Advances in Data Analysis and Classification (2023) 17:701–724
https://doi.org/10.1007/s11634-022-00521-7

REGULAR ARTICLE

Benchmarking distance-based partitioning methods for mixed-type data

Efthymios Costa[1] · Ioanna Papatsouma[1] · Angelos Markos[2]

Received: 11 April 2022 / Revised: 23 August 2022 / Accepted: 30 August 2022 /
Published online: 22 September 2022

Abstract

Clustering mixed-type data, that is, observation by variable data that consist of both continuous and categorical variables poses novel challenges. Foremost among these challenges is the choice of the most appropriate clustering method for the data. This paper presents a benchmarking study comparing eight distance-based partitioning methods for mixed-type data in terms of cluster recovery performance. A series of simulations carried out by a full factorial design are presented that examined the effect of a variety of factors on cluster recovery. The amount of cluster overlap, the percentage of categorical variables in the data set, the number of clusters and the number of observations had the largest effects on cluster recovery and in most of the tested scenarios. KAMILA, K-Prototypes and sequential Factor Analysis and K-Means clustering typically performed better than other methods. The study can be a useful reference for practitioners in the choice of the most appropriate method.

Keywords Cluster benchmarking · Partitioning · Mixed-type data · Heterogeneous data · K-Means

Mathematics Subject Classification 62H30

✉ Efthymios Costa
efthymios.costa17@imperial.ac.uk

Ioanna Papatsouma
i.papatsouma@imperial.ac.uk

Angelos Markos
amarkos@eled.duth.gr

[1] Department of Mathematics, Imperial College London, London, United Kingdom

[2] Department of Primary Education, Democritus University of Thrace, Alexandroupoli, Greece

1 Introduction

Benchmarking studies of clustering are increasingly important for guiding users and practitioners in choosing appropriate clustering approaches among an increasing number of alternatives (Van Mechelen et al. 2018). The objective of the present study is to contribute to the benchmarking literature by evaluating the performance of clustering methods for mixed-type data, that is, observation by variable data that consist of both continuous and categorical variables. Indeed, research in a variety of domains usually relies on heterogeneous or mixed-type data. In social science research, for example, data sets typically include demographic background characteristics (usually categorical variables) together with socioeconomic or psychological measures (usually continuous variables). Such heterogeneity urges for ways to guide users and practitioners in choosing appropriate clustering approaches for mixed-type data sets in order to identify distinct profiles of individuals and/or generate hypotheses, thereby contributing to a quantitative empirical methodology for the discipline.

Cluster analysis of mixed-type data sets can be a particularly challenging task because it requires to weigh and aggregate different variables against each other (Hennig and Liao 2013). One of the main issues is the choice of the most appropriate distance or model to simultaneously process both data types. Among the most simple and intuitive strategies for clustering mixed-type data is to convert all variables to a single type, continuous or categorical, via discretization, dummy-coding or fuzzy-coding. Such a strategy may lead to a significant loss of information from the original data and may consequently lead to increased bias (Foss and Markatou 2018). Another approach is to cluster observations separately for continuous and categorical variables, and then match the clusters in the two clusterings. This approach, however, ignores any dependencies that might exist between variables of different types (Hunt and Jorgensen 2011). Fortunately, a wide range of clustering algorithms has been specifically developed to deal with mixed-type data. A taxonomy of available methods can be found in Ahmad and Khan (2019) and overviews of distance or dissimilarity-based methods are given by Foss et al. (2019) and van de Velden et al. (2019).

Benchmarking studies may be performed by independent groups interested in systematically comparing existing methods or by authors of new methods to demonstrate performance improvements or other advantages over existing competitors. With regard to studies performed by independent groups, there have been several benchmarking studies of clustering for continuous data only or categorical data only (e.g., Milligan 1980; Meilă and Heckerman 2001; Ferreira and Hitchcock 2009; Saraçli et al. 2013; Boulesteix and Hatz 2017; Javed et al. 2020; Hennig 2022), whereas benchmarking studies of clustering for mixed-type data are scarce (Jimeno et al. 2021; Preud'Homme et al. 2021). We can also distinguish benchmarking studies of clustering for mixed-type data that are part of original papers where new methods are proposed (e.g., Ahmad and Dey 2007; Hennig and Liao 2013; Foss et al. 2016).

In this paper, we concern ourselves with distance or dissimilarity-based partitioning methods for mixed-type data, that is, methods that rely on explicit distances or dissimilarities between observations or between observations and cluster centroids. In the authors' understanding, these methods span three general approaches. The first approach involves computing an appropriate dissimilarity measure for mixed-type data, followed by a partitioning algorithm on the resulting dissimilarity matrix. Typically, such a dissimilarity measure can be constructed by defining and combining dissimilarity measures for each type of variable. A popular choice in this category involves the computation of Gower's (dis)similarity measure (Gower 1971) among observations and then applying K-Medoids or hierarchical clustering on the dissimilarity matrix. Instead of using Gower's dissimilarity, Hennig and Liao (2013) proposed a specific weighting scheme to more appropriately balance continuous against categorical variables. The second approach involves conducting a partitional clustering of the observation by variable data with the distances between observations and cluster centroids calculated separately for categorical and continuous variables, and combine them into a single objective function. Representative methods in this category are K-Prototypes (Huang 1997), Modha-Spangler K-Means (Modha and Spangler 2003) and Mixed K-Means (Ahmad and Dey 2007). The third approach comprises factor analysis of the variables and partitional clustering of the observations in the low-dimensional space. Factor analysis and clustering can be performed sequentially, i.e., in a two-step approach where clustering is applied to the resulting factor scores (see e.g., Dolnicar and Grün 2008) or simultaneously, where the two objectives are combined by optimizing a single convex objective function (Vichi et al. 2019). In the sequential approach, the first step usually involves Factor Analysis for Mixed Data or PCAMIX (Pagès 2014; Kiers 1991) to obtain the observation scores in a low-dimensional space and then K-Means clustering to partition the observation scores. Simultaneous approaches include extensions of Reduced K-Means (De Soete and Carroll 1994) and Factorial K-Means (Vichi and Kiers 2001) to deal with the general relevant case of mixed variables (Vichi et al. 2019).

A simulation study was conducted to compare eight distance-based partitioning methods in terms of cluster recovery performance, following recommendations provided in Boulesteix et al. (2013) and Van Mechelen et al. (2018). The involved clustering methods represent a major class of methods listed in Murtagh (2015), are well established and widely used - at least the less recent ones - while the most recently proposed have been shown in previous studies to outperform others. The study attempts to provide a neutral comparison of the methods since none of the authors have 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. The current study goes beyond previous work by considering different aspects that might affect performance (number of clusters, number of observations, number of variables, percentage of categorical variables in the data set, cluster overlap, cluster density and cluster sphericity), according to a full factorial design. The result is a concrete description of the method performance, with the goal of providing researchers with a guide to selecting the most suitable method for their study.

The remainder of this paper is structured as follows: Sect. 2 reviews benchmarking studies of clustering for mixed-type data, Sect. 3 presents the methods under comparison, Sect. 4 describes the simulation study design, Sect. 5 presents the results and Sect. 6 discusses the results and concludes the paper.

2 Related work

Most comparison studies of clustering algorithms for mixed-type data have been performed within original articles presenting new methods, usually in order to establish their superiority over classical approaches. Foss et al. (2016) conducted a small scale simulation study and analyses of real-world data sets to illustrate the effectiveness of KAMILA, a newly proposed semi-parametric method, versus Modha-Spangler K-Means, K-Means with a weighting scheme described in Hennig and Liao (2013) and two finite mixture models. They considered both normal and non-normal data sets (data following a p-generalized normal-multinomial distribution or the lognormal-multinomial distribution), with varying sample sizes (250, 500, 1000, and 10000), number of continuous variables (2, 4), number of categorical variables (1, 2, 3, and 4), level of continuous overlap (1%, 15%, 30%, and 45%), level of categorical overlap (1%, 15%, 30%, 45%, 60%, 75%, and 90%), number of clusters (2, and 4), and number of categorical levels (2, and 4). The authors verified that the findings from the artificial data set analysis are generalizable in the context of real-world applications. KAMILA performed well across all conditions.

In line with Foss et al. (2016), Markos et al. (2020) investigated the performance of three sequential dimensionality reduction and clustering approaches versus KAMILA, Modha-Spangler K-Means and Gower's (dis)similarity measure followed by Partitioning Around Medoids on simulated data with varying degree of cluster separation. More precisely, the study focused on three different scenarios; first, a scenario where both continuous and categorical variables have approximately comparable cluster overlap (i.e., contain equally useful information regarding the cluster structure), second, a scenario where continuous variables have substantially more overlap compared to categorical ones (i.e., the categorical variables are more useful for clustering purposes) and third, a case where categorical variables have substantially more overlap compared to continuous ones (i.e., the continuous variables are more useful for clustering purposes). They generated 500 data sets with 200 observations, two continuous and two categorical variables with four categories/levels each. Results showed that dimensionality reduction followed by clustering in the reduced space is an effective strategy for clustering mixed-type data when categorical variables are more informative than continuous ones with regard to the cluster structure.

More recently, Jimeno et al. (2021) compared KAMILA, K-Prototypes and Multiple Correspondence Analysis (MCA) followed by K-Means, fuzzy C-Means, Probabilistic Distance Clustering or a mixture of Student's t distributions, under different simulated scenarios. The study considered 27 simulated scenarios based on three parameters: the number of clusters (2, 5, and 7), the amount of overlap in each cluster (30%, 60%, and 80%), and the ratio of continuous variables to nominal variables (1:3, 1:1, and 3:1). The total numbers of variables and observations were fixed in each case (128 variables and 1920 observations). K-Prototypes and KAMILA performed consistently well for spherical clusters. As the number of clusters increased, the performance of MCA followed by fuzzy C-means or PD clustering worsened. MCA followed by a mixture of Student's t distributions performed well in all cases.

A recent benchmarking study by Preud'Homme et al. (2021) compared the performance of four model-based methods (KAMILA, Latent Class Analysis, Latent Class

Model, and Clustering by Mixture Modeling) and five distance/dissimilarity-based methods (Gower's dissimilarity or Unsupervised Extra Trees dissimilarity followed by hierarchical clustering or Partitioning Around Medoids, K-prototypes) on both simulated and real data. The parameters used for the simulations were the number of observations (300, 600, and 1200), the number of clusters (2, 6, and 10), the ratio between the number of continuous and categorical variables in the data, the proportion of relevant or non-noisy variables (20%, 50%, and 90%), and the degree of relevance of the variables with regard to the cluster structure (low, mild, and high, as defined by a cluster separation index in the case of continuous variables and a noise proportion introduced via resampling in the categorical variables). The authors considered seven scenarios and 1000 data sets were generated for each scenario. Results revealed the dominance of model-based over most distance or dissimilarity-based methods; this was somewhat expected since the simulated data matched the assumptions of model-based methods. K-Prototypes was the only efficient distance-based method, outperforming all other techniques for larger numbers of clusters.

It is important to outline that none of the aforementioned studies made use of a full factorial design to enhance inferential capacity in terms of disentangling the effects of each of the manipulated parameters and their interactions.

3 Benchmark methods

The present study constrains its scope to distance or dissimilarity-based methods for partitioning mixed-type data, i.e., methods that rely on explicit distances or dissimilarities between observations or between observations and cluster centroids. The methods under comparison produce crisp partitions, allow to fix the number of clusters in advance and have an R-implementation.

Two of the methods considered in the study involve the conversion of observation by variable data into observation by observation proximities. A popular choice is to calculate pairwise Gower's dissimilarities among observations (Gower 1971):

$$d_{Gower}(\boldsymbol{X}_i, \boldsymbol{X}_{i'}) = 1 - \frac{\sum_{j=1}^{p} w_j(\boldsymbol{X}_i, \boldsymbol{X}_{i'}) s_j(\boldsymbol{X}_i, \boldsymbol{X}_{i'})}{\sum_{j=1}^{p} w_j(\boldsymbol{X}_i, \boldsymbol{X}_{i'})}, \quad 1 \le i, i' \le n, \; i \neq i', \tag{1}$$

where $\boldsymbol{X}_i$, $\boldsymbol{X}_{i'}$ are distinct observations, therefore rows of an $(n \times p)$-dimensional data matrix. We denote the weight of the jth variable for the two observations by w_j; this is typically set to 1, assuming equal weight for all variables, but can also take different values for different variables based on their subject matter importance (Hennig and Liao 2013). Finally, s_j is a coefficient of similarity between the jth components of $\boldsymbol{X}_i$ and $\boldsymbol{X}_{i'}$, defined as the range-normalised Manhattan distance for continuous variables and the Kronecker delta for categorical ones. Gower's dissimilarity is very general and covers most applications of dissimilarity-based clustering to mixed-type variables.

In a discussion regarding the formal definition of ‘dissimilarity’, Hennig and Liao (2013) argue that a proper dissimilarity measure between data objects should not only aggregate the variables, but it should also make use of some weighting scheme that controls variable importance, especially for nominal variables. In fact, they propose standardising the continuous variables to unit variance, contrary to the range standardisation that is used for Gower’s dissimilarity, while they introduce a more sophisticated weighting for nominal variables. More precisely, they claim that since it holds that for two independent and identically distributed continuous random variables X_1 and X_2, standardisation to unit variance implies $\mathrm{E}\left\{(X_1 - X_2)^2\right\} = 2$, standardisation of a nominal variable should be done in such a way that the dissimilarity between the two categories of a binary variable is about equal to the aforementioned expression or less than that, if the variable has more than two categorical levels. Based on this rationale, Hennig and Liao (2013) suggest setting $\sum\limits_{i=1}^{I} \mathrm{E}\left\{(Z_{i1} - Z_{i2})^2\right\} = \mathrm{E}\left\{(X_1 - X_2)^2\right\} = 2\xi$, where Z_{i1}, Z_{i2} represent the values of the first and second data points on the dummy variables $\boldsymbol{Z}_i$, obtained after dummy coding of a nominal variable with I levels. The coefficient, ξ, is set to be equal to $1/2$ in order to avoid a clustering output that is highly dependent on the levels of the categorical variables. Thus, dummy-coded nominal variables are scaled so that they are comparable to unit variance scaled continuous variables. This is followed by constructing the Euclidean distance matrix between the observations, which is equivalent to the notion of a ‘dissimilarity matrix’, as for Gower’s dissimilarity.

Once the pairwise dissimilarities are calculated (either using Gower’s or Hennig-Liao’s measure), Partitioning Around Medoids or PAM (Kaufman and Rousseeuw 1990) is applied to the proximity matrix obtained from the previous step. The objective of PAM is to find K observations that will be representative, in the sense that they will minimise the average dissimilarity with all other points in each cluster. These are called ‘medoids’ and are analogous to the ‘centroids’ in the widely-used K-Means algorithm. In this study, the R package `cluster` (Maechler et al. 2021) was used to calculate Gower’s dissimilarity (function daisy()) and subsequently carry out PAM clustering (function pam()). The function distancefactor() of the R package `fpc` (Hennig 2020) was used for the standardisation of nominal variables based on Hennig and Liao (2013). Clustering mixed-type data with Gower’s dissimilarity and the weighting scheme of Hennig and Liao (2013) followed by PAM are herein referred to as ‘Gower/PAM’ and ‘HL/PAM’ respectively.

K-Prototypes is a clustering method introduced by Huang (1997) for dealing with data of mixed type. The K-Means algorithm can be seen as a ‘special case’ of this method, since the rationale behind it is that it seeks for a minimisation of the trace of the within cluster dispersion matrix cost function, defined as:

$$E = \sum_{l=1}^{K} \sum_{i=1}^{n} y_{il}\, d(\boldsymbol{X}_i, \boldsymbol{Q}_l). \tag{2}$$

The term y_{il} in Eq. (2) denotes the (i, l)th element of an $(n \times K)$ partition matrix, taking values 0 and 1 (1 indicating cluster membership), while $\boldsymbol{Q}_l$ is the prototype for the lth cluster. A prototype is the equivalent to a medoid for PAM or a centroid for K-Means. The distance between $\boldsymbol{X}_i$ and $\boldsymbol{Q}_l$ is denoted by $d(\boldsymbol{X}_i, \boldsymbol{Q}_l)$ and it is calculated as a combination of the squared Euclidean distance for continuous and the weighted binary indicator for categorical variables. Assuming, without loss of generality, that the first $p_r < p$ variables in our data set are continuous and the rest are categorical, this may be expressed as:

$$d(\boldsymbol{X}_i, \boldsymbol{Q}_l) = \sum_{j=1}^{p_r}(x_{ij} - q_{lj})^2 + \gamma_l \sum_{j=p_r+1}^{p} \delta(x_{ij}, q_{lj}). \tag{3}$$

In the expression above, γ_l is a weight coefficient for categorical variables in the lth cluster; setting it to zero (thus indicating the absence of categorical variables), one can recover the K-Means algorithm. For computational reasons, the value of γ_l is chosen to be the same for all clusters and it is calculated as the ratio of the variance of continuous variables to the variance of the categorical variables in the data set. For the jth categorical variable, the variance is defined as $1 - \sum_h p_{jh}^2$, where p_{jh} is the frequency of hth categorical level of the jth variable divided by n. It can be shown that the components q_{lj} $(j = 1, \ldots, p)$ of $\boldsymbol{Q}_l$ upon minimisation of (2) are given by the mean or the mode of values that the jth variable takes in the lth cluster, for j being a continuous or a categorical variable respectively. K-Prototypes was conducted in this study using the kproto() function in the R package `clustMixType` (Szepannek 2018).

Some of the shortcomings of K-Prototypes, as argued by Ahmad and Dey (2007), include the use of the mode of categorical variables while ignoring other frequent categories, the fact that the distance for categorical variables is not weighted and the need for a more 'refined' notion of categorical distance. Therefore, Ahmad and Dey (2007) proposed another K-Means-based algorithm for mixed-type data, that scales the Euclidean distance and calculates categorical distances based on the co-occurrence of categorical values (herein referred to as 'Mixed K-Means').

More precisely, the distance between two distinct categories is first calculated with respect to the rest of the variables. Say A and B are two categories of the same jth categorical variable, then the categorical distance between the two with respect a j'th categorical variable is defined as the sum of the conditional probability that the jth component of an observation $\boldsymbol{X}_i$ takes the value A, given that $x_{ij'}$ is in some subset σ of possible values of a j'th variable and the conditional probability that $x_{ij} = B$ given $x_{ij'} \notin \sigma$. In order for this to be a distance metric satisfying that the distance between two identical categorical values is zero, we subtract a unit from the sum we obtain. Notice that σ is chosen carefully among all possible subsets of values of a j'th variable so that it maximises the aforementioned sum. Then, the distance between A and B is calculated as the average of all the categorical distances with respect to all other variables. Since this notion is defined for a j'th categorical variable, Ahmad and Dey (2007) suggested a simple algorithm for the discretization of continuous variables in intervals of equal width. This discretization is also used for determining the weight of continuous variables, which is defined to be the average categorical distance between all possible combinations of the categorical levels introduced.

To conduct Mixed K-Means in this study, a distance matrix was first computed using the function distmix() from the R package `kmed` and was then supplied as the input to a K-Medoids algorithm, implemented in the function fastkmed() of the same package.

Another clustering method, that is based on K-Prototypes, is Modha-Spangler K-Means (Modha and Spangler 2003). Once again, the Euclidean distance is used for continuous variables (not scaled, unlike for Mixed K-Means), while the cosine dissimilarity is used for categorical variables. The objective function is thus given by:

$$d_{MS}(\boldsymbol{X}_i, \boldsymbol{Q}_l) = \sum_{j=1}^{p_r}(x_{ij} - q_{lj})^2 + \gamma_l \left(1 - \frac{\sum_{j=p_r+1}^{p^*} x_{ij} q_{lj}}{\sqrt{\sum_{j=p_r+1}^{p^*} x_{ij}^2} \sqrt{\sum_{j=p_r+1}^{p^*} q_{lj}^2}} \right) \tag{4}$$

which is really similar to the cost function (2) but uses a different categorical distance. For the cosine dissimilarity to be used as in Eq. (4), we need to make sure that our categorical variables are first dummy-coded, so that inner products and norms of vectors can be calculated. We also denote the total number of columns for continuous and dummy-coded categorical variables by p^*.

Modha-Spangler K-Means is a convex algorithm as both distance functions used are convex. One of its main strengths is that the coefficient γ_l is automatically determined by the algorithm, by trying to minimise the ratio of the product of the average within-cluster dispersion for continuous and categorical variables to the product of the average between-cluster dispersion for continuous and categorical variables. A much more detailed description can be found in Modha and Spangler (2003). However, a weakness of this algorithm is that it requires a brute-force approach for determining the optimal value of γ_l; usually a greedy search over a grid of values specified by the user is employed. In our case, due to computational constraints, we consider only five candidate values of γ_l, which are the values of the set $\Gamma_l = \{\frac{i}{6} : i \in [1, 5]\}$. The element of Γ_l that yields the smallest value for the objective function (4) is the one that is eventually used for γ_l. Modha-Spangler K-Means was applied in this study using the function gmsClust() of the R package `kamila` (Foss and Markatou 2018). The number of distinct cluster weightings evaluated in the brute-force search was set to 10 (the default option).

The five aforementioned clustering methods all work with the full data, in perhaps very high dimensions. Another approach to cluster analysis is the so-called ‘tandem analysis’, a term coined by Arabie (1994), which consists of a dimensionality reduction step via factor analysis, followed by a clustering of the observations in the resulting low-dimensional space (see also Dolnicar and Grün 2008). One such dimensionality reduction method, suitable for mixed-type data, is Factor Analysis for Mixed Data or FAMD (also known as Principal Component Analysis for Mixed Data) (Pagès 2014).

Dimensionality reduction in FAMD is seen as a compromise between Principal Component Analysis and Multiple Correspondence Analysis (Markos et al. 2020). The idea is that the data matrix is partitioned in such a way that all columns consisting of continuous variables are ‘stacked’ right next to an ‘indicator matrix’ or ‘complete disjunctive table' that is constructed from the categorical variables. This partition matrix is constructed by recoding the categorical variables using dummy variables. The usual standardisation process of subtracting from each column its mean and dividing by its standard deviation is used for continuous variables. Standardisation of the indicator matrix is achieved by dividing the elements of each of its columns by the square root of the proportion of observations possessing the respective category that the column represents. Then, the two standardised matrices are concatenated and standard Principal Component Analysis (PCA) is performed on the resulting matrix. As in PCA, when applying FAMD it is important to decide on the number of factors to retain.

Audigier et al. (2016) state that if the ith principal component obtained is denoted by $\boldsymbol{F}_i$, then the first principal component $\boldsymbol{F}_1$ maximises the expression:

$$\sum_{j=1}^{p_r} R^2\left(\boldsymbol{F}_i, \boldsymbol{X}_{con_j}\right) + \sum_{j=p_r+1}^{p} \eta^2\left(\boldsymbol{F}_i, \boldsymbol{X}_{cat_j}\right). \tag{5}$$

Here, R^2 represents the coefficient of determination and η^2 is the squared correlation ratio, also known as the 'Intraclass Correlation Coefficient'. Moreover, $\boldsymbol{X}_{con_j}$ and $\boldsymbol{X}_{cat_j}$ denote the jth continuous and categorical variables respectively, with j being the column index. Maximising Eq. (5) is therefore equivalent to maximising the link between continuous and categorical variables, so one may view $\boldsymbol{F}_1$ as the synthetic variable that is most correlated with both continuous and categorical variables. Similarly, $\boldsymbol{F}_2$ will be the synthetic variable orthogonal to $\boldsymbol{F}_1$ maximising Eq. (5). Once the dimensionality reduction step has been implemented, the final step consists of applying K-Means clustering on the lower-dimensional representation that has been obtained. This two-step procedure is herein referred to as 'FAMD/K-Means'. FAMD is conducted using the function FAMD() of the R package `FactoMineR` and K-Means using the base R function kmeans().

While FAMD followed by K-Means and generally tandem analysis seems like a reasonable approach to the clustering problem, De Soete and Carroll (1994) raise the point that variables with little contribution to the cluster structure can potentially 'mask' this structure, thus leading to unreliable results. This problem of 'cluster masking' is described in more detail by Vichi et al. (2019), who provide an illustration via a toy example. The idea of performing dimensionality reduction via PCA and K-Means clustering simultaneously, known as Reduced K-Means, was introduced in De Soete and Carroll (1994) as a potential solution to the cluster masking problem, with van de Velden et al. (2017) giving a concise description of a similar algorithm suitable for categorical data. Vichi et al. (2019) generalized Reduced K-Means algorithm in the case of mixed-type data.

This joint dimensionality reduction and clustering technique is referred to as Mixed Reduced K-Means, where 'Mixed' indicates the presence of mixed-type data (van de Velden et al. 2019). Its objective function is given by:

$$\phi_{RKM}\left(\boldsymbol{B}, \boldsymbol{Z}_K, \boldsymbol{G}\right) = \left\|\boldsymbol{X} - \boldsymbol{Z}_K \boldsymbol{G} \boldsymbol{B}^{\mathsf{T}}\right\|_F^2, \tag{6}$$

with $\boldsymbol{X}, \boldsymbol{Z}_K, \boldsymbol{G}, \boldsymbol{B}$ indicating the data matrix is centered and standardised in the exact same way as described for FAMD, the $(n \times K)$-dimensional cluster membership matrix, the $(K \times d)$-dimensional matrix of cluster centroids in the reduced d-dimensional space and a $(p^* \times d)$-dimensional columnwise orthonormal loadings matrix respectively. We use $\|\cdot\|_F$ to refer to the Frobenius norm and Eq. (6) is minimised via an Alternating Least Squares algorithm. In fact, it can be shown that there exists a certain expression for $\boldsymbol{G}$ that minimises (6), from which one can derive an expression for ϕ_{RKM} that only depends on $\boldsymbol{Z}_K$ and $\boldsymbol{B}$. The ALS algorithm used will first update the loadings matrix $\boldsymbol{B}$ while keeping $\boldsymbol{Z}_K$ fixed and this corresponds to a dimensionality reduction step. Once $\boldsymbol{B}$ has been updated, it is kept fixed and $\boldsymbol{Z}_K$ is updated accordingly, which can be seen as a K-Means problem. This also explains the intuition behind this algorithm performing joint dimensionality reduction and clustering. The choice of the number of dimensions retained, namely d, is set to be equal to $K-1$, where K is the number of clusters. This follows from the recommendation of Vichi and Kiers (2001), who argue that keeping more than $K-1$ dimensions is wasteful in joint dimensionality reduction and clustering algorithms, as this corresponds to describing a low-dimensional configuration of centroids in more dimensions than necessary. We have used the same number of dimensions in FAMD as in Mixed RKM for consistency. Notice that while Mixed RKM, as well as Mixed Factorial K-Means, which we have not implemented in our study, seem like reasonable methods for one to implement, Yamamoto and Hwang (2014) warn that these joint dimensionality reduction and clustering techniques are prone to giving inaccurate results if there exist variables irrelevant to the cluster structure in the data, which also happen to have high correlations between each other. Reduced K-Means was conducted in the current study using the function cluspca() of the R package `clustrd` (Markos et al. 2019).

The final method considered in the study is the KAMILA (KAy-means for MIxed LArge data) algorithm, that was introduced in Foss et al. (2016). The method attempts to cluster mixed-type data while balancing the level of contribution of continuous and categorical variables in a flexible way, such that no strong parametric assumptions are made. KAMILA can be seen as a combination of the K-Means algorithm with the Gaussian-Multinomial mixture model that is commonly used in model-based clustering (Hunt and Jorgensen 2011). In fact, the crucial assumption of Gaussianity of the continuous variables $\boldsymbol{X}_{con} = \left(\boldsymbol{X}_{con_1}, \ldots, \boldsymbol{X}_{con_{p_r}}\right)^\intercal$ is relaxed in KAMILA by considering the following more general probability density function for spherically symmetric distributions centered at the origin:

$$f_{\boldsymbol{X}_{con}}(\boldsymbol{x}_{con}) = \frac{f_R(r)\Gamma\left(\frac{p_r}{2}+1\right)}{p_r r^{p_r-1} \pi^{\frac{p_r}{2}}}. \tag{7}$$

Equation (7) requires the evaluation of the density of pairwise distances between continuous variables $r = \sqrt{\boldsymbol{x}_{con}^\intercal \boldsymbol{x}_{con}}$, which is replaced by a univariate kernel density estimate that uses the Radial Basis Function (RBF) kernel. The assignment of data points in the clusters is made according to the sum of the estimated $\hat{f}_{\boldsymbol{X}_{con}}$ log likelihood value at the Euclidean distance of each data point to each cluster centroid and of the log probability of observing the ith categorical vector given population membership. This quantity needs to be calculated for each cluster separately, with the cluster maximising the quantity being the one that the data point will be assigned to. Notice that this formulation is valid under the assumption of independence of the $p - p_r$ categorical variables within each cluster. The iterative scheme of KAMILA updates the cluster centroids and the parameters of the assumed underlying multinomial and spherically symmetric distributions until these remain unchanged, thus yielding the same partitions. Although KAMILA cannot be strictly considered as a distance-based partitioning approach, it was included in the comparison as a model-based alternative of K-Means. Moreover, in a series of studies it was consistently found to outperform model-based and distance-based methods. A more detailed mathematical description of the algorithm can be found in Foss et al. (2016), while the R implementation of KAMILA is available in the `kamila` package (function kamila()) (Foss and Markatou 2018).

4 Simulation study

A simulation study was conducted to evaluate the performance of the eight clustering methods presented in the previous Section in terms of cluster recovery. The data were generated using the function MixSim() of the R package `MixSim` (Melnykov et al. 2012), which allows for the determination of pairwise overlap between any pair of clusters. The notion of pairwise overlap is defined as the sum of two misclassification probabilities for a pair of weighted Gaussian distributions (Melnykov and Maitra 2010). More precisely, if $\boldsymbol{X}$ is a random variable originating from cluster l', the probability that it is misclassified to be originating from the lth cluster is given by $\omega_{l|l'} = \mathbb{P}_{\boldsymbol{X}} \left(\pi_{l'} \phi \left(\boldsymbol{X}; \boldsymbol{\mu}_{l'}, \boldsymbol{\Sigma}_{l'} \right) < \pi_l \phi \left(\boldsymbol{X}; \boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l \right) | \boldsymbol{X} \sim \mathcal{N}_p \left(\boldsymbol{\mu}_{l'}, \boldsymbol{\Sigma}_{l'} \right) \right)$. Defining $\omega_{l'|l}$ analogously, the overlap between the two clusters is given by $\omega_{ll'} = \omega_{l|l'} + \omega_{l'|l}$. Notice that $\boldsymbol{\mu}_l$, $\boldsymbol{\mu}_{l'}$, $\boldsymbol{\Sigma}_l$ and $\boldsymbol{\Sigma}_{l'}$ denote the mean vectors and the covariance matrices of the lth and the l'th components. However, this notion is only determined for continuous variables. In order to generate a categorical variable, a continuous variable was discretized by dividing it into c classes with the 100/c% quantile as the cut point. For simplicity, we will be assuming an equal number of categorical levels for all categorical variables (set equal to 4).

Seven factors that are typical of data sets collected in real-world scenarios and commonly encountered in benchmarking studies of clustering, were systematically manipulated for data generation. The first factor, the number of clusters in the data set, was examined at three levels, $K = 3$, 5, and 8. The second factor, the number of observations, was evaluated at three levels, $n = 100$, 600 and 1000, corresponding to a small, moderately large and large sample size in the social and behavioral sciences. The third factor, number of variables, was tested at three levels, $p = 8$, 12 and 16. The fourth factor, overlap of clusters, assumed values of 0.1%, 0.5%, 1.0%, 1.5% and 2.0%, corresponding to very small, small, moderate, high and very high overlap (specified via the argument `BarOmega` of the function MixSim()). These values correspond to smaller overlap, similar overlap, and much more overlap than in real data sets typically used to demonstrate clustering algorithms (see Shireman et al. (2016), for a justification and discussion of cluster overlap in real data sets compared to simulated data sets).

The fifth factor, percentage of categorical variables in the data (versus continuous variables), was tested at three levels, 20%, 50% and 80%. The sixth factor, density of the clusters, was tested at two levels: (a) an equal number of observations in each cluster and (b) 10% of the observations in one cluster and the remaining observations equally divided across the remaining clusters. The seventh factor considered is cluster sphericity, defined by the covariance matrix structure with two levels: (a) mixtures of heteroscedastic spherical components, that is spherical covariance matrices with nonhomogeneous variances among the mixture components (arguments `sph` and `hom` of the function MixSim() were set to `TRUE` and `FALSE`, respectively) and (b) mixtures of heteroscedastic non-spherical components, that is ellipsoidal covariance matrices with nonhomogeneous variances among the mixture components (arguments `sph` and `hom` were both set to `FALSE`). This resulted in $3 \times 3 \times 3 \times 5 \times 3 \times 2 \times 2 = 1620$ distinct data scenarios. Fifty replications were made for each scenario, resulting in a total of 81000 data sets. Each clustering procedure was fit 100 times using random starting values. For each data set, the number of clusters was always correctly specified. For FAMD and Mixed Reduced K-Means the number of dimensions (factors) was set to the number of clusters minus one. The ability of each procedure to return the true cluster structure was measured by the Adjusted Rand Index (ARI; Hubert and Arabie 1985) and the Adjusted Mutual Information (AMI; Vinh et al. 2010). The ARI measures the agreement between two different partitions of the same set of observations, by looking at pairs of observations in the original data set and counting and comparing how many pairs were assigned to the same cluster in both partitions, and how many pairs were not assigned to the same clusters in both partitions. The maximum value of the ARI is 1 and its expected value in the case of random partitions is 0. Steinley (2004) has provided some guidelines for interpreting ARI values in simulation experiments, with thresholds of .90, .80, and .65 corresponding to excellent, good, and fair cluster recovery, respectively. Values of the ARI below .65 reflect poor recovery. The AMI is an information-theoretic index that measures the amount of “shared information" between two clusterings and is expected to be less susceptible to cluster size imbalance than ARI (Van der Hoef and Warrens 2019). Both ARI and AMI measure the similarity between ground truth class assignments and those of the clustering method, adjusted for chance groupings. The simulated data sets, the resulting ARI and AMI values and the R code used for analyses are publicly available in an OSF repository at https://rb.gy/rgpdyu.

5 Results

Table 1 reports average cluster recovery of the eight methods across all factors. The best performing methods, on average, are KAMILA, FAMD/K-Means and K-Prototypes, followed by Modha-Spangler K-Means and Mixed Reduced K-Means. The worst performing methods were HL/PAM, Mixed K-Means and Gower/PAM. Table 1 also presents the degree of agreement in cluster recovery between methods in terms of ARI/AMI, based on Pearson’s correlation. Some pairwise correlations are large enough, the largest being Cor(FAMD/K-Means, Mixed Reduced K-Means)

Table 1 Agreement between methods based on Pearson's correlation and mean cluster recovery (ARI/AMI values) in the analysis of simulated data sets

Method ARI/AMI	(1)	(2)	(3)	(4)	(5)	(6)	(7)	Mean
(1) KAMILA	-							.366/.404
(2) FAMD/K-Means	.90/.93	-						.340/.381
(3) K-Prototypes	.93/.95	.90/.92	-					.336/.377
(4) M-S K-Means	.90/.91	.89/.91	.88/.91	-				.313/.357
(5) Mixed RKM	.87/.90	.94/.96	.88/.89	.92/.93	-			.309/.357
(6) HL/PAM	.79/.82	.75/.78	.81/.86	.78/.82	.76/.79	-		.193/.246
(7) Mixed K-Means	.78/.82	.70/.74	.81/.86	.75/.80	.69/.74	.81/.88	-	.191/.246
(8) Gower/PAM	.73/.72	.75/.75	.73/.74	.74/.75	.77/.77	.65/.71	.77/.82	.136/.182

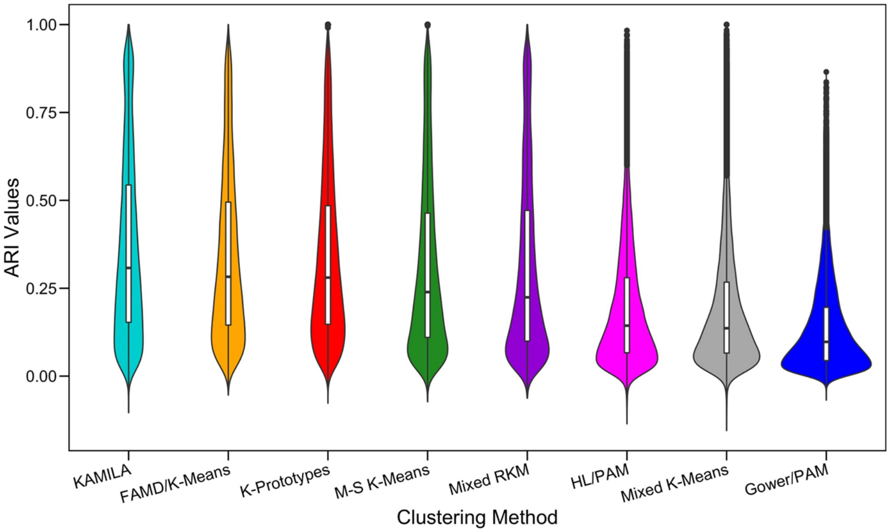

Fig. 1 Violin/box plots of Adjusted Rand Index values by method

= .94/.96; that is both FAMD/K-Means and Mixed Reduced K-Means account for 88%/92% of the variance in the other.

The violin/box plots in Figs. 1 and 2 show the corresponding distributions of ARI and AMI values, respectively, for the eight methods computed on the true clusterings, confirming that KAMILA, FAMD/K-Means and K-Prototypes perform somewhat better than Modha-Spangler K-Means and Mixed Reduced K-Means, and much better than HL/PAM, Mixed K-Means and Gower/PAM. For instance, more than 50% of the ARI values for HL/PAM, Mixed K-Means and Gower/PAM are less than .14. Also notice that more than 75% of the ARI values for Gower/PAM are less than .20, indicating poor cluster recovery.

After examining the overall performance of the methods, it is informative to determine if method performance is dependent upon specific situations (i.e., performance

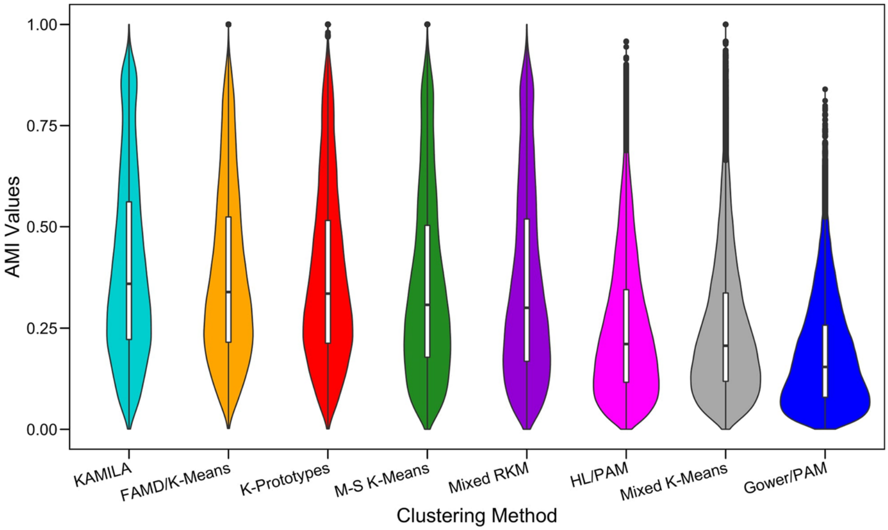

Fig. 2 Violin/box plots of Adjusted Mutual Information values by method

varies with the factor levels). The individual performances are examined by the levels of each factor. For this purpose, two separate repeated-measures ANOVAs were conducted on ARI and AMI scores (see Table 2). All main effects and interactions were modeled. Given the large sample size, it was expected that most factors would be statistically significant; therefore, all effects were evaluated with respect to their estimated effect sizes, partial eta-squared (η^2). Main effects and interactions were presented and discussed further only if they reached at least a moderate effect size (partial $\eta^2 \geq .01$).

The between data sets effects in Table 2 can be thought of as the influence of the design factors across all clustering methods. Cluster overlap had the largest effect on cluster recovery. Overall, and as expected, as the overlap of clusters increased from .01 to .20, the average recovery in terms of both ARI/AMI decreased, going from .55/.59 to .11/.16. Cluster sphericity had also a large effect on cluster recovery with mean ARI/AMI values of .32/.36 for spherical and .23/.27 for non-spherical clusters. The number of clusters had a large and negative effect on cluster recovery based on ARI, with .33, .27 and .22, for 3, 5 and 8 clusters, respectively; the effect was negligible in the case of AMI (.32, .31, .32). As the number of variables increased, the clustering performance deteriorated, as indicated by ARI/AMI values of .30/.35, .27/.31 and .25/.29 for 8, 12 and 16 variables, respectively. The percentage of categorical variables in the data set had also a moderate and negative effect, with mean ARI/AMI values equal to .29/.34, .28/.33 and .24/.29, for 20%, 50% and 80%, respectively. The number of observations had a more profound effect in the case of ARI, with mean values of .25/.33, .28/.31 and .29/.31 for 100, 600 and 1000 observations, respectively. Last, cluster density had a small effect on cluster recovery, with mean ARI/AMI values equal to .26/.31 and .28/.33 for equally-sized clusters and a 10% of the observations in a single cluster, respectively.

Table 2 Repeated measures ANOVAs for eight clustering methods on ARI (top half) and AMI (bottom half). Factors are ordered by decreasing effect size, partial η^2

	Effect	Source	df	SS	F	Partial η^2
ARI	Between data sets effects	Overlap	4	16112.64	76435.11	.794
		Sphericity	1	1321.89	25083.02	.240
		# clusters	2	1089.54	10337.15	.207
		# vars	2	283.06	2685.57	.063
		% categorical	2	282.99	2684.87	.063
		# obs	2	185.60	1760.91	.042
		Density	1	68.60	1301.74	.016
	Within data sets effects (univariate tests)	Method (M)	7	4200.76	122234.65	.606
		M*overlap	28	1113.81	8102.46	.290
		M*categorical	14	413.61	6017.59	.132
		M*clusters	14	328.71	4782.51	.108
		M*obs	14	157.04	2284.78	.054
		M*sphericity	7	149.53	4350.96	.052
		M*vars	14	141.70	2061.73	.049
		M*overlap*categorical	56	92.58	336.75	.033
		M*overlap*density	28	82.16	597.71	.029
		M*overlap*vars	56	73.77	268.31	.026
		M*density	7	34.48	1555.65	.019
AMI	Between data sets effects	Overlap	4	15745.59	103229.18	.839
		Sphericity	1	1217.15	31918.95	.287
		# vars	2	412.21	5405.00	.120
		% categorical	2	291.69	3824.67	.088
		# obs	2	65.27	855.89	.021
		Density	1	61.67	1617.29	.020
		# clusters	2	9.20	120.58	.003
	Within data sets effects (univariate tests)	Method (M)	7	3778.28	170484.85	.682
		M*overlap	28	885.97	9994.29	.335
		M*categorical	14	345.71	7799.54	.164
		M*clusters	14	220.68	4978.78	.111
		M*obs	14	145.00	3271.27	.076
		M*vars	14	116.96	2638.72	.062
		M*sphericity	7	91.60	4133.20	.049
		M*overlap*categorical	56	71.71	404.48	.039
		M*overlap*vars	56	60.9	343.37	.033
		M*overlap*density	28	45.48	513.09	.025
		M*density	7	34.48	1555.65	.019

Based on the within data sets effects (Table 2) we determine which methods are effective under which conditions. We start by considering two-way interactions first. Then we discuss three-way interactions. In the presence of a significant interaction, main effects and lower order effects were ignored.

Table 3 shows the ARI/AMI values of the eight methods by all factors. In general, both measures yield similar results. The two-way interaction between method and number of clusters shows that the number of clusters negatively affects the performance of all methods, but the effect is more profound for Mixed Reduced K-Means, when the number of clusters is other than 3 (Fig. 6 and Table 3). The two-way interaction between method and number of observations reveals that K-Prototypes performs slightly better than other methods for the small sample-size scenario, $n = 100$ (Fig. 6 and Table 3). KAMILA's performance greatly improves for $n = 600$. The performance of HL/PAM, Mixed K-Means and Gower/PAM does not appear to be significantly affected by n, but their performance remains poor compared to other methods. Non-sphericity of the clusters seems to affect all methods but not in a uniform manner (Fig. 6 and Table 3). The difference in performance between KAMILA and other methods is less profound when clusters are non-spherical. This is not surprising, since in KAMILA continuous variables are assumed to follow a mixture distribution with arbitrary spherical clusters.

The heat maps in Fig. 3 visualize the three-way interaction of method by cluster overlap and percentage of categorical variables (mean ARI values). In the presence of categorical variables, there are differences between methods. KAMILA and K-Prototypes outperform other methods when the percentage of categorical variables is low (20%), whereas FAMD/K-Means and Mixed Reduced K-Means perform best when the percentage of categorical variables is high (80%). The performance of Mixed Reduced K-Means deteriorates at a faster rate than other methods with increasing overlap. Modha-Spangler K-Means performs best when the number of categorical and continuous variables in the data set is equal (50%). The interaction of method by cluster overlap and density is illustrated in Fig. 4. Some methods are affected more than others by the presence of a small-size cluster and at different levels of overlap. When cluster overlap is very low (.01) all methods perform better in the case of clusters with equal size, but for higher levels of overlap ($>$ .01) there is negligible difference in cluster recovery or cluster recovery is slightly better in the small-size cluster scenario. The interaction of method by cluster overlap and the number of variables (Fig. 5) reveals that going from 8 to 16 variables, deteriorates the performance of KAMILA, K-Prototypes, HL/PAM, Mixed K-Means and Gower/PAM, whereas cluster recovery of Mixed Reduced K-Means and FAMD/K-Means is improved. This improvement, however, is observed only when cluster overlap is low ($\leq$ .05).

Last, it is worth underlining that in terms of absolute performance, the mean ARI/AMI values in Table 3 and Figs. 3 to 6 suggest poor cluster recovery in the vast majority of cases (ARI/AMI values below .65) for all methods under comparison. Cluster recovery was found to be fair, albeit not good or excellent, for certain methods and conditions only (values above .65 but less than .80). In particular, for three out of eight methods (Mixed K-Means, HL/PAM, Gower/PAM) average cluster recovery is poor across all factors, even for well-separated and equally sized clusters. For the top-performing methods, there are cases when cluster recovery can be considered fair

Table 3 Cluster recovery (ARI/AMI) of eight clustering methods by cluster overlap, cluster sphericity, number of clusters, percentage of categorical variables, number of variables, cluster density and number of observations

Factor	Level	KAMILA	FAMD /KM	K-Proto types	M-S KM	Mixed RKM	HL /PAM	Mixed KM	Gower /PAM
Overlap	.01	.69/.72	.65/.68	.64/.67	.64/.68	.64/.70	.41/.47	.39/.46	.29/.34
	.05	.45/.49	.42/.46	.41/.46	.39/.43	.39/.44	.23/.29	.23/.29	.16/.21
	.10	.31/.35	.28/.33	.28/.32	.24/.29	.24/.29	.15/.20	.15/.21	.10/.15
	.15	.22/.26	.20/.25	.20/.24	.17/.21	.15/.20	.10/.15	.11/.16	.07/.12
	.20	.16/.20	.15/.19	.15/.19	.12/.17	.10/.15	.07/.12	.08/.13	.06/.10
Sphericity	No	.30/.35	.28/.33	.29/.33	.26/.31	.17/.31	.17/.22	.17/.22	.11/.15
	Yes	.43/.46	.40/.44	.39/.43	.37/.41	.22/.41	.22/.28	.22/.27	.17/.21
# clusters	3	.42/.41	.44/.42	.39/.38	.38/.37	.40/.39	.21/.22	.21/.22	.15/.16
	5	.37/.40	.34/.38	.34/.38	.30/.35	.29/.34	.19/.25	.19/.25	.14/.18
	8	.31/.40	.25/.35	.28/.38	.26/.36	.24/.35	.17/.28	.17/.28	.12/.21
% categorical	20%	.42/.45	.35/.39	.41/.44	.30/.35	.30/.35	.21/.27	.24/.29	.12/.17
	50%	.38/.41	.34/.38	.34/.38	.35/.39	.31/.36	.19/.25	.20/.26	.13/.18
	80%	.30/.35	.33/.37	.26/.31	.28/.32	.32/.32	.18/.36	.14/.22	.15/.20
# vars	8	.40/.44	.34/.39	.36/.41	.33/.38	.31/.37	.25/.31	.24/.30	.17/.22
	12	.36/.40	.34/.38	.33/.37	.30/.35	.31/.36	.19/.24	.18/.24	.13/.18
	16	.34/.37	.33/.37	.31/.35	.30/.34	.31/.35	.14/.20	.15/.20	.11/.15
Density	10%	.35/.39	.32/.37	.31/.36	.30/.34	.29/.34	.20/.25	.19/.25	.14/.18
	Equal	.38/.42	.36/.40	.36/.39	.33/.37	.33/.37	.19/.25	.19/.25	.14/.18
# obs	100	.30/.37	.29/.37	.31/.39	.29/.37	.28/.37	.20/.29	.18/.27	.14/.22
	600	.39/.41	.36/.38	.35/.37	.32/.35	.32/.35	.19/.23	.19/.23	.14/.16
	1000	.40/.42	.37/.39	.35/.37	.32/.35	.32/.35	.19/.23	.20/.23	.13/.16

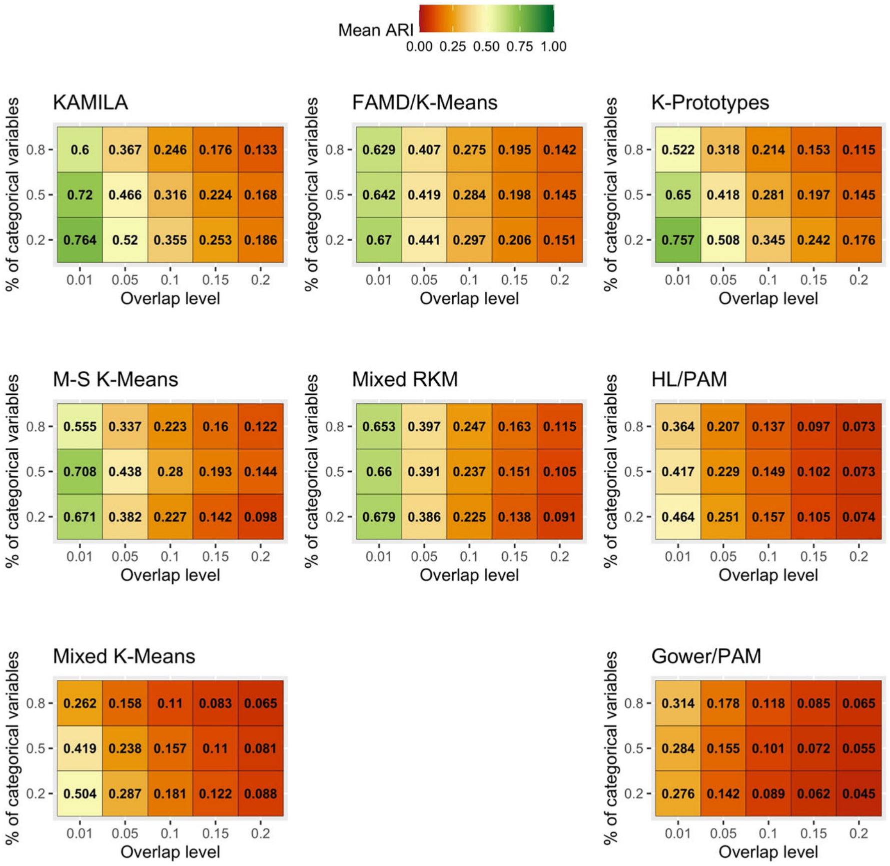

Fig. 3 Three-way interaction of method by overlap level and percentage of categorical variables. The numbers indicate the mean ARI for each combination of overlap level and percentage of categorical variables

to good, especially when clusters are well-separated, equally sized and the percentage of categorical variables is low or moderate.

6 Discussion

This paper reports benchmark test results from applying distance-based partitioning methods on simulated data sets with different characteristics. Eight methods were selected to cover three general strategies of distance or dissimilarity-based partitioning of mixed-type data (i.e., constructing a dissimilarity matrix between observations given as input to K-Medoids, extending K-Means to mixed-type data and reducing the number of variables and clustering of the observations in the reduced space).

One essential goal of the benchmark is to make the results available and reusable to other researchers. Benchmark results revealed both similarities and differences in the overall performance of the eight algorithms, as well as across different criteria. A group of top-performing methods with similar performance can be distinguished,

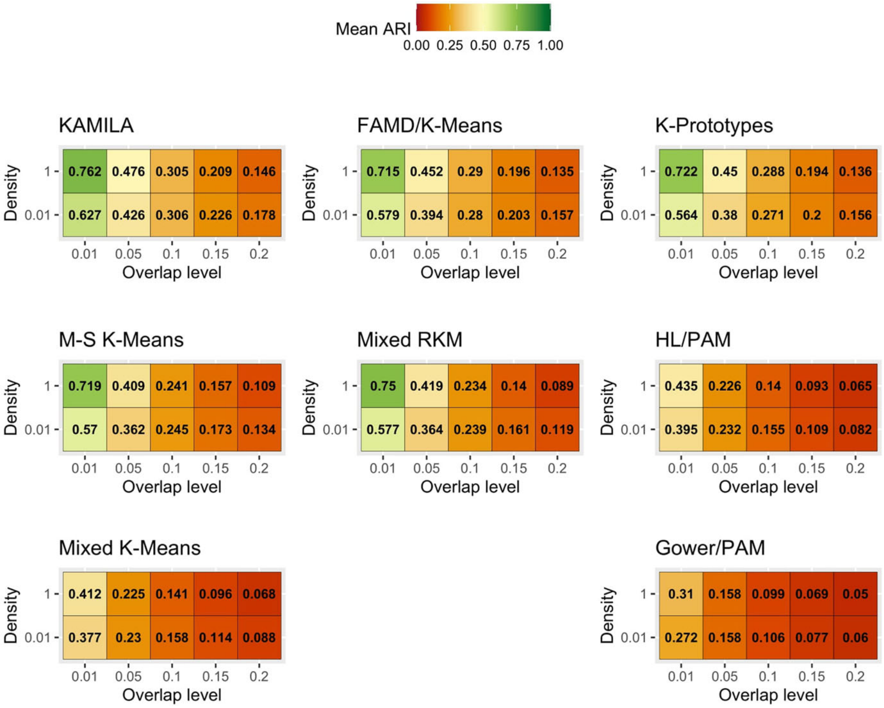

Fig. 4 Three-way interaction of method by overlap level and cluster density. The numbers indicate the mean ARI for each combination of overlap level and cluster density

consisting of KAMILA, FAMD/K-Means and K-Prototypes. KAMILA was the best method in about half of the 1620 different data scenarios (based on both ARI and AMI). These are mostly data sets with moderate or large sample size and more continuous than categorical variables. The deterioration of KAMILA's performance in the small sample size scenario was somewhat expected, since the method's reliance on a multinomial model for categorical variables requires a commensurate sample size, as has been previously indicated in Foss et al. (2019). Therefore, we recommend the use of KAMILA when the sample size is reasonably large and the categorical variables are not dominant in the data set. FAMD/K-Means was the top-performing approach in about one fifth of the cases. This method performed well for data sets with moderate or large sample size and more categorical than continuous variables. In contrast to the other two methods, FAMD/K-Means additionally involves a dimensionality reduction step, which can be convenient for visualizing and interpreting the clusters in the reduced space. This means that it depends heavily on the amenability of the data set to dimensionality reduction where a few principal components account for a high percentage of variability in the data set. Where this is not the case, the FAMD/K-Means method cannot be reasonably applied. K-Prototypes was the best approach in about 13% of the distinct scenarios and is recommended in cases when the sample size is relatively small and there are more continuous than categorical variables.

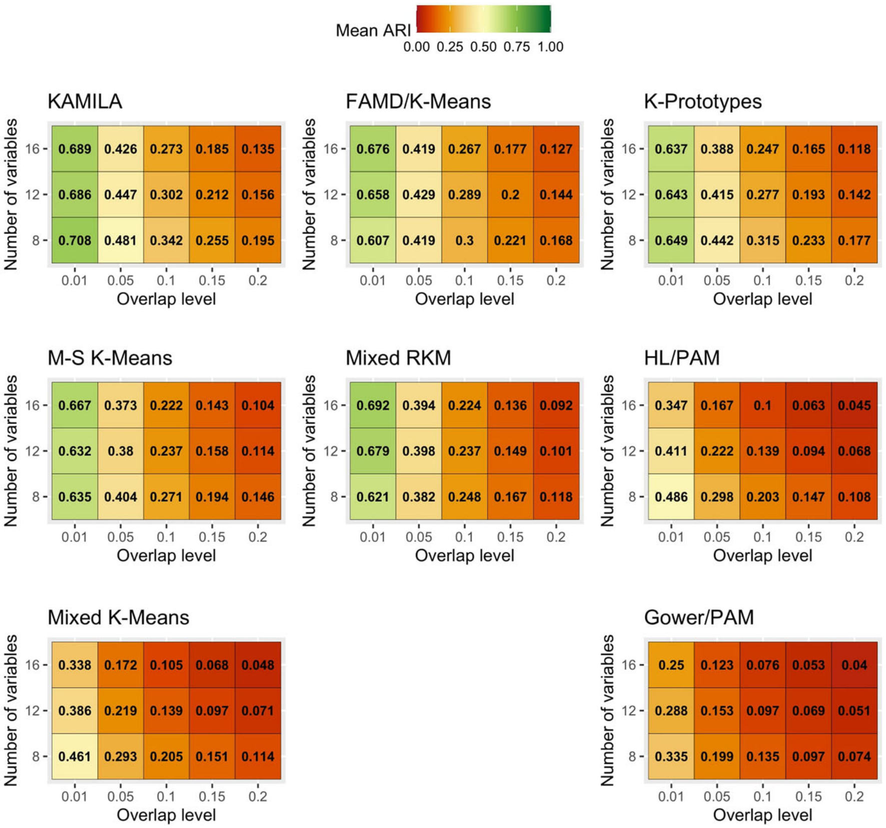

Fig. 5 Three-way interaction of method by overlap level and number of variables. The numbers indicate the mean ARI for each combination of overlap level and number of variables

Modha-Spangler K-Means and Mixed Reduced K-Means form a second group of methods with similar performance, not far from the first group in terms of cluster recovery. Modha-Spangler K-Means was the best method in 11% of the different scenarios, mainly when the number of continuous variables is equal or greater than that of categorical variables. This is also demonstrated in Foss et al. (2016), where Modha-Spangler K-Means was found to underperform relative to competing methods when there are more categorical than continuous variables because of its over-reliance on continuous variables. Mixed Reduced K-Means was the best approach in 7% of the cases, performing well for moderate or large samples sizes, more categorical than continuous variables and low levels of cluster overlap. Although Mixed Reduced K-Means was expected to improve upon FAMD/K-Means, in the sense that it was developed to address the cluster masking problem by optimizing a single objective function (Vichi et al. 2019), this hypothesis was not confirmed by the study results. Therefore, when dimensionality reduction is an additional goal, FAMD/K-Means seems to be a more reasonable choice than Mixed Reduced K-Means. However, both FAMD/K-Means and Mixed Reduced K-Means have in common that they perform worse for a sam-

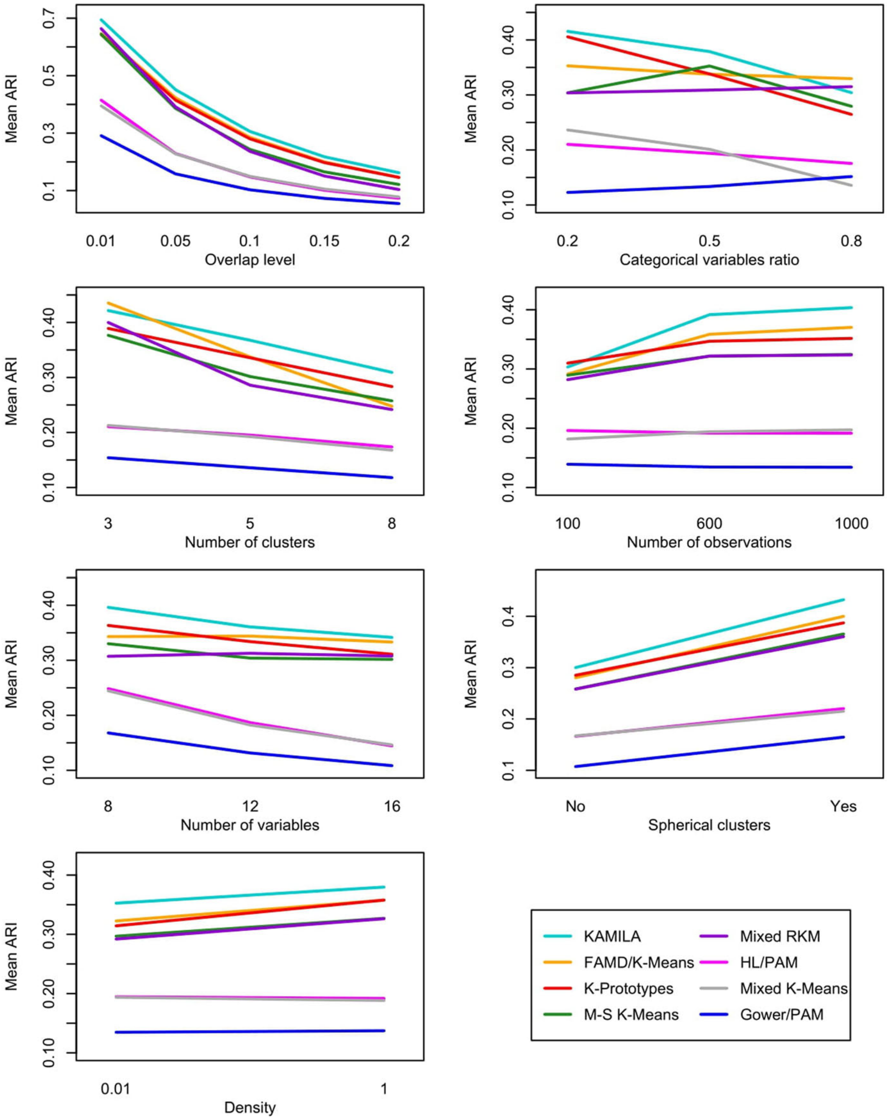

Fig. 6 Two-way interactions of method by overlap level, percentage of categorical variables, number of clusters, number of observations, number of variables, density, and cluster sphericity (mean ARI values). Subplots/factors are arranged, from left to right, by decreasing effect size, partial η^2

ple size of 100 compared to larger sample sizes; a sample size of 100 is usually not sufficient for PCA-based methods (see e.g., Saccenti and Timmerman 2016).

A third group of methods, clearly distinct from the other two in terms of cluster recovery, contains HL/PAM, Mixed K-Means and Gower/PAM. These methods demonstrated poor performance in our experiments, with Gower/PAM being the worst performing method across all criteria, even for well-separated clusters. This could be seen as a surprising finding, considering that Gower/PAM is among the most popular choices in the literature for clustering mixed-type data. A potential explanation for this could be that the task of clustering multivariate normal distributions, as the objective of the simulations conducted in the current study, can be expected to favour K-Means-like approaches that use a squared loss function; PAM-based approaches instead are known to be more robust against non-normality (Kaufman and Rousseeuw 1990, p.117). Also notice that HL/PAM performed better than Gower/PAM in all tested scenarios but did not reach the performance of K-Means-based methods.

There are some limitations with the current study. First, to generate mixed-type data for the simulations, continuous variables were generated by drawing from finite mixtures of multivariate normal distributions; categorical variables were generated via discretization of such continuous variables. Ideally, for our experiments we would need to generate purely mixed-type data, that is, purely categorical variables and purely continuous variables with a cluster structure. However, controlling the overlap between clusters in mixed-type data sets with more than two clusters is not straightforward (see, e.g., Maitra and Melnykov 2010). In addition, the covariance structure between the variables was not user-defined. Controlling the correlation structure between the variables could have been useful, so as to draw conclusions on how correlated variables affect the performance of clustering algorithms. Also, the simulations could be extended to mixtures of non-Gaussian distributions. Second, clustering performance depends on the software implementation used; different implementations of a method often lead to different results. 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. Furthermore, in a study by Steinley (2006), the K-Means algorithm with 100 random initializations has been shown to produce the same solution for well-separated clusters, but the algorithm produced different solutions in the case of overlapping clusters. The author recommended using several thousand random initializations. Based on this observation, the 100 random starts used in our study might not be sufficiently high for K-Means-based clustering methods to avoid local minima. However, from a computational viewpoint, this would result in a much higher and prohibitive computational cost. Another limitation is that for the methods under comparison, the number of clusters was always correctly specified, instead of incorporating cluster number estimation in the clustering task. Although this is a challenging and interesting endeavour, it requires additional decisions, such as the choice of the index/method to be used for cluster estimation and was beyond the scope of the current study. Last, the results of the current study were based on simulated data sets only. An empirical comparison of cluster analysis methods on real mixed-type data (see, e.g., Hennig 2022) is expected to further highlight 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.

Funding: Efthymios Costa is funded by EPSRC's StatML CDT.

References

Ahmad A, Dey L (2007) A k-mean clustering algorithm for mixed numeric and categorical data. Data & Knowledge Engineering 63(2):503–527

Ahmad A, Khan SS (2019) Survey of state-of-the-art mixed data clustering algorithms. IEEE Access 7:31,883–31,902

Arabie P (1994) Cluster analysis in marketing research. Blackwell, Oxford, pp 160–189

Audigier V, Husson F, Josse J (2016) A principal component method to impute missing values for mixed data. Adv Data Anal Classif 10(1):5–26

Boulesteix AL, Hatz M (2017) Benchmarking for clustering methods based on real data: A statistical view. In: Palumbo F, Montanari A, Vichi M (eds) Data Science. Springer International Publishing, Cham, pp 73–82

Boulesteix AL, Lauer S, Eugster MJ (2013) A plea for neutral comparison studies in computational sciences. PLoS ONE 8(e61):562

De Soete G, Carroll JD (1994) K-means clustering in a low-dimensional Euclidean space, Springer, 212–219

Dolnicar S, Grün B (2008) Challenging "factor-cluster segmentation". J Travel Res 47(1):63–71

Ferreira L, Hitchcock DB (2009) A comparison of hierarchical methods for clustering functional data. Communications in Statistics - Simulation and Computation 38(9):1925–1949

Foss A, Markatou M, Ray B et al (2016) A semiparametric method for clustering mixed data. Mach Learn 105(3):419–458

Foss AH, Markatou M (2018) kamila: Clustering mixed-type data in R and Hadoop. J Stat Softw 83:1–44

Foss AH, Markatou M, Ray B (2019) Distance metrics and clustering methods for mixed-type data. Int Stat Rev 87(1):80–109

Gower JC (1971) A general coefficient of similarity and some of its properties. Biometrics 27:857–871

Hennig C (2020) Package 'fpc'. URL https://cran.r-project.org/web/packages/fpc/fpc.pdf

Hennig C (2022) An empirical comparison and characterisation of nine popular clustering methods. Adv Data Anal Classif 16:201–229

Hennig C, Liao TF (2013) How to find an appropriate clustering for mixed-type variables with application to socio-economic stratification. J Roy Stat Soc: Ser C (Appl Stat) 62(3):309–369

Huang Z (1997) Clustering large data sets with mixed numeric and categorical values. In: Proceedings of the 1st Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD), Citeseer, 21–34

Hubert L, Arabie P (1985) Comparing partitions. J Classif 2(2):193–218

Hunt L, Jorgensen M (2011) Clustering mixed data. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 1(4):352–361

Javed A, Lee BS, Rizzo DM (2020) A benchmark study on time series clustering. Machine Learning with Applications 1(100):001

Jimeno J, Roy M, Tortora C (2021) Clustering mixed-type data: A benchmark study on KAMILA and K-Prototypes. In: Chadjipadelis T, Lausen B, Markos A et al (eds) Data Analysis and Rationality in a Complex World. Springer International Publishing, Cham, pp 83–91

Kaufman L, Rousseeuw PJ (1990) Finding Groups in Data: An Introduction to Cluster Analysis. John Wiley & Sons, chap 2:68–125

Kiers HA (1991) Simple structure in component analysis techniques for mixtures of qualitative and quantitative variables. Psychometrika 56(2):197–212

Maechler M, Rousseeuw P, Struyf A et al (2021) cluster: Cluster Analysis Basics and Extensions. URL https://CRAN.R-project.org/package=cluster, R package version 2.1.2)

Maitra R, Melnykov V (2010) Simulating data to study performance of finite mixture modeling and clustering algorithms. J Comput Graph Stat 19(2):354–376

Markos A, Iodice D'Enza A, van de Velden M (2019) Beyond tandem analysis: Joint dimension reduction and clustering in R. J Stat Softw 91:1–24

Markos A, Moschidis O, Chadjipantelis T (2020) Sequential dimension reduction and clustering of mixed-type data. International Journal of Data Analysis Techniques and Strategies 12(3):228–246

Meilă M, Heckerman D (2001) An experimental comparison of model-based clustering methods. Mach Learn 42:9–29

Melnykov V, Maitra R (2010) Finite mixture models and model-based clustering. Statistics Surveys 4:80–116

Melnykov V, Chen WC, Maitra R (2012) MixSim: An R package for simulating data to study performance of clustering algorithms. J Stat Softw 51(12):1–25

Milligan GW (1980) An examination of the effect of six types of error perturbation on fifteen clustering algorithms. Psychometrika 45:325–342

Modha DS, Spangler WS (2003) Feature weighting in k-means clustering. Mach Learn 52(3):217–237

Murtagh F (2015) A Brief History of Cluster Analysis. In: Hennig C, Meila M, Murtagh F et al (eds) Handbook of Cluster Analysis. Chapman & Hall/CRC, 21–33

Pagès J (2014) Multiple Factor Analysis By Example Using R. Chapman and Hall/CRC, chap 3:67–78

Preud'Homme G, Duarte K, Dalleau K et al (2021) Head-to-head comparison of clustering methods for heterogeneous data: a simulation-driven benchmark. Sci Rep 11(1):1–14

Saccenti E, Timmerman ME (2016) Approaches to sample size determination for multivariate data: Applications to PCA and PLS-DA of omics data. J Proteome Res 15(8):2379–2393

Saraçli S, Doğan N, Doğan İsmet (2013) Comparison of hierarchical cluster analysis methods by cophenetic correlation. Journal of Inequalities And Applications 2013:1–8

Shireman EM, Steinley D, Brusco MJ (2016) Local optima in mixture modeling. Multivar Behav Res 51(4):466–481

Steinley D (2004) Properties of the Hubert-Arabie Adjusted Rand Index. Psychol Methods 9(3):386–396

Steinley D (2006) Profiling local optima in k-means clustering: developing a diagnostic technique. Psychol Methods 11(2):178–192

Szepannek G (2018) clustMixType: User-Friendly Clustering of Mixed-Type Data in R. The R Journal 10(2):200–208

Van der Hoef H, Warrens MJ (2019) Understanding information theoretic measures for comparing clusterings. Behaviormetrika 46:353–370

Van Mechelen I, Boulesteix AL, Dang R et al (2018) Benchmarking in cluster analysis: A white paper arxiv:1809.10496v2

van de Velden M, Iodice D'Enza A, Palumbo F (2017) Cluster correspondence analysis. Psychometrika 82(1):158–185

van de Velden M, Iodice D'Enza A, Markos A (2019) Distance-based clustering of mixed data. Wiley Interdisciplinary Reviews: Computational Statistics 11(3):e1456

Vichi M, Kiers HA (2001) Factorial k-means analysis for two-way data. Computational Statistics & Data Analysis 37(1):49–64

Vichi M, Vicari D, Kiers HA (2019) Clustering and dimension reduction for mixed variables. Behaviormetrika 46(2):243–269

Vinh NX, Epps J, Bailey J (2010) Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance. The Journal of Machine Learning Research 11:2837–2854

Yamamoto M, Hwang H (2014) A general formulation of cluster analysis with dimension reduction and subspace separation. Behaviormetrika 41(1):115–129