EURASIP Journal on Applied Signal Processing 2004:1, 64–80

Cluster Structure Inference Based on Clustering Stability with Applications to Microarray Data Analysis

Ciprian Doru Giurcăneanu

Institute of Signal Processing, Tampere University of Technology, P.O. Box 553, FIN-33101 Tampere, Finland
Email: cipriand@cs.tut.fi

Ioan Tăbuş

Institute of Signal Processing, Tampere University of Technology, P.O. Box 553, FIN-33101 Tampere, Finland
Email: tabus@cs.tut.fi

Received 28 February 2003; Revised 7 July 2003

This paper focuses on the stability-based approach for estimating the number of clusters K in microarray data. The cluster stability approach amounts to performing clustering successively over random subsets of the available data and evaluating an index which expresses the similarity of the successive partitions obtained. We present a method for automatically estimating K by starting from the distribution of the similarity index. We investigate how the selection of the hierarchical clustering (HC) method, respectively, the similarity index, influences the estimation accuracy. The paper introduces a new similarity index based on a partition distance. The performance of the new index and that of other well-known indices are experimentally evaluated by comparing the “true” data partition with the partition obtained at each level of an HC tree. A case study is conducted with a publicly available Leukemia dataset.

Keywords and phrases: clustering stability, number of clusters, hierarchical clustering methods, similarity indices, partition-distance, microarray data.

1. INTRODUCTION

The clustering algorithms are frequently used for analyzing the microarray data. While various clustering methods help the practitioner in bioinformatics to ascertain different characteristics in structural organization of microarray datasets, the task of selecting the most appropriate algorithm for solving a particular problem is nontrivial. While various clustering methods are applied in hundreds of microarray research papers, a question arises frequently, namely, how to compare two different partitions of the same dataset obtained by two different algorithms. The comparison becomes more difficult when the two partitions do not contain the same number of clusters. The accurate estimation for the number of clusters K is essential because most of the existing clustering procedures request K as input.

The robustness of the clustering algorithms is usually studied by investigating their stability with respect to perturbations changing the original dataset, for example, by drawing random subsets or by artificially adding noise [1]. The stability methods can be also used in exploratory data analysis when little prior information is available regarding the dataset, which is generally the case with microarray data. The main principle is to randomly split the dataset and cluster each subset independently, and then to check the stability (or degree of agreement) of the two obtained partitions. The clustering is stable if the cluster memberships inferred in the two subsets are similar to the memberships in the entire sample [1]. The following two different approaches have been considered when applying the stability methods for finding structure in microarray data.

(1) After randomly splitting the dataset into two subsets, select one subset for learning and another for test. Firstly, a clustering algorithm C_A is applied to the learning set, and the resulting classes are used to classify the samples which belong to the test set. Then the test set is clustered with the same algorithm C_A, and a similarity measure (index) is computed between the labels produced by classification, respectively, clustering [2, 3, 4].

(2) Apply the same clustering algorithm C_A to both subsets and calculate the similarity index on the samples belonging to the intersection of subsets [5]. A modified variant is

introduced in [6]: C_A is applied to the whole dataset (reference clustering) and to a randomly chosen subset. The similarity index is computed for the samples contained in the selected subset.

In both approaches, it is assumed that the number of clusters is $k \in \{2, 3, \dots, k_{\max}\}$, and for each value allowed for k, after running the algorithm many times, the empirical distribution of the similarity index is collected. In [3], the number of clusters K is estimated based on the median of similarity index values. Evaluating the degree of agreement is rephrased in [4] as a prediction problem: their index ("prediction strength") $ps(k)$ measures how well the cluster centroids from the training set predict "co-memberships" in the test set. The index $ps(k)$ is averaged over several random splittings of the original data (into training set and test set), and the estimated number of clusters is given by $\hat{K} = \arg\max_{2 \le k \le k_{\max}} \text{mean}[ps(k)]$ when $\max(\text{mean}[ps(k)])$ is larger than a given threshold. The approach in [6] evaluates the stability for individual patterns and clusters relying on a different similarity score called optimal association. In [5], $\hat{K}$ is chosen as the value, where there is a transition from a similarity index distribution that is concentrated near one to a wider distribution: $\hat{K}$ is visually estimated by using the empirical cumulative distribution function or, alternatively, based on the value of the 90th percentile. In consensus clustering (CC) [7], the central role is played by the consensus matrix that records, for every pair of objects, the proportion of clustering runs in which the two objects are clustered together. Based on the histogram of the consensus matrix entries, an empirical cumulative distribution function is defined, and the selection of the appropriate number of clusters proceeds by inspection of the shape of this function when $k \in \{2, 3, \dots, k_{\max}\}$.

We propose to improve the algorithm described in [5] such that $\hat{K}$ can be automatically estimated without resorting to visual inspection or other heuristic methods. To evaluate the importance of index selection on the accuracy of the estimation, we revisit various similarity indices. Then we define and analyze a new similarity index, which is connected to the recently introduced partition distance [8]. In [3, 5], the Fowlkes-Mallows index [9] is recommended for stability-based methods, but we show experimentally that our newly introduced index and the Jaccard index [10] perform better. We also show in this paper that partition distance is useful in designing a visualization tool which helps consistently the interpretation of clustering results for microarray data.

Potentially, any clustering algorithm can be used in our settings, and we investigate the impact of the algorithm selection on the estimated $\hat{K}$. We restrict our investigation to the agglomerative hierarchical clustering (HC) algorithms [10] mainly because this class of clustering methods is very popular in microarray data analysis [11]. These algorithms are computationally efficient since the same tree can be used for all values of $k \in \{2, 3, \dots, k_{\max}\}$ by looking at different levels of the tree each time. In [7], when evaluating the performances of CC with various microarray datasets, it was concluded that CC based on HC produces slightly better results than CC based on self-organizing maps (SOM). We remark that in [5, 6, 7] the HC is done by the group-average algorithm [10, 12]. In our simulated experiments, the group-average shows modest results when compared with complete-linkage and Ward's methods [10, 12].

The remainder of this paper is organized as follows. Section 2 includes a discussion of some results on the estimation of the number of clusters, previously reported for the publicly available Leukemia dataset [13]. In Section 3, we introduce the similarity indices. Relying on the revisited properties of the partition distance [8], a new similarity index $s(\cdot, \cdot)$ is defined, and a lower bound is found under the hypothesis of generalized hypergeometric distribution for the contingency table. In Section 4, we evaluate experimentally $s(\cdot, \cdot)$ by comparing the "true" clustering of a dataset with the partition obtained at each level of a HC tree. In Section 5, we introduce the stability-based method for finding the data structure by extending the approach proposed in [5]. Comparisons with other methods are reported for simulated data, and a case study is conducted on Leukemia dataset [13].

2. MOTIVATION OF THE WORK

In order to illustrate the challenge of structure estimation for microarray data, we consider the leukemia dataset described in [13], publicly available at http://www-genome.wi.mit.edu/cgi-bin/cancer/datasets.cgi, which comes from a study of gene expression in two types of acute leukemias, acute lymphoblastic leukemia (ALL) and acute myeloid leukemia (AML). The true number of classes may be considered three since the biological labeling of the patient samples is ALL-B, ALL-T, and AML [13]. The dataset consists of 6817 human genes measured for 72 patients: 47 cases of ALL (38 B-cell ALL and 9 T-cell ALL) and 25 cases of AML.

We note that the clustering of Leukemia dataset was already investigated in several studies. In [13], the SOM are applied to cluster measurements from 38 patients (out of 72), relying on 50 "informative" genes selected based on a supervised procedure. We emphasize here that the "informative" genes selection relies on the gene correlation with different types of Leukemia. In two recent publications [14, 15], various validation techniques based on computing internal indices are used to estimate the number of clusters in the 38×50 dataset when SOM is the clustering algorithm. The paper [15] concludes that the estimated number of clusters is $\hat{K} = 2$ and mentions, as a second best choice, $\hat{K} = 4$.

The whole set of measurements from the 72 patients is clustered in [16] by k-means, fuzzy c-means networks, SOM, fuzzy SOM, and growing cell structure (GCS) algorithm. When varying the number of clusters between 2 and 16, all the resulting clusterings are evaluated based on the distribution of Leukemia types within the clusters, the highest degree of intracluster homogeneity being obtained when samples are divided into 9 clusters by fuzzy SOM. A procedure for gene selection is applied.

In [3], the 72 tumors from Leukemia dataset are clustered by partitioning around medoids (PAM) [10] after selecting 100 genes which have the largest variance across tumor samples: for $\hat{K} = 3$, one ALL B-cell sample is clustered with the ALL T-cell samples, and the rest of the observations are allocated correctly. Results on estimating the number of clusters are also reported: applying *clest*, *kl* [17], *hart* [18], or silhoutte (*sil*) [10] leads to $\hat{K} = 3$; *ch* [19] estimates $\hat{K} = 2$. The estimated number of clusters is $\hat{K} = 10$ when using *gap* [20], and $\hat{K} = 5$ when employing *gapPC* [20]. Note that *clest* was originally introduced in [3] and extends the stability-based approach from [2]. Another method relying on stability principle, CC, was formalized and tested in [7]. Since their settings allow to apply various clustering methods, results on estimated number of clusters for 38 (out of 72) samples of Leukemia dataset are reported when using HC and SOM. The method CC in conjunction with HC leads to $\hat{K} = 5$, and to $\hat{K} = 4$ when employing CC in combination with SOM.

In light of these results reported for the Leukemia dataset, we can better understand the importance and difficulty of validation of the number of clusters. It becomes apparent that every method for structure estimation must be deeply analyzed and validated with simulated data for which the true nature is known before applying it to analyze the microarray data. Leukemia dataset is also a good example for illustrating the paradigm of "high dimension and small sample size" which is common in microarray data analysis. It was pointed out in [7] that this paradigm prevents the use of some clustering algorithms, and we show in this paper how stability methods can circumvent this difficulty.

3. SIMILARITY MEASURES

Given an N-object set $T = \{O_1, O_2, \dots, O_N\}$, suppose that $P = \{P_1, P_2, \dots, P_r\}$ and $P' = \{P'_1, P'_2, \dots, P'_c\}$ represent two distinct partitions of T, that is, $\bigcup_{i=1}^{r} P_i = \bigcup_{i=1}^{c} P'_i = T$, where $P_i \bigcap P_j = \varnothing$ for $1 \le i \ne j \le r$ and $P'_i \bigcap P'_j = \varnothing$ for $1 \le i \ne j \le c$. We name, in the sequel, any nonempty subset of T *cluster*. So, any partition of T is a set of mutually exclusive clusters whose reunion is T.

The partitions P and P' are identical if and only if every cluster in P is a cluster in P'. Let M be an $r \times c$ matrix where the quantity m_{ij} is the number of objects in common between the ith cluster of P and the jth cluster of P'. The contingency table is represented in Table 1, where $m_{i\cdot} \triangleq \sum_{j=1}^{c} m_{ij}$ for $1 \le i \le r$ and $m_{\cdot j} \triangleq \sum_{i=1}^{r} m_{ij}$ for $1 \le j \le c$. It is easy to observe that $m_{\cdot\cdot} \triangleq \sum_{i=1}^{r} m_{i\cdot} = \sum_{j=1}^{c} m_{\cdot j} = N$.

3.1. Rand, Jaccard, and Fowlkes-Mallows similarity indices

We introduce the following function relative to an arbitrary partition P of T: for any pair of distinct objects $(O_\ell, O_m) \in T^2$, $1 \le \ell < m \le N$,

$$\mathbf{1}_P(O_\ell, O_m) \triangleq \begin{cases} 1, & \exists i \in \{1, 2, \dots, |P|\} \text{ such that } \{O_\ell, O_m\} \subseteq P_i, \\ 0, & \text{otherwise}, \end{cases} \tag{1}$$

which indicates if two objects belong to the same cluster in the partition P.

Table 1: The contingency table for the partitions P and P' of the N-object set T.

	Cluster	Partition P'				Sums
		P'_1	P'_2	$\cdots$	P'_c	
Partition P	P_1	m_{11}	m_{12}	$\cdots$	m_{1c}	$m_{1\cdot}$
	P_2	m_{21}	m_{22}	$\cdots$	m_{2c}	$m_{2\cdot}$
	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
	P_r	m_{r1}	m_{r2}	$\cdots$	m_{rc}	$m_{r\cdot}$
	Sums	$m_{\cdot 1}$	$m_{\cdot 2}$	$\cdots$	$m_{\cdot c}$	$m_{\cdot\cdot} = N$

Following a classic procedure, we firstly define four sets:

$$\begin{aligned}
\mathscr{W}_1 &\triangleq \{(O_\ell, O_m) \in T^2 \mid \mathbf{1}_P(O_\ell, O_m) = 1,\ \mathbf{1}_{P'}(O_\ell, O_m) = 1\},\\
\mathscr{W}_2 &\triangleq \{(O_\ell, O_m) \in T^2 \mid \mathbf{1}_P(O_\ell, O_m) = 1,\ \mathbf{1}_{P'}(O_\ell, O_m) = 0\},\\
\mathscr{W}_3 &\triangleq \{(O_\ell, O_m) \in T^2 \mid \mathbf{1}_P(O_\ell, O_m) = 0,\ \mathbf{1}_{P'}(O_\ell, O_m) = 1\},\\
\mathscr{W}_4 &\triangleq \{(O_\ell, O_m) \in T^2 \mid \mathbf{1}_P(O_\ell, O_m) = 0,\ \mathbf{1}_{P'}(O_\ell, O_m) = 0\},
\end{aligned} \tag{2}$$

and denote the cardinalities of these sets, $w_i \triangleq |\mathscr{W}_i|$ for $i \in \{1, 2, 3, 4\}$. Then we recall the definitions for three well-known similarity indices:

(1) Rand [21]: $(w_1 + w_4)/\sum_{i=1}^{4} w_i$,
(2) Jaccard [22]: $w_1/\sum_{i=1}^{3} w_i$,
(3) Fowlkes-Mallows [9]: $w_1/\sqrt{(w_1 + w_2)(w_1 + w_3)}$.

Since w_i $(1 \le i \le 4)$ are nonnegative numbers, all three indices take values in the interval $[0, 1]$. The partitions P and P' are identical if and only if $w_2 = w_3 = 0$; when they are identical and $w_1 \ne 0$, then all indices are equal to their maximum value 1. Observe for the denominator of Rand index that $\sum_{i=1}^{4} w_i = \binom{N}{2}$. The Jaccard index is not defined for the trivial case when each cluster in P and P' contains at most 1 object, which is equivalent to $w_1 = w_2 = w_3 = 0$. The Fowlkes-Mallows index is not defined when $w_1 = w_2 = 0$ (each cluster in P contains at most 1 object) or $w_1 = w_3 = 0$ (each cluster in P' contains at most 1 object). Formulae for fast computing w_i $(1 \le i \le 4)$ are available [23].

To each similarity measure $sm(P, P')$, bounded by zero and unity, we can associate a dissimilarity $d(P, P') \triangleq 1 - sm(P, P')$; in some cases, $d(P, P')$ could be a metric on the set of all partitions of a given set of objects T [12]. In the next section, we start from the definition given in [8] for the partition distance (which is a metric) and define a new similarity index.

3.2. A similarity index defined as complement of a partition distance

In [8], the following definition is introduced for the partition distance $D(P, P')$ between P and P': "$D(P, P')$ is the minimum number of elements that must be deleted from

T, so that the two induced partitions (P and P' restricted to the remaining elements) are identical." It was pointed out in [24] that the partition distance is also equal to the minimum number of elements that must be moved between clusters in P, so that the resulting partition equals P' (with the convention that any set which becomes empty is no longer a cluster).

Proposition 1. *The partition distance $D(P, P')$ is a metric on the set of all partitions of a given set of objects T.*

Proof. See Appendix A. □

An *assignment* is a selection of entries of the contingency matrix M such that no row or column contains more than one selected entry and is called *optimal* when the sum of the selected cell values is the largest over all possible assignments [24]. Let $A(P, P')$ denote the value of the optimal assignment for the contingency matrix M.

Theorem 1 [24]. *Two properties of partition distance:*

(a) *The relationship between the partition distance and the optimal assignment is given by $D(P, P') = N - A(P, P')$.*

(b) *The elements to be removed from T to induce identical partitions on P and P', are all those objects not associated with any selected cells of the optimal assignment.*

The proof of the theorem is given in [24] where the theorem is further used to show how the partition distance can be computed in $\mathbb{O}((r + c)^3)$ time after creating the matrix M in $\mathbb{O}(N)$ time. Note that the initial algorithm proposed in [8] to compute $D(P, P')$ for any pair of partitions (P, P') is an exponential-time algorithm, and the algorithm in [24] reduces dramatically the computational complexity.

Proposition 2. *The maximum of the partition distance is* $\max_{(P,P')} D(P, P') = N - 1$ *and is achieved if and only if one partition consists of a single cluster and the other one consists only of clusters containing single-objects.*

Proof. See Appendix A. □

The above results suggest the definition of the following index of similarity between any two partitions P and P':

$$s(P, P') \triangleq 1 - \frac{D(P, P')}{N - 1} = \frac{A(P, P') - 1}{N - 1}. \tag{3}$$

The new index is a measure of similarity ranging from $s(P, P') = 0$ when the two partitions have no similarities (i.e., when one consists of a single cluster and the other only of clusters containing single-objects) to $s(P, P') = 1$ when the partitions are identical.

Any injective mapping $\sigma : \{1, 2, \ldots, |P'|\} \to \{1, 2, \ldots, |P|\}$ ($|P'| \leq |P|$) is called association [6] and is useful for comparing two partitions P and P' defined over an N object set T. The measure of similarity between P and P' is computed as $s^*(P, P') \triangleq \max_{\sigma(\cdot)}(1/N) \sum_{j=1}^{|P'|} m_{\sigma(j),j}$ where $m_{\cdot,\cdot}$ denotes the entries of the contingency matrix. Observe that $s^*(P, P') = A(P, P')/N$ and is close to the similarity index defined in (3); $A(P, P') \leq N$ implies that $s(P, P') \leq s^*(P, P')$. It is noticed in [6] that the computation of $s^*(P, P')$ by brute-force enumeration is exponential in the number of clusters, and therefore an approximative greedy heuristic was used there for finding a suboptimal association $\sigma(\cdot)$. Since then, the fast algorithm was introduced in [24], and hence we are going to use the fast, nonapproximative evaluation of $s(P, P')$.

Table 2: The contingency table for the Leukemia dataset: the true partition given by a priori knowledge on the type of disease for each patient is compared with the partition produced by complete-linkage algorithm when $\hat{K} = 3$. All the 3571 genes are used for clustering. The entries associated to the optimal assignment are represented in bold.

Cluster	ALL B-cell	ALL T-cell	AML
C_1	**26**	8	8
C_2	7	**0**	2
C_3	5	1	**15**

We observe that the definition of both $s(P, P')$ and $s^*(P, P')$ relies on the optimal assignment $A(P, P')$, and the main difference between these similarity indices is given by the normalization procedure. Since in [6] $s^*(P, P')$ was successfully applied for detecting stable clusters in microarray data, we are encouraged to employ $s(P, P')$ in stability-based methods for analyzing data produced by microarray technology. The superiority of our approach consists in using nonapproximative algorithms for computing the similarity index.

The use of $s(\cdot, \cdot)$ in validation of microarray data clustering is appealing since the optimal assignment lends itself to be employed as a visualization tool. Assume that we depict the contingency matrix defined by two partitions P and P', where P corresponds to the classes in a microarray dataset already known from medical evidence while P' contains classes found for the same dataset after running a clustering algorithm. Representing in bold the entries associated to the optimal assignment will allow the investigator to assess very easily the memberships. The procedure does not require the number of clusters to be the same in the compared partitions. Moreover, the number of clusters can be visually assessed by checking that all entries in the optimal assignment are larger than zero. Examples of such representations are given in Section 5.2, Tables 2 and 8. When the true state of the nature is not known, the same graphical representation can be used for comparing the results of two different clustering algorithms.

3.3. Similarity indices "corrected for chance"

A similarity index is "corrected for chance" when the expectation of the index takes some constant value (e.g., zero) under an appropriate null model for the contingency table. The property is discussed in [25], and the following general formula is proposed to correct an index:

$$\frac{\text{Index} - \text{Expected Index}}{\text{Maximum Index} - \text{Expected Index}}. \tag{4}$$

The most popular null model assumes that the $r \times c$ contingency table ($r \geq c$) is constructed from the generalized hypergeometric distribution. The main hypothesis is that the two partitions are mutually independent and subject to the condition that the cluster sizes are fixed at $(\alpha_1, \alpha_2, \ldots, \alpha_r)$ and $(\beta_1, \beta_2, \ldots, \beta_c)$, respectively. The α_i and β_j are the marginal totals of m_{ij}, namely, $m_{i\cdot}$ and $m_{\cdot j}$, respectively. Then the expectation of m_{ij} is $E[m_{ij}] = \alpha_i\beta_j/N$ [9, 25]. For example, correcting the Rand index under this hypothesis leads to the expression

$$\text{Adjusted Rand} = \frac{w_1 + w_4 - N_c}{\sum_{i=1}^{4} w_i - N_c}, \tag{5}$$

where two different formulae were proposed for N_c in [25, 26]. We use in the sequel the notations Rand_{HA} and Rand_{MA} for the adjusted index defined in [25], respectively, [26].

We investigate in Appendix B the existence of a lower bound for the expectation of the similarity index defined by (3) when the hypothesis of generalized hypergeometric distribution is verified.

It was already pointed out in [3] that the assumption on the statistical independence of the two compared clusterings does not hold for stability methods since the same data are used to produce both partitions. To gain more insights on the possibility of using $s(\cdot,\cdot)$ in practical applications, we study in Appendix C the asymptotic and finite characteristics of $s(\cdot,\cdot)$ and compare them with the characteristics of other similarity indices.

4. USING THE SIMILARITY INDEX $s(\cdot,\cdot)$ IN HIERARCHICAL CLUSTER ANALYSIS

The aim of this section is to evaluate experimentally $s(\cdot,\cdot)$ when we assume that the "true" structure of the data (the number of clusters and the membership) is known and compare this partition with the partition obtained at each level of a HC tree.

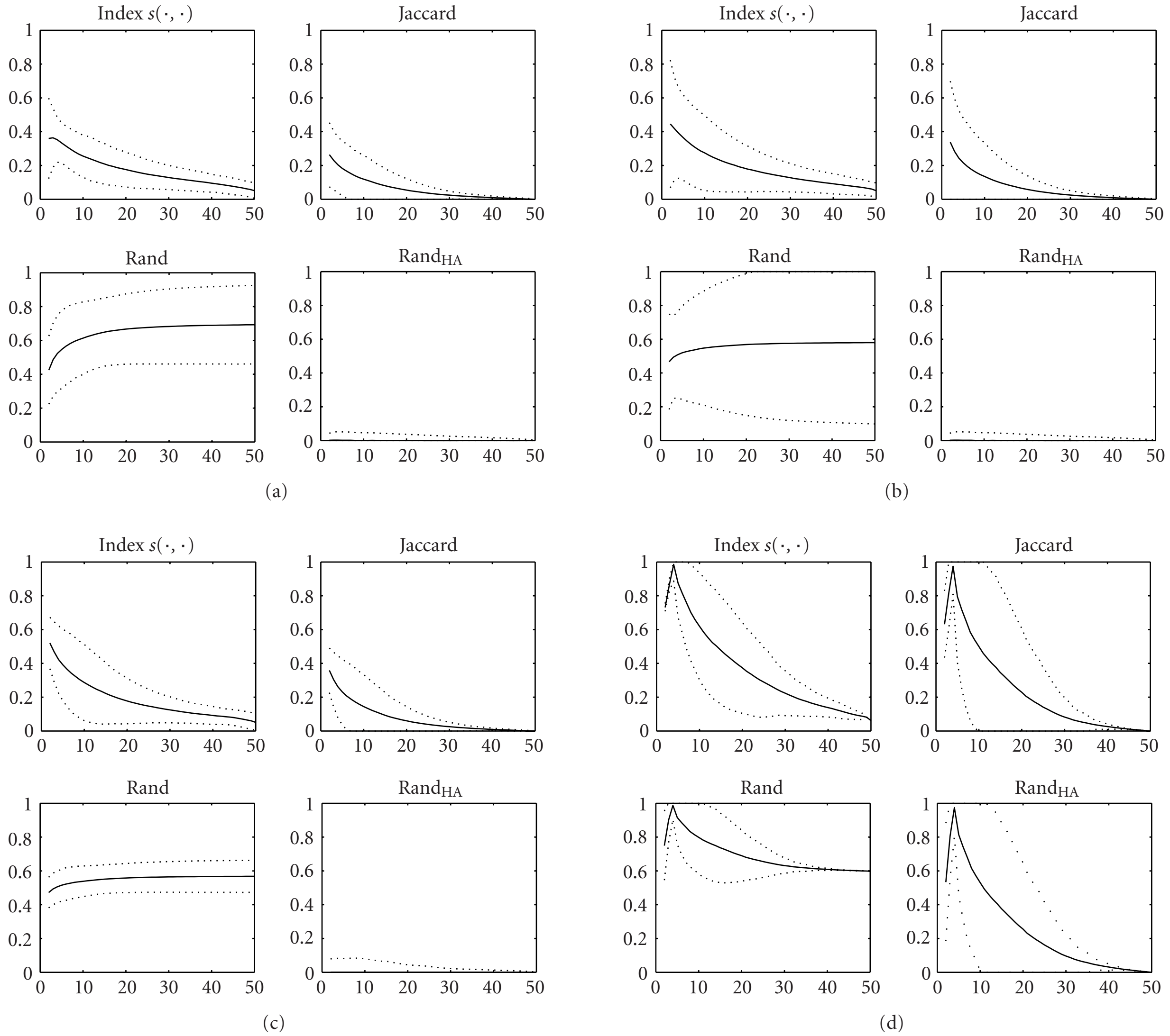
It is a well-known fact that the HC does not yield a discrete number of clusters, but rather a hierarchical arrangement between objects. For better understanding of the behavior of similarity indices, assume that the "true" structure of the data is known and compare this partition with the partition obtained at each level of the HC solution. This approach was originally used in [27] to compare Rand, Rand_{HA}, Rand_{MA}, Fowlkes-Mallows, and Jaccard indices.

We reconsider the experiments described in [27] to evaluate the newly introduced index $s(\cdot,\cdot)$, and for comparison, we compute also Rand, Rand_{HA}, and Jaccard indices. For the first set of experiments, each generated dataset consists of 50 points uniformly distributed in a hypercube in 4-, 6-, or 8-dimensional Euclidean space. There is no significant cluster structure in the data, but a "criterion" solution is assumed: a hypothetical number of clusters (set at either 2, 3, 4, or 5) and a particular distribution pattern of the points to the clusters. Three density patterns are used: equal density (objects are uniformly assigned across the clusters), 10% density condition (one cluster contains 10% of the total number of objects, while 90% of objects are uniformly assigned across the other clusters), and 60% density condition (one cluster contains 60% of the total number of objects, while 40% of objects are uniformly assigned across the other clusters). For example, when the number of clusters is 5 for 10% density, the points are assigned to the clusters as follows: 5, 11, 11, 11, 12. For each selected number of clusters and for each pattern distribution, 15 datasets are generated. The HC is performed by using the single link, the complete link, the group average, and the Ward method [12]. The computed similarity index is averaged over the datasets and over the HC methods, and the mean statistics (with limits at two standard deviation) are plotted in Figures 1a, 1b, and 1c versus the hierarchy level for each of the three density conditions. The two-standard deviation limit is omitted for those levels where the values would be negative or larger than 1.0. The only index for which the mean plot is flat and close to zero is Rand_{HA}. For $s(\cdot,\cdot)$ and Jaccard, the computed mean is decreasing when the number of clusters in HC is increasing. Rand takes values larger than the other indices, and the mean is increasing slowly when the number of clusters in HC is increasing. For $s(\cdot,\cdot)$ and Jaccard, the variance is larger when the partition contains a small cluster; in the same situation, we observe a serious increase in the variance of Rand.

In the second set of experiments, the test data are generated according to the algorithm described in [28]; the clusters contained in the data are separated in the variable space and are internally cohesive. It was observed that the mean of similarity indices is close to 1.0 when the number of clusters in HC solution is equal to the true number of clusters for all considered structures. We plot in Figure 1d the mean statistics for the similarity indices in the case of 60% density condition for four clusters.

All plots in Figure 1 for Rand, Rand_{HA}, and Jaccard are very close to similar plots in [27]. The new index $s(\cdot,\cdot)$ has almost the same performance pattern as Jaccard; generally, the variance of $s(\cdot,\cdot)$ is smaller than the variance of Jaccard index, while the mean is larger. Extending the conclusions from [27], we can observe that a value larger than 0.9 for the Rand, 0.7 for the Jaccard, and 0.8 for $s(\cdot,\cdot)$ is likely to reflect the recovery of some part of the true structure.

For all structured datasets, the clusters contained in the data have been crafted to be disjointed, separated in the variable space, and internally cohesive. Relying on these properties to obtain grouping in k clusters ($2 \leq k \leq k_{\max}$), we choose the clusters at kth depth in the dendrogram. In microarray cluster analysis, the datasets contain outliers which do not belong to any group. Consequently, the dendrogram resulting after running a certain HC algorithm could have at kth depth a singleton (a cluster containing only an outlier). In that case, we move down the HC tree until k distinct clusters are identified, each of them containing at least two objects. It was shown in [29] that the similarity with the true partition is larger when considering the k distinct clusters (and ignoring the outliers) than simply taking all clusters at kth depth in the dendrogram. Since we aim to identify structures in data, we prefer an algorithm which can accurately



(a)

(b)

(c)

(d)

Figure 1: Mean of the similarity indices versus the number of clusters (solid line) with limits at two-standard deviation (dotted line). (a) The equal density condition when no structure exists in the data. (b) The 10% density condition when no structure exists in the data. (c) The 60% density condition when no structure exists in the data. (d) The 60% density condition when data contains four distinct clusters The x-axes denote the number of clusters while the y-axes denote the similarity index value.

estimate $\hat{K}$ relying on a subset of the original dataset instead of one that clusters all objects with an increased risk of misclassification.

5. STABILITY-BASED METHOD FOR ESTIMATING THE NUMBER OF CLUSTERS

First, we briefly revisit the algorithm introduced in [5] when the dataset contains N points embedded in p-dimensional space. Assume that the maximum number of clusters is $k_{\max}$, and for each allowable value of k, except the trivial case ($k = 1$), select from the data two subsets such that each of them contains $f = 80\%$ of the original samples. Use the average-link HC algorithm [12] to cluster every subset in k nonsingleton groups, and then compute the Fowlkes-Mallows similarity index [9] on the intersection of subsets. The number of pairs of solutions compared for each k is $N_t = 100$. It was pointed out in [5] that the histogram of similarity indices is concentrated near one only for values of k smaller than or equal to the "true" number of clusters. Relying on this observation, the number of clusters has been visually evaluated by inspecting the plot of the empirical cumulative distribution function of similarity index. We extend the algorithm from [5] for any similarity index and any HC algorithm.

In the rest of the section, we introduce and analyze the method for automatic selection of the number of clusters. Let $sm_{k,t}$ be the value of the similarity index for the tth pair of solutions compared under the hypothesis of k nonsingleton clusters. For a given k, we are interested in the histogram obtained from the values $sm_{k,1}, sm_{k,2}, \ldots, sm_{k,N_t}$. A good indicator for the location of the histogram is the *mean* of the values, but since the *median* is more robust to the presence of outliers, we compute

$$m_k \triangleq \text{median}\left(sm_{k,1}, sm_{k,2}, \ldots, sm_{k,N_t}\right). \quad (6)$$

We decide that there is no significant structure in the analyzed data ($\hat{K} = 1$) if

$$\max_{2 \le k \le k_{\max}} m_k < Th, \quad (7)$$

where Th depends on the similarity index and the HC algorithm. The threshold Th is determined under a suitable null hypothesis: the *uniformity hypothesis* states that the data are sampled from a uniform distribution in p-dimensional space, while under the *unimodality hypothesis*, the data are thought to be random sample from a multivariate normal distribution [3]. We use in the sequel the uniformity hypothesis for the null case.

When $\max_{2 \le k \le k_{\max}} m_k \ge Th$, let $\nu : \{2, 3, \ldots, k_{\max}\} \to \{2, 3, \ldots, k_{\max}\}$ be a permutation such that $m_{\nu(2)}, m_{\nu(3)}, \ldots, m_{\nu(k_{\max})}$ are the elements of the set $\{m_2, m_3, \ldots, m_{k_{\max}}\}$, decreasingly ordered. Calculate

$$i^* \triangleq \arg\max_i \left(m_{\nu(i)} - m_{\nu(i+1)}\right), \quad (8)$$

which is a "border" between values of k yielding stable, respectively, unstable clustering. The estimated number of clusters is given by the maximum value of k for which the resulting clustering is still stable, or equivalently $\hat{K} = \max(\nu(2), \nu(3), \ldots, \nu(i^*))$.

The improvement proposed for the algorithm described in [5] leads to an automatic procedure for estimating $\hat{K}$ without resorting to any heuristic method. The accuracy of the new algorithm is tested next using artificial and microarray data.

5.1. Performance evaluation with simulated data

We investigate the performances of the algorithm by using artificially generated data for which the true state of the nature is known. The experiments are intended for studying the influence of the HC algorithm and the similarity index on the accuracy of estimation. In [3, 5], the use of Fowlkes-Mallows similarity index is recommended. Due to this reason, we report estimation results when applying it in conjunction with group-average, complete-linkage, Ward's method, centroid, and single-linkage clustering, while for other considered indices, the comparisons are restricted to three clustering algorithms. A complete description of the clustering algorithms could be found in [10, 12]. In all cases, the distance between two clustered objects is taken to be the Euclidean distance.

The artificial data are generated according to Models 1–8 introduced in [3]: Model 1 obeys the uniformity hypothesis and Models 2–8 assume the presence of various number of clusters. For each model, $N_d = 50$ datasets are simulated, and the results are reported in Tables 3, 4, and 5, where $k_{\max} = 7$ is assumed. In Tables 3, 4, and 5, the maximum of the distribution for $\hat{K}$ over $N_d = 50$ estimations is represented in bold for each method. For every dataset, the number of pairs of solutions compared for each k ($2 \le k \le k_{\max}$) is $N_t = 100$. We note that for Models 1–8, the number of samples in every dataset varies between 100 and 200 [3] and during the subsampling process we select from the data two subsets such that each of them contains $f = 80\%$ of original samples.

For each model, the best solution corresponds to the method having the highest percentage of simulations for which the number of clusters is correctly recovered and is marked with an arrow (⇐) in Tables 3, 4, and 5. The only clustering algorithms that lead to good results are complete-linkage and Ward's method; the former gives 4 and the latter 8 "best solutions." The group-average clustering is recommended in [5, 6], but we remark the modest performances of the algorithm for the actual tests. Only one similarity index "corrected for chance" is considered in these experiments, namely, Rand_{HA}. Unsurprisingly, Rand_{HA} distinguishes very well between structured and unstructured datasets; when applied in conjunction with complete-linkage or Ward's method, it identifies the lack of structure for all files generated according to Model 1 ($K = 1$) and for the files associated to Models 2–8, the estimated $\hat{K}$ is always larger than 1. When the HC is based on group-average and the similarity index is Rand_{HA}, five false positive results are reported ($\hat{K} > 1$ five times for Model 1), respectively, five false negative results ($\hat{K} = 1$ five times for Model 7). The values of the threshold Th used in (7) to decide for the Models 1–8 if there is no significant structure in the analyzed dataset ($\hat{K} = 1$) are given in Table 6.

For structured Models 2–8, the best solution is associated only once to the algorithm which measures the similarity with Rand_{HA}, and this occurs for Model 5 (Table 4). Comparing the performances of various similarity indices over all models, we observe that $s(\cdot, \cdot)$ leads to the best solution five times (Models 1, 2, 3, 6, 8), Jaccard three times (Models 1, 4, 7), Rand_{HA} three times (Models 1, 5), while Fowlkes-Mallows only once (Model 1). We remark that the newly introduced index $s(\cdot, \cdot)$ is best ranked. When clustering is done by group-average, measuring the similarity with Fowlkes-Mallows index leads to poor results.

We dub *sw*, the stability-based method, for estimating the number of clusters when Ward's algorithm is used in conjunction with $s(\cdot, \cdot)$ and compare it, for the Models 1–8, with seven methods analyzed in [3]: prediction-based resampling *clest*, *gap* and *gapPC* [20], *sil* [10], *ch* [19], *kl* [17], and *hart* [18]. A description for all seven methods can be found in [3]. The bar plots in Figure 2 represent the percentage of simulations for which the number of clusters was correctly estimated by each considered method according to Tables 3, 4, and 5, respectively [3, Table 3]. By their design, *sil*, *ch*, and *kl* cannot detect the lack of structure, so for these methods, $\hat{K} \ge 2$. The plots in Figure 2 show that excepting *sw* and *clest*, all methods fail to estimate the number of clusters for at least

Table 3: Estimated number of clusters in simulated data. Results for the Models 1, 2, 3.

Similarity index	Hierarchical clustering method	Number of clusters							
	Model 1 (1 cluster in 10 dimensions)								
		1*	2	3	4	5	6	7	
$s(\cdot,\cdot)$	Group-average	**21**	12	16	1	0	0	0	
	Complete-linkage	**44**	6	0	0	0	0	0	
	Ward's method	**50**	0	0	0	0	0	0	⇐
Jaccard	Group-average	16	**23**	10	1	0	0	0	
	Complete-linkage	**45**	5	0	0	0	0	0	
	Ward's method	**50**	0	0	0	0	0	0	⇐
Fowlkes-Mallows	Group-average	15	**17**	16	2	0	0	0	
	Complete-linkage	**44**	6	0	0	0	0	0	
	Ward's method	**50**	0	0	0	0	0	0	⇐
	Centroid method	0	**14**	7	7	11	11	0	
	Single-linkage	**19**	7	2	9	5	8	0	
Rand_{HA}	Group-average	**45**	4	1	0	0	0	0	
	Complete-linkage	**50**	0	0	0	0	0	0	⇐
	Ward's method	**50**	0	0	0	0	0	0	⇐
	Model 2 (3 clusters in 2 dimensions)								
		1	2	3*	4	5	6	7	
$s(\cdot,\cdot)$	Group-average	0	1	13	**21**	14	0	1	
	Complete-linkage	0	0	**38**	10	2	0	0	⇐
	Ward's method	0	0	**25**	19	5	1	0	
Jaccard	Group-average	0	1	13	**20**	13	2	1	
	Complete-linkage	0	0	**35**	9	6	0	0	
	Ward's method	0	2	**35**	11	1	1	0	
Fowlkes-Mallows	Group-average	0	1	11	**20**	15	2	1	
	Complete-linkage	0	0	**31**	10	7	1	1	
	Ward's method	0	1	**34**	13	1	1	0	
	Centroid method	0	0	12	14	**15**	9	0	
	Single-linkage	3	4	**14**	9	7	5	8	
Rand_{HA}	Group-average	0	0	15	**20**	13	1	1	
	Complete-linkage	0	0	**34**	9	5	0	2	
	Ward's method	0	1	**35**	12	1	1	0	
	Model 3 (4 clusters in 10 dimensions, 7 noise variables)								
		1	2	3	4*	5	6	7	
$s(\cdot,\cdot)$	Group-average	0	2	7	**17**	10	14	0	
	Complete-linkage	0	1	10	**21**	12	6	0	
	Ward's method	0	1	4	**39**	5	1	0	⇐
Jaccard	Group-average	0	4	13	**15**	9	9	0	
	Complete-linkage	0	1	14	**15**	10	10	0	
	Ward's method	0	2	9	**35**	4	0	0	
Fowlkes-Mallows	Group-average	0	4	**13**	12	10	11	0	
	Complete-linkage	0	1	12	**13**	12	12	0	
	Ward's method	0	2	7	**33**	6	2	0	
	Centroid method	0	3	10	11	**14**	10	2	
	Single-linkage	0	4	4	10	**14**	8	10	
Rand_{HA}	Group-average	0	4	13	**16**	9	8	0	
	Complete-linkage	0	1	12	**14**	11	12	0	
	Ward's method	0	2	9	**30**	5	2	2	

Table 4: Estimated number of clusters in simulated data. Results for the Models 4, 5, 6.

Similarity index	Hierarchical clustering method	Number of clusters							
		Model 4 (4 clusters in 10 dimensions)							
		1	2	3	4*	5	6	7	
$s(\cdot,\cdot)$	Group-average	0	0	1	**23**	12	12	2	
	Complete-linkage	0	0	0	**34**	12	4	0	
	Ward's method	0	0	0	**36**	14	0	0	
Jaccard	Group-average	0	2	4	**20**	8	12	4	
	Complete-linkage	0	0	3	**24**	15	7	1	
	Ward's method	0	0	0	**41**	8	1	0	⇐
Fowlkes-Mallows	Group-average	0	2	4	**18**	8	14	4	
	Complete-linkage	0	0	3	10	**20**	14	3	
	Ward's method	0	0	0	**31**	16	2	1	
	Centroid method	0	2	2	**19**	14	7	6	
	Single-linkage	2	0	3	**14**	11	6	**14**	
Rand$_{HA}$	Group-average	0	2	4	**17**	10	13	4	
	Complete-linkage	0	0	3	14	**16**	10	7	
	Ward's method	0	1	0	**33**	12	2	2	
		Model 5 (2 elongated clusters in 3 dimensions)							
		1	2*	3	4	5	6	7	
$s(\cdot,\cdot)$	Group-average	0	**17**	5	6	2	7	13	
	Complete-linkage	0	**26**	10	11	0	1	2	
	Ward's method	0	**17**	4	7	5	7	10	
Jaccard	Group-average	0	**24**	15	3	2	4	2	
	Complete-linkage	0	**27**	21	2	0	0	0	
	Ward's method	0	**26**	14	4	3	1	2	
Fowlkes-Mallows	Group-average	0	**21**	12	6	2	6	3	
	Complete-linkage	0	22	**26**	2	0	0	0	
	Ward's method	0	**21**	16	5	4	2	2	
	Centroid method	0	**20**	13	6	3	5	3	
	Single-linkage	0	10	**15**	10	7	7	1	
Rand$_{HA}$	Group-average	0	**20**	13	2	2	3	10	
	Complete-linkage	0	**33**	15	2	0	0	0	⇐
	Ward's method	0	**25**	14	3	2	1	5	
		Model 6 (2 elongated clusters in 10 dimensions, 7 noise variables)							
		1	2*	3	4	5	6	7	
$s(\cdot,\cdot)$	Group-average	1	**12**	10	7	5	4	11	
	Complete-linkage	3	**47**	0	0	0	0	0	⇐
	Ward's method	0	**42**	6	2	0	0	0	
Jaccard	Group-average	0	**14**	7	9	5	3	12	
	Complete-linkage	4	**46**	0	0	0	0	0	
	Ward's method	0	**42**	6	1	1	0	0	
Fowlkes-Mallows	Group-average	0	**12**	7	9	6	4	**12**	
	Complete-linkage	4	**45**	1	0	0	0	0	
	Ward's method	0	**39**	9	1	1	0	0	
	Centroid method	0	**14**	9	6	10	11	0	
	Single-linkage	3	12	7	1	5	**15**	7	
Rand$_{HA}$	Group-average	0	1	0	0	1	1	**47**	
	Complete-linkage	0	0	0	0	0	0	**50**	
	Ward's method	0	**35**	7	1	0	0	7	

Table 5: Estimated number of clusters in simulated data. Results for the Models 7 and 8.

Similarity index	Hierarchical clustering method	Number of clusters							
Model 7 (2 overlapping clusters in 10 dimensions, 9 noise variables)									
		1	2*	3	4	5	6	7	
$s(\cdot,\cdot)$	Group-average	**15**	13	**15**	7	0	0	0	
	Complete-linkage	2	**43**	5	0	0	0	0	
	Ward's method	0	**47**	3	0	0	0	0	
Jaccard	Group-average	14	13	**18**	5	0	0	0	
	Complete-linkage	2	**46**	2	0	0	0	0	
	Ward's method	0	**48**	2	0	0	0	0	⇐
Fowlkes-Mallows	Group-average	13	12	**17**	7	1	0	0	
	Complete-linkage	2	**46**	2	0	0	0	0	
	Ward's method	0	**47**	3	0	0	0	0	
	Centroid method	6	**16**	8	6	5	4	5	
	Single-linkage	**19**	13	0	1	1	5	11	
Rand$_{HA}$	Group-average	5	5	3	1	0	0	**36**	
	Complete-linkage	0	5	3	0	1	0	**41**	
	Ward's method	0	**32**	3	0	0	0	15	
Model 8 (3 overlapping clusters in 13 dimensions, 10 noise variables)									
		1	2	3*	4	5	6	7	
$s(\cdot,\cdot)$	Group-average	10	6	7	3	2	0	**22**	
	Complete-linkage	0	**39**	10	1	0	0	0	
	Ward's method	0	0	**38**	11	1	0	0	⇐
Jaccard	Group-average	**21**	10	7	3	1	0	8	
	Complete-linkage	0	**37**	6	1	5	1	0	
	Ward's method	0	10	**31**	8	1	0	0	
Fowlkes-Mallows	Group-average	**19**	10	8	3	1	0	9	
	Complete-linkage	0	**35**	6	2	6	1	0	
	Ward's method	0	7	**27**	13	3	0	0	
	Centroid method	**38**	4	1	1	2	2	2	
	Single-linkage	4	7	3	1	1	0	**34**	
Rand$_{HA}$	Group-average	0	7	7	2	1	0	**33**	
	Complete-linkage	0	**16**	6	9	4	6	9	
	Ward's method	0	6	**25**	10	3	0	6	

Table 6: The threshold Th used in (7) to decide for the Models 1–8 if there is no significant structure in the analyzed dataset ($\hat{K} = 1$). Remark that the value of Th depends on the HC algorithm and the similarity index.

	$s(\cdot,\cdot)$	Jaccard	Fowlkes-Mallows	Rand$_{HA}$
Group-average	0.9350	0.8600	0.9220	0.3750
Complete-linkage	0.6260	0.4285	0.6040	0.1938
Ward's method	0.7234	0.4532	0.6255	0.3156
Centroid	—	—	0.9620	—
Single-linkage	—	—	0.9750	—

one model: *gap* for Models 5 and 6, *gapPC* for Model 6, *sil* for Model 8, *ch* for Models 5 and 8, while *hart* for Models 1, 2, 5, and 6. Since *hart* fails in four models out of eight, it is concluded in [3] that it performs the worst; *kl* does not really fail in any model, but the results are poor for Models 6–8. In all these cases the percentage of correct estimation is lower than 40%. The methods *sw* and *clest* prove to be robust; the worst result of *sw* occurs in Model 5, while the worst result of

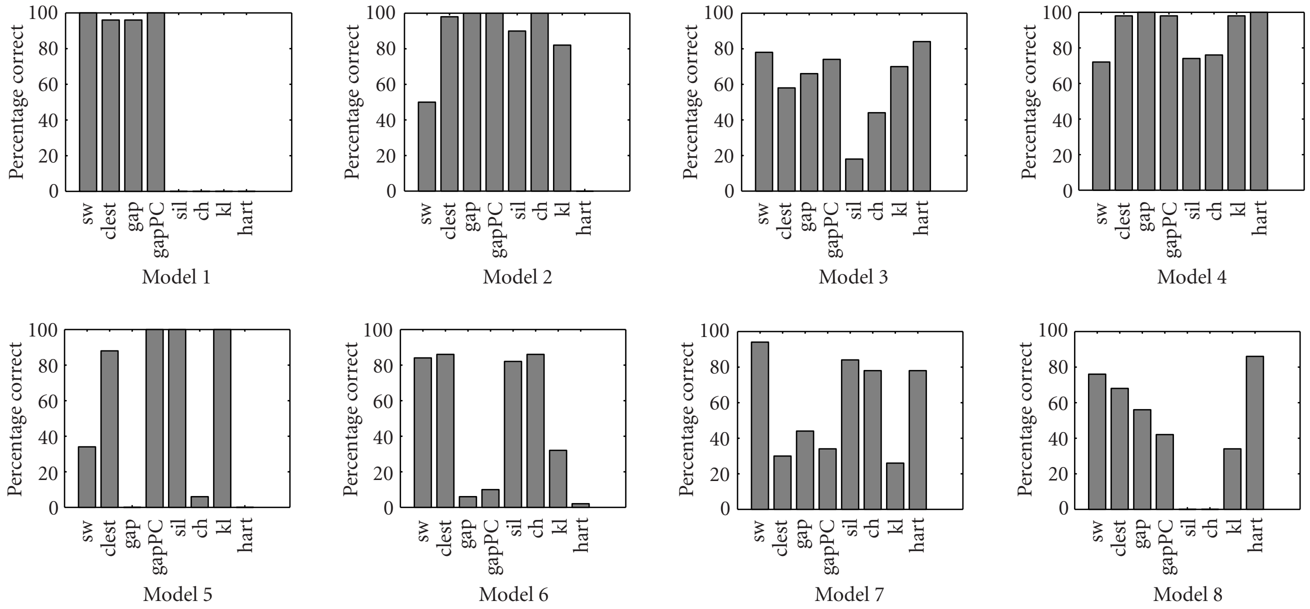


Figure 2: Models 1–8: bar plots representing the percentage of simulations for which the number of clusters is correctly estimated by each method.

clest occurs in Model 7. We emphasize the excellent results of *sw* in finding the two overlapping clusters of Model 7; *clest*, *gap*, and *gapPC* are not able to distinguish between one and two clusters and *kl* overestimates the number of clusters. The behavior of *sw* in Model 2 is surprisingly bad; it is peculiar for Model 2 that the true number of clusters, three, exceeds the dimension of variable space, two. Our interest is on clustering samples from microarray data when N samples (objects) are observed, and each object is associated with a vector of p attributes. Generally, p exceeds by far N, so the number of clusters K is much smaller than the variable space dimension p.

To have a complete picture on the performances of *sw* algorithm, we list in a decreasing order the percentage of simulations for which the number of clusters is correctly estimated by *sw* for every considered model: 100% (Model 1), 94% (Model 7), 84% (Model 6), 78% (Model 3), 76% (Model 8), 72% (Model 4), 50% (Model 2), and 34% (Model 5). The algorithm identifies successfully the lack of structure for Model 1, and for other five structured models, the percentage of correct estimation is larger than 70% which recommends the use of *sw* for a wide family of input data distributions, even if some variables are noisy.

5.2. Clustering the Leukemia dataset

The Leukemia dataset consists of 6817 human genes measured for 72 patients: 47 cases of ALL (38 B-cell ALL and 9 T-cell ALL) and 25 cases of AML. After applying the preprocessing steps described in [3], the measurements for some genes are discarded, and the data are summarized by $N = 72$ vectors in p-dimensional space where $p = 3571$. The results reported in the sequel are obtained without applying any normalization procedure to the data.

We compare the three clusters found by complete-linkage and use all the 3571 genes with the true clusters by displaying in Table 2 the contingency table. We gain more insights by computing the optimal assignment, $A = 26 + 0 + 15 = 41$, according to the definition introduced in Section 3.1. Theorem 1 claims that for inducing identical partitions, we have to remove 31 objects from the dataset, namely, all those objects not associated with any selected entries of the optimal assignment. Since the entry associated to the optimal assignment in the second row has the value zero, the identical induced partitions contain two clusters. This shows that complete-linkage HC amalgamates in C_1 almost all ALL T-cell samples with many ALL B-cell samples, and some AML samples, while in C_2 ALL B-cell samples are grouped with AML samples.

This inability to correctly group the data leads to the conclusion that clustering based on measurements from all genes produces modest results. Therefore we resort to a simple *unsupervised* feature selection method which was also used in [3]: only 100 genes (out of 3571) having the largest variance across tumor samples are employed for clustering. We restrict our investigations to three HC algorithms (group-average, complete-linkage, Ward's method), respectively, three similarity indices ($s(\cdot,\cdot)$, Jaccard, Fowlkes-Mallows), and apply the proposed algorithm when the newly defined space dimension is $p = 100$.

From the dataset consisting of $N = 72$ vectors with length $p = 100$, we select randomly two subsets such that each of them contains 80% of the samples. Then we run the chosen HC algorithm for both subsets and measure the clustering agreement for the samples belonging to the intersection of the subsets. For every hypothesized number of clusters $k \in \{2, 3, \ldots, k_{\max}\}$, the clustering agreement is measured by

Table 7: The estimated number of clusters for Leukemia dataset when measurements from $p = 100$ genes having the largest variance across tumor samples are used. The hypothesized number of clusters varies between 1 and 10. We represent in bold the maximum value over a row.

Hierarchical clustering method	Similarity index	Number of clusters									
		1	2	3*	4	5	6	7	8	9	10
Group-average	$s(\cdot,\cdot)$	0	3	56	40	0	0	0	**184**	17	0
	Jaccard	0	0	0	0	0	0	0	**290**	10	0
	Fowlkes-Mallows	0	0	0	0	0	0	0	**282**	18	0
Complete-linkage	$s(\cdot,\cdot)$	0	75	**212**	12	1	0	0	0	0	0
	Jaccard	0	36	56	**188**	20	0	0	0	0	0
	Fowlkes-Mallows	0	17	31	**207**	37	8	0	0	0	0
Ward's method	$s(\cdot,\cdot)$	0	0	**176**	118	6	0	0	0	0	0
	Jaccard	0	0	95	**202**	3	0	0	0	0	0
	Fowlkes-Mallows	0	0	59	**233**	8	0	0	0	0	0

calculating the value of the selected similarity index *sm*. Repeating the procedure $N_t = 30000$ times, we obtain for every similarity index *sm* and for every allowed value of k, a large set $\mathcal{S}_k \triangleq \{sm_{k,1}, sm_{k,2}, \dots, sm_{k,N_t}\}$.

The histogram drawn for each k from the corresponding set $\mathcal{S}_k$ plays the key role in the automatic estimation method introduced at the beginning of Section 5. To improve the accuracy, we base the estimation on several histograms for every k. This is performed by splitting each set $\mathcal{S}_k$ into $N_b = 300$ non-overlapping blocks and drawing a different histogram for every block. Observe that the length of a block is $N_\ell = 100$. More precisely, we can write $\mathcal{S}_k = \mathcal{B}_{k,1} \bigcup \mathcal{B}_{k,2} \bigcup \cdots \bigcup \mathcal{B}_{k,N_b}$, where $\mathcal{B}_{k,i} = \{sm_{k,(i-1)\times N_\ell+1}, \dots, sm_{k,i\times N_\ell}\}$ for $1 \le i \le N_b$. Applying the newly introduced method, we estimate the number of clusters, which is assumed to lie between 1 and $k_{\max}$, using only the blocks $\mathcal{B}_{2,1}, \mathcal{B}_{3,1}, \dots, \mathcal{B}_{k_{\max},1}$. This is done by computing $m_k = \text{median}(\mathcal{B}_{k,1})$ for $2 \le k \le k_{\max}$ and then applying (7) and (8). Similarly, we obtain another estimation from the blocks $\mathcal{B}_{2,2}, \mathcal{B}_{3,2}, \dots, \mathcal{B}_{k_{\max},2}$. Continuing the procedure, N_b estimations of the number of clusters are resulting for every pair (*clustering method, similarity index*). For the case $k_{\max} = 10$, we show in Table 7 the distributions of estimated number of clusters when various similarity indices and HC algorithms are applied. For each distribution, we decide that $\hat{K}$ is the value corresponding to the maximum number of occurrences (represented in bold).

According to the existing biological knowledge, the number of clusters for Leukemia dataset is three. Following the procedure described above, we obtain from Table 7 that $\hat{K} = 3$ only when complete-linkage, respectively, Ward's method are used in conjunction with the new similarity index $s(\cdot,\cdot)$. Recall that for the simulated data, only complete-linkage and Ward's method have produced good results. For Leukemia dataset, when these two HC algorithms are applied in combination with Jaccard or Fowlkes-Mallows index, the estimated number of clusters is $\hat{K} = 4$. A possible explanation for $\hat{K} = 4$ relies on the remark, from [7], that ALL B-lineage type samples can be further split into two clusters. Surprisingly, the group-average is leading to $\hat{K} = 8$, which is hard to be given a plausible biological interpretation. The experiments with Leukemia dataset reconfirm that the estimated number of clusters $\hat{K}$ depends strongly on the HC algorithm and on the similarity index. The newly introduced index $s(\cdot,\cdot)$ is the only one that leads to correct estimations.

We further investigate how various HC algorithms cluster the 72×100 Leukemia dataset in classes when Euclidian distance is used to measure the distance between objects. We show in Table 8 the contingency tables when the true partition is compared with partitions produced by clustering algorithms for $\hat{K} \in \{3, 4, 8\}$. In each case, we measure the degree of agreement between the compared partitions by computing the optimal assignment (A^*) as defined in Section 3.1: the larger the value of A^*, the better the degree of agreement. Remark that only the entries of the contingency matrix associated with the optimal assignment (bold represented in Table 8) correspond to samples reliably clustered. The values of A^* reported in Table 8 vary between 45 (group-average) and 53 (complete-linkage), or equivalently, the proportion of reliably clustered samples varies between 63% and 74%.

As expected, A^* declines when the estimated number of clusters $\hat{K}$ is larger than three. For $\hat{K} = 3$, the complete-linkage method clusters properly 30 samples from ALL B-cell class, 8 samples from ALL T-cell class, and 15 samples from AML class. When $\hat{K}$ raises from 3 to 4, only the number of samples from AML class, well classified by complete-linkage method, changes; namely, it decreases from 15 to 12. For $\hat{K} \in \{3, 4\}$, the number of ALL B-cell samples correctly grouped by Ward's method is 28, and 14 AML samples are also well classified. In the case of Ward's method, the number of correctly grouped ALL T-cell samples drops from 8 to 6 when $\hat{K}$ increases from 3 to 4. It is obvious that the smallest A^* is obtained for group-average for which $\hat{K} = 8$; remark in this case that 8 ALL T-cell samples are assigned to the same group.

The importance of feature selection is revealed when comparing the results reported, in Tables 2 and 8, for $\hat{K} = 3$. Using the measurements of only variance-based selected 100 genes improves significantly the clustering. The issue of feature selection for stability-based algorithms is addressed in

Table 8: The contingency tables for the Leukemia dataset: the true partition given by a priori knowledge on the type of disease for each patient is compared with partitions produced by HC algorithms when $\hat{K} \in \{3, 4, 8\}$. Only 100 genes having the largest variance across tumor samples are used for clustering. For each contingency table, the entries associated to the optimal assignment are represented in bold.

Hierarchical clustering method	Cluster	ALL B-cell	ALL T-cell	AML	A^*
Group-average $\hat{K} = 8$	C_1	3	0	**11**	45
	C_2	1	1	3	
	C_3	**26**	0	5	
	C_4	3	0	2	
	C_5	2	0	0	
	C_6	1	0	3	
	C_7	1	0	0	
	C_8	1	**8**	1	
Complete-linkage $\hat{K} = 3$	C_1	**30**	0	8	53
	C_2	3	**8**	2	
	C_3	5	1	**15**	
Complete-linkage $\hat{K} = 4$	C_1	5	1	**12**	50
	C_2	0	0	3	
	C_3	**30**	0	8	
	C_4	3	**8**	2	
Ward's method $\hat{K} = 3$	C_1	**28**	0	7	50
	C_2	5	**8**	4	
	C_3	5	1	**14**	
Ward's method $\hat{K} = 4$	C_1	5	2	4	48
	C_2	0	**6**	0	
	C_3	**28**	0	7	
	C_4	5	1	**14**	

[30], and also a special form, namely, leading principal components selection is investigated in [6]. In this paper, we focus on the choice of the HC algorithm, respectively, the similarity index, and we refer for the feature selection problem to the rich literature on this topic.

6. CONCLUSION

In this study, we present a stability-based method applied for the estimation of the number of clusters in microarray data. To gain insights into the choice of the similarity index and HC algorithm, a careful study on simulated and real data is performed.

A new similarity index $s(\cdot,\cdot)$ is introduced, and its capabilities are evaluated against other well-known similarity indices, based on a benchmark originally proposed in [21]. In this framework, $s(P, P')$ takes small values when partition P' is obtained from partition P after *severe* modifications, which recommends the use of $s(\cdot,\cdot)$ in practical applications. The index $s(\cdot,\cdot)$ is further evaluated in standard experimental conditions when measuring the agreement between the true partition and the partition obtained at each level of an HC solution. We draw the conclusion that a value of 0.8 for $s(\cdot,\cdot)$ is likely to reflect the recovery of some part of the true structure. Moreover, since microarray data are noisy, when necessary to obtain grouping in k clusters, we do not choose automatically the clusters at kth depth in the dendrogram, but move down the hierarchical tree until k nonsingleton clusters are identified.

We note the superiority of $s(\cdot,\cdot)$ and Jaccard when compared to Fowlkes-Mallows index. In experiments with simulated data, the use of $s(\cdot,\cdot)$ was leading to the highest percentage of recovering the true number of clusters five times, while Jaccard index three times and Fowlkes-Mallows index only once. Also for the Leukemia dataset, $s(\cdot,\cdot)$ is the only index which leads to the correct estimation of the number of clusters ($\hat{K} = 3$). We emphasize that the definition of $s(\cdot,\cdot)$ relies on optimal assignment, which is the core of a visualization tool newly proposed in this paper for the interpretation of microarray data clustering.

The good performances of complete-linkage algorithm and Ward's method, observed in Section 5.1 for artificial data, have been reconfirmed for Leukemia data. Even when basing the clustering only on 100 selected genes, the results in Table 8 show the presence of misclassified samples for $\hat{K} = 3$. A major drawback of agglomerative HC was already pointed out in [12]: the fusions once made are irrevocable, so when an algorithm has joined two individuals, they cannot subsequently be separated. A similar drawback occurs for divisive HC algorithms, while partition methods can reconsider, at every stage of clustering, to which group to assign an object [12]. We conclude that agglomerative HC algorithms like complete-linkage or Ward's method are well suited to be used with the newly introduced method for the estimation of the number of clusters. The resulting method will offer reliable estimates for K and at the same time will be very fast since the HC is computationally efficient; the same tree can be used for all values of $k \in \{2, 3, \dots, k_{\max}\}$ by looking at different levels of the tree each time. Once K is estimated, partition methods can be further employed for assigning the objects to the clusters.

APPENDICES

A. PROOFS OF PROPOSITIONS

Proof of Proposition 1. It results from the definition that $D(P, P') = D(P', P) \geq 0$ for each pair of partitions (P, P'), while $D(P, P') = 0$ if and only if P and P' are identical. It remains to verify the triangle inequality. Consider three partitions P, P' and P'' of the objects in T. Let U (and V) denote the minimal subset of T which must be removed such that the induced partitions on P and P' (resp., on P' and P'') are identical. Removing $U \bigcup V$ from T induces identical partitions on P, P', and P'', which leads to the chain of inequalities and equalities: $D(P, P'') \leq |U \bigcup V| \leq |U| + |V| = D(P, P') + D(P', P'')$. □

Proof of Proposition 2. Since all entries of M are nonnegative integers and their sum is $N > 0$, there exist at least one entry m_{ij} such that $m_{ij} > 0$. This leads to $A(P, P') \geq 1$ which is equivalent to $D(P, P') \leq N - 1$. When $P = \{P_1, P_2, \dots, P_N\}$ with $|P_1| = |P_2| = \cdots = |P_N| = 1$ and $P' = \{T\}$, the matrix M reduces to a column vector having only ones as entries, which implies that $D(P, P') = N - 1$. Conversely, when $D(P, P') = N - 1$ and $|P| = r \geq c = |P'|$, let $m_{ij} = 1$ be the only entry of M which is considered in the computation of the optimal assignment $A(P, P') = N - D(P, P') = 1$. Since no entry of the columns with indexes different of j is considered in $A(P, P')$, it follows that all the columns contain only zeros, so, M is essentially a column vector. Because this column vector does not have any entry larger than one, the partition P' consists of a single cluster and the partition P consists only of clusters containing single-objects. □

B. A LOWER BOUND FOR $E[s(\cdot, \cdot)]$ UNDER THE HYPOTHESIS OF GENERALIZED HYPERGEOMETRIC DISTRIBUTION

Proposition B.1. *Under the assumption of fixed margins $m_{i\cdot}$ and $m_{\cdot j}$, and random allocation of matching counts to m_{ij},*

$$E[s(P, P')] \geq \frac{1}{N-1}\left(\frac{\sum_{i=1}^{c} \alpha_{(i)}\beta_{(i)}}{N} - 1\right) \geq \frac{1}{N-1}\left(\frac{\sum_{i=1}^{c} \alpha_{(i)}}{c} - 1\right), \tag{B.1}$$

where $\alpha_{(1)}, \alpha_{(2)}, \dots, \alpha_{(r)}$ and $\beta_{(1)}, \beta_{(2)}, \dots, \beta_{(c)}$ are the elements of the set $\{\alpha_1, \alpha_2, \dots, \alpha_r\}$, respectively, the set $\{\beta_1, \beta_2, \dots, \beta_c\}$ decreasingly ordered.

Proof. Consider the particular assignment value $a(P, P') \triangleq \sum_{i=1}^{c} m_{(i),(i)}$. By definition, $A(P, P') \geq a(P, P')$, and consequently, $E[A(P, P')] \geq E[a(P, P')]$. This observation, together with definition (3) and $E[a(P, P')] = \sum_{i=1}^{c} \alpha_{(i)}\beta_{(i)}/N$, proves the first inequality in (B.1). The second inequality results from the Chebyshev inequality [31] applied for the sequences $(\alpha_{(1)}, \alpha_{(2)}, \dots, \alpha_{(c)})$ and $(\beta_{(1)}, \beta_{(2)}, \dots, \beta_{(c)})$; we also used the identity $\sum_{i=1}^{c} \beta_{(i)} = N$. Note that the equality occurs if and only if $\alpha_{(1)} = \alpha_{(2)} = \cdots = \alpha_{(c)}$ or $\beta_{(1)} = \beta_{(2)} = \cdots = \beta_{(c)}$. □

Corollary B.1. (a) *When $r > c$, the maximum value of the lower bound,*

$$\max_{\alpha_1, \alpha_2, \dots, \alpha_r} \frac{1}{N-1}\left(\frac{\sum_{i=1}^{c} \alpha_{(i)}}{c} - 1\right) = \frac{1}{c}\frac{N-r}{N-1}, \tag{B.2}$$

is achieved whenever $\alpha_{(c+1)} = \alpha_{(c+2)} = \cdots = \alpha_{(r)} = 1$.

(b) *When $r = c$, the expression of the lower bound becomes* $(1/(N-1))(N/c - 1)$.

C. ASYMPTOTIC AND FINITE SAMPLE CHARACTERISTICS FOR THE SIMILARITY INDICES

We illustrate the computation of $s(P, P')$ by considering an example from [21]: two partitions of six objects, $P = \{\{x_1, x_2, x_3\}, \{x_4, x_5, x_6\}\}$ and $P' = \{\{x_1, x_2\}, \{x_3, x_4, x_5\}, \{x_6\}\}$. Elementary calculations lead to $s(P, P') = 0.6$ which is equal to the Rand index value reported in [21]. The same example was used in [25] to compare Rand_{HA}, which takes the value $2/17 \approx 0.1176$, with $\text{Rand}_{MA} = 1/3 \approx 0.3333$. For this particular case, $s(\cdot, \cdot)$ and Rand index take the same value which is larger than the adjusted forms of Rand. We consider in this section more comparisons of the newly introduced index with Rand, Rand_{HA}, Rand_{MA}, Jaccard, and Fowlkes-Mallows indices.

To study the finite and asymptotic characteristics, assume that the original data partition P consists of K clusters with n objects each; ten cases when P' is obtained from P after various simple and major modifications are considered. This approach was firstly proposed in [21] to establish some formal properties of Rand index and further used in [9] when evaluating the performances of the Fowlkes-Mallows index. The expressions of Rand, Rand_{MA}, Jaccard, and Fowlkes-Mallows indices for all the ten cases are given in [23]. We compute in Table C.1 the close forms for the partition distance and the index $s(P, P')$ when P' is obtained by modifying P as described in [21].

We compute also the asymptotics when the number of objects in each cluster increases without bound ($n \to \infty$), while the number of clusters is fixed (K fixed). We observe from the fourth column in Table C.1 that the index asymptotics for the fourth and fifth scenarios are equal to 1.0, which is also true for all similarity indices analyzed in [23]. As it was already pointed out in [23], this is reasonable since P and P' are different in, at most, K points; differences of this magnitude are not very serious if an infinite number of the other points are clustered identically by P and P'.

Table C.1: Expressions for the partition distance $D(\cdot,\cdot)$ and the index $s(\cdot,\cdot)$ between two similar partitions, given an initial partition P which has K clusters of n objects each.

P' is a simple modification of the original partition P				
Modification of P	$D(P,P')$	$s(P,P')$	$\lim_{n\to\infty} s(P,P')$ K fixed	$\lim_{n\to\infty} s(P,P')$ $K=\lambda n$
Two clusters joined	n	$\frac{n(K-1)-1}{nK-1}$	$\frac{K-1}{K}$	1.0
One cluster splits into two equal parts (n even)	$n/2$	$\frac{n(K-1/2)-1}{nK-1}$	$\frac{K-1/2}{K}$	1.0
One cluster splits into single-object clusters	$n-1$	$\frac{n(K-1)}{nK-1}$	$\frac{K-1}{K}$	1.0
One object taken from each cluster to form a new cluster of K objects	K	$\frac{(n-1)K-1}{nK-1}$	1.0	1.0
P' and P'' are similar modifications of the original partition P				
Differences between P' and P''	$D(P',P'')$	$s(P',P'')$	$\lim_{n\to\infty} s(P',P'')$ K fixed	$\lim_{n\to\infty} s(P',P'')$ $K=\lambda n$
Movement of an object to different clusters	1	$\frac{nK-2}{nK-1}$	1.0	1.0
Different clusters split into two equal parts (n even)	n	$\frac{n(K-1)-1}{nK-1}$	$\frac{K-1}{K}$	1.0
Different pairs of clusters joined	$2n$	$\frac{n(K-2)-1}{nK-1}$	$\frac{K-2}{K}$	1.0
P' is a major modification of the original partition P				
Modification of P	$D(P,P')$	$s(P,P')$	$\lim_{n\to\infty} s(P,P')$ K fixed	$\lim_{n\to\infty} s(P,P')$ $K=\lambda n$
All clusters joined into one large cluster	$n(K-1)$	$\frac{n-1}{nK-1}$	$1/K$	0.0
All clusters split into single-object clusters	$(n-1)K$	$\frac{K-1}{nK-1}$	0.0	0.0
n clusters are formed with K objects in each, one object from each original cluster	$nK-\min(n,K)$	$\frac{\min(n,K)-1}{nK-1}$	0.0	0.0

The asymptotic values for $s(P,P')$ and the Jaccard index coincide for seven out of ten evaluated situations, while the asymptotics of Jaccard index never exceed the asymptotics of Fowlkes-Mallows index [23]. Comparing the expressions of Jaccard and Fowlkes-Mallows indices given in Section 3.1, it is easy to prove that the Jaccard index cannot be larger than the Fowlkes-Mallows index when both are well defined. We pay particular attention to the behavior of the similarity indices for the last three scenarios (severe cases). The asymptotics for $s(P,P')$ are 0.0 in the last two cases (identical with the values of Rand_{MA} and Jaccard), which shows the superiority of $s(\cdot,\cdot)$ when comparing with the Rand index. The value reported for Rand in [21] in both cases is $(K-1)/K$, and seems unacceptable since it is too close to 1.0. For the nineth modification, the Fowlkes-Mallows index is not defined, while for the tenth modification, it is equal to 0.0. The asymptotic value of $s(P,P')$ is $1/K$ when the modification is such that all clusters are joined into one large cluster, and being smaller than 0.5 for any $K \geq 2$, it may be considered acceptable. For that case, Rand and Jaccard are also equal to $1/K$, while Fowlkes-Mallows is larger $(1/\sqrt{K})$ and $\text{Rand}_{MA} = 0.0$.

When K is allowed to increase without bound ($K \to \infty$), n must also increase without limit, and the solution is to let K increase as a simple proportion of n ($K = \lambda n$) [9]. The results are reported in the last column of Table C.1: only in the severe cases, the computed value is 0.0, while for other situations is 1.0. The behavior is identical for Rand_{MA}, Jaccard, and Fowlkes-Mallows indices, while Rand is equal to 1.0 for the last two severe cases.

Considering an example based on fixed values for n and K, Table C.2 compares different indices when $n = K = 4$. The severity of the modification from the true clustering is ranked as in [23], where Rand, Rand_{MA}, Fowlkes-Mallows and Jaccard similarity measures have been compared for $n = K = 4$. Rand takes values close to one in many cases, while

Table C.2: Six criteria measures computed for two similar partitions, given an initial partition which has $K = 4$ clusters with $n = 4$ objects each (the largest and second largest values are framed).

Modification of true clustering	Rand	Rand$_{HA}$	Rand$_{MA}$	Fowlkes Mallows	Jaccard	$s(\cdot,\cdot)$
Two clusters joined (Serious)	$\boxed{\mathbf{0.8667}}$	0.6667	0.7143	$\boxed{0.7746}$	0.6000	0.7333
One cluster splits into two equal parts (Not Serious)	$\boxed{\mathbf{0.9667}}$	0.8889	$\boxed{0.9130}$	0.9129	0.8333	0.8667
One cluster splits into single-object clusters (Slight Severity)	$\boxed{\mathbf{0.9500}}$	0.8276	$\boxed{0.8667}$	0.8660	0.7500	0.8000
One object taken from each cluster to form a new cluster of k objects (Serious)	$\boxed{\mathbf{0.8500}}$	0.4828	0.6000	0.5774	0.4000	$\boxed{0.7333}$
Movement of an object to different clusters (Not Serious)	$\boxed{\mathbf{0.9333}}$	0.7979	0.8367	$\boxed{0.8400}$	0.7241	$\boxed{\mathbf{0.9333}}$
Different clusters split into two equal parts (Slight)	$\boxed{\mathbf{0.9333}}$	0.7600	$\boxed{0.8171}$	0.8000	0.6667	0.7333
Different pairs of clusters joined (Serious)	$\boxed{\mathbf{0.7333}}$	0.4000	0.4667	$\boxed{0.6000}$	0.4286	0.4667
All clusters joined into one large cluster (Severe)	$\boxed{0.2000}$	0.0000	0.0000	$\boxed{\mathbf{0.4472}}$	$\boxed{0.2000}$	$\boxed{0.2000}$
All clusters split into single-object clusters (Severe)	$\boxed{\mathbf{0.8000}}$	0.0000	$\boxed{0.3333}$	undefined	0.0000	0.2000
n clusters are formed with k objects in each, one object from each original cluster (Severe)	$\boxed{\mathbf{0.6000}}$	−0.2500	0.0000	0.0000	0.0000	$\boxed{0.2000}$

the indices "corrected for chance" (Rand$_{HA}$ and Rand$_{MA}$) have always smaller values. We observe that in all cases, Rand$_{HA}$ is smaller than Rand$_{MA}$. The value of Rand$_{HA}$ in the last row of the table is negative. In general, Rand$_{HA}$ takes values between -1 and 1, but negative values of the index have no substantive use [25]. When the compared partitions are chosen as described in the last row of Table C.2, for any $n = K \geq 2$, the contingency table is an $n \times n$ matrix with all entries equal to one. Simple calculations show that Rand$_{HA} = -1/n < 0$, which leads to Rand$_{HA} = -0.25$ for $n = 4$. When n (and implicitly K) is allowed to increase without bound, Rand$_{HA}$ has the limit 0.0.

When the similarity index takes small values for severe cases, then it is recommended to be used in practical applications [23]. Among the considered indices, $s(\cdot,\cdot)$ is the largest only for a modification ranked *not serious*, and the second largest for a *serious* modification and two *severe* modifications. For the case $n = K = 4$, the only index which shows a better behavior is Jaccard.

ACKNOWLEDGMENT

This work has been supported by Academy of Finland Project no. 44876, 80743, and 80745.

REFERENCES

[1] S. P. Smith and R. Dubes, "Stability of a hierarchical clustering," *Pattern Recognition*, vol. 12, pp. 177–187, 1980.

[2] J. N. Breckenridge, "Replicating cluster analysis: method, consistency, and validity," *Multivariate Behav. Res.*, vol. 24, no. 2, pp. 147–161, 1989.

[3] S. Dudoit and J. Fridlyand, "A prediction-based resampling method for estimating the number of clusters in a dataset," *Genome Biology*, vol. 3, no. 7, pp. research 0036.1–0036.21, 2002, http://genomebiology.com/2002/3/7/research/0036.

[4] R. Tibshirani, G. Walther, D. Botstein, and P. Brown, "Cluster validation by prediction strength," Tech. Rep., Department of Biostatistics, Stanford University, September 2001.

[5] A. Ben-Hur, A. Elisseeff, and I. Guyon, "A stability based method for discovering structure in clustered data," in *Biocomputing 2002: Proc. Pacific Symposium*, R. B. Altman and K. Lauderdalc, Eds., vol. 7, pp. 6–17, Kauai, Hawaii, USA, 2002.

[6] A. Ben-Hur and I. Guyon, "Detecting stable clusters using principal component analysis," in *Methods in Molecular Biology*, M. J. Brownstein and A. Kohodursky, Eds., pp. 159–182, Humana Press, Totowa, NJ, USA, 2003.

[7] S. Monti, P. Tamayo, J. Mesirov, and T. R. Golub, "Consensus clustering: a resampling-based method for class discovery and vizualization of gene expression microarray data," *J. Mach. Learn. Res.*, vol. 52, no. 1-2, pp. 91–118, 2003.

[8] A. Almudevar and C. Field, "Estimation of single-generation sibling relationships based on DNA markers," *J. Agric. Biol. Environ. Stat.*, vol. 4, no. 1, pp. 136–165, 1999.

[9] E. B. Fowlkes and C. L. Mallows, "A method for comparing two hierarchical clusterings," *J. Amer. Statist. Assoc.*, vol. 78, no. 383, pp. 553–569, 1983.

[10] L. Kaufman and P. J. Rousseeuw, *Finding Groups in Data: an Introduction to Cluster Analysis*, John Wiley, NY, USA, 1990.

[11] M. B. Eisen, P. T. Spellman, P. O. Brown, and D. Botstein, "Cluster analysis and display of genome-wide expression patterns," *Proc. Natl. Acad. Sci. USA*, vol. 95, no. 25, pp. 14863–14868, 1998.

[12] B. S. Everitt, *Cluster Analysis*, Edward Arnold, London, UK, 3rd edition, 1993.

[13] T. R. Golub, D. K. Slonim, P. Tamayo, et al., "Molecular classification of cancer: class discovery and class prediction by gene

expression monitoring," *Science*, vol. 286, no. 5439, pp. 531–537, 1999.

[14] F. Azuaje and N. Bolshakova, "Clustering genomic expression data: design and evaluation principles," in *A Practical Approach to Microarray Data Analysis*, D. P. Berrar, W. Dubitzky, and M. Granzow, Eds., pp. 230–245, Kluwer Academic Publishers, Mass, USA, 2002.

[15] N. Bolshakova and F. Azuaje, "Cluster validation techniques for genome expression data," *Signal Processing*, vol. 83, no. 4, pp. 825–833, 2003.

[16] M. Granzow, D. Berrar, W. Dubitzky, A. Schuster, F. Azuaje, and R. Elis, "Tumor classification by gene expression profiling: comparison and validation of five clustering methods," *ACM-SIGBIO Newsletters*, vol. 21, no. 1, pp. 16–22, 2001.

[17] W. Krzanowski and Y. Lai, "A criterion for determining the number of groups in a dataset using sum of squares clustering," *Biometrics*, vol. 44, pp. 23–34, 1985.

[18] J. A. Hartigan, "Statistical theory in clustering," *J. Classification*, vol. 2, pp. 63–76, 1985.

[19] R. Calinski and J. Harabasz, "A dendrite method for cluster analysis," *Commun. Statistics*, vol. 3, no. 1, pp. 1–27, 1974.

[20] R. Tibshirani, G. Walther, and T. Hastie, "Estimating the number of clusters in a dataset via the Gap statistic," *J. R. Stat. Soc. Ser. B Stat. Methodol.*, vol. 63 (pt 2), pp. 411–423, 2001.

[21] W. M. Rand, "Objective criteria for the evaluation of clustering methods," *J. Amer. Statist. Assoc.*, vol. 66, no. 336, pp. 846–850, 1971.

[22] P. Jaccard, "Nouvelles recherches sur la distribution florale," *Bul. Soc. Vaud. Sci. Nat.*, vol. 44, pp. 223–270, 1908.

[23] G. W. Milligan and D. A. Schilling, "Asymptotic and finite sample characteristics of four external criterion measures," *Multivariate Behav. Res.*, vol. 20, pp. 97–109, 1985.

[24] D. Gusfield, "Partition-distance: a problem and class of perfect graphs arising in clustering," *Inform. Process. Lett.*, vol. 82, no. 3, pp. 159–164, 2002.

[25] L. Hubert and P. Arabie, "Comparing partitions," *J. Classification*, vol. 2, pp. 193–218, 1985.

[26] L. Morey and A. Agresti, "The measurement of classification agreement: an adjustment to the Rand statistic for the chance agreement," *Educ. Psychol. Meas.*, vol. 44, pp. 33–37, 1984.

[27] G. W. Milligan and M. C. Cooper, "A study of the comparability of external criteria for hierarchical cluster analysis," *Multivariate Behav. Res.*, vol. 21, pp. 441–458, 1986.

[28] G. W. Milligan, "An algorithm for generating artificial test clusters," *Psychometrika*, vol. 50, no. 1, pp. 123–127, 1985.

[29] C. Edelbrock, "Mixture model tests of hierarchical clustering algorithms: the problem of classifying everybody," *Multivariate Behav. Res.*, vol. 14, pp. 367–384, 1979.

[30] M. Smolkin and D. Ghosh, "Cluster stability scores for microarray data in cancer studies," *BMC Bioinformatics*, vol. 4, no. 1, pp. 36, 2003.

[31] D. S. Mitrinovic and P. M. Vasic, *Analytic Inequalities*, Springer-Verlag, NY, USA, 1970.

Ciprian Doru Giurcăneanu was born in Birlad, Romania, in 1968. He received the M.S. degree in control engineering from the Department of Control and Computers, "Politehnica" University of Bucharest, Romania, in 1993, and the Ph.D. degree in signal processing (with honors) from the Department of Information Technology, Tampere University of Technology, Finland, in 2001. From 1993 to 1997, he was a Junior Assistant at "Politehnica" University of Bucharest. Since 1997, he has been with the Institute of Signal Processing, Tampere University of Technology, where he is currently a Senior Researcher. From September 2002 to August 2003, he was a Research Fellow with the Academy of Finland. His current research interests include genomics and lossless signal compression.

Ioan Tăbuş received the M.S. degree in control systems and computers in 1982 from "Politehnica" University of Bucharest, Romania, the Ph.D. degree in control systems in 1993 from "Politehnica" University of Bucharest, and the Ph.D. degree in signal processing (with honors) in 1995 from Tampere University of Technology (TUT), Finland. He was a Teaching Assistant from 1984 to 1990, Lecturer from 1990 to 1993, and Associate Professor from 1994 to 1995 in the Department of Control and Computers, "Politehnica" University of Bucharest. From 1996 to 1999, he was a Senior Researcher at TUT. Since January 2000, he has been a Professor in the Institute of Signal Processing at TUT. His research interests include genomic signal processing, speech, audio, image, and data compression, joint source and channel coding, nonlinear signal processing, and image processing. He is the coauthor of two books and more than 90 publications in the fields of signal compression, image processing, and system identification. He is a Senior Member of IEEE and Associate Editor for IEEE Transactions on Signal Processing. He was Chair of IEEE SP/CAS Chapter of Finland Section. Dr. Tabus is a corecipient of 1991 "Traian Vuia" Award of the Romanian Academy and corecipient of the NSIP 2001 Best Paper Award.