07 May 2024

ARCHIVIO ISTITUZIONALE
DELLA RICERCA

Alma Mater Studiorum Università di Bologna
Archivio istituzionale della ricerca

An empirical comparison and characterisation of nine popular clustering methods

This is the final peer-reviewed author's accepted manuscript (postprint) of the following publication:

Published Version:
An empirical comparison and characterisation of nine popular clustering methods / Christian Hennig. - In: ADVANCES IN DATA ANALYSIS AND CLASSIFICATION. - ISSN 1862-5355. - STAMPA. - 16:1 (March)(2022), pp. 201-229. [10.1007/s11634-021-00478-z]

This version is available at: https://hdl.handle.net/11585/898140 since: 2022-10-29

Published:

DOI: http://doi.org/10.1007/s11634-021-00478-z

Terms of use:

Some rights reserved. The terms and conditions for the reuse of this version of the manuscript are specified in the publishing policy. For all terms of use and more information see the publisher's website.

(Article begins on next page)

This item was downloaded from IRIS Università di Bologna (https://cris.unibo.it/).
When citing, please refer to the published version.

Noname manuscript No.
(will be inserted by the editor)

An empirical comparison and characterisation of nine popular clustering methods

Christian Hennig

Received: date / Accepted: date

Abstract Nine popular clustering methods are applied to 42 real data sets. The aim is to give a detailed characterisation of the methods by means of several cluster validation indexes that measure various individual aspects of the resulting clusters such as small within-cluster distances, separation of clusters, closeness to a Gaussian distribution etc. as introduced in [29]. 30 of the data sets come with a "true" clustering. On these data sets the similarity of the clusterings from the nine methods to the "true" clusterings is explored. Furthermore, a mixed effects regression relates the observable individual aspects of the clusters to the similarity with the "true" clusterings, which in real clustering problems is unobservable. The study gives new insight not only into the ability of the methods to discover "true" clusterings, but also into properties of clusterings that can be expected from the methods, which is crucial for the choice of a method in a real situation without a given "true" clustering.

Keywords Cluster benchmarking · internal cluster validation · external cluster validation · mixed effects model

1 Introduction

This work compares cluster analysis methods empirically on 42 real data sets. 30 of these data sets come with a given "true" classification. The principal aim is to explore how different clustering methods produce solutions with different data analytic characteristics, which can help a user choosing an appropriate method for the research question of interest. This does not require the knowledge of a "true" clustering. The performance of the methods regarding recovery of the "truth" is reported, but is not the main focus.

C. Hennig
Dipartimento di Scienze Statistiche "Paolo Fortunati"
Universitá di Bologna
E-mail: christian.hennig@unibo.it

Cluster analysis plays a central role in modern data analysis and is applied in almost every field where data arise, be it finance, marketing, genetics, medicine, psychology, archaeology, social and political science, chemistry, engineering, or machine learning. Cluster analysis can have well-defined research aims such as species delimitation in biology, or be applied in a rather exploratory manner to learn about potentially informative structure in a data set, for example when clustering the districts of a city. New cluster analysis methods are regularly developed, often for new data formats, but also to fix apparent defects of already existing methods. One reason for this is that cluster analysis is difficult, and all methods, or at least those with which enough experience has been collected, are known to "fail" in certain, even fairly regular and non-pathological, situations, where "failing" is often taken to mean that a certain pre-specified "true" clustering in data is not recovered.

A key problem with clustering is that there is no unique and generally accepted definition of what constitutes a cluster. This is not an accident, but rather part of the nature of the clustering problem. In real applications there can be different requirements for a good clustering, and different clusterings can qualify as "true" on the same data set. For example, crabs can be classified according to species, or as male or female; paintings can be classified according to style of the painter or according to the motif; a data set of customers of a company may not show any clusters that are clearly separated from each other, but may be very heterogeneous, and the company may be interested in having homogeneous subgroups of customers in order to better target their campaigns, but the data set may allow for different groupings of similar quality; in many situations with given "true" classes, such as companies that go bankrupt in a given period vs. those that do not, there is no guarantee that these "true" classes correspond to patterns in the data that can be found at all. One could even argue that in a data set that comes with a supposedly "true" grouping a clustering that does *not* coincide with that grouping is of more scientific interest than reproducing what is already known.

Rather than being generally better or worse, different cluster analysis methods can be seen as each coming with their own implicit definition of what a cluster is, and when cluster analysis is to be applied, the researchers have to decide which cluster concepts are appropriate for the application at hand. Cluster analysis can have various aims, and these aims can be in conflict with each other. For example, clusters that are well separated by clear density gaps may involve quite large within-cluster distances, which may be tolerable in some applications but unacceptable in others. Clusters that can be well represented by cluster centroids may be different from those that correspond to separable Gaussian distributions with potentially different covariance matrices, which in some applications are interpreted as meaningfully different data subsets. See [3, 43, 27, 26] for the underlying philosophy of clustering.

The starting point of this work is the collection of cluster validation indexes presented in [29]. These are indexes defined in order to provide a multivariate characterisation of a clustering, individually measuring aspects such as between-cluster separation, within-cluster homogeneity, or representation of

the overall dissimilarity structure by the clustering. They are applied here in order to give general information about how the characteristics of clusterings depend on the clustering method.

Many cluster validation indexes have been proposed in the literature, often in order to pick an optimal clustering in a given situation, e.g., by comparing different numbers of clusters, see [24] for an overview. Most of them (such as the Average Silhouette Width, [38]) attempt to assess the quality of a clustering overall by defining a compromise of various aspects, particularly within-cluster homogeneity and between-cluster separation. Following [29] and [5], the present work deviates from this approach by keeping different aspects separate in order to inform the user in a more detailed way what a given clustering achieves.

A number of benchmark studies for cluster analysis have already been published. Most of them focus on evaluating the quality of clusterings by comparing them to given "true" clusterings. This has been done for artificially generated data (e.g., [50,15,60,56,55]; see [52] for an overview of earlier work), for real data, mostly focusing on specific application areas or types of data (e.g., [59,40,13,42]), or a mixed collection of real and artificial data, sometimes generating artificial data from models closely derived from a real application (e.g., [49,45,18,10,36]). An exception is [34], where different clustering methods were mapped according to the similarity of their clusterings on various data sets (something similar is done here, see Section 3.1). [7] contrasted recovery of a "true" classification in artificial data sets with the requirement of having homogeneous clusters. [51] ran a study to compare different internal validation indexes according to their ability to correlate with the similarity between a clustering of the data produced by a clustering method and a "true" clustering, which has some connection to Section 3.3 here.

All of these studies attempt to provide a neutral comparison of clustering methods, which is to be distinguished from the large number of studies, using real and artificial data, that have been carried out by method developers in order to demonstrate that their newly proposed method compares favourably with existing methods. Due to selection effects, the results of such work, although of some value in their own right, cannot be taken as objective indicators of the quality of methods ([14,28]). The study presented here is meant to be neutral; I have not been involved in the development of any of the compared methods, and have no specific interest to portray any of them as particularly good or bad. No selections have been made depending on results ([12]); the 42 data sets from which results are reported are all that were involved.

A different line of work focuses on elaborating characteristics of different clustering methods theoretically, see [35,22,39,1,3,2,17].

Section 2 explains the design of the study, i.e., the clustering methods, the data sets, and the validation indexes. Section 3 presents the results, starting with the characterisation of the methods in terms of the internal indexes, then results regarding the recovery of the "true" clusters, and ultimately connecting "true" cluster recovery with the characteristics of the clustering solutions using a mixed effects regression model. A discussion concludes the paper.

2 Study design

For the study design, recommendations for benchmark studies as given, e.g., in [12,61] have been taken into account. One important issue is a definition of the scope of the study. There is an enormous amount of clustering methods, and clustering is applied to data of very different formats. It is not even remotely possible to cover everything that could potentially be of interest. Therefore the present study constrains its scope in the following way:

- Only clustering methods for $2 \leq p$-dimensional Euclidean data that can be treated as continuous are used. Methods that work with dissimilarities are run using the Euclidean distance.
- Accordingly, data sets contain numerical variables only. Some data sets include discrete variables, which are treated as admissible for the study if they carry numerical information and take at least three different values (variables taking a small number of values, particularly three or four, are very rare in the study).
- The number of clusters is always treated as fixed. Only methods that allow to fix the number of clusters are used; methods to estimate the number of clusters are not involved. For data sets with a given "true" clustering, the corresponding number of clusters was taken. For data sets without such information, a number of clusters was chosen subjectively considering data visualisation and, where possible, subject matter information.
- The included clustering methods were required to have an R-implementation that can be used in a default way without additional tuning in order to allow for a comparison that is not influenced by different tuning flexibilities.
- No statistical structure (such as time series or regression clustering) is taken into account, and neither is any automatic dimension reduction involved as part of any method. All data is treated as plain p-dimensional Euclidean.
- Methods are only admissible for the study if they always produce crisp partitions. Every observation always is classified (also in the given "true" clusterings) to belong to one and only one cluster.

2.1 Clustering methods

The involved clustering methods are all well established and widely used, as far as my knowledge goes. They represent the major classes of clustering methods listed in [31] with the exception of density-based clustering, which was excluded because standard density-based methods such as DBSCAN ([20]) do not accept the number of clusters as input and often do not produce partitions. Another popular method that was not involved was Ward's method, as this is based on the same objective function as K-means and can be seen as just another technique to optimise this function locally ([21]). On the other hand, including mixtures of t- and skew t-distributions means that mixture model-based clustering is strongly represented. The motivation for this is that

the other included methods are not meant to fit distributional shapes including outliers and skewness, which may be widespread in practice; alternatives would be methods that have the ability to not include observations classified as outliers in any cluster, but this is beyond the scope of the present study. Here are the included methods.

K-means as implemented in the R-function `kmeans` using the algorithm by [25].

Partitioning Around Medoids (clara) ([38]) as implemented in the R-function `claraCBI` (therefore abbreviated "clara" in the results) in R-package `fpc` ([30]), which calls function `pam` in R-package `cluster` ([44]) using (unsquared) Euclidean distances.

Gaussian mixture model (mclust) fitted by Maximum Likelihood using the EM-algorithm, where the best of various covariance matrix models is chosen by the Bayesian Information Criterion (BIC) ([23]) as implemented in the R-function `mclustBIC` in R-package `mclust` ([57]).

Mixture of skew t-distributions (emskewt) fitted by Maximum Likelihood using the EM-algorithm ([41]), including fully flexible estimation of the degrees of freedom and the shape matrix, as implemented in the function `EmSkew` with parameter `distr="mst"` in the R-package `EMMIXskew` ([62]).

Mixture of t-distributions (teigen) fitted by Maximum Likelihood using the EM-algorithm ([46]), where the best of various covariance matrix models is chosen by the BIC ([8]) as implemented in the R-function `teigen` in R-package `teigen` ([9]).

Single linkage hierarchical clustering as implemented in the R-function `hclust` and the dendrogram cut at the required number of clusters to produce a partition, as is done also for the other hierarchical methods. See [21] for an explanation and historical references for all involved hierarchical methods.

Average linkage hierarchical clustering as implemented in the R-function `hclust`.

Complete linkage hierarchical clustering as implemented in the R-function `hclust`.

Spectral clustering ([53]) as implemented in the R-function `specc` in R-package `kernlab` ([37]).

The functions were mostly run using the default settings. In some cases, e.g., `hclust`, parameters had to be provided in order to determine which exact method was used. Some amendments were required. In particular, all methods were run in such a way that they would always deliver a valid partition as a result. See Supplementary material S1 for more computational detail.

2.2 Data sets

The data sets used in this study are a convenience sample, collected from mostly well known benchmark data sets in widespread use together with some data sets that I have come across in my work. 21 data sets are from the UCI repository ([19]), further ones are from Kaggle, `www.openml.org`, example data sets of R-packages, open data accompanying books and research papers, and

Table 1 Numbers of observations for the 42 data sets.

Observations	Number of data sets
$n \leq 100$	5
$100 < n \leq 200$	6
$200 < n \leq 300$	8
$300 < n \leq 500$	5
$500 \leq n < 1000$	7
$1000 \leq n < 2000$	6
$n > 2000$	5

some were collected by myself or provided to me by collaborators and advisory clients with permission to use them. Details about the data sets are given in Supplementary material S2.

There were some criteria on top of those stated above according to which data sets have been selected, which define the scope of the study. There was a target number of collecting at least 30 data sets with and at least 10 data sets without given "true" classes; ultimately there are 30 data sets with and 12 data sets without true classes. The aim was to cover a large range of application areas, although due to the availability of data sets, this has not been perfectly achieved. 17 of the data sets come from the related areas of biology, genetics, medicine, and chemistry. Eight are from the social sciences, two from finance, eight can be classified as engineering including typical pattern recognition tasks, the remaining seven data sets come from miscellaneous areas.

As some of the clustering methods cannot handle data with a smaller number of observations n than the number of variables p within clusters, all data sets have p substantially smaller than n. The calibration of validation indexes requires repeated computations based on $n \times n$ distance matrices (see Section 2.3), for this reason the biggest data set has $n = 4601$, and generally data sets with $n < 3000$ were preferred. The maximum p is 72. $p = 1$ is excluded, as it could not be handled by two methods. The maximum number of "true" clusters K is 100. Data sets without missing values were preferred, but some data sets with a very small number of missing values were admitted. In these cases mean imputation was used. Tables 1, 2, and 3 show the distributions of n, p, and K, respectively, over the data sets.

The variables were scaled to mean 0 and variance 1 before clustering, except for data sets in which the variables have compatible units of measurement and there seems to be a subject matter justification to make their impact for clustering proportional to the standard deviation. See Supplementary material S2 for details on the preprocessing for some data sets.

An issue with the representativity of these data sets for real clustering problems is that the availability of "true" clusterings constitutes a difference to the real unsupervised problems to which clustering is usually applied. This is an issue with almost all collections of data sets for benchmarking clustering algorithms. In particular, several such data sets have been constructed in order to have all clusters represented by the same number of observations. This is the case for eight of the 30 data sets with "true" clusterings used here

Table 2 Numbers of variables for the 42 data sets

Variables	Number of data sets
$p = 2$	2
$p = 4$	5
$p = 5$	5
$6 \leq p \leq 8$	6
$9 \leq p \leq 11$	11
$12 \leq p \leq 20$	6
$21 \leq p \leq 50$	4
$p > 50$	3

Table 3 Numbers of clusters for the 30 data sets with given "true" clusterings, and for the 12 data sets without "true" clusterings, as chosen by the author.

Number of clusters	With "true" clustering	Without "true" clustering
$k = 2$	8	1
$k = 3$	3	3
$k = 4$	3	1
$k = 5$	2	6
$6 \leq k \leq 7$	5	1
$8 \leq k \leq 11$	6	0
$k > 11$	3	0

(seven of these have exactly equal cluster sizes). This is not possible for unsupervised problems in practice. Such data sets will favour methods that tend to produce clusters of about equal sizes. Furthermore, there is no guarantee that the "true" clusterings correspond to data subsets that can be clearly distinguished, for example by separation. The "clusterability" of some of them may be questioned ([4]), but this is arguably anyway often the case with real clustering applications.

2.3 Internal validation indexes

Internal validation indexes are used here with the aim of measuring various aspects of a clustering that can be seen as desirable, depending on the specific application. It is then investigated to what extent the different clustering methods work well according to these aspects. [26] lists and discusses a number of aspects that can be relevant. [29] and [5] formalised many of these aspects, partly using already existing indexes, partly introducing new ones. Here the indexes used in the present study are listed. For more background and discussion, including possible alternatives, see [29] and [5]. The indexes attempt to formalise clustering aspects in a direct intuitive manner, without making reference to specific models (unless it is of interest whether data look like generated by a particular probability model, see below). The indexes as defined here do not allow comparison between or aggregation over different data sets. In order to do this, they need to be calibrated, which is treated in Section 2.4.

The data set is denoted as $\mathcal{D} = \{x_1, \ldots, x_n\}$. Here the observations $x_1, \ldots, x_n$ are assumed to be $\in \mathbb{R}^p$, and $d(x, y)$ is the Euclidean distance between x and

y, although the indexes can be applied to more general types of data and distances. A clustering is a set $\mathcal{C} = \{C_1, \ldots, C_K\}$ with $C_j \subseteq \mathcal{D},\ j = 1, \ldots, K$. For $j = 1, \ldots, K$, $n_j = |C_j|$ is the number of objects in C_j. Assume $\mathcal{C}$ to be a partition, e.g., $j \neq k \Rightarrow C_j \cap C_k = \emptyset$ and $\bigcup_{j=1}^{K} C_j = \mathcal{D}$. Let $\gamma :\ \{1, \ldots, n\} \mapsto \{1, \ldots, K\}$ be the assignment function, i.e., $\gamma(i) = j \Leftrightarrow x_i \in C_j$.

Average within-cluster distances (avewithin; aw; [5]). This index measures homogeneity in the sense of small distances within clusters. Smaller values are better.

$$I_{avewithin}(\mathcal{C}) = \frac{1}{n}\sum_{k=1}^{K}\frac{1}{n_k - 1}\sum_{x_i \neq x_j \in C_k} d(x_i, x_j).$$

Representation of cluster members by centroids. In some applications cluster centroids are used in order to represent the clustered objects, and an important aim is that this representation is good for all cluster members. This is directly formalised by the objective functions of K-means (sum of squared distances from the cluster mean) and Partitioning Around Medoids (sum of distances from the cluster medoid). Both of these criteria have been used as internal validation indexes in the present study, however results are not presented, because over all results both of these turn out to have a correlation of larger than 0.95 with $I_{avewithin}$, so $I_{avewithin}$ can be taken to measure this clustering aspect as well.

Maximum diameter (maxdiameter; md). In some applications there may be a stricter requirement that large distances within clusters cannot be tolerated, rather than having only the distance average small. This can be formalised by

$$I_{maxdiameter}(\mathcal{C}) = \max_{C \in \mathcal{C}; x_i, x_j \in C} d(x_i, x_j).$$

Smaller values are better.

Widest within-cluster gap (widestgap; wg; [29]). Another interpretation of cluster homogeneity is that there should not be different parts of the same cluster that are separated from each other. This can be formalised by

$$I_{widestgap}(\mathcal{C}) = \max_{C \in \mathcal{C}, D, E:\ C = D \cup E} \min_{x \in D, y \in E} d(x, y).$$

Smaller values are better.

Separation index (sindex; si; [29]). This index measures whether clusters are separated in the sense that the closest distances between clusters are large. For every object $x_i \in C_k$, $i = 1, \ldots, n$, $k \in 1, \ldots, K$, let $d_{k:i} = \min_{x_j \notin C_k} d(x_i, x_j)$. Let $d_{k:(1)} \leq \ldots \leq d_{k:(n_k)}$ be the values of $d_{k:i}$ for $x_i \in C_k$ ordered from the smallest to the largest, and let $[pn_k]$ be the largest integer $\leq pn_k$. p is a parameter tuning what proportion of observations counts as close to the border of a cluster with another. Here, $p = 0.1$. Then,

$$I_{sindex}(\mathcal{C}; p) = \frac{1}{\sum_{k=1}^{K}[pn_k]}\sum_{k=1}^{K}\sum_{i=1}^{[pn_k]} d_{k:(i)}.$$

Larger values are better.

Analogously to the maximum diameter, the minimum separation, i.e., the minimum distance between any two clusters may also be of interest. In the present study, this has a correlation of 0.93 with I_{sindex}, and results for the minimum separation are omitted for reasons of redundancy.

Pearson-version of Hubert's Γ (pearsongamma; pg; [33]). This index measures to what extent the clustering corresponds or represents the distance structure in the data. Let $\mathbf{d} = \text{vec}([d(x_i, x_j)]_{i<j})$ be the vector of pairwise distances. Let $\mathbf{c} = \text{vec}([c_{ij}]_{i<j})$, where $c_{ij} = 1(\gamma(i) \neq \gamma(j))$, and $1(\bullet)$ denotes the indicator function, be a vector of clustering induced dissimilarities. With r denoting the sample Pearson correlation,

$$I_{Pearson\Gamma}(\mathcal{C}) = r(\mathbf{d}, \mathbf{c}).$$

Larger values are better. This is one version of a family of indexes introduced in [33], sometimes referred to as "Hubert's Γ".

Density mode index (dmode; dm). An intuitive idea of a cluster is that it is associated with a density mode, and that the density goes down toward the cluster border. This is formalised by the dmode index. It is based on a simple kernel density estimator h that assigns a density value $h(x)$ to every observation. Let $q_{d,p}$ be the p-quantile of the vector of dissimilarities $\mathbf{d}$, e.g., for $p = 0.1$, the 10% smallest dissimilarities are $\leq q_{d,0,1}$. Define the kernel and density as

$$\kappa(d) = \left(1 - \frac{1}{q_{d,p}} d\right) 1(d \leq q_{d,p}), \qquad h(x) = \sum_{i=1}^{n} \kappa(d(x, x_i)).$$

The following algorithm constructs a sequence of neighbouring observations from the mode in such a way that the density should always go down, and penalties are incurred if the density goes up. It also constructs a set T that collects information about high dissimilarities between high density observations used below. $I_{densdec}$ collects the penalties.

Initialisation $I_{d1} = 0$, $T = \emptyset$. For $j = 1, \ldots, K$:

Step 1 $S_j = \{x\}$, where $x = \arg\max_{y \in C_j} h(y)$.

Step 2 Let $R_j = C_j \setminus S_j$. If $R_j = \emptyset$: $j = j + 1$, if $j \leq K$ go to Step 1, if $j + K = 1$ then go to Step 5. Otherwise:

Step 3 Find $(x, y) = \arg\min_{(z_1, z_2): z_1 \in R_j, z_2 \in S_j} d(z_1, z_2)$. $S_j = S_j \cup \{x\}$, $T = T \cup \{\max_{z \in R_j} h(z) d(x, y)\}$.

Step 4 If $h(x) > h(y)$: $I_{d1} = I_{d1} + (h(x) - h(y))^2$, back to Step 2.

Step 5 $I_{densdec}(\mathcal{C}) = \sqrt{\frac{I_{d1}}{n}}$.

It is possible that there is a large gap between two observations with high density, which does not incur penalties in $I_{densdec}$ if there are no low-density observations in between. This can be picked up by

$$I_{highdgap}(\mathcal{C}) = \max T.$$

These two indexes, which are both better for smaller values, were defined in [29], but they can be seen as contributing to the measurement of the same aspect, with $I_{highdgap}$ just adding information missed by $I_{densdec}$. An aggregate version, which is used here, can be defined as

$$I_{dmode}(\mathcal{C}) = 0.75 I^*_{densdec}(\mathcal{C}) + 0.25 I^*_{highdgap}(\mathcal{C}),$$

where $I^*_{densdec}$ and $I^*_{highdgap}$ are suitably calibrated versions of $I_{densdec}$, $I_{highdgap}$, respectively, see Section 2.4. The weights 0.75 and 0.25 in the definition of I_{dmode} can be interpreted as the relative impact of the two sub-indexes.

Cluster boundaries cutting through density valleys (denscut; dc; [29]). A complementary aspect of the idea that clusters are associated with high density regions is that cluster boundaries should run through density valleys rather than density mountains. The denscut-index penalises a high contribution of points from different clusters to the density values in a cluster (measured by h_o below).

$$\text{For } x_i,\ i = 1, \ldots, n: \ h_o(x_i) = \sum_{k=1}^{n} \kappa(d(x_i, x_k)) 1(\gamma(k) \neq \gamma(i)).$$

A penalty is incurred if for observations with a large density $h(x)$ there is a large contribution $h_o(x)$ to that density from other clusters:

$$I_{denscut}(\mathcal{C}) = \frac{1}{n} \sum_{j=1}^{K} \sum_{x \in C_j} h(x) h_o(x).$$

Smaller values are better.

Entropy (en; [58]). In many applications very small clusters are not very useful, and cluster sizes should optimally be close to uniform. This is measured by the well known entropy:

$$I_{entropy}(\mathcal{C}) = -\sum_{k=1}^{K} \frac{n_k}{n} \log(\frac{n_k}{n}).$$

Large values are good.

Gaussianity of clusters (kdnorm; nor; [16]). Due to the Central Limit Theorem and a widespread belief that the Gaussian distribution approximates many real random processes, it may be of interest in its own right to have clusters that are approximately Gaussian. The index I_{kdnorm} is defined, following [16], as the Kolmogorov distance between the empirical distribution of within-cluster Mahalanobis distances to the cluster means, and a χ^2_p-distribution, which is the distribution of Mahalanobis distances in perfectly Gaussian clusters.

Coefficient of variation of distances to within-cluster neighbours (cvnnd; cvn; [29]). Another within-cluster distributional shape of potential interest is uniformity, where clusters are characterised by a uniform within-cluster density level. This can be characterised by the coefficient of variation (CV) of the dissimilarities to the kth nearest within-cluster neighbour $d_w^k(x)$ ($k = 2$ is used here). Define for $j = 1, \ldots, k$, assuming $n_j > k$:

$$m(C_j; k) = \frac{1}{n_j} \sum_{x \in C_j} d_w^k(x), \qquad \mathrm{CV}(C_j) = \frac{\sqrt{\frac{1}{n_j - 1} \sum_{x \in C_j} (d_w^k(x) - m(C_j; k))^2}}{m(C_j; k)}.$$

Using this,

$$I_{cvdens}(\mathcal{C}) = \frac{\sum_{j=1}^{K} n_j \mathrm{CV}(C_j) 1(n_j > k)}{\sum_{j=1}^{K} n_j 1(n_j > k)}.$$

Smaller values are better.

Average Silhouette Width (asw; [38]). This is a popular internal validation index that deviates somewhat from the philosophy behind the collection of indexes presented here, because it attempts to balance two aspects of cluster quality, namely homogeneity and separation. It has been included in the study anyway, because it also uses an intuitive direct formalisation of clustering characteristics of interest.

2.4 Calibrating the indexes

For aggregating the indexes introduced in Section 2.3 over different data sets and to compare the performance of a clustering method over the indexes in order to characterise it, it is necessary to calibrate the values of the indexes, so that they become comparable. This is done as in [29, 5]. The idea is to generate a large number m of random clusterings $\mathcal{C}_{R1}, \ldots, \mathcal{C}_{Rm}$ on the data. Denote the clusterings of the $q = 9$ methods from Section 2.1 by $\mathcal{C}_1, \ldots, \mathcal{C}_q$. For a given data set $\mathcal{D}$ and index I, first change I to $-I$ in case that smaller values are better according to the original definition of I, so that for all calibrated indexes larger values are better. Then use these clusterings to standardise I:

$$m(I, \mathcal{D}) = \frac{1}{m+q} \left(\sum_{i=1}^{m} I(\mathcal{C}_{Ri}) + \sum_{i=1}^{q} I(\mathcal{C}_i) \right),$$

$$s^2(I, \mathcal{D}) = \frac{1}{m+q-1} \left(\sum_{i=1}^{m} [I(\mathcal{C}_{Ri}) - m(I, \mathcal{D})]^2 + \sum_{i=1}^{q} [I(\mathcal{C}_i) - m(I, \mathcal{D})]^2 \right),$$

$$I^*(\mathcal{C}_i) = \frac{I(\mathcal{C}_i) - m(I, \mathcal{D})}{s(I, \mathcal{D})}, \; i = 1, \ldots, q.$$

I^* is therefore scaled so that its values can be interpreted as expressing the quality (larger is better) compared to what the collection of clusterings $\mathcal{C}_{R1}, \ldots, \mathcal{C}_{Rm}, \mathcal{C}_1, \ldots, \mathcal{C}_q$ achieves on the same data set. The approach depends

on the definition of the random clusterings. These should generate enough random variation in order to work as a tool for calibration, but they also need to be reasonable as clusterings, because if all random clusterings are several standard deviations away from the clusterings provided by the standard clustering methods, the exact distance may not be very meaningful.

Four different algorithms are used for generating the random clusterings, for details see [5]. For clusterings with K clusters, these are:

Random K-centroids: Draw K observations from $\mathcal{D}$. Assign every observation to the nearest centroid.

Random nearest neighbour: Draw K observations as starting points for the K clusters. At every stage, of the observations that are not yet clustered, assign the observation x to the cluster of its nearest already clustered neighbour, where x is the observation that has the smallest distance to this neighbour.

Random farthest neighbour: As random nearest neighbour, but x is the observation that has the smallest distance to the minimum farthest cluster member.

Random average distances: As random nearest neighbour, but x is the observation that has the smallest average distance to the closest cluster.

Experience shows that these methods generate a range of clusterings that have sufficient variation in characteristics and are mostly reasonably close to the proper clustering methods (as can be seen in [5] as well as from the results of the present study). Here, 50 random clusterings from each algorithm are generated, i.e., $m = 200$. All results in Section 3 are given in terms of calibrated indexes I^*.

2.5 External validation indexes

"Truth" recovery is measured by external validation indexes that quantify the similarity between two clusterings on the same data, here the "true" one and a clustering generated by one of the clustering methods.

The probably most popular external validation index is the Adjusted Rand Index (ARI; [32]). This index is based on the relative number of pairs of points that are in the same cluster in both clusterings or in different clusters in both clusterings, adjusted for the number of clusters and the cluster sizes in such a way that its expected value under random cluster labels with the same number and sizes of clusters is 0. The maximum value is 1 for perfect agreement. Values can be negative, but already a value of 0 can be interpreted as indicating that the two clusterings have nothing to do with each other.

In some work, the ARI has been criticised, often in the framework of an axiomatic approach where it can be shown that it violates some axioms taken to be desirable, e.g., [47,6]. Alternative indexes have been proposed that fulfill the presented axioms. [47] introduced the Variation of Information (VI), which is a proper metric between partitions. This means that, as opposed to the ARI,

smaller values are better. In Section 3, the negative VI is considered so that for all considered indexes larger values are better. The VI is defined by comparing the entropies of the two clusterings with the so-called mutual information, which is based on the entropy of the intersections between two clusters from the two different clusterings. If the two clusterings are the same, the entropy of the intersections between clusters is the same as the entropy of the original clusterings, meaning that the VI is zero, its minimum value.

[6] show their axioms for an index called BCubed first proposed in [11]. This index is based on observation-wise concepts of precision and recall, i.e., what percentage of observations in the same cluster are from the same "true" class, and what percentage of observations in a different cluster is "truly" different. It takes values between 0 and 1, 1 corresponding to a perfect agreement. See [48] for further discussion and some more alternatives.

3 Results

Three issues are addressed:

- How can the clusters produced by the methods be characterised in terms of the external validation indexes?
- How do the methods perform regarding the recovery of the "true" clusterings?
- Can the recovery of the "true" clusterings be related to the internal validation indexes?

3.1 Characterisation of the methods in terms of the internal indexes

The methods can be characterised by the distribution of values of the calibrated internal validation indexes, highlighting the dominating features of the clusterings that they produce. In order to do this, parallel coordinate plots will be used that show the full results including how results belonging to the same data set depend on each other.

I decided against running null hypothesis tests due to issues of multiple testing and model assumptions; the plots allow a good assessment of to what extent differences between methods are meaningful, dominated by random variation, or borderline.

Average within-cluster distances (left side of Figure 1): The two centroid-based methods K-means and clara achieve the best results. The Gaussian and t-mixture are about at the same level as spectral clustering; complete linkage and the mixture of skew t-distributions are worse. Average linkage is behind these, and single linkage is the worst by some distance.

Maximum diameter (right side of Figure 1): Unsurprisingly, complete linkage is best; at each step it merges clusters so that the maximum diameter is the smallest possible, although it is not optimal for every single data set (the

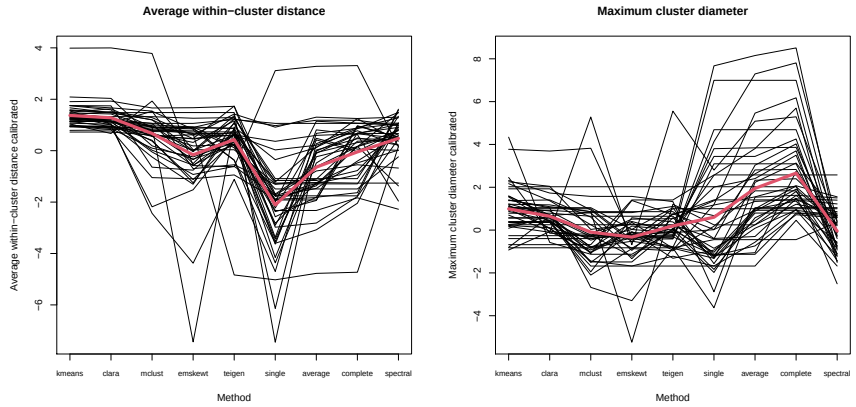

Fig. 1 Calibrated values of $I^*_{avewithin}$ and $I^*_{maxdiameter}$. Values belonging to the same data set are connected by lines. The thick red line gives the average values.

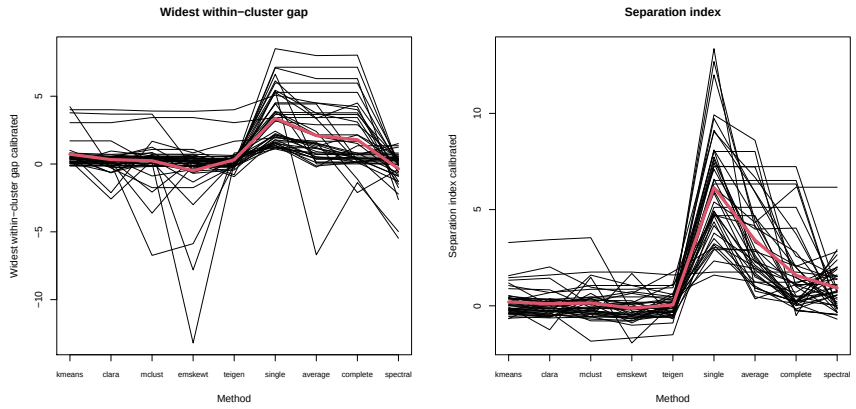

Fig. 2 Calibrated values of $I^*_{widestgap}$ and I^*_{sindex}. Values belonging to the same data set are connected by lines. The thick red line gives the average values.

hierarchical scheme will not normally produce a global optimum). Average linkage is second best, followed by K-means, clara, and single linkage, which somewhat surprisingly avoids large distances within clusters more than spectral clustering and the three mixture models. Another potential surprise is that the Gaussian mixture does not do better than the t-mixture in this respect; a flexible covariance matrix can occasionally allow for very large within-cluster distances.

Widest within-cluster gap (left side of Figure 2): The three linkage methods are best at avoiding large within-cluster gaps, with single linkage in the first place, which will not join sets between which there is a large gap. The

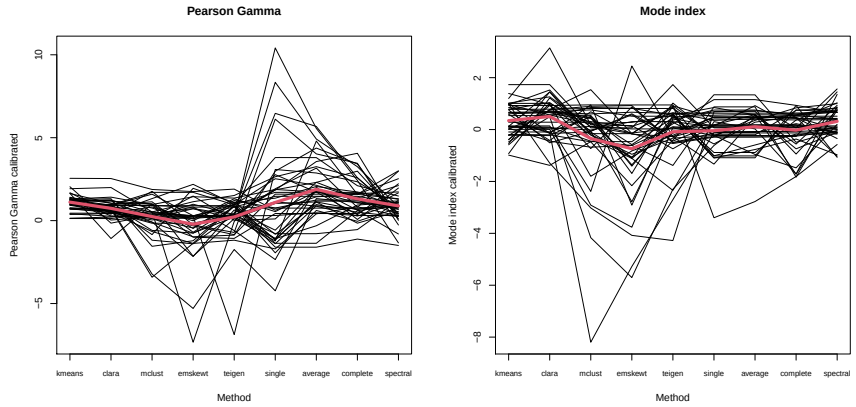

Fig. 3 Calibrated values of $I^*_{Pearson\Gamma}$ and I^*_{dmode}. Values belonging to the same data set are connected by lines. The thick red line gives the average values.

two centroid-based methods follow, however differences between them, the three mixture models, and spectral clustering look small compared to the variance, and dominated by outliers. The skew t-mixture produces very large within-cluster gaps for a number of data sets. With strong skewness there can be large distances in a tail of a cluster.

Separation index (right side of Figure 2): Single linkage achieves the best results here. Its clustering process keep separated subsets in distinct clusters (often one-point clusters with strongly separated outliers). The two other linkage methods follow. Complete linkage is sometimes portrayed as totally prioritising within-cluster homogeneity over separation, but in fact regarding separation it does better than spectral clustering, which is still a bit better than the centroid-based and the mixture models, between which differences look insignificant.

Pearson-Γ (left side of Figure 3): The average results for the methods regarding the representation of the distance structure by the clustering vary relatively little compared to the variation over data sets. Average linkage is overall best, and the skew t-mixture worst, even if the latter has good results in some data sets. Single linkage does occasionally very well, but also worse than the others for a number of data sets.

Density mode index (right side of Figure 3): Results here are dominated by variation between data sets as well. Interestingly, the methods based on mixtures of unimodal distributions do not do best here, but rather clara and spectral clustering. Once more the mixture of skew t-distributions does worst, with outliers in both directions.

Density cutting (left side of Figure 4): Due to its focus on cluster separation, single linkage is best at avoiding cutting through density mountains. The skew t- and t-mixture have the strongest tendency to put cluster boundaries in high density areas, but differences between methods are not large.

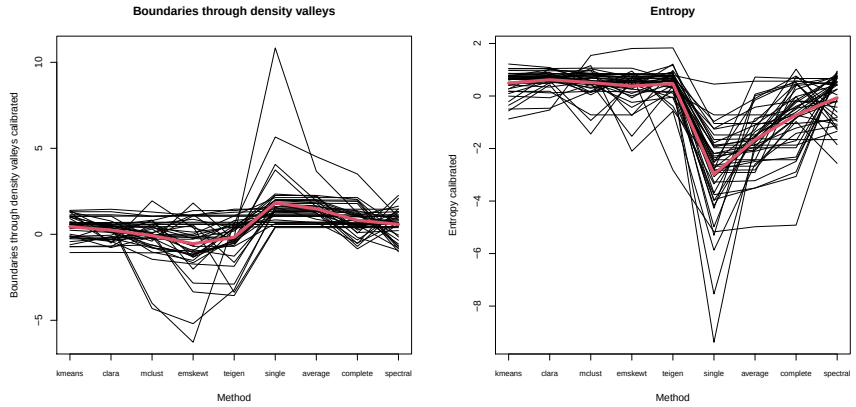

Fig. 4 Calibrated values of $I^*_{denscut}$ and $I^*_{entropy}$. Values belonging to the same data set are connected by lines. The thick red line gives the average values.

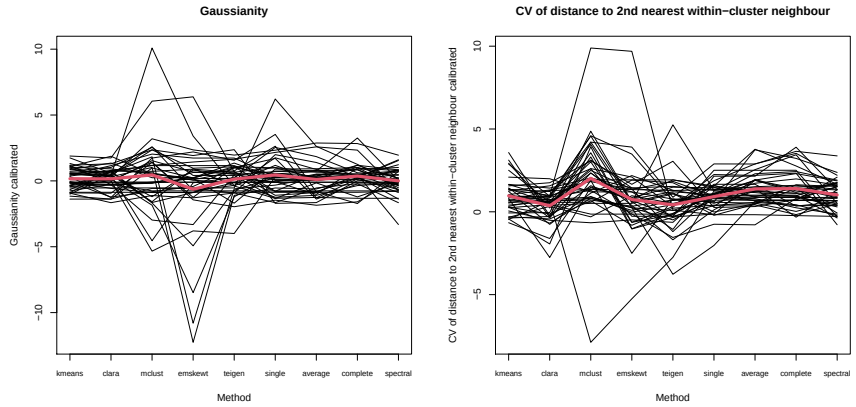

Fig. 5 Calibrated values of I^*_{kdnorm} and I^*_{cvdens}. Values belonging to the same data set are connected by lines. The thick red line gives the average values.

Entropy (right side of Figure 4): clara yields the highest average entropy followed by K-means, but differences between these and the three mixture models do not seem significant. This runs counter to the idea, sometimes found in the literature, that K-means favours similar cluster sizes more than mixtures, or even implicitly assumes them. The other four methods have a clear tendency to produce less balanced clusters, particularly single linkage, but also average and complete linkage, and to some lesser extent spectral clustering.

Gaussianity (left side of Figure 5): Although the Gaussian mixture produces on average the most Gaussian-looking clusters, as was to be expected,

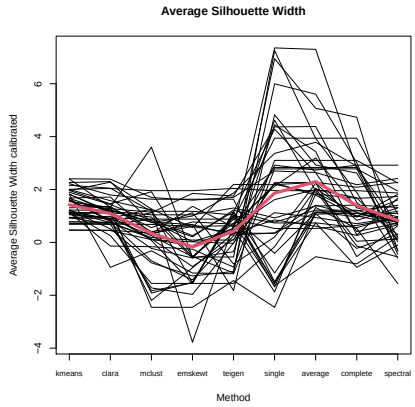

Fig. 6 Calibrated values of the ASW. Values belonging to the same data set are connected by lines. The thick red line gives the average values.

the differences between all nine methods look largely insignificant. The Gaussian mixture has positive and negative outliers, the skew t-mixture only negative ones.

CV of distances to within-cluster neighbours (right side of Figure 5): Despite one lower outlier, the Gaussian mixture tends to produce the largest cvnnd, i.e., the lowest within-cluster CVs. It probably helps that large variance clusters can bring together observations that have large distances between each other and to the rest. clara and the t-mixture produce the lowest cvnnd values. Differences between the other methods are rather small.

Average silhouette width (left side of Figure 6): Average linkage is a method that explicitly balances separation and homogeneity, and consequently it achieves the best ASW values. K-means achieves higher values than complete linkage, but the remaining methods do worse than the linkage methods. ASW had been originally proposed for use with clara ([38]), but clara does not produce particularly high ASW values, if better than the mixture models and spectral clustering.

These results characterise the clustering methods as follows:

kmeans clearly favours within-cluster homogeneity over separation. It does not favour entropy as strongly as some literature suggests; in this respect it is in line with clara and the mixture models, ahead of the remaining methods. It should be noted that entropy is treated here as a potentially beneficial feature of a clustering, whereas some literature makes it seem like a defect of kmeans that such solutions are favoured (as far as this in fact happens).

clara has largely similar characteristics to kmeans. It is slightly worse regarding the representation of the distance structure and the ASW. It is slightly better regarding clusters with density decrease from the mode. This may

have to do with the fact that the density goes down faster from the mode for the multivariate Laplace distribution (where the log-likelihood sums up unsquared distances) than for the Gaussian distribution (which corresponds to squared distances).

mclust produces clusters with the highest Gaussianity, but only by a rather insignificant distance. It is best regarding uniformity as measured by cvnnd. The reason for this is probably its ability to build clusters with large within-cluster variation collecting observations that have large distances to all or most other points, whereas other methods either need to isolate such observations in one-point clusters, or integrate them in clusters with denser cores. Mixtures of t- and skew t-distributions could in principle also produce large variance clusters, but the shapes of t- and skew t-distributions allow to integrate outlying observations more easily with denser regions.
mclust often tolerates large within-cluster distances, whereas its clusters are not on average better separated than those from K-means. On the other hand, its cluster sizes are not significantly less well balanced. Its ability to produce clusters with strongly different within-cluster variance makes it less suitable regarding Pearson-Γ and the ASW, which treat distances in the same way in all clusters.

emskewt looks bad on almost all internal indexes. It is not particularly bad regarding recovery of the "true" clusters though, see Section 3.2. This means that the current collection of internal indexes does not capture favourable characteristics of skewly distributed clusters appropriately; it also means that emskewt is not an appropriate method for finding clusters with the characteristics that are formalised by the internal indexes.

teigen has a profile that is by and large very similar to the one of mclust, apart from being slightly better regarding the maximum diameter, and slightly worse regarding Gaussianity and uniformity.

single linkage has a very distinct profile. It is best regarding separation, avoiding wide within-cluster gaps, and cluster boundaries through density valleys, and worst by some distance regarding within-cluster homogeneity and entropy. Despite the many positive features of single linkage, in practice it will often be a bad choice, as within-cluster homogeneity is desirable in most applications (if together with other features), and disregarding it completely, as single linkage does, is rarely acceptable. Also the weakness regarding entropy is of practical importance; single linkage tends to produce one very large and many very small clusters, which is often not useful.

average linkage has similar strengths and weaknesses as single linkage, but not as extreme. It is the best method regarding Pearson-Γ and the ASW, both of which balance homogeneity and separation and measure therefore how much the clustering is in line with the distance structure.

complete linkage is best regarding the maximum diameter. In most other respects it stands between single and average linkage on one side and the centroid- and mixture-based methods on the other side.

spectral is another method that provides a compromise between the rather separation-oriented single and average linkage on one side and the rather homogeneity-oriented centroid- and mixture-based methods. Its maximum cluster diameter is rather high on average. Its mode index value is good if not clearly different from the one of clara. Its mid-range entropy value may look attractive in applications in which a considerable degree of imbalance in the cluster sizes may seem realistic but the tendency of the linkage methods to produce one-point clusters should be avoided.

The multivariate characterisation of the clustering methods also allows to map them, using a principal components analysis (PCA). Results of this are shown in Figure 7. On the left side, PCs are shown using every index value for every data set as a separate value, i.e., 42*11 variables. The first two PCs carry 30.9% and 16.6% of the variance, respectively. On the right side, the PCA is performed on 11 variables that give average index values over all data sets. While this reduces information, it allows to show the indexes as axes in a biplot. The first two PCs here carry 50.0% and 19.7% of the variance, respectively. After rotation, the maps are fairly similar. Using the more detailed data set information, spectral seems much closer to kmeans and clara than to mclust and teigen, but the apparent similarity to the latter ones using average index values is an effect of dimension reduction; involving information from the third PC (not shown), the similarity structure is more similar to that of the plot using all 42*11 variables. The biplot on the right side shows the opposite tendencies of separation on one hand and entropy and average within distances on the other hand when characterising the methods, with indexes such as maximum diameter, density mode, Pearson-Γ, and the ASW opening another dimension, rather corresponding to kmeans, average, and complete linkage. Qualitative conclusions from these maps agree roughly with those in [34], where more clustering algorithms, but fewer data sets, were involved.

3.2 Recovery of "true" clusterings

The quality of the recovery of the "true" clusterings is measured by the ARI, BCubed, and the VI. Figure 8 shows the ARI-values achieved by the different clustering methods. On average, there is a clear advantage of the centroid- and mixture-based methods compared with the linkage methods (single linkage is clearly the worst), and spectral clustering is in between. Every method achieves good results on some data sets, but the linkage methods produce an ARI around zero on many data sets. Differences between kmeans, clara, mclust, emskewt, and teigen do not seem significant but are clearly dominated by variation. On some data sets all methods produce very low values, and no method achieves an ARI larger than 0.5 on more than half of the data sets. The mean ARI is 0.28, the mean ARI of the best clusterings for every data set is 0.46. Interpreting these numbers, it has to be kept in mind that the given "true" clustering does not necessarily qualify as the best clustering of the data from a data analytic point of view; some of these are neither homogeneous

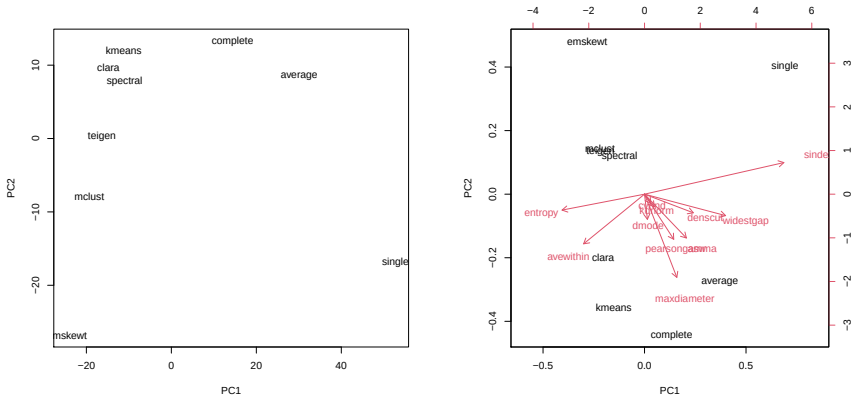

Fig. 7 Clustering methods mapped on first two principal components from using all data sets separately (left side), and from using mean values over the data sets (right side).

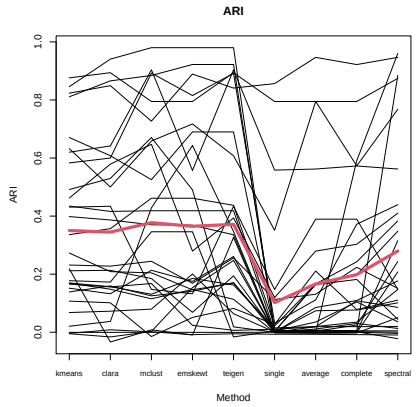

Fig. 8 Adjusted Rand Index values by method. Values belonging to the same data set are connected by lines. The thick red line gives the average values.

nor separated. Furthermore there may be meaningful clusters in the data that differ from those declared as "truth". A better recovery does not necessarily mean that a method delivers the most useful clustering that can be found. On the other hand, some given "true" clusterings correspond to clearly visible patterns in the data, and at least some methods manage to find them. Overall, the variation is quite high.

The picture changes strongly looking at the results regarding BCubed and particularly VI, see Figure 9. BCubed still shows single linkage as the weakest method, but otherwise differences look hardly significant, and according to the VI, the average quality of the methods is almost uniform.

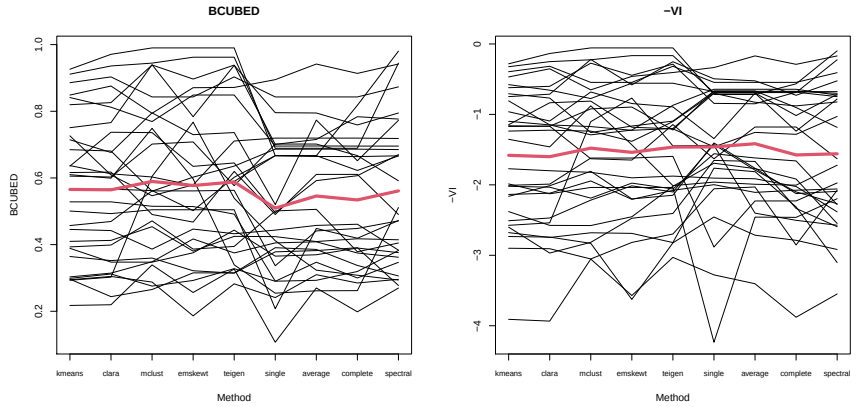

Fig. 9 BCubed and negative Variation of Information values by method. Values belonging to the same data set are connected by lines. The thick red line gives the average values.

	Clustering methods								
Index	kmeans	clara	mclust	mskewt	teigen	single	average	complete	spectral
ARI	3	4	8	5	5	0	3	1	1
BCubed	2	2	7	5	3	4	4	2	1
VI	2	1	6	3	3	11	2	1	1

Table 4 Number of times that a method comes out best according to the three external indexes.

Table 4 shows how often the different methods come out as the best according to the indexes. This portrays mclust as very successful at recovering the "truth". Spectral clustering is hardly ever on top, but it has values very close to the best for a number of data sets. Given that emskewt looks so bad regarding the internal indexes in Section 3.1, its performance regarding the external indexes looks surprisingly good. The most striking difference between the indexes is that single linkage is not the best method for a single data set with respect to the the ARI, but it is the best for 11 data sets with respect to the VI. The latter result is not to be interpreted as an endorsement of single linkage, it rather points to a weakness of the VI (and to a lesser extent the BCubed) index. This is explored in the following.

Figure 10 shows how the three indexes are related to each other over all nine clustering methods applied to the 30 data sets with "true" clusterings. VI and BCubed have a correlation ρ of -0.94, but the ARI is correlated substantially weaker to both, $\rho = 0.75$ with BCubed and $\rho = -0.57$ with VI. BCubed can therefore be seen as a compromise between the two. In order to explore what causes the differences between ARI and VI, in Figure 10 it can be seen that the major issue is that the VI can produce fairly good values close to zero for some situations in which the ARI is around zero, indicating unrelated clusterings, or only slightly better. Generally these situations tend

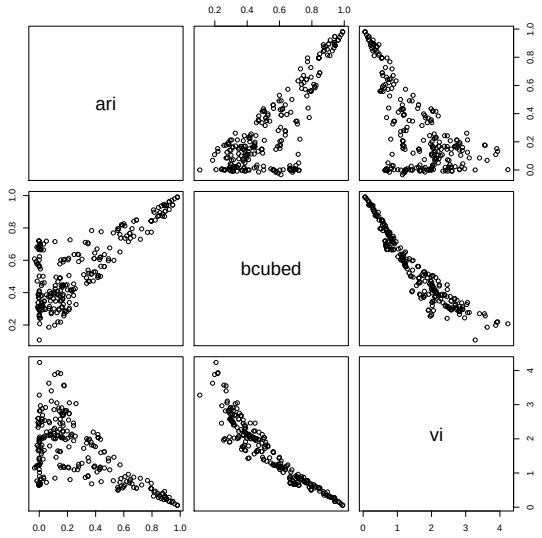

Fig. 10 Pairs plot of ARI, BCubed, and VI

Table 5 Contingency table of "true" clustering and single linkage clustering for data set "22 - Wholesale"

	Single linkage cluster	
Truth	1	2
1	297	1
2	142	0

to occur where one clustering is very imbalanced, mostly with one or more one-point clusters, whereas the other one (more often the "true" one) is not. The VI involves cluster-wise percentages of points occurring together in the same cluster in the other clustering, and therefore assesses one-point clusters favourably, whereas the random labels model behind ARI indicates that what happens with the object in a one-point cluster in another (potentially "true") clustering is random and therefore not meaningful as long as it appears in a substantially bigger cluster there.

For example, consider the data set "22 - Wholesale" (see Supplementary material S2). According to the VI, the single linkage clustering is optimal (VI= 0.64), but this has an ARI-value of about 0. It is second best according to BCubed with a value of 0.72. Table 5 shows how this is related to the "true" clustering. It is clear that any random clustering that fixes one cluster size as 1 will be about equally good. This is a rather extreme case, however most of the assessment differences between ARI and VI (and to a lesser extent BCubed) are of a similar kind. This makes the ARI look like the more appropriate index here.

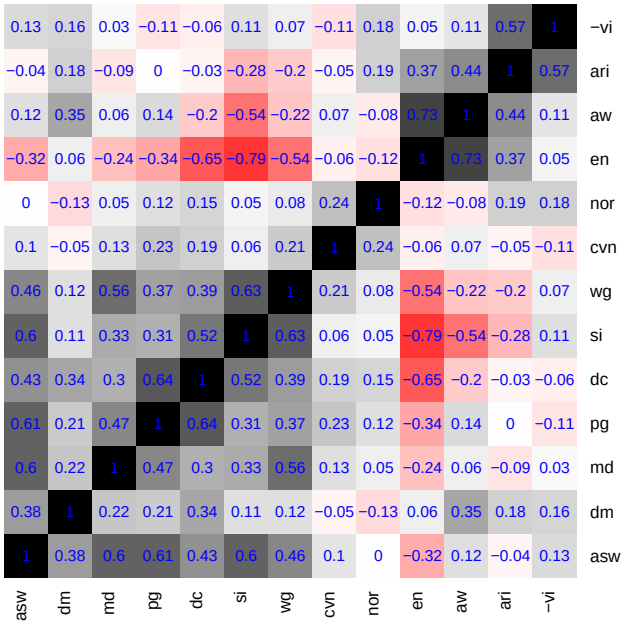

Fig. 11 Correlation matrix of internal and external validation indexes

3.3 Relating "true" cluster recovery to the internal indexes

It is of interest whether the internal index values, which are observable in a real situation, can explain to some extent the performance regarding the "true" cluster recovery. A tool to assess this is a linear regression with an external index as response, and the internal indexes as explanatory variables. There is dependence between the different clusterings on the same data set, and this can be appropriately handled using a random data set effect.

An important issue is that the internal indexes are correlated, which can make the interpretation of the regression results difficult. Figure 11 shows the correlation structure among the internal indexes, ARI and -VI (BCubed is not taken into account in this section due to the high correlation with VI). The order of indexes in Figure 11 was determined by a hierarchical clustering using correlation dissimilarity, however -VI and ARI were put on top due to their different role in the regression, and the ASW was put at the bottom. The ASW is not involved in the regression, as it is defined in order to compromise between homogeneity and separation, which themselves are represented by other internal indexes. It is involved in Figure 11 because its correlation to the other indexes may be of interest anyway. One thing that can be seen is that it is fairly strongly correlated to a number of other indexes, particularly

Table 6 Mixed-effects regression results regressing ARI, -VI, respectively, on the internal indexes excluding the ASW.

Response	ARI			-VI		
Indexes	Coefficient	t	p	Coefficient	t	p
Intercept	.324	6.91	.000	-1.54	-10.11	.000
avewithin	-.019	-1.34	.181	0.03	0.88	.377
maxdiameter	-.025	-4.03	.000	0.01	0.64	.520
widestgap	.014	2.00	.047	-0.00	-0.21	.814
sindex	-.010	-1.65	.101	0.05	3.84	.000
pearsongamma	.020	2.43	.016	-0.04	-1.86	.064
dmode	.009	0.89	.374	0.05	1.92	.056
denscut	.000	0.03	.978	-0.05	-1.80	.074
entropy	.088	4.69	.000	0.00	0.01	.990
kdnorm	.024	3.51	.001	0.02	1.44	.151
cvnnd	-.006	-0.86	.388	-0.01	-0.48	.633
random eff. (data set)			.000			.000

maximum diameter, Pearson-Γ, and the separation index, but rather weakly to the average within-cluster distances meant to formalise homogeneity.

Considerable correlation occurs between the average within-cluster distances and the entropy. Both of these are the internal indexes with the highest correlation to the ARI. This is a problem for interpretation because this means that entropy and homogeneity are confounded when explaining recovery success. Furthermore, both, entropy in particular, are strongly negatively correlated with separation, which may explain the negative correlation between separation and the ARI. There is no further high (> 0.2) correlation between either -VI or ARI and other internal indexes. It is obvious that the ARI is closer connected to entropy and homogeneity, whereas the -VI is more positively connected to separation. There are a number of further correlations among the internal indexes; separation, the density mode and cut indexes, Pearson-Γ, the maximum cluster diameter, and the absence of large within-cluster gaps are all positively connected. The Gaussianity index and the nearest neighbours CV are correlated 0.24 to each other; all their other correlations are lower.

Table 6 gives the results of two regression analyses, with ARI and -VI as responses, with a random data set effect. This has been obtained by the R-package `lme`, [54]. p-values are interpreted in an exploratory manner, as they are not precise. However, the null hypotheses of zero effect of a variable given all other variables are in the model are of interest here.

The ARI regression has maximum diameter, entropy, and Gaussianity as highly significant effects; Pearson-Γ is clearly significant at 5%-level. widestgap is borderline significant, which is potentially not meaningful given the number of tests.

The interpretation of entropy (which has the clearly largest t-value) is problematic for two reasons. Firstly, due to correlation, its coefficient may partly carry information due to avewithin. Secondly, eight data sets have artificially balanced classes, which may favour entropy among good clusterings. The regression was re-run excluding those data sets (not shown), yielding by

and large the same significances including entropy, but its t-value fell to 2.75. Even in this scenario it cannot be excluded that the collection of data sets with known "true" clusters favours entropy artificially. Gaussianity seems to be a valuable predictor for recovery of "true" classes. The maximum diameter has a negative coefficient, meaning that on average and controlled for all other indexes, a larger (therefore worse) maximum cluster diameter went with a better "truth" recovery regarding the ARI. It is however clearly correlated with Pearson-Γ and widestgap, which have positive effects.

Despite a positive relationship between ARI and -VI, the results of the VI-regression are very different, mainly because -VI can achieve high values for clusterings with very low entropy even if the "true" clustering is balanced. This means that there is no bias in favour of entropy by the data set sample; rather the VI seems biased against entropy by definition, see above. The only clearly significant index for -VI is the separation index, with a positive coefficient, which was not significant in the ARI-regression.

Plots of the fitted values of both regressions against their response variable (not shown) look satisfactorily linear. In principle, the regressions could be used to predict the ARI or VI for data sets with unknown "truth" from the observable internal indexes, but this will not work very well, due the strong data set effect.

Overall these results do not allow clear cut conclusions, due to correlation, issues with the representativity of the data sets, and the very different patterns observed for ARI and VI. The character of the "true" clusterings may just be so diverse that no general statement about which clustering characteristics allow for good recovery can be made. Preferring the ARI as external index, the only safely interpretable significance seems to be the one of Gaussianity, due to its low correlation with other indexes. Separation seems to help in terms of the VI, but this includes favouring clusterings that separate outliers as one-point clusters, arguably an issue with the VI.

4 Discussion

The aim of this study is to characterise the clustering methods in terms of the internal indexes, to learn about the recovery of "true" clusterings, both regarding the methods, and regarding characteristics that could be connected to recovery.

Regarding the characterisation of the clustering methods, the right side of Figure 7 is probably most expressive, locating the clustering methods relative to the internal indexes. Some indexes do not separate the methods very strongly. Single linkage stands out as being quite different to most other methods in many respects, although its weaknesses regard key characteristics that are normally very important in practice. On the other hand, the centroid-based methods, the mixture-based methods and spectral clustering have much in common; one surprising result is that K-means does not favour balanced cluster sizes particularly strongly, compared to the mixture-based methods.

Another result is that single and complete linkage are not opposite extremes, but rather that on most characteristics of single linkage, complete linkage is closer to single linkage, with average linkage in between, than the centroid-based and mixture-based methods. This is in line with some theoretical work, see, e.g., [3]. Gaussian mixture-based clustering stands out more by its good value regarding uniformity (cvnnd) than regarding Gaussianity of the clusters.

Regarding the recovery of "true" clusterings, there is large variation between the data sets. According to the ARI and BCubed, the Gaussian mixture is the best for the largest number of data sets. Single Linkage does badly regarding the ARI. Differences between the other methods are not that pronounced, and all of them did best in some data sets. This includes the skew t-mixture, which does not look good according to the internal indexes but better regarding the external indexes. There is currently no index, at least in the collection used here, that formalises in which sense such a mixture can yield a good clustering. This is a topic for further work. According to the VI (and to some extent BCubed), single linkage does much better, but this rather indicates a problem with the indexes than a good performance of single linkage.

Explaining the "true" cluster recovery by the internal indexes does not deliver very clear results, except that Gaussianity seems to help, which is sometimes achieved by the Gaussian mixture, but only insignificantly more often than by some other methods. A critical interpretation could be that quality according to the internal indexes does not really measure what is important for recovery. On the other hand one could argue that this shows the heterogeneity of "true" clusterings, and that there is no "one fits it all approach", neither for clustering, nor for measuring clustering quality. The given "true" clusterings are of such a nature that their recovery cannot be reliably predicted from observable cluster characteristics.

Some problems were exposed with the non-representativity of the data sets, with "true" clusterings, and with the VI (and somewhat less extreme the BCubed) index. These problems are not exclusive to the present study, and it can be hoped that these issues are on the radar whenever such benchmark studies are run. These problems affect analyses involving the "true clusterings" in particular. There is no reason to believe that the results regarding the internal validation indexes are biased for these reasons.

Key limitations of the present study are that it focuses on Euclidean distances, that the number of clusters is treated as fixed and existing methods to estimate it were not used, and that no dimension reduction methods were used. Involving these issues in a similar study would be a worthwhile project.

References

1. Ackerman, M., Ben-David, S.: Measures of clustering quality: A working set of axioms for clustering. Advances in Neural Information Processing Systems (NIPS) **22**, 121–128 (2008)

2. Ackerman, M., Ben-David, S., Branzei, S., Loker, D.: Weighted clustering. In: Proc. 26th AAAI Conference on Artificial Intelligence, pp. 858–863 (2012)
3. Ackerman, M., Ben-David, S., Loker, D.: Towards property-based classification of clustering paradigms. In: Advances in Neural Information Processing Systems (NIPS), pp. 10–18 (2010)
4. Adolfsson, A., Ackerman, M., Brownstein, N.C.: To cluster, or not to cluster: An analysis of clusterability methods. Pattern Recognition **88**, 13–26 (2019)
5. Akhanli, S.E., Hennig, C.: Comparing clusterings and numbers of clusters by aggregation of calibrated clustering validity indexes. Statistics and Computing **30**(5), 1523–1544 (2020)
6. Amigo, E., Gonzalo, J., Artiles, J., Verdejo, F.: A comparison of extrinsic clustering evaluation metrics based on formal constraints. Information retrieval **12**, 461–486 (2009)
7. Anderlucci, L., Hennig, C.: Clustering of categorical data: a comparison of a model-based and a distance-based approach. Communications in Statistics - Theory and Methods **43**, 704–721 (2014)
8. Andrews, J.L., McNicholas, P.D.: Model-based clustering, classification, and discriminant analysis via mixtures of multivariate t-distributions. Statistics and Computing **22**(5), 1021–1029 (2012)
9. Andrews, J.L., Wickins, J.R., Boers, N.M., McNicholas, P.D.: teigen: An R package for model-based clustering and classification via the multivariate t distribution. Journal of Statistical Software **83**(7), 1–32 (2018)
10. Arbelaitz, O., Gurrutxaga, I., Muguerza, J., Pérez, J.M., Perona, I.: An extensive comparative study of cluster validity indices. Pattern Recognition **46**(1), 243–256 (2013)
11. Bagga, A., Baldwin, B.: Entity-basedcross-document coreferencing using the vector space model. In: Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and the 17th International Conference on Computational Linguistics (COLING-ACL 98), pp. 79–85. ACL, Stroudsburg PE (1998)
12. Boulesteix, A.L.: Ten simple rules for reducing overoptimistic reporting in methodological computational research. Plos Computational Biology **11**, e1004191 (2015)
13. Boulesteix, A.L., Hatz, M.: Benchmarking for clustering methods based on real data: A statistical view. In: Data Science: Innovative Developments in Data Analysis and Clustering, pp. 73–82. Springer, Berlin (2017)
14. Boulesteix, A.L., Lauer, S., Eugster, M.J.A.: A plea for neutral comparison studies in computational sciences. PlosOne **8**, e61562 (2013)
15. Brusco, M.J., Steinley, D.: A comparison of heuristic procedures for minimum within-cluster sums of squares partitioning. Psychometrika **72**, 583–600 (2007)
16. Coretto, P., Hennig, C.: Robust improper maximum likelihood: tuning, computation, and a comparison with other methods for robust Gaussian clustering. Journal of the American Statistical Association **111**, 1648–1659 (2016)
17. Correa-Morris, J.: An indication of unification for different clustering approaches. Pattern Recognition **46**, 2548—-2561 (2013)
18. Dimitriadou, E., Barth, M., Windischberger, C., Hornik, K., Moser, E.: A quantitative comparison of functional mri cluster analysis. Artificial Intelligence in Medicine **31**, 57–71 (2004)
19. Dua, D., Graff, C.: UCI machine learning repository (2017). URL http://archive.ics.uci.edu/ml
20. Ester, M., Kriegel, H.P., Sander, J., Xu, X.: A density-based algorithm for discovering clusters in large spatial databases with noise. In: E. Simoudis, J. Han, U.M. Fayyad (eds.) KDD 96: Proceedings of the Second International Conference on Knowledge Discovery and Data Mining, pp. 226–231. AAAI Press, Menlo Park CA (1996)
21. Everitt, B.S., Landau, S., Leese, M., Stahl, D.: Cluster Analysis, 5th ed. Wiley, New York (2011)
22. Fisher, L., Van Ness, J.: Admissible clustering procedures. Biometrika **58**, 91–104 (1971)
23. Fraley, C., Raftery, A.E.: Model-based clustering, discriminant analysis and density estimation. Journal of the American Statistical Association **97**, 611–631 (2002)
24. Halkidi, M., Vazirgiannis, M., Hennig, C.: Method-independent indices for cluster validation and estimating the number of clusters. In: C. Hennig, M. Meila, F. Murtagh, R. Rocci (eds.) Handbook of Cluster Analysis, pp. 595–618. CRC Press (2015)

25. Hartigan, J.A., Wong, M.A.: Algorithm as 136: A k-means clustering algorithm. Applied Statistics **28**, 100–108 (1979)
26. Hennig, C.: Clustering strategy and method selection. In: C. Hennig, M. Meila, F. Murtagh, R. Rocci (eds.) Handbook of Cluster Analysis, pp. 703–730. CRC Press (2015)
27. Hennig, C.: What are the true clusters? Pattern Recognition Letters **64**, 53–62 (2015)
28. Hennig, C.: Some thoughts on simulation studies to compare clustering methods. Archives of Data Science, Series A (Online First) **5**(1), 1–21 (2018)
29. Hennig, C.: Cluster validation by measurement of clustering characteristics relevant to the user. In: C.H. Skiadas, J.R. Bozeman (eds.) Data Analysis and Applications 1: Clustering and Regression, Modeling - Estimating, Forecasting and Data Mining, pp. 1–24. ISTE Ltd., London (2019)
30. Hennig, C.: fpc: Flexible Procedures for Clustering (2020). R package version 2.2-8
31. Hennig, C., Meila, M.: Cluster analysis: An overview. In: C. Hennig, M. Meila, F. Murtagh, R. Rocci (eds.) Handbook of Cluster Analysis, pp. 1–19. CRC Press (2015)
32. Hubert, L., Arabie, P.: Comparing partitions. Journal of Classification **2**(2), 193–218 (1985)
33. Hubert, L.J., Schultz, J.: Quadratic assignment as a general data analysis strategy. British Journal of Mathematical and Statistical Psychology **29**, 190–241 (1976)
34. Jain, A.K., Topchy, A., Law, M.H.C., Buhmann, J.M.: Landscape of clustering algorithms. In: Proceedings of the 17th International Conference on Pattern Recognition (ICPR04) , Vol. 1, pp. 260–263. IEEE Computer Society Washington, DC (2004)
35. Jardine, N., Sibson, R.: Mathematical Taxonomy. Wiley, London and New York (1971)
36. Javed, A., Lee, B.S., Rizzo, D.M.: A benchmark study on time series clustering. Machine Learning with Applications **1**, 100001 (2020)
37. Karatzoglou, A., Smola, A., Hornik, K., Zeileis, A.: kernlab – an S4 package for kernel methods in R. Journal of Statistical Software **11**(9), 1–20 (2004)
38. Kaufman, L., Rousseeuw, P.J.: Finding groups in data: an introduction to cluster analysis, vol. 344. Wiley, New York (1990)
39. Kleinberg, J.: An impossibility theorem for clustering. Advances in Neural Information Processing Systems (NIPS) **15**, 463–470 (2002)
40. Kou, G., Peng, Y., Wang, G.: Evaluation of clustering algorithms for financial risk analysis using mcdm methods. Information Sciences **275**, 1–12 (2014)
41. Lee, S.X., McLachlan, G.J.: On mixtures of skew normal and skew t-distributions. Advances in Data Analysis and Classification **7**, 241–266 (2013)
42. Liu, X., Song, W., Wong, B.Y., Zhang, T., Yu, S., Lin, G.N., Di, X.: A comparison framework and guideline of clustering methods for mass cytometry data. Genome Biology **20**, 297 (2019)
43. von Luxburg, U., Williamson, R., Guyon, I.: Clustering: Science or art? JMLR Workshop and Conference Proceedings **27**, 65–79 (2012)
44. Maechler, M., Rousseeuw, P., Struyf, A., Hubert, M., Hornik, K.: cluster: Cluster Analysis Basics and Extensions (2019). R package version 2.1.0
45. Maulik, U., Bandyopadhyay, S.: Performance evaluation of some clustering algorithms and validity indices. IEEE Transactions on Pattern Analysis and Machine Intelligence **24**(12), 1650–1654 (2002)
46. McLachlan, G.J., Peel, D.: Finite Mixture Models. Wiley, New York (2000)
47. Meila, M.: Comparing clusterings—an information based distance. Journal of Multivariate Analysis **98**(5), 873 – 895 (2007)
48. Meila, M.: Criteria for comparing clusterings. In: C. Hennig, M. Meila, F. Murtagh, R. Rocci (eds.) Handbook of Cluster Analysis, pp. 619–635. CRC Press (2015)
49. Meila, M., Heckerman, D.: An experimental comparison of model-based clustering methods. Machine Learning **42**, 9–29 (2001)
50. Milligan, G.W.: An examination of the effect of six types of error perturbation on fifteen clustering algorithms. Psychometrika **45**, 325–342 (1980)
51. Milligan, G.W.: A monte carlo study of thirty internal criterion measures for cluster analysis. Psychometrika **46**, 187–199 (1981)
52. Milligan, G.W.: Clustering validation: results and implications for applied analyses. In: P. Arabie, L.J. Hubert, G.D. Soete (eds.) Clustering and Classification, pp. 341–375. World Scientific, Singapore (1996)

53. Ng, A.Y., Jordan, M.I., Weiss, Y.: On spectral clustering: Analysis and an algorithm. In: T. Dietterich, S. Becker, Z. Ghahramani (eds.) Advances in Neural Information Processing Systems 14 (NIPS 2001), pp. 1–8. NIPS (2001)
54. Pinheiro, J.C., Bates, D.M.: Mixed-Effects Models in S and S-PLUS. Springer, New York (2000)
55. Rodriguez, M.Z., Comin, C.H., Casanova, D., Bruno, O.M., Amancio, D.R., Costa, L., Rodrigues, F.A.: Clustering algorithms: A comparative approach. PloS one **14**, e0210236 (2019)
56. Saracli, S., Dogan, N., Dogan, I.: Comparison of hierarchical cluster analysis methods by cophenetic correlation. Journal of Inequalities and Applications (electronic publication) **203** (2013)
57. Scrucca, L., Fop, M., Murphy, T.B., Raftery, A.E.: mclust 5: clustering, classification and density estimation using Gaussian finite mixture models. The R Journal **8**(1), 289–317 (2016)
58. Shannon, C.E.: A mathematical theory of communication. The Bell System Technical Journal **27**(3), 379–423 (1948)
59. de Souto, M.C., Costa, I.G., de Araujo, D.S., Ludermir, T.B., Schliep, A.: Clustering cancer gene expression data: a comparative study. BMC Bioinformatics **9**, 497 (2008)
60. Steinley, D., Brusco, M.J.: Evaluating the performance of model-based clustering: Recommendations and cautions. Psychological Methods **16**, 63–79 (2011)
61. Van Mechelen, I., Boulesteix, A.L., Dangl, R., Dean, N., Guyon, I., Hennig, C., Leisch, F., Steinley, D.: Benchmarking in cluster analysis: A white paper. arXiv:1809.10496 [stat] (2018)
62. Wang, K., Ng, A., McLachlan., G.: EMMIXskew: The EM Algorithm and Skew Mixture Distribution (2018). R package version 1.0.3