.24}$	$0.41_{0.28}$
DI	$0.24_{0.37}$	$0.17_{0.35}$	$0.29_{0.25}$	$0.25_{0.38}$	$0.43_{0.28}$	$0.20_{0.16}$	$0.26_{0.30}$	$0.45_{0.29}$
DRI	$0.67_{0.21}$	$0.20_{0.20}$	$0.54_{0.25}$	$0.69_{0.20}$	$0.36_{0.20}$	$0.13_{0.24}$	$0.37_{0.25}$	$0.38_{0.23}$
HI	$0.62_{0.39}$	$0.20_{0.34}$	$0.16_{0.25}$	$0.58_{0.37}$	$0.43_{0.27}$	$0.28_{0.22}$	$0.27_{0.27}$	$0.45_{0.29}$
KDI	$0.56_{0.35}$	$0.19_{0.34}$	$0.44_{0.31}$	$0.59_{0.35}$	$0.21_{0.25}$	$0.13_{0.13}$	$0.21_{0.27}$	$0.29_{0.26}$
LDRI	$0.67_{0.21}$	$0.20_{0.20}$	$0.54_{0.25}$	$0.69_{0.20}$	$0.36_{0.20}$	$0.13_{0.24}$	$0.37_{0.25}$	$0.38_{0.23}$
LSRI	$0.66_{0.21}$	$0.20_{0.20}$	$0.54_{0.26}$	$0.69_{0.20}$	$0.23_{0.22}$	$0.07_{0.19}$	$0.30_{0.26}$	$0.26_{0.24}$
McRao	$0.66_{0.21}$	$0.21_{0.19}$	$0.55_{0.25}$	$0.69_{0.19}$	$0.37_{0.23}$	$0.08_{0.11}$	$0.30_{0.26}$	$0.40_{0.23}$
PBM	$0.26_{0.32}$	$0.21_{0.34}$	$0.16_{0.17}$	$0.27_{0.34}$	$0.41_{0.26}$	$0.18_{0.17}$	$0.21_{0.27}$	$0.42_{0.28}$
SI	$0.24_{0.34}$	$0.08_{0.17}$	$\mathbf{0.10_{0.14}}$	$0.31_{0.34}$	$0.40_{0.26}$	$0.19_{0.19}$	$0.14_{0.24}$	$0.41_{0.28}$
WGSS	$0.66_{0.21}$	$0.20_{0.20}$	$0.54_{0.26}$	$0.69_{0.20}$	$0.23_{0.22}$	$0.07_{0.19}$	$0.30_{0.26}$	$0.26_{0.24}$
WG	$0.21_{0.23}$	$0.29_{0.19}$	$0.54_{0.26}$	$\mathbf{0.22_{0.21}}$	$0.40_{0.26}$	$0.18_{0.23}$	$0.28_{0.25}$	$0.41_{0.28}$
XBI	$0.28_{0.31}$	$0.19_{0.33}$	$0.15_{0.19}$	$0.34_{0.38}$	$0.40_{0.26}$	$0.21_{0.24}$	$0.16_{0.23}$	$0.41_{0.28}$
H	$0.40_{0.32}$	$0.10_{0.17}$	$0.41_{0.27}$	$0.43_{0.31}$	$0.23_{0.25}$	$0.12_{0.13}$	$0.23_{0.23}$	$0.15_{0.21}$
KL	$0.40_{0.32}$	$0.10_{0.17}$	$0.42_{0.28}$	$0.43_{0.31}$	$0.23_{0.25}$	$0.12_{0.13}$	$0.23_{0.23}$	$0.15_{0.21}$
TV	$0.40_{0.32}$	$0.10_{0.17}$	$0.41_{0.27}$	$0.43_{0.31}$	$0.23_{0.25}$	$0.12_{0.13}$	$0.23_{0.23}$	$0.15_{0.21}$
Binary	$\mathbf{0.14_{0.20}}$	$\mathbf{0.06_{0.14}}$	$0.17_{0.21}$	$\mathbf{0.22_{0.24}}$	$\mathbf{0.11_{0.16}}$	$\mathbf{0.05_{0.06}}$	$\mathbf{0.09_{0.12}}$	$\mathbf{0.10_{0.14}}$

Table 17: Extended results for the regret on the ARI of the model selected$_{\text{std}}$ (↓) by each ranking metric when the base clusterings are restricted to as many clusters as targets.

	FCPS		UCI	
Model	Agglomerative	Kmeans	Agglomerative	Kmeans
AARI	$0.16_{0.28}$	$0.06_{0.10}$	$0.31_{0.22}$	$0.05_{0.05}$
ANMI	$0.16_{0.28}$	$0.06_{0.10}$	$0.31_{0.22}$	$0.03_{0.03}$
BHI	$0.24_{0.33}$	$0.05_{0.09}$	$\mathbf{0.00_{0.00}}$	$0.06_{0.05}$
BRI	$\mathbf{0.14_{0.21}}$	$0.07_{0.12}$	$\mathbf{0.00_{0.00}}$	$\mathbf{0.02_{0.03}}$
CHI	$0.24_{0.33}$	$0.05_{0.09}$	$\mathbf{0.00_{0.00}}$	$0.06_{0.05}$
CI	$0.22_{0.27}$	$0.05_{0.09}$	$0.29_{0.23}$	$0.11_{0.11}$
DBI	$0.36_{0.41}$	$0.02_{0.04}$	$0.43_{0.28}$	$0.09_{0.08}$
DHI	$0.48_{0.43}$	$0.02_{0.04}$	$0.43_{0.28}$	$0.18_{0.16}$
DI	$0.09_{0.27}$	$\mathbf{0.01_{0.01}}$	$0.38_{0.25}$	$0.08_{0.06}$
DRI	$0.30_{0.41}$	$0.05_{0.10}$	$0.23_{0.27}$	$0.05_{0.07}$
HI	$0.36_{0.41}$	$0.02_{0.04}$	$0.43_{0.28}$	$0.15_{0.16}$
KDI	$0.57_{0.40}$	$0.14_{0.25}$	$0.17_{0.18}$	$0.17_{0.18}$
LDRI	$0.30_{0.41}$	$0.05_{0.10}$	$0.23_{0.27}$	$0.05_{0.07}$
LSRI	$0.24_{0.33}$	$0.05_{0.09}$	$\mathbf{0.00_{0.00}}$	$0.06_{0.05}$
McRao	$0.30_{0.32}$	$0.05_{0.09}$	$0.28_{0.25}$	$0.09_{0.06}$
PBM	$0.37_{0.38}$	$0.05_{0.09}$	$0.35_{0.27}$	$0.12_{0.17}$
SI	$0.31_{0.37}$	$0.02_{0.04}$	$0.38_{0.25}$	$0.07_{0.08}$
WGSS	$0.24_{0.33}$	$0.05_{0.09}$	$\mathbf{0.00_{0.00}}$	$0.06_{0.05}$
WG	$0.57_{0.45}$	$0.01_{0.04}$	$0.43_{0.28}$	$0.16_{0.16}$
XBI	$0.31_{0.37}$	$0.02_{0.04}$	$0.42_{0.28}$	$0.06_{0.07}$
H	$0.26_{0.37}$	$0.06_{0.10}$	$0.32_{0.21}$	$0.05_{0.05}$
KL	$0.26_{0.37}$	$0.06_{0.10}$	$0.32_{0.21}$	$0.05_{0.05}$
TV	$0.26_{0.37}$	$0.06_{0.10}$	$0.32_{0.21}$	$0.05_{0.05}$
Binary	$0.15_{0.24}$	$0.05_{0.10}$	$0.17_{0.18}$	$0.05_{0.06}$